BIOE 505:
Computational Bioengineering

What this class is all about?

Instructor

- Name: Sergei Maslov

- Professor of Bioengineering, Physics, Carl R. Woese Institute for Genomic Biology, and National Center for Supercomputing Applications

- Office: 3103 Carl Woese Institute for Genomic Biology and sometimes 3146C Everitt Laboratory (both by appointment)

- E-mail: maslov@illinois.edu

- Phone: 217-265-5705

Questions and Suggestions:

maslov@Illinois.edu

Start subject with [BIOE505]

Grading

- Midterm exam 40%
- Final exam 60%
- Homework (ungraded) will be posted online. Solutions will be posted in a week.
- Homework will build on topics covered in lectures and will consist of problem sets related to topics covered in lectures
- Useful to prepare for exams

Course Website

https://courses.engr.illinois.edu/bioe505

Grades will be on

https://my.bioen.illinois.edu/gradebook

courses.grainger.illinois.edu/bioe505/fa2019/index.html

Apps Recent Group Library@RC Ludwig • Find your... WebEdit BIOE

BIOE 505 - Compu

Schedule

Instructor

Sergei Maslov: maslov@illinois.edu
Office: IGB 3406
Office hours: by appointment

Logistics

Tuesdays: 12:00AM - 1:50AM
Thursdays: 12:00AM - 1:50PM

106B8 Engineering Hall

I WANT YOU TO BRING YOUR i-CL

I WANT YOU TO BRING YOUR OWN
LAPTOPS SHOULD HAVE MATLAB
STATISTICS AND MACHINE LEARN

Description

courses.grainger.illinois.edu/bioe505/fa2019/schedule.html

Apps Recent Group Library@RC Ludwig • Find your... WebEdit BIOE 505 | Prof. Ser... BIOE 505 | Course E...

BIOE 505 - Computational Bioengineering

#	Date	Topics	Slides	Matlab	Homework	Exams
1	Aug 27					
2	Aug 29					

Bring your iClickers to my lectures

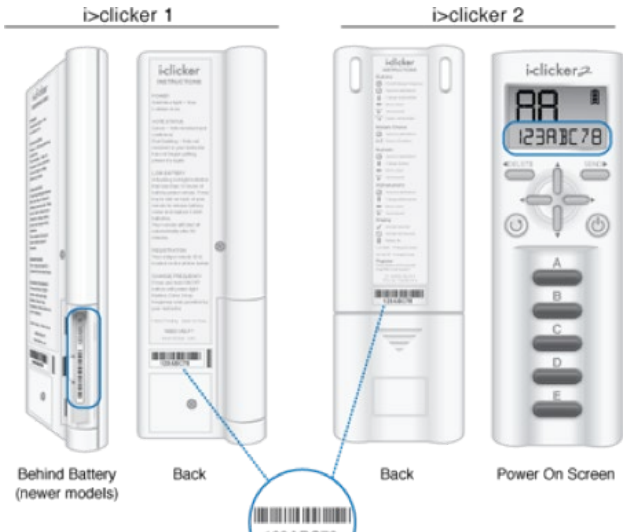
- **Who knows what is an iClicker?**

- **Show of hands: who has an iClicker?**

- **I would like you all to have an iClicker and bring it to every class. On amazon.com a new iClicker (1st generation is OK) costs around $40. It is also sold at UIUC Bookstore. The used ones are cheaper.**

- **An alternative solution is using a mobile app: https://www.iclicker.com/students/apps-and-remotes/apps**

- **Your answers WILL NOT be used for grading.**
  I need them to see if I lost some of you and what could I rephrase to better explain the material



Who has Matlab?

A. Already have it installed on my laptop
B. Will install it (starting this year **it is free**!)
C. Plan to access it on EWS via CITRIX
D. I don’t know yet
E. I will never use Matlab!

Why don’t we use Python?

Get your i-clickers

We will use Matlab in class

- Bring your laptops to class
- Need to have Matlab installed and know the basic user interface (inline commands, plotting)
- We will use Statistics and Machine Learning Toolbox and Bioinformatics Toolboxes
- Good news! Now all faculty and graduate students get Matlab **for free.** See offering on the WebStore site and follow the detailed instructions.
- .m files and .mat with Matlab commands and data will be on the website after the lecture

Possible alternative to purchasing Matlab and toolboxes is to use campus resources.

Both Engineering Workstations (EWS) and ACES computers have Matlab.
I don't think all of them offer the statistics and bioinformatics toolboxes
(EWS should, ACES computers may not..).

See the following to access:

Citrix for EWS, Matlab, and ACES computers -- links for all

https://it.engineering.illinois.edu/ews/lab-information/remote-connections/connecting-citrix
https://it.engineering.illinois.edu/services/instructional-services/remote-connections-citrix

Accessing Engineering Workstations (EWS)

https://it.engineering.illinois.edu/ews

Accessing ACES Academic Computing Workstations

http://acf.aces.illinois.edu/remote/
http://acf.aces.illinois.edu/remote/pc.html

To access off campus use:

CISCO Virtual Private Network -- For off-campus access to campus computer and network resources (software programs, files saved on the network, etc.)

https://techservices.illinois.edu/services/virtual-private-networking-vpn/download-and-set-up-the-vpn-client

CISCO VPN CLIENT

https://webstore.illinois.edu/shop/product.aspx?zpid=2600

CISCO AnyConnect VPN

https://webstore.illinois.edu/shop/product.aspx?zpid=1222

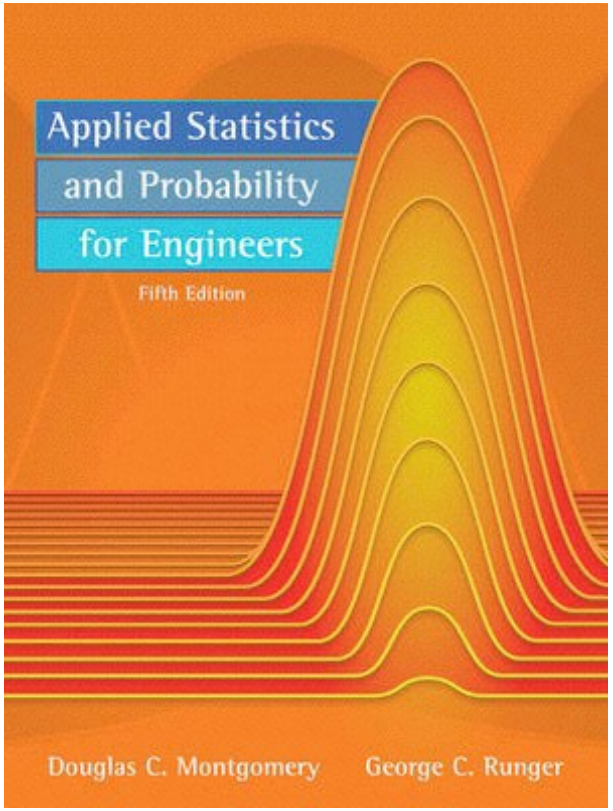
What will you learn in this course?

- Basics of probability and statistics
  - Basic concepts of probability, Bayes theorem
  - Discrete and continuous probability distributions
  - Multivariate statistics
  - Sampling distributions
  - Parameter estimation
  - Hypothesis testing
  - Regression
- How it is applied to biological data
  - Basics of genomics
  - Systems biology (gene expression, networks)

The main Probability/Statistics Textbook

Applied Statistics and Probability for Engineers, 5th Edition
D. C. Montgomery and G. C. Runger
John Wiley & Sons, Inc. (2011)

You can also use other editions from 4th (2007) to 6th (2014)



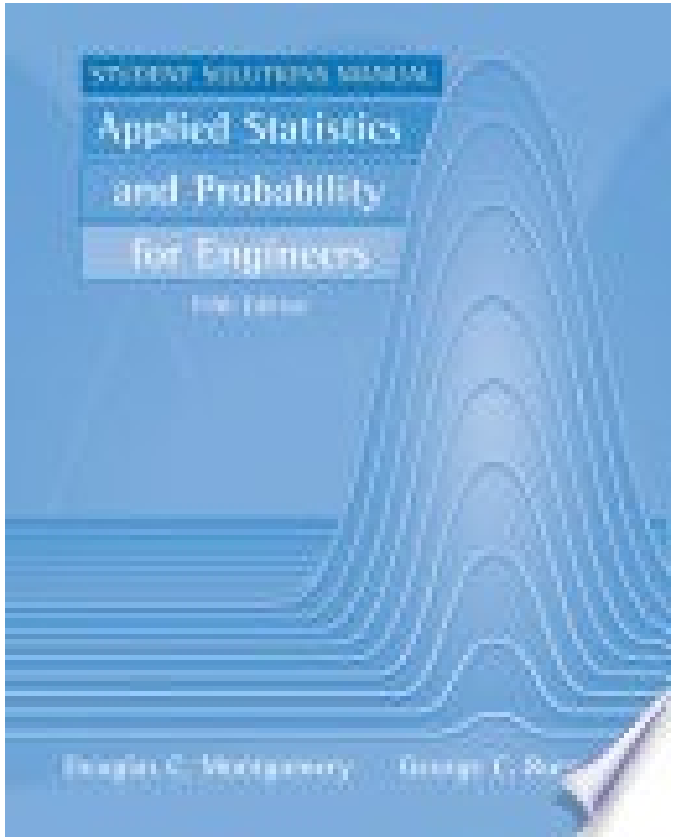
5th edition is available for free at our library

Problems for our main Probability/Statistics Textbook

Student Solutions Manual Applied Statistics and Probability for Engineers, 5th Edition
D. C. Montgomery and G. C. Runger
John Wiley & Sons, Inc. (2010)

You can also use other editions from 4th (2007) to 6th (2014)

5th edition is available for free at our library



Probability/Statistics for Bioengineering with Matlab exercises

Statistics for Bioengineering Sciences

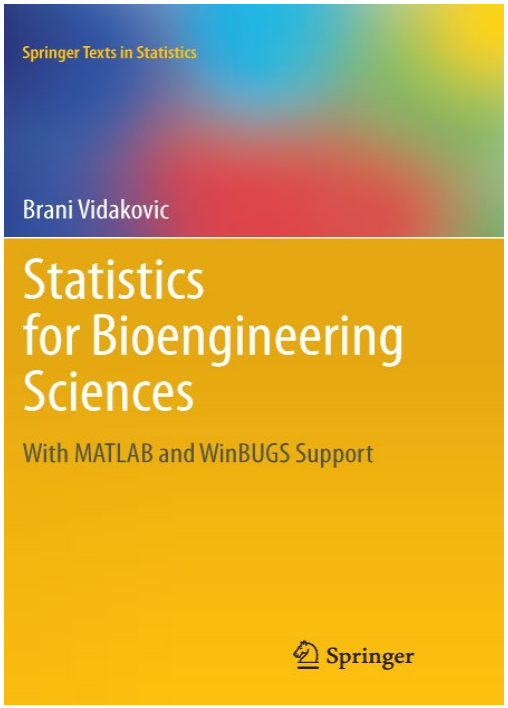
with MATLAB and WinBUGS Support

Brani Vidakovic

Department of Biomedical Engineering, Georgia Tech

(2011) Springer, New York

It is constantly updated with the newest version at the link below.



Free as a PDF eBook at

http://statbook.gatech.edu/statb4.pdf

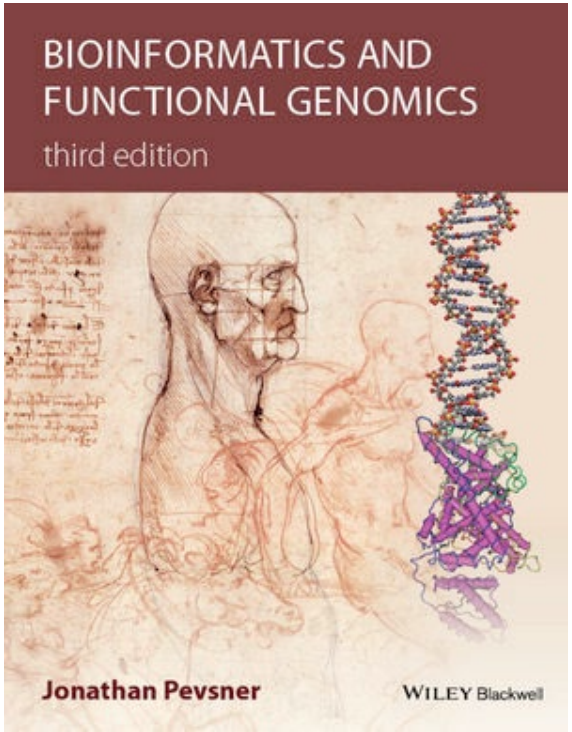
Matlab exercises and datasets are at

http://springer.bme.gatech.edu

Genomics/Systems Biology Textbook

- *J Pevsner*
  Bioinformatics and functional genomics
  Wiley-Blackwell,
  2nd edition [2009] exists in electronic form
  3rd edition [2015] has up-to-date information on NGS: RECOMMENDED
  (about $60 on amazon)

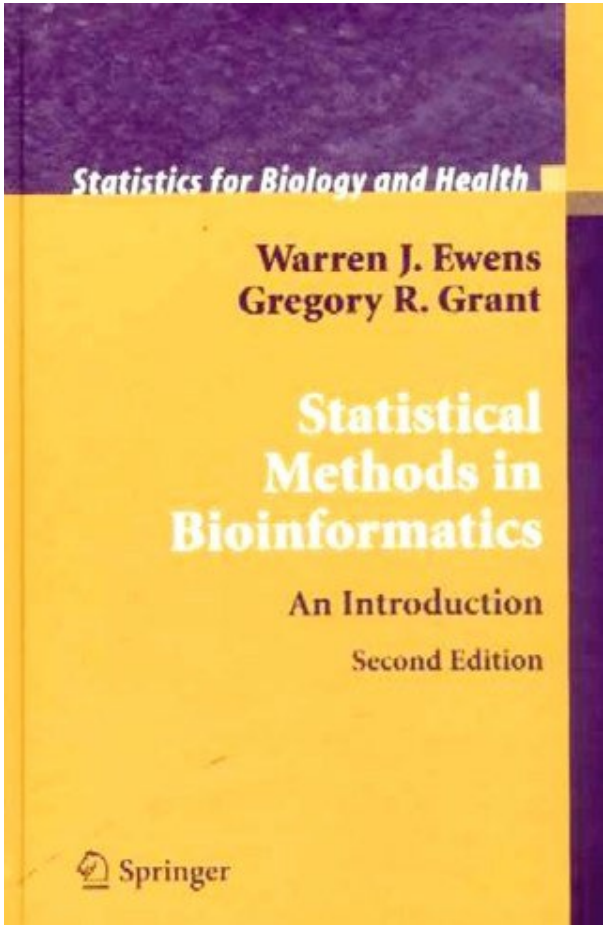
- *2nd edition is available for free in electronic form in our library*



Another Bioinformatics/Statistics Textbook

- *Ewens, WJ and Grant, GR* **Statistical Methods in Bioinformatics: An Introduction,** *2nd ed, Springer, 2005.*

- *2nd edition as PDF eBook*

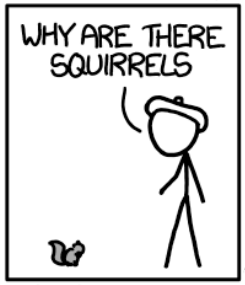


Credit: XKCD comics

QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT



WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT

WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY ARE MY BOOBS ITCHY
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD

WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

This course is about biological data and probability theory, and statistics concepts needed for its analysis

What biological data will be discussed?

Will be covered in lectures or Matlab exercises:

- Genomic data: strings of letters ACGT
- Gene Expression data: messenger RNA copy numbers transcribed from genes
- Proteomic data: protein abundances
- Network data: pairs of interacting genes or proteins and protein-protein interaction strengths

Will not be covered:

- Imaging data such as e.g. fMRI brain scans,
  Brain connectome data, Ecosystem dynamics data

Why do you need
probability and statistics
to analyze
modern biological data?

Definition of probability theory by Encyclopedia Britannica

a branch of mathematics concerned with the analysis of random phenomena

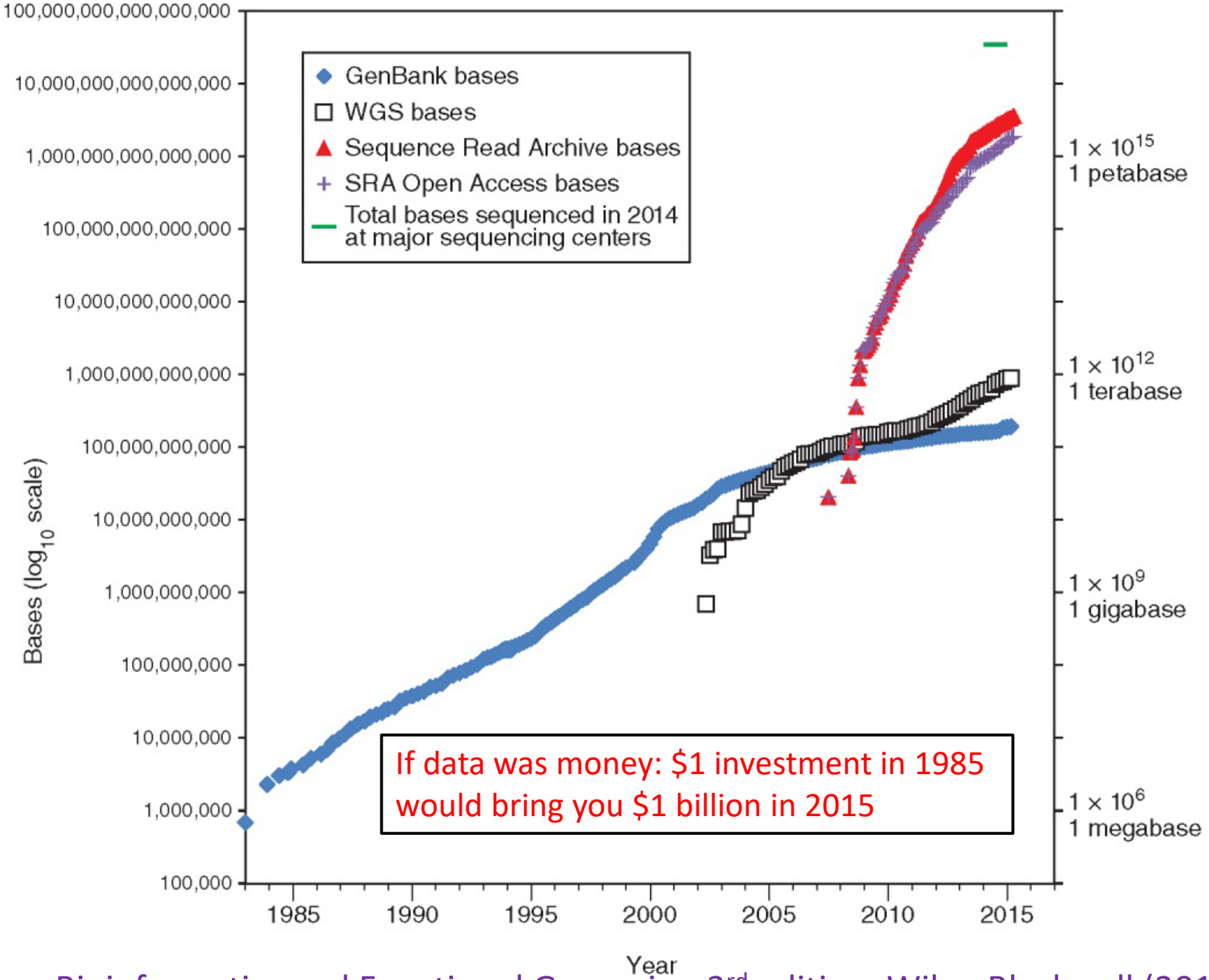
Definition of *statistics* by Merriam-Webster

1 : a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data

…

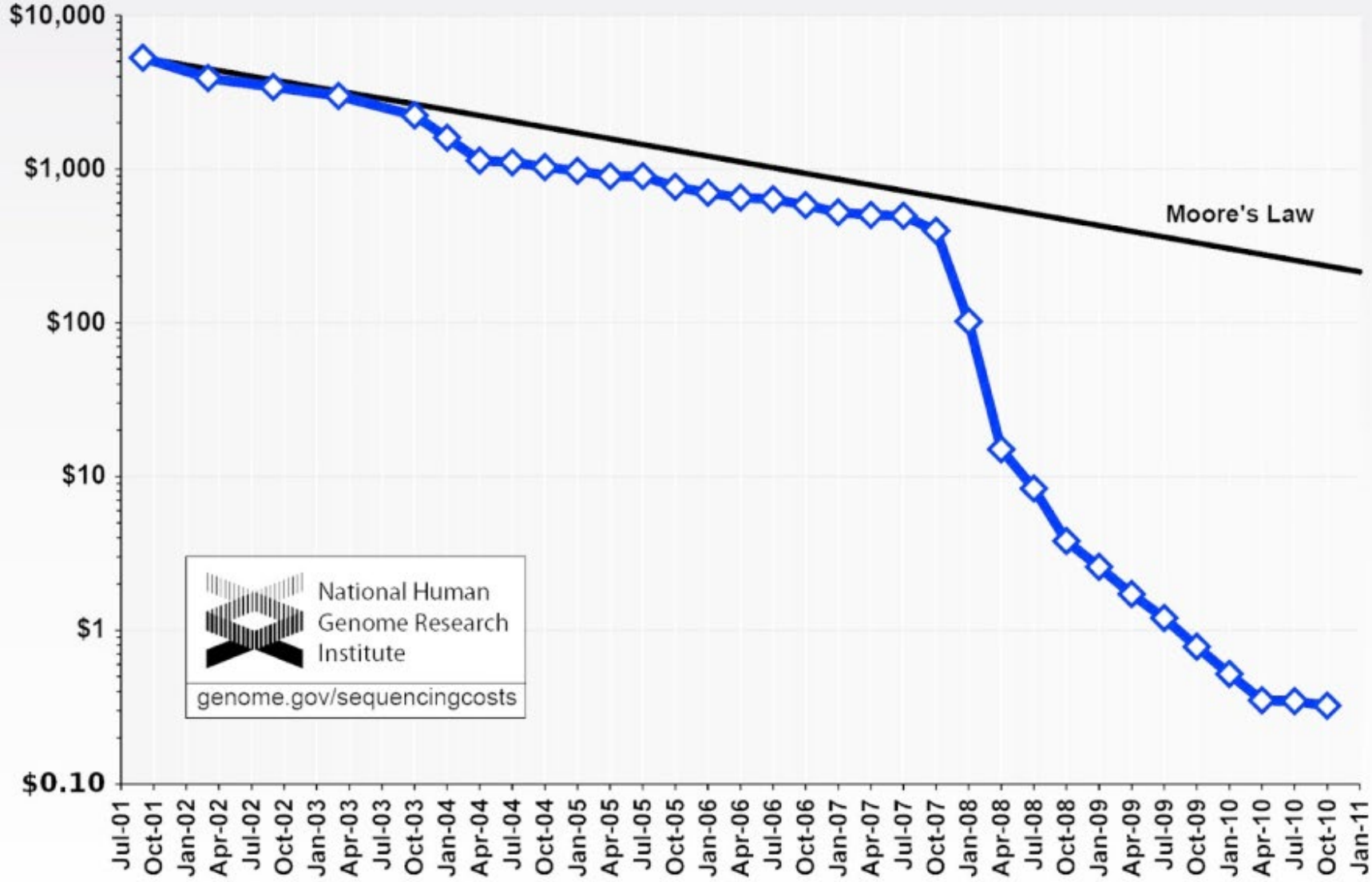
Why do you need
probability and statistics
to analyze
modern biological data?

Reason 1:
Biology now has <u>Lots of Data</u>



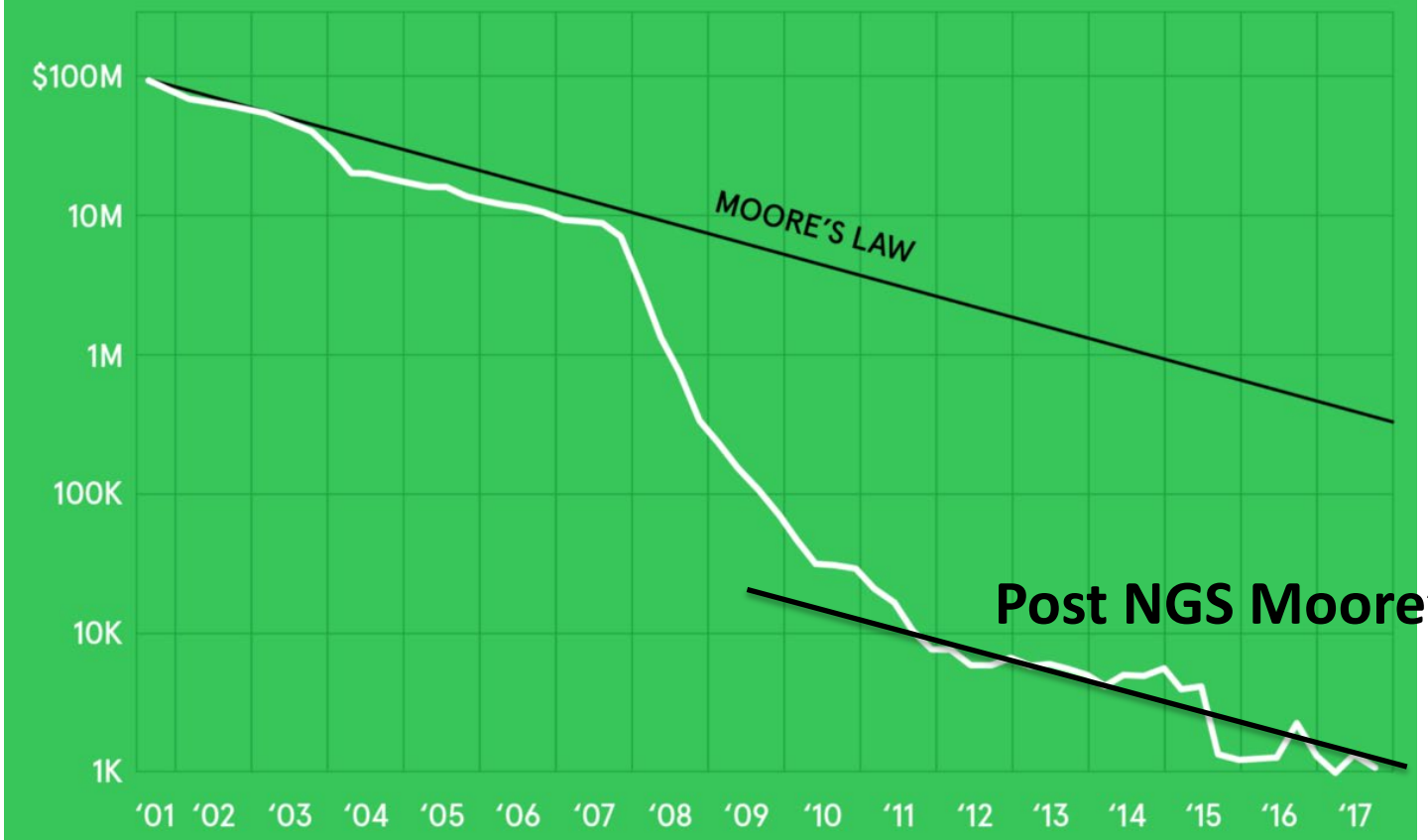
J. Pevsner, Bioinformatics and Functional Genomics, 3rd edition, Wiley-Blackwell (2015)

Cost per Megabase of DNA Sequence



Cost per Genome Sequenced

The cost of sequencing a human genome compared with the reductions that would be expected at the rate Moore's law predicts for computer chips. Over the past decade, next-generation sequencing and cloud computing drove the figure down. The average bumped higher in recent years because of brief slowdowns in production.



Source: NIH

NEO LIFE

Who will have bigger data by 2025?

Data Phase	Astronomy	Twitter
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year
Storage	1 EB/year	1–17 PB/year

Peta=10^{15} Exa=10^{18} Zetta=10^{21}

YouTube	Genomics
500–900 million hours/year	1 zetta-bases/year
1–2 EB/year	2–40 EB/year

Z. Stephens, S. Lee, F. Faghri, R. Campbell, C. Zhai, M. Efron, R. Iyer, M. Schatz, S. Sinha, and G. Robinson (2015) PLoS Biol 13: e1002195.

Base pairs	Unit	Abbreviation	Example
1	1 base pair	1 bp	A, C, G, T = 2 bits = 0.25 bytes
1000	1 kilobase pair	1 kb	
1,000,000	1 megabase pair	1 Mb	
10^9	1 gigabase pair	1 Gb	
10^{12}	1 terabase pair	1 Tb	
10^{15}	1 petabase pair	1 Pb	

Size	Abbreviation	No. bytes	Examples
Bytes	–	1	1 byte is typically 8 bits, used to encode a single character of text
Kilobytes	1 kb	10^3	Size of a text file with up to 1000 characters
Megabytes	1 MB	10^6	Size of a text file with 1 million characters
Gigabytes	1 GB	10^9	600 GB: size of GenBank (uncompressed flat files) ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt (WebLink 2.84)
Terabytes	1 TB	10^{12}	385 TB: United States Library of Congress web archive (http://www.loc.gov/webarchiving/faq.html) (WebLink 2.85)
			464 TB: Data generated by the 1000 Genomes Project (http://www.1000genomes.org/faq/how-much-disk-space-used-1000-genomes-project) (WebLink 2.86)
Petabytes	1 PB	10^{15}	1 PB: size of dataset available from The Cancer Genome Atlas (TCGA)
			5 PB: size of SRA data available for download from NCBI
			15 PB: amount of data produced each year at the physics facility CERN (near Geneva) (http://home.web.cern.ch/about/computing) (WebLink 2.87)
Exabytes	1 EB	10^{18}	2.5 exabytes of data are produced worldwide (Lampitt, 2014)

J. Pevsner, Bioinformatics and Functional Genomics, 3rd edition, Wiley-Blackwell (2015)

What makes genomic data so big?

- There are ~9 millions species each with its own genome
- Each of us humans (7.5 billions and counting) has unique DNA: we want to compare them all to each other
- Each cell has just 1 genome (DNA) but multitude of transcriptomes (RNA levels) and proteomes (protein levels)
- Cancer cells acquire mutations in their genomes: need to track multiple lineages in a tumor vs time to understand cancer
- DNA was proposed as a long-term storage medium of information

Farfetched? Storage standards evolve fast but DNA standard remained unchanged for 4 billion years

Note: Nature article started the comparison with a hard drive and flash memory skipping the floppy disk

nature International weekly journal of science

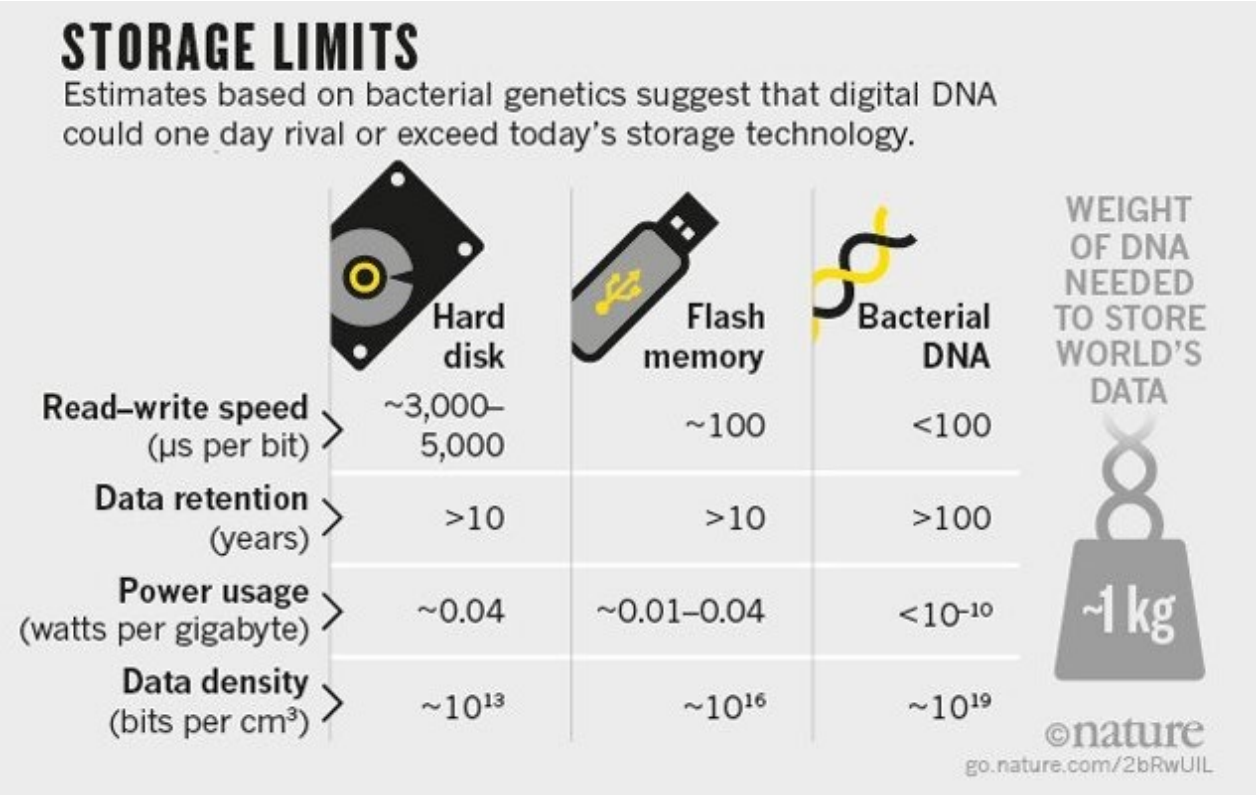
NATURE | NEWS FEATURE

How DNA could store all the world's data

Modern archiving technology
may hold an answer to that pr

Andy Extance

31 August 2016

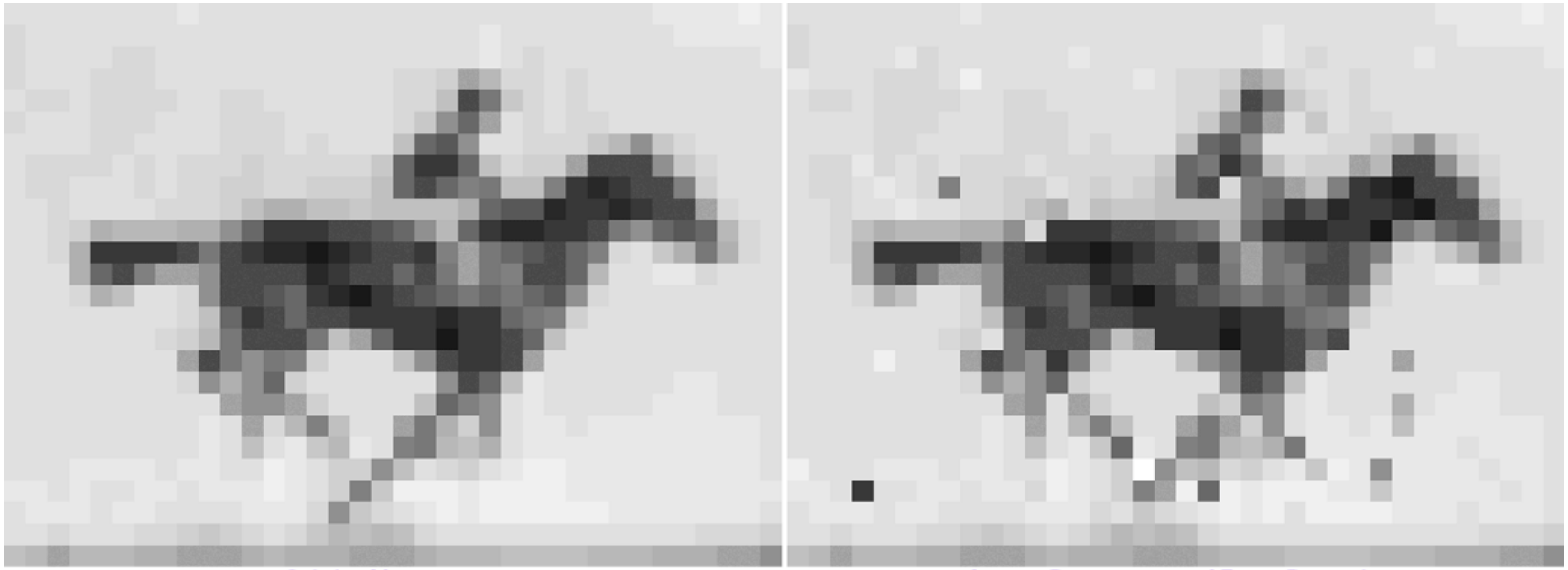


- Prof Olgica Milenkovic from Electrical and Computer Engineering UIUC is a local expert on this topic

- Profs. George Church and Sri Kosuri (Harvard Medical School) explains a potential use of DNA as storage medium in 2012

- https://www.youtube.com/watch?v=IJAdqAVjQqY

Fast-forward from 2012 to 2017



Original Image

Image Reconstructed From Bacteria

Shipman SL, Nivala J, Macklis JD, Church GM.
CRISPR–Cas encoding of a digital movie into the genomes
of a population of living bacteria. Nature. 2017;547: 345–349. doi:10.1038/nature23017

Why do you need
probability and statistics
to analyze
modern biological data?

Reason 2:
Life is random and messy

Show video
“Cell organelles”

- Made at the Walter and Eliza Hall Institute of Medical Research at Victoria, Australia
- Animated by award-winning artist Dr. Drew Berry
- Go to https://www.wehi.edu.au/wehi-tv for other videos

Life is messy, random, and noisy

Yet it is beautifully complex
and has many parts
(see statistics)

Why life is so random?

- Biomolecules are very small (nano- to micro-meters) → Brownian noise
- # molecules/cell is often small → Large cell-to-cell variations
- Genomic data comes from biological evolution
  - the Mother of all random processes
- Genomic data involves (random) samples
  - We have genomes of some (not all) organisms
  - We have tissue samples of some (not all) cancer patients

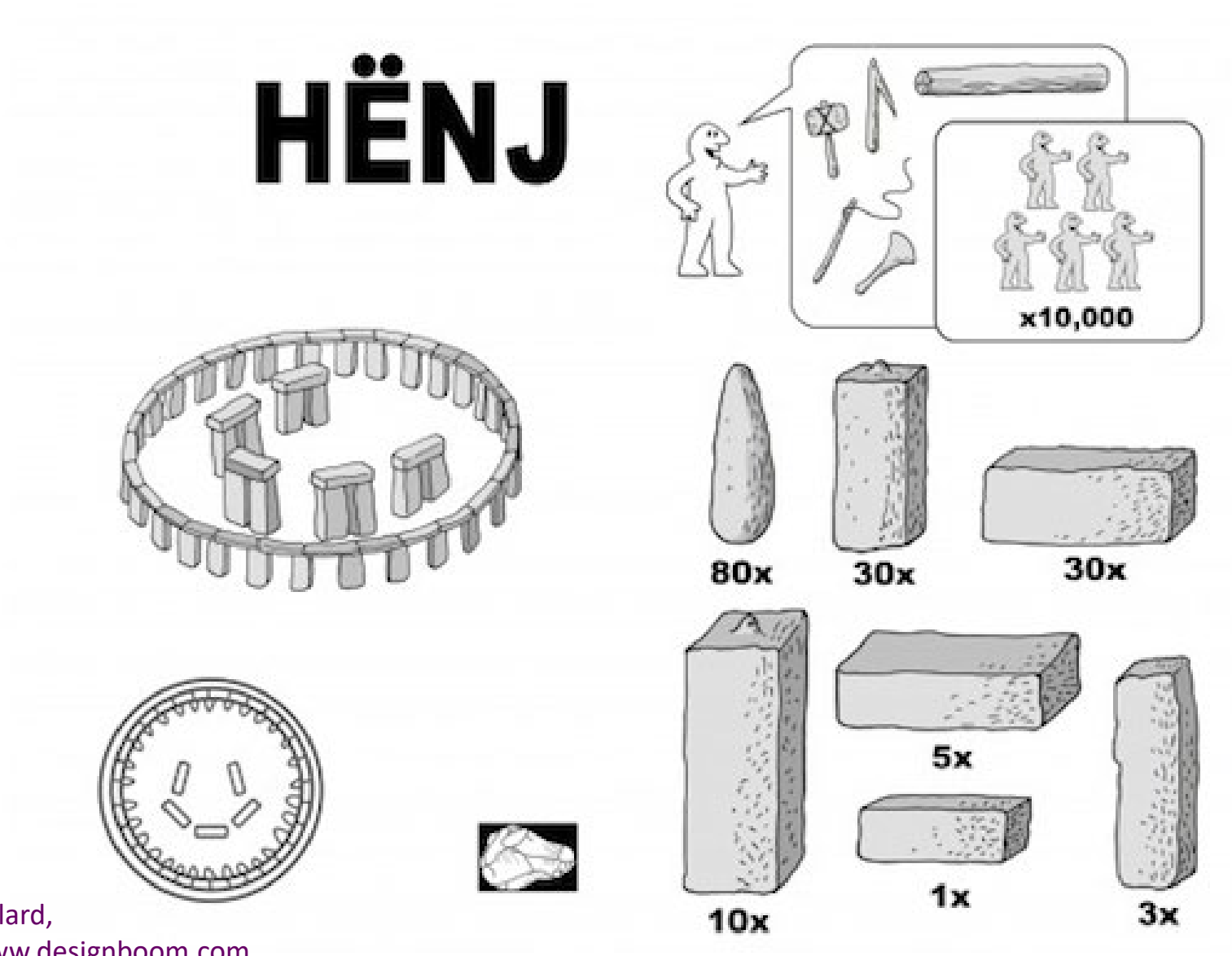
Why life is so complex?

Primer on complex system

Complex systems have many interacting parts

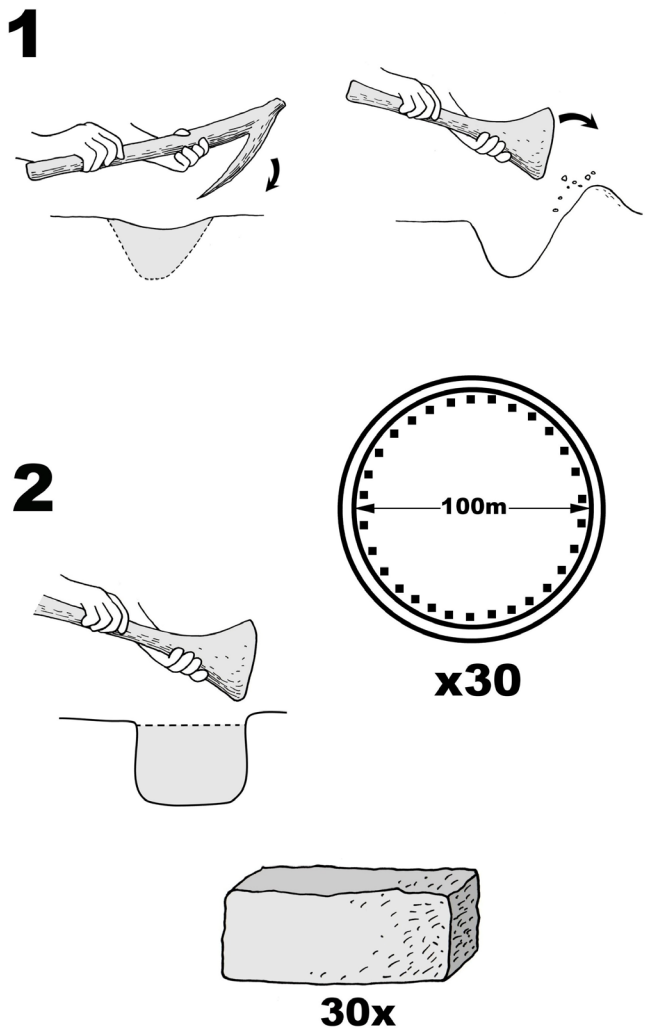
- All parts are different from each other
  - 10s thousands (10^4) types of proteins in an organism
  - 100 thousands (10^5) organizations (AS) in the Internet
  - 1 billion (10^9) people on Facebook
  - 10 billion (10^{10}) web pages in the WWW
  - 100 billion (10^{11}) neurons in a human brain
  - NOT 10^{23} electrons or quarks studied by physics: they are all the same and boring!
- Yet they share the same basic design
  - All proteins are strings of the same 20 amino acids
  - All WWW pages use HTML, JavaScript, etc.
  - All neurons generate and receive electric spikes

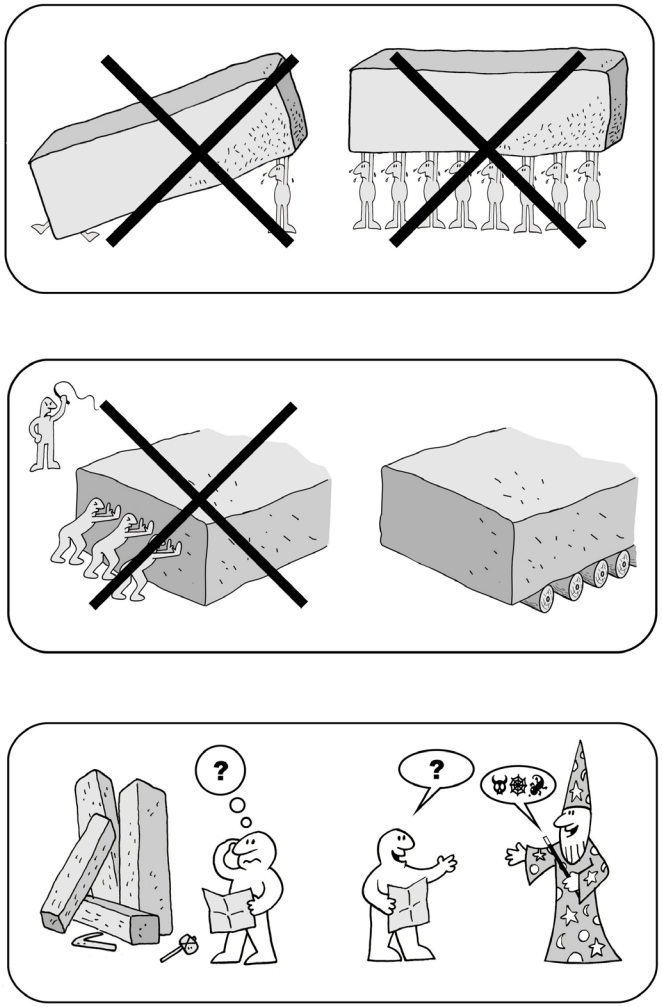
Example: a complex system with many parts



Justin Pollard,
http://www.designboom.com

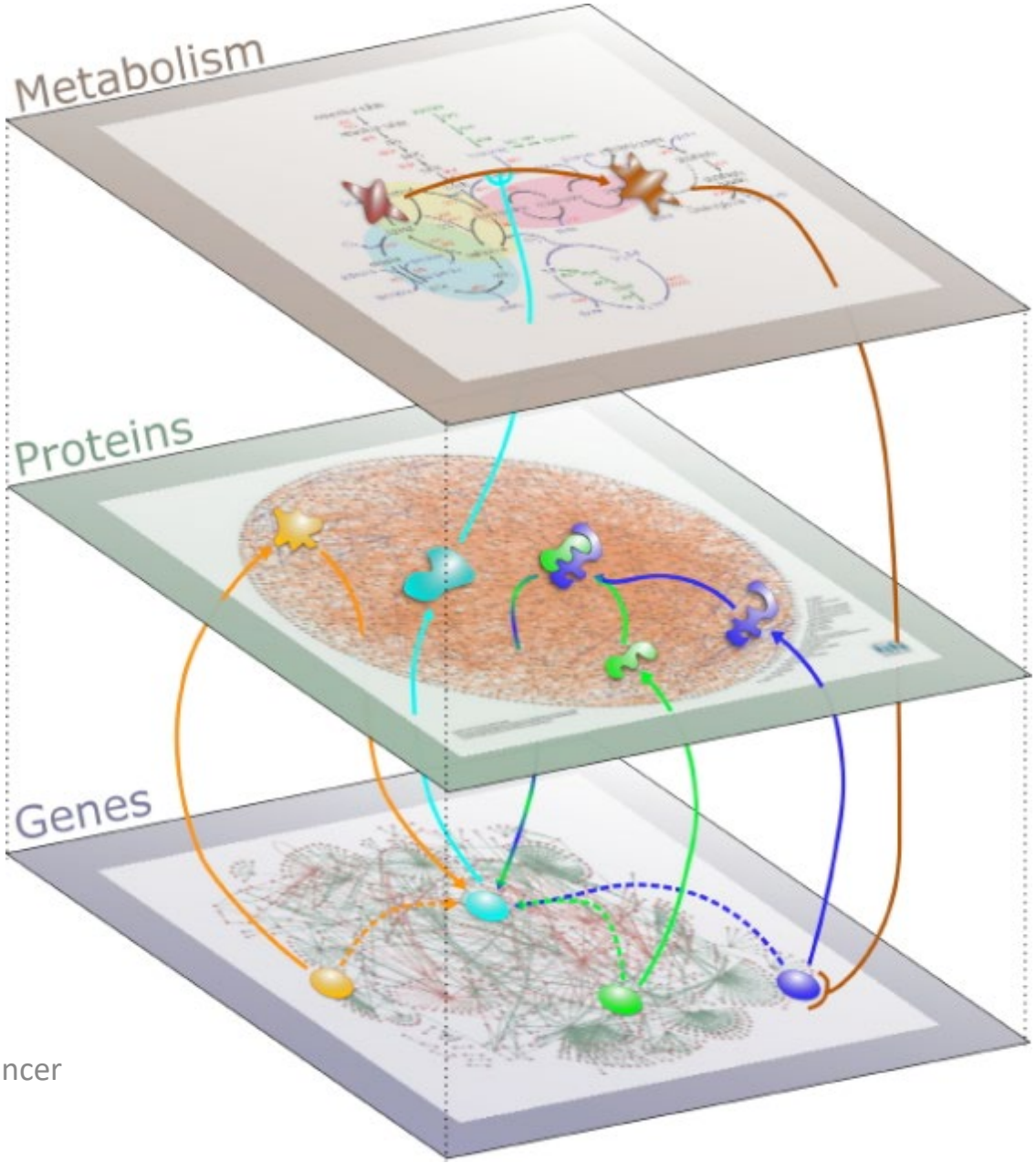
Parts interact →
they need to be assembled to work





Justin Pollard,
http://www.designboom.com

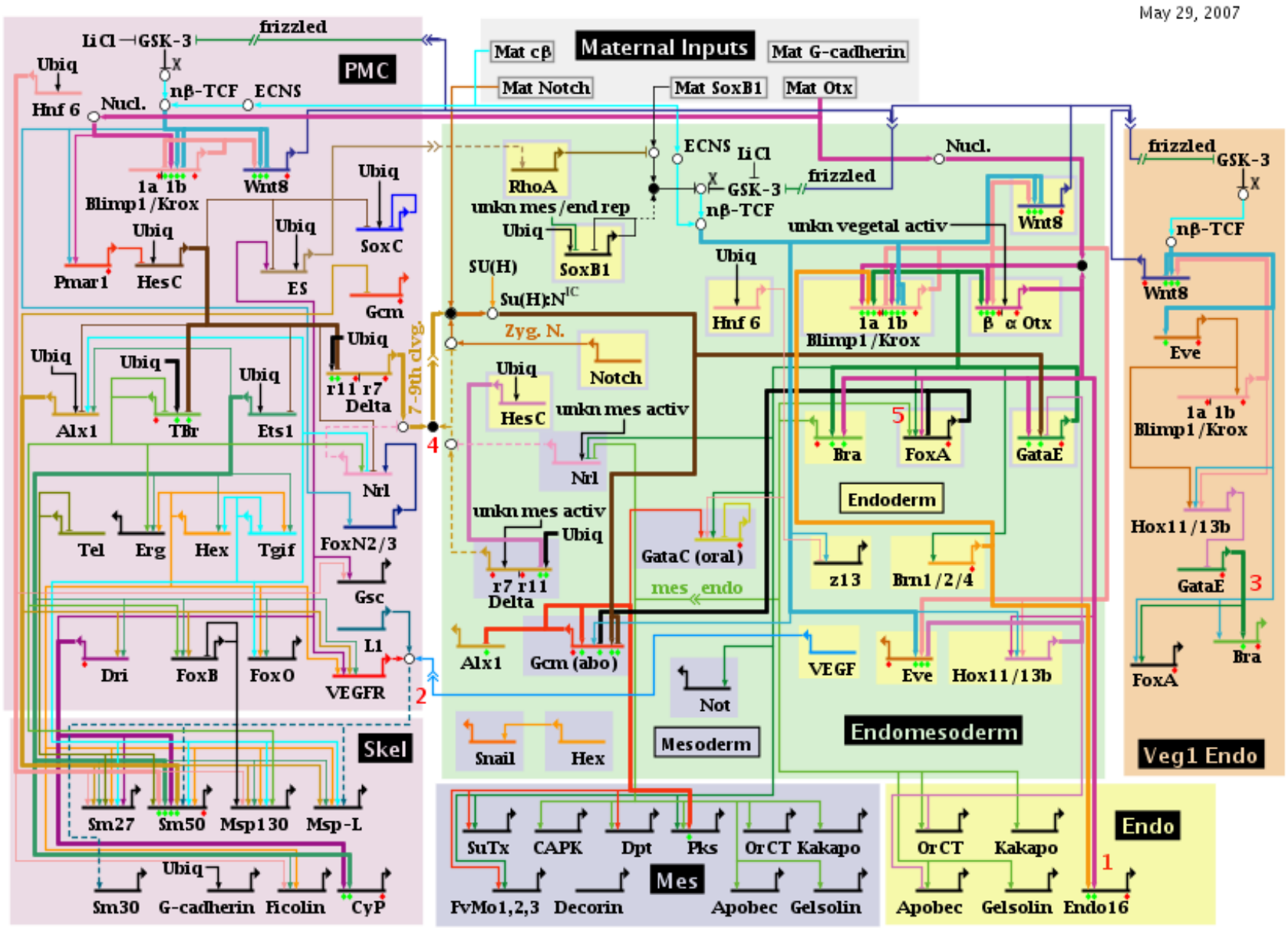
Intra-cellular Networks operate on multiple levels



Slides by Amitabh Sharma, PhD

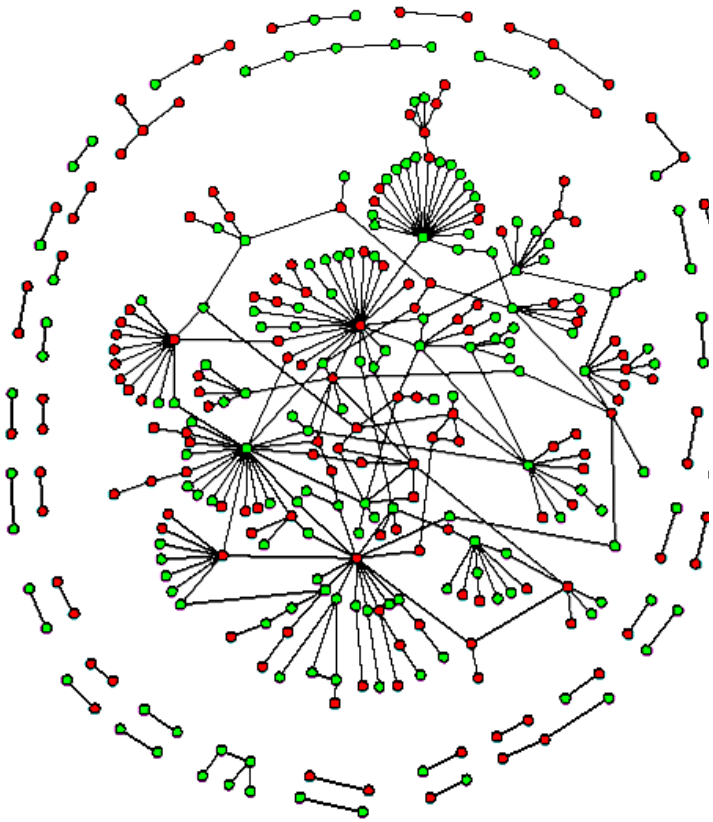
Northeastern University & Dana Farber Cancer Institute

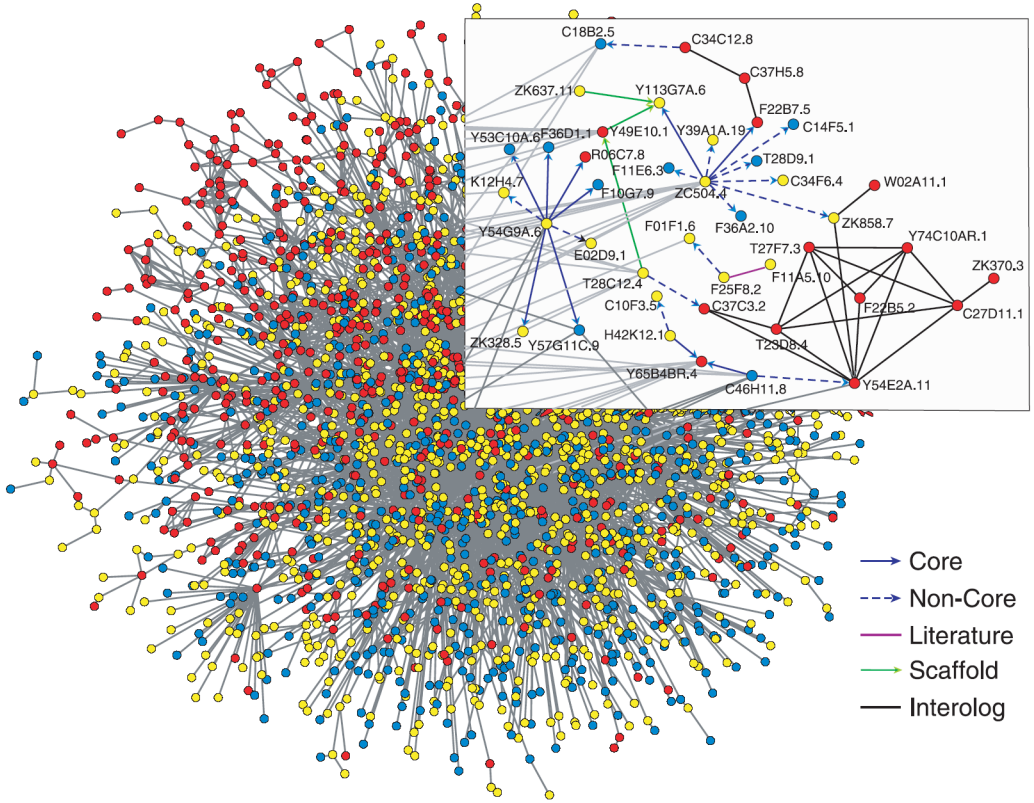
Sea urchin embryonic development (from endomesoderm up to 30 hours) by Davidson's lab



Ubiq=ubiquitous; Mat = maternal; activ = activator; rep = repressor;
unkn = unknown; Nucl. = nuclearization; χ = β-catenin source;
nβ-TCF = nuclearized b-β-catenin-Tcf1; ES = early signal;
ECNS = early cytoplasmic nuclearization system; Zyg. N. = zygotic Notch

Protein-Protein binding
IntAct Database (Dec 2015)
Interactions: 577,297 Proteins: 89,716

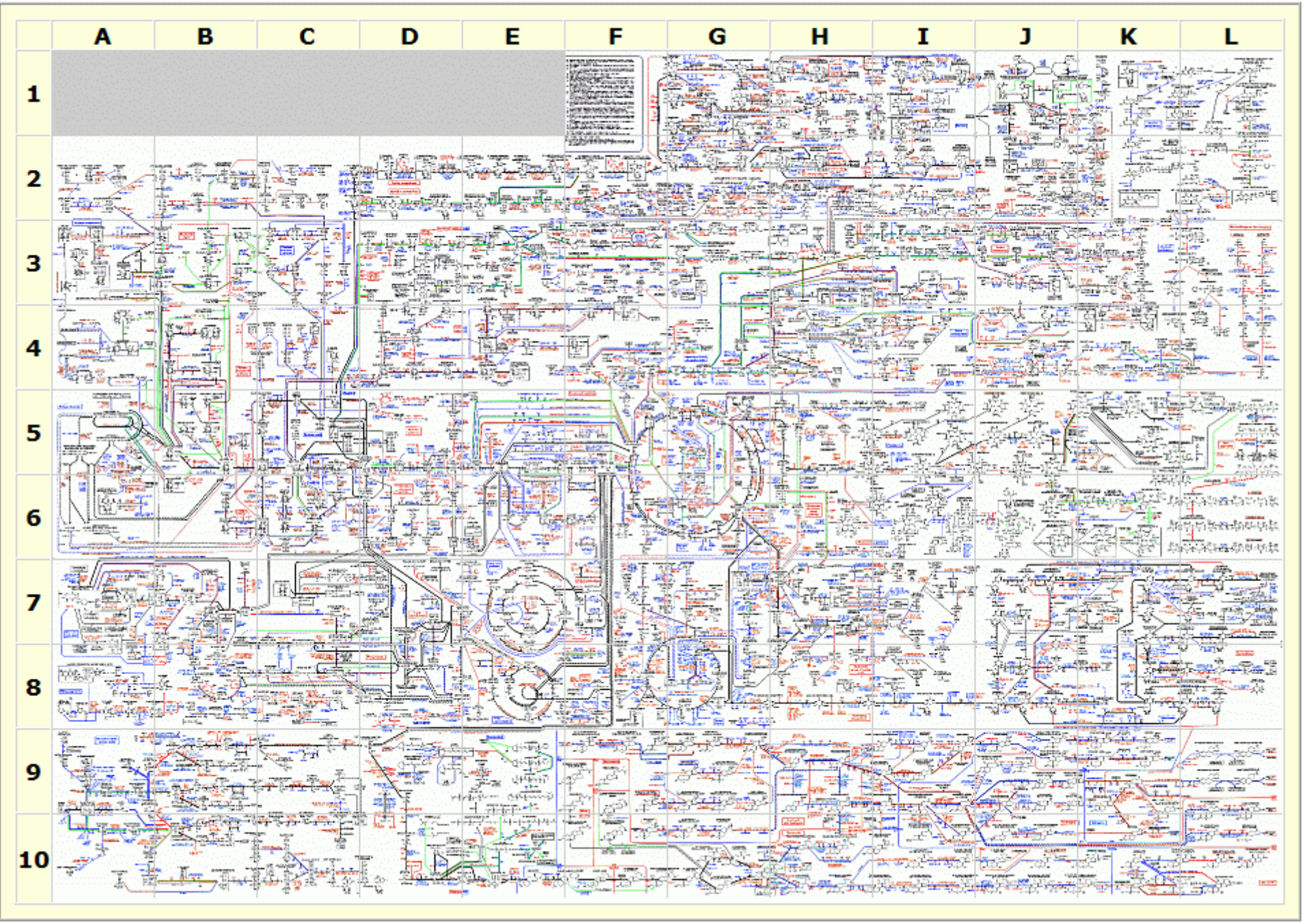




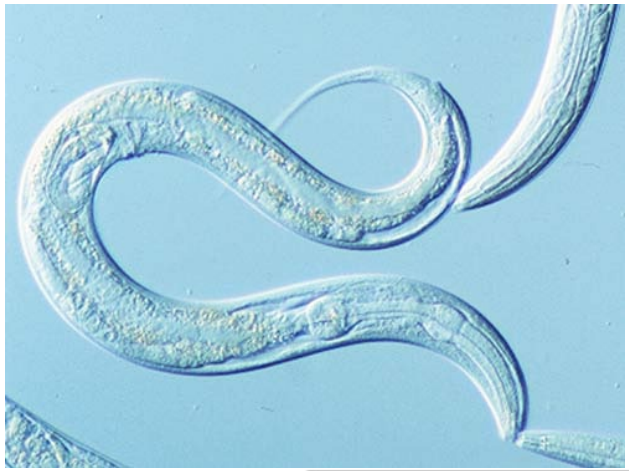
Baker's yeast *S. cerevisiae* (only nuclear proteins shown)
From S. Maslov, K. Sneppen, Science 2002

