

BIOE 310:

Computational Tools for Biological Data

What this class is all about?

Instructor

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

Teaching Assistants:

Novia Gonsalves
ngons2@illinois.edu

Dalton Jayne
djayne2@illinois.edu

Questions and Suggestions:

maslov@illinois.edu

ngons2@illinois.edu

djayne2@illinois.edu

Start subject with [BIOE310]

Homework and Exams

- **Homework assignments (4-5)**

Due at the beginning of the class on the designated day

- **2 Midterms**

- **Final exam**

- **Grading:**

Attendance 10%

Homework 20%

Midterm 1 20%

Midterm 2 20%

Final 30%

How attendance score is calculated

- Attendance is worth 10% of your final grade
- Each student receives 5 no-questions-asked excused absences (out of ~26 classes)
- After that, each additional absence reduces your attendance score by 0.5% per missed class, up to a minimum of zero
- Because this policy already provides flexibility, I do not review or grant individual absence exceptions
- Attendance on iClicker website may not reflect true attendance: responding to a single poll means you attended

Will you use iClicker remote

- A. Yes
- B. No, I have an app
- C. I am not using iClicker



Get your i-clickers

Course Website

<https://courses.engr.illinois.edu/bioe310>

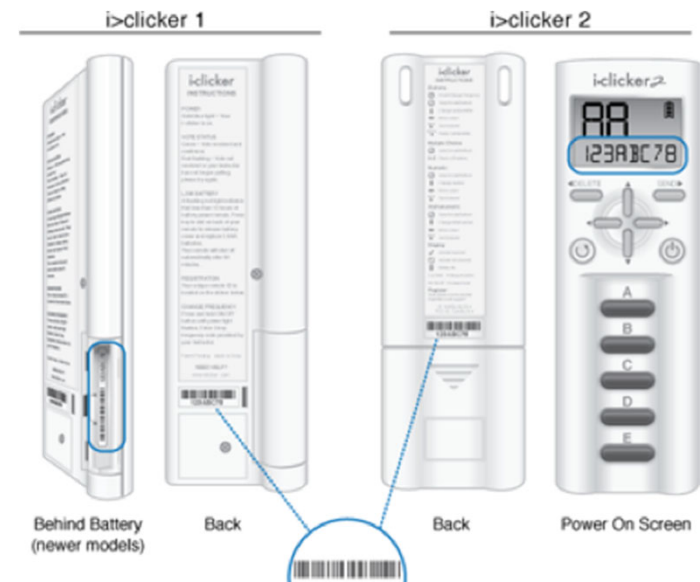
BIOE 310 - Computational Tools for Biological Data

[Return to syllabus](#)

#	Date	Topics	Slides	Matlab	Homework	Exams
1	Jan 26					
2	Jan 28					

Bring your iClickers to my lectures

- **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



We will use Matlab in class

- Bring **your laptops to class**
- **Poll: who has Matlab?**
- 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**
- You can use CITRIX for UIUC students and connect to EWS Windows Lab Software
- **.m files and .mat** with Matlab commands and data **will be on the website** after the lecture

Who has Matlab?

- A. Have on my own laptop
- B. Plan to use CITRIX
- C. I don't have Matlab
- D. I don't know yet
- E. I will never use Matlab

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>

Accessing Engineering Workstations (EWS)

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

Accessing ACES Academic Computing Workstations

<http://acf.aces.illinois.edu/remote/>

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 genome assembly, phylogenetic trees
 - Systems biology (gene expression, networks)

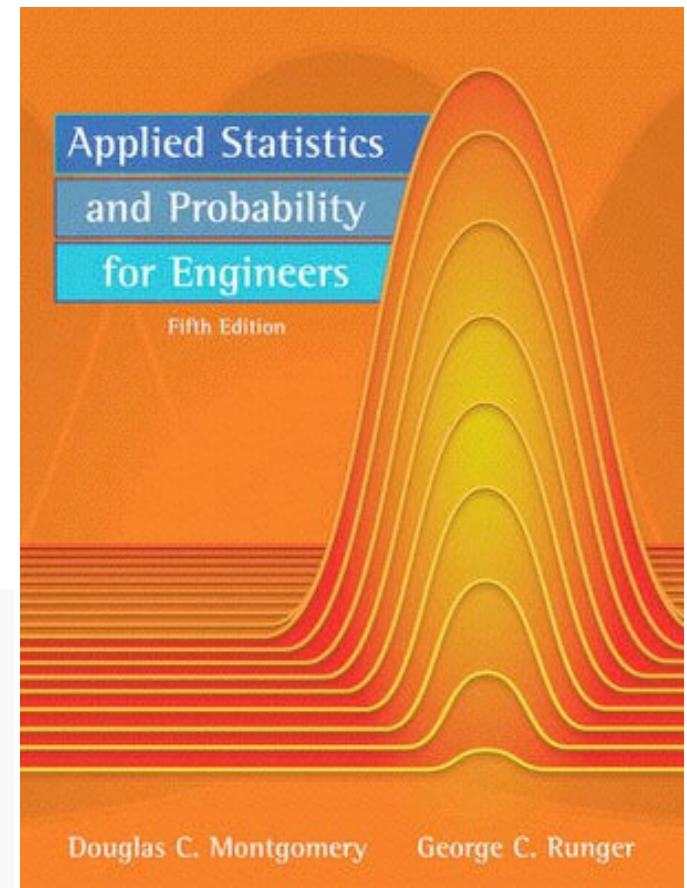
The main Probability/Statistics Textbook