Worm *C. elegans*
From S. Lee et al , Science 2004

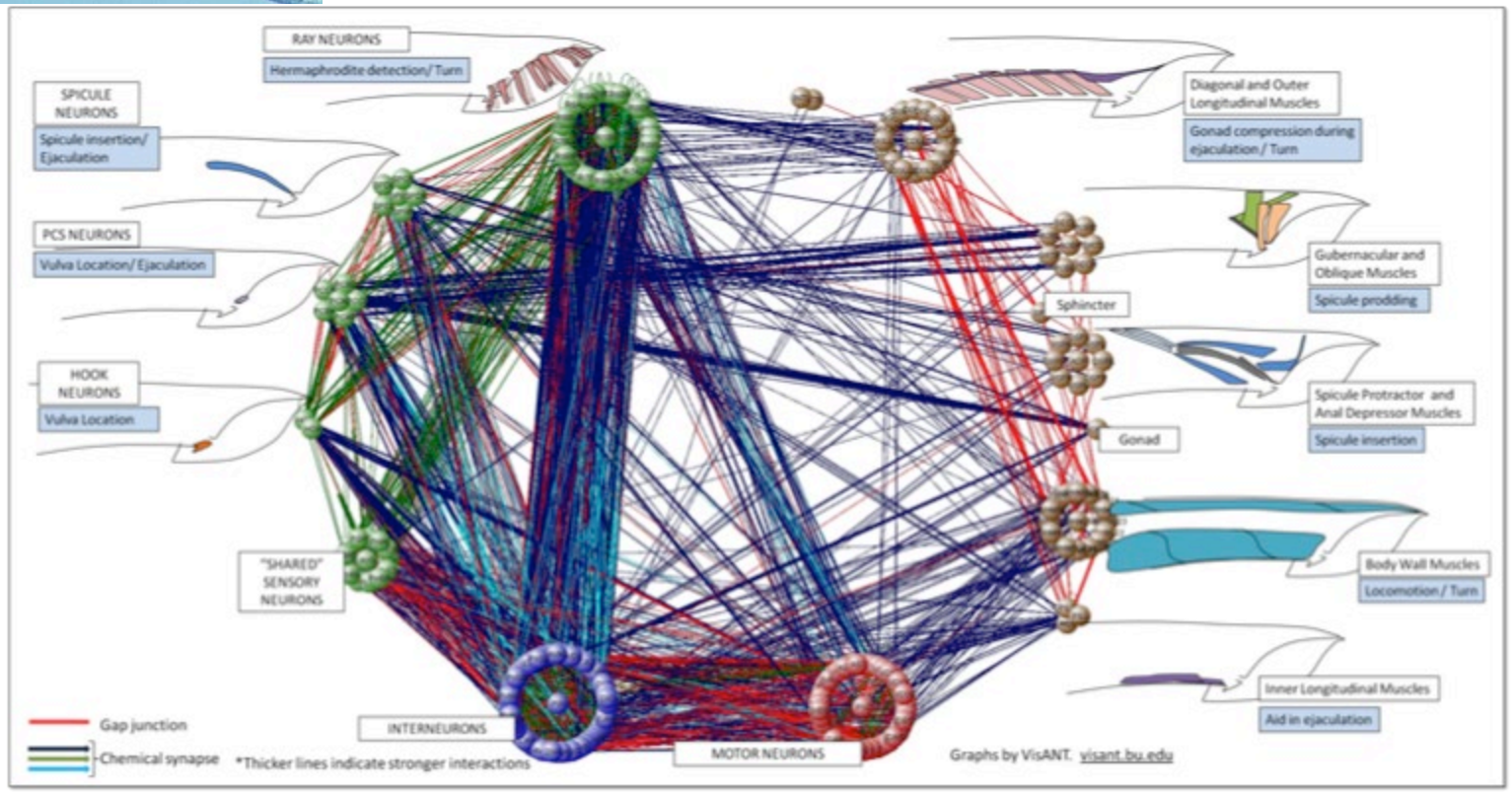
Metabolic pathway chart by ExPASy: **5702 reactions as of December 2015**

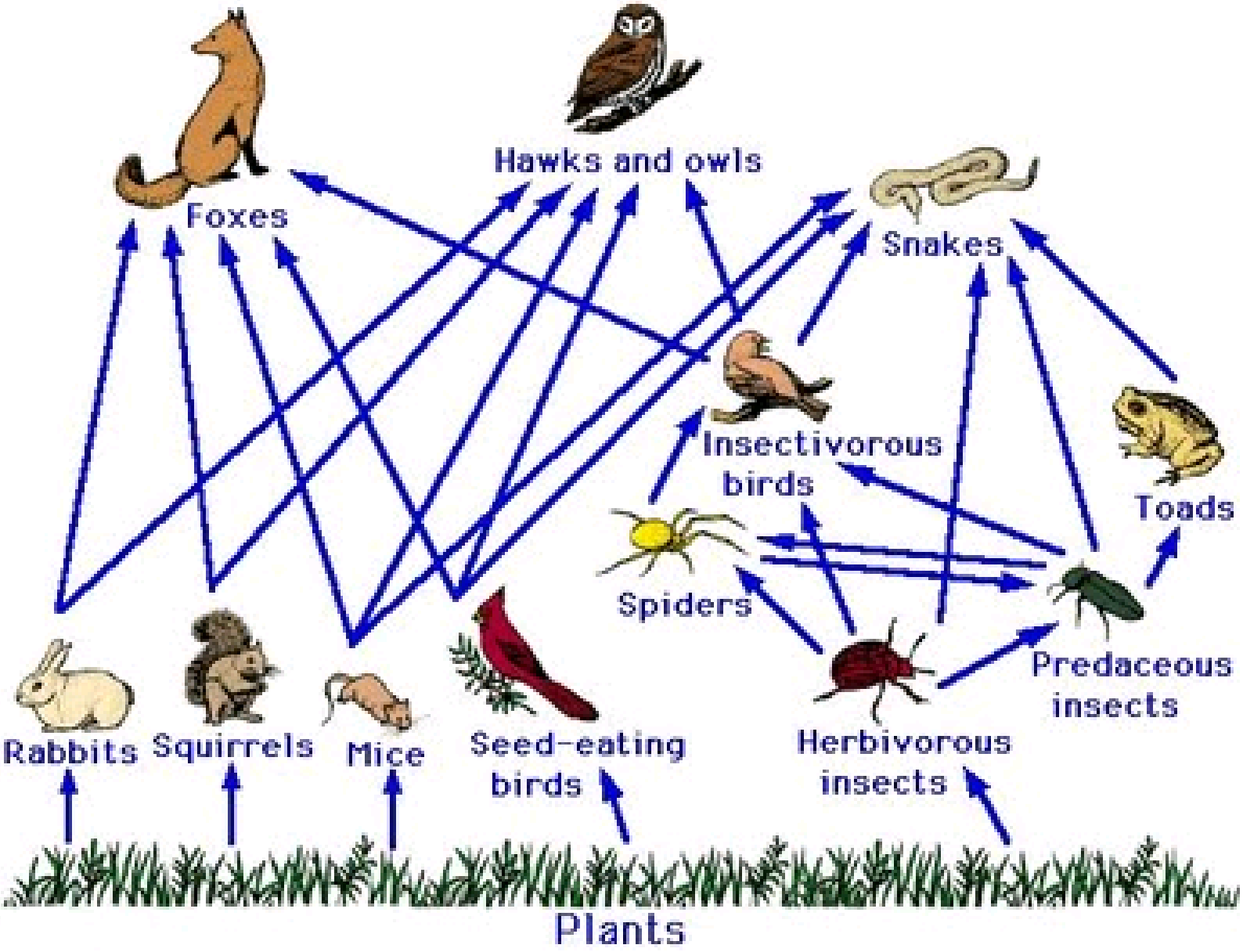


Brain and nerves of a worm

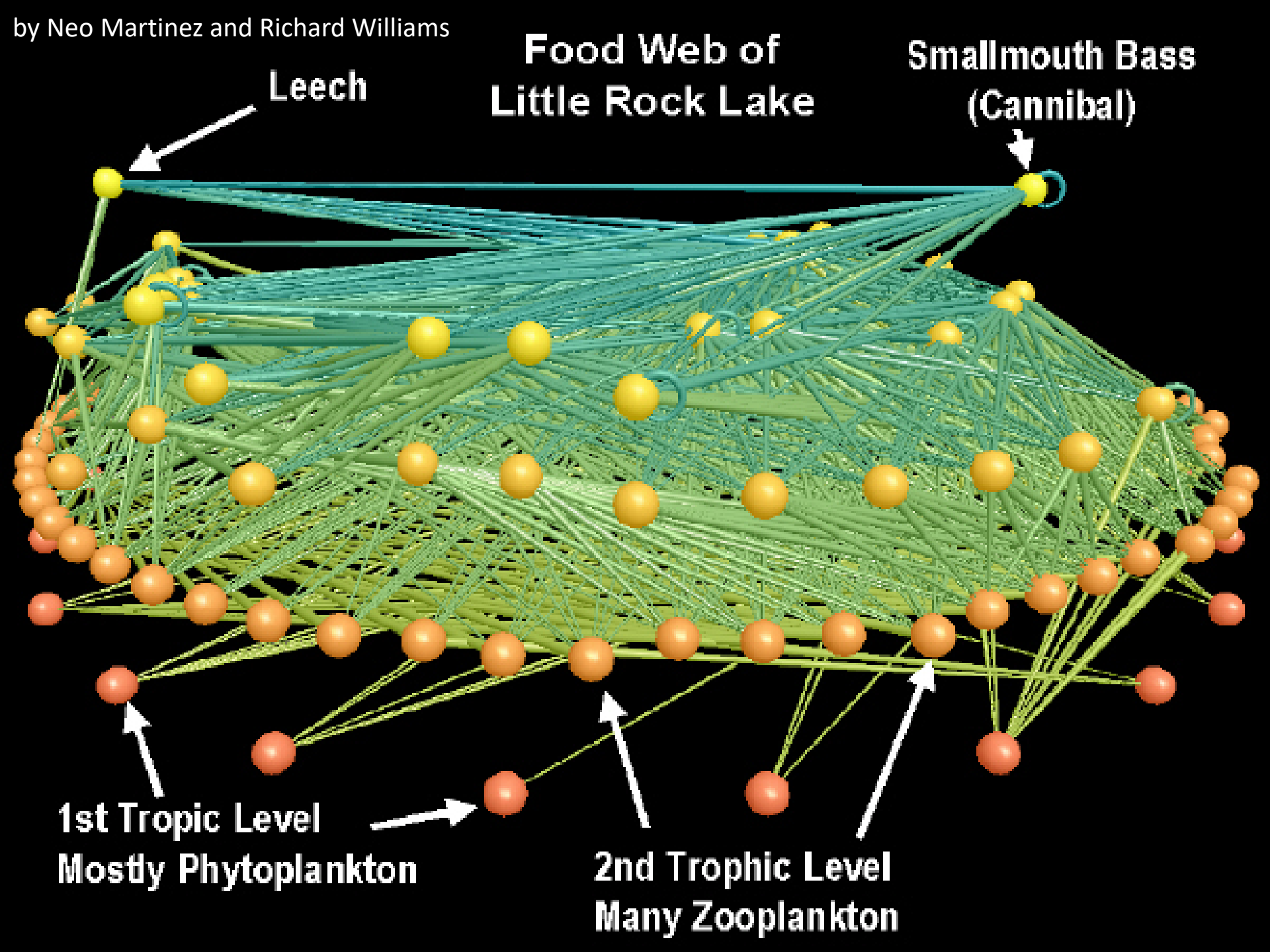


- Worm (C. elegans) has 302 neurons
- Our brain has 100 billion (10^{11}) neurons




Foxes
Hawks and owls
Snakes
Insectivorous birds
Toads
Spiders
Predaceous insects
Rabbits
Squirrels
Mice
Seed-eating birds
Herbivorous insects
Plants

by Neo Martinez and Richard Williams



QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT

WHY AREN'T THERE DINOSAUR GHOSTS

WHY DO TESTICLES MOVE
WHY ARE THERE PSYCHICS
WHY ARE HATS SO EXPENSIVE
WHY IS THERE CAFFEINE IN MY SHAMPOO
WHY DO YOUR BOOBS HURT
WHY ARE TREES TALL
WHY DO IGUANAS DIE

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY ARE MY BOOBS ITCHY
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY IS LIFE SO BORING
WHY ARE THERE FEMALE MR MIMES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA

WHY IS THERE HELL IF GOD FORGIVES
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

Credit: XKCD comics

Foundations of Probability

Random experiments

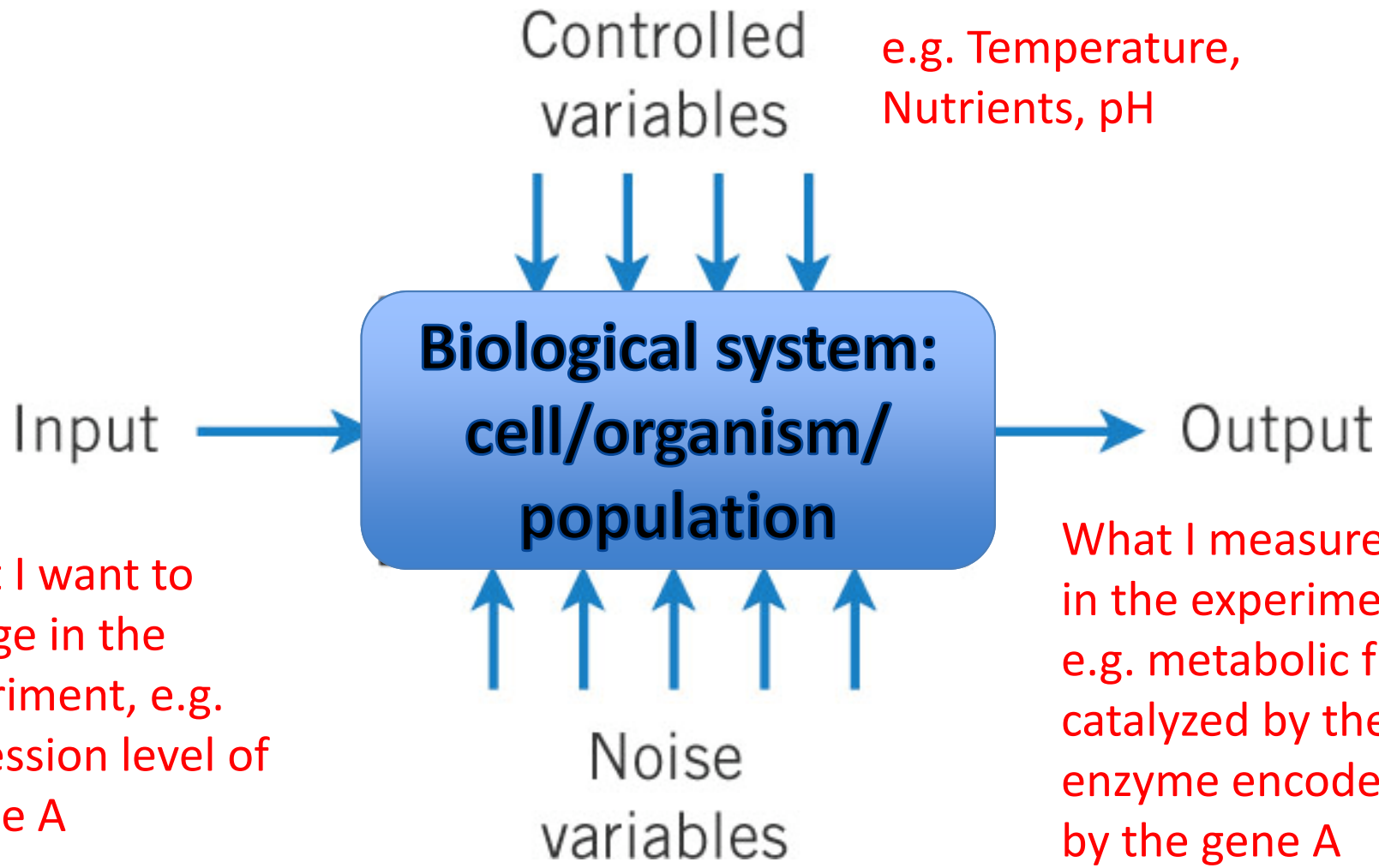
Sample spaces

Venn diagrams of random events

Random Experiments

- An experiment is an operation or procedure, carried out under controlled conditions
  - Example: measure the metabolic flux through a reaction catalyzed by the enzyme A
- An experiment that can result in different outcomes, even if repeated in the same manner every time, is called a random experiment
  - Cell-to-cell variability due to history/genome variants
  - Noise in external parameters such as temperature, nutrients, pH, etc.
- Evolution offers ready-made random experiments
  - Genomes of different species
  - Genomes of different individuals within a species
  - Individual cancer cells

Variability/Noise Produce Output Variation



e.g. Temperature, Nutrients, pH

What I want to change in the experiment, e.g. expression level of a gene A

What I measure in the experiment, e.g. metabolic flux catalyzed by the enzyme encoded by the gene A

Internal state of individual cells, Signals from neighbors

Sample Spaces

- Random experiments have unique outcomes.
- The set of all possible outcomes of a random experiment is called the sample space, *S*.
- *S* is discrete if it consists of a finite or countable infinite set of outcomes.
- *S* is continuous if it contains an interval (either a finite or infinite width) of real numbers.

Examples of a Sample Space

- Experiment measuring the abundance of mRNA expressed from a single gene
  $S = \{x | x >= 0\}$: continuous.
- Bin it into four groups
  $S = \{$*below 10, 10-30, 30-100, above 100*$\}$: discrete.
- Is gene "on" (mRNA above 30)?
  $S = \{$*true, false*$\}$: logical/Boolean/discrete.

Event

An event (*E*) is a subset of the sample space of a random experiment, i.e., one or more outcomes of the sample space.

- The union of two events is the event that consists of all outcomes that are contained in either of the two events. We denote the union as $E_1 \cup E_2$

- The intersection of two events is the event that consists of all outcomes that are contained in both of the two events. We denote the intersection as $E_1 \cap E_2$

- The complement of an event in a sample space is the set of outcomes in the sample space that are not in the event. We denote the complement of the event E as E' (sometimes E^c or $\bar{E}$)

Examples

Discrete

1. Assume you toss a coin once. The sample space is $S = \{H, T\}$, where H = head and T = tail and the event of a head is $\{H\}$.

2. Assume you toss a coin twice. The sample space is $S = \{(H, H), (H, T), (T, H), (T, T)\}$, and the event of obtaining exactly one head is $\{(H, T), (T, H)\}$.

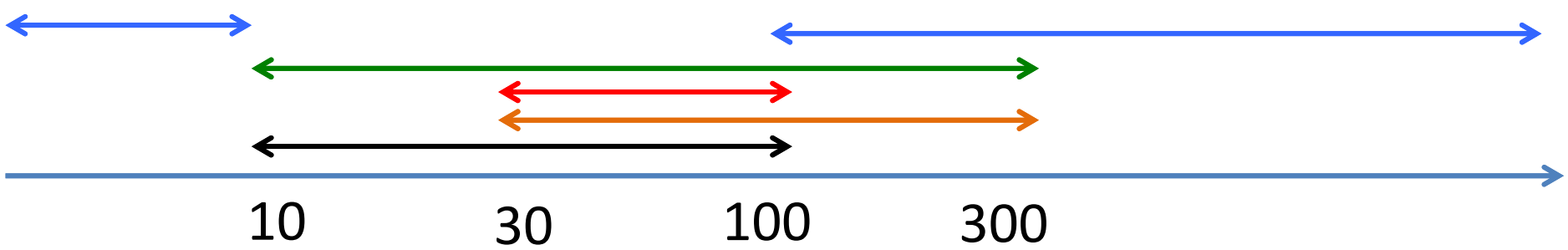
Continuous

Sample space for the expression level of a gene: $S = \{x \mid x \geq 0\}$

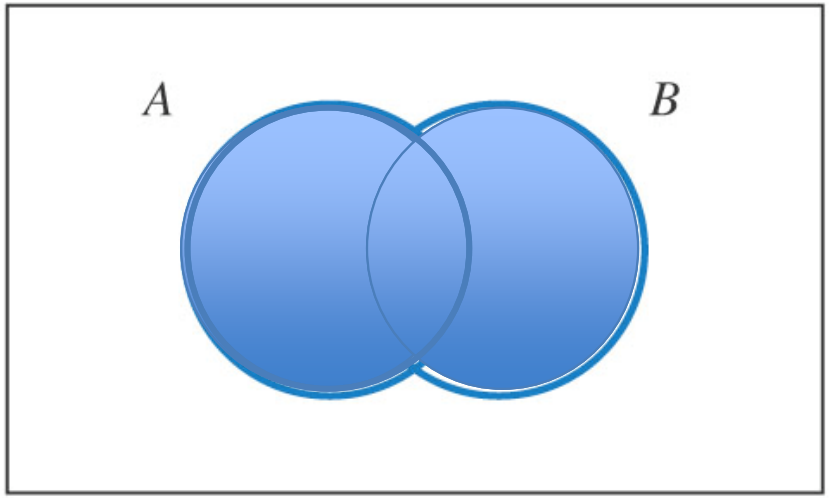
Two events:

- $E1 = \{x \mid 10 < x < 100\}$
- $E2 = \{x \mid 30 < x < 300\}$

- $E1 \cap E2 = \{x \mid 30 < x < 100\}$
- $E1 \cup E2 = \{x \mid 10 < x < 300\}$
- $E1' = \{x \mid x \leq 10 \text{ or } x \geq 100\}$



Venn diagrams



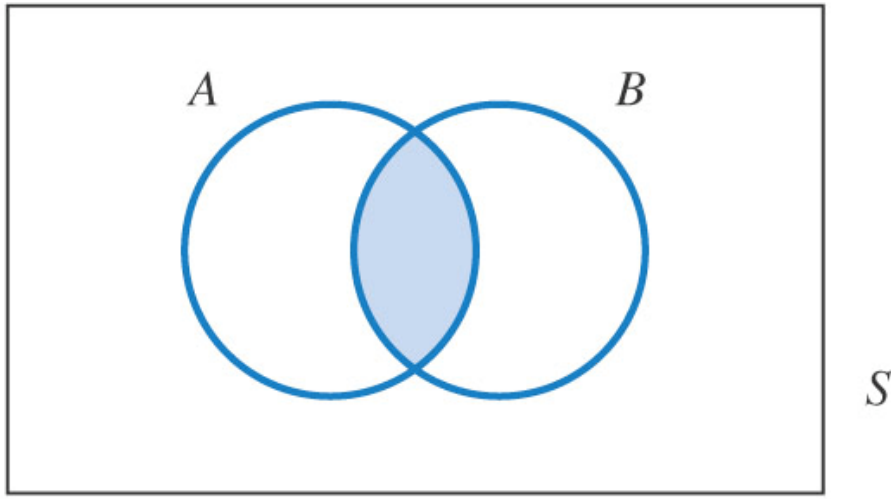
A U B

John Venn (1843-1923)
British logician

Find
5 differences
in beard and
hairstyle

John Venn (1990-)
Brooklyn hipster

Venn diagrams



Which formula describes the blue region?

A. A ∪ B

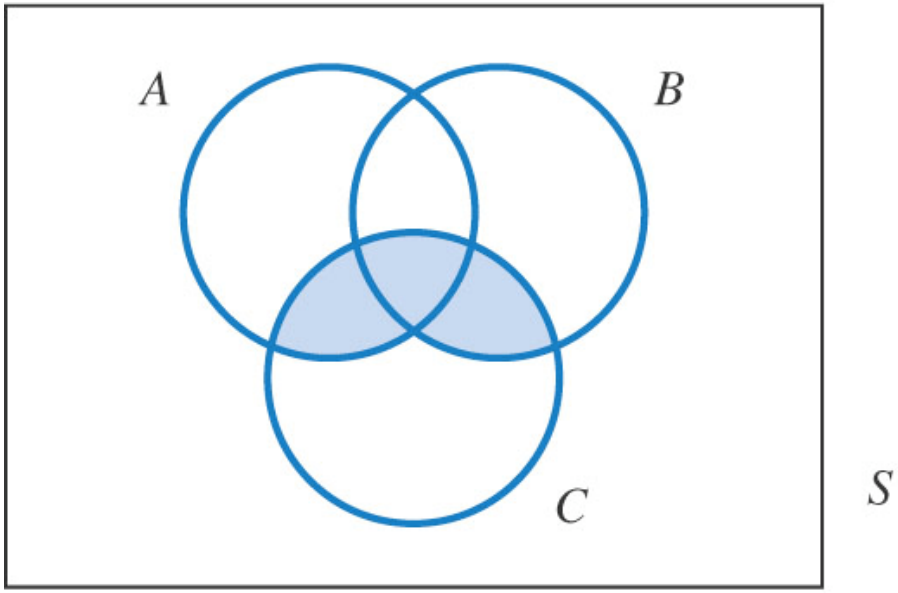
B. A ∩ B

C. A’

D. B’

Get your i-clickers

Venn diagrams

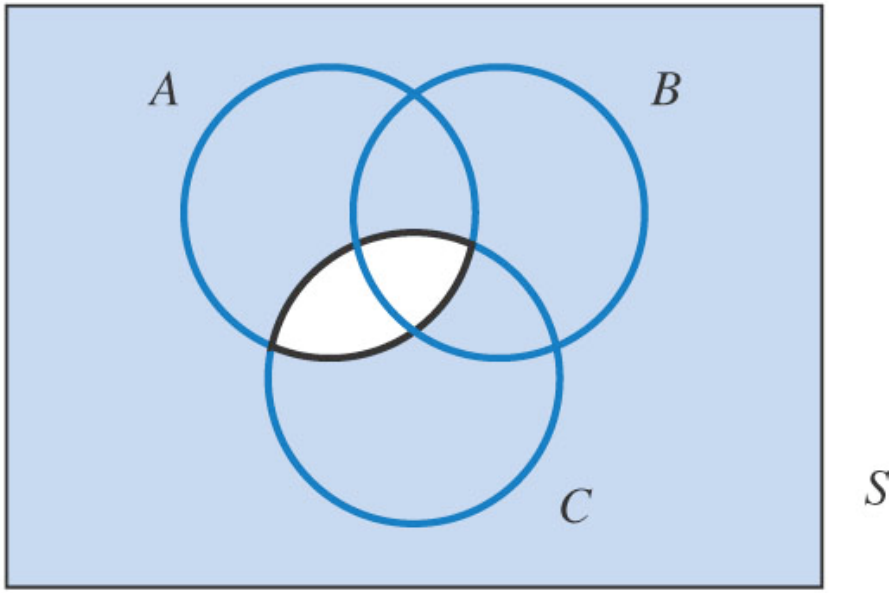


Which formula describes the blue region?

A. (A U B) ∩ C
B. (A ∩ B) ∩ C
C. (A U B) U C
D. (A ∩ B) U C

Get your i-clickers

Venn diagrams



Which formula describes the blue region?

A. $A \cap C$

B. $A' \cup C'$

C. $(A \cap B \cap C)'$

D. $(A \cap B) \cap C$

Get your i-clickers

Credit: XKCD comics

QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY ARE THERE FEMALE MR MIMES
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING

WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE THERE SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE HELL IF GOD FORGIVES
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE

Definitions of Probability

Two definitions of probability

- (1) STATISTICAL PROBABILITY: the relative frequency with which an event occurs in the long run

- (2) INDUCTIVE PROBABILITY: the degree of belief which it is reasonable to place in a proposition on given evidence

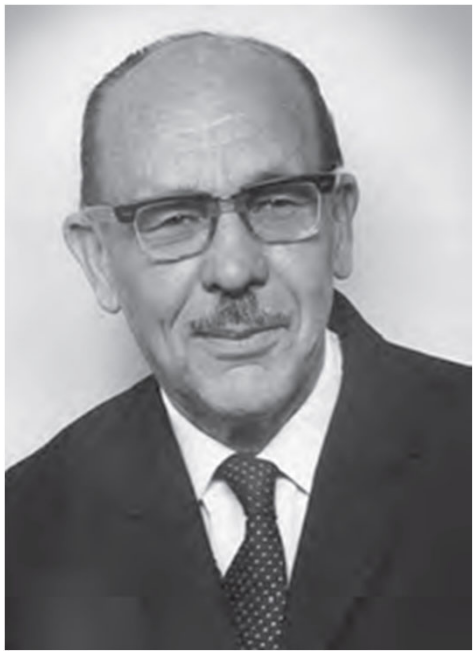
Bulmer, M. G.. Principles of Statistics (Dover Books on Mathematics)

Statistical Probability

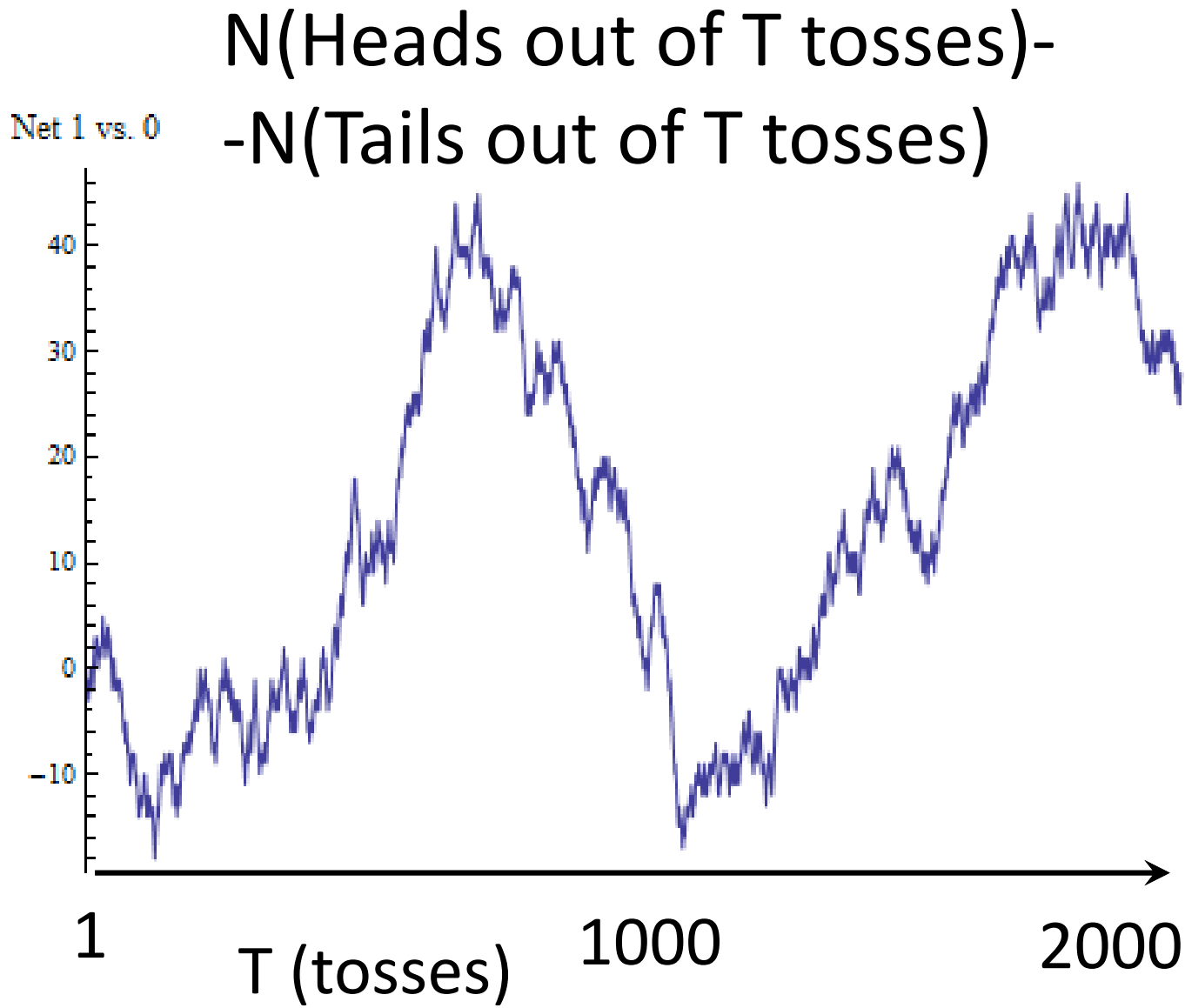
A statistical probability of an event is the limiting value of the relative frequency with it occurs in a very large number of independent trials

Empirical

Statistical Probability of a Coin Toss

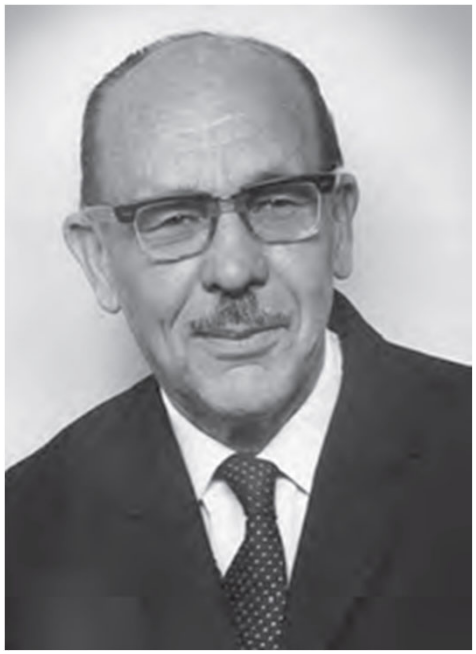


John Edmund Kerrich
(1903–1985)
British/South African
mathematician

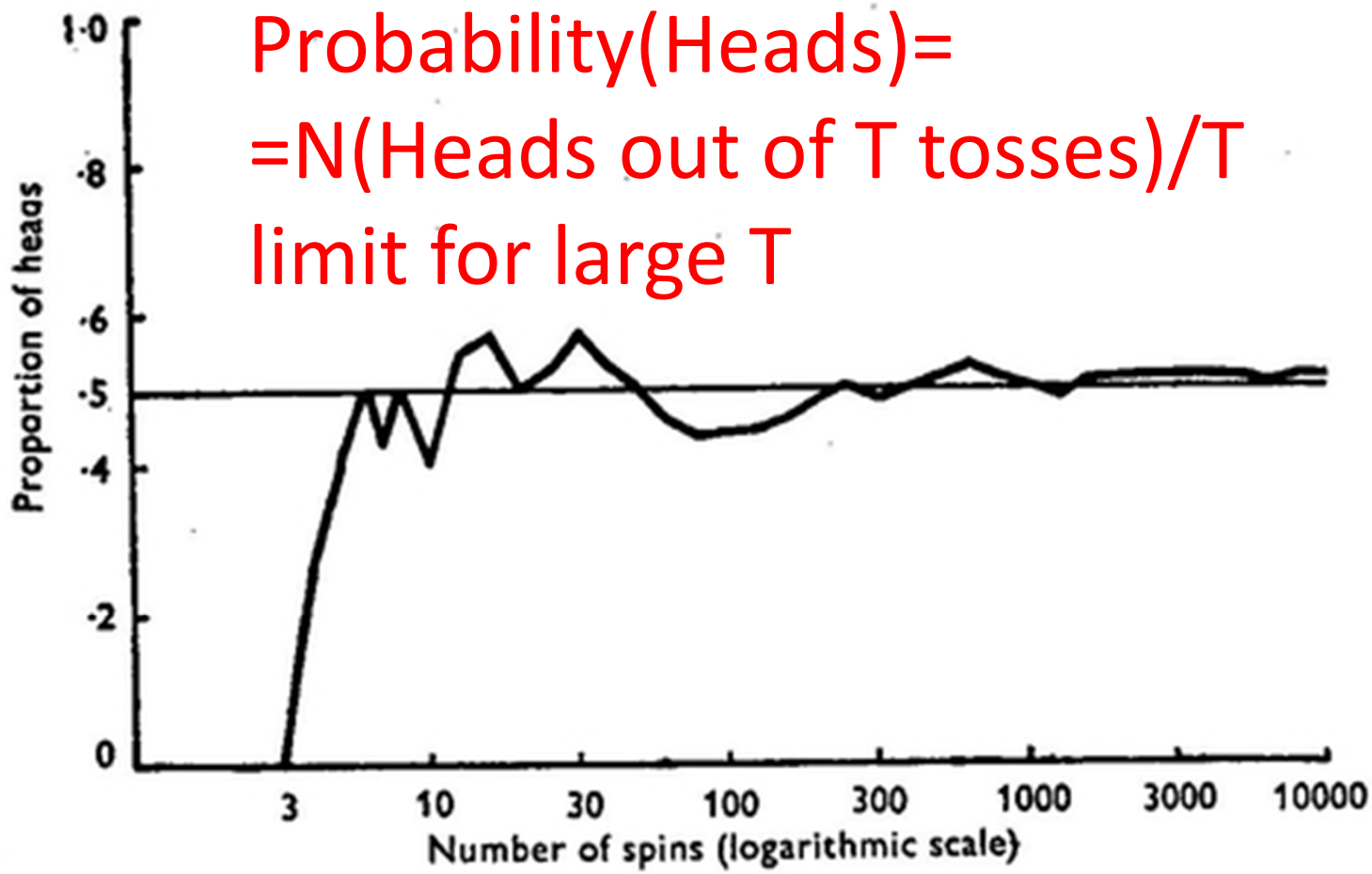


Excess of heads among 2,000 coin tosses (Kerrich 1946)

Statistical Probability of a Coin Toss



John Edmund Kerrich
(1903–1985)
British/South African
mathematician



Proportion of heads among 10,000 coin tosses (Kerrich 1946)

Who is ready to use Matlab?

A. I have Matlab installed on my laptop
B. I am ready to use Matlab on EWS
C. I don’t have it ready but plan to install it
D. I am not ready but plan to use EWS
E. I plan to use other software (Python, R, etc.)

Get your i-clickers

Matlab is easy to learn

- Matlab is the lingua franca of all of **engineering**
- Use online tutorials e.g.: https://www.youtube.com/watch?v=82TGgQApFIQ
- Matlab is designed to work with Matrices → symbols * and / are understood as matrix multiplication and division
- Use .* and ./ for regular (non-matrix) multiplication
- Add ; in the end of the line to avoid displaying the output on the screen
- Loops: for i=1:100; f(i)=floor(2.*rand); end;
- Conditional statements: if rand>0.5; count=count+1; end;
- Plotting: plot(x,y,'ko-'); or semilogx(x,y,'ko-'); or loglog(x,y,'ko-'); .
  To keep adding plots onto the same axes use: hold on;
  To create a new axes use figure;
- Generating matrices: rand(100) – generates square matrix 100x100. **Confusing!** Use rand(100,1) or zeros(30,20), or randn(1,40) (Gaussian);
- If Matlab complains multiplying matrices check sizes using whos and if needed use transpose operation: x=x';

A Matlab Cheat-sheet (MIT 18.06, Fall 2007)

Basics:

`save 'file.mat'` save variables to *file.mat*
`load 'file.mat'` load variables from *file.mat*
`diary on` record input/output to file *diary*
`diary off` stop recording
`whos` list all variables currenly defined
`clear` delete/undefine all variables
`help command` quick help on a given *command*
`doc command` extensive help on a given *command*

Defining/changing variables:

`x = 3` define variable x to be 3
`x = [1 2 3]` set x to the 1×3 row-vector (1,2,3)
`x = [1 2 3];` same, but don't echo x to output
`x = [1;2;3]` set x to the 3×1 column-vector (1,2,3)
`A = [1 2 3 4;5 6 7 8;9 10 11 12];`
set A to the 3×4 matrix with rows 1,2,3,4 etc.
`x(2) = 7` change x from (1,2,3) to (1,7,3)
`A(2,1) = 0` change $A_{2,1}$ from 5 to 0

Arithmetic and functions of numbers:

`3*4, 7+4, 2-6 8/3` multiply, add, subtract, and divide numbers
`3^7, 3^(8+2i)` compute 3 to the 7th power, or 3 to the 8+2i power
`sqrt(-5)` compute the square root of –5
`exp(12)` compute e^{12}
`log(3), log10(100)` compute the natural log (ln) and base-10 log ($\log_{10}$)
`abs(-5)` compute the absolute value $|-5|$
`sin(5*pi/3)` compute the sine of $5\pi/3$
`besselj(2,6)` compute the Bessel function $J_2(6)$

Arithmetic and functions of vectors and matrices:

`x * 3` multiply every element of x by 3
`x + 2` add 2 to every element of x
`x + y` element-wise addition of two vectors x and y
`A * y` product of a matrix A and a vector y
`A * B` product of two matrices A and B
`x * y` not allowed if x and y are two column vectors!
`x .* y` element-wise product of vectors x and y
`A^3` the square matrix A to the 3rd power
`x^3` not allowed if x is not a square matrix!
`x.^3` every element of x is taken to the 3rd power
`cos(x)` the cosine of every element of x
`abs(A)` the absolute value of every element of A
`exp(A)` e to the power of every element of A
`sqrt(A)` the square root of every element of A
`expm(A)` the matrix exponential e^A
`sqrtm(A)` the matrix whose square is A

Constructing a few simple matrices:

`rand(12,4)` a 12×4 matrix with uniform random numbers in [0,1)
`randn(12,4)` a 12×4 matrix with Gaussian random (center 0, variance 1)
`zeros(12,4)` a 12×4 matrix of zeros
`ones(12,4)` a 12×4 matrix of ones
`eye(5)` a 5×5 identity matrix I ("eye")
`eye(12,4)` a 12×4 matrix whose first 4 rows are the 4×4 identity
`linspace(1.2,4.7,100)`
row vector of 100 equally-spaced numbers from 1.2 to 4.7
`7:15` row vector of 7,8,9,…,14,15
`diag(x)` matrix whose diagonal is the entries of x (and other elements = 0)

Portions of matrices and vectors:

`x(2:12)` the 2nd to the 12th elements of x
`x(2:end)` the 2nd to the last elements of x
`x(1:3:end)` every third element of x, from 1st to the last
`x(:)` all the elements of x
`A(5,:)` the row vector of every element in the 5th row of A
`A(5,1:3)` the row vector of the first 3 elements in the 5th row of A
`A(:,2)` the column vector of every element in the 2nd column of A
`diag(A)` column vector of the diagonal elements of A

Solving linear equations:

`A \ b` for A a matrix and b a column vector, the solution x to $Ax=b$
`inv(A)` the inverse matrix A^{-1}
`[L,U,P] = lu(A)` the LU factorization $PA=LU$
`eig(A)` the eigenvalues of A
`[V,D] = eig(A)` the columns of V are the eigenvectors of A, and
the diagonals `diag(D)` are the eigenvalues of A

Plotting:

`plot(y)` plot y as the y axis, with 1,2,3,… as the x axis
`plot(x,y)` plot y versus x (must have same length)
`plot(x,A)` plot columns of A versus x (must have same # rows)
`loglog(x,y)` plot y versus x on a log-log scale
`semilogx(x,y)` plot y versus x with x on a log scale
`semilogy(x,y)` plot y versus x with y on a log scale
`fplot(@(x) …expression…,[a,b])`
plot some expression in x from $x=a$ to $x=b$
`axis equal` force the x and y axes of the current plot to be scaled equally
`title('A Title')` add a title *A Title* at the top of the plot
`xlabel('blah')` label the x axis as *blah*
`ylabel('blah')` label the y axis as *blah*
`legend('foo','bar')` label 2 curves in the plot *foo* and *bar*
`grid` include a grid in the plot
`figure` open up a new figure window

Transposes and dot products:

`x.', A.'` the transposes of x and A
`x', A'` the complex-conjugate of the transposes of x and A
`x' * y` the dot (inner) product of two *column* vectors x and y
`dot(x,y), sum(x.*y)` …two other ways to write the dot product
`x * y'` the *outer* product of two *column* vectors x and y

http://web.mit.edu/18.06/www/Spring09/matlab-cheatsheet.pdf

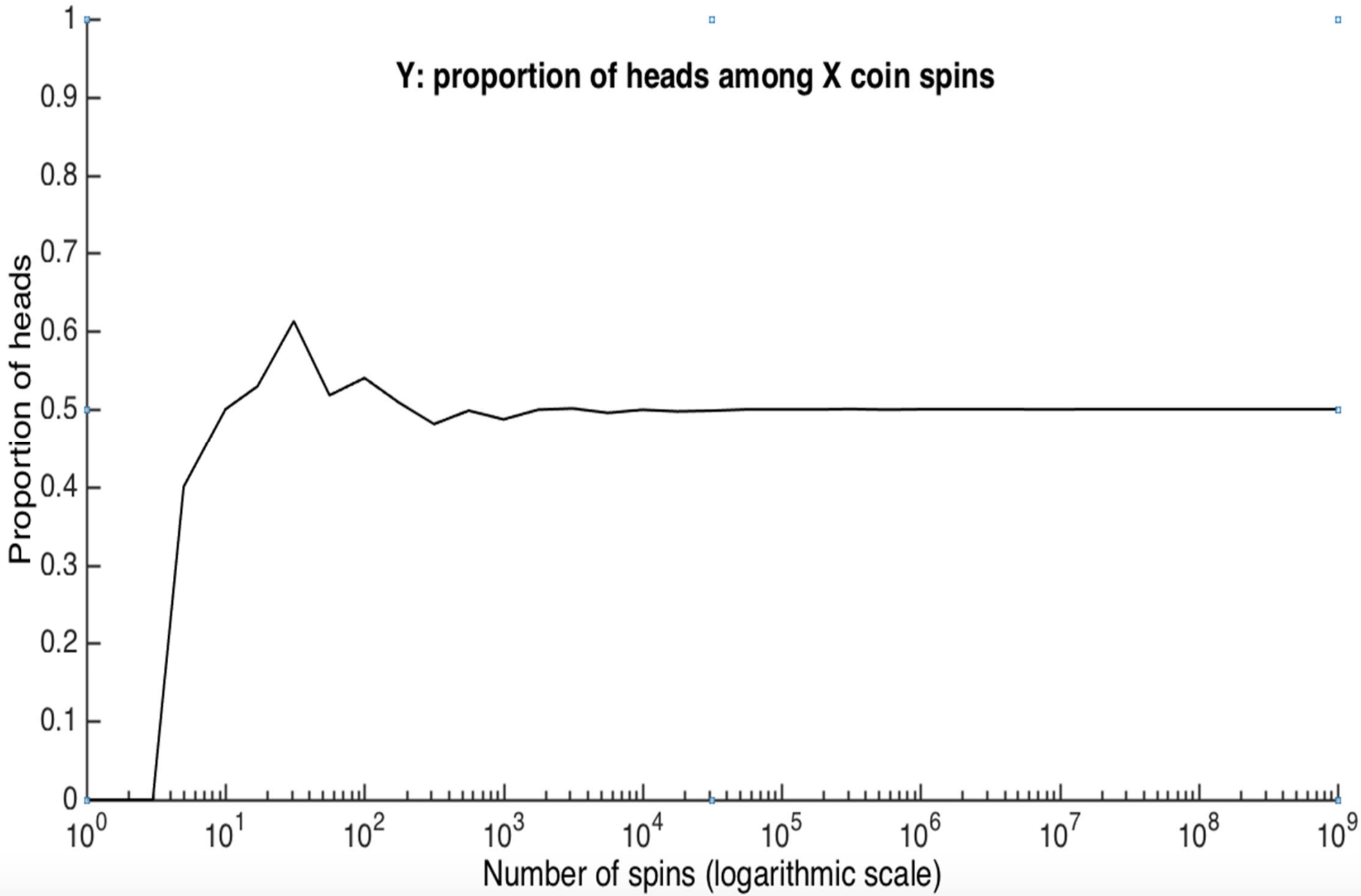
Matlab group exercise

Each table to edit the file coin_toss_template.m (replace all ?? with commands/variables/operations) or writes a new Matlab (Python, R, or anything else) script to:

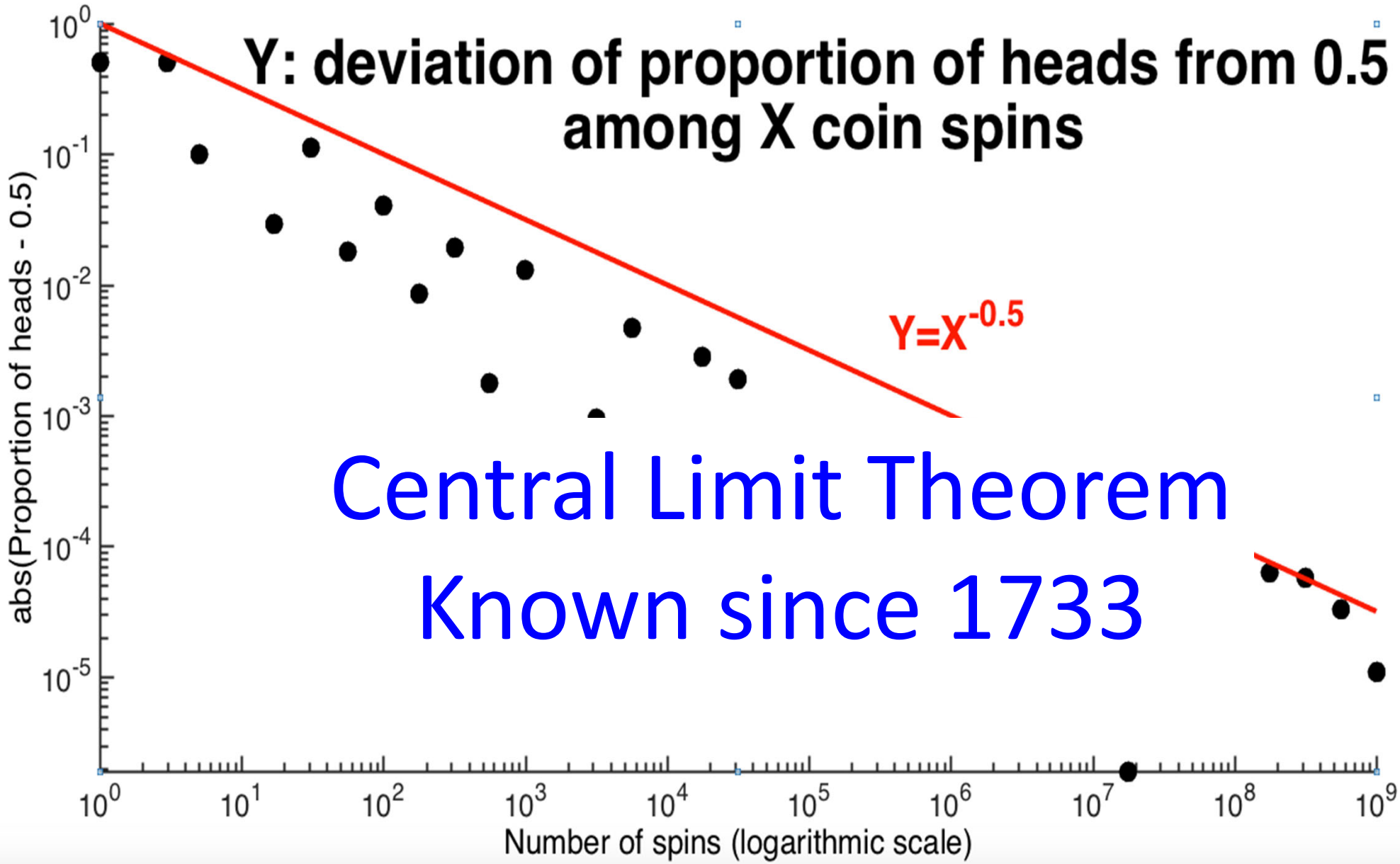
- Simulate a fair coin toss experiment
- Generate multiple tosses of a fair coin:
  1 – heads, 0 - tails
- Calculate the fraction of heads (f_heads(t)) at timepoints:
  t=10; 100; 1000; 10,000; 100,000; 1,000,000;10,000,000 coin tosses
- Plot fraction of heads f_heads(t) vs t with a logarithmic t-axis
- Plot abs(f_heads(t)-0.5) vs t on a log-log plot (both axes are logarithmic)

How I did it

- Stats=1e7;
- r0=rand(Stats,1); r1=floor(2.*r0);
- n_heads(1)=r1(1);
- for t=2:Stats; n_heads(t)=n_heads(t-1)+r1(t); end;
- tp=[1, 10,100,1000, 10000, 100000, 1000000, 10000000]
- np=n_heads(tp); fp=np./tp
- figure; semilogx(tp,fp,'ko-');
- hold on; semilogx([1,10000000],[0.5,0.5],'r--');
- figure; loglog(tp,abs(fp-0.5),'ko-');
- hold on; loglog(tp,0.5./sqrt(tp),'r--');



Proportion of heads among 1,000,000,000 coin tosses
(10^5 more than Kerrich) took me 33 seconds on my Surface Book

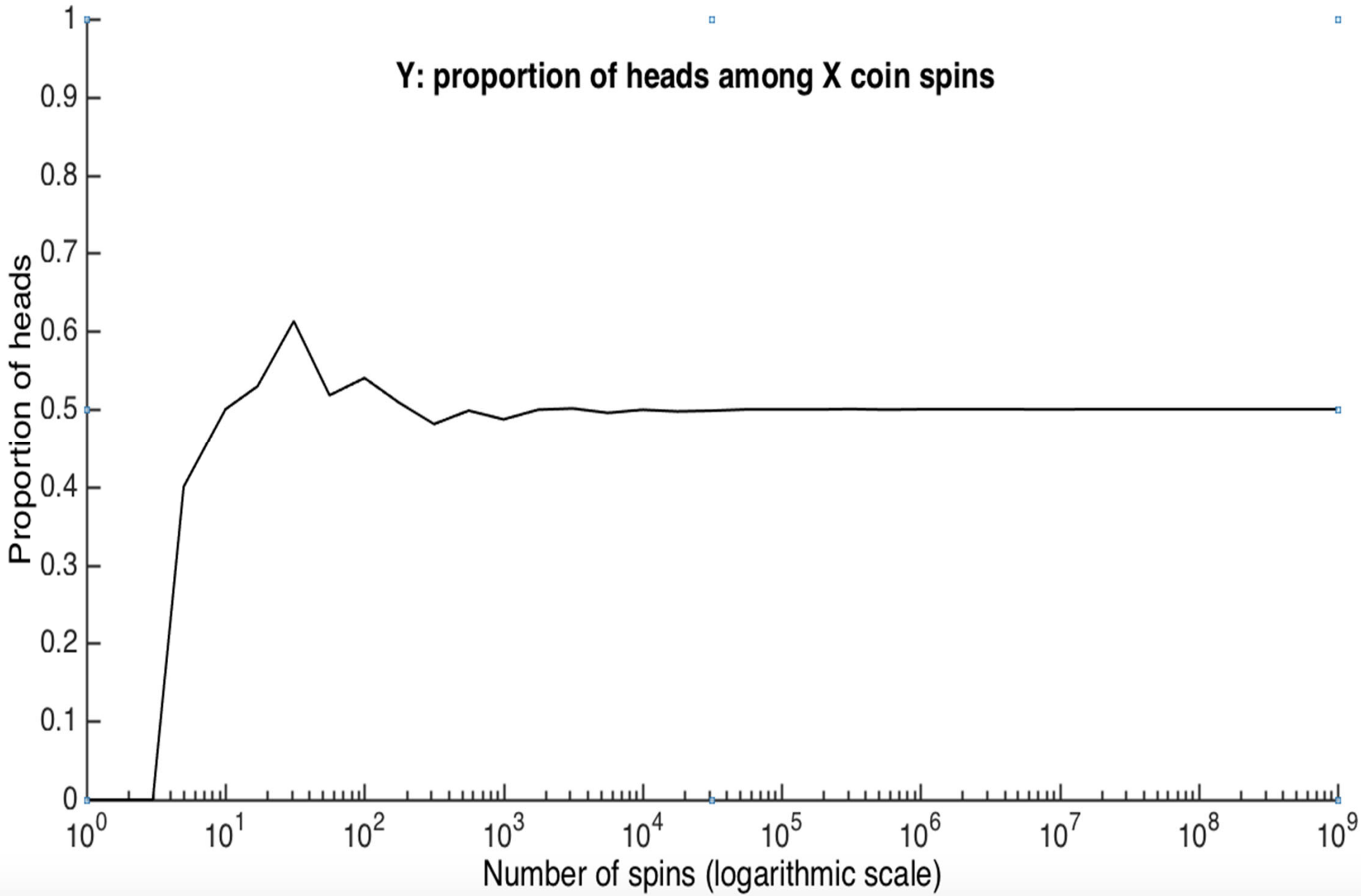


ABS(Proportion of heads-0.5)
among 100,000,000 coin tosses

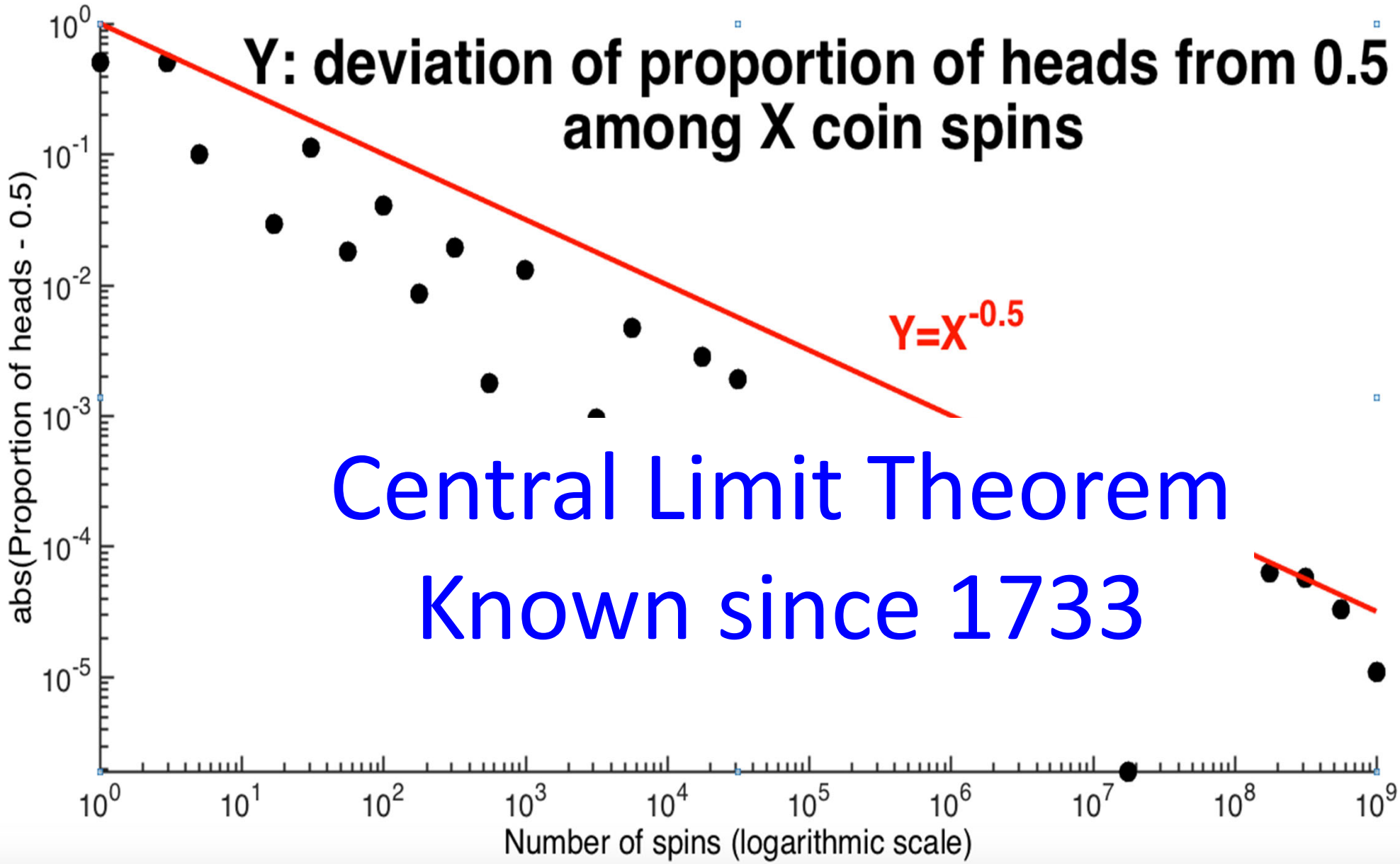
Matlab group exercise

Each table to edit the file coin_toss_template.m (replace all ?? with commands/variables/operations) or writes a new Matlab (Python, R, or anything else) script to:

- Simulate a fair coin toss experiment
- Generate multiple tosses of a fair coin:
  1 – heads, 0 - tails
- Calculate the fraction of heads (f_heads(t)) at timepoints:
  t=10; 100; 1000; 10,000; 100,000; 1,000,000;10,000,000 coin tosses
- Plot fraction of heads f_heads(t) vs t with a logarithmic t-axis
- Plot abs(f_heads(t)-0.5) vs t on a log-log plot (both axes are logarithmic)



Proportion of heads among 1,000,000,000 coin tosses
(10^5 more than Kerrich) took me 33 seconds on my Surface Book



ABS(Proportion of heads-0.5)
among 100,000,000 coin tosses

Definitions of Probability

Two definitions of probability

- (1) STATISTICAL PROBABILITY: the relative frequency with which an event occurs in the long run

- (2) INDUCTIVE PROBABILITY: the degree of belief which it is reasonable to place in a proposition on given evidence

Bulmer, M. G.. Principles of Statistics (Dover Books on Mathematics)

Inductive Probability

An inductive probability of an event the degree of belief which it is rational to place in a hypothesis or proposition on given evidence.

Logical

Principle of indifference

- Principle of Indifference states that two events are equally probable if we have no reason to suppose that one of them will happen rather than the other. (Laplace, 1814)

Pierre-Simon, marquis de Laplace
(1749 –1827)
French mathematician, physicist, astronomer

- Unbiased coin:

probability Heads = probability Tails = ½

- Symmetric die:

probability of each side = 1/6

Inductive = Naïve probability

- If space S is finite and all outcomes are equally likely, then

  $$\text{Prob(Event E)} = \frac{\text{\# of outcomes in E}}{\text{\# of all outcomes in S}}$$

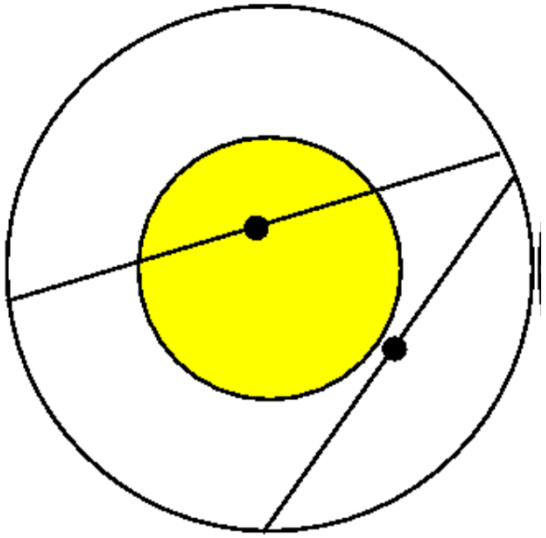
- Can also work with continuous is # is replaced with Area or Volume
- Unbiased coin: Prob(Heads) = Prob(Tails) = 1/2
- Symmetric die: probability of each side = 1/6
- Lottery outcomes are not symmetric: It is not a 50%-50% chance to win or loose in a lottery

Inductive probability can lead to trouble

- Glass contains a mixture of wine and water and proportion of water to wine can be anywhere between 1:1 and 2:1
- (i) We can argue that the proportion of water to wine is equally likely to lie between 1 and 1.5 as between 1.5 and 2.
- (ii) Consider now ratio of wine to water. It is between 0.5 and 1. Based on the same argument it is equally likely in [1/2, 3/4] as it is in [3/4, 1]. But then water to wine ratio is equally likely to lie between 1 and 4/3=1.333… as it is to lie between 1.333.. and 2. This is clearly inconsistent with the previous calculation…
- Paradox solved by clearly defining the experimental design:
  - For (i) use fixed amount of wine (1 liter) and select a uniformly-distributed random number between 1 and 2 for water.
  - For (ii) use 1 liter of water and select uniformly-distributed a random number between 0.5 and 1 for wine.
  - Different experiments – different answers
- Paradox is old. It is attributed to (among others) Joseph Bertrand

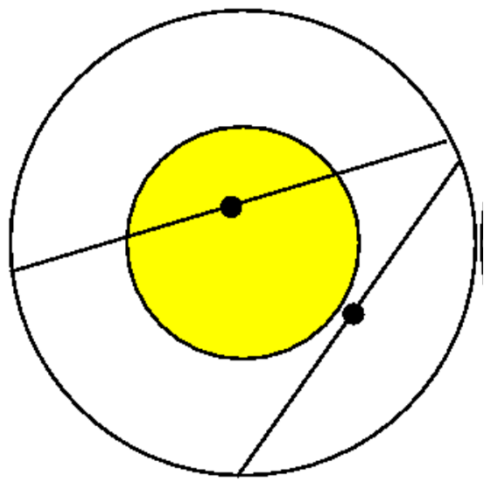
Better known Bertrand's paradox

- Take a circle of radius 2 and randomly draw a line segment through the circle. What is the probability P that the line intersects a concentric circle of radius 1?



Joseph Bertrand
(1822 –1900)
French mathematician

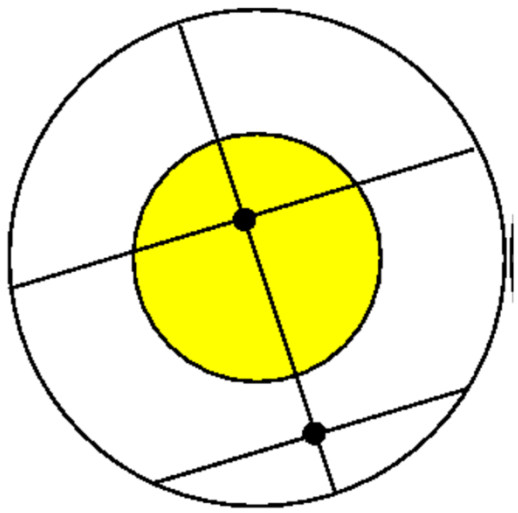
Solution #1



1. **Random point in 2D**: Each line has a unique midpoint, and a line will intersect the inner circle if its midpoint lies inside inner circle. Thus, P = probability that a randomly chosen midpoint lies in the inner circle:

$$P = \frac{\text{Area of the inner circle}}{\text{Area of the outer circle}} = \frac{\pi}{\pi 2^2} = \boxed{\frac{1}{4}}.$$

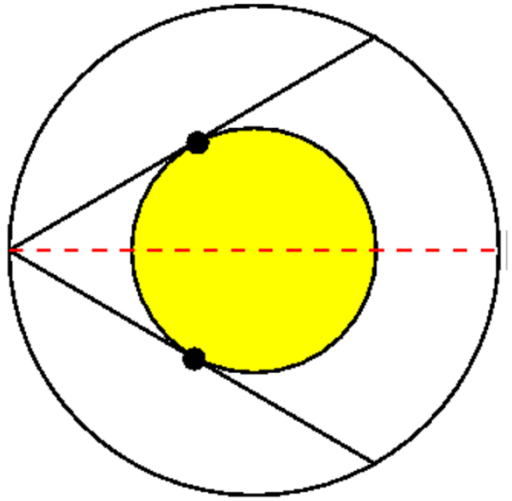
Solution #2



2. **Random point along the diameter**: Each line has a unique perpendicular bisector of length 4. So, P = probability that the midpoint lies on the inner part of the diameter:

$$P = \frac{\text{Length of the inner part of the diameter}}{\text{Length of the diameter}} = \frac{2}{4} = \boxed{\frac{1}{2}}.$$

Solution #3

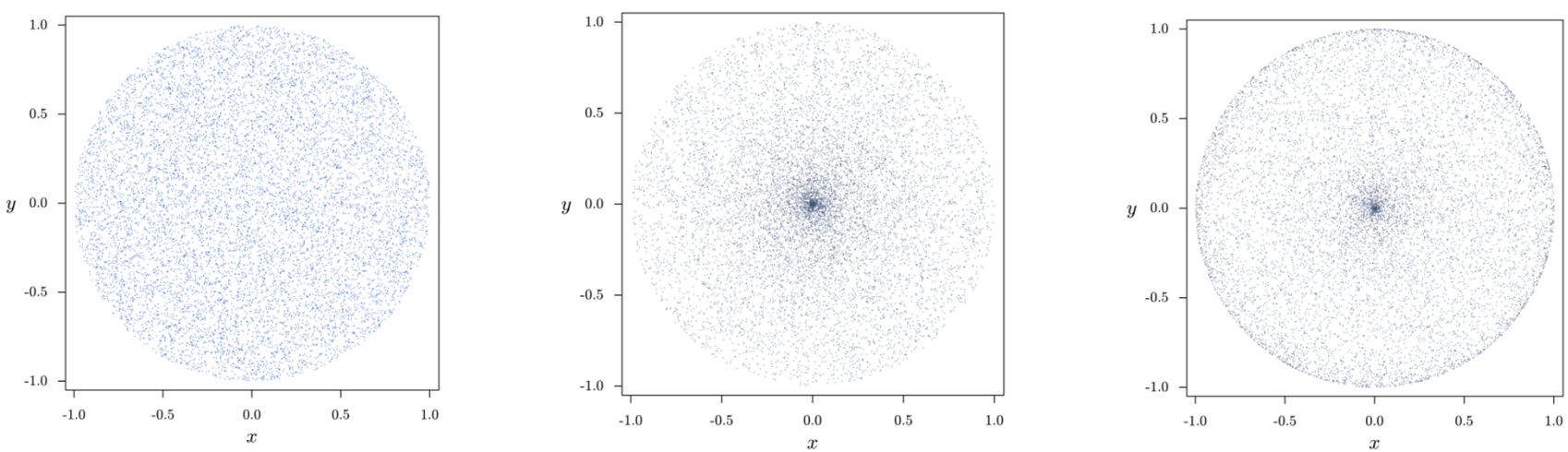


3. **Random angle**: Whether a line intersects the inner circle is determined by the angle it makes with the diameter intersecting the line on the outer circle:

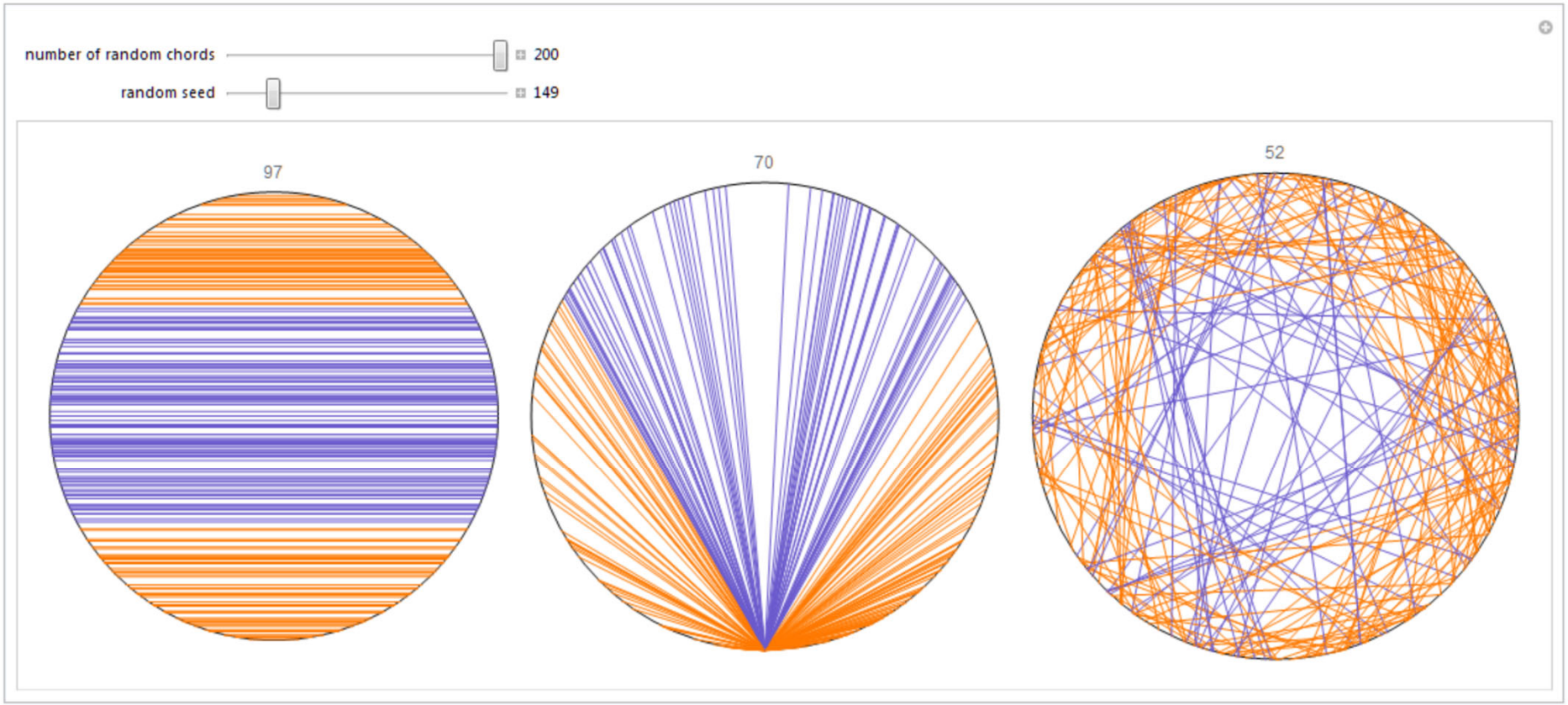
$$P = \frac{\pi/6}{\pi/2} = \boxed{\frac{1}{3}}.$$

So, is probability 1/4, 1/2 , or 1/3 ?

- Depends on how a “random” arc is selected:
  - For #1: select a point inside big circle and then draw an arc with this point as the center. Prob=1/4
  - For #2: select a diameter and a point on this diameter, then draw an arc. Prob=1/2
  - For #3: select a point on the circle and random angle. Prob=1/3



Mathematica visualization



BertrandRandomChordParadox.cdf

I have two children.

One of them is a boy born on Tuesday.

What is the probability I have two boys?

A. 1/2
B. 1/3
C. 2/3
D. 13/27
E. I don't know

Get your i-clickers

Inductive probability
relies on combinatorics
or the art of counting
combinations

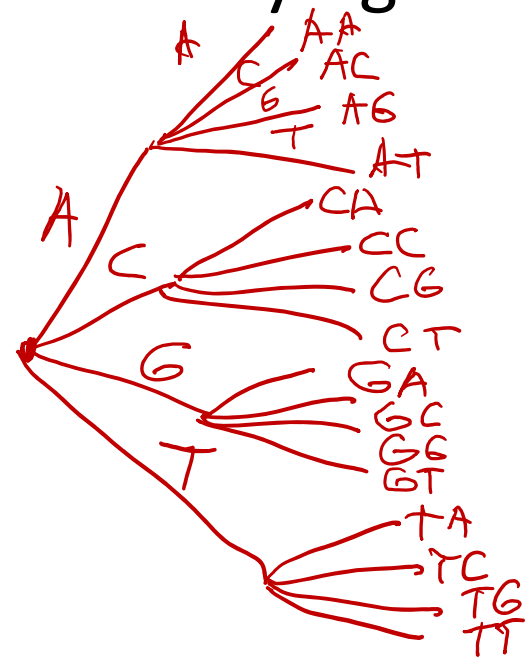
Counting – Multiplication Rule

- Multiplication rule:
  - Let an operation consist of k steps and
    - n_1 ways of completing the step 1,
    - n_2 ways of completing the step 2, … and

    …….
    - n_k ways of completing the step k.
  - Then, the total number of ways of carrying the entire operation is:
    - $n_1 * n_2 * \ldots * n_k$

$n_1 = n_2 = 4$

Example: DNA 2-mer



- $S = \{A, C, G, T\}$ the set of 4 DNA bases
  - Number of k-mers is 4^k=4*4*4…*4 (k –times)
  - There are 4^3=64 triplets in the genetic code
  - There are only 20 amino acids (AA)+1 stop codon
  - There is redundancy: same AA coded by 1-3 codons
  - Evidence of natural selection: "silent" changes of bases are more common than AA changing ones
- A protein-coding part of the gene is typically 1000 bases long
  - There are $4^{1000}= 2^{2000} \sim 10^{600}$ possible sequences of just one gene
  - Or $(10^{600})^{25,000}=10^{15,000,000}$ of 25,000 human genes.
  - For comparison, the Universe has between 10^{78} and 10^{80} atoms and is $4*10^{17}$ seconds old.

Counting – Permutation Rule

- A permutation is a unique sequence of distinct items.
- If $S = \{a, b, c\}$, then there are 6 permutations
  - Namely: abc, acb, bac, bca, cab, cba (order matters)
- # of permutations for a set of n items is n!
- n! (factorial function) = n*(n-1)*(n-2)*…*2*1
- 7! = 7*6*5*4*3*2*1 = 5,040
- By definition: 0! = 1

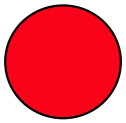
Multiplication and permutation rules are two examples of a general problem, where a sample of size k is drawn from a population of n distinct objects

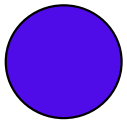
Balls drawn from an urn (or bowl)

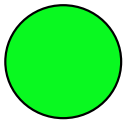
1 ball is red

1 ball is blue

1 ball is green





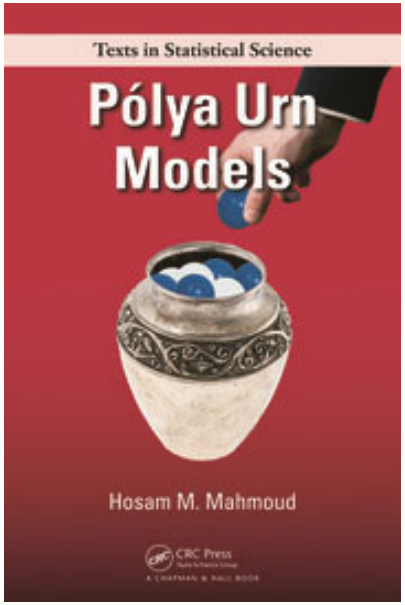


n=3 balls of different colors in an urn from which I draw k=2 balls one at a time

- Do I put each ball back to the bag after drawing it?
  - Yes: problem with replacement
  - No: problem without replacement
- Do I keep track of the order in which balls are drawn?
  - Yes: the order matters
  - No: the order does not matter

George Pólya

- George Pólya (December 13, 1887 – September 7, 1985) was a Hungarian mathematician. He was a professor of mathematics from 1914 to 1940 at ETH Zürich and from 1940 to 1953 at Stanford University. He made fundamental contributions to combinatorics, number theory, numerical analysis and probability theory.



How many ways to choose a sample of K objects out of a population of n objects

	Order matters	Order does not matter
Replace	$n \times n \times n \times \dots \times n = n^K$	$\cancel{\frac{n^K}{K!}}$ not all objects are different
Do not replace	$n \times (n-1) \times (n-2) \times \dots \times (n-K+1) = \frac{n!}{(n-K)!}$	All objects are different $\rightarrow$ $\frac{n!}{(n-K)!} \times \frac{1}{K!} = \binom{n}{K}$

How to solve the problem of K out of n with replacement but where order does not matter?

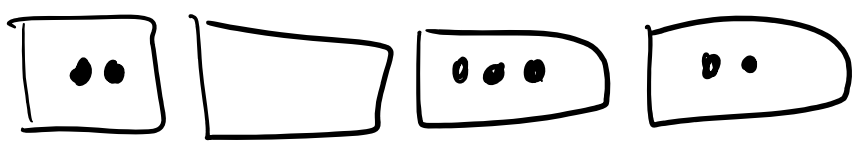
Let's solve $n=2$ problem first:

object 1 object 2

$K=3$

4 possibilities

(1) (2) (3) (4)



$n=4$, $K=7$

•• || ••• | •• K dots, $n-1$ box boundaries

$$\binom{K+n-1}{K} = \frac{(K+n-1)!}{K!\,(n-1)!}$$ ways to distribute

Sampling table

How many ways to choose a sample of k objects out of population of n objects?

	Order matters	Order does not matter
Replacement	$(n)^k$	Difficult: $\binom{n+k-1}{k}=\frac{(n+k-1)!}{(n-1)!k!}$
No replacement	$n(n-1)(n-2)..(n-k+1)=$ $=\frac{n!}{(n-k)!}$	$\binom{n}{k}=\frac{n!}{(n-k)!k!}$

Inductive probability
relies on combinatorics
or the art of counting
combinations

How many ways to choose a sample of K objects out of a population of n objects

	Order matters	Order does not matter
Replace	$n \times n \times n \times \dots \times n = n^K$	$\cancel{\frac{n^K}{K!}}$ not all objects are different
Do not replace	$n \times (n-1) \times (n-2) \times \dots \times (n-K+1) = \frac{n!}{(n-K)!}$	All objects are different $\rightarrow$ $\frac{n!}{(n-K)!} \times \frac{1}{K!} = \binom{n}{K}$

How to solve the problem of K out of n with replacement but where order does not matter?

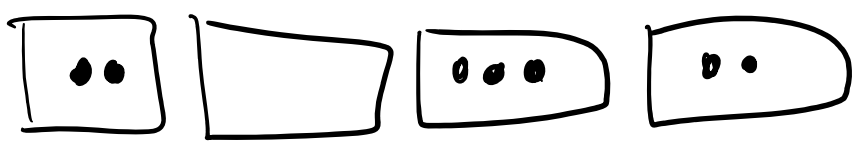
Let's solve $n=2$ problem first:

object 1 object 2

$K=3$

4 possibilities

(1) (2) (3) (4)



$n=4$, $K=7$

•• || ••• | •• K dots, $n-1$ box boundaries

$$\binom{K+n-1}{K} = \frac{(K+n-1)!}{K!\,(n-1)!}$$ ways to distribute

Sampling table

How many ways to choose a sample of k objects out of population of n objects?

	Order matters	Order does not matter
Replacement	$(n)^k$	Difficult: $\binom{n+k-1}{k}=\frac{(n+k-1)!}{(n-1)!k!}$
No replacement	$n(n-1)(n-2)..(n-k+1)=$ $=\frac{n!}{(n-k)!}$	$\binom{n}{k}=\frac{n!}{(n-k)!k!}$

Example

- A DNA of 100 bases is characterized by its numbers of 4 nucleotides: d_A, d_C, d_G, and d_T ($d_A+d_C+d_G+d_T=100$)

- **<u>I don't care about the sequence</u>** (only about the total numbers of A,C,G, and T

- How many distinct combinations of d_A, d_C, d_G, and d_T are out there?

Probability Axioms, Conditional Probability, Statistical (In)dependence, Circuit Problems

Axioms of probability

Probability is a number that is assigned to each member of a collection of events from a random experiment that satisfies the following properties:

If S is the sample space and E is any event in a random experiment,

(1) $P(S) = 1$

(2) $0 \leq P(E) \leq 1$

(3) For two events E_1 and E_2 with $E_1 \cap E_2 = \varnothing$

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

These axioms imply that:

$$P(\varnothing) = 0$$

$$P(E') = 1 - P(E)$$

if the event E_1 is contained in the event E_2

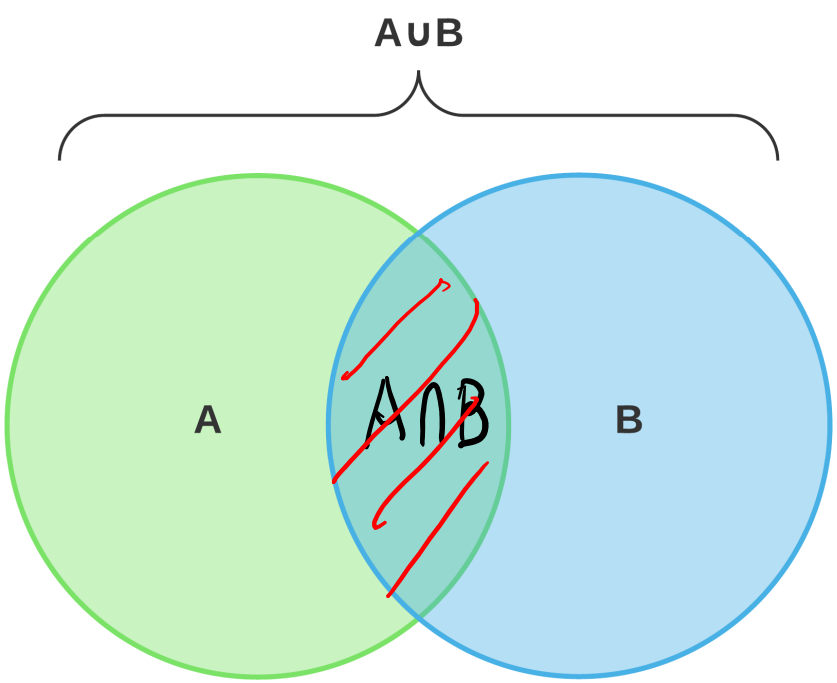
$$P(E_1) \leq P(E_2)$$

Addition rules following from the Axiom (3)

If A and B are mutually exclusive events, i.e. $A \cap B = \emptyset$

$$P(A \cup B) = P(A) + P(B) \tag{2-2}$$

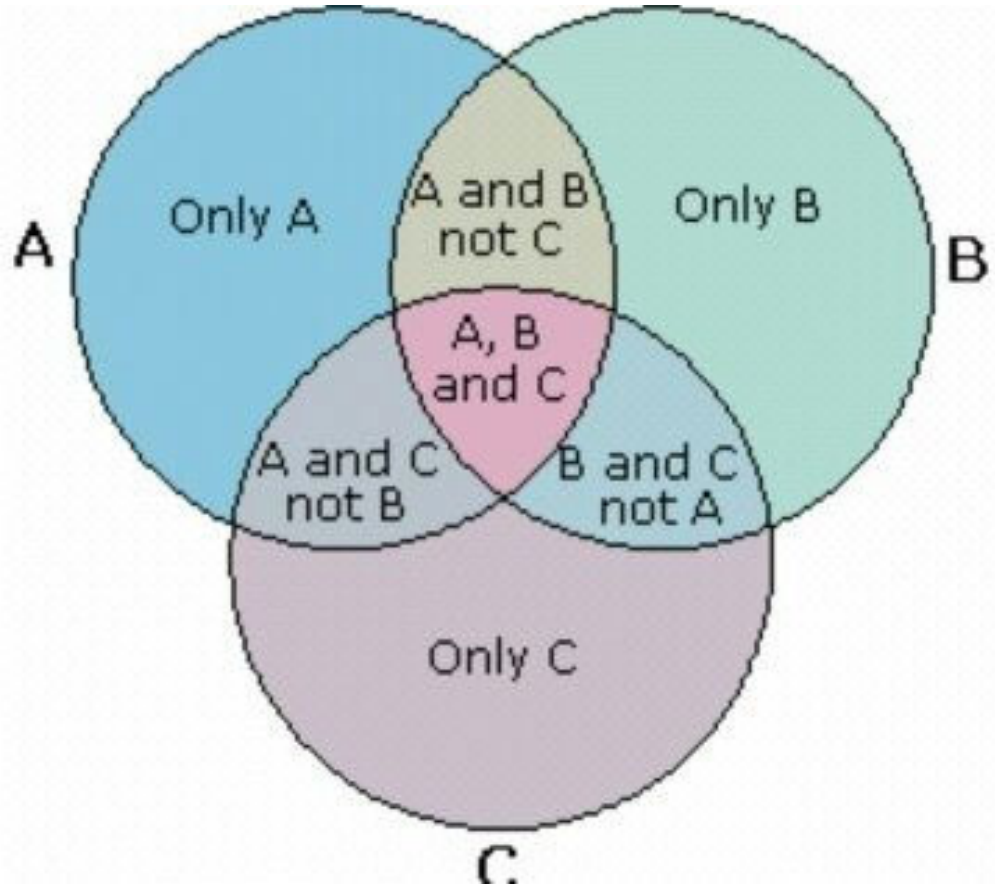
$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{2-1}$$



$$P(A \cup B \cup C) = P(A) + P(B) + P(C) -$$

$$- P(A \cap B) - P(A \cap C) - P(B \cap C) +$$

$$+ P(A \cap B \cap C).$$



Conditional probability

The **conditional probability** of an event B given an event A, denoted as $P(B|A)$, is

$$P(B|A) = P(A \cap B)/P(A)$$

for $P(A) > 0$.

This definition can be understood in a special case in which all outcomes of a random experiment are equally likely. If there are n total outcomes,

$$P(A) = (\text{number of outcomes in } A)/n$$

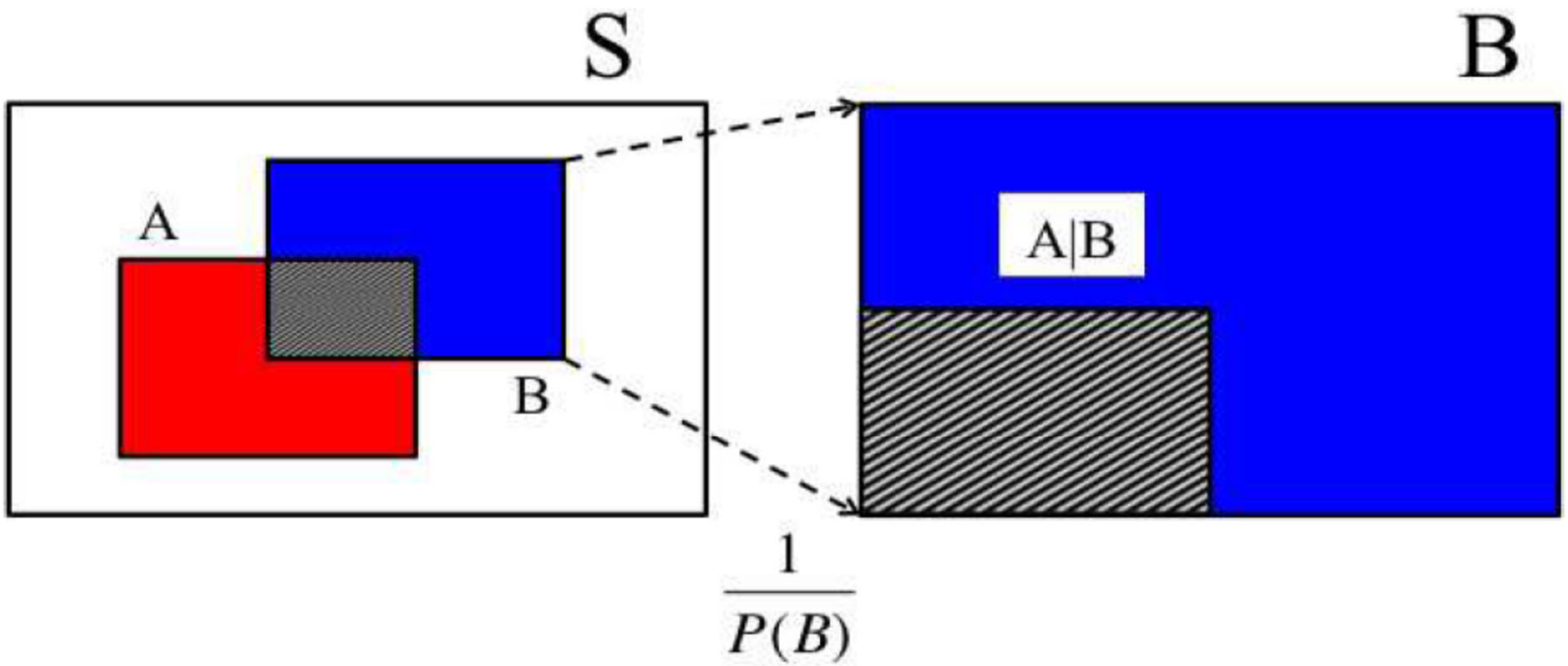
Also,

$$P(A \cap B) = (\text{number of outcomes in } A \cap B)/n$$

Consequently,

$$P(A \cap B)/P(A) = \frac{\text{number of outcomes in } A \cap B}{\text{number of outcomes in } A}$$

Therefore, $P(B|A)$ can be interpreted as the relative frequency of event B among the trials that produce an outcome in event A.


S
A
B
B
A|B
$\frac{1}{P(B)}$

Multiplication rule

is just definition of conditional probability

$$P(B|A) = P(B \cap A)/P(A) \rightarrow$$

$$P(B \cap A) = P(B|A) \cdot P(A)$$

Drake equation

$$N = R^{*} \cdot f_p \cdot n_e \cdot f_l \cdot f_i \cdot f_c \cdot L$$

- N = The number of civilizations in The Milky Way Galaxy whose electromagnetic emissions are detectable.
- R* = The rate of formation of stars suitable for the development of intelligent life.
- f_p = The fraction of those stars with planetary systems.
- n_e = The number of planets, per solar system, with an environment suitable for life.
- f_l = The fraction of suitable planets on which life actually appears.
- f_i = The fraction of life bearing planets on which intelligent life emerges.
- f_c = The fraction of civilizations that develop a technology that releases detectable signs of their existence into space.
- L = The length of time such civilizations release them

Statistically independent events

Always true: $P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$

- **Two events**

Two events are **independent** if any one of the following equivalent statements is true:

(1) $P(A|B) = P(A)$

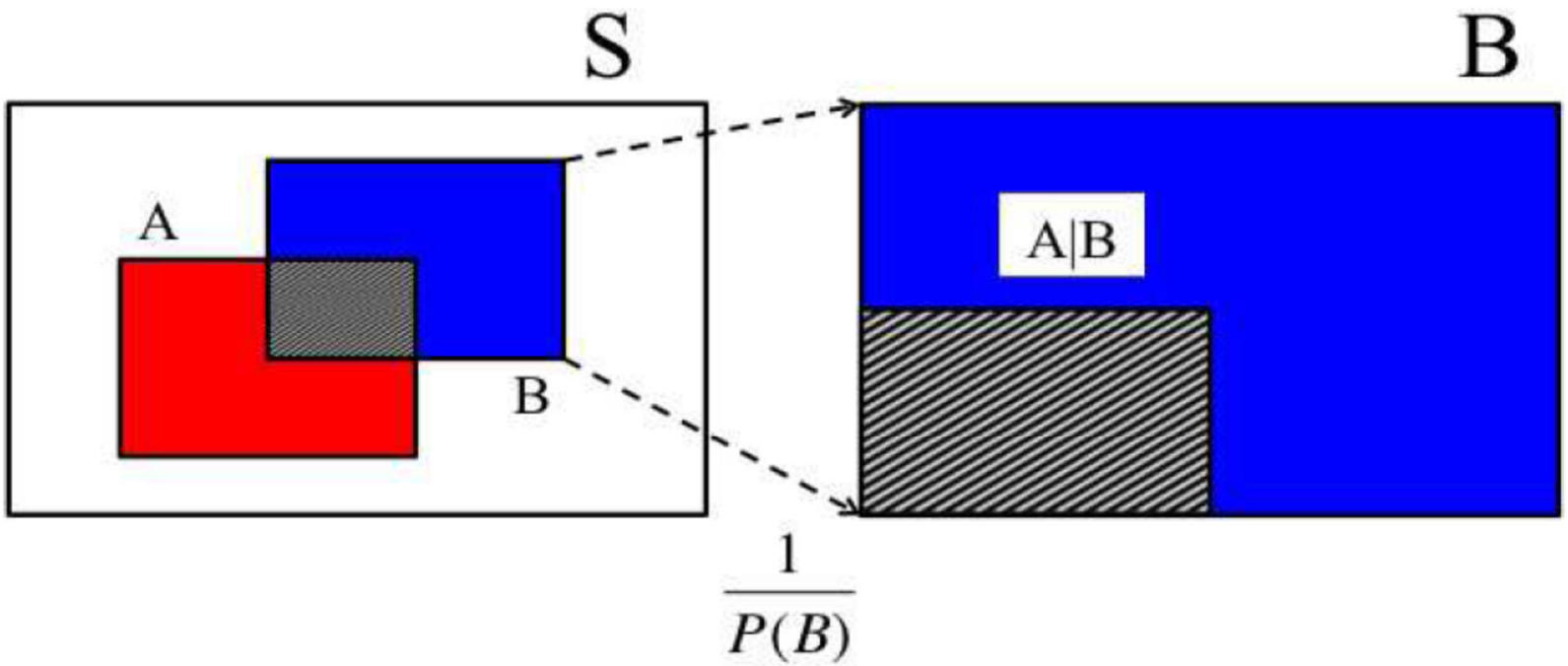
(2) $P(B|A) = P(B)$

(3) $P(A \cap B) = P(A)P(B)$

- **Multiple events**

The events $E_1, E_2, \ldots, E_n$ are independent if and only if for any subset of these events $E_{i_1}, E_{i_2}, \ldots, E_{i_k}$,

$$P(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_k}) = P(E_{i_1}) \times P(E_{i_2}) \times \cdots \times P(E_{i_k})$$


S
A
B
B
A|B
$\frac{1}{P(B)}$

Example 3.10. Let an experiment consist of drawing a card at random from a standard deck of 52 playing cards. Define events A and B as "the card is a ♠" and "the card is a queen." Are the events A and B independent? By definition, $P(A \cdot B) = P(Q\spadesuit) = \frac{1}{52}$. This is the product of $P(\spadesuit) = \frac{13}{52}$ and $P(Q) = \frac{4}{52}$, and events A and B in question are independent. In this situation, intuition provides no help. Now, pretend that the $2\heartsuit$ is drawn and excluded from the deck prior to the experiment. Events A and B become dependent since

$$\mathbb{P}(A) \cdot \mathbb{P}(B) = \frac{13}{51} \cdot \frac{4}{51} \neq \frac{1}{51} = \mathbb{P}(A \cdot B).$$

Credit: XKCD comics

QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

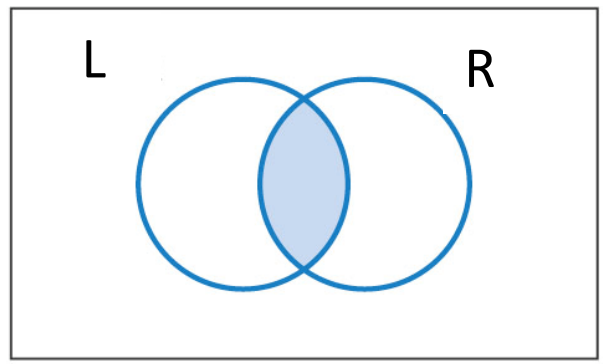
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY ARE THERE FEMALE MR MIMES
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING

WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE THERE SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE HELL IF GOD FORGIVES
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE

Series Circuit

This circuit operates only if there is at least one path of functional devices from left to right. The probability that each device functions is shown on the graph. Assume that the devices fail independently. What is the probability that the circuit operates?

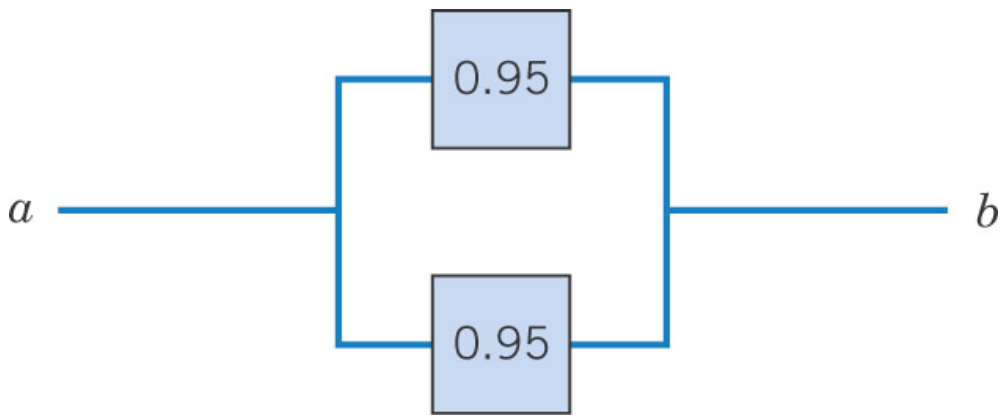


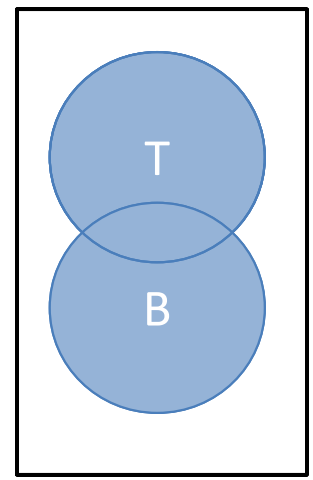
Let L & R denote the events that the left and right devices operate. The probability that the circuit operates is:

$$P(L \text{ and } R) = P(L \cap R) = P(L) * P(R) = 0.8 * 0.9 = 0.72.$$

Parallel Circuit

This circuit operates only if there is a path of functional devices from left to right. The probability that each device functions is shown. Each device fails independently.



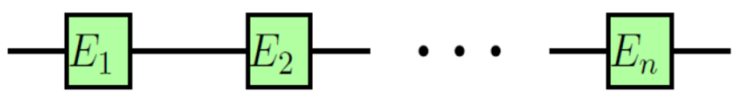


Let T & B denote the events that the top and bottom devices operate. The probability that the circuit operates is:

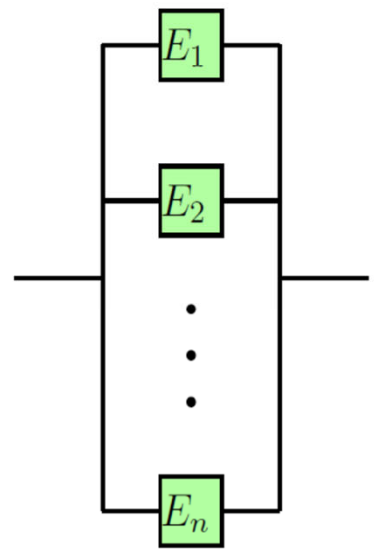
$P(T \cup B) = 1 - P(T' \cap B') = 1 - P(T')*P(B') = 1 - 0.05^2 = 1 - 0.0025 - 0.9975.$

Duality between parallel and series circuits

$$q_i = 1 - p_i.$$



(a)

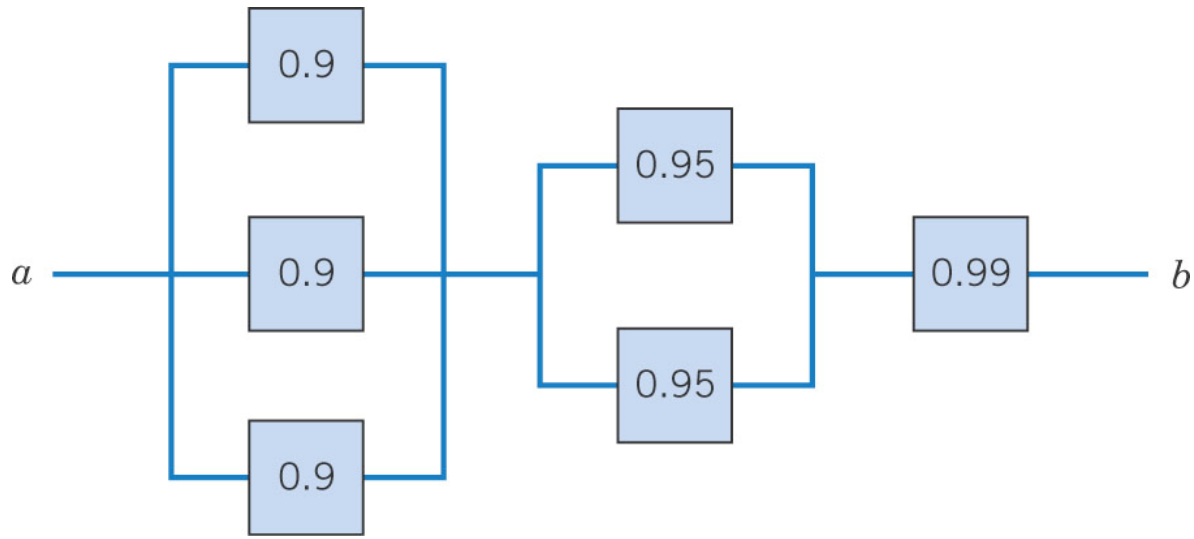


(b)

Connection	Notation	Works with prob	Fails with prob
Serial	$E_1 \cap E_2 \cap \cdots \cap E_n$	$p_1 p_2 \ldots p_n$	$1 - p_1 p_2 \ldots p_n$
Parallel	$E_1 \cup E_2 \cup \cdots \cup E_n$	$1 - q_1 q_2 \ldots q_n$	$q_1 q_2 \ldots q_n$

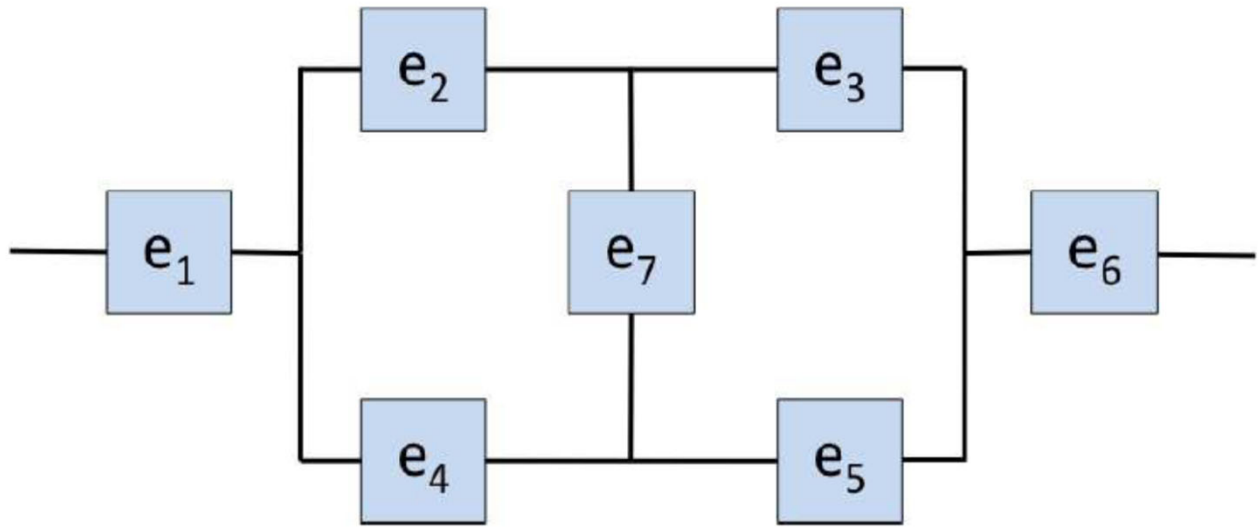
Advanced Circuit

This circuit operates only if there is a path of functional devices from left to right. The probability that each device functions is shown. Each device fails independently.

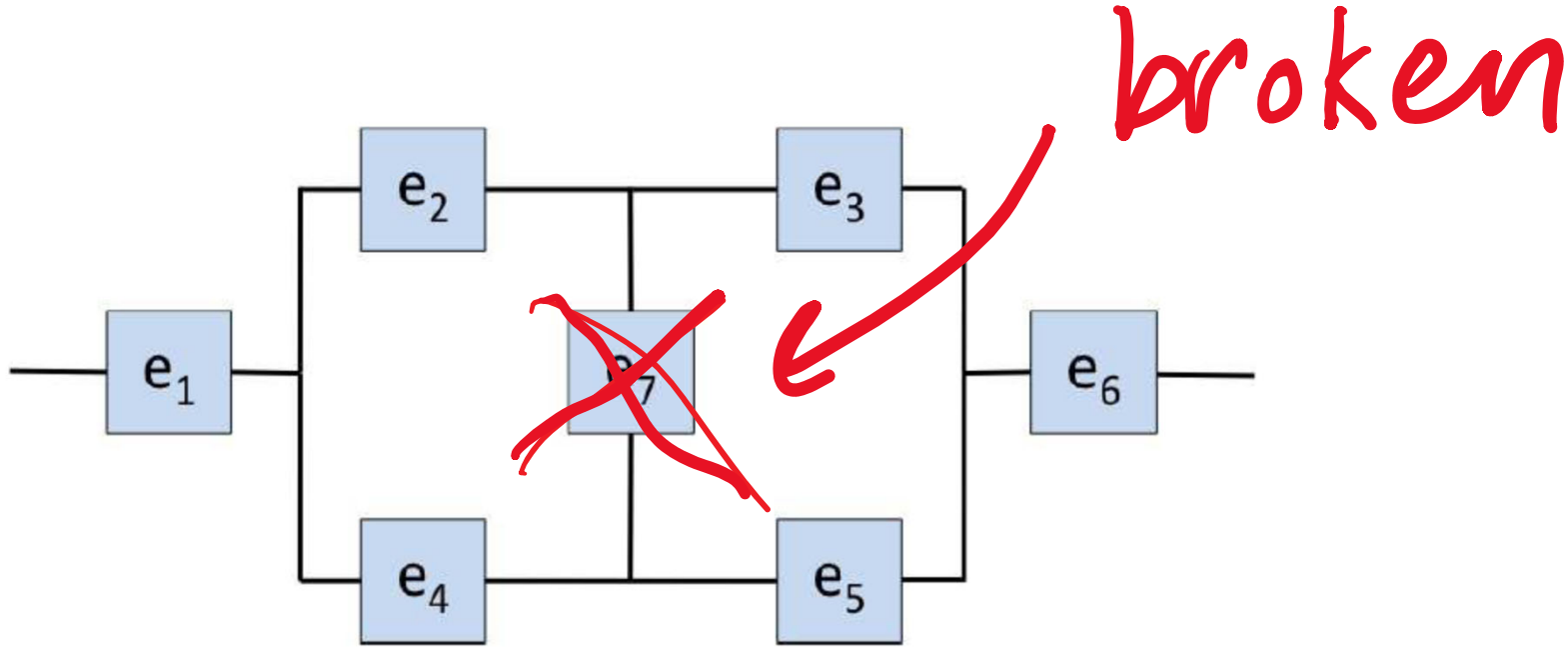


Partition the graph into 3 columns with L & M denoting the left & middle columns.

P(L) = 1- 0.1^3 , and P(M) = 1- 0.05^2, so the probability that the circuit operates is: $(1 – 0.1^3)(1-0.05^2)(0.99) = 0.9875$ (this is a series of parallel circuits).



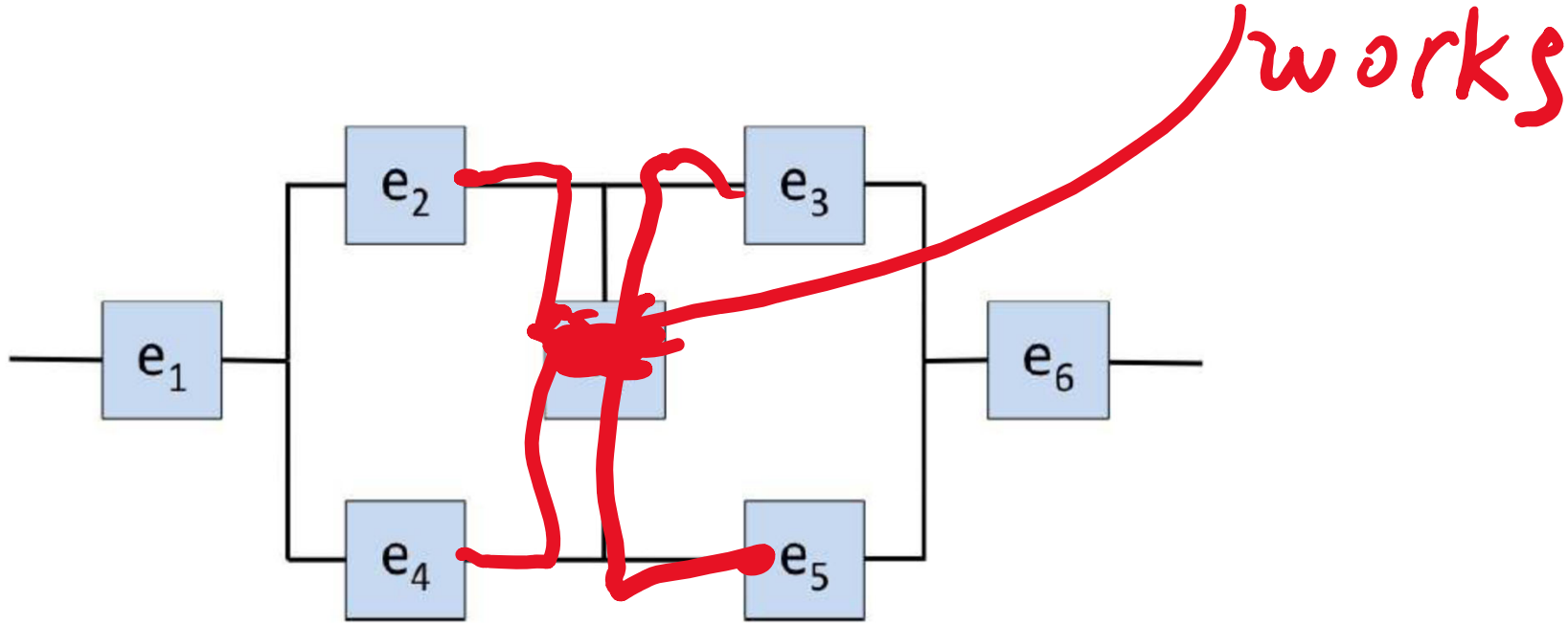
Component	e_1	e_2	e_3	e_4	e_5	e_6	e_7
Probability of component working	0.3	0.8	0.2	0.2	0.5	0.6	0.4



Component	e_1	e_2	e_3	e_4	e_5	e_6	e_7
Probability of component working	0.3	0.8	0.2	0.2	0.5	0.6	0.4

P(circuit works |e7 is broken)=P(e1 works)*
[1-(1-P(e2 works)*P(e3 works))*(1-P(e4 works)*P(e5 works))]*
P(e6 works)=0.3*(1-(1-0.8*0.2)*(1-0.2*0.5))*0.6=0.0439

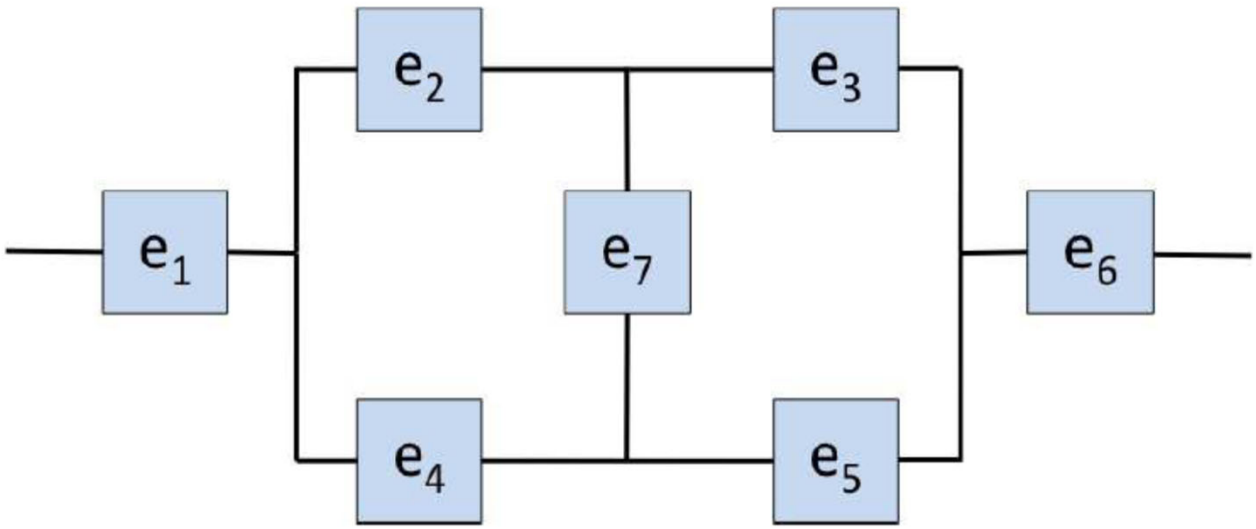
The contribution to total probability:
P(circuit works |e7 is broken)*P(e7 is broken)=0.6*0.0439=0.0264



Component	e_1	e_2	e_3	e_4	e_5	e_6	e_7
Probability of component working	0.3	0.8	0.2	0.2	0.5	0.6	0.4

P(circuit works |e7 works)=P(e1 works)*
[1-(1-P(e2 works))*(1-P(e3 works))]
[1-(1-P(e4 works))(1-P(e5 works))]*
P(e6 works)=0.3*(1-(1-0.8)*(1-0.2))*(1-(1-0.2)*(1-0.5)))*0.6=0.0907

The contribution to total probability:
P(circuit works |e7 works)*P(e7 works)=0.4*0.0907=0.0363

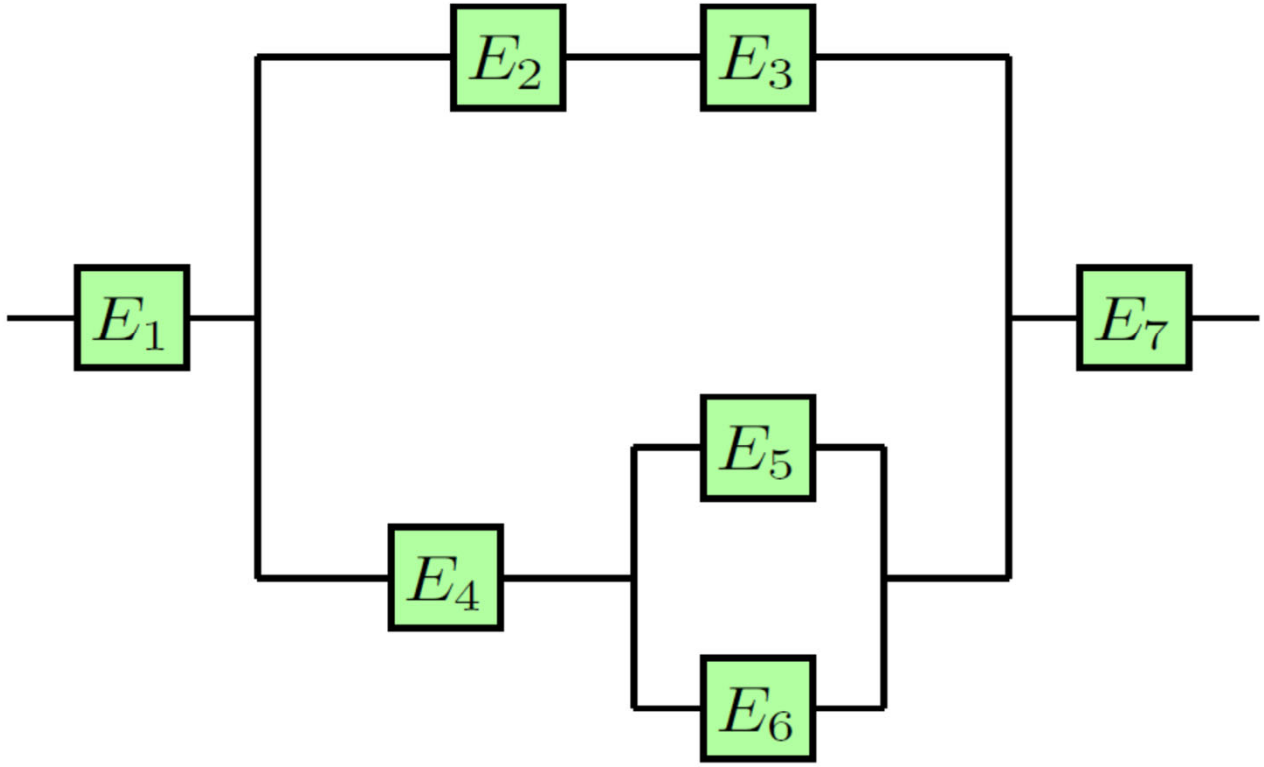


Component	e_1	e_2	e_3	e_4	e_5	e_6	e_7
Probability of component working	0.3	0.8	0.2	0.2	0.5	0.6	0.4

P(circuit works)=
P(circuit works |e7 works)*P(e7 works)+
P(circuit works |e7 is broken)*P(e7 is broken)=
=0.0264+0.0363=0.0627

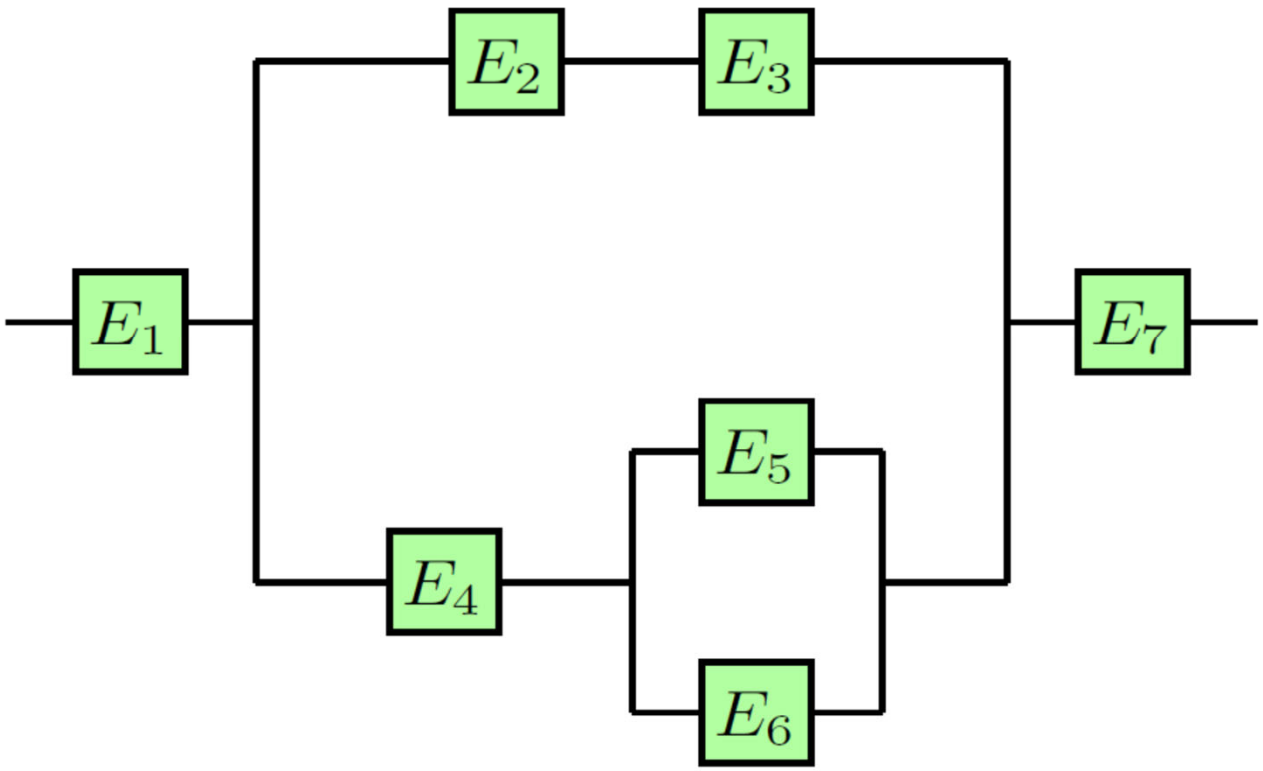
Answer: 6.27%

Circuit → Set equation



Component	E_1	E_2	E_3	E_4	E_5	E_6	E_7
Probability of functioning well	0.9	0.5	0.3	0.1	0.4	0.5	0.8

Circuit → Set equation



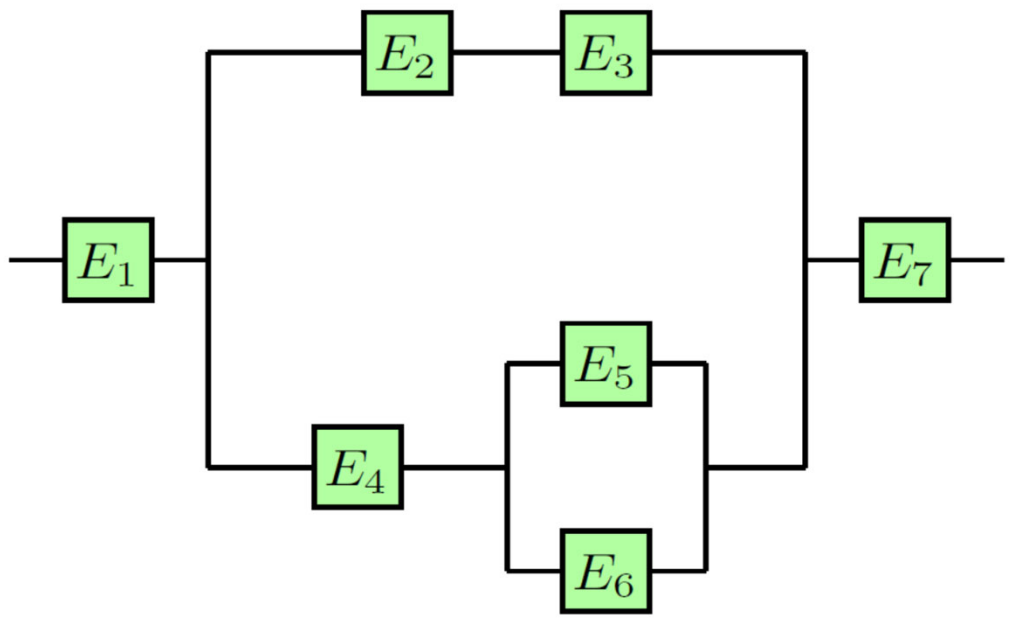
Component	E_1	E_2	E_3	E_4	E_5	E_6	E_7
Probability of functioning well	0.9	0.5	0.3	0.1	0.4	0.5	0.8

$$E_1 \cap [(E_2 \cap E_3) \cup (E_4 \cap (E_5 \cup E_6))] \cap E_7.$$

P(Works) = 0.9.*(1-(1-0.5.*0.3).*(1-0.1.*(1-0.6.*0.5))).*0.8=0.15084

Matlab group exercise

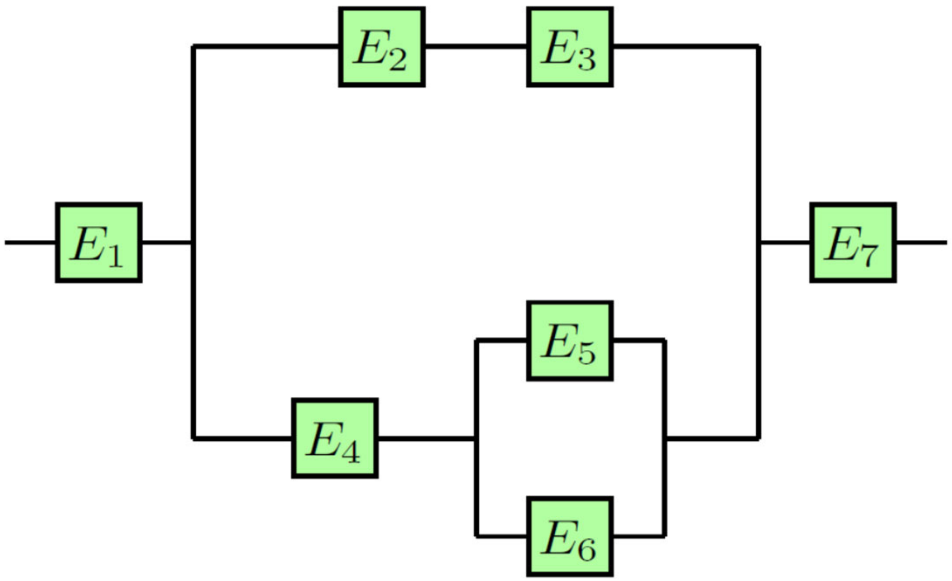
- Test our result for this circuit.
- Use circuit_template.m on the website



Component	E_1	E_2	E_3	E_4	E_5	E_6	E_7
Probability of functioning well	0.9	0.5	0.3	0.1	0.4	0.5	0.8

Matlab group exercise

- Test our result for this circuit.
- Download circuit_template.m from the website



Component	E_1	E_2	E_3	E_4	E_5	E_6	E_7
Probability of functioning well	0.9	0.5	0.3	0.1	0.4	0.5	0.8

P(Works) = 0.9.*(1-(1-0.5.*0.3).*(1-0.1.*(1-0.6.*0.5))).*0.8=0.15084

Credit: XKCD comics

QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE

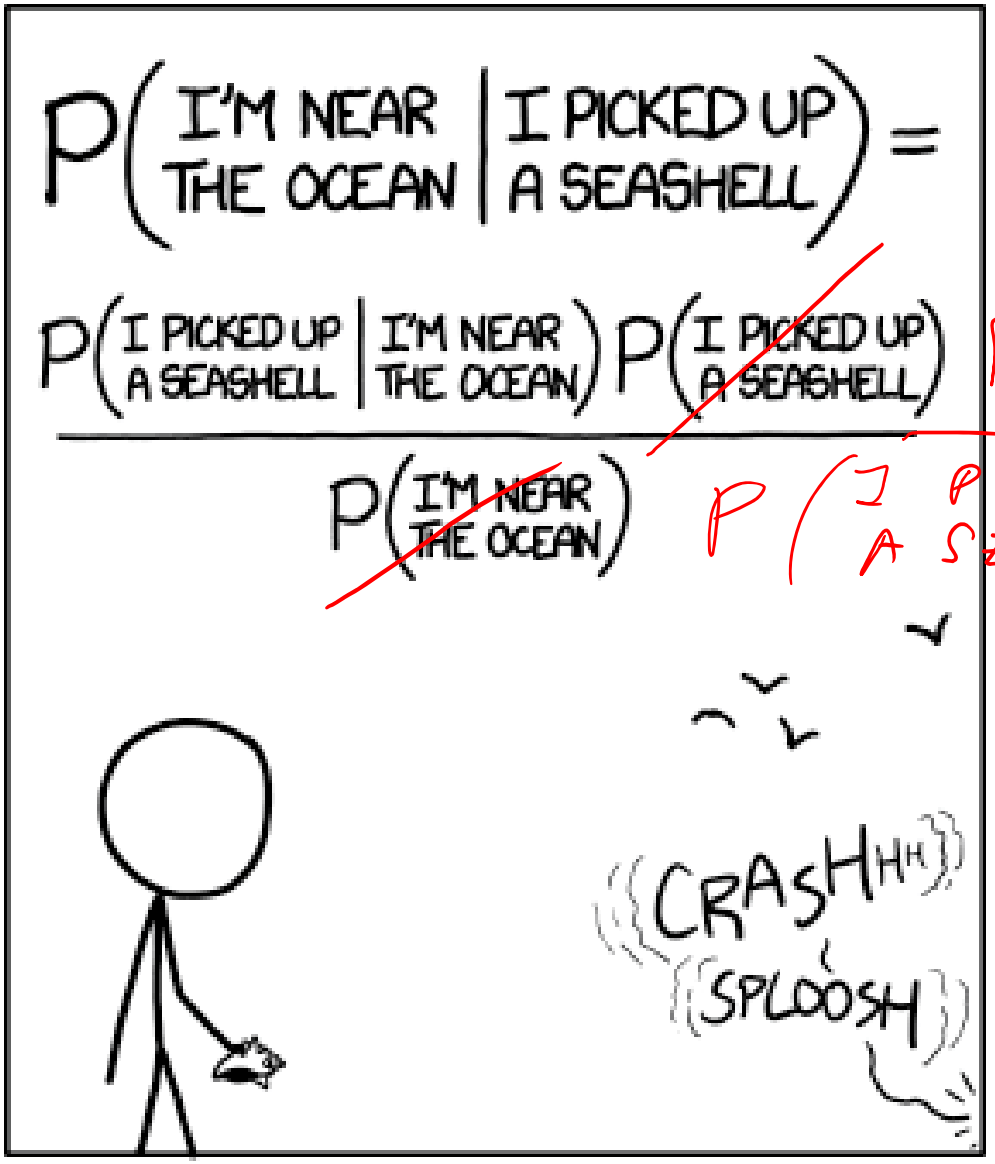
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY ARE THERE FEMALE MR MIMES
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING

WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE THERE SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE HELL IF GOD FORGIVES
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE

Reminder:
Conditional probability



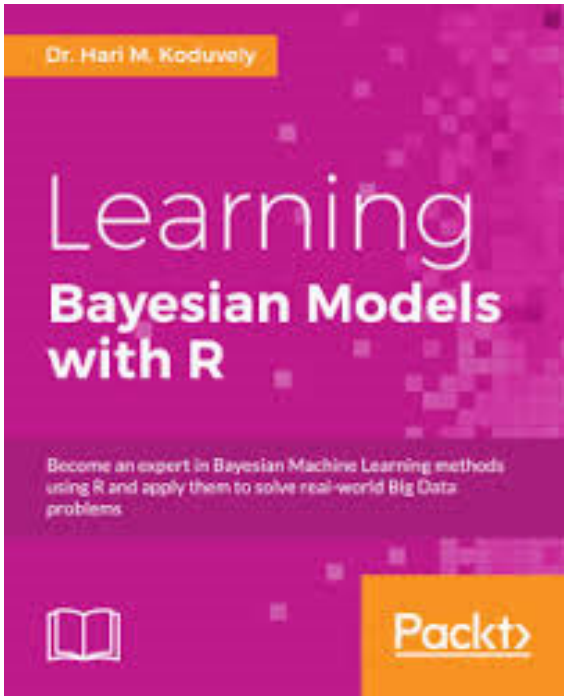
STATISTICALLY SPEAKING, IF YOU PICK UP A SEASHELL AND *DON'T* HOLD IT TO YOUR EAR, YOU CAN PROBABLY HEAR THE OCEAN.