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



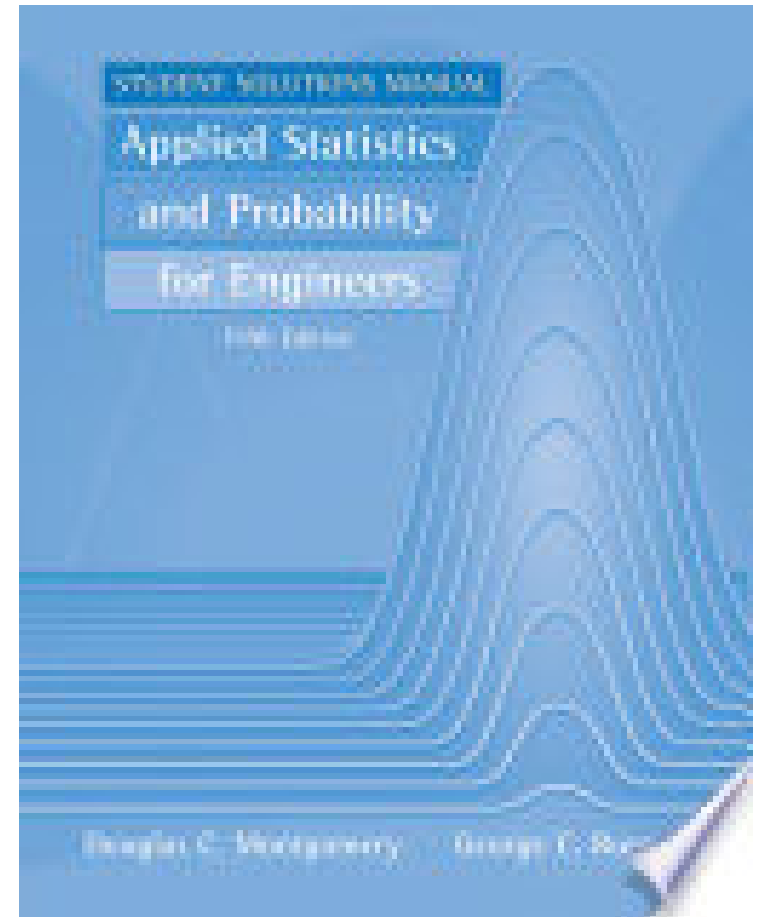
Problems for our main Probability/Statistics Textbook

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

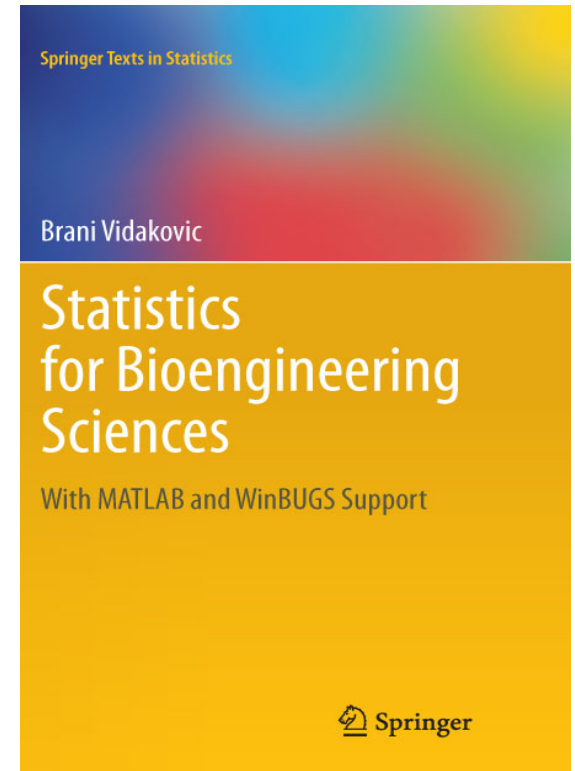
with MATLAB and WinBUGS Support

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

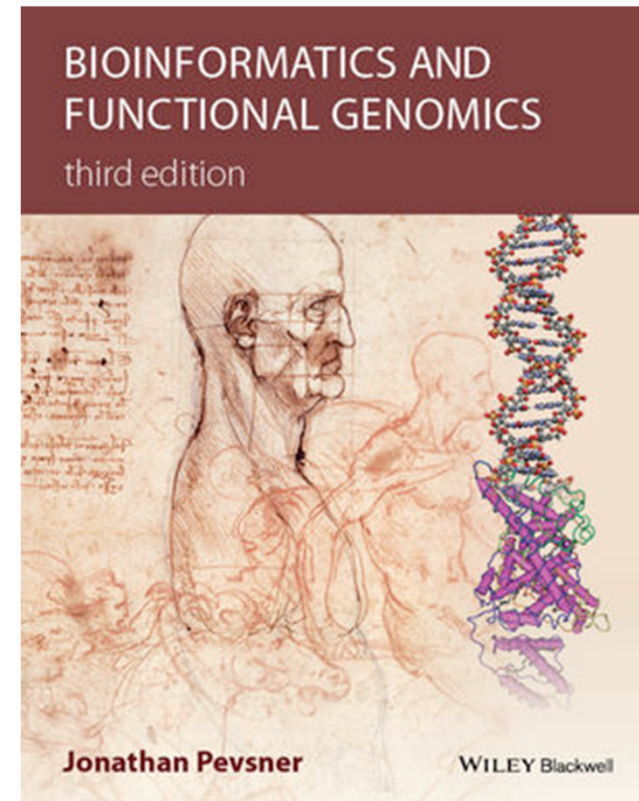
<http://statbook.gatech.edu/statb4.pdf>

Matlab exercises and datasets are at

<http://springer.bme.gatech.edu>