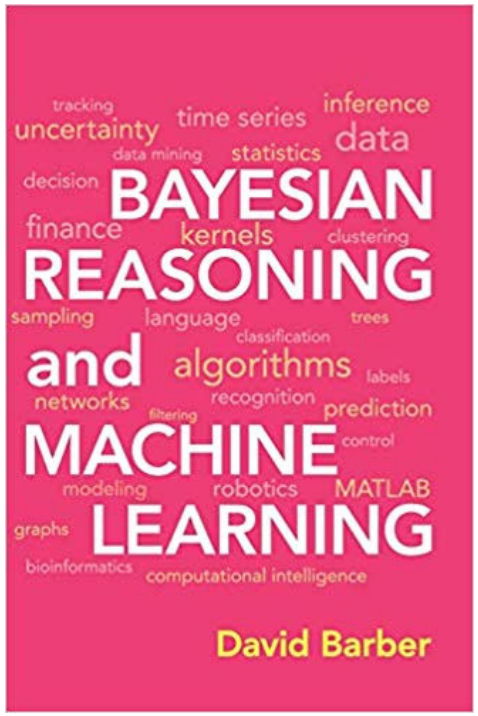
What is wrong in this comics?

If you are not yet reading XKCD comics https://xkcd.com/ you should start

Bayes Theorem

Bayes' theorem





Thomas Bayes (1701-1761)
English statistician, philosopher, and Presbyterian minister

Bayes' theorem was presented in "An Essay towards solving a Problem in the Doctrine of Chances" which was read to the Royal Society in 1763 already after Bayes' death.

Bayes' theorem (simple)

$$P(A \cap B) = \underline{P(A|B)P(B)} = P(B \cap A) = \underline{P(B|A)P(A)}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- In Science we often want to know:
  "How much faith should I put into hypothesis, given the data?"
  or $P(H|D)$ (see also the inductive definition of probability)
- What we usually can calculate if the hypothesis/model is OK:
  "Assuming that this hypothesis is true, what is the probability of the observed data?" or $P(D|H)$
- Bayes' theorem can help: $P(H|D)=P(D|H)\bullet P(H)/P(D)$
- The problem is $P(H)$ (so-called prior) is often not known

Bayes' theorem (continued)

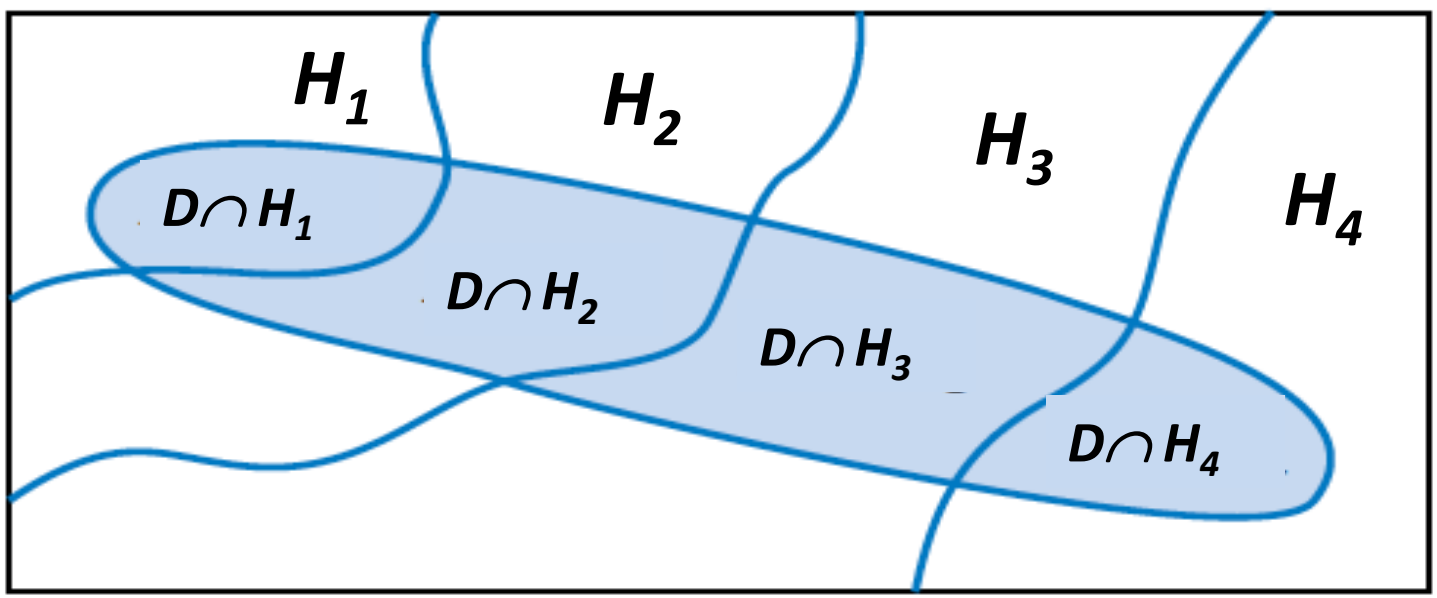
Works best with exhaustive and mutually-exclusive hypotheses:

$H_1, H_2, \ldots H_n$ such that $H_1 \cup H_2 \cup H_3 \ldots \cup H_n = S$ and $H_i \cap H_j = \emptyset$ for $i \neq j$

$P(H_k|D) = P(D|H_k) \bullet P(H_k)/P(D)$

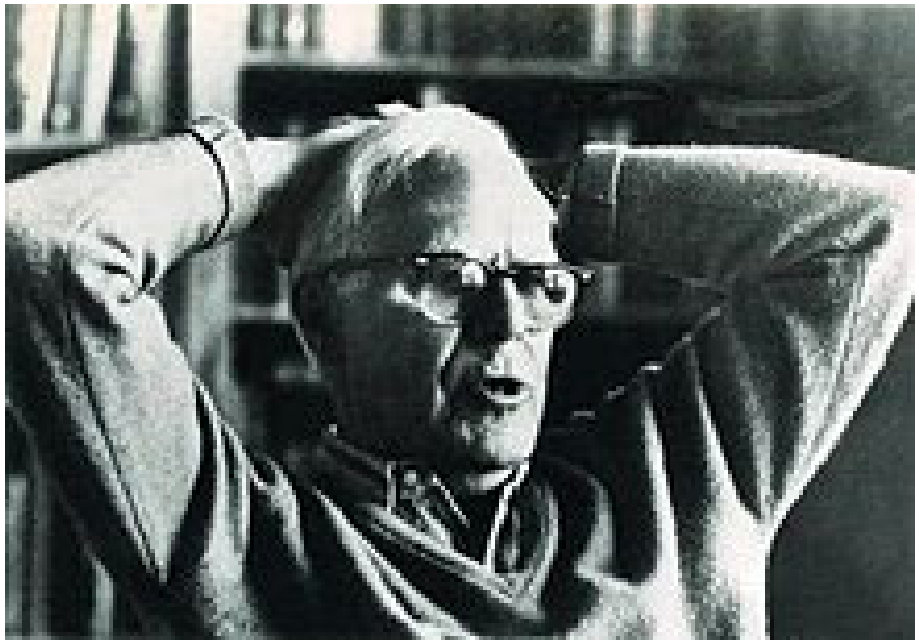
where:

$P(D) = P(D|H_1) \bullet P(H_1) + P(D|H_2) \bullet P(H_2) + \ldots P(D|H_n) \bullet P(H_n)$



Secretary problem

- An employer has a known number – n – of applicants for a secretary position, whom are interviewed one at a time
- Employer can easily evaluate and rank applicants relative to each other but has no idea of the overall distribution of their quality
- Employer has only one chance to choose the secretary: gives yes/no answer in the end of each interview and cannot go back to rejected applicants
- How can employer maximize the probability to choose the best secretary among all applicants?



Martin Gardner (1914 – 2010)
Described the “secretary problem” in Scientific American 1960.
was an American popular mathematics and popular science writer. Best known for “recreational mathematics”: He was behind the “Mathematical Games” section in Scientific American.

Eugene Dynkin (1924 – 2014)
solved this problem in 1963.
He referred to it as a “picky bride problem”
was a Soviet and later American mathematician, member of the US National Academy of Science. He has made contributions to the fields of probability and algebra. The Dynkin diagram, the Dynkin system, and Dynkin's lemma are all named after him.

Who solved the secretary problem?

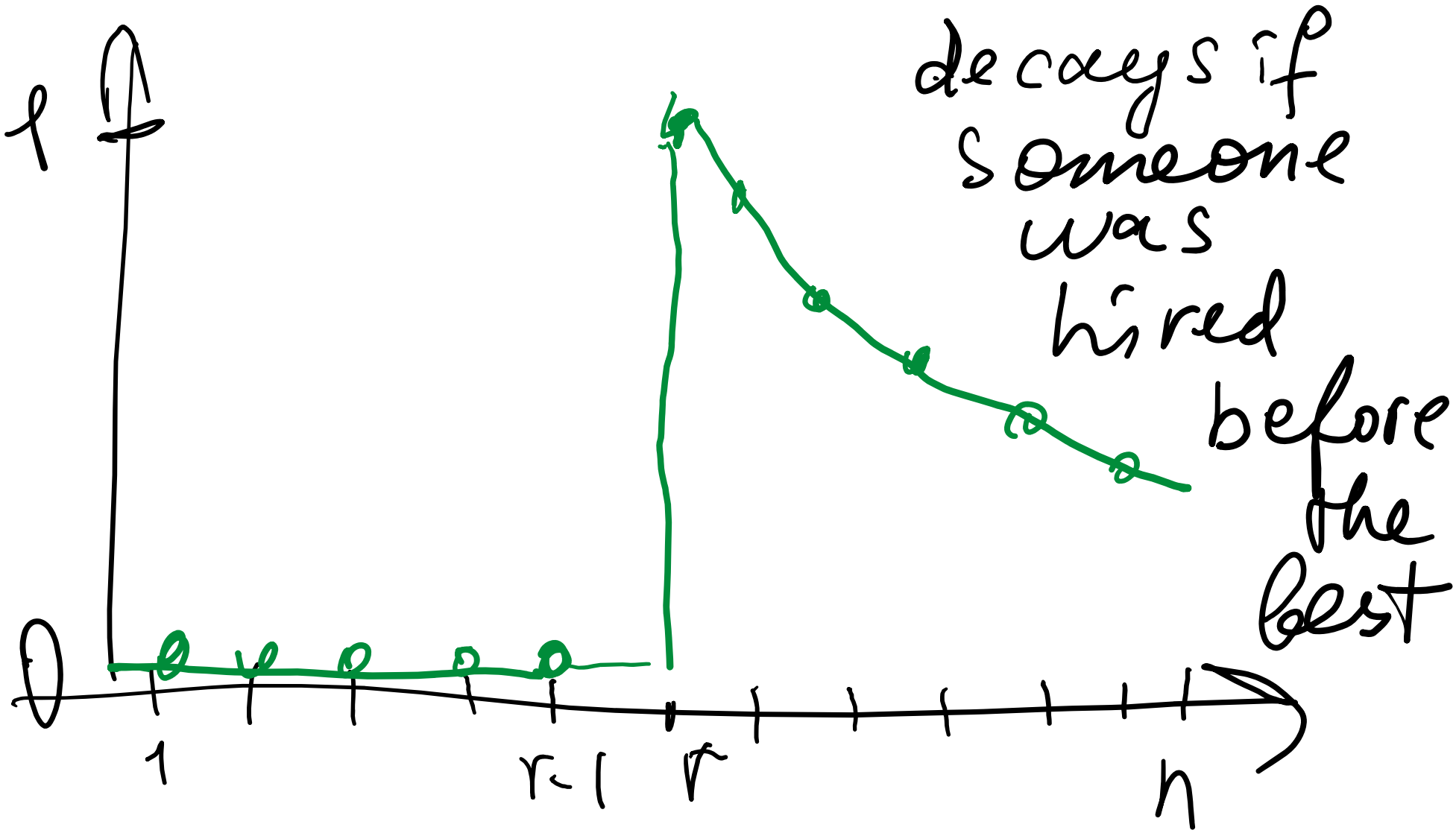
- Gardner outlined the solution in Sci Am 1960 but gave no formal proof
- Solution by Lindey was published in 1961:
  Lindey, D. V. (1961). Dynamic programming and decision theory. Appl. Statist. 10 39-51
- Dynkin's paper was published in 1963:
  Dynkin, E. B. (1963). The optimum choice of the instant for stopping a Markov process. Soviet Math. Dokl. 4 627-629
- When the celebrated German astronomer, Johannes Kepler (1571-1630), lost his first wife to cholera in 1611, he set about finding a new wife
- He spent 2 years on the process, had 11 candidates and married the 5th candidate (11/e~4 so he married the first after)

Thomas S Ferguson, Statistical Science 1989, Who Solved the Secretary Problem?

What should the employer do?

- Employer does not know the distribution of the quality of applicants and has to learn it on the fly
- Algorithm: look at the first ***r-1*** applicants, remember the best among them
- Hire the first among next ***n-r+1*** applicants who is better than the best among the first ***r*** applicants
- How to choose ***r***?
- When ***r*** is too small – not enough information: the best among *r* is not very good. You are likely to hire a bad secretary
- When ***r*** is too large (e.g. ***r=n-1***) – you procrastinated for too long! You have almost all the information, but you will have to hire the last applicant who is (likely) not particularly good

Probability of hiring the best candidate if he/she has #i in the queue



Look at $i-1$ candidates before the best

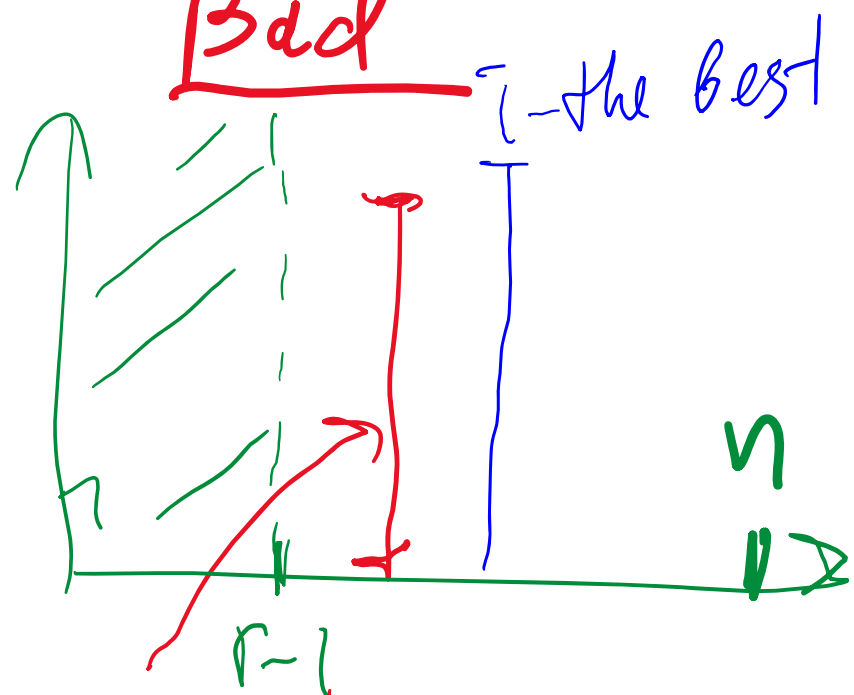
$$\text{Prob} = \frac{r-1}{i-1}$$

$$\text{Prob} = \frac{i-r}{i-1}$$

Good

the best among $i-1$

Bad



the best among $i-1$

$$
\begin{aligned}
P(r) &= \sum_{i=1}^{n} P\left(\text{applicant } i \text{ is selected} \cap \text{applicant } i \text{ is the best}\right) \\
&= \sum_{i=1}^{n} P\left(\text{applicant } i \text{ is selected} \middle| \text{applicant } i \text{ is the best}\right) \times P\left(\text{applicant } i \text{ is the best}\right) \\
&= \left[\sum_{i=1}^{r-1} 0 + \sum_{i=r}^{n} P\left(\begin{matrix}\text{the best of the first } i-1 \text{ applicants} \\ \text{is in the first } r-1 \text{ applicants}\end{matrix} \middle| \text{applicant } i \text{ is the best}\right)\right] \times \frac{1}{n} \\
&= \sum_{i=r}^{n} \frac{r-1}{i-1} \times \frac{1}{n} \quad = \quad \frac{r-1}{n} \sum_{i=r}^{n} \frac{1}{i-1}.
\end{aligned}
$$

$$P(r) = \frac{r-1}{n} \sum_{i=r}^{n} \frac{1}{i-1}.$$

Letting *n* tend to infinity, writing x as the limit of *r*/*n*, using *t* for *i*/*n* and *dt* for 1/*n*,

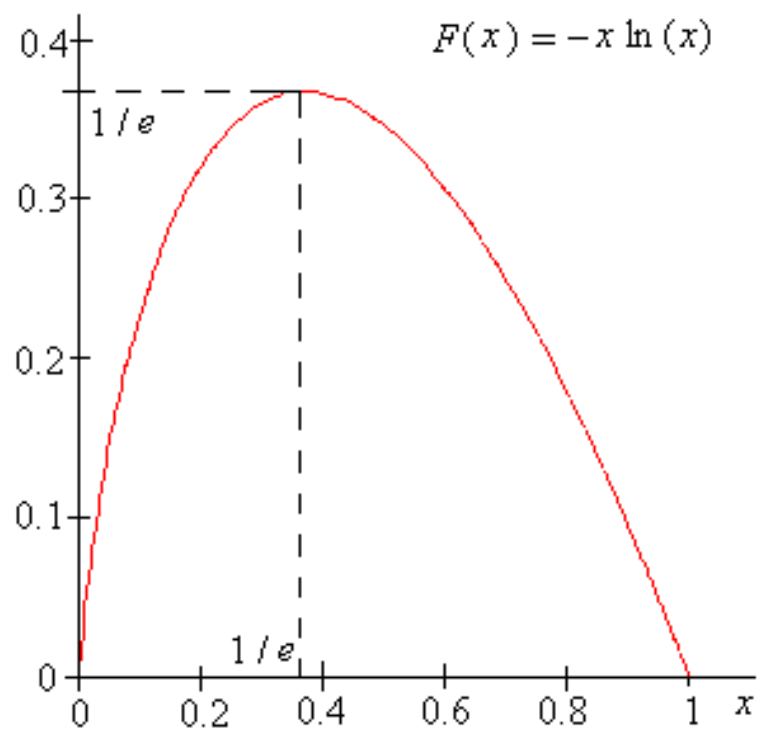
$$P(x) = x \int_x^1 \frac{1}{t}\, dt = -x \ln(x)\ .$$

dP(x)/dx=-ln(x)-1

-ln(x)-1=0*

x=1/e=0.3679*

Probability of picking the best applicant is also *1/e=0.3679*



Credit: XKCD comics

QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY ARE THERE FEMALE MR MIMES
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING

WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE THERE SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE HELL IF GOD FORGIVES
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE

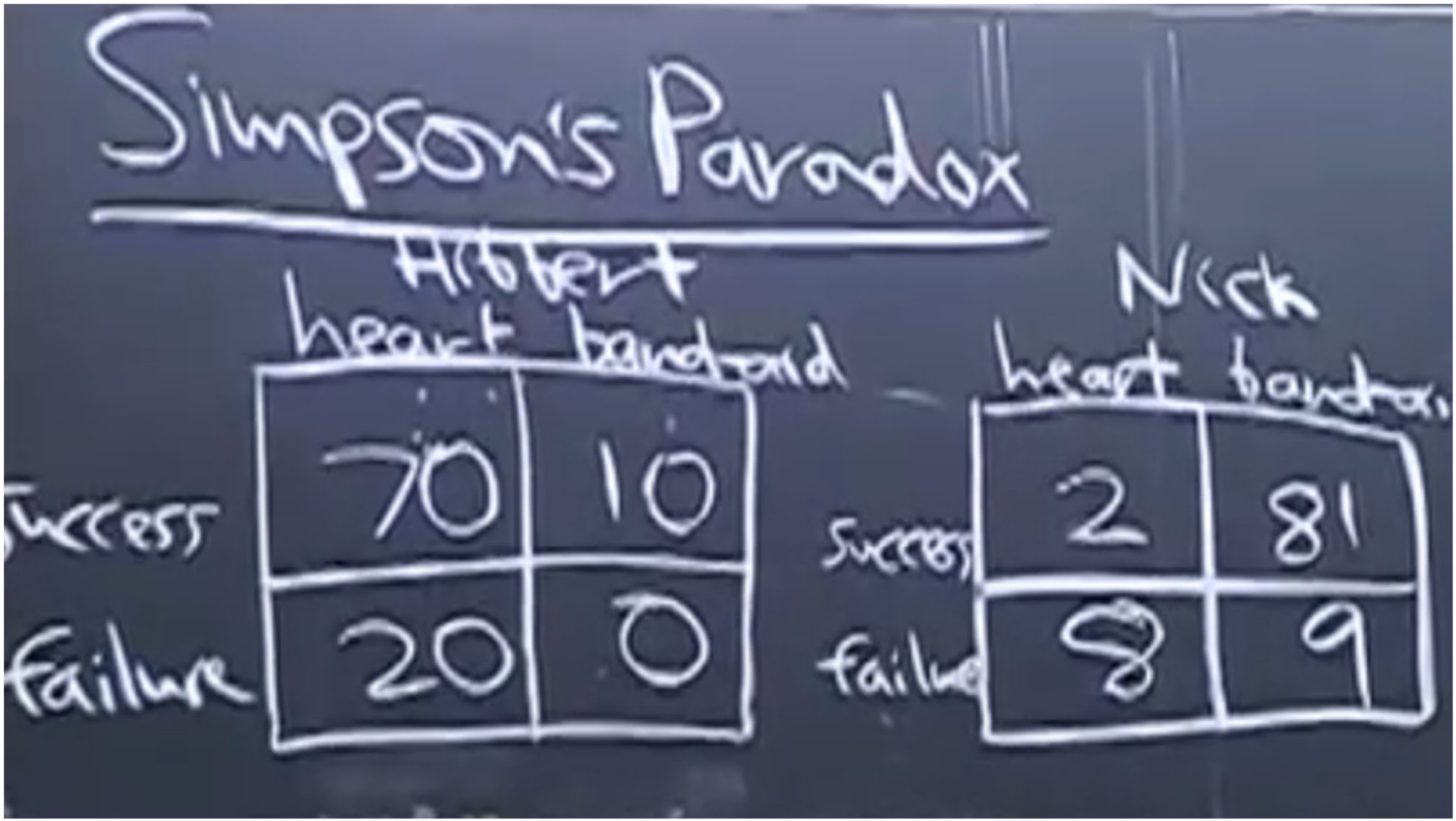
Simpson's paradox

Edward Hugh Simpson
(10 December 1922 – 5 February 2019)
was a British codebreaker, statistician
and civil servant.
"The Interpretation of Interaction
in Contingency Tables", Journal of the Royal Statistical Society, 1951

Is it possible for one doctor to have a **higher** success rate than another doctor in **every** type of treatment he performs but to have a **lower** **overall** success rate across all treatment types?

Dr. Hibbert

Dr. Nick



Dr. Hibbert: success rate =80%

Dr. Nick: success rate =83%

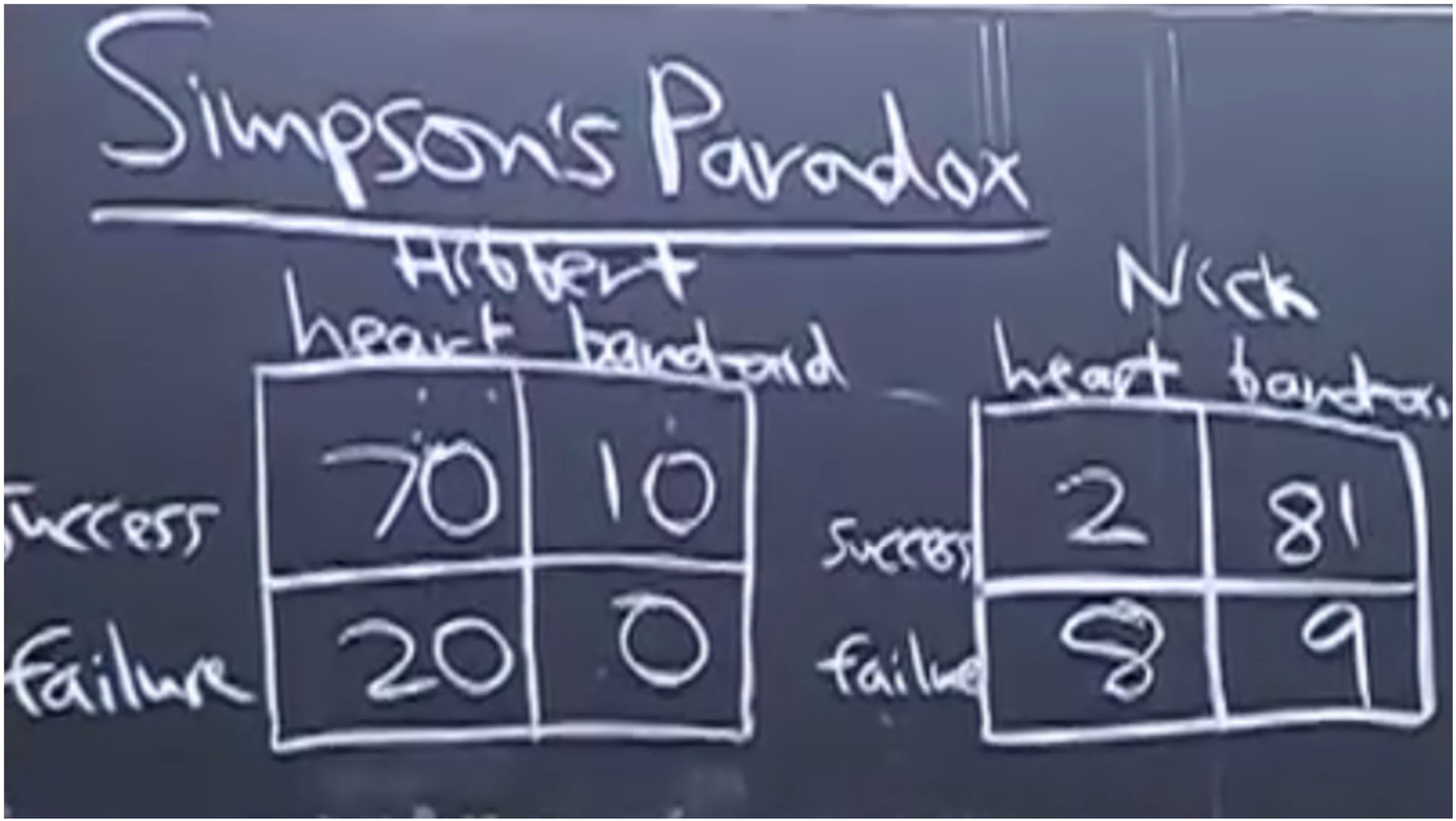
Simpson's paradox

Edward Hugh Simpson
(10 December 1922 – 5 February 2019)
was a British codebreaker, statistician
and civil servant.
"The Interpretation of Interaction
in Contingency Tables", Journal of the Royal Statistical Society, 1951

Is it possible for one doctor to have a **higher** success rate than another doctor in **every** type of treatment he performs but to have a **lower** **overall** success rate across all treatment types?

Dr. Hibbert

Dr. Nick

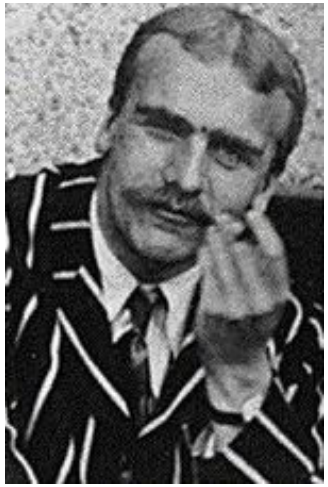


Dr. Hibbert: success rate =80%

Dr. Nick: success rate =83%

Simpson's paradox might explain altruism

- Darwinian evolution has a problem with altruism
- "Selfish genes" do not care about others
- J. B. S. Haldane, (1892-1964) British geneticist, evolutionary biologist



- When asked if he would give his life to save a drowning brother answered: "No, but I would to save two brothers or eight cousins"
- Altruism in some insect colonies like ants is because they are all genetically similar.

Altruism in bacteria

- Bacteria live in communities in close proximity to each other
- Individual bugs spend significant resources to produce extracellular molecules, excrete them outside of the cell to share with others. **That slows their growth**
  - Examples: extracellular enzymes, biofilm components, antimicrobial and anti-immune agents
- Cheaters have faster growth rate
  - They can take over by not producing any shared molecules
- **Evolutionary paradox: how bacteria can be altruistic?**

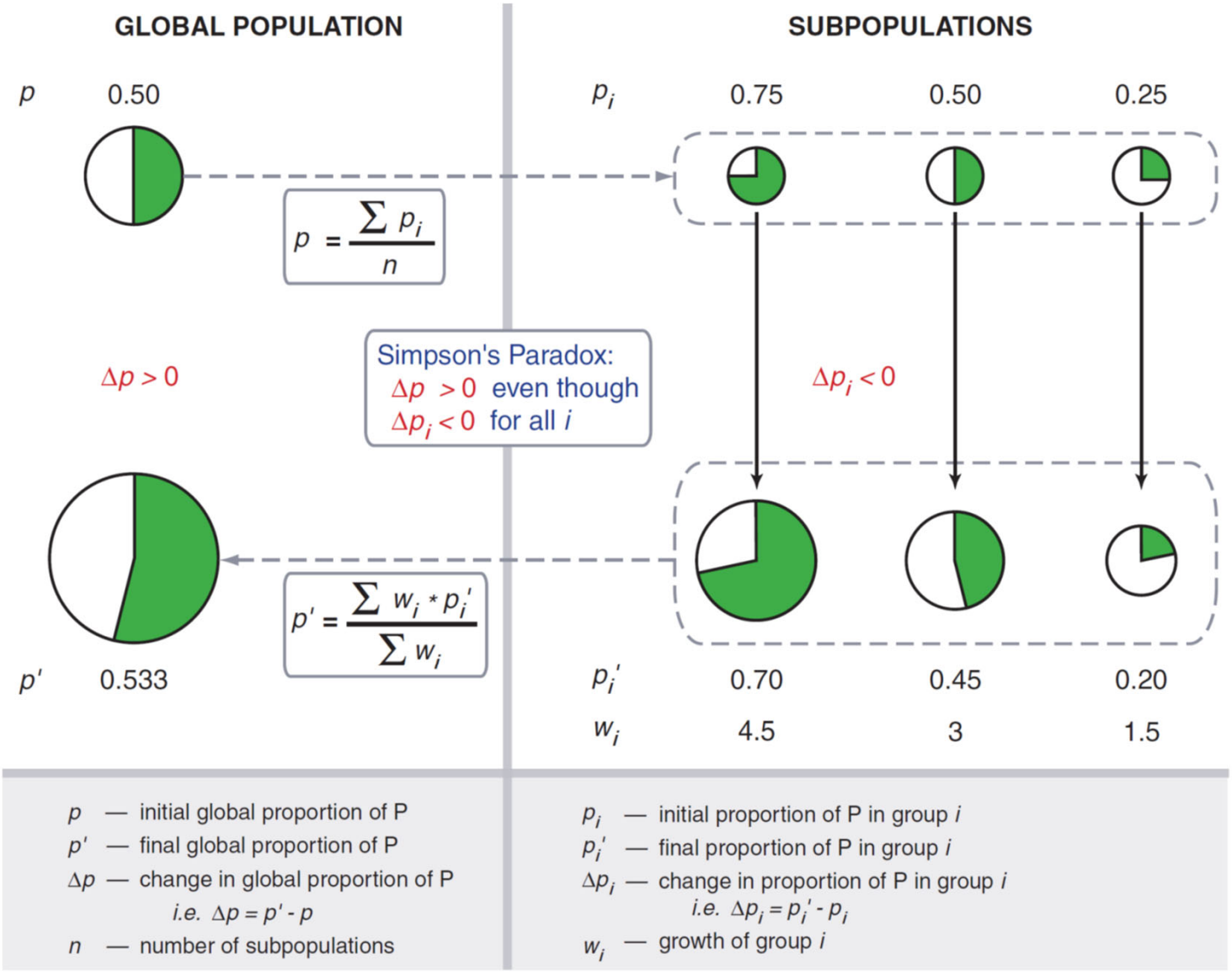
Chuang, Rivoire, and Leibler's answer

Simpson's Paradox in a Synthetic Microbial System

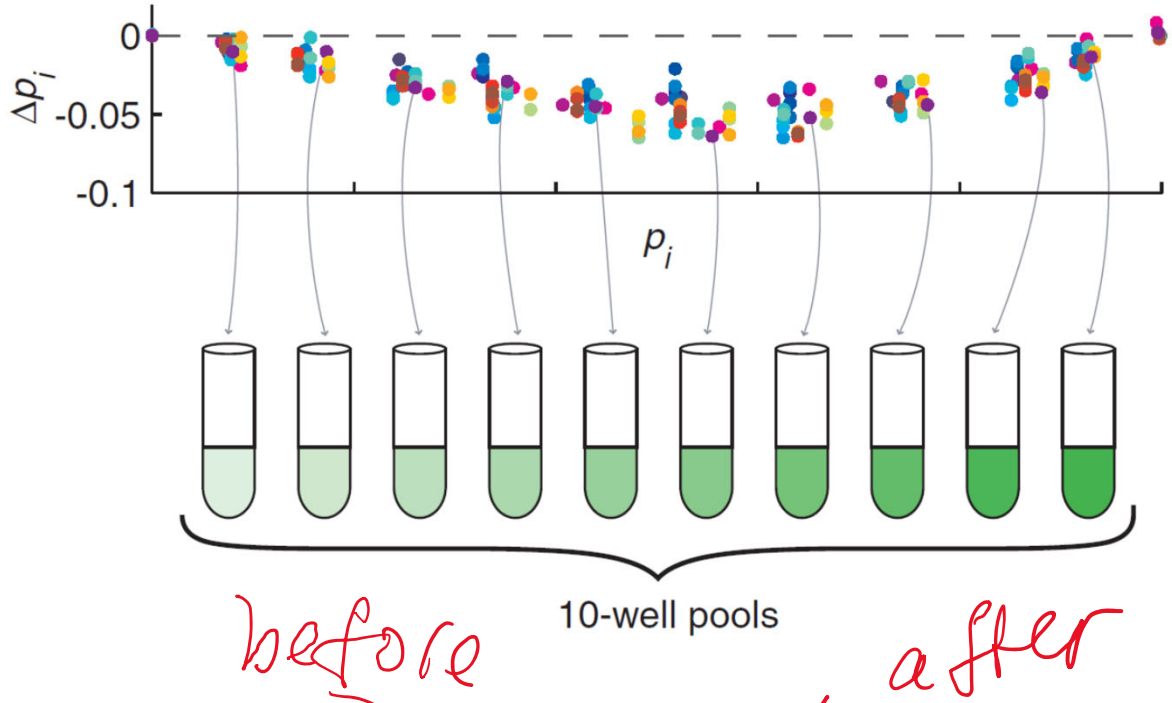
John S. Chuang,* Olivier Rivoire, Stanislas Leibler

The maintenance of "public" or "common good" producers is a major question in the evolution of cooperation. Because nonproducers benefit from the shared resource without bearing its cost of production, they may proliferate faster than producers. We established a synthetic microbial system consisting of two *Escherichia coli* strains of common-good producers and nonproducers. Depending on the population structure, which was varied by forming groups with different initial compositions, an apparently paradoxical situation could be attained in which nonproducers grew faster within each group, yet producers increased overall. We show that a simple way to generate the variance required for this effect is through stochastic fluctuations via population bottlenecks. The synthetic approach described here thus provides a way to study generic mechanisms of natural selection.

- The common good was a membrane-permeable Rhl autoinducer molecule rewired to activate antibiotic (chloramphenicol; Cm) resistance gene expression.



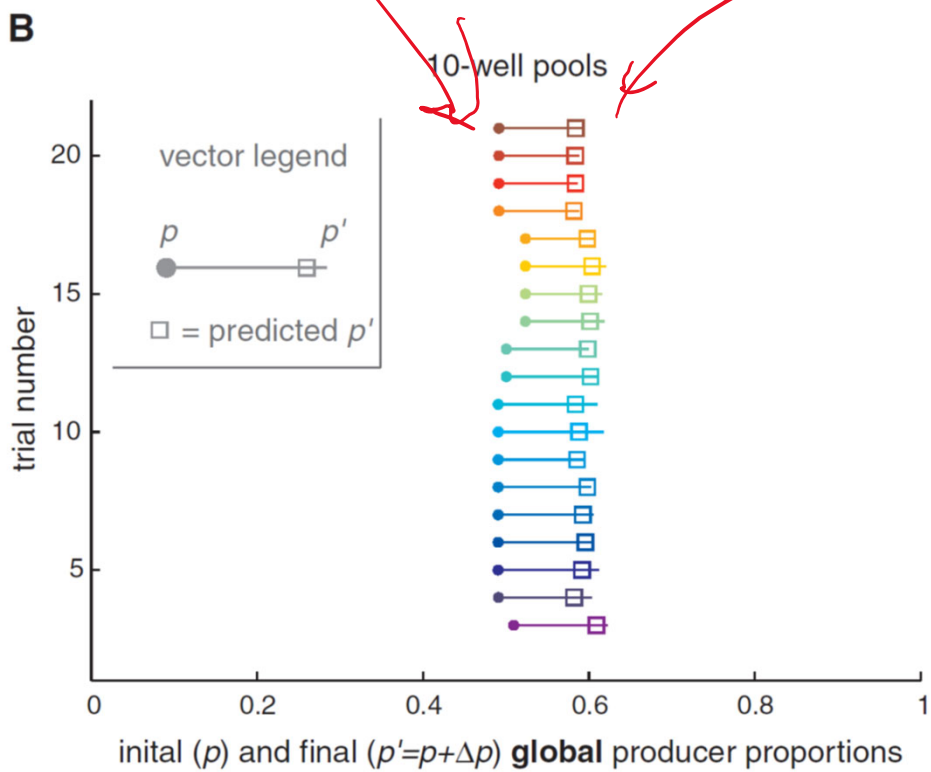
Fraction of altruists in
each of individual
test tubes dropped



before

after

Yet the overall fraction of
altruists in
all test tubes combined
increased



Credit: XKCD comics

QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY ARE THERE FEMALE MR MIMES
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING

WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE THERE SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE HELL IF GOD FORGIVES
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE

Let’s check the theory by playing the dame

Go to
https://dacalderon.shinyapps.io/montyhall/

- Tables 1,3,5 will play “switch the door” strategy
- Tables 2,4,6 will play “same door” strategy
- Play at least 30 rounds (more is better)
- In the end we will **add up the numbers from all tables**

Let’s check with more random experiments

```
Stats=??;
%set Stats large...
switch_count=0; noswitch_count=0; %set 0 at the beginning
for n = 1:Stats
a = randperm(3); %Monty places two goats and the car at random
%a(1) -goat, a(2) -goat, a(3) - car
i= floor(3.*rand)+1; %you select the door!
% SWITCH STRATEGY
if(i == a(1)) switch_count=switch_count+??; %a(2)-opened, switch to a(3), car!
elseif (i == a(2)) switch_count = switch_count + ??;%a(1) opened, switch to a(3), car!
else switch_count = switch_count + ??; %a(1)/a(2) opened, switch to a(2)/a(1), no car :-(
end
% NO SWITCH STRATEGY
if(i == a(1)) noswitch_count = noswitch_count + ??; %a(2)-opened, no car :-(
elseif (i==a(2)) noswitch_count = noswitch_count + ?? %a(1)-opened, no car :-(
else  noswitch_count = noswitch_count + ??; %a(1) or a(2)-opened, car!
endend;
disp('probability to win a car if switched doors=');
disp(num2str(switch_count./??)); %# of cars with switching
disp('probability to win a car if did not switch doors=');
disp(num2str(noswitch_count./??)); %# of cars w/o switching
```

Credit: XKCD comics

QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY ARE THERE FEMALE MR MIMES
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING

WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE THERE SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE HELL IF GOD FORGIVES
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE

Discrete Probability Distributions

Random Variables

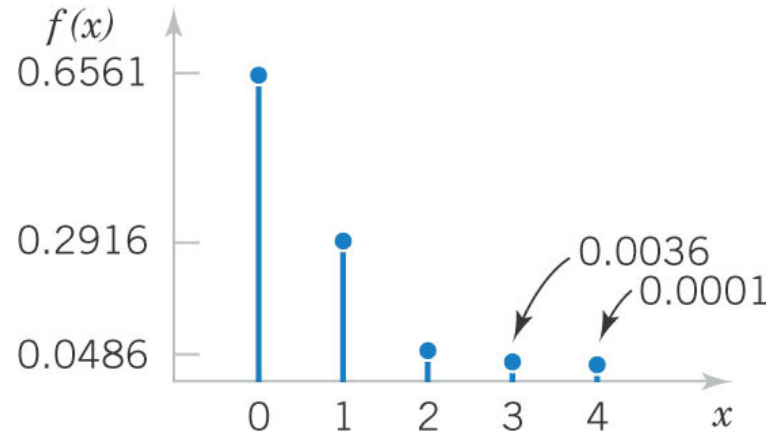
- A variable that associates a number with the outcome of a random experiment is called a random variable.
- Notation: random variable is denoted by an uppercase letter, such as X. After the experiment is conducted, the measured value is denoted by a lowercase letter, such a x. Both X and x are shown in italics, e.g., $P(X=x)$.

Continuous & Discrete Random Variables

- A discrete random variable is usually integer number
  - N - the number of p53 proteins in a cell
  - D - the number of nucleotides different between two sequences
- A continuous random variable is a real number
  - C=N/V – the concentration of p53 protein in a cell of volume V
  - Percentage (D/L)*100% of different nucleotides in protein sequences of different lengths L (depending on the set of L's may be discrete but dense)

Probability Mass Function (PMF)

- I want to compare all 4-mers in a pair of human genomes
- *X* – random variable: the number of nucleotide differences in a given 4-mer
- Probability Mass Function: f(x) or P(X=x) – the probability that the # of SNPs is exactly equal to x



Probability Mass Function for the # of mismatches in 4-mers

P(*X* =0) =	0.6561
P(*X* =1) =	0.2916
P(*X* =2) =	0.0486
P(*X* =3) =	0.0036
P(*X* =4) =	0.0001
$\sum_x$ P(X=x)=	1.0000

Cumulative Distribution Function (CDF)

x	$P(X=x)$	$P(X\leq x)$	$P(X>x)$
-1	0.0000	0.0000	1.0000
0	0.6561	0.6561	0.3439
1	0.2916	0.9477	0.0523
2	0.0486	0.9963	0.0037
3	0.0036	0.9999	0.0001
4	0.0001	1.0000	0.0000

Cumulative Distribution Function CDF: $F(x)=P(X\leq x)$

Example:

$F(3)=P(X \leq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3) = 0.9999$

Complementary Cumulative Distribution Function (tail distribution) or CCDF: $F_{>}(x)=P(X>x)$

Example: $F_{>}(0)= P(X >0) =1- P(X \leq 0) =1-0.6561=0.3439$

Mean or Expected Value of X

The mean or expected value of the discrete random variable X, denoted as μ or $E(X)$, is

$$\mu = E(X) = \sum_x x \cdot P(X = x) = \sum_x x \cdot f(x)$$

- The mean = the weighted average of all possible values of *X*. It represents its "center of mass"

- The mean may, or may not, be an allowed value of X

- It is also called the arithmetic mean (to distinguish from e.g. the geometric mean discussed later)

- Mean may be infinite if X any integer and tail $P(X=x)>c/x^2$

Outcomes of 6 random experiments

0, 1, 0, 0, 2, 1

$$\text{Mean} = \frac{0+1+0+0+2+1}{6} =$$

$$= \frac{3\times 0 + 2\times 1 + 1\times 2}{6} =$$

$$= 0\times\frac{3}{6} + 1\times\frac{2}{6} + 2\times\frac{1}{6} = \sum_{x=0}^{2} x\,P(X=x)$$

- $E[X] = \sum_{x} x \cdot P(X = x)$
- $E[X^2] = \sum x^2 \cdot P(X = x)$
- $E[a \cdot X + b \cdot X^2] = \sum (ax + bx^2) \times$
  $\times P(X = x) = a \cdot \sum x P(X = x) +$
  $+ b \sum x^2 P(X = x)$
- $E[e^X] = \sum e^x P(X = x)$

Variance $V(X)$: square of a typical deviation from the mean $\mu = E(X)$

$V(X) = \sigma^2$, where σ is called standard deviation

$$\sigma^2 = V(X) = E((X-\mu)^2) =$$

$$= E(X^2 - 2\mu X + \mu^2) = E(X^2) -$$

$$- 2\mu E(X) + \mu^2 = E(X^2) - 2\mu^2 + \mu^2 =$$

$$= E(X^2) - \mu^2 = E(X^2) - (E(X))^2$$

Variance of a Random Variable

If X is a discrete random variable with probability mass function $f(x)$,

$$E\left[h(X)\right] = \sum_x h(x) \cdot P(X = x) = \sum_x h(x) f(x) \qquad (3\text{-}4)$$

If $h(x) = (X - \mu)^2$, then its expectation, $V(x)$, is the variance of X.

$\sigma = \sqrt{V(x)}$, is called standard deviation of X

$\sigma^2 = V(X) = \sum_x (x - \mu)^2 f(x)$ is the definitional formula

$$= \sum_x \left(x^2 - 2\mu x + \mu^2\right) f(x)$$

$$= \sum_x x^2 f(x) - 2\mu \sum x f(x) + \mu^2 \sum_x f(x)$$

$$= \sum_x x^2 f(x) - 2\mu^2 + \mu^2$$

$= \sum_x x^2 f(x) - \mu^2$ is the computational formula

Variance can be infinite
if X can be any integer
and tail of P(X=x) $\geq c/x^3$

Skewness of a random variable

- Want to quantify how asymmetric is the distribution around the mean?
- Need any odd moment: $E[(X-\mu)^{2n+1}]$
- Cannot do it with the first moment: $E[X-\mu]=0$
- Normalized 3-rd moment is skewness: $\gamma 1=E[(X-\mu)^3/\sigma^3]$
- Skewness can be infinite if X takes unbounded integer values and tail $P(X=x) \geq c/x^4$

Geometric mean of a random variable

- Useful for very broad distributions (many orders of magnitude)?
- Mean may be dominated by very unlikely but very large events. Think of a lottery
- Exponent of the mean of *log X*: *Geometric mean=exp(E[log X])*
- Geometric mean usually is not infinite

Summary: Parameters of a Probability Distribution

- Probability Mass Function (PMF): $f(x)=\text{Prob}(X=x)$
- Cumulative Distribution Function (CDF): $F(x)=\text{Prob}(X\leq x)$
- Complementary Cumulative Distribution Function (CCDF): $F_{>}(x)=\text{Prob}(X>x)$
- The mean, $\mu=E[X]$, is a measure of the center of mass of a random variable
- The variance, $V(X)=E[(X-\mu)^2]$, is a measure of the dispersion of a random variable around its mean
- The standard deviation, $\sigma=[V(X)]^{1/2}$, is another measure of the dispersion around mean. Has the same units as X
- The skewness, $\gamma 1=E[(X-\mu)^3/\sigma^3]$, a measure of asymmetry around mean
- The geometric mean, $\exp(E[\log X])$ is useful for very broad distributions

Skewness of a random variable

- Want to quantify how asymmetric is the distribution around the mean?
- Need any odd moment: $E[(X-\mu)^{2n+1}]$
- Cannot do it with the first moment: $E[X-\mu]=0$
- Normalized 3-rd moment is skewness: $\gamma_1=E[(X-\mu)^3/\sigma^3]$
- Skewness can be infinite if X takes unbounded positive integer values and the tail $P(X=x) \geq c/x^4$ for large x

Geometric mean of a random variable

- Useful for very broad distributions (many orders of magnitude)?
- Mean may be dominated by very unlikely but very large events. Think of a lottery
- Exponent of the mean of *log X*: *Geometric mean=exp(E[log X])*
- Geometric mean usually is not infinite

Summary: Parameters of a Probability Distribution

- Probability Mass Function (PMF): $f(x)=\text{Prob}(X=x)$
- Cumulative Distribution Function (CDF): $F(x)=\text{Prob}(X\leq x)$
- Complementary Cumulative Distribution Function (CCDF): $F_{>}(x)=\text{Prob}(X>x)$
- The mean, $\mu=E[X]$, is a measure of the center of mass of a random variable
- The variance, $V(X)=E[(X-\mu)^2]$, is a measure of the dispersion of a random variable around its mean
- The standard deviation, $\sigma=[V(X)]^{1/2}$, is another measure of the dispersion around mean. Has the same units as X
- The skewness, $\gamma_1=E[(X-\mu)^3/\sigma^3]$, a measure of asymmetry around mean
- The geometric mean, $\exp(E[\log X])$ is useful for very broad distributions

A gallery of useful
discrete probability distributions

Discrete Uniform Distribution

- Simplest discrete distribution.
- The random variable X assumes only a finite number of values, each with equal probability.
- A random variable X has a discrete uniform distribution if each of the n values in its range, say $x_1, x_2, \ldots, x_n$, has equal probability.

$$f(x_i) = 1/n$$

Uniform Distribution of Consecutive Integers

- Let X be a discrete uniform random variable all integers from a to b (inclusive). There are $b - a + 1$ integers. Therefore each one gets:

$$f(x) = 1/(b-a+1)$$

- Its measures are:

$$\mu = E(x) = (b+a)/2$$

$$\sigma^2 = V(x) = [(b-a+1)^2 - 1]/12$$

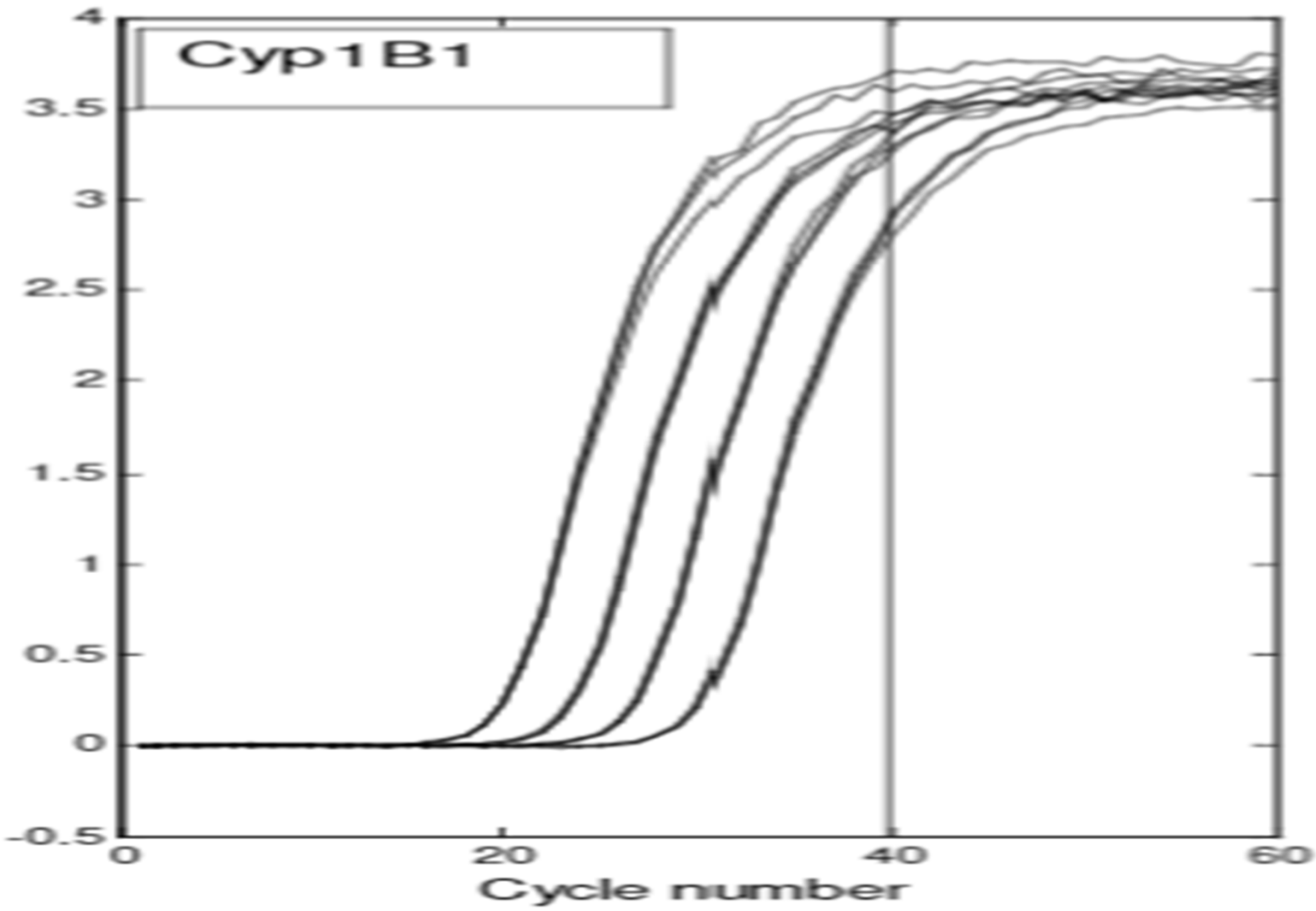
Note that the mean is the midpoint of a & b.

An example of the uniform distribution

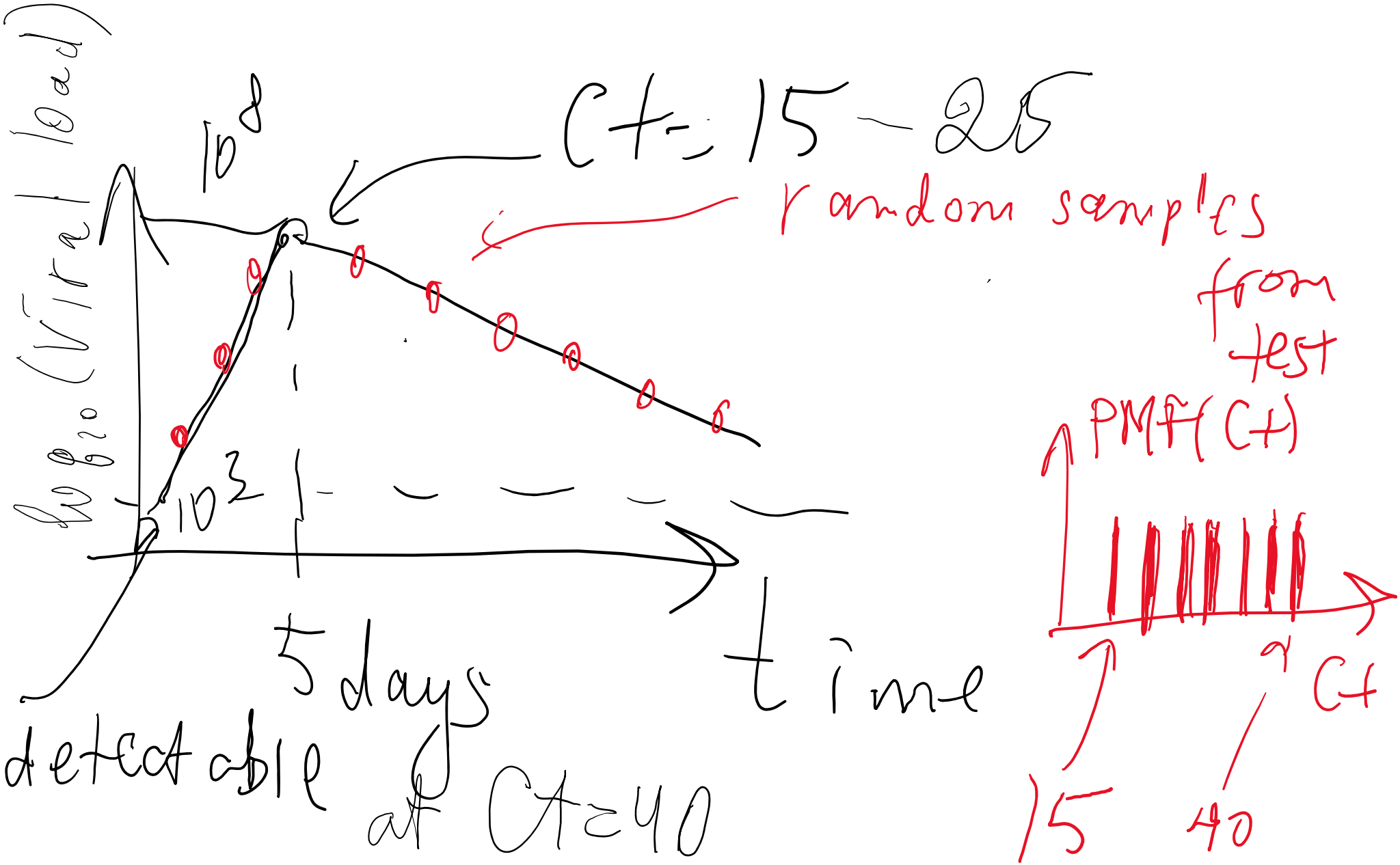
Cycle threshold (Ct) value in COVID-19 infection

What is the Ct value of a PCR test?

Ct = const – log2(viral DNA concentration)



Why Ct distribution should it be uniform?



Examples of uniform distribution: Ct value of PCR test of a virus

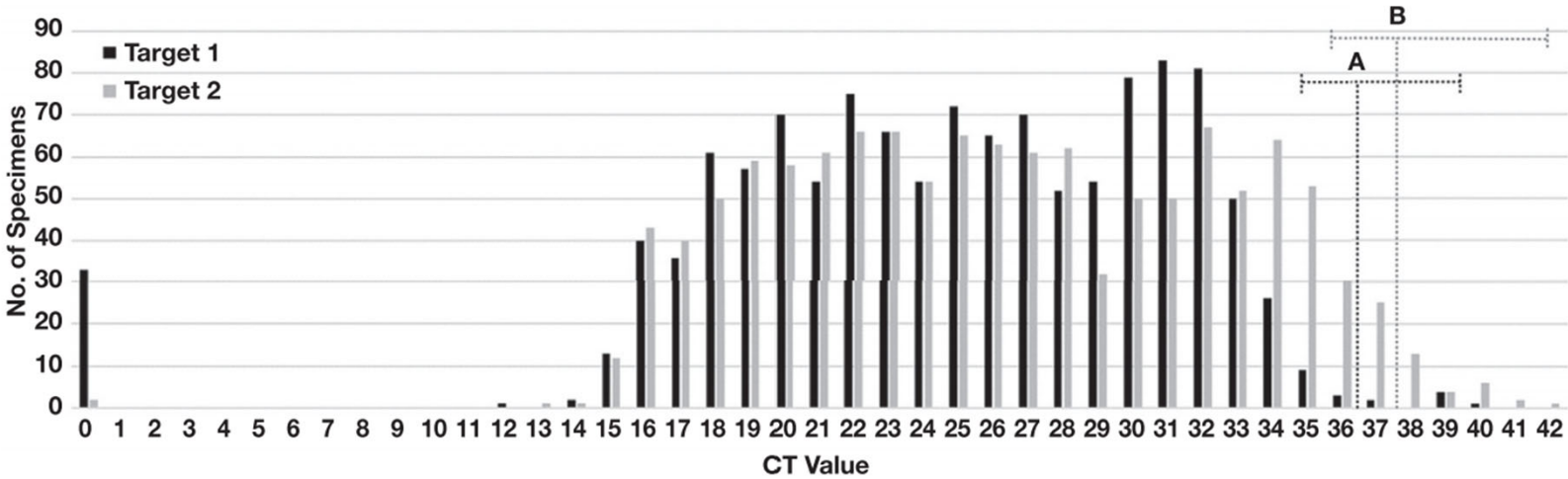


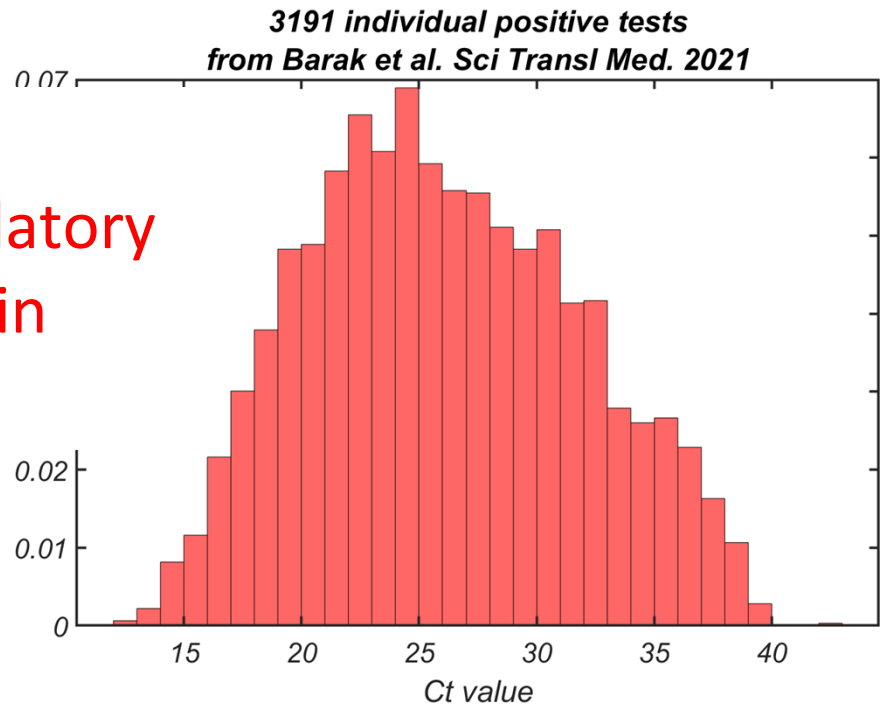
Figure 3 Distribution of cycle threshold (CT) values. The total number of specimens with indicated CT values for Target 1 and 2 are plotted. The estimated limit of detection for (A) Target 1 and (B) Target 2 are indicated by vertical dotted lines. Horizontal dotted lines encompass specimens with CT values less than 3× the LoD for which sensitivity of detection may be less than 100%. This included 19/1,180 (1.6%) reported CT values for Target 1 and 81/1,211 (6.7%) reported CT values for Target 2. Specimens with Target 1 or 2 reported as "not detected" are denoted as a CT value of "0."

Distribution of SARS-CoV-2 PCR Cycle Threshold Values Provide Practical Insight Into Overall and Target-Specific Sensitivity Among Symptomatic Patients

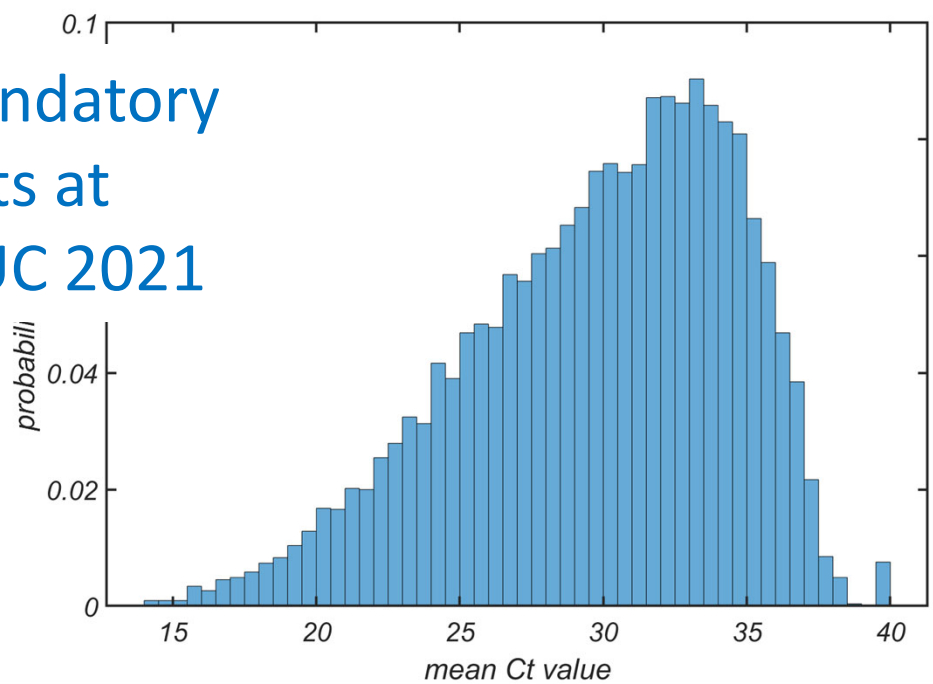
Blake W Buchan, PhD, Jessica S Hoff, PhD, Cameron G Gmehlin, Adriana Perez, Matthew L Faron, PhD, L Silvia Munoz-Price, MD, PhD, Nathan A Ledeboer, PhD *American Journal of Clinical Pathology*, Volume 154, Issue 4, 1 October 2020, https://academic.oup.com/ajcp/article/154/4/479/5873820

Why should we care?

Non-mandatory tests in Israel



Mandatory tests at UIUC 2021



- High Ct value means we identified the infected individual early, hopefully before transmission to others
- When testing is mandatory, and people are tested frequently – Ct value is skewed towards high values

Matlab exercise: Uniform distribution

- Generate a sample of size 100,000 for uniform random variable X taking values 1,2,3,…10
- Plot the approximation to the probability mass function based on this sample
- Calculate mean and variance of this sample and compare it to infinite sample predictions: $E[X]=(a+b)/2$ and $V[X]=((a-b+1)^2-1)/12$

Matlab template: Uniform distribution

- **b=10; a=1; % b= upper bound; a= lower bound (inclusive)'**
- **Stats=100000; % sample size to generate**
- **r1=rand(Stats,1);**
- **r2=floor(??*r1)+??;**
- **mean(r2)**
- **var(r2)**
- **std(r2)**
- **[hy,hx]=hist(r2, 1:10); % hist generates histogram in bins 1,2,3...,10**
- **% hy - number of counts in each bin; hx - coordinates of bins**
- **p_f=hy./??; % normalize counts to add up to 1**
- **figure; plot(??,p_f, 'ko-'); ylim([0, max(p_f)+0.01]); % plot the PMF**

Bernoulli distribution

The simplest non-uniform distribution

p – probability of success (1)

1-p – probability of failure (0)

$$f(x) = P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

Jacob Bernoulli
(1654-1705)
Swiss mathematician (Basel)

- Law of large numbers
- Mathematical constant e=2.718…

Bernoulli distribution

$$f(x) = P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

$$E(X) = 0 \times P(X = 0) + 1 \times P(X = 1) = 0(1 - p) + 1(p) = p$$

$$\mathrm{Var}(X) = E(X^2) - (EX)^2 = [0^2(1 - p) + 1^2(p)] - p^2 = p - p^2 = p(1 - p)$$

Bernoulli distribution

The simplest non-uniform distribution

p – probability of success (1)

1-p – probability of failure (0)

$$f(x) = P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

Jacob Bernoulli
(1654-1705)
Swiss mathematician (Basel)

- Law of large numbers
- Mathematical constant e=2.718…

Bernoulli distribution

$$f(x) = P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

$$E(X) = 0 \times P(X = 0) + 1 \times P(X = 1) = 0(1 - p) + 1(p) = p$$

$$\mathrm{Var}(X) = E(X^2) - (EX)^2 = [0^2(1 - p) + 1^2(p)] - p^2 = p - p^2 = p(1 - p)$$

Refresher: Binomial Coefficients

$$\binom{n}{k} = C_k^n = \frac{n!}{k!(n-k)!}\ ,\ \text{called } n \text{ choose } k$$

$$\binom{10}{3} = C_3^{10} = \frac{10!}{3!7!} = \frac{10\cdot 9\cdot 8\cdot 7!}{3\cdot 2\cdot 1\cdot 7!} = 120$$

Number of ways to choose k objects out of n
without replacement and where the order does not matter.
Called binomial coefficients because of the binomial formula

$$(p+q)^n = (p+q)\times(p+q)\ldots\times(p+q) = \sum_{x=0}^{n} C_x^n p^x q^{n-x}$$

Binomial Distribution

- Binomially-distributed random variable *X* equals sum (number of successes) of n independent Bernoulli trials
- The probability mass function is:

$q = 1 - p$

$$f(x) = C_x^n p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, \ldots n \qquad (3\text{-}7)$$

- Based on the binomial expansion:

$$1 = (p+q)^n = \sum_{x=0}^{n} C_x^n p^x q^{n-x}$$

Binomial variance and standard deviation

Let X be a binomial random variable with parameters p and n

Variance:

$\sigma^2 = V(X) = np(1\text{-}p)$

Standard deviation:

$\sigma = \sqrt{np(1-p)}$

Poisson Distribution

- Limit of the binomial distribution when
  - n , the number of attempts, is very large
  - p , *the probability of success* is very small
  - $E(X)=np=\lambda$ is O(1)

The annual numbers of deaths from horse kicks in 14 Prussian army corps between 1875 and 1894

Number of deaths	Observed frequency	Expected frequency
0	144	139
1	91	97
2	32	34
3	11	8
4	2	1
5 and over	0	0
Total	280	280

From von Bortkiewicz 1898

Siméon Denis Poisson (1781–1840) French mathematician and physicist

Let $\lambda = np = E(x)$, so $p = \dfrac{\lambda}{n}$

$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$

$$= \frac{n(n-1)\ldots(n-x+1)}{x!}\left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x} \sim \frac{n^x}{x!}\left(\frac{\lambda}{n}\right)^x = \frac{\lambda^x}{x!};$$

$$\sum_x \frac{\lambda^x}{x!} = e^{\lambda}.$$

Normalization requires $\sum_x P(X = x) = 1$.

Thus $P(X = x) = \dfrac{\lambda^x}{x!} e^{-\lambda}$

Poisson Mean & Variance

If X is a Poisson random variable, then:

- Mean: $\mu = E(X) = \lambda = n \cdot p$
- Variance: $\sigma^2 = V(X) = \lambda = n \cdot p \cdot (1-p) \approx n \cdot p$
- Standard deviation: $\sigma = \lambda^{1/2}$

Note: Variance = Mean

Note: Standard deviation/Mean = $\lambda^{-1/2}$
decreases with λ

Matlab exercise: Poisson distribution

- Generate a sample of size 100,000 for Poisson-distributed random variable X with $\lambda = 2$
- Plot the approximation to the Probability Mass Function based on this sample
- Calculate the mean and variance of this sample and compare it to theoretical calculations:
  $E[X] = \lambda$ and $V[X] = \lambda$

Credit: XKCD comics

QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

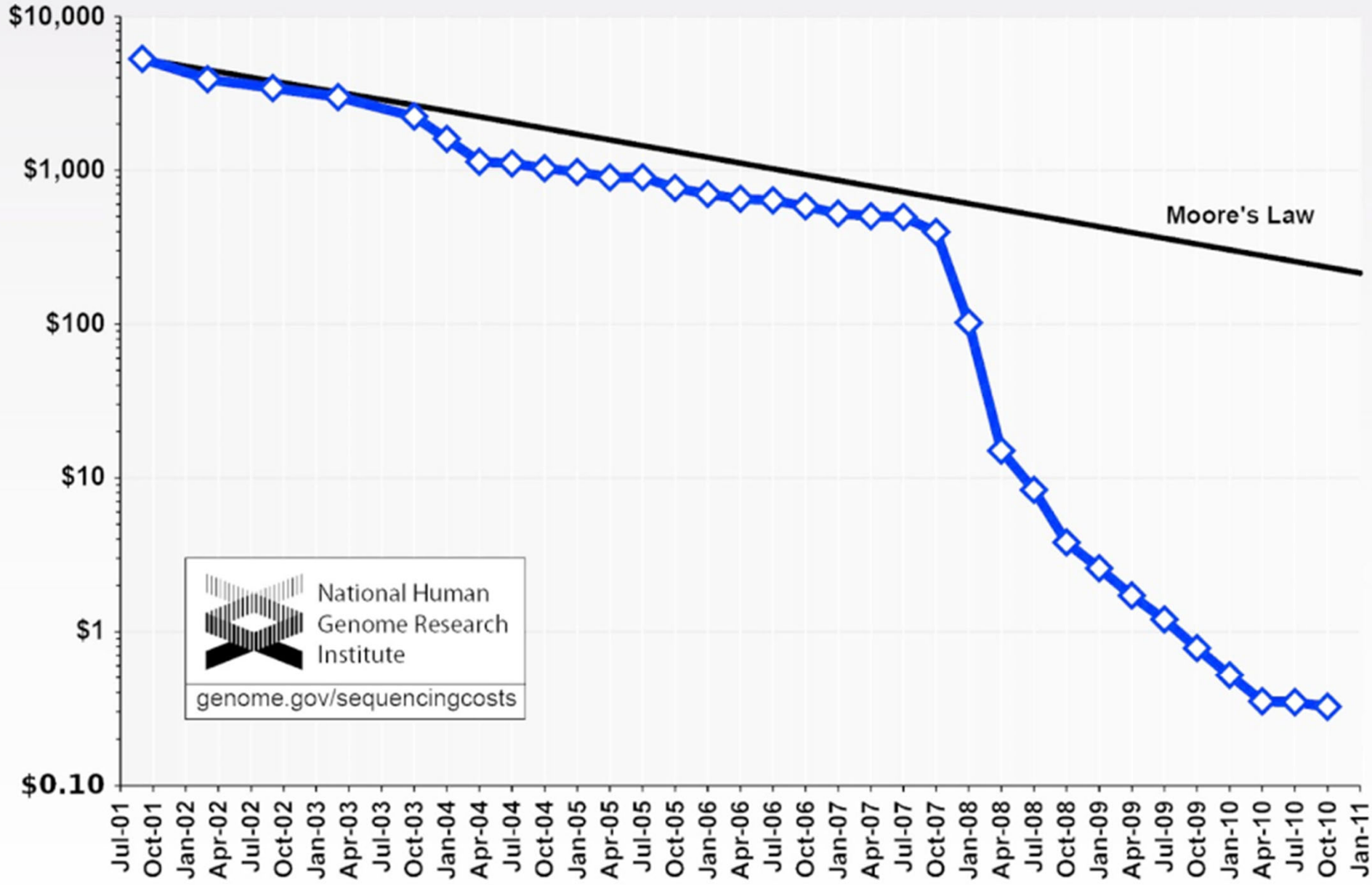
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY ARE THERE FEMALE MR MIMES
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING

WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE THERE SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE HELL IF GOD FORGIVES
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE

Poisson Distribution in Genome Assembly

Cost per Megabase of DNA Sequence



Poisson Example: Genome Assembly

- Goal: figure out the sequence of DNA nucleotides (ACTG) along the entire genome
- Problem: Sequencers generate random short reads

TABLE 9.1 Next-generation sequencing technologies compared to Sanger sequencing. Adapted from the companies' websites, ⊕ http://en.wikipedia.org/wiki/DNA_sequencer, and literature cited for each technology.

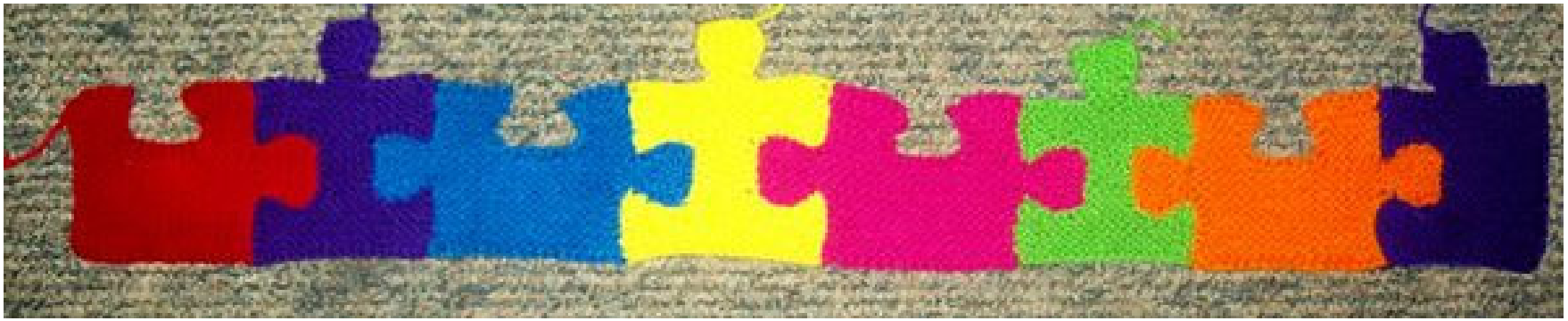
Technology	Read length (bp)	Reads per run	Time per run	Cost per megabase (US$)	Accuracy (%)
Roche 454	700	1 million	1 day	10	99.90
Illumina	50–250	<3 billion	1–10 days	~0.10	98
SOLiD	50	~1.4 billion	7–14 days	0.13	99.90
Ion Torrent	200	<5 million	2 hours	1	98
Pacific Biosciences	2900	<75,000	<2 hours	2	99
Sanger	400–900	N/A	<3 hours	2400	99.90

- Solution: assemble genome from short reads using computers. Whole Genome Shotgun Assembly.

MinION, a palm-sized gene sequencer made by UK-based Oxford Nanopore Technologies

Short Reads assemble into Contigs

Figure 5.1.

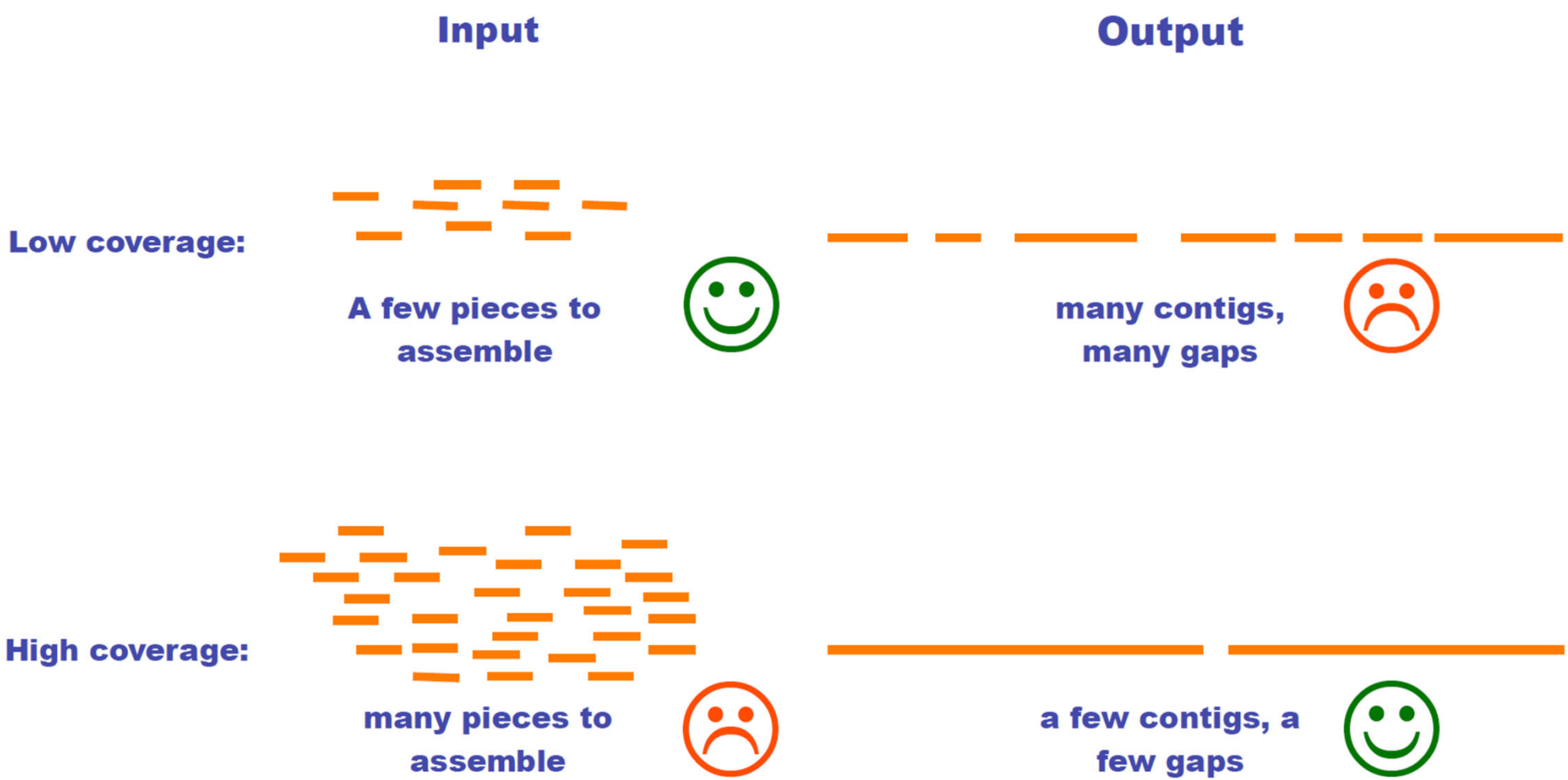


Promise of Genomics

Drew Sheneman, New Jersey -- The Newark Star Ledger, E-mail Drew.

I think I found the corner piece!

How many short reads do we need?



Genome Assembly

Whole-genome "shotgun" sequencing starts by copying and fragmenting the DNA

("Shotgun" refers to the random fragmentation of the whole genome; like it was fired from a shotgun)

Input: GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
35bp

Copy by PCR:
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT

Fragment:
GGCGTCTA TATCTCGG CTCTAGGCCCTC ATTTTTT
GGC GTCTATAT CTCGGCTCTAGGCCCTCA TTTTTT
GGCGTC TATATCT CGGCTCTAGGCCCT CATTTTTT
GGCGTCTAT ATCTCGGCTCTAG GCCCTCA TTTTTT

http://www.langmead-lab.org/teaching-materials/

Assembly

Assume sequencing produces such a large # fragments that almost all genome positions are *covered* by many fragments...

...but we don't know what came from where

Reconstruct this

CTAGGCCCTCAATTTTT
GGCGTCTATATCT
CTCTAGGCCCTCAATTTTT
TCTATATCTCGGCTCTAGG
GGCTCTAGGCCCTCATTTTTT
CTCGGCTCTAGCCCCTCATTTT
TATCTCGACTCTAGGCCCTCA
GGCGTCGATATCT
TATCTCGACTCTAGGCC
GGCGTCTATATCTCG

From these

GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT

http://www.langmead-lab.org/teaching-materials/

Assembly

Overlaps between short reads help to put them together

```
                    CTAGGCCCTCAATTTTT
                  CTCTAGGCCCTCAATTTTT
                GGCTCTAGGCCCTCATTTTTT
            CTCGGCTCTAGCCCCTCATTTT
        TATCTCGACTCTAGGCCCTCA
        TATCTCGACTCTAGGCC
    TCTATATCTCGGCTCTAGG
GGCGTCTATATCTCG
GGCGTCGATATCT
GGCGTCTATATCT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
```

177 nucleotides

35 nucleotides

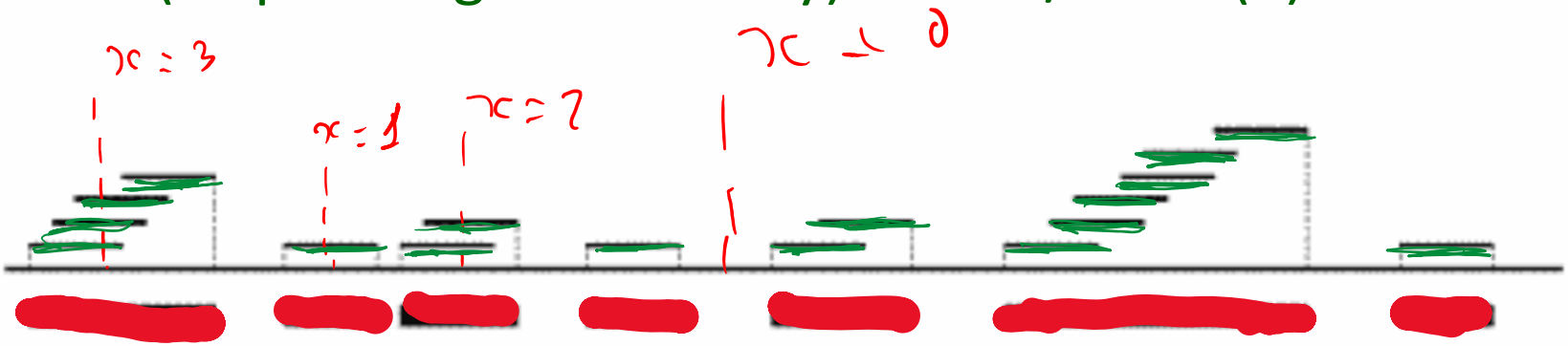
http://www.langmead-lab.org/teaching-materials/

Where is the Poisson?

- *G - genome length (in bp)*
- *L - short read average length*
- *N – number of short read sequenced*
- *λ – sequencing coverage redundancy = LN/G*
- *x- number of short reads covering a given site on the genome*

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Poisson as a limit of Binomial: For a given site on the genome for each short read Prob(site covered): p=L/G is very small. Number of attempts (short reads): N is very large. Their product (sequencing redundancy): λ = NL/G is O(1).



Ewens, Grant, Chapter 5.1

What fraction of genome is covered?

- Coverage: $\lambda = NL/G$,
  X – random variable equal to the number of times a given site is covered by short reads.
  Poisson: $P(X=x) = \lambda^x \exp(-\lambda)/x!$
  $P(X=0) = \exp(-\lambda)$, $P(X>0) = 1 - \exp(-\lambda)$
- *Total length covered:* $G*[1 - \exp(-\lambda)]$

λ	2	4	6	8	10	12
Mean proportion of genome covered	.864665	.981684	.997521	.999665	.999955	.999994

Table 5.1. The mean proportion of the genome covered for different values of λ

How many contigs?

- A given short read is the right end of a contig if and only if no left ends of other short reads fall within it.
- The left end of another short read has the probability $p=(L-1)/G$ to fall within a given read. There are $N-1$ other reads. Hence the expected number of left ends inside a given shot read is $p\cdot (N-1)=(N-1)\cdot(L-1)/G \approx \lambda$
- If significant overlap required to merge two short reads is L_{ov}, modified λ is given by $(N-1)\cdot(L-L_{ov})/G$
- Probability that no left ends fall inside a short read is $exp(-\lambda)$. Thus the Number of contigs is $N_{contigs}=Ne^{-\lambda}$:

λ	0.5	0.75	1	1.5	2	3	4	5	6	7
Mean number of contigs	60.7	70.8	73.6	66.9	54.1	29.9	14.7	6.7	3.0	1.3

Table 5.2. The mean number of contigs for different levels of coverage, with $G = 100{,}000$ and $L = 500$.

Poisson Example: Genome Assembly

- Goal: DNA sequence (ACTG) of the entire genome
- Problem: Sequencers generate random short reads

Sequencer	Sanger 3730xl	454 GS	Ion Torrent	SOLiDv4	Illumina HiSeq 2000	Pac Bio
Mechanism	Dideoxy chain termination	Pyrosequencing	Detection of hydrogen ion	Ligation and two-base coding	Reversible Nucleotides	Single molecule real time
Read length	400-900 bp	700 bp	~400 bp	50 + 50 bp	100 bp PE	>10000 bp
Error Rate	0.001%	0.1%	2%	0.1%	2%	10-15%
Output data (per run)	100 KB	1 GB	100 GB	100 GB	1 TB	10 GB
Approx cost per GB		10,000	1000	100	10	1000

- Solution: assemble genome from short reads using computers. Whole Genome Shotgun Assembly.

Table from the course EE 372 taught by David Tse at Stanford

Current sequencing technologies

	Second gen. (Illumina)	Oxford Nanopore (MinIon)	PacBio
read length (bases)	100-500	10K-100K	10K-20K
error rates	< 1%	10-15%	10-15%
speed (time/base)	6 mins/base/strand	250 bases/s	3 bases/s
# of reads in parallel	10^9	2000	150K
throughput (total # of bases/s)	3M	500K	450K

Table from the course EE 372: Data Science for High-Throughput Sequencing. taught by David Tse at Stanford

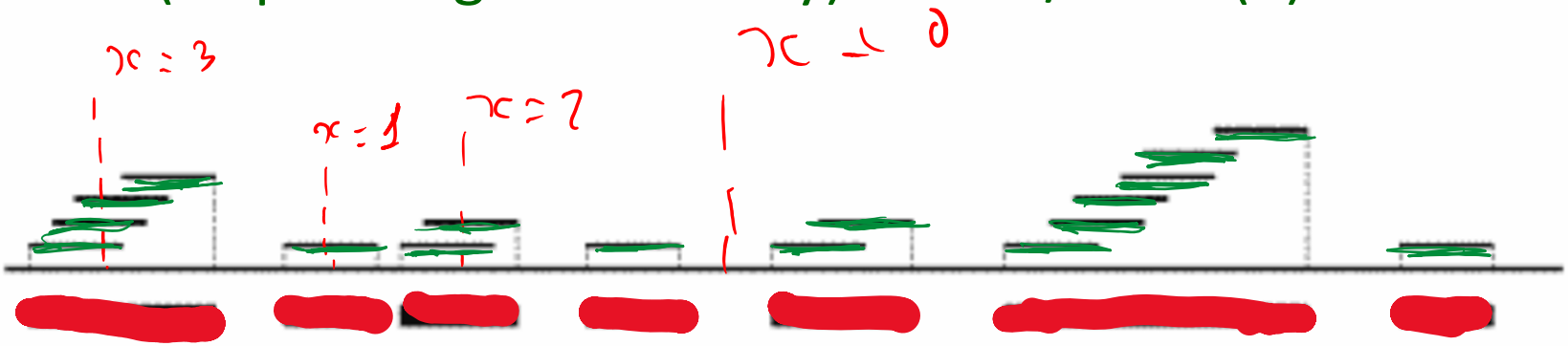
MinION, a palm-sized gene sequencer made by UK-based Oxford Nanopore Technologies

Where is the Poisson?

- *G - genome length (in bp)*
- *L - short read average length*
- *N – number of short read sequenced*
- *λ – sequencing coverage redundancy = LN/G*
- *x- number of short reads covering a given site on the genome*

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Poisson as a limit of Binomial: For a given site on the genome for each short read Prob(site covered): p=L/G is very small. Number of attempts (short reads): N is very large. Their product (sequencing redundancy): λ = NL/G is O(1).



Ewens, Grant, Chapter 5.1

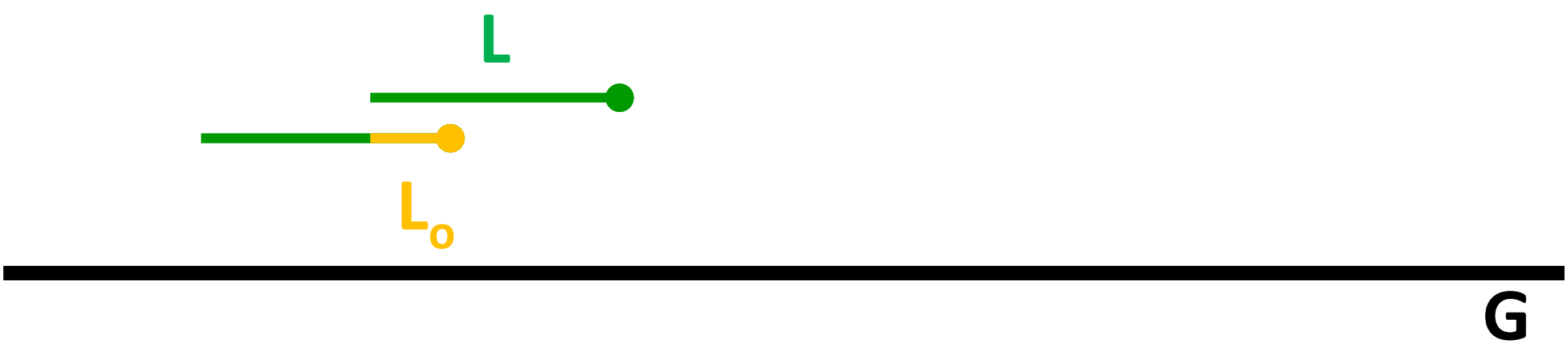
What fraction of genome is covered?

- Coverage: $\lambda = NL/G$,
  X – random variable equal to the number of times a given site is covered by short reads.
  Poisson: $P(X=x) = \lambda^x \exp(-\lambda)/x!$
  $P(X=0) = \exp(-\lambda)$, $P(X>0) = 1 - \exp(-\lambda)$
- *Total length covered:* $G*[1 - \exp(-\lambda)]$

λ	2	4	6	8	10	12
Mean proportion of genome covered	.864665	.981684	.997521	.999665	.999955	.999994

Table 5.1. The mean proportion of the genome covered for different values of λ

How long should the overlap be to connect two short reads?



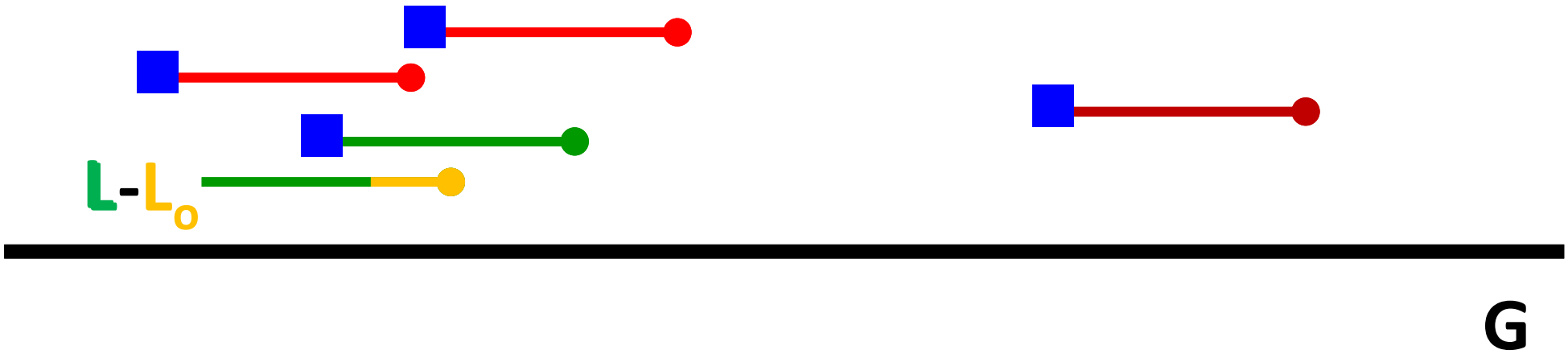
If DNA was a random chain with $p_A = p_C = p_G = p_T = 1/4$

L_o~16-20 would be enough

$$2 \cdot G \cdot 4^{-Lo} = 2 \cdot 3x10^9 \cdot 4^{-16} = 1.4$$

$$2 \cdot 3x10^9 \cdot 4^{-20} = 0.0055 << 1$$

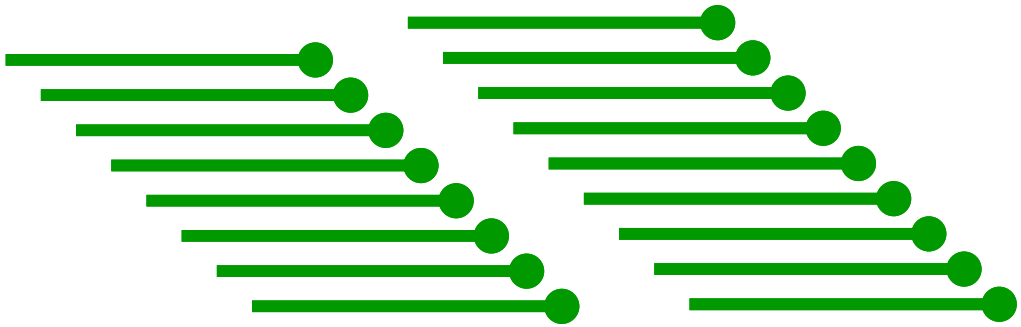
How many contigs?



$$\text{P(short read can be extended by another short read)} = \frac{L - L_o}{G} = \text{p}$$

$$\text{P(short read cannot be extended by any short reads)} = e^{-pN} \approx Ne^{-\lambda}$$

$$\text{number of contigs} = Ne^{-pN} \approx Ne^{-\lambda}$$



How many contigs?

- A given short read is the right end of a contig if and only if no left ends of other short reads fall within first $L\text{-}L_{overlap}$ base pairs
- The left end of another short read has the probability $p=(L\text{-}L_{overlap})/G$ to fall within a given read. There are $N\text{-}1$ other reads.
- The expected number of left ends inside a given short read is $p\cdot (N\text{-}1)=(N\text{-}1)\cdot(L\text{-}L_{overlap})/G \approx \lambda$ (if $L>>L_{overlap}$)
- Probability that no left ends fall inside a given short read is $exp(-\lambda)$. Thus, the Number of contigs is $N_{contigs}=Ne^{-\lambda}$:

λ	0.5	0.75	1	1.5	2	3	4	5	6	7
Mean number of contigs	60.7	70.8	73.6	66.9	54.1	29.9	14.7	6.7	3.0	1.3

Table 5.2. The mean number of contigs for different levels of coverage, with $G = 100{,}000$ and $L = 500$.

Average length of a contig?

- Length of a genome covered:
  $G_{covered}=G\cdot P(X>0)=G\cdot(1-\exp(-\lambda))$
- *Number of contigs* $N_{contigs}=N\cdot e^{-\lambda}$
- Average length of a contig =
  $\langle L\rangle=\sum_i L_i/N_{contigs}=G_{covered}/N_{contigs}=$

$G\cdot(1-\exp(-\lambda))/N\cdot e^{-\lambda}=L\cdot(1-\exp(-\lambda))/\lambda\cdot e^{-\lambda}$

λ	2	4	6	8	10
Mean contig size	1,600	6,700	33,500	186,000	1,100,000

Table 5.3. The mean contig size for different values of a for the case $L = 500$.

Estimate

- Human genome is $3x10^9$ bp long
- Chromosome 1 is about $G=0.25x10^9$ bp
- Illumina generates short reads $L=100$ bp long
- What number of reads N are needed to completely assemble the 1st chromosome?
- The formula to use is: $1=N_{contigs}=Ne^{-\lambda}=Ne^{-NL/G}$
- Answer: $N=4.4x10^7$ short (100bp) reads
  Test: 4.4e7*exp(-4.4e7*100/0.25e9)=0.9997
- What coverage redundancy λ will it be?
  Answer: $\lambda=NL/G=17.6$ coverage redundancy

How much would it cost to assemble human genome now?

- Human Genome Project: $2.7 billion in 1991 dollars.
- Now a de novo full assembly of the whole human genome would now cost $3 \times 10^9 \times 17.6 / 10^6 \times 0.1\$/MB = \$5300$
- 2nd genome (and after) would be even cheaper as we would already have a reference genome to which we can map short reads. (Puzzle: picture on the box)
- But this is a naïve estimate. In reality, there are complications. See next slides:

What spoils these estimates?

```
>gi|224514922|ref|NT_024477.14| Homo sapiens chromosome 12 genomic
contig, GRCh37.p13 Primary Assembly (displaying 3′ end)
CGGGAAATCAAAAGCCCCTCTGAATCCTGCGCACCGAGATTCTCCCCAGCCAAGGTGAGGCGGCAGCAGT
GGGAGATCCACACCGTAGCATTGGAACACAAATGCAGCATTACAAATGCAGACATGACACCGAAAATATA
ACACACCCCATTGCTCATGTAACAAGCACCTGTAATGCTAATGCACTGCCTCAAAACAAAATATTAATAT
AAGATCGGCAATCCGCACACTGCCGTGCAGTGCTAAGACAGCAATGAAAATAGTCAACATAATAACCCTA
ATAGTGTTAGGGTTAGGGTCAGGGTCCCGGTCCGGGTCGGGGTCCGGGTCCGGGGTCCGGGTCAGGGTGA
GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
GTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA
GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
GTTAGGGTTAGGGTTAGGGTTAG
```

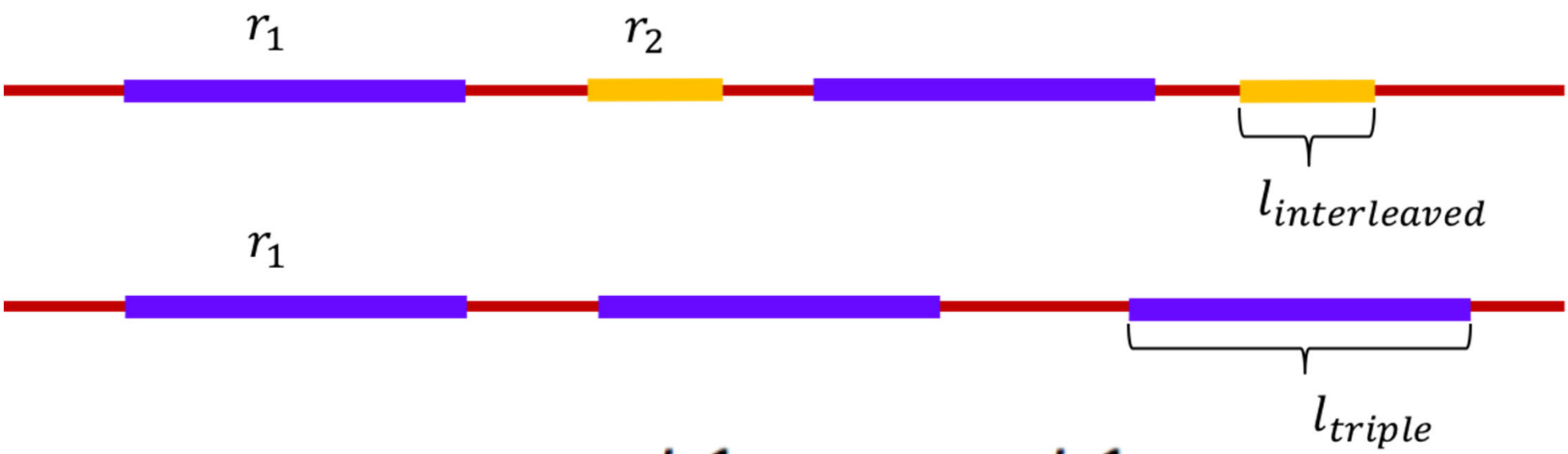
FIGURE 8.11 A BLASTN search of the human genome (all assemblies) database was performed at the NCBI website using TTAGGGTTAGGGTTAGGG as query (i.e., three TTAGGG repeats). There were matches to hundreds of genomic scaffolds. This figure shows an example (NT_024477.14) assigned to the telomere of chromosome 12q having many dozens of TTAGGG repeats. These occurred at the 3′ end of the genomic contig sequence.

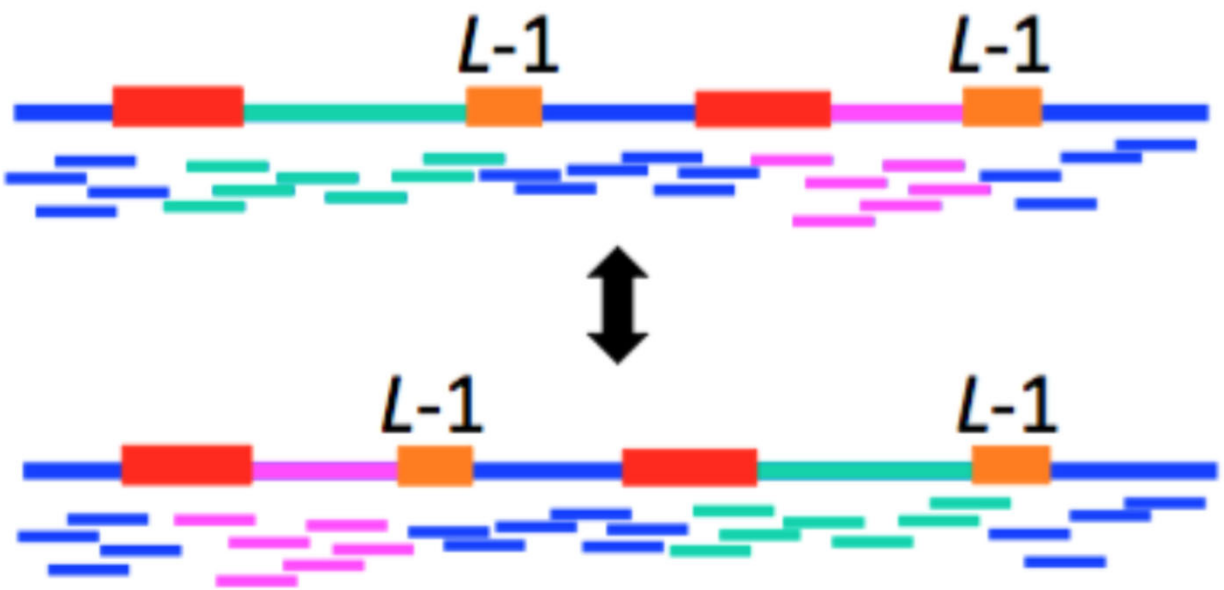
There were 100s of matches while one expects << 1 match:

$$2 \cdot 3\text{x}10^{9} \cdot 4^{-18}=0.08<<1$$

DNA repeats make assembly difficult

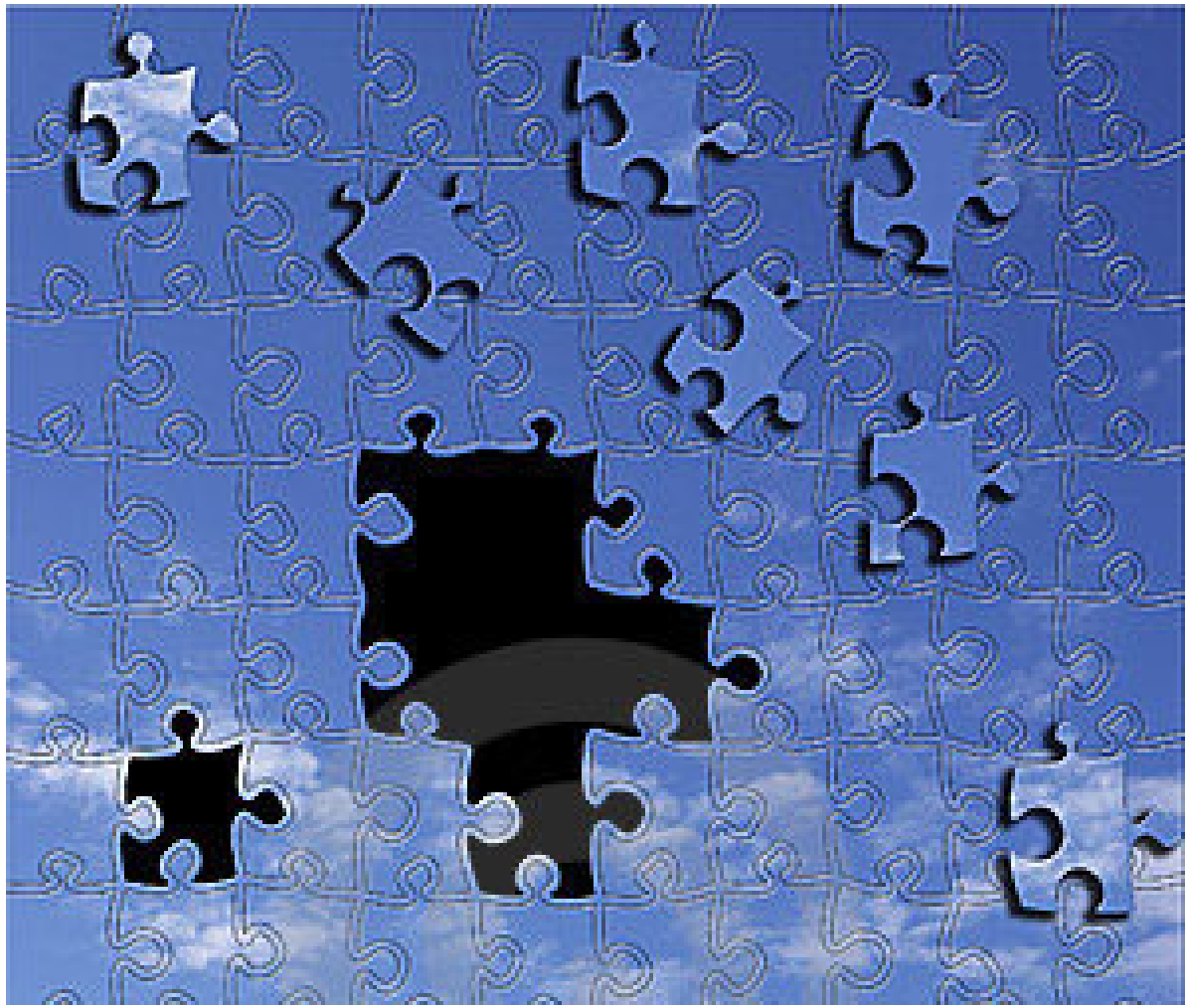
Why repeats make assembly difficult?



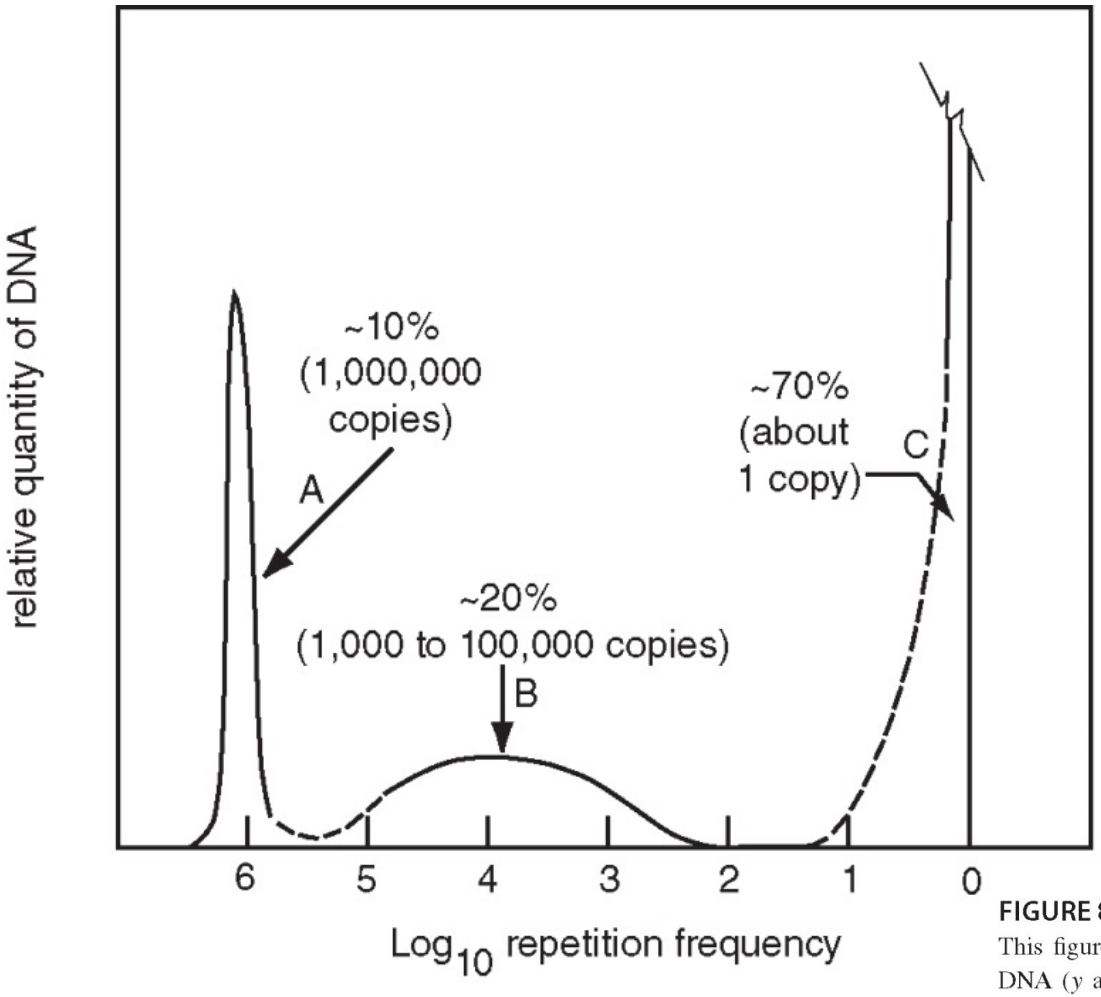


Images from the course EE 372: Data Science for High-Throughput Sequencing.
taught by David Tse at Stanford

Repeats are like sky puzzle pieces

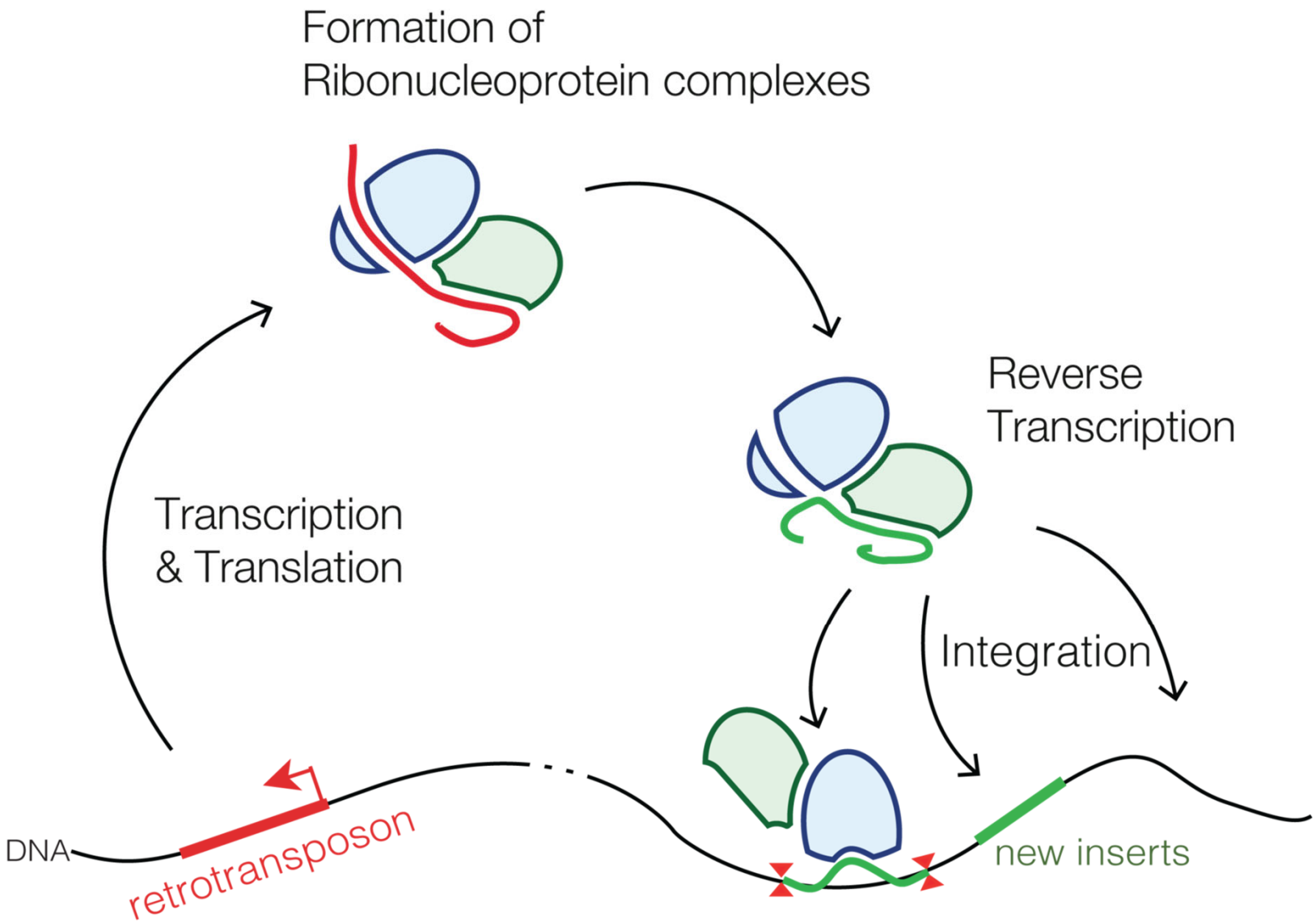


How many repeats are in eukaryotic genomes?

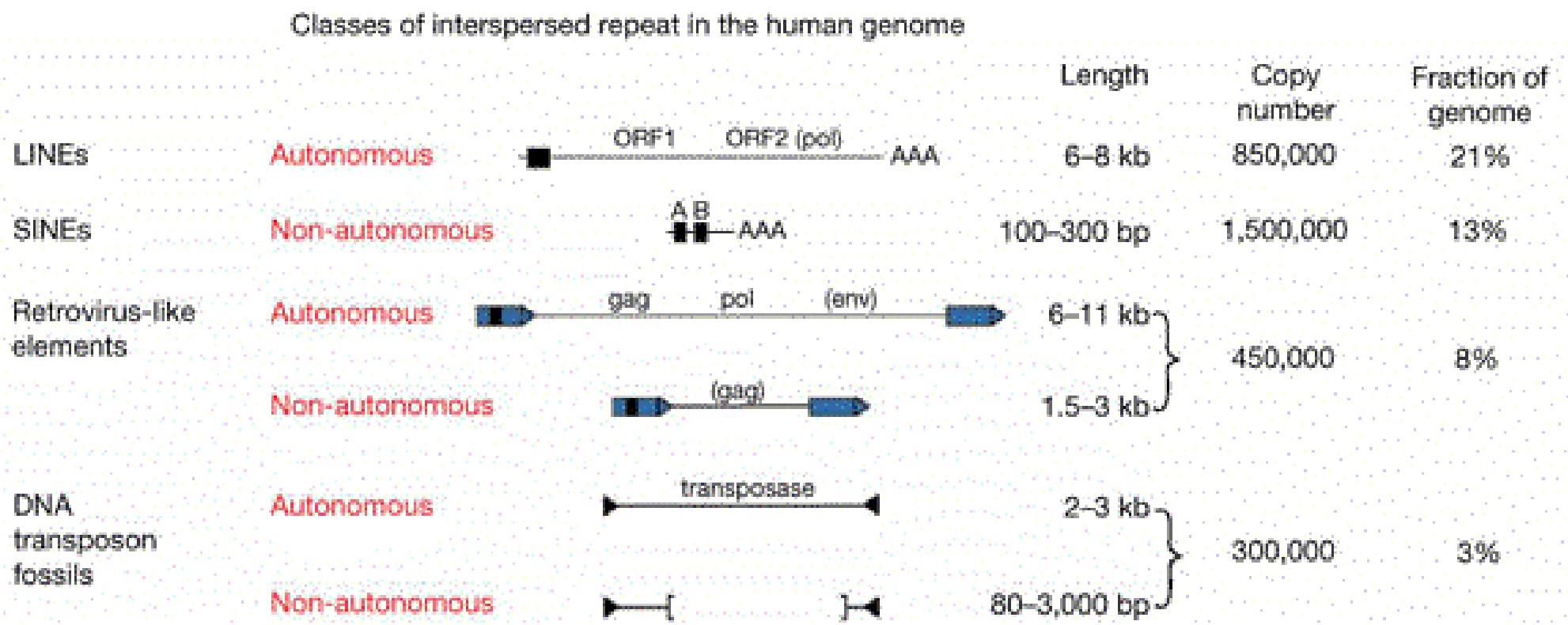


Data for mouse genome obtained in 1961 (sic!) using DNA denaturation and renaturation curves

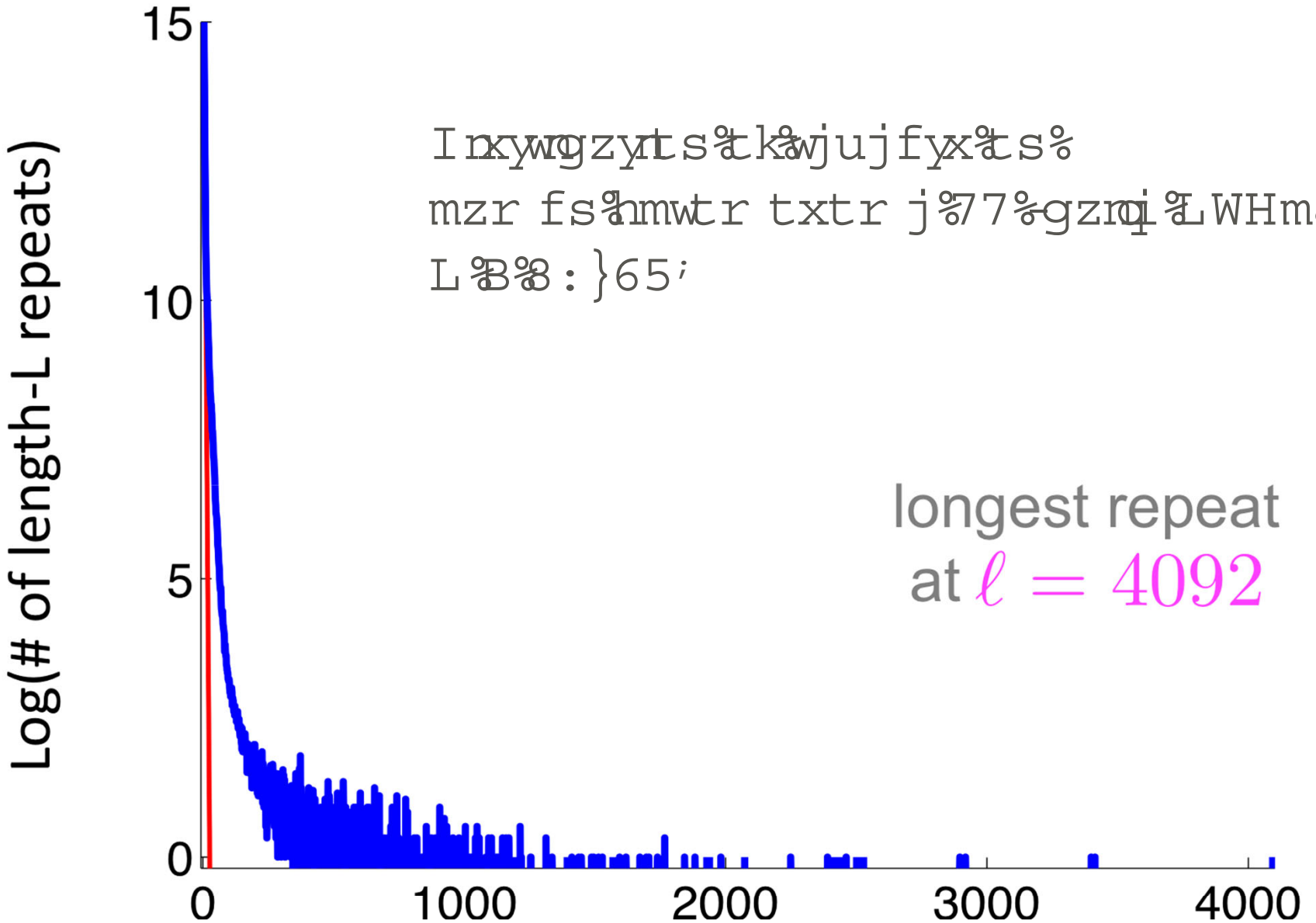
FIGURE 8.6 The complexity of genomic DNA can be estimated by denaturing then renaturing DNA. This figure (redrawn from Britten and Kohne, 1968) depicts the relative quantity of mouse genomic DNA (*y* axis) versus the logarithm of the frequency with which the DNA is repeated. The data are derived from a $C_0\,t_{1/2}$ curve, which describes the percent of genomic DNA that reassociates at particular times and DNA concentrations. A large $C_0\,t_{1/2}$ value implies a slower reassociation reaction. Three classes are apparent. The fast component accounts for 10% of mouse genomic DNA (arrow **A**), and represents highly repetitive satellite DNA. An intermediate component accounts for about 20% of mouse genomic DNA and contains repeats having from 1000 to 100,000 copies. The slowly reassociating component, comprising 70% of the mouse genome, corresponds to unique, single-copy DNA. Britten and Kohne (1968) obtained similar profiles from other eukaryotes, although distinct differences were evident between species. Used with permission.


Formation of
Ribonucleoprotein complexes
Reverse
Transcription
Integration
Transcription
& Translation
DNA
retrotransposon
new inserts

Almost all transposable elements in mammals fall into one of four classes



Slide by Ross Hardison, Penn State U.



Images from the course EE 372: Data Science for High-Throughput Sequencing.
taught by David Tse at Stanford

How to assemble a real genome with repeats?

Here we assume a "de novo" assembly without help from the previously assembled genomes

Nicolaas Govert de Bruijn (1918 – 2012) was a Dutch mathematician, noted for his many contributions in the fields of graph theory, analysis, number theory, combinatorics and logic

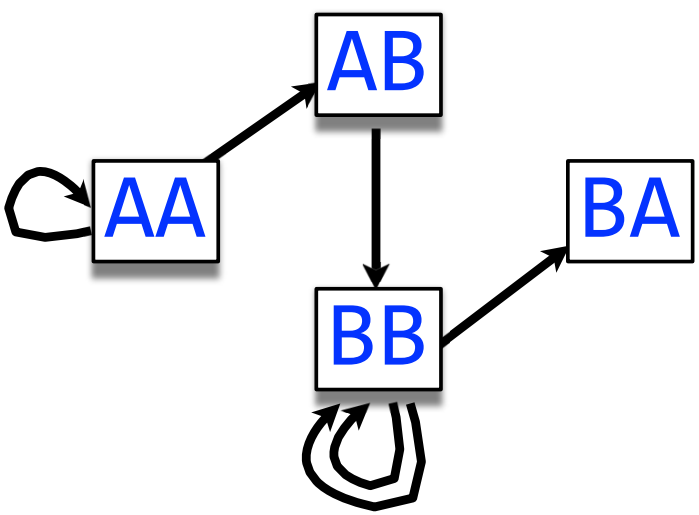
http://www.langmead-lab.org/teaching-materials/

De Bruijn graph

genome: AAABBBBA

3-mers: AAA, AAB, ABB, BBB, BBB, BBA

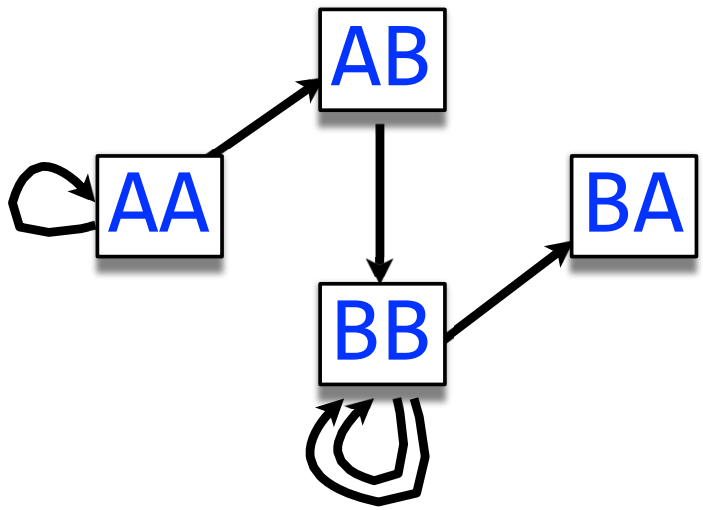
L/R 2-mers: AA, AA AA, AB AB, BB BB, BB BB, BB BB, BA



One edge per every *k*-mer

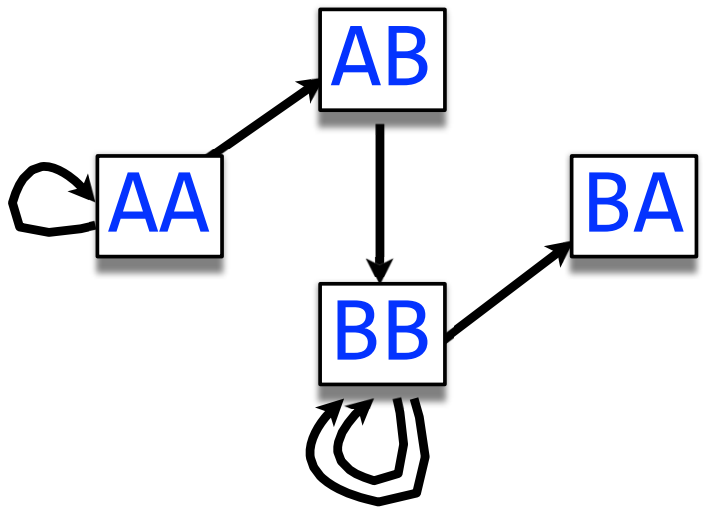
One node per distinct *k*-1-mer

De Bruijn graph



Walk crossing each edge exactly once gives a reconstruction of the genome

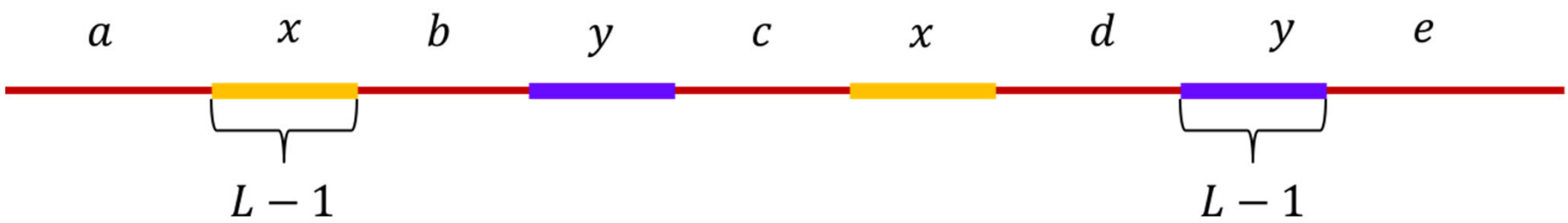
Assembly = Eulerian walk on De Bruijn graph

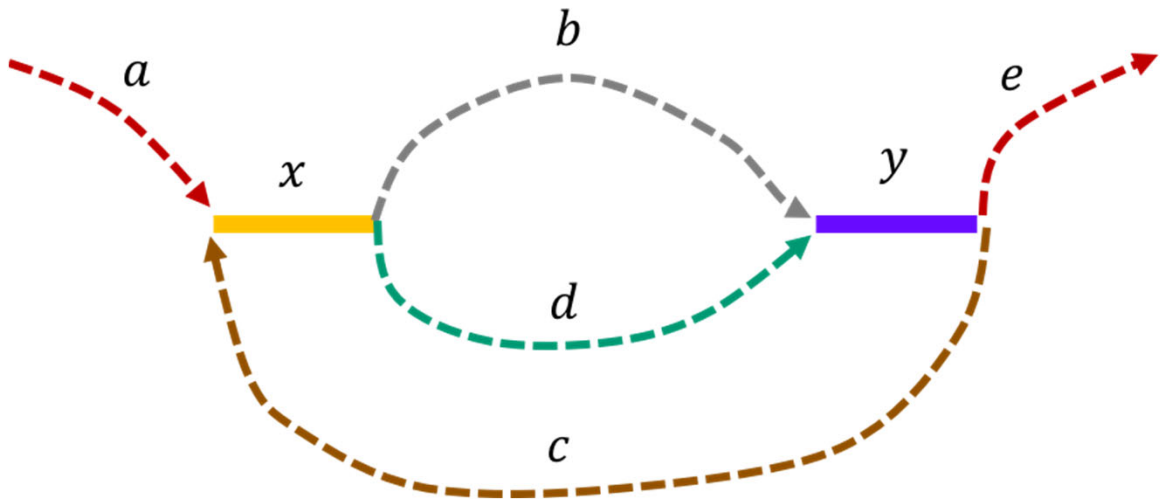


AAABBBBA

Walk crossing each edge exactly once gives a reconstruction of the genome. This is an *Eulerian walk*.

Why interleaved repeats are dangerous?



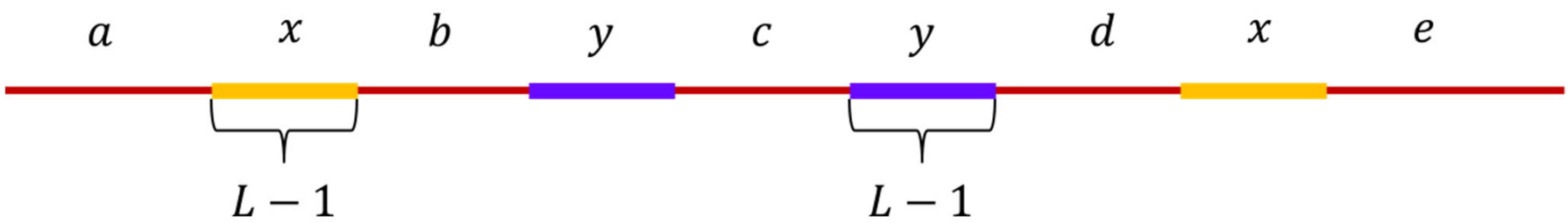


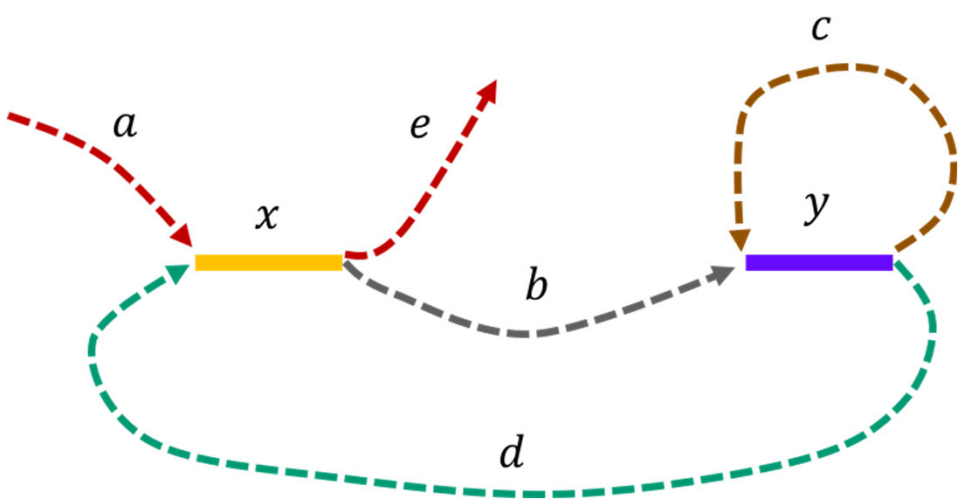
The two Eulerian paths that are on the graph:

a–x–b–y–c–x–d–y–e and a–x–d–y–c–x–b–y–e

Images from the course EE 372: Data Science for High-Throughput Sequencing.
taught by David Tse at Stanford

Why non-interleaved repeats are safe?

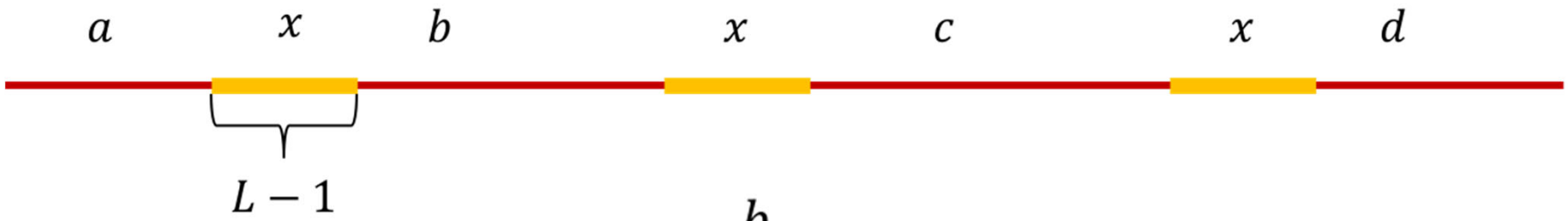


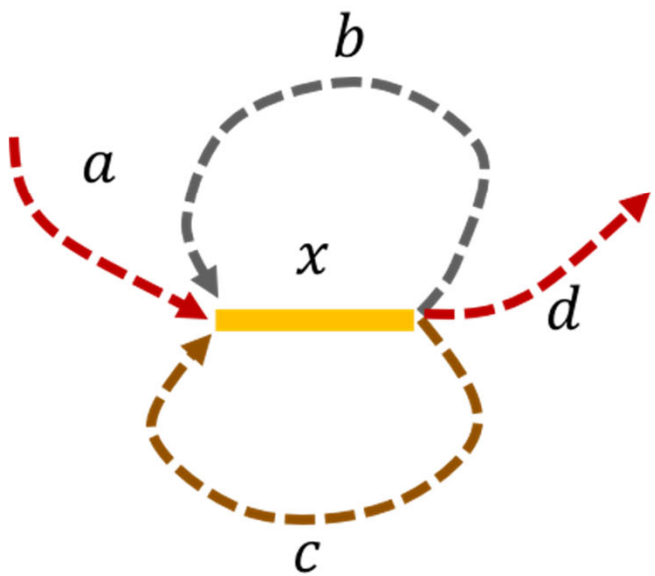


The only Eulerian path is: a–x–b–y–c–y–d–x–e

Images from the course EE 372: Data Science for High-Throughput Sequencing.
taught by David Tse at Stanford

Why triple repeats are dangerous?



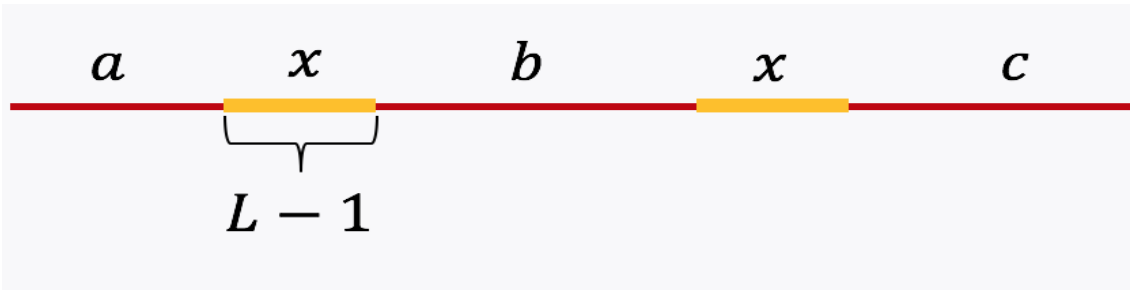


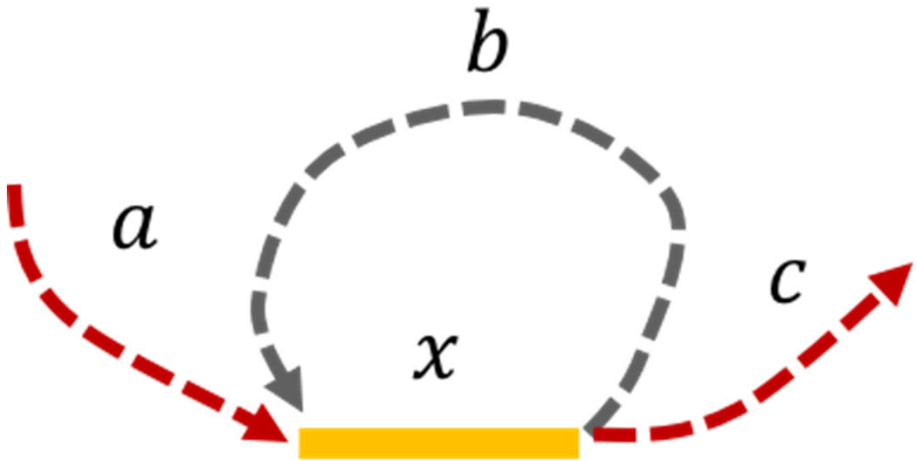
The two Eulerian paths that are on the graph:

a–x–b–x–c–x–d and a–x–c–y–b–x–d

Images from the course EE 372: Data Science for High-Throughput Sequencing.
taught by David Tse at Stanford

Why double repeats are safe?





The only Eulerian path is: a−x−b−x−c

Images from the course EE 372: Data Science for High-Throughput Sequencing.
taught by David Tse at Stanford

Pavel Pevzner's theorem

- **Theorem [Pevzner 1995]**:
  If L, the read length, is strictly greater than $\max(\ell_{interleaved}, \ell_{triple})$, then the de Bruijn graph has a unique Eulerian path corresponding to the original genome.

Pavel Pevzner
is the Ronald R. Taylor Chair and
Distinguished Professor of
Computer Science and Engineering
at University of California, San Diego.
His Alma Mater is
Moscow Institute of
Physics and Technology
in Russia.

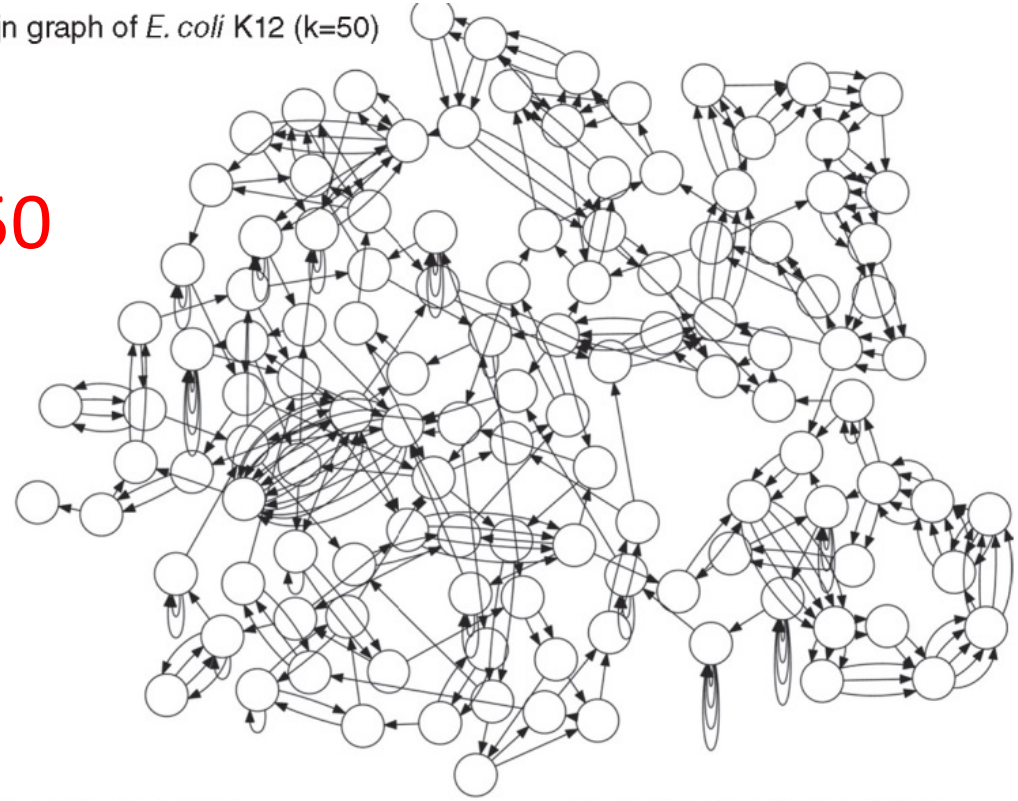
How to assemble a genome with repeats?

- Answer:
  longer reads
- But:
  cheap sequencing
  =
  short reads

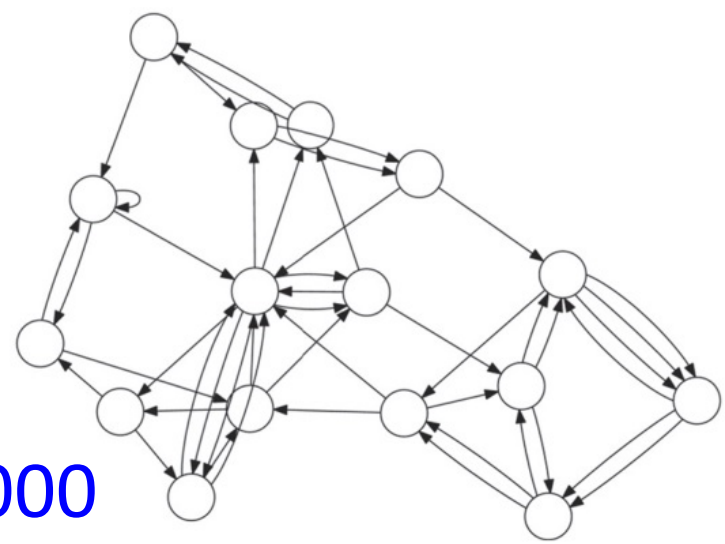
Technology	Read length (bp)
Roche 454	700
Illumina	50–250
SOLiD	50
Ion Torrent	400
Pacific Biosciences	>10,000

(a) de Bruijn graph of *E. coli* K12 (k=50)

k=50

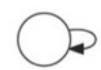


(b) de Bruijn graph (k=1,000)



k=1000

(c) de Bruijn graph (k=5,000)



k=5000

Credit: XKCD comics

QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY ARE THERE FEMALE MR MIMES
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING

WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE THERE SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE HELL IF GOD FORGIVES
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE

Geometric Distribution

- A series of Bernoulli trials with probability of success =*p*. continued **<u>until the first success</u>**. X is the number of trials.
- Compare to: Binomial distribution has:
  - Fixed number of trials =n.
  - Random number of successes = x.

$$P(X = x) = C_x^n p^x (1-p)^{n-x}$$

- Geometric distribution has reversed roles:
  - Random number of trials, *x*
  - Fixed number of successes, in this case 1.
  - Success always comes in the end: so no combinatorial factor C_x^n
  - $P(X=x) = p(1-p)^{x-1}$ where:

    $x-1$ = 0, 1, 2, … , the number of failures until the 1st success.
- NOTE OF CAUTION: Matlab, Mathematica, and many other sources use x to denote the number of failures until the first success. We stick with Montgomery-Runger notation

Geometric Mean & Variance

$$P(X=x) = p(1-p)^{x-1} = p \cdot q^{x-1}$$

$$S(p,q) = \sum_{x=1}^{\infty} P(X=x) = \frac{p}{1-q} = \frac{p}{p} = 1$$

$$q\frac{\partial S}{\partial q} = \sum (x-1)P(X=x) = \frac{pq}{(1-q)^2} = \frac{q}{p}$$

$$\langle x \rangle = \sum (x-1)P(X=x) + 1 = \frac{1-p}{p} + 1 = \frac{1}{p}$$

Geometric Mean & Variance

- If X is a geometric random variable (according to Montgomery-Bulmer) with parameter p,

$$\mu = E(X) = \frac{1}{p} \quad \text{and} \quad \sigma^2 = V(X) = \frac{(1-p)}{p^2} \qquad (3-10)$$

- For small p the standard deviation ~= mean

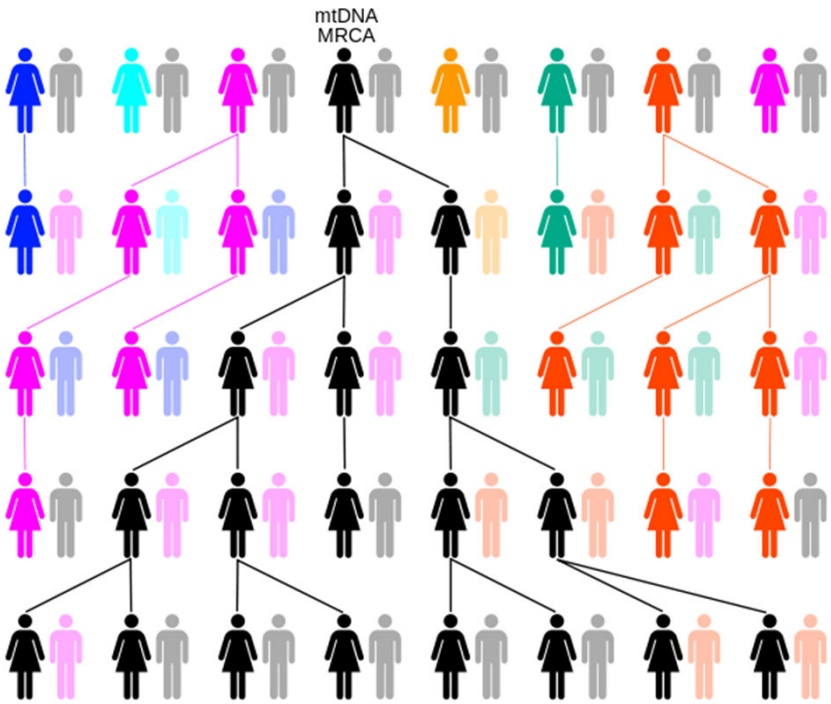
- Very different from Poisson, where it is

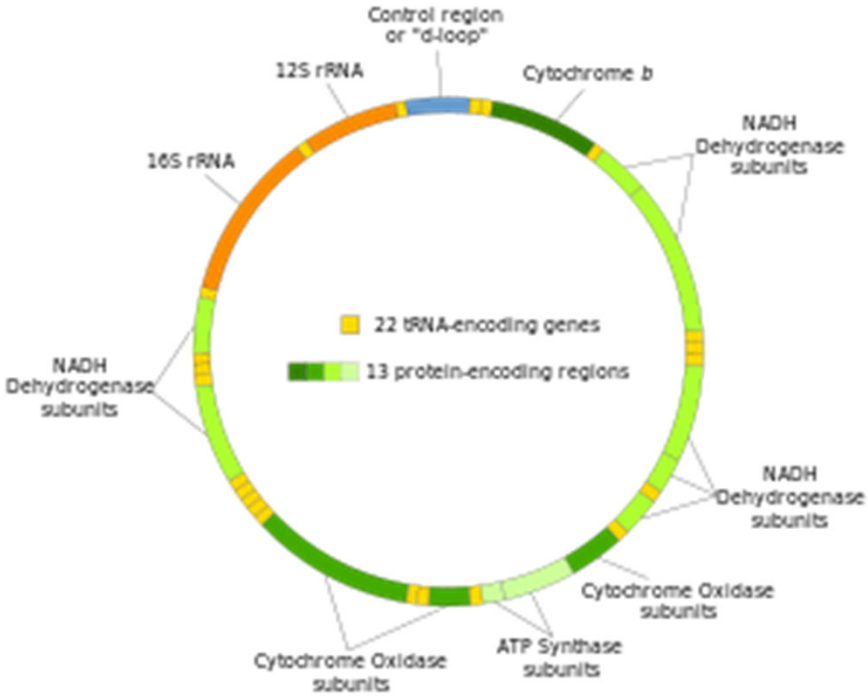
variance = mean and standard deviation = mean$^{1/2}$

Matlab exercise

- Find mean, variance, and histogram of 100,000 geometrically-distributed numbers with p=0.1
- Hint: Use help page for **<u>random</u>** command on how to generate geometrically-distributed random numbers

Geometric distribution in biology

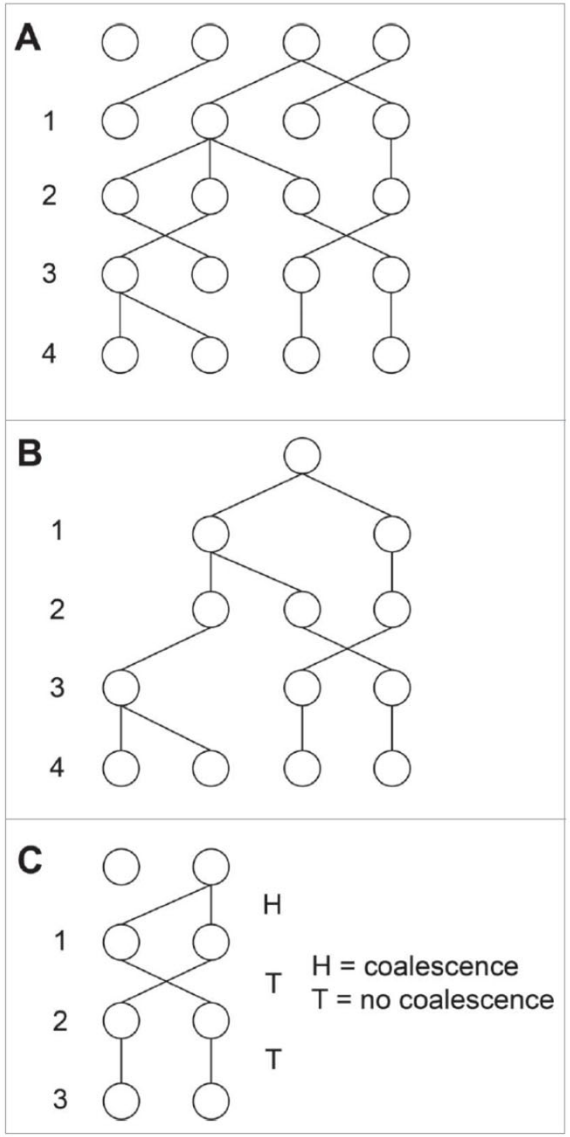




- Each of our cells has mitochondria with 16.5kb of mtDNA **inherited only from our mother**
- Human mtDNA has 37 genes encoding 13 proteins, 22+2 tRNA & rRNA
- Mitochondria appeared 1.5-2 billion years ago as a symbiosis between an alpha-proteobacterium (1000s of genes) and an archaeaon (of UIUC's Carl R. Woese fame)
- Since that time most mitochondrial genes were transferred into the nucleus
- Plants also have plastids with genomes related to cyanobacteria

Time to the last common (maternal) ancestor follows geometric distribution

- Constant population of N women
- Random number of (female) offsprings. Average is 1 (but can be 0 or 2)
- Randomly pick two women.
  Question: how many generations T since their last maternal ancestor?
- T is a random variable What is its PMF: P(T=t)?
  Answer: P(T=t) follows a geometric distribution
- Do these two women have the same mother? Yes: "success" in finding their last common ancestor (p=1/N). P(T=1)=1/N.
- No? "failure" (1-p=1-1/N). Go to their mothers and repeat the same question.
- $P(T=t)=(1-1/N)^{t-1}(1/N) \approx (1/N)\exp(-(t-1)/N)$
- t can be inferred from the density of differences on mtDNA =2μt



Maddamsetti R, MOBILE GENETIC ELEMENTS, 6, e1137380(2016)

A gallery of useful
discrete probability distributions

Geometric Distribution

- A series of Bernoulli trials with probability of success =*p*. continued **<u>until the first success</u>**. X is the number of trials.
- Compare to: Binomial distribution has:
  - Fixed number of trials =n.
  - Random number of successes = x.

$$P(X=x)=C_x^n p^x (1-p)^{n-x}$$

- Geometric distribution has reversed roles:
  - Random number of trials, *x*
  - Fixed number of successes, in this case 1.
  - Success always comes in the end: so no combinatorial factor C_x^n
  - $P(X=x) = p(1-p)^{x-1}$ where:

    $x-1$ = 0, 1, 2, … , the number of failures until the 1st success.
- NOTE OF CAUTION: Matlab, Mathematica, and many other sources use x to denote the number of failures until the first success. We stick with Montgomery-Runger notation

Geometric Mean & Variance

- If *X* is a geometric random variable (according to Montgomery-Bulmer) with parameter *p*,

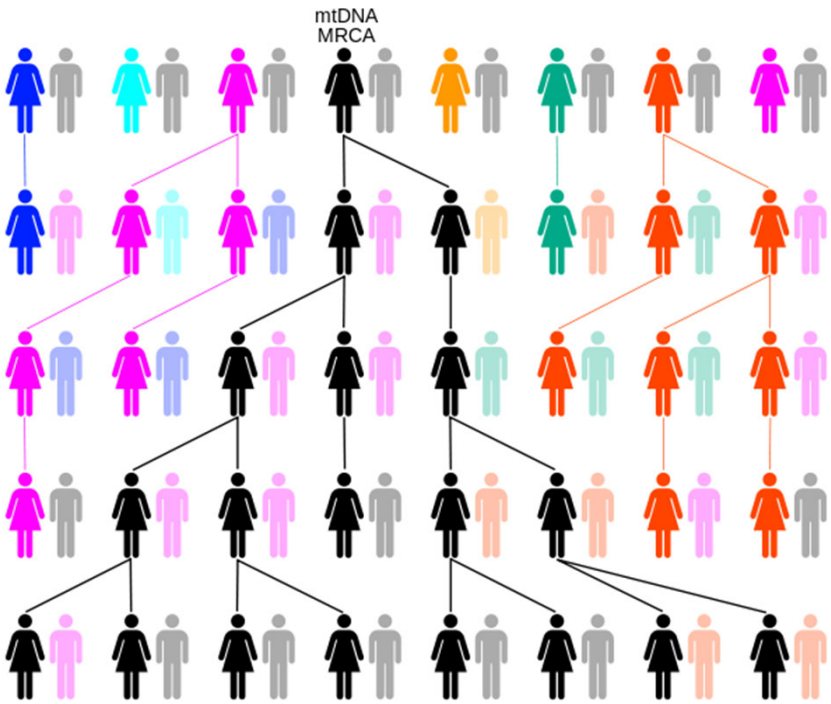
$$\mu = E(X) = \frac{1}{p} \quad \text{and} \quad \sigma^2 = V(X) = \frac{(1-p)}{p^2} \qquad (3\text{-}10)$$

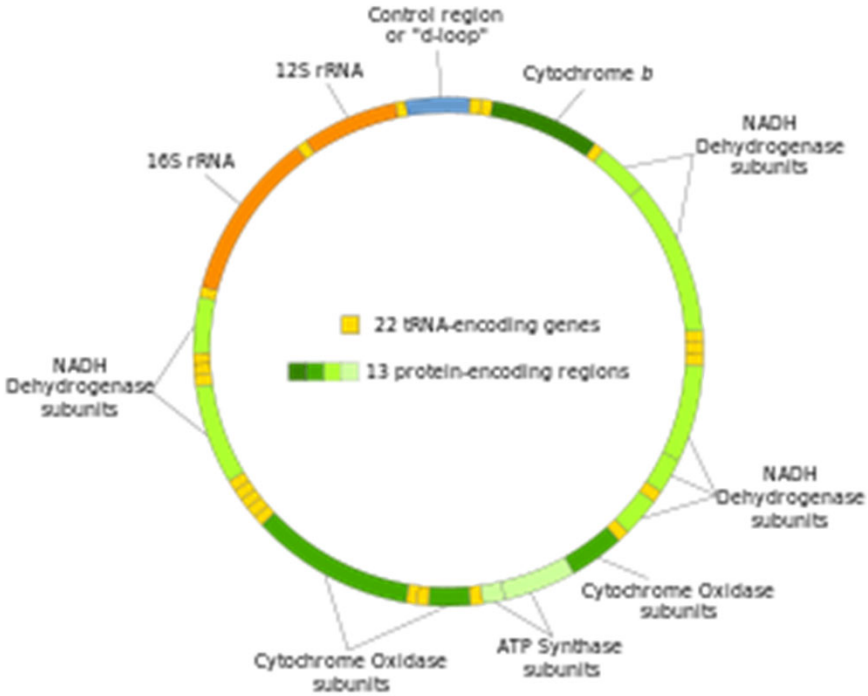
- For small *p* the standard deviation $=(1-p)^{0.5}/p$ ~= mean=1/p

- Very different from Binomial and Poisson, where

variance = mean and standard deviation = $\text{mean}^{1/2}$

Geometric distribution in biology

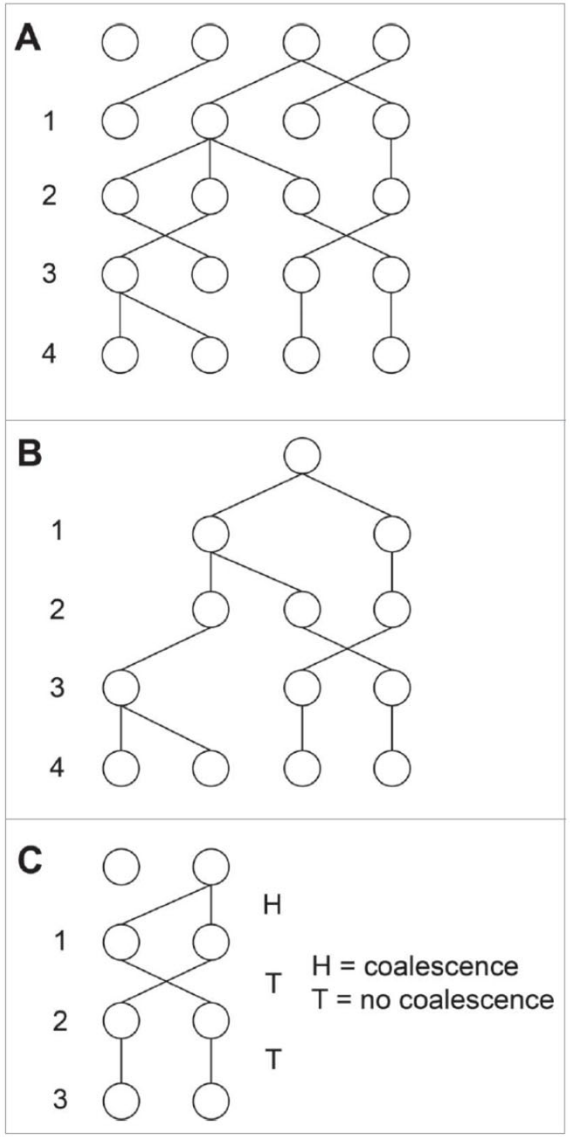




- Each of our cells has mitochondria with 16.5kb of mtDNA **inherited only from our mother**
- Human mtDNA has 37 genes encoding 13 proteins, 22+2 tRNA & rRNA
- Mitochondria appeared 1.5-2 billion years ago as a symbiosis between an alpha-proteobacterium (1000s of genes) and an archaeaon (of UIUC's Carl R. Woese fame)
- Since that time most mitochondrial genes were transferred into the nucleus
- Plants also have plastids with genomes related to cyanobacteria

Time to the last common (maternal) ancestor follows geometric distribution

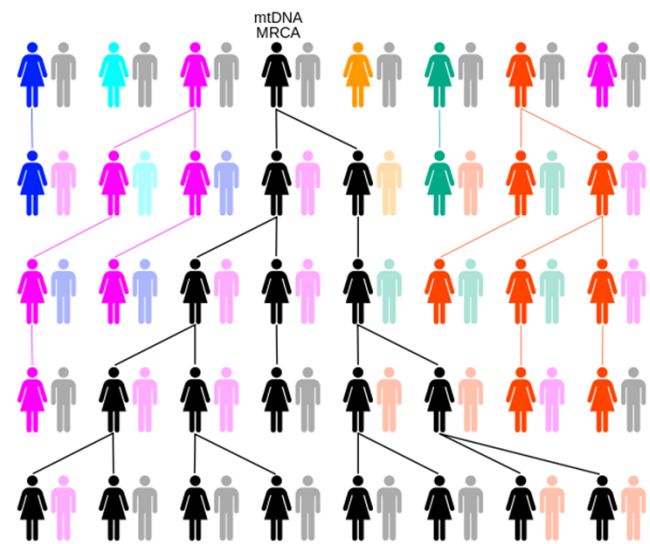
- Constant population of N women
- Random number of (female) offsprings. Average is 1 (but can be 0 or 2)
- Randomly pick two women.
  Question: how many generations T since their last maternal ancestor?
- T is a random variable What is its PMF: P(T=t)?
  Answer: P(T=t) follows a geometric distribution
- Do these two women have the same mother? Yes: "success" in finding their last common ancestor (p=1/N). P(T=1)=1/N.
- No? "failure" (1-p=1-1/N). Go to their mothers and repeat the same question.
- $P(T=t)=(1-1/N)^{t-1}(1/N) \approx (1/N)\exp(-(t-1)/N)$
- t can be inferred from the density of differences on mtDNA =2μt



Maddamsetti R, MOBILE GENETIC ELEMENTS, 6, e1137380(2016)

Most Recent Common Ancestor (MRCA)

- Start with N individuals. Unit of time is N generations (time for one pair to merge) since $E(T)=\sum_{t=1}^{\infty} t \cdot (1/N)\, exp(-t/N)=N$
- Any of $\frac{N(N-1)}{2}$ pairs can merge first. The average time for the first pair to merge is $\frac{2}{N(N-1)}$
- After merger $N \to N-1$,
- so time until the next
- merger is $\frac{2}{(N-1)(N-2)}$



Most Recent Common Ancestor (MRCA)

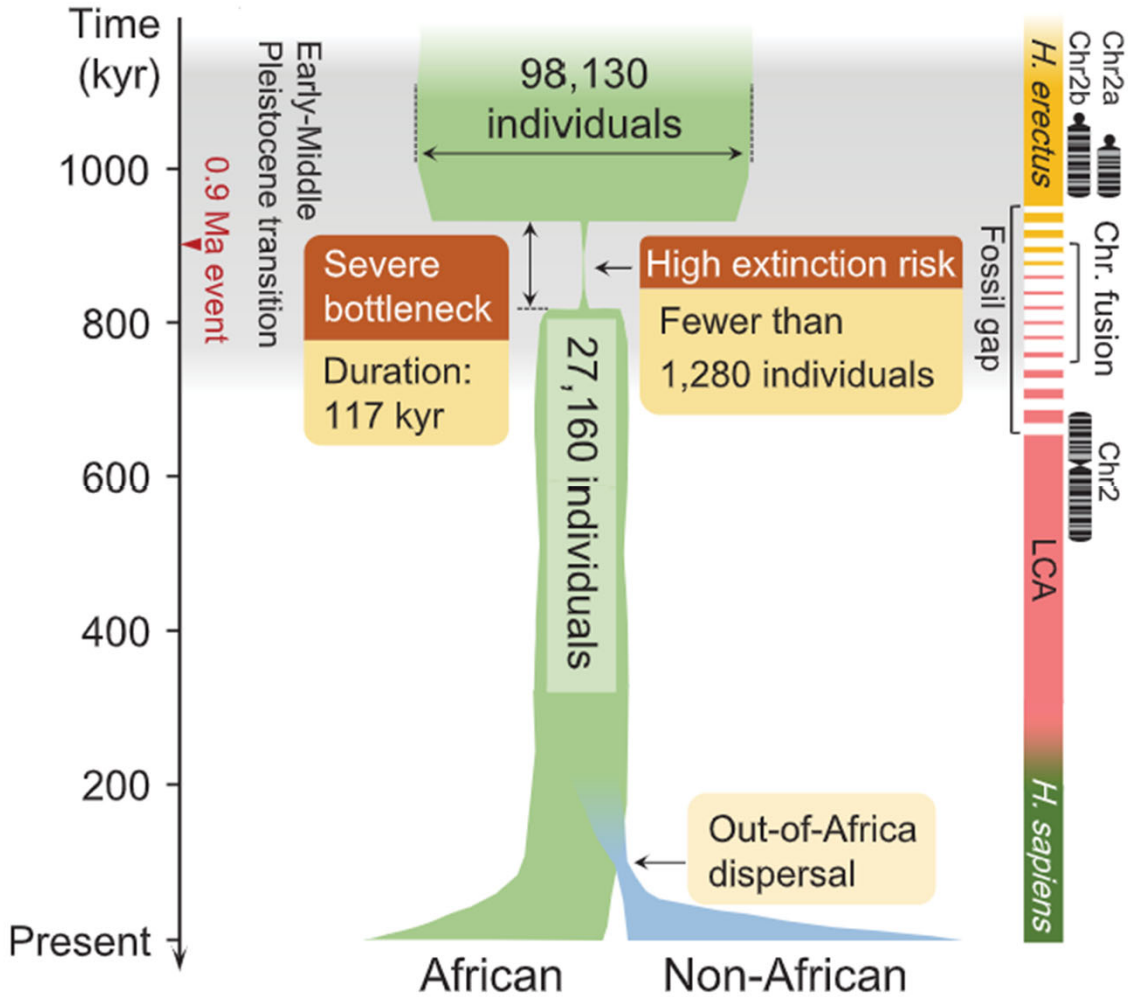
Total time until the MRCA

$$T_{MRCA} = N \cdot \sum_{k=2}^{N} \frac{2}{k(k-1)}$$

$$= 2N \sum_{k=2}^{N} \left(\frac{1}{k-1} - \frac{1}{k} \right) = 2N \left(1 - \frac{1}{N} \right) \approx 2N$$

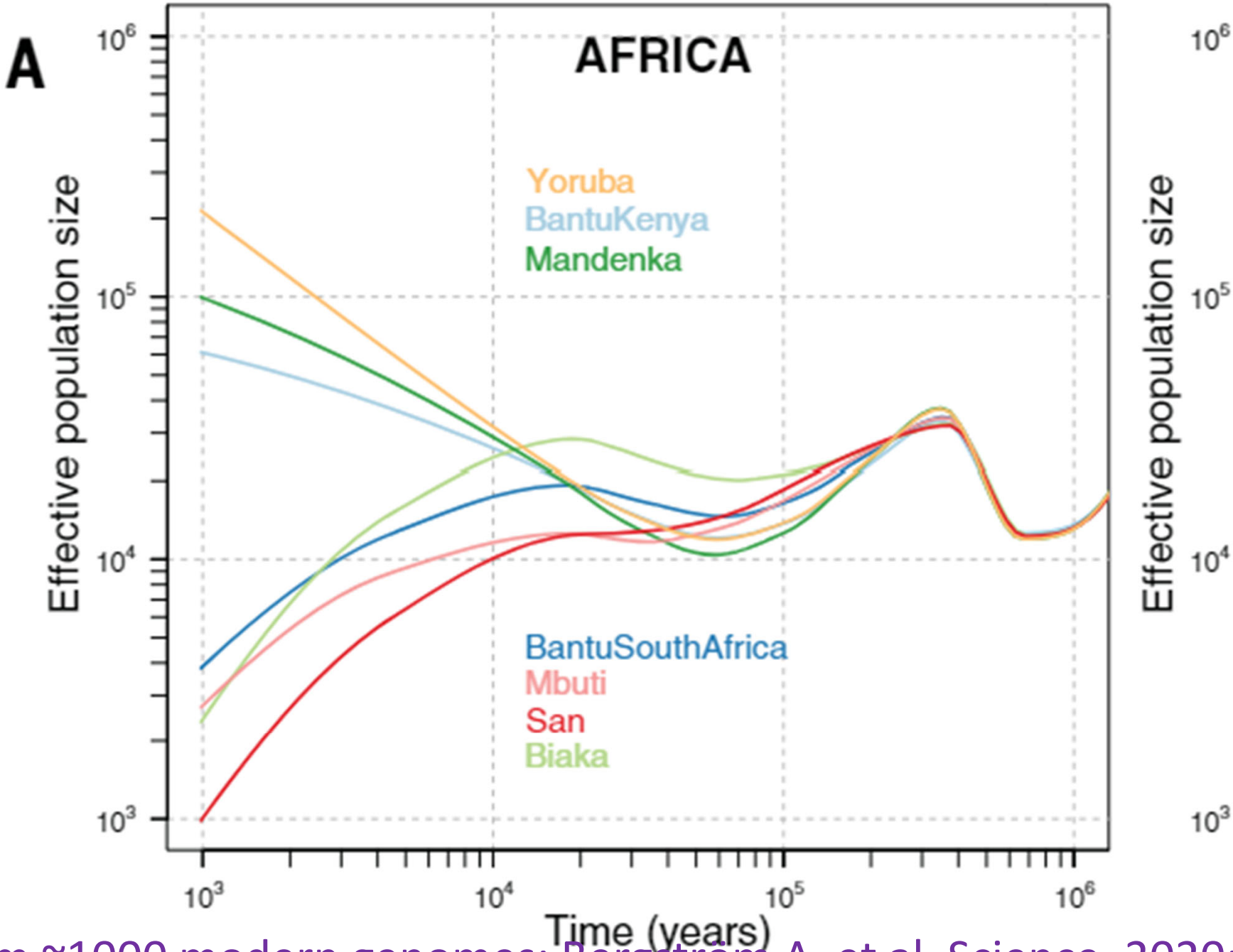
- There are about $N=3.5x10^{9}$ women living today
- Most Recent maternal Common Ancestor (MRCA)
  of all people living today lived $T_{MRCA}=2N$ generations ago
- $T_{MRCA} = 2 \cdot 3.5x10^{9}$ generations
- If the generation time 20 years it is 140 billion years > 10 times the time since the Big Bang.
- Something is wrong here!

Hot off the press: human ancestors almost got extinct about 1M years ago



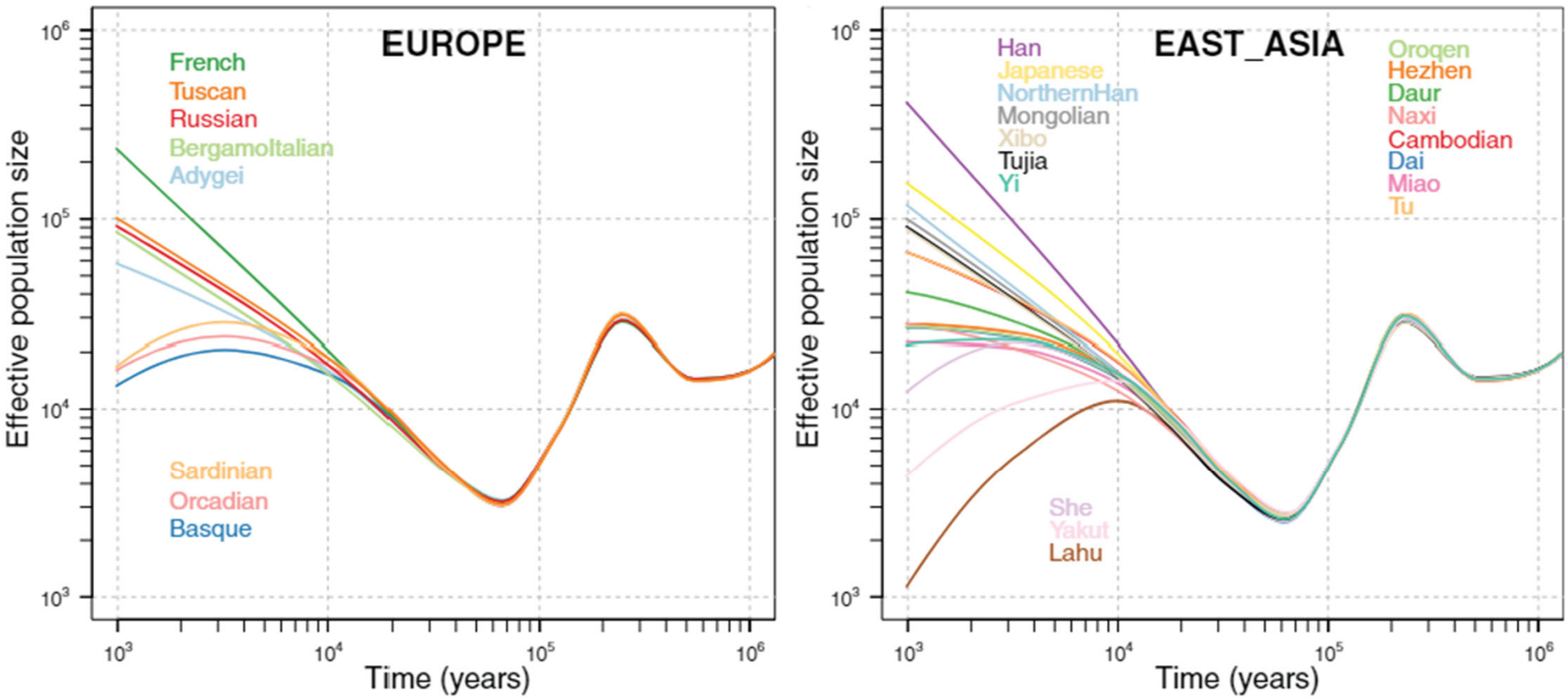
Hu W, et al. Science. 2023;381: 979–984

Effective human population size ~10,000



From ~1000 modern genomes: Bergström A, et al. Science. 2020;367

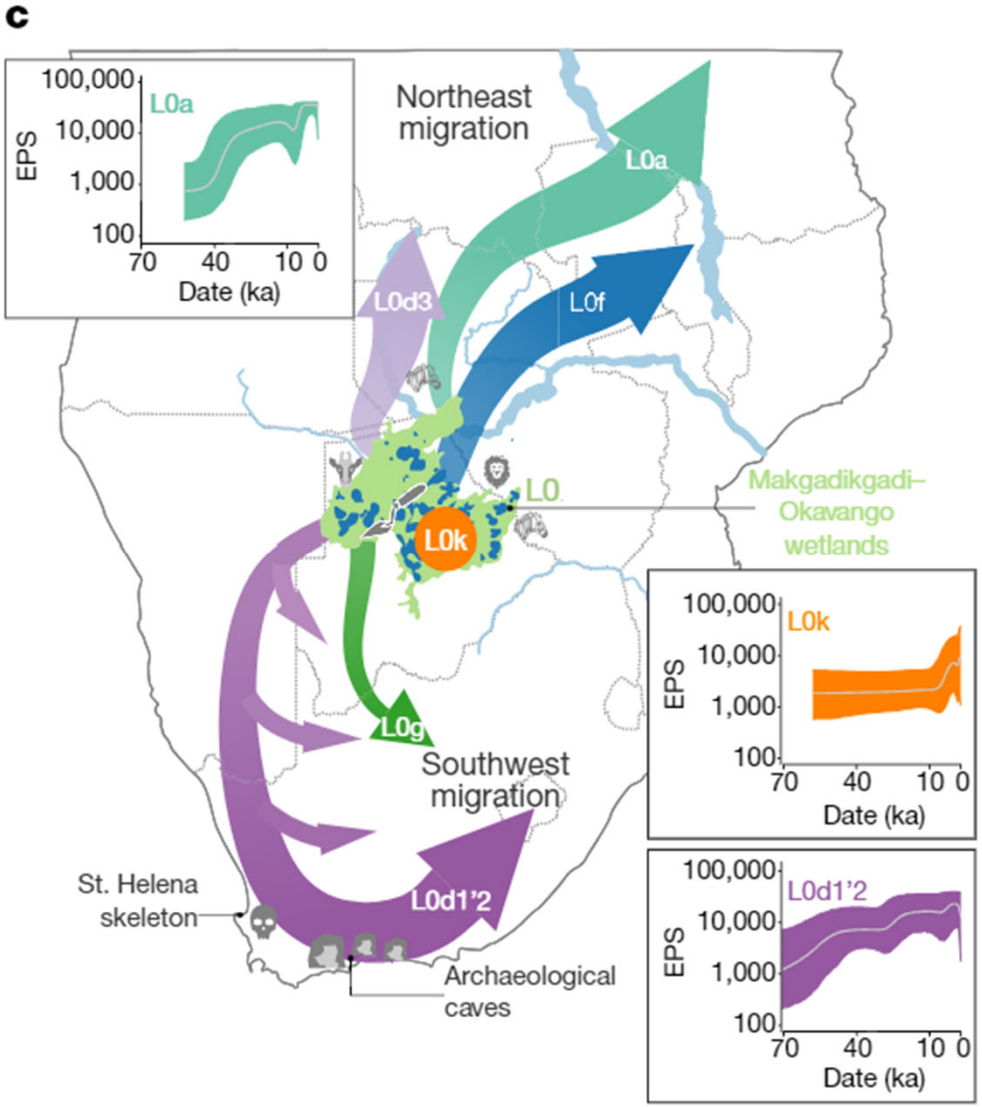
Effective human population size in Europe and Asia ~3000 people ~60,000 years ago



From ~1000 modern genomes: Bergström A, et al. Science. 2020;367

- Population is not constant and for a long time was very low
- Change N to the "effective" size N_e
- Current thinking is that for all of us including people of African ancestry N_e~10,000 people
- For humans of European + Asian ancestry N_e~ 3000 people
- Mito Eve lived in Africa ~2*(Ne/2)*20 years=10,000*20 years= 200,000 years ago

“Mitochondrial Eve” lived in Africa

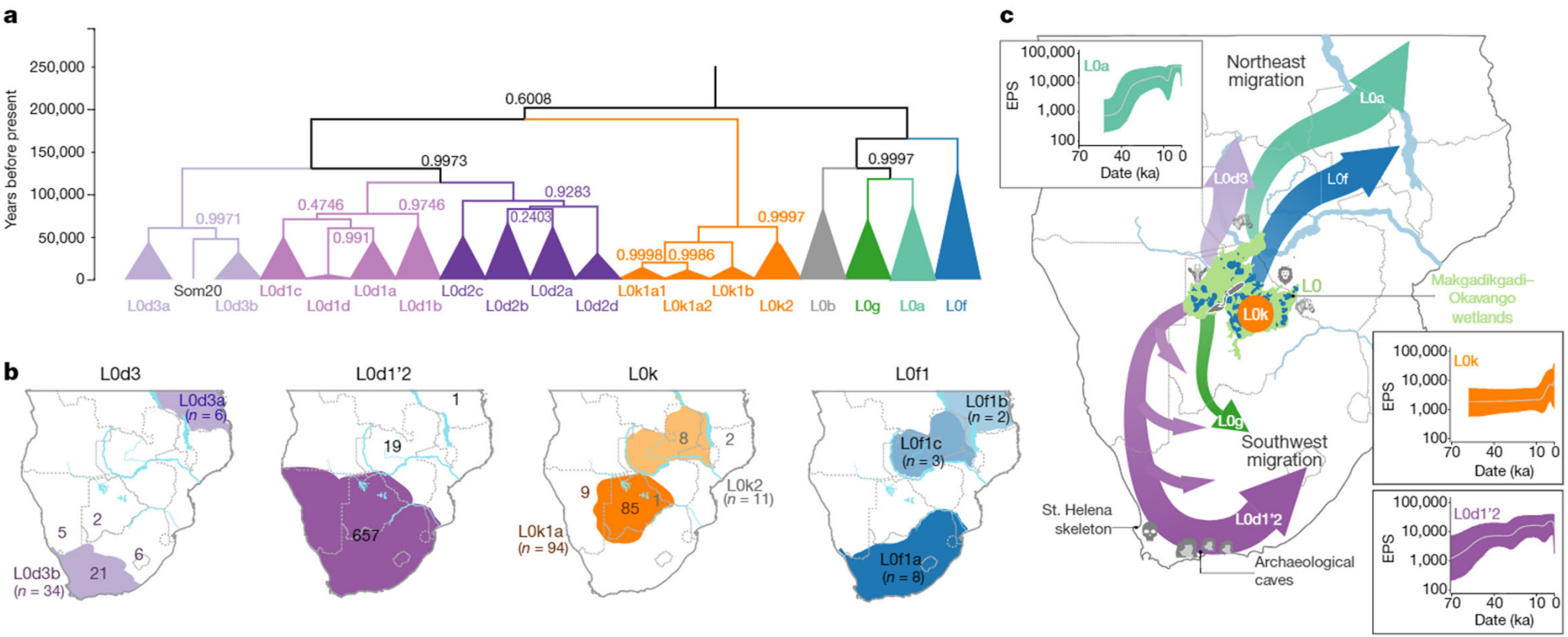


“Mitochondrial Eve” lived in Makgadikgadi–Okavango paleo-wetland of southern Africa ~200,000 years ago (between 165,000 and 240,000 years ago)

Chan EKF, et al. Nature. 2019; 575: 185–189.

Okavango Delta now

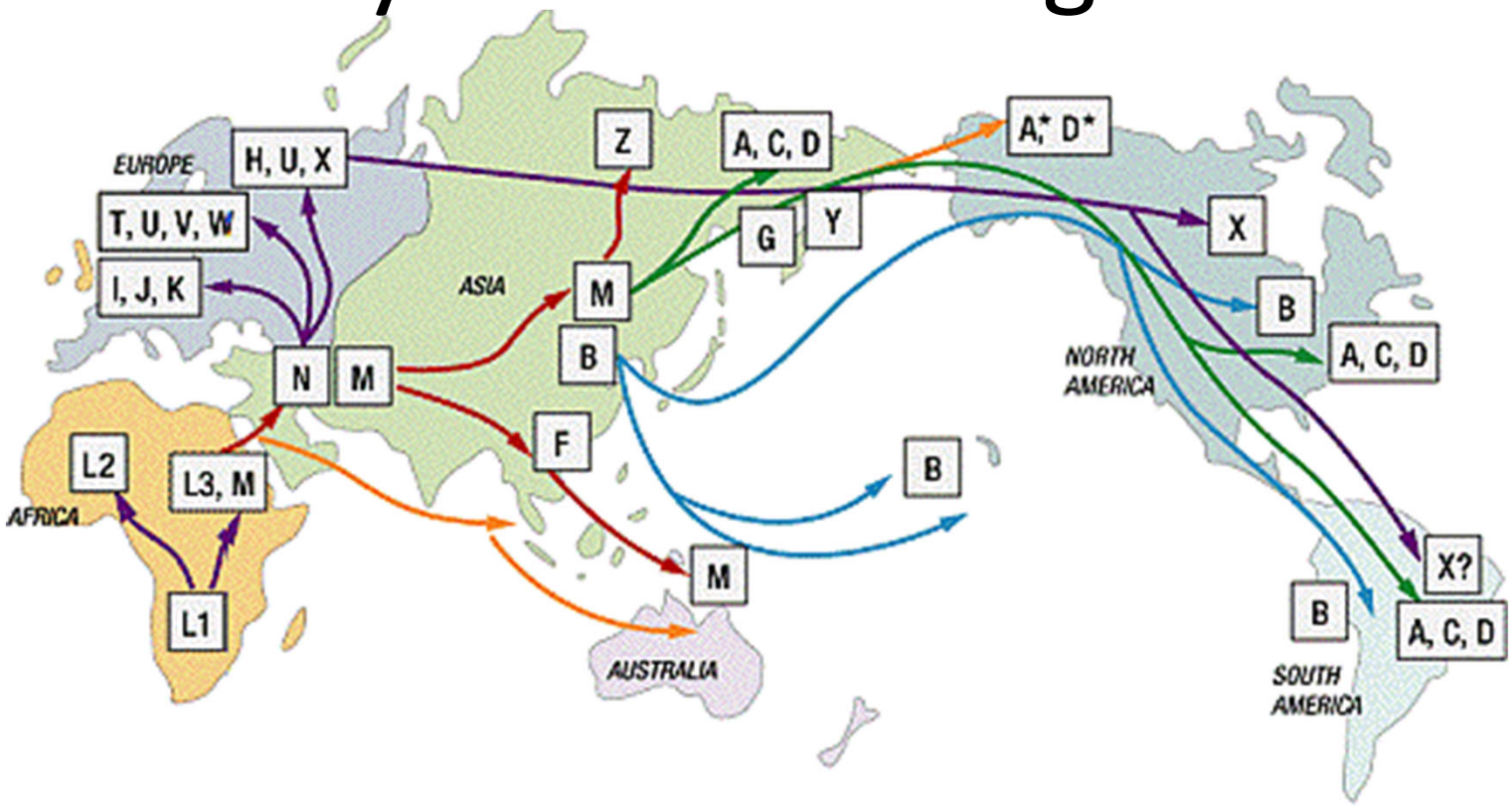
“Mitochondrial Eve” lived in Africa



“Mitochondrial Eve” lived in Makgadikgadi–Okavango paleo-wetland of southern Africa ~200,000 years ago (between 165,000 and 240,000 years ago)

Chan EKF, et al. Nature. 2019; 575: 185–189.

Modern mitochondrial DNA contains history of human migrations



EXPANSION TIMES *(years ago)*	
Africa	120,000 - 150,000
Out of Africa	55,000 - 75,000
Asia	40,000 - 70,000
Australia/PNG	40,000 - 60,000
Europe	35,000 - 50,000
Americas	15,000 - 35,000
Na-Dene/Esk/Aleuts	8,000 - 10,000

FamilyTreeDNA
mtDNA Migrations Map

*Poznik GD, et al (Carlos Bustamante lab in Stanford), Science **341**: 562 (August 2013).*

What about men?

- Y-chromosome is transferred from father to son
- Like mitochondria it can be used to trace ancestry of all men to the "Y-chromosome Adam"
- Where did "Adam'" live? Did he meet the "mitochondrial Eve"?

Y-chromosomal Adam also lived in Africa

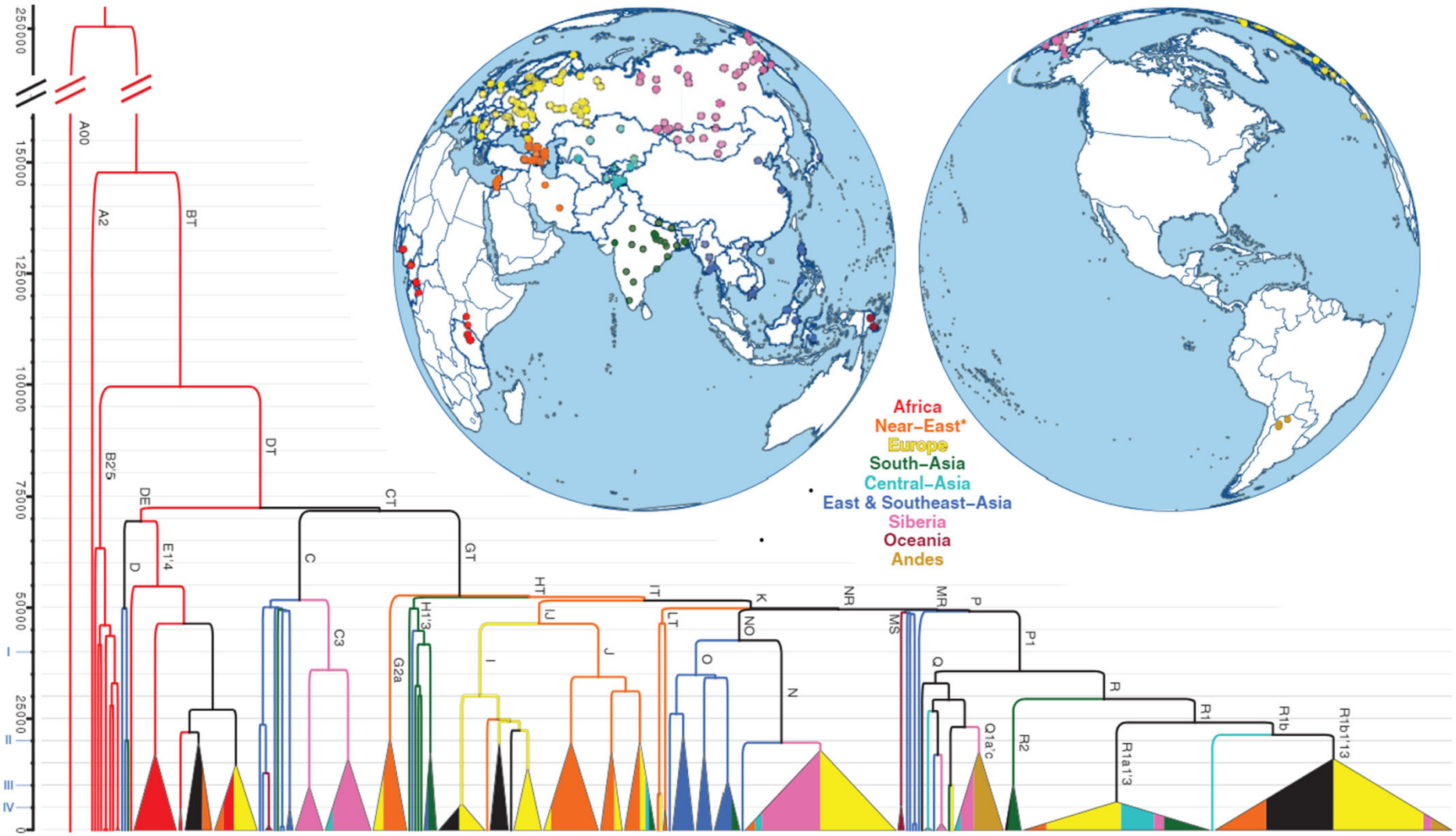
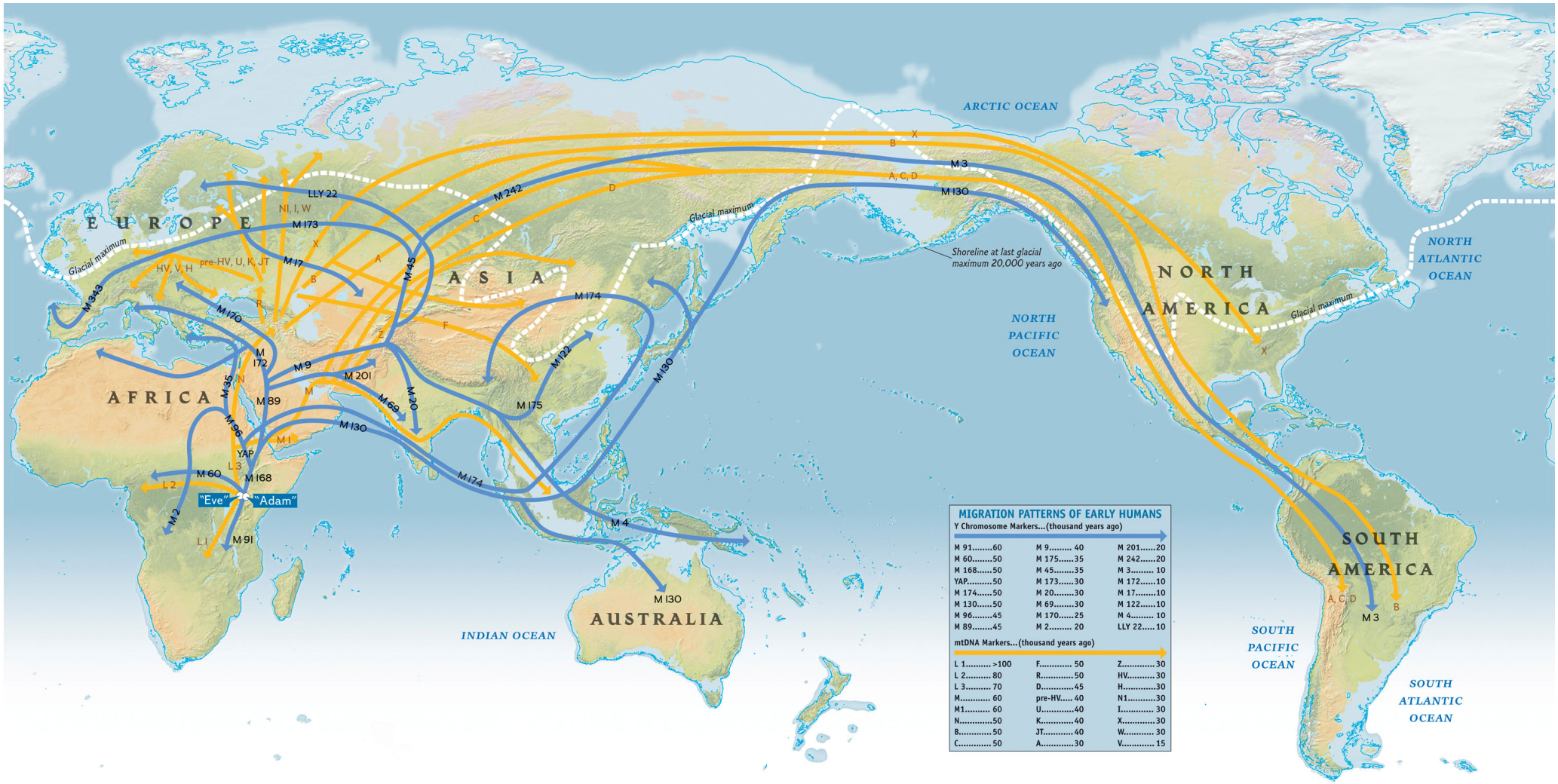


Figure 1. The phylogenetic tree of 456 whole Y chromosome sequences and a map of sampling locations. The phylogenetic tree is reconstructed using BEAST. Clades coalescing within 10% of the overall depth of the tree have been collapsed. Only main haplogroup labels are shown (details are provided in Supplemental Information 6). Colors indicate geographic origin of samples (Supplemental Table S1), and fill proportions of the collapsed clades represent the proportion of samples from a given region. Asterisk (*) marks the inclusion of samples from Caucasus area. Personal Genomes Project (http://www.personalgenomes.org) samples of unknown and mixed geographic/ethnic origin are shown in black. The proposed structure of Y chromosome haplogroup naming (Supplemental Table S5) is given in Roman numbers on the *y*-axis.

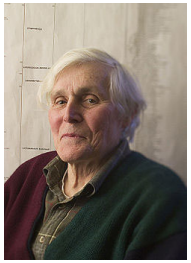
Karmin M, Saag L, Vicente M, Sayres MAW, Järve M, Talas UG, et al. Genome Res. 2015;25: 459–466.

“Adam” and “Eve” both lived in Africa

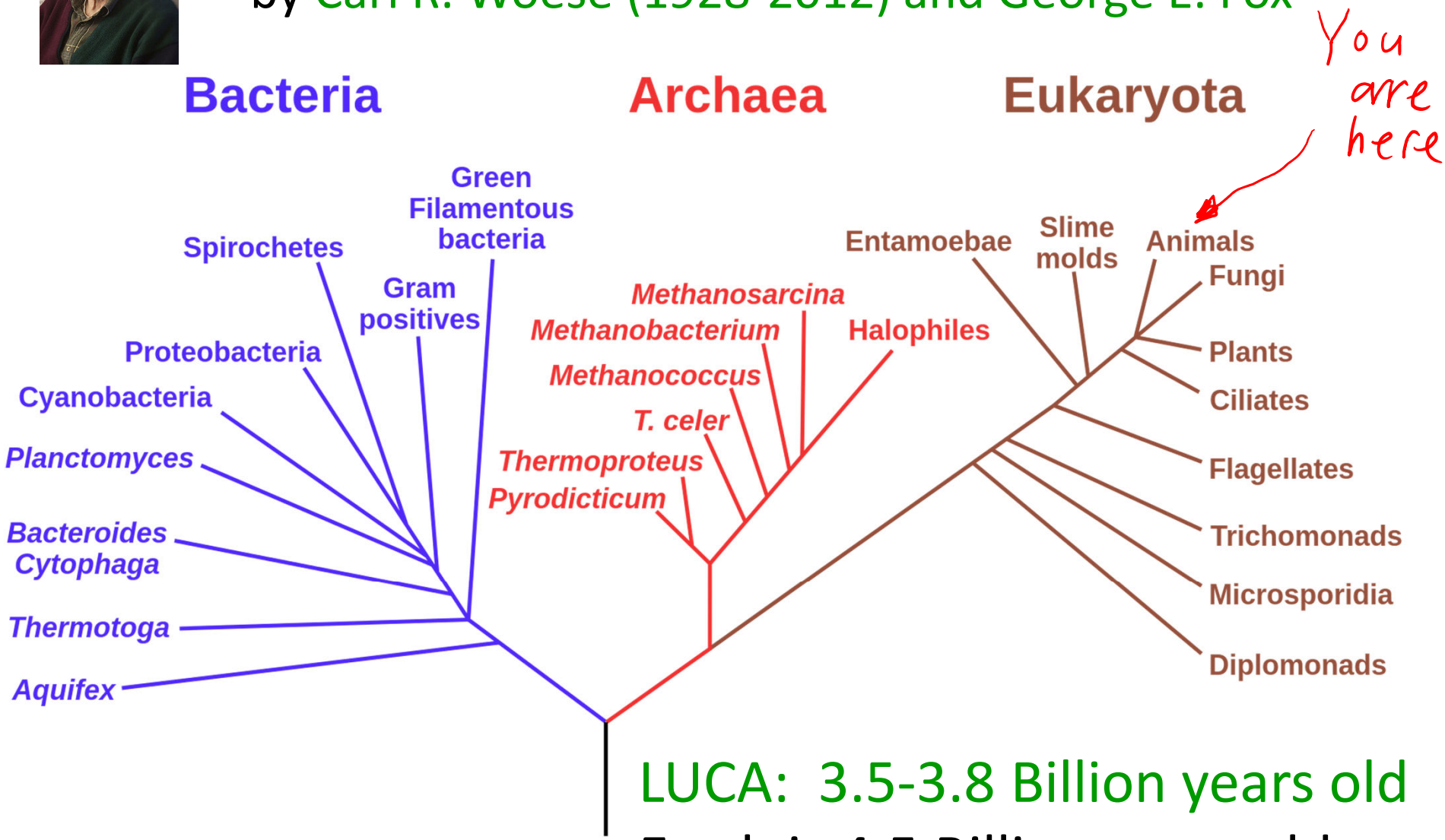


- “Mitochondrial Eve” lived in Africa between 100,000 and 240,000 years ago
- “Y-chromosome Adam” also lived in Africa between 120,000 and 160,000 years ago
- *Poznik GD, et al (Carlos Bustamante lab in Stanford), Science **341**: 562 (August 2013).*

Last Universal Common Ancestor (LUCA)



Archaea were discovered here at UIUC in 1977
by Carl R. Woese (1928-2012) and George E. Fox



LUCA: 3.5-3.8 Billion years old

Earth is 4.5 Billion years old

Credit: XKCD comics

QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY ARE THERE FEMALE MR MIMES
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING

WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE THERE SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE HELL IF GOD FORGIVES
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE

Negative Binomial Definition

- In a series of independent trials with constant probability of success, p, let the random variable X denote the number of trials until r successes occur. Then X is a negative binomial random variable with parameters:
  $0 < p < 1$ and r = 1, 2, 3, ...
- The probability mass function is:

$$f(x) = C_{r-1}^{x-1} p^r (1-p)^{x-r} \quad \text{for } x = r, r+1, r+2 \ldots \qquad (3\text{-}11)$$

- Compare it to binomial

$$f(x) = C_x^n p^x (1-p)^{n-x} \quad \text{for } x = 1, 2, \ldots \text{ n}$$

NOTE OF CAUTION: Matlab, Mathematica, and many other sources use x to denote the number of failures until one gets r successes.
We stick with Montgomery-Runger.

Negative Binomial Mean & Variance

- If X is a negative binomial random variable with parameters p and r,

$$\mu = E(X) = \frac{r}{p} \quad \text{and} \quad \sigma^2 = V(X) = \frac{r(1-p)}{p^2} \qquad (3\text{-}12)$$

- Compare to geometric distribution:

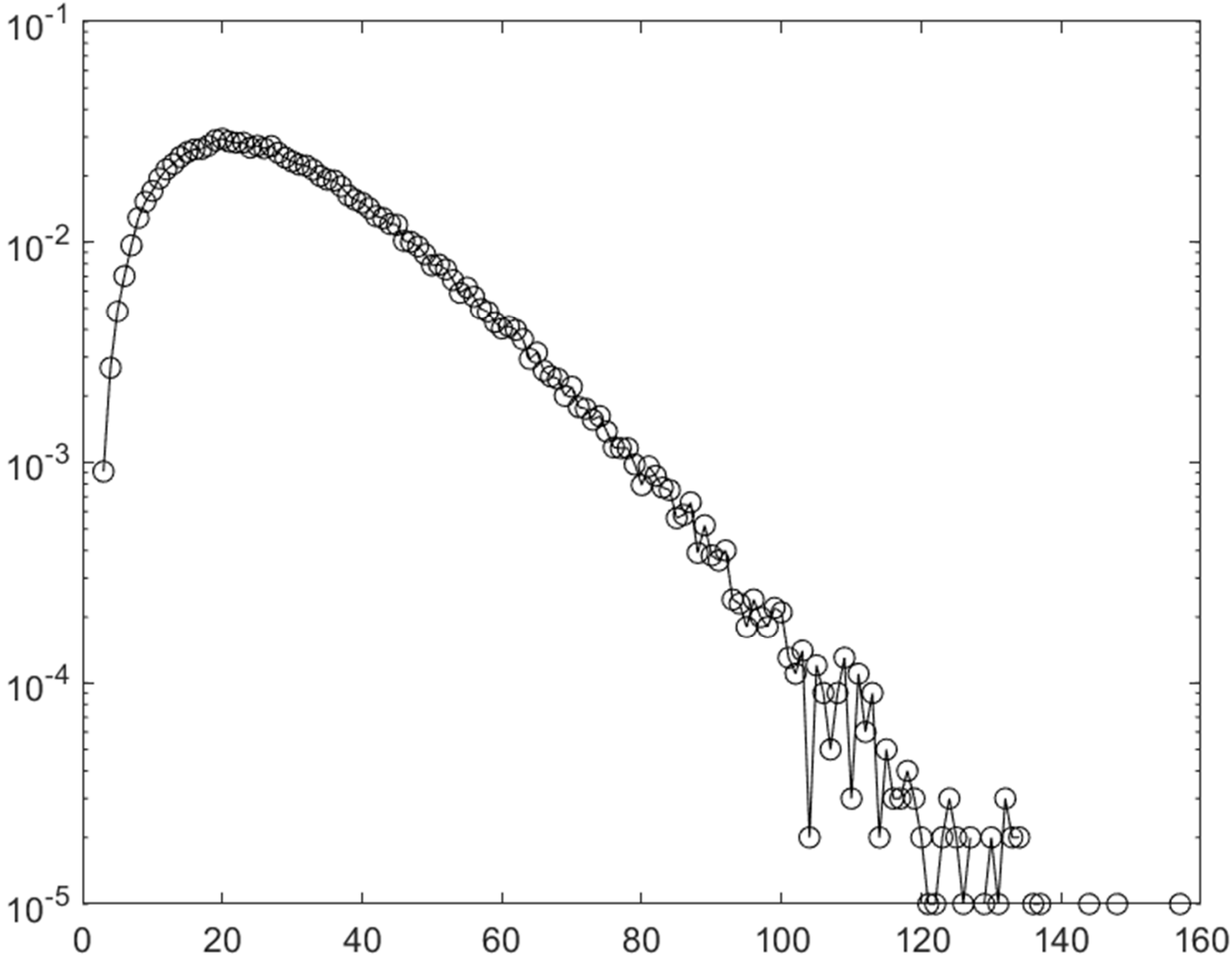
$$\mu = E(X) = \frac{1}{p} \quad \text{and} \quad \sigma^2 = V(X) = \frac{(1-p)}{p^2} \qquad (3\text{-}10)$$

Matlab exercise

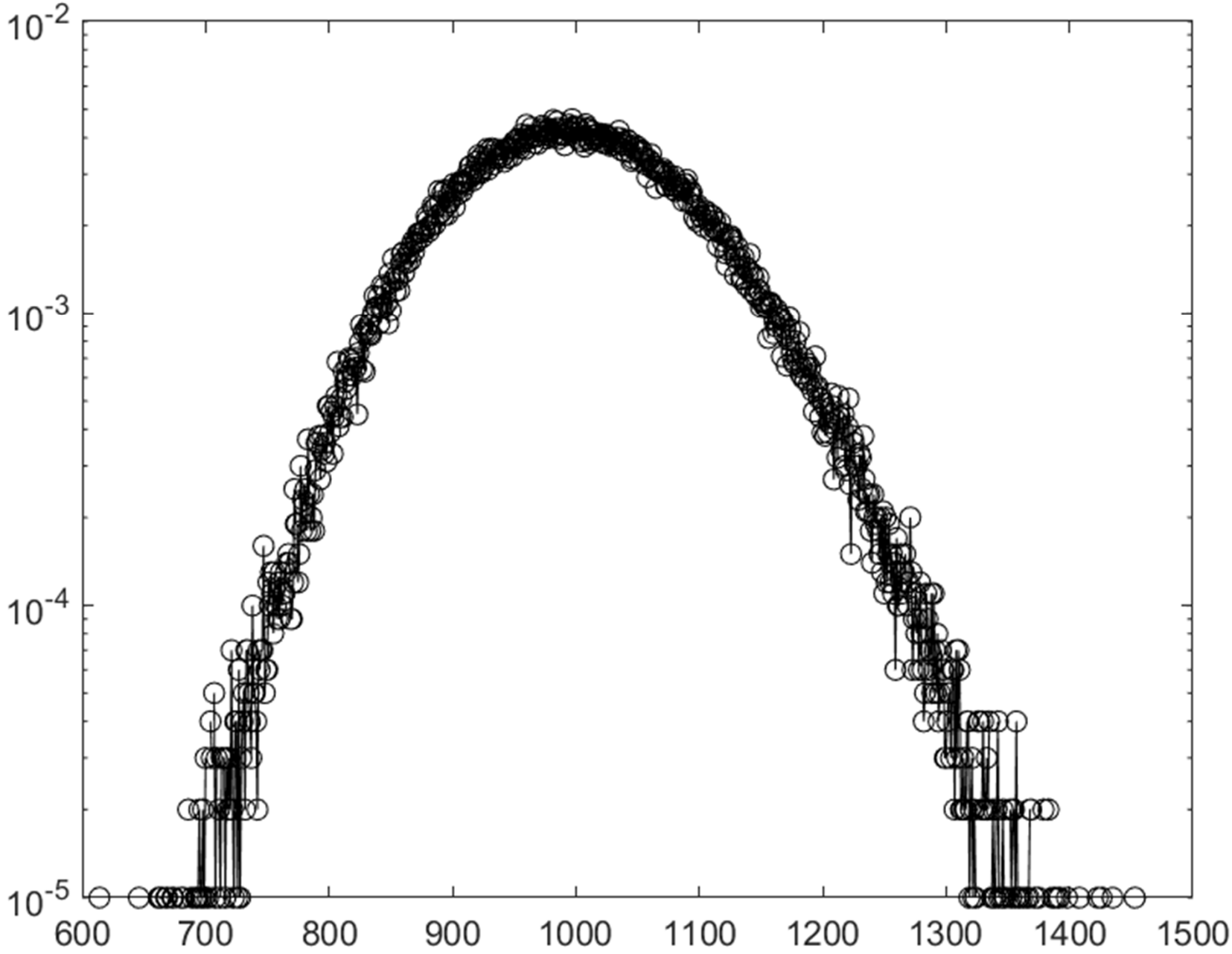
- Estimate mean, variance, and PMF based on 100,000 random variables drawn from a negative binomial distribution with p=0.1, r=3

- Repeat with negative binomial distribution with p=0.1, r=100

Negative binomial PMF, p=0,1 r=3



Negative binomial PMF, p=0,1 r=100



Cancer is scary!

- Approximately 40% of men and women will be diagnosed with cancer at some point during their lifetimes (source: NCI website)

TABLE 21.2 Leading causes of death in United States in 2010. Cause of death is based on the International Classification of Diseases, Tenth Revision, 1992.

Rank	Cause of death	Number	Percent of all deaths
–	All causes	2,468,435	100.0
1	Diseases of heart	597,689	24.2
2	Malignant neoplasms	574,743	23.3
3	Chronic lower respiratory diseases	138,080	5.6
4	Cerebrovascular diseases	129,476	5.2
5	Accidents (unintentional injuries)	120,859	4.9
6	Alzheimer's disease	83,494	3.4
7	Diabetes mellitus	69,071	2.8
8	Nephritis, nephrotic syndrome, and nephrosis	50,476	2.0
9	Influenza and pneumonia	50,097	2.0
10	Intentional self-harm (suicide)	38,364	1.6

Source: National Vital Statistics Reports, 62(6) (http://www.cdc.gov/nchs/data/nvsr/nvsr62/nvsr62_06.pdf)

Table from J. Pevsner 3rd edition

- "War on Cancer" – president Nixon 1971.
  "Moonshot to Cure Cancer" – vice-president Joe Biden 2016

“War on Cancer” progress report

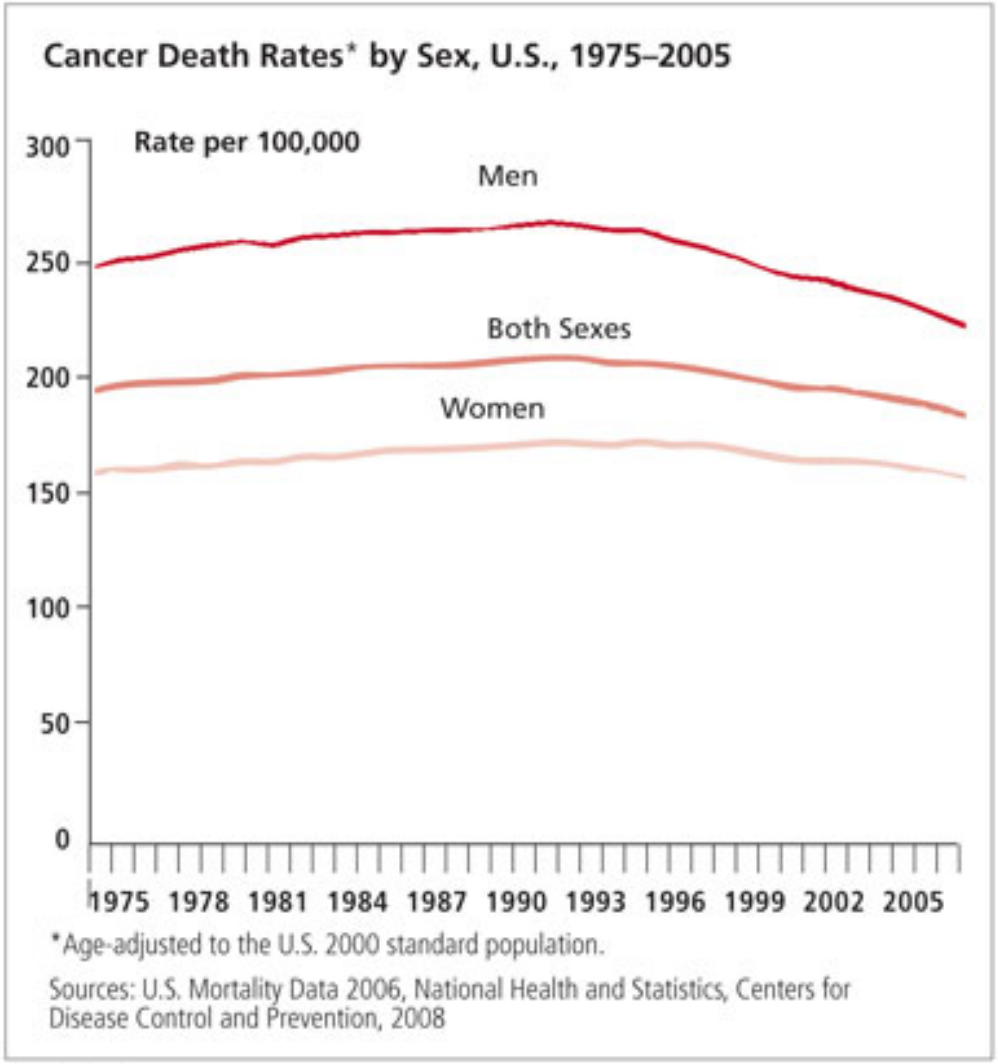


Figure 2

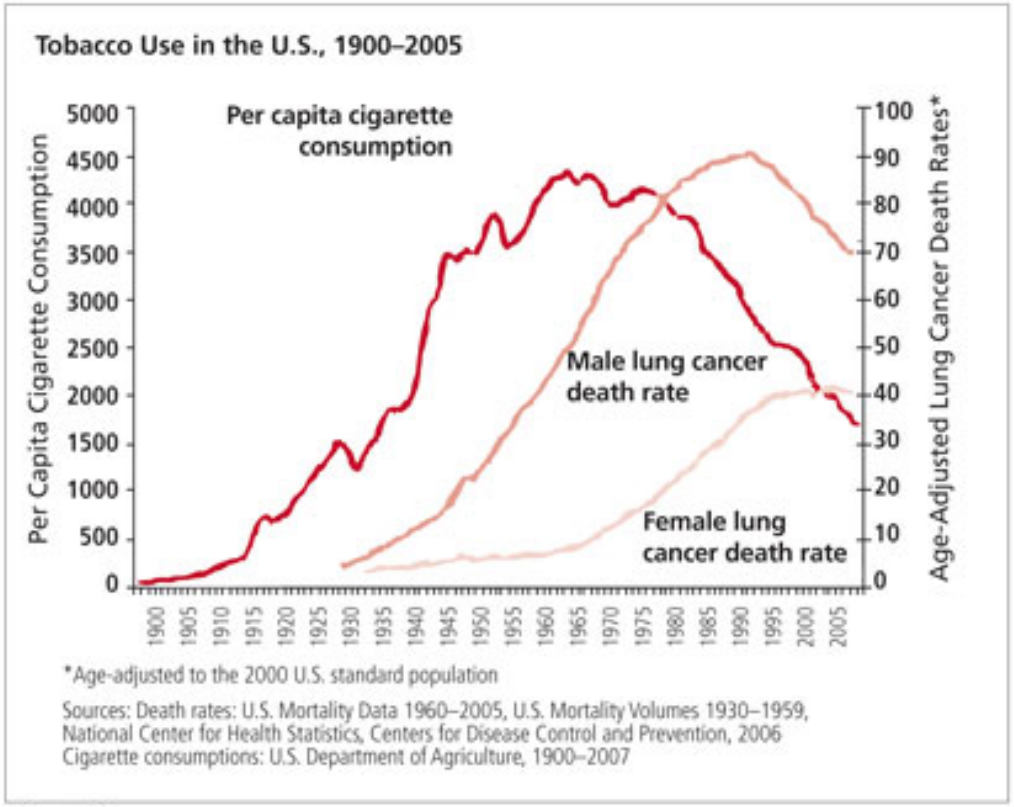


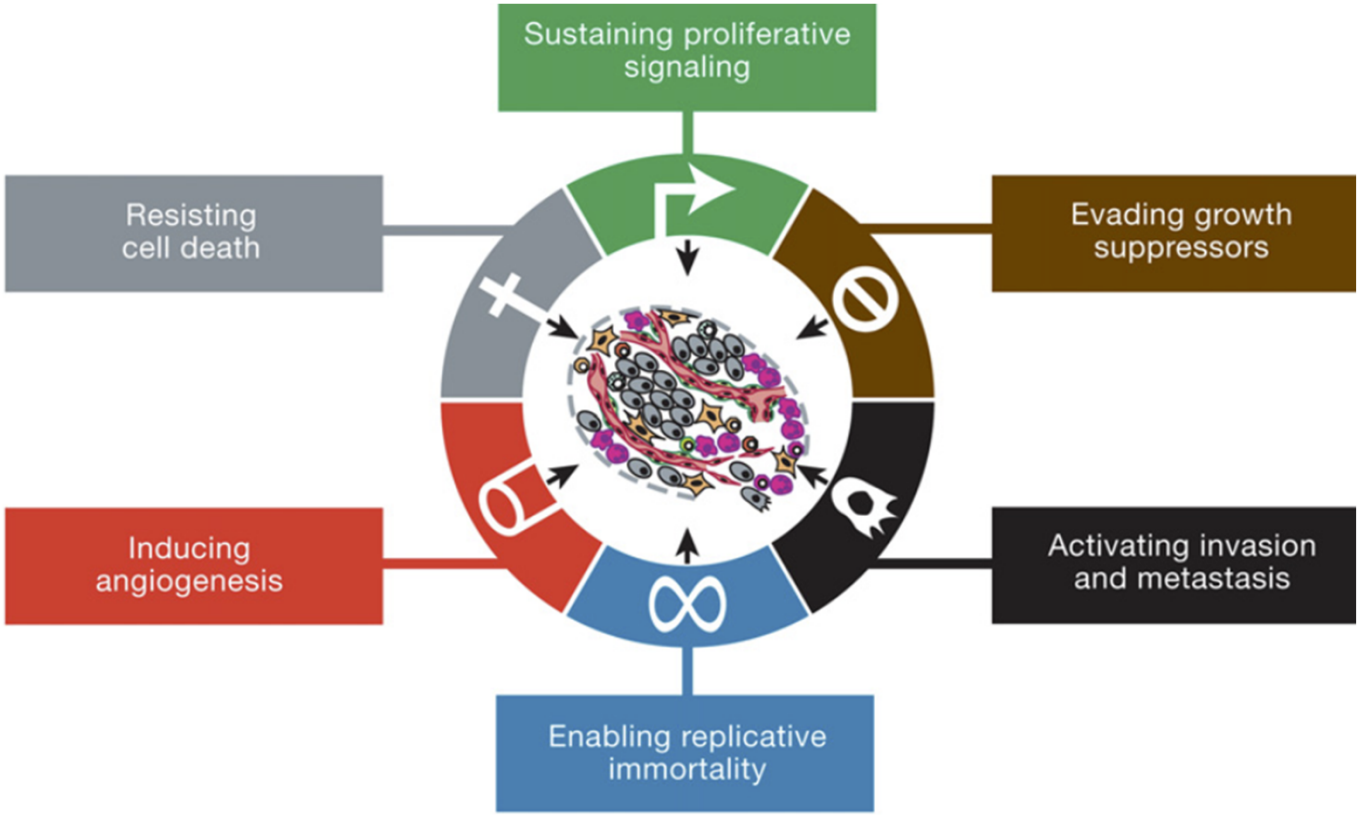
Figure 3

Probability theory and statistics
is a powerful tool to
learn new cancer biology

“Driver genes” theory

- Progression of cancer is caused by accumulation of mutations in a handful of “driver” genes
- Mutations in driver genes boost the growth of a tumor
- Oncogenes: expression needs to be elevated for cancer
- Tumor suppressors (e.g. p53) need to be turned off in cancer

Douglas Hanahan and Robert A. Weinberg
Hallmarks of Cancer: The Next Generation
Cell 144, 2011



Statistics of cancer incidence vs age

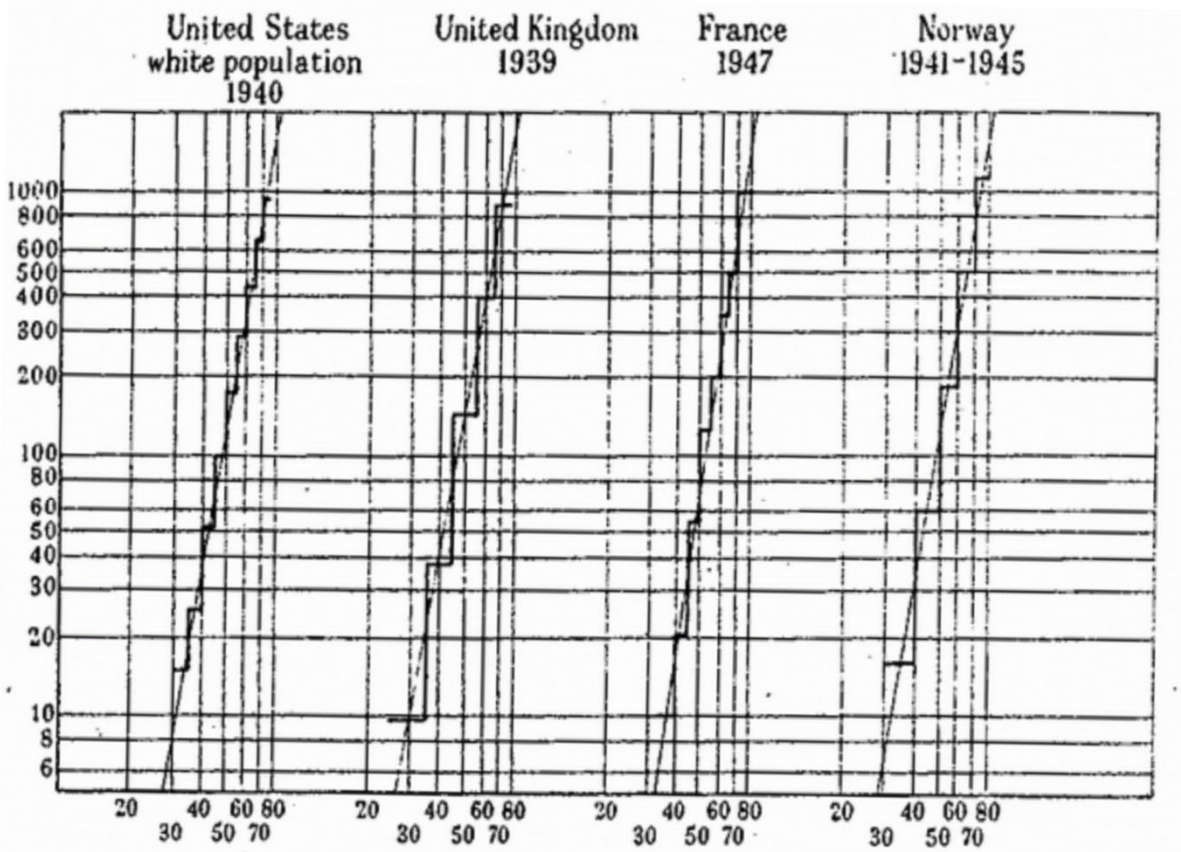


FIG. 1.—Diagram drawn to double logarithmic (log/log) scale showing the cancer death-rate (in the case of the United Kingdom, the carcinoma death-rate) in males at different ages. Deaths per 100,000 males are shown on the vertical scale, age figures on the horizontal scale.

Multi-mutation theory of cancer:
Carl O. Nordling (British J. of Cancer, March 1953):

Cancer death rate
~ (patient age)^6

It suggests the existence of
k=7 driver genes

$$P(T_{cancer} \leq t) \sim (u_1 t)(u_2 t)..(u_k t) \sim u_1 u_2 .. u_k \, t^k$$

$$P(T_{cancer} = t) \sim \frac{d}{dt}(u_1 t)(u_2 t)..(u_k t) \sim k \, u_1 u_2 .. u_k \, t^{k-1}$$

How many driver gene mutations for different types of cancer?

PNAS

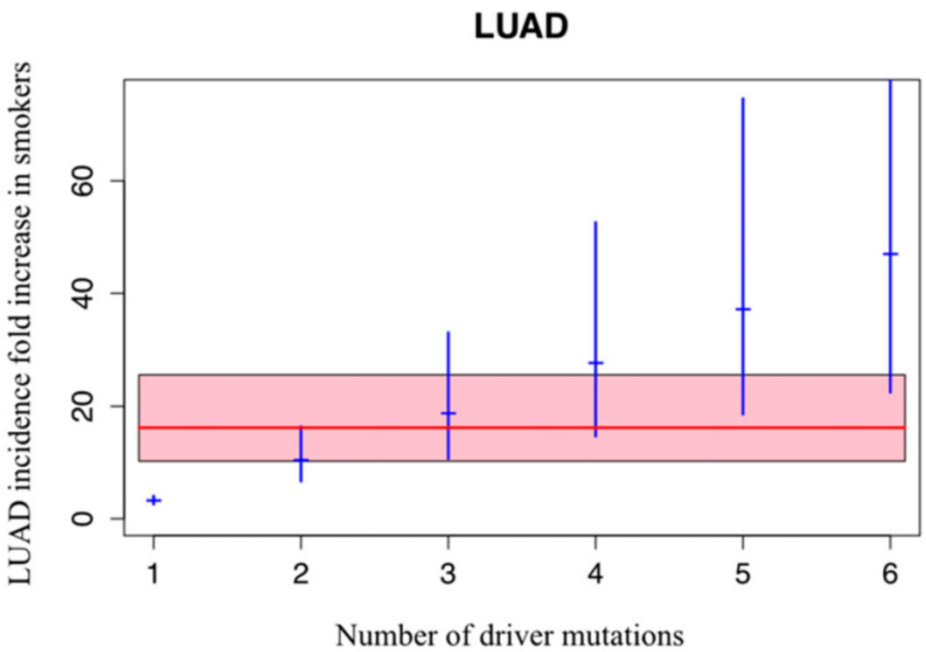
Only three driver gene mutations are required for the development of lung and colorectal cancers

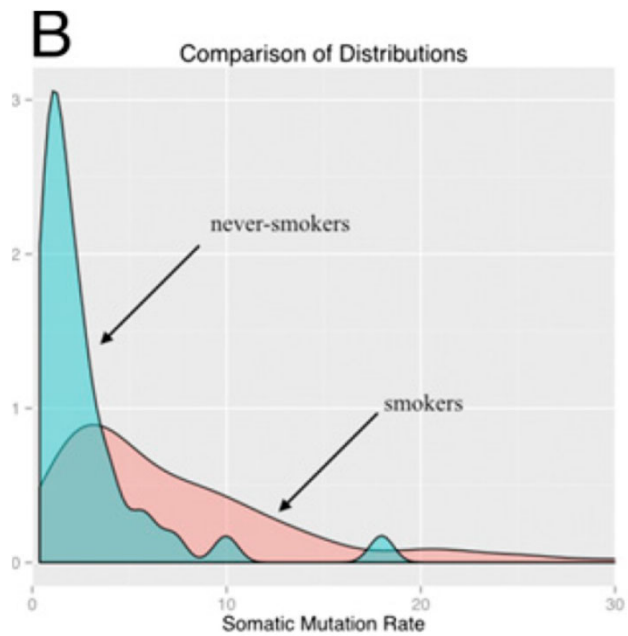
Cristian Tomasetti[a,b,1], Luigi Marchionni[c], Martin A. Nowak[d], Giovanni Parmigiani[e], and Bert Vogelstein[f,g,1]

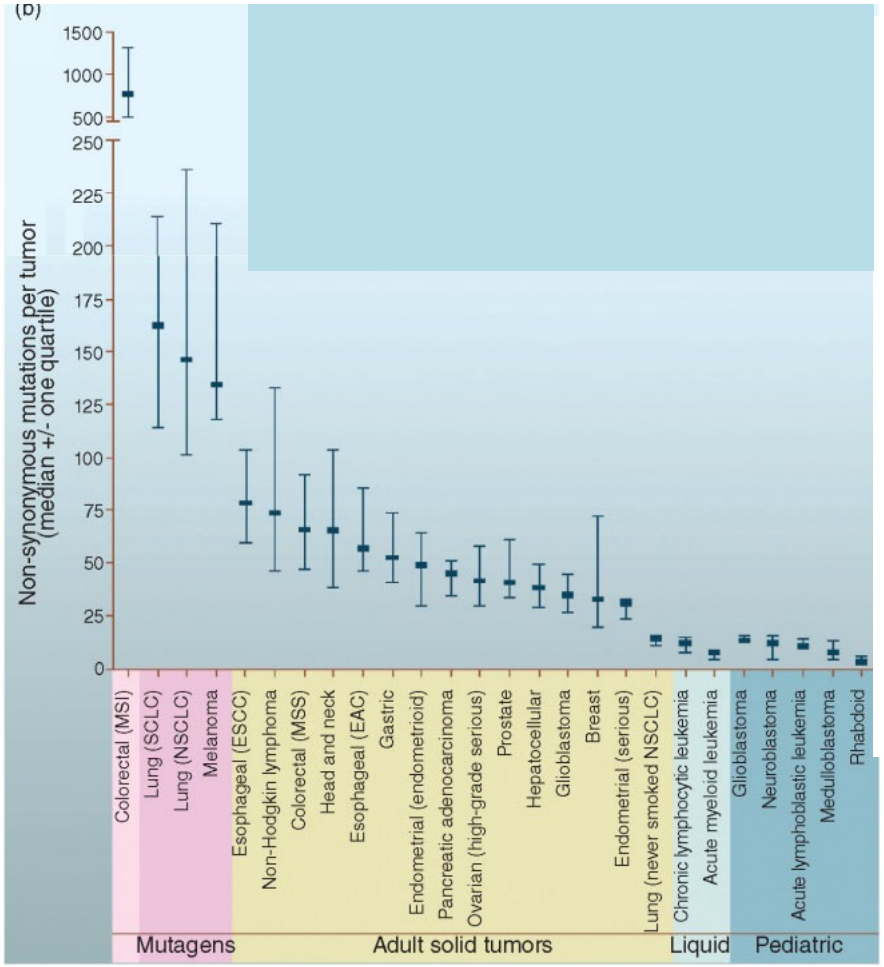
[a]Division of Biostatistics and Bioinformatics, Department of Oncology, Sidney Kimmel Cancer Center, Johns Hopkins University School of Medicine, and [b]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205; [c]Cancer Biology Program, Sidney Kimmel Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD 21205; [d]Program for Evolutionary Dynamics, Department of Mathematics, Harvard University, Cambridge, MA 02138; [e]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, MA 02215; and [f]Ludwig Center for Cancer Genetics and Therapeutics and [g]Howard Hughes Medical Institute, Sidney Kimmel Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD 21205

Contributed by Bert Vogelstein, November 21, 2014 (sent for review July 31, 2014; reviewed by Zvia Agur)

Smokers have 3.23 times more mutations in lungs







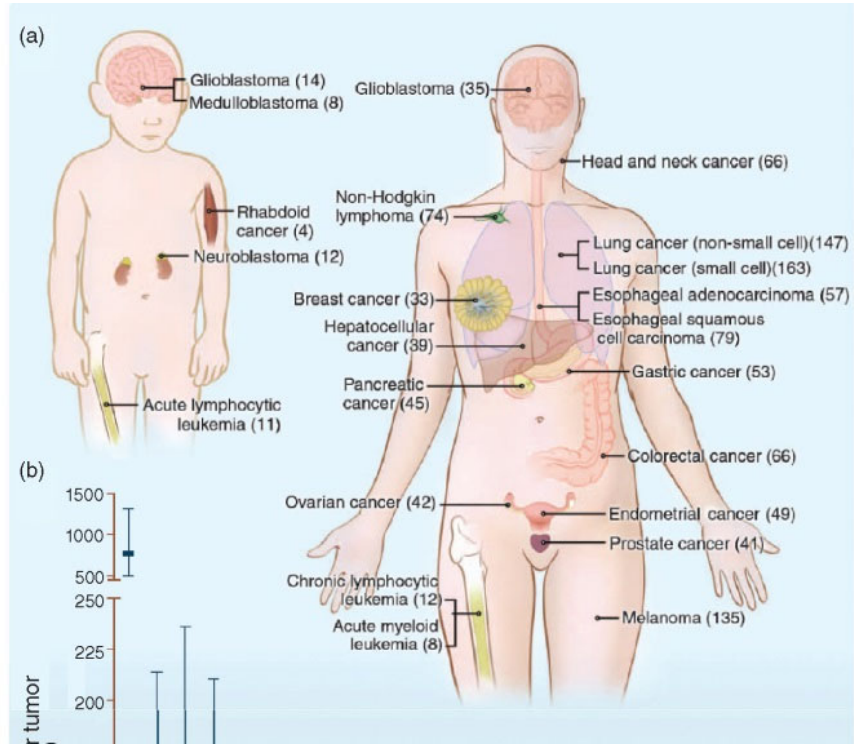


FIGURE 21.10 Somatic mutations in representative human cancers, based on genome-wide sequencing studies. (a) The genomes of adult (right) and pediatric (left) cancers are represented. Numbers in parentheses are the median number of nonsynonymous mutations per tumor. Redrawn from Vogelstein *et al*. (2013). Reproduced with permission from AAAS. (b) Median number of nonsynonymous substitutions per tumor. Horizonal bars indicate the 25% and 75% quartiles. MSI: microsatellite instability; SCLC: small cell lung cancers; NSCLC: non-small cell lung cancers; ESCC: esophageal squamous cell carcinomas; MSS: microsatellite stable; EAC: esophageal adenocarcinomas.

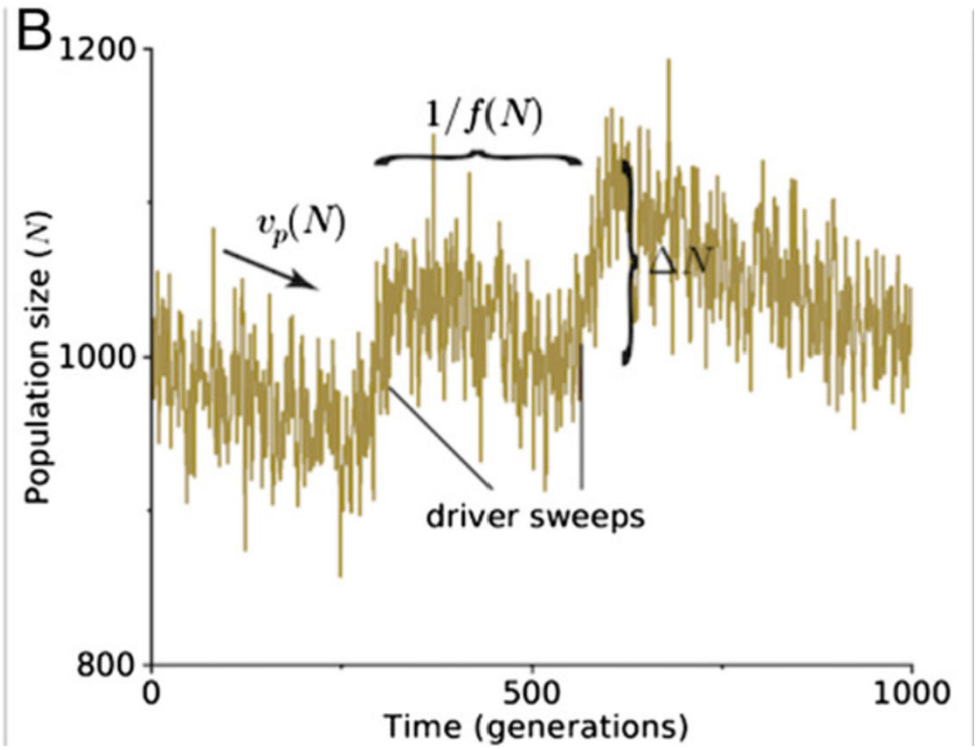
- Cancer cells carry both "Driver" and "Passengers" mutations
- Passenger mutations cause little to no harm (see later for how even little harm matters)
- Both are common as cancers elevate mutation rate

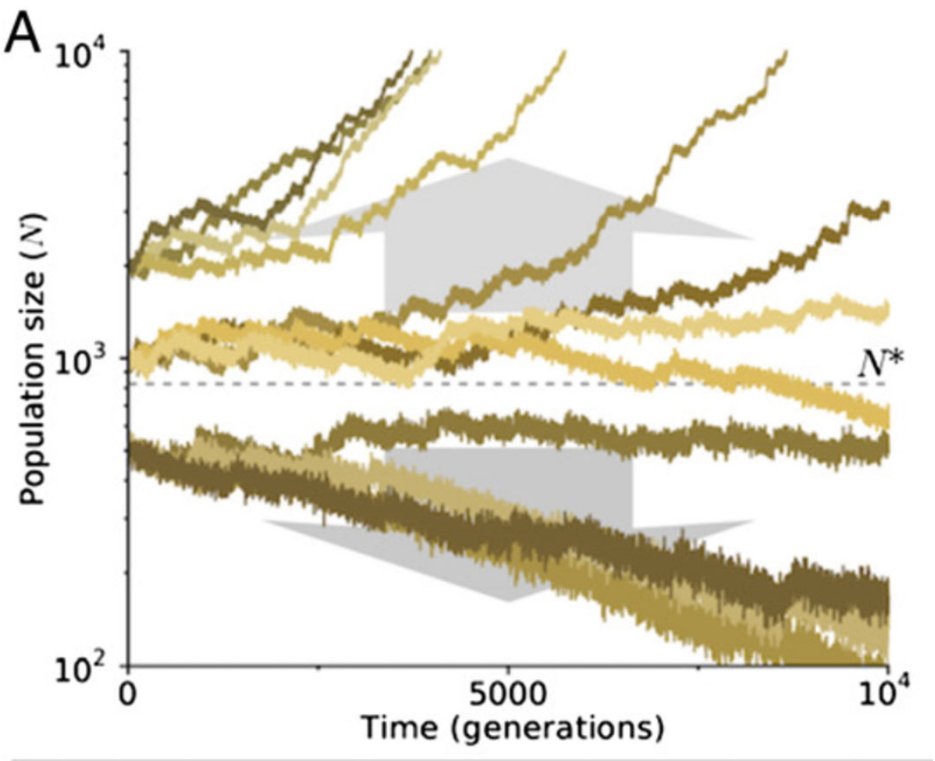
Number of passenger+driver mutations follows negative binomial distribution

- What is the probability to have n_p passenger mutations or (n_p+k) total mutations by the time you are diagnosed with cancer requiring k driver mutations?
- Let p is the probability that a mutation is a driver (p= Genome_target_of_driv/ (Genome_target_of_driv+Genome_target_of_pass) (1-p) – it is a passenger mutation

$$P(n_p + k \mid p, k) = \binom{n_p + k - 1}{n_p} (1-p)^{n_p} p^k$$

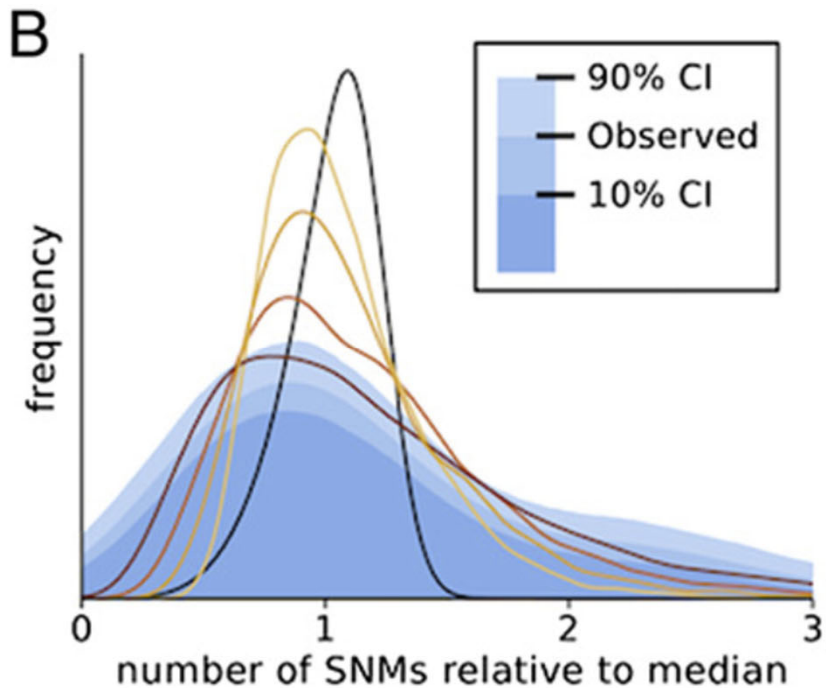
What if passenger mutations slow down the growth of cancer tumors?

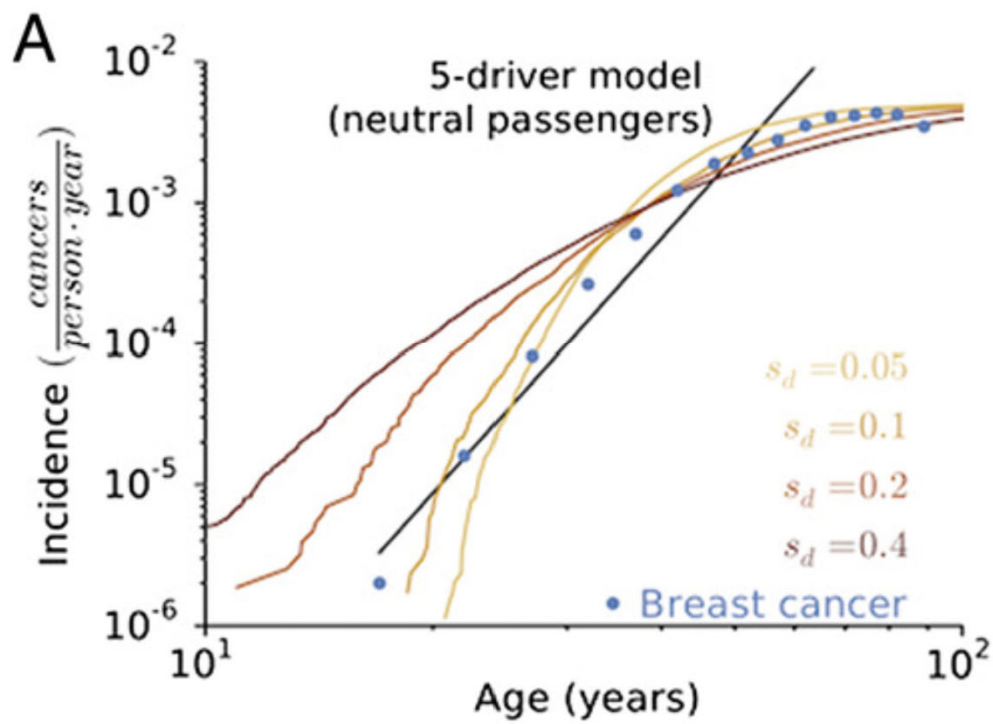




McFarland CD, Mirny L, Korolev KS, PNAS 2014

Can we prove/quantify it using statistics?





Assume: growth rate of cancer=$(1+s_d)^{Nd}/(1+s_p)^{Np}$

$\mu=10^{-8}$, $Target_d$=1,400, $Target_p=10^7$, s_d=0.05 to 0.4, s_p=0.001

s_p/s_d for breast: 0.0060±0.0010;

melanoma: 0.016±0.003; lung: 0.0094±0.0093;

Blue - data on breast cancer: incidence; non-synonymous mutations

Credit: XKCD comics

QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY ARE THERE FEMALE MR MIMES
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING

WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE THERE SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE HELL IF GOD FORGIVES
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE

Important terms & concepts for **discrete random variables**

- Probability Mass Function (PMF)
- ***Cumulative Distribution Function (CDF)***
- ***Complementary Cumulative Distribution Function (CCDF)***
- ***Expected value***
- ***Mean***
- ***Variance***
- ***Standard deviation***

Boldface and underlined are the same for continuous distributions

Name	Probability Distribution	Mean	Variance
Discrete			
Uniform	$\frac{1}{n}, a \leq b$	$\frac{(b+a)}{2}$	$\frac{(b-a+1)^2-1}{12}$
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, 1, \ldots, n, 0 \leq p \leq 1$	np	$np(1-p)$
Geometric	$(1-p)^{x-1}p$, $x = 1, 2, \ldots, 0 \leq p \leq 1$	$1/p$	$(1-p)/p^2$
Negative binomial	$\binom{x-1}{r-1}(1-p)^{x-r}p^r$ $x = r, r+1, r+2, \ldots, 0 \leq p \leq 1$	r/p	$r(1-p)/p^2$
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots, 0 < \lambda$	λ	λ

What distributions we learn

- Uniform distribution
- Bernoulli distribution/trial
- Binomial distribution
- Poisson distribution
- Geometric distribution
- Negative binomial distribution

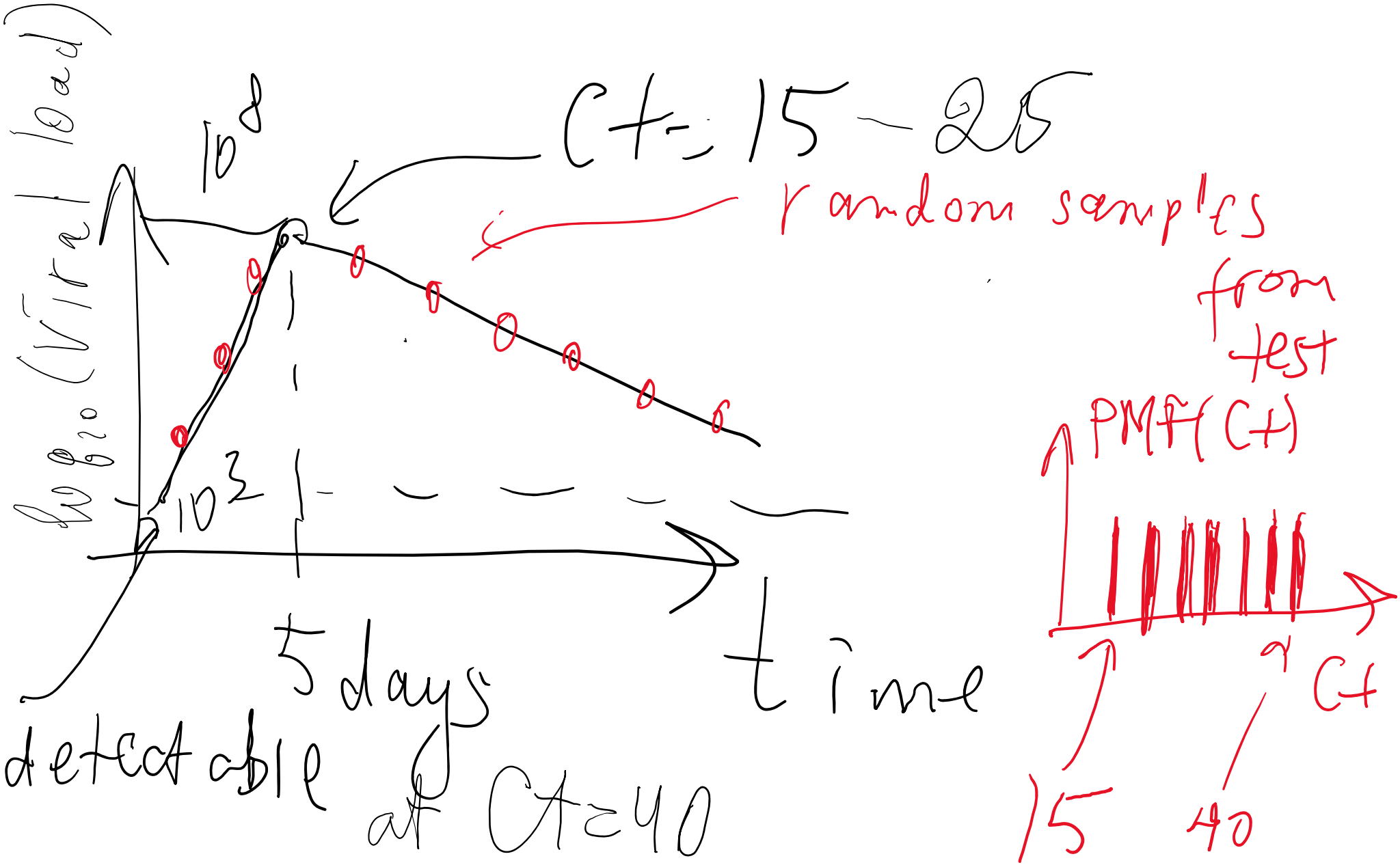
Why do we need to know these simple distributions?

Ways to use statistics

- To process your experimental data
  - What do you need? Mean, Variance, Standard deviation. No need to know any textbook distributions
- To plan experiments
  - Need to know distributions, e.g., Poisson to plan how much redundancy to use for genome assembly
- To learn biological processes behind your data
  - Need to know distributions to compare empirical distributions in your data to what you expect based on a simple hypothesis

Uniform distribution

Why Ct distribution should it be uniform?



Examples of uniform distribution: Ct value of PCR test of a virus

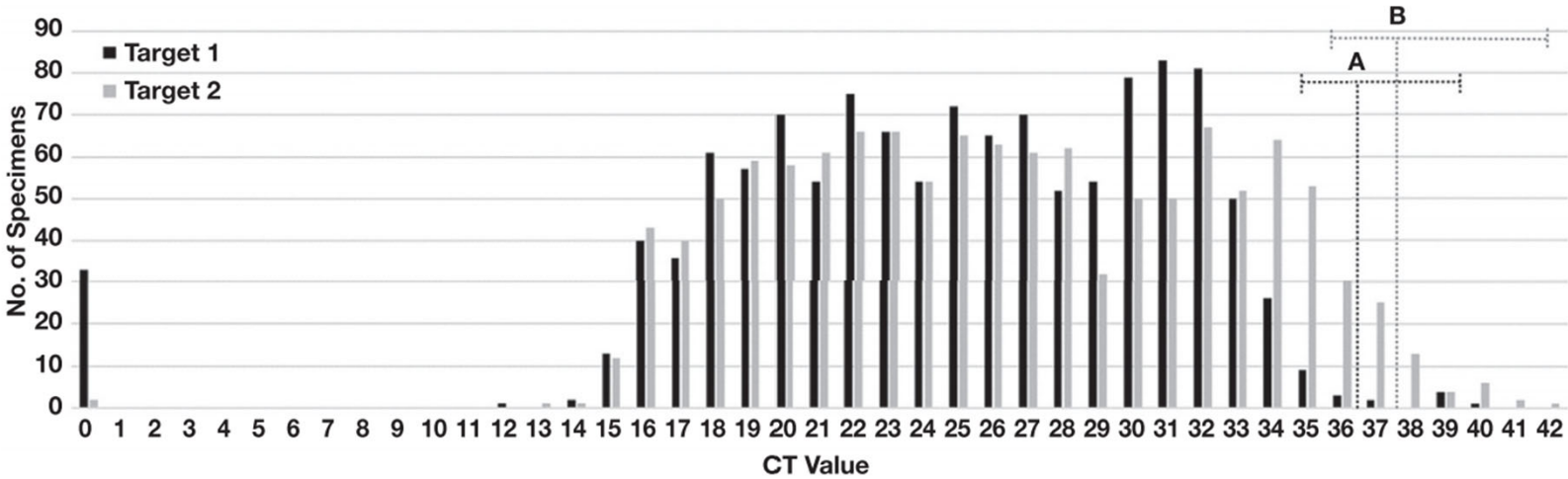


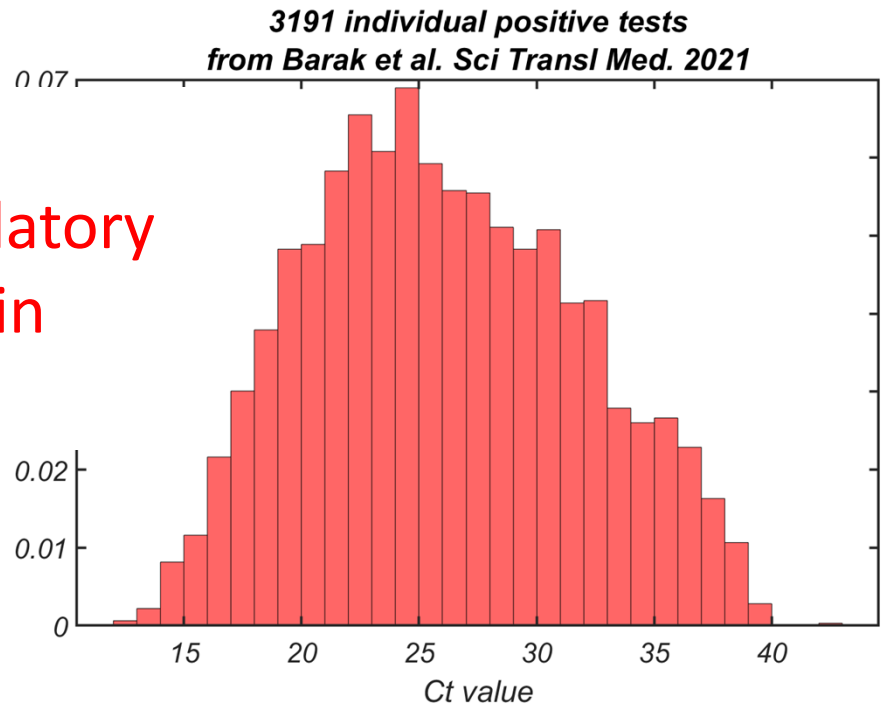
Figure 3 Distribution of cycle threshold (CT) values. The total number of specimens with indicated CT values for Target 1 and 2 are plotted. The estimated limit of detection for (A) Target 1 and (B) Target 2 are indicated by vertical dotted lines. Horizontal dotted lines encompass specimens with CT values less than 3× the LoD for which sensitivity of detection may be less than 100%. This included 19/1,180 (1.6%) reported CT values for Target 1 and 81/1,211 (6.7%) reported CT values for Target 2. Specimens with Target 1 or 2 reported as "not detected" are denoted as a CT value of "0."

Distribution of SARS-CoV-2 PCR Cycle Threshold Values Provide Practical Insight Into Overall and Target-Specific Sensitivity Among Symptomatic Patients

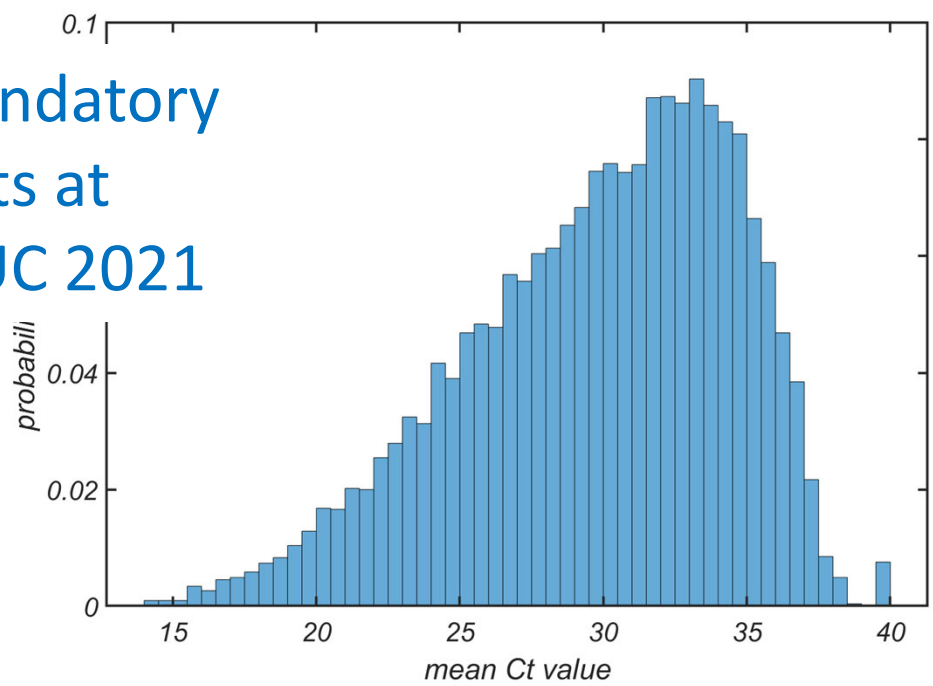
Blake W Buchan, PhD, Jessica S Hoff, PhD, Cameron G Gmehlin, Adriana Perez, Matthew L Faron, PhD, L Silvia Munoz-Price, MD, PhD, Nathan A Ledeboer, PhD *American Journal of Clinical Pathology*, Volume 154, Issue 4, 1 October 2020, https://academic.oup.com/ajcp/article/154/4/479/5873820

Why should we care?

Non-mandatory tests in Israel



Mandatory tests at UIUC 2021



- High Ct value means we identified the infected individual early, hopefully before transmission to others
- When testing is mandatory, and people are tested frequently – Ct value is skewed towards high values

Negative binomial distribution

Statistics of cancer incidence vs age

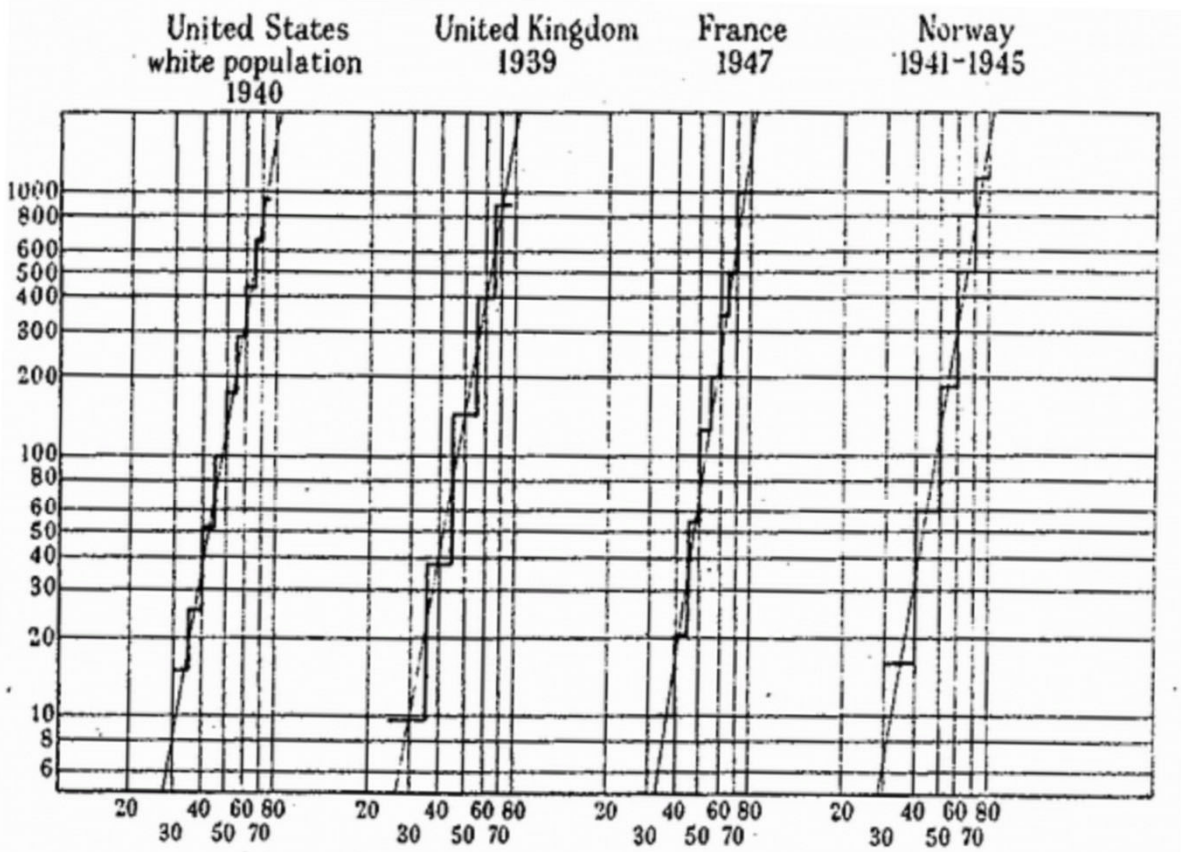


FIG. 1.—Diagram drawn to double logarithmic (log/log) scale showing the cancer death-rate (in the case of the United Kingdom, the carcinoma death-rate) in males at different ages. Deaths per 100,000 males are shown on the vertical scale, age figures on the horizontal scale.

Multi-mutation theory of cancer:
Carl O. Nordling (British J. of Cancer, March 1953):

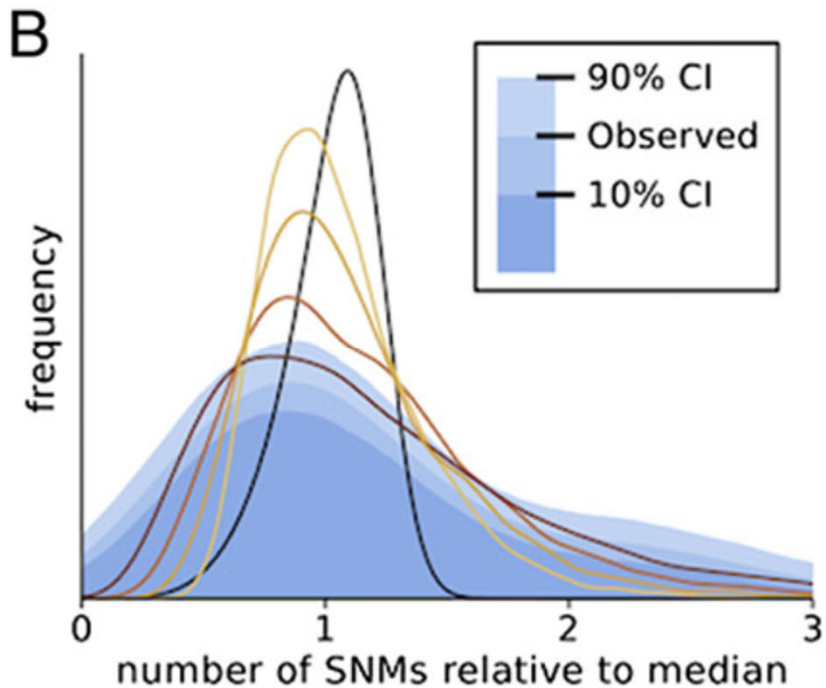
Cancer death rate
~ (patient age)^6

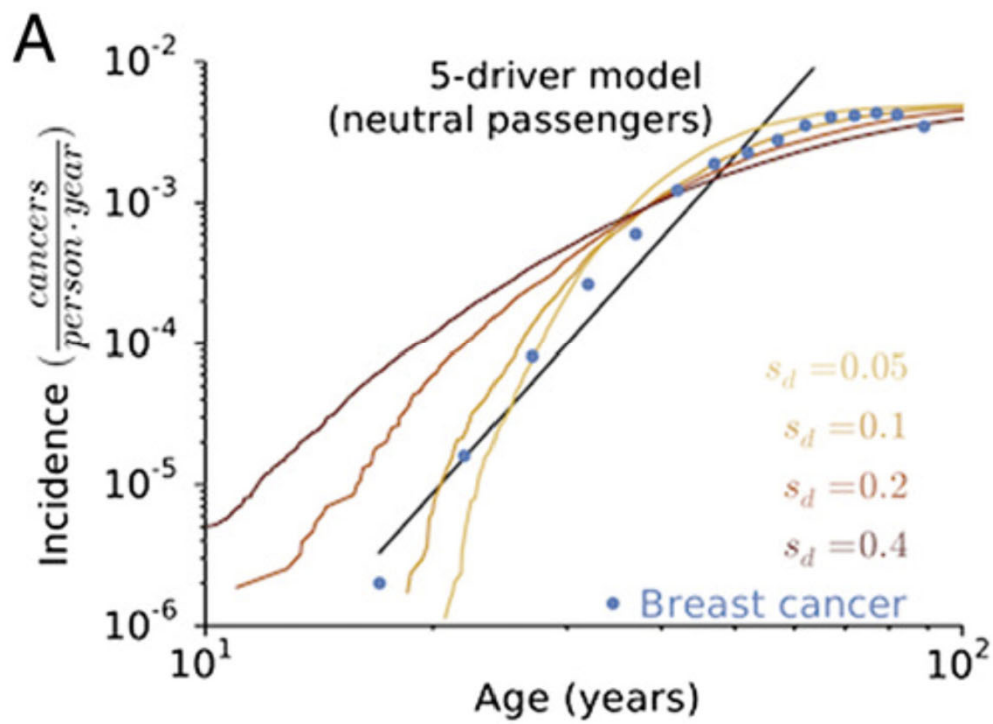
It suggests the existence of
k=7 driver genes

$$P(T_{cancer} \leq t) \sim (u_1 t)(u_2 t)..(u_k t) \sim u_1 u_2 .. u_k \, t^k$$

$$P(T_{cancer} = t) \sim \frac{d}{dt}(u_1 t)(u_2 t)..(u_k t) \sim k \, u_1 u_2 .. u_k \, t^{k-1}$$

Can we prove/quantify it using statistics?





Assume: growth rate of cancer=$(1+s_d)^{Nd}/(1+s_p)^{Np}$

$\mu=10^{-8}$, $Target_d$=1,400, $Target_p=10^7$, s_d=0.05 to 0.4, s_p=0.001

s_p/s_d for breast: 0.0060±0.0010;

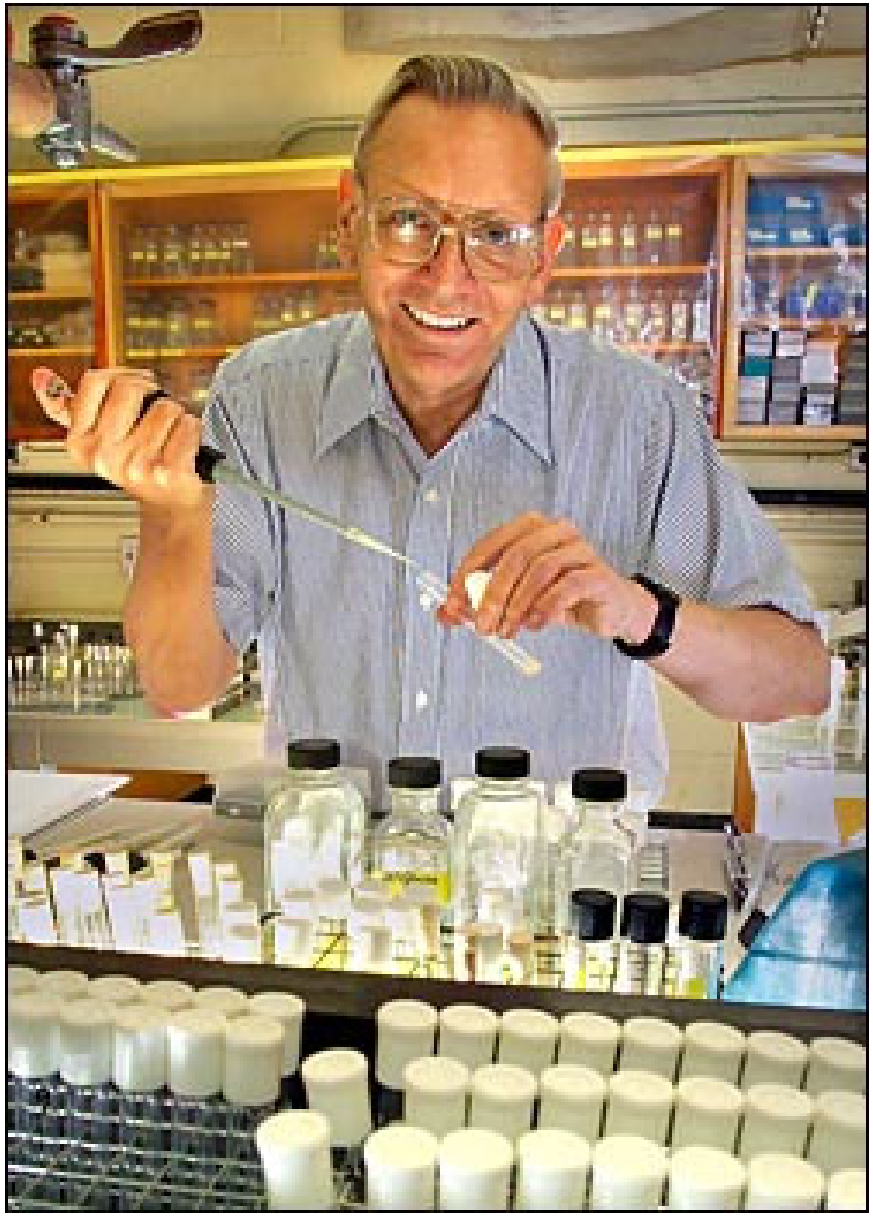
melanoma: 0.016±0.003; lung: 0.0094±0.0093;

Blue - data on breast cancer: incidence; non-synonymous mutations

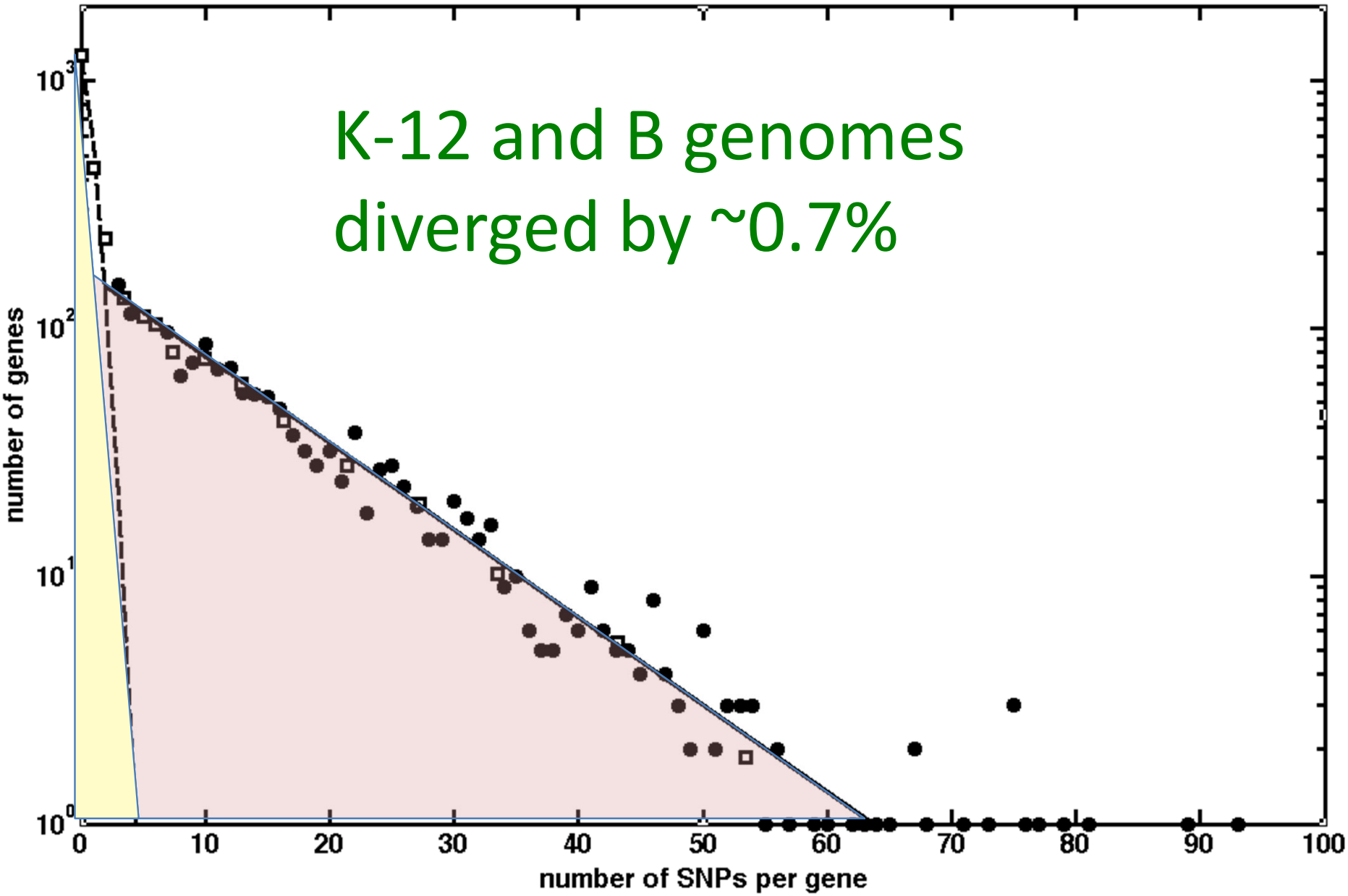
Poisson and Exponential Distributions

F. William Studier

- Worked at Brookhaven National Laboratory, Long Island, NY since 1964
- Inventor of **slab gel electrophoresis in 1970** (not patented- back then no incentive to patent work if you are supported by the US givernment)
- Inventor of T7 phage expression system for **fast production of proteins**. Licensed by over 900 companies, generated over $55 million for the lab https://en.wikipedia.org/wiki/T7_expression_system

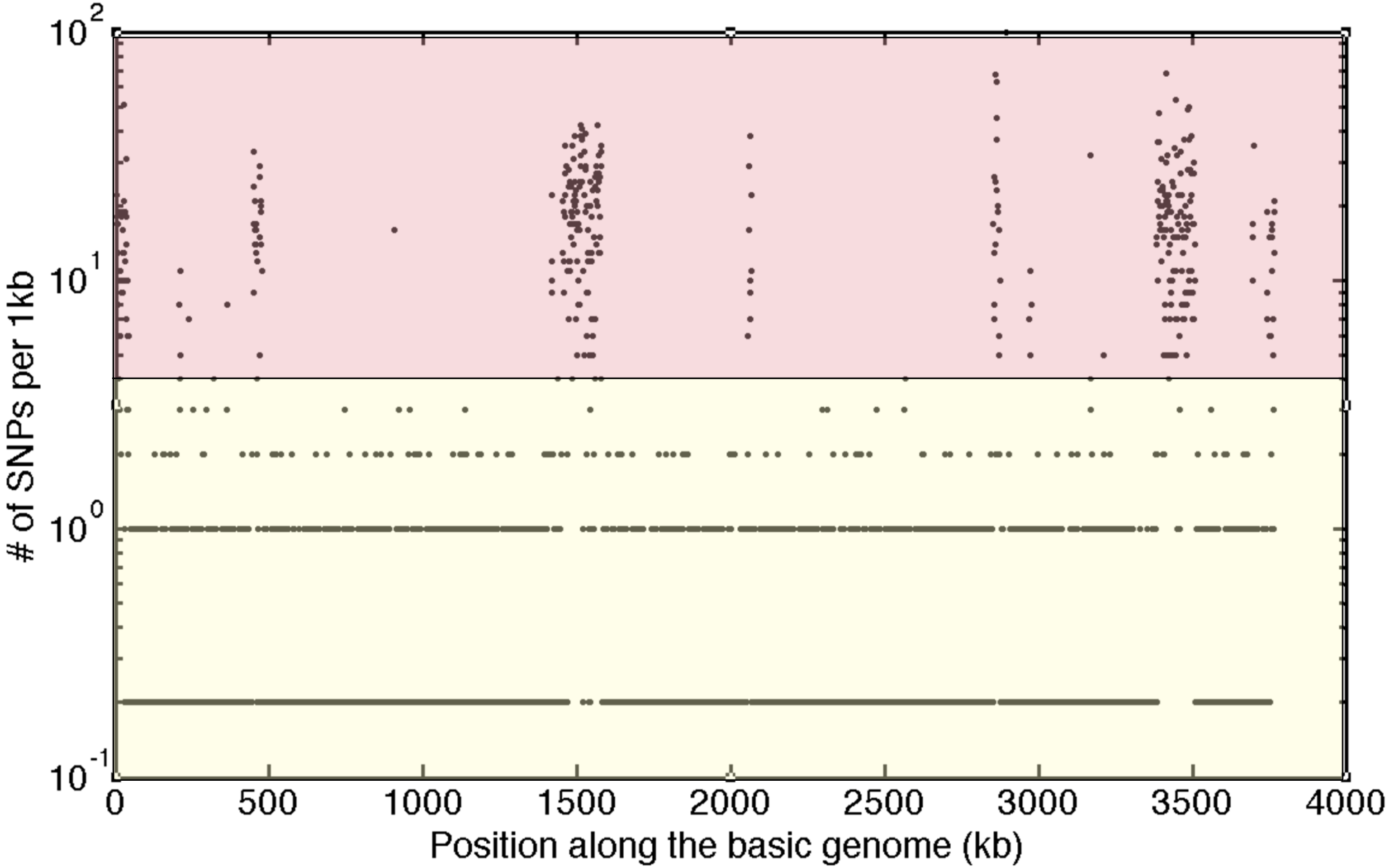


K-12 vs BL21(DE3) strains of E. coli



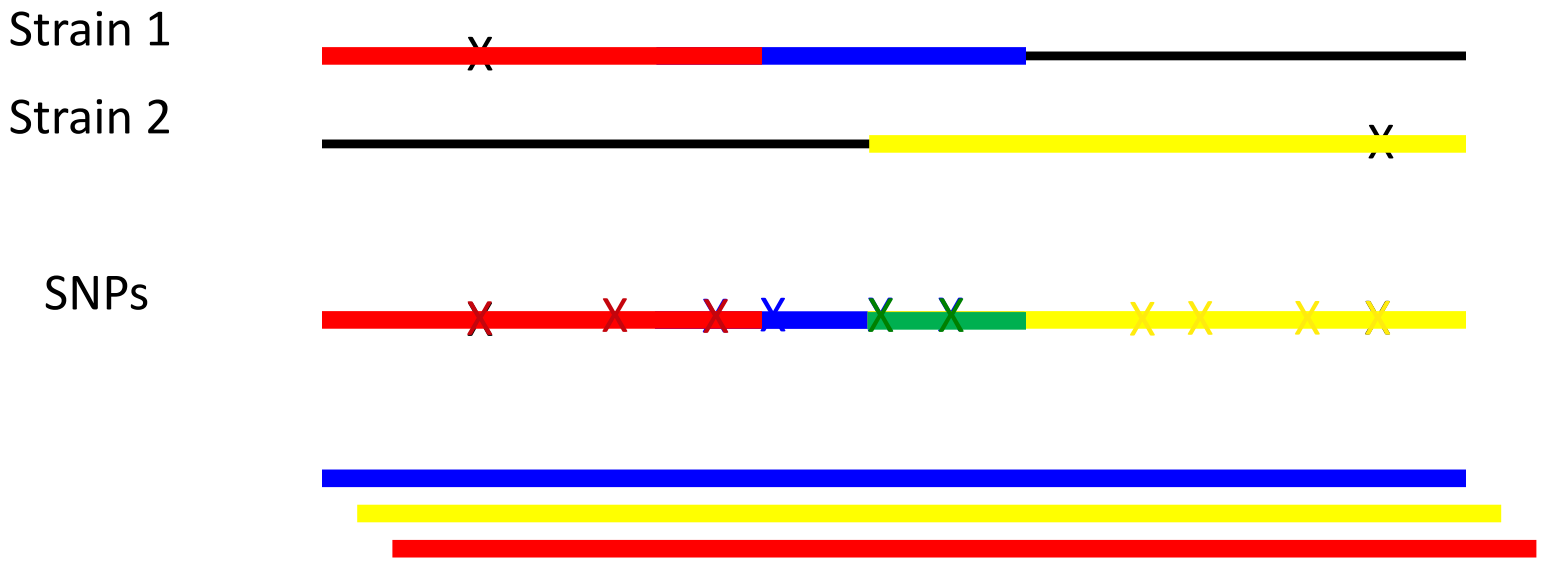
Studier FW, Daegelen P, Lenski RE, **Maslov S**, Kim JF, J. Mol Biol. (2009)

Highly variable segments are clustered



K-12 vs UMNF18 diverged by ~0.18%

Model of bacterial evolution by mutations and homologous recombination



- Mutation rate μ (bp/generation)
- Recombination rate ρ (bp/generation)
- l_R- average length of recombined segments
- $\theta=2\mu N_e$ depending on N_e – (effective) population size
- δ_{TE} transfer efficiency: Prob(successful transfer + recombination): ~ $\exp(-\delta/\delta_{TE})$

Dixit P, Pang TY, Studier FW, **Maslov S,** PNAS (2015); arXiv:1405.2548

Why exponential tail?

- Empirical data for E. coli: Prob(δ)=exp(-δ/0.01)
  Similar slopes in other species as distant as B. subtilis
- Theory 1: PopGen 101 coalescence time distribution:
  - Prob(T) ~ exp(-T/N_e) →
    Prob(δ) ~ exp(- δ/ 2μN_e) = exp(- δ/θ)
    θ =2μN_e~0.01, μ~10^{-10} → N_e~10^8
- Theory 2: biophysics of homologous recombination:
  - Requires perfect matches of L=30bp on each side →
    Prob(δ)=$(1- \delta)^{2L}$=exp(-60•δ)=exp(-δ/0.016)=exp (-δ/δ_{TE})
- Both mechanisms likely to work together:
  biophysics of recombination affects the effective population size

Continuous Probability Distributions

Uniform Distribution

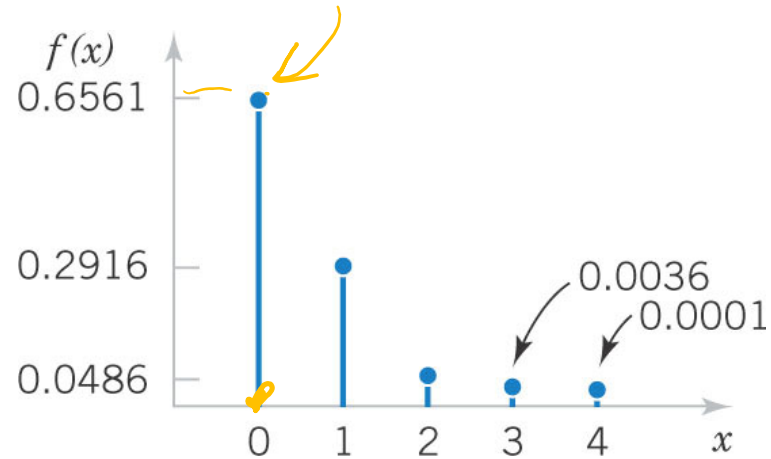
Continuous & Discrete Random Variables

- A discrete random variable is usually integer number
  - N – the number of proteins in a cell
  - D- number of nucleotides different between two sequences
- A continuous random variable is a real number
  - C=N/V – the concentration of proteins in a cell of volume V
  - Percentage D/L*100% of different nucleotides in protein sequences of different lengths L (depending on set of L's may be discrete but dense)

Probability Mass Function (PMF)

- X – discrete random variable

- Probability Mass Function: f(x)=P(X=x) – the probability that X is exactly equal to x



Probability Mass Function for the # of mismatches in 4-mers

P(X =0) =	0.6561
P(X =1) =	0.2916
P(X =2) =	0.0486
P(X =3) =	0.0036
P(X =4) =	0.0001
$\sum_x$ P(X=x)=	1.0000

Probability Density Function (PDF)

Density functions, in contrast to mass functions, distribute probability continuously along an interval

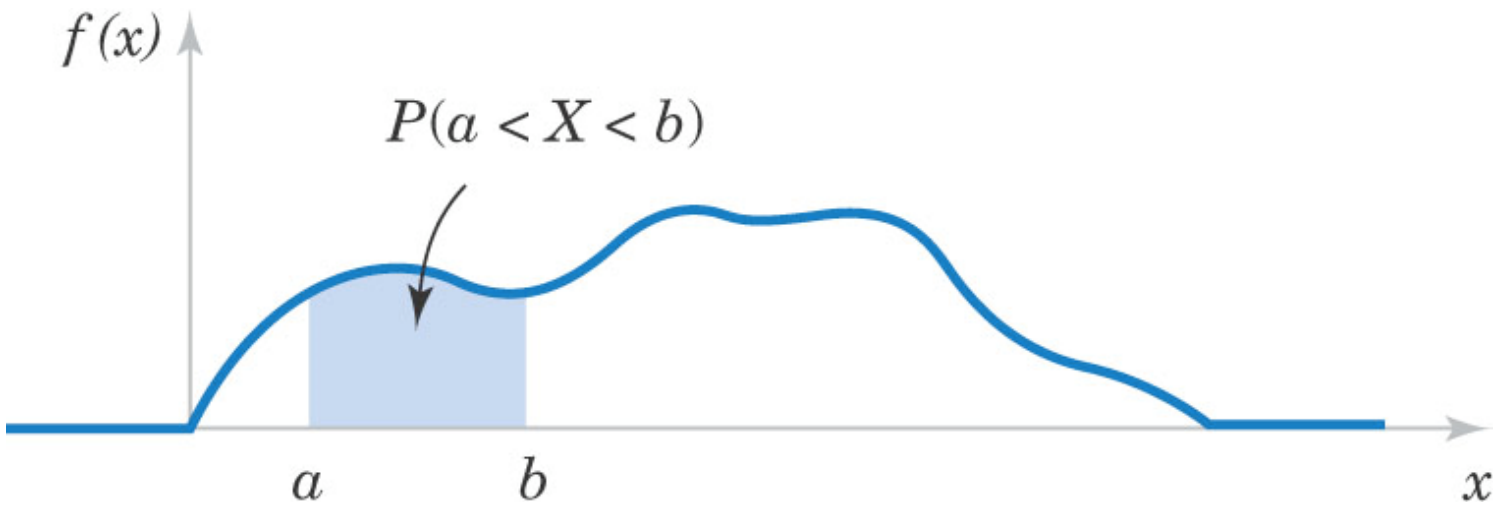


Figure 4-2 Probability is determined from the area under $f(x)$ from a to b.

Probability Density Function

For a continuous random variable X,
a probability density function is a function such that

(1) $f(x) \geq 0$ means that the function is always non−negative.

(2) $$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

(3) $$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx = \text{ area under } f(x)dx \text{ from } a \text{ to } b$$

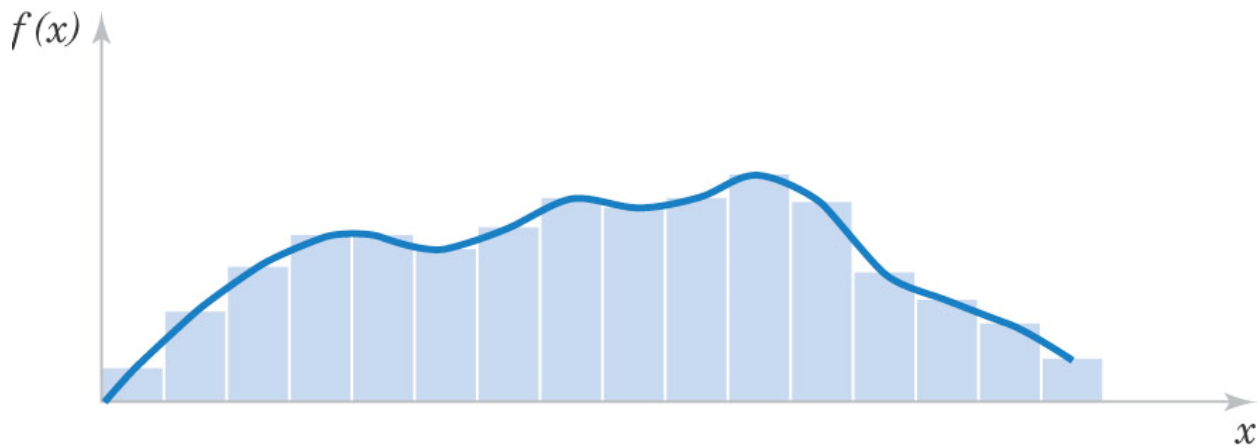
Normalized histogram approximates PDF

A histogram is graphical display of data showing a series of adjacent rectangles. Each rectangle has a base which represents an interval of data values. The height of the rectangle is a number of events in the sample within the base.

When base length is narrow, the histogram could be normalized to approximate PDF (f(x)):

height of each rectangle =

=(# of events within base)/(total # of events)/width of its base.



Normalized histogram approximates a probability density function.

Cumulative Distribution Functions (CDF & CCDF)

The cumulative distribution function (CDF)
of a continuous random variable X is,

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du \text{ for } -\infty < x < \infty \quad (4-3)$$

One can also use the inverse cumulative distribution function
or complementary cumulative distribution function (CCDF)

$$F_{>}(x) = P(X > x) = \int_{x}^{\infty} f(u)du \text{ for } -\infty < x < \infty$$

Definition of CDF for a continous variable is the same
as for a discrete variable

Density vs. Cumulative Functions

- The probability density function (PDF) is the derivative of the cumulative distribution function (CDF).

$$f(x) = \frac{dF(x)}{dx} = -\frac{dF_{>}(x)}{dx}$$

as long as the derivative exists.

Mean & Variance

Suppose X is a continuous random variable with probability density function $f(x)$. The mean or expected value of X, denoted as μ or $E(X)$, is

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx \qquad (4-4)$$

The variance of X, denoted as $V(X)$ or σ^2, is

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2$$

The standard deviation of X is $\sigma = \sqrt{\sigma^2}$.

Gallery of Useful Continuous Probability Distributions

Continuous Uniform Distribution

- This is the simplest continuous distribution and analogous to its discrete counterpart.
- A continuous random variable X with probability density function

$$f(x) = 1 / (b-a) \text{ for } a \le x \le b \qquad (4\text{-}6)$$

Compare to discrete

$$f(x) = 1/(b-a+1)$$

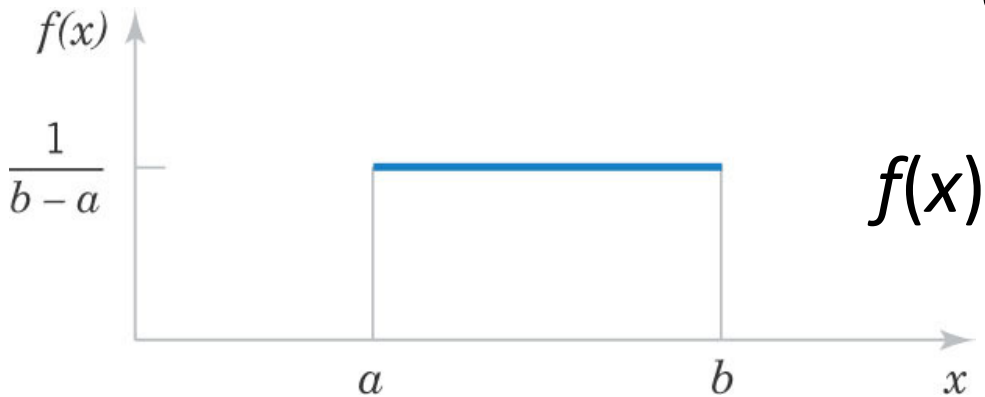


Figure 4-8 Continuous uniform PDF

Comparison between Discrete & Continuous Uniform Distributions

Discrete:

- PMF: $f(x) = 1/(b-a+1)$
- Mean and Variance:

  $\mu = E(x) = (b+a)/2$

  $\sigma^2 = V(x) = [(b-a+1)^2-1]/12$

Continuous:

- PMF: $f(x) = 1/(b-a)$
- Mean and Variance:

  $\mu = E(x) = (b+a)/2$

  $\sigma^2 = V(x) = (b-a)^2/12$

Credit: XKCD comics

QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

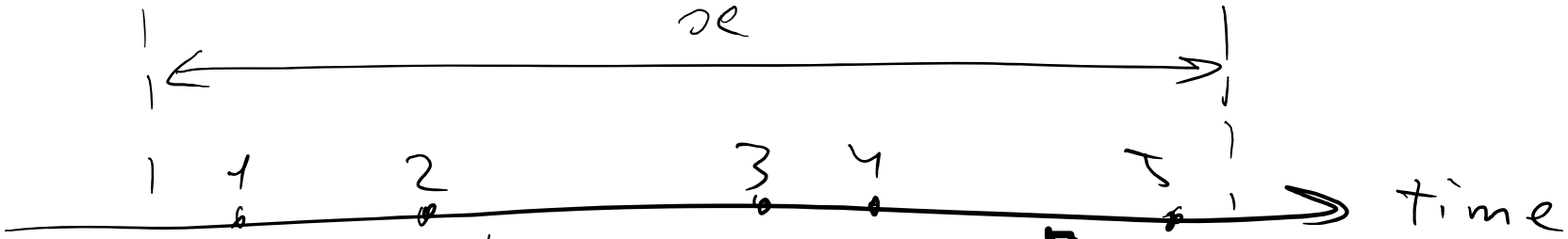
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY ARE THERE FEMALE MR MIMES
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING

WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE THERE SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE HELL IF GOD FORGIVES
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE

Constant rate (Poisson) process

Constant rate (Poisson) process



discrete events happen at rate Γ

Expected number of events in time x is Γx

The actual number of events N_x is a Poisson distributed discrete random variable

$$\boxed{P(N_x = n) = \frac{(\Gamma x)^n}{n!} e^{-\Gamma x}}$$

Why Poisson? Divide x into many tiny intervals of length Δx

$p = \Gamma \Delta x$

$L = x/\Delta x$

$E(N_x) = pL = \Gamma x$

$$\text{Prob}(N = n) = \binom{L}{n} p^n (1-p)^{L-n}$$

$\downarrow$

Poisson

$p \sim \Delta x \to 0, \quad L \sim 1/\Delta x \to \infty$

Constant rate (AKA Poisson) processes

- Let's assume that proteins are produced by ribosomes in the cell at a rate r per second.
- The expected number of proteins produced in x seconds is $r \cdot x$.
- The actual number of proteins N_x is a discrete random variable following a Poisson distribution with mean $r \cdot x$:

$$P_N(N_x=n)=\exp(-r\cdot x)(r\cdot x)^n/n! \qquad E(N_x)= rx$$

- Why Discrete Poisson Distribution?
  - Divide time into many tiny intervals of length $\Delta x \ll 1/r$
  - The probability of success (protein production) per internal is small: $p\_success=r\Delta x \ll 1$,
  - The number of intervals is large: $n= x/\Delta x \gg 1$
  - Mean is constant: $r=E(N_x)=p\_success \cdot n= r\Delta x \cdot x/\Delta x = r\cdot x$
  - In the limit $\Delta x \ll x$, p_success is small and n is large, thus Binomial distribution → Poisson distribution

Exponential Distribution Definition

Exponential random variable X describes interval between two successes of a constant rate (Poisson) random process with success rate r per unit interval.

The probability density function of X is:

$$f(x) = re^{-rx} \quad \text{for} \quad 0 \le x < \infty$$

Closely related to the discrete geometric distribution

$$f(x) = p(1-p)^{x-1} = p\, e^{(x-1)\ln(1-p)} \approx pe^{-px} \quad \text{for small p}$$

To summarize constant rate processes:

r – rate per unit of time length 1D

$N(x)$ – discrete number of events in time x

Poisson: $\boxed{P(N(x)=n) = \frac{(r \cdot x)^n}{n!} e^{-r \cdot x}}$

Time interval X between successive events is a continuously distributed random variable

Its PDF if $\boxed{f(x) = e^{-rx}}$

What is the interval X between two successes of a constant rate process?

- X is a continuous random variable
- CCDF: $P_X(X>x) = P_N(N_X=0)=exp(-r\cdot x)$.
  - Remember: $P_N(N_X=n)=exp(-r\cdot x)\ (r\cdot x)^n/n!$
- PDF: $f_X(x)=-dCCDF_X(x)/dx = r\cdot exp(-r\cdot x)$
- We started with a discrete Poisson distribution where time x was a parameter
- We ended up with a continuous exponential distribution

Exponential Mean & Variance

If the random variable X has an exponential distribution with rate r,

$$\mu = E(X) = \frac{1}{r} \quad \text{and} \quad \sigma^2 = V(X) = \frac{1}{r^2} \qquad (4-15)$$

Note that, for the:

- Poisson distribution: mean= variance
- Exponential distribution: mean = standard deviation = variance$^{0.5}$

Biochemical Reaction Time

- The time x (in minutes) until all enzymes in a cell catalyze a biochemical reaction and generate a product is approximated by this CCDF:

$$F_{>}(x) = e^{-2x} \text{ for } 0 \leq x$$

Here the rate of this process is r=2 min^{-1} and 1/r=0.5 min is the average time between successive products of these enzymes

- What is the PDF?

$$f(x) = -\frac{dF_{>}(x)}{dx} = -\frac{d}{dx}e^{-2x} = 2e^{-2x} \text{ for } 0 \leq \text{x}$$

- What proportion of reactions will not generate another product within 0.5 minutes of the previous product?

$$P(X > 0.5) = F_{>}(0.5) = e^{-2*\,0.5} = 0.37$$

We observed our cell for 1 minute and no product has been generated: The product is "overdue"

What is the probability that a product will not appear during the next 0.5 minutes?

$$F_>(x) = e^{-2x}$$
$$F_>(0.5) \approx 0.37$$
$$F_>(1.5) \approx 0.05$$
$$F_>(1.0) \approx 0.13$$

A. 0.32
B. 0.37
C. 0.08
D. 0.24
E. I have no idea

Get your i-clickers

Memoryless property of the exponential distribution

$$P(X>t+s \mid X>s) = P(X>t)$$

$$P(X>t+s \mid X>s) = \frac{P(X>t+s,\ X>s)}{P(X>s)} =$$

$$= \frac{\exp(-\lambda(t+s))}{\exp(-\lambda s)} = \exp(-\lambda t) =$$

$$= P(X>t)$$

Exponential is <u>the only</u> memoryless distribution

Matlab exercise:

- Generate a sample of 100,000 variables from Exponential distribution with r =0.1
- Calculate mean and compare it to 1/r
- Calculate standard deviation and compare it to 1/r
- Plot semilog-y plots of PDFs **and CCDFs.**
- Hint: read the help page (better yet documentation webpage) for random('Exponential'…) one of their parameters is different than r

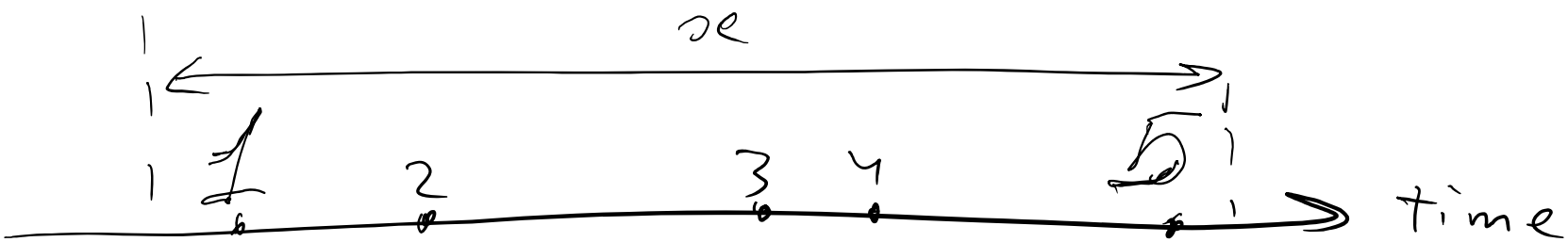
Matlab exercise: Exponential

- **Stats=100000; r=0.1;**
- **r2=random('Exponential', 1./r, Stats,1);**
- **disp([mean(r2),1./r]); disp([std(r2),1./r]);**
- **step=1; [a,b]=hist(r2,0:step:max(r2));**
- **pdf_e=a./sum(a)./step;**
- **subplot(1,2,1); semilogy(b,pdf_e,'rd-’);**
- **x=0:0.01:max(r2);**
- **for m=1:length(x);**
- ** ccdf_e(m)=sum(r2>x(m))./Stats;**
- **end;**
- **subplot(1,2,2); semilogy(x,ccdf_e,'ko-');**

Erlang Distribution

- The Erlang distribution is a generalization of the exponential distribution.
- The exponential distribution models the time interval to the 1st event, while the
- Erlang distribution models the time interval to the kth event, i.e., a sum of k exponentially distributed variables.
- The exponential, as well as Erlang distributions, is based on the constant rate (or Poisson) process.

Constant rate (Poisson) process



Events happen <u>independently</u> from each other at <u>constant rate = r</u> : $E[N_x] = rx$

- X follows Erlang distribution

$$P(X > x) = \sum_{n=0}^{r-1} P(N_x = n) =$$

$$= \sum \frac{(rx)^n}{n!} e^{-rx}$$

Erlang Distribution

Generalizes the Exponential Distribution:

waiting time until k's events

(constant rate process with rate=r)

$$P(X > x) = \sum_{m=0}^{k-1} \frac{e^{-rx}(rx)^m}{m!} = 1 - F(x)$$

Differentiating $F(x)$ we find that all terms in the sum except the last one cancel each other:

$$f(x) = \frac{r^k x^{k-1} e^{-rx}}{(k-1)!} \quad \text{for } x > 0 \text{ and } \mathrm{k} = 1, 2, 3, \ldots$$

Gamma Distribution

The random variable *X* with a probability density function:

$$f(x) = \frac{r^k x^{k-1} e^{-rx}}{\Gamma(k)}, \text{ for } x > 0 \qquad (4\text{-}18)$$

has a gamma random distribution with parameters r > 0 and *k* > 0. If *k* is a positive integer, then *X* has an Erlang distribution.

$$f(x) = \frac{r^k x^{k-1} e^{-rx}}{\Gamma(k)}, \text{ for } x > 0$$

$$\int_0^{+\infty} f(x)\,dx = 1\,, \quad \text{Hence}$$

$$\Gamma(k) = \int_0^{+\infty} r^k x^{k-1} e^{-rx}\,dx = \int_0^{+\infty} y^{k-1} e^{-y}\,dy$$

Comparing with Erlang distribution

for integer k one gets

$$\Gamma(k) = (k-1)!$$

Gamma Function

The gamma function is the generalization of the factorial function for r > 0, not just non-negative integers.

$$\Gamma(k) = \int_0^\infty y^{k-1} e^{-y} dy, \quad \text{for } r > 0 \qquad (4\text{-}17)$$

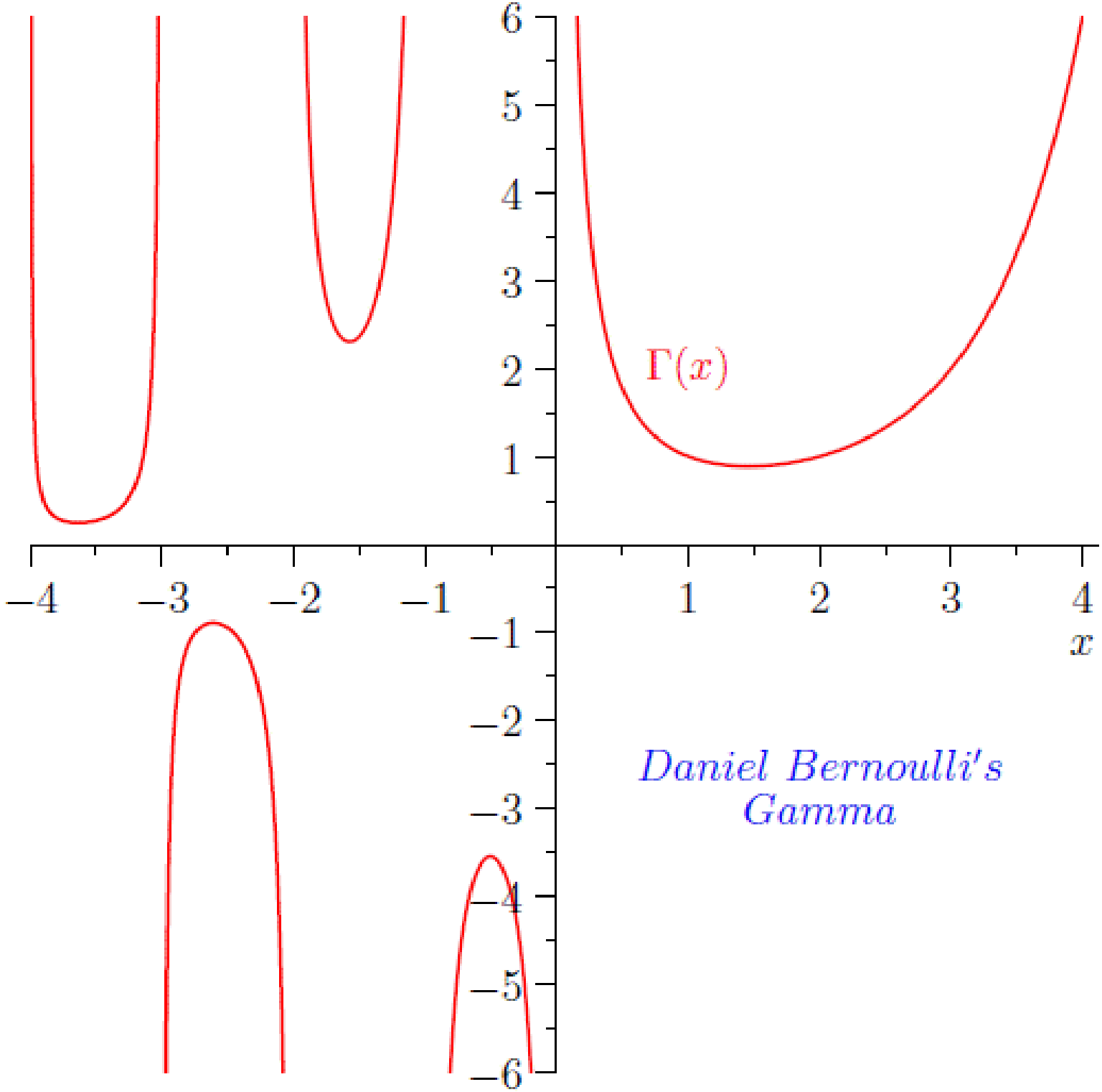
Properties of the gamma function

$$\Gamma(1) = 1$$

$$\Gamma(k) = (k-1)\Gamma(k-1) \quad \text{recursive property}$$

$$\Gamma(k) = (k-1)! \quad \text{factorial function}$$

$$\Gamma(1/2) = \pi^{1/2} = 1.77 \quad \text{interesting fact}$$


6
5
4
3
2
1
−1
−2
−3
−4
−5
−6
−4
−3
−2
−1
1
2
3
4
x
Γ(x)
Daniel Bernoulli's
Gamma

SOLO

BERNOULLI FAMILY

Bernoulli trials

SOLO HERMELIN

http://www.solohermelin.com

The Bernoulli family

Nicolaus
1623-1708

Jacob 1654-1705 | Nicolaus 1662-1716 | **Johann** 1667-1748

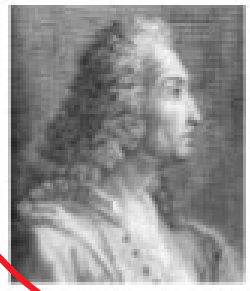
Nicolaus (I)
1687-1759

Nicolaus (II) 1695-1726 | **Daniel** 1700-1782 | **Johann (II)** 1710-1790

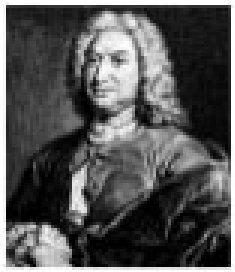
Johann (III) 1744-1807 | Daniel (II) 1751-1834 | **Jacob (II)** 1759-1789

Those shown in **bold** above are in our archive

Run This

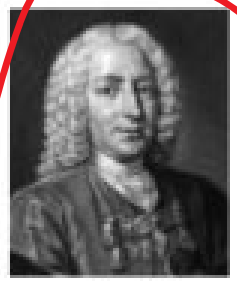


Jacob
1654-1705



Johann
1667-1748

Nicolaus II
1695-1720



Daniel
1700-1782

Johann II
1710-1790

Johann III
1744-1807

Jacob II
1759-1789

Gamma function

Mean & Variance of the Erlang and Gamma

- If X is an Erlang (or more generally Gamma) random variable with parameters r and k,

  $\mu = E(X) = k/r$ and $\sigma^2 = V(X) = k/r^2$ (4-19)

- Generalization of exponential results:

  $\mu = E(X) = 1/r$ and $\sigma^2 = V(X) = 1/r^2$ or

  Negative binomial results:

  $\mu = E(X) = k/p$ and $\sigma^2 = V(X) = k(1-p) / p^2$

Matlab exercise:

- Generate a sample of 100,000 variables with "Harry Potter" Gamma distribution with r =0.1 and k=9 ¾ (9.75)
- Calculate mean and compare it to k/r (Gamma)
- Calculate standard deviation and compare it to sqrt(k)/r (Gamma)
- Plot semilog-y plots of PDFs **and CCDFs.**
- Hint: read the help page (better yet documentation webpage) for random('Gamma'…): one of their parameters is different than r

Credit: XKCD comics

QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE

WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY IS YKK ON ALL ZIPPERS

WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY ARE THERE FEMALE MR MIMES
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING

WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE THERE SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE HELL IF GOD FORGIVES
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE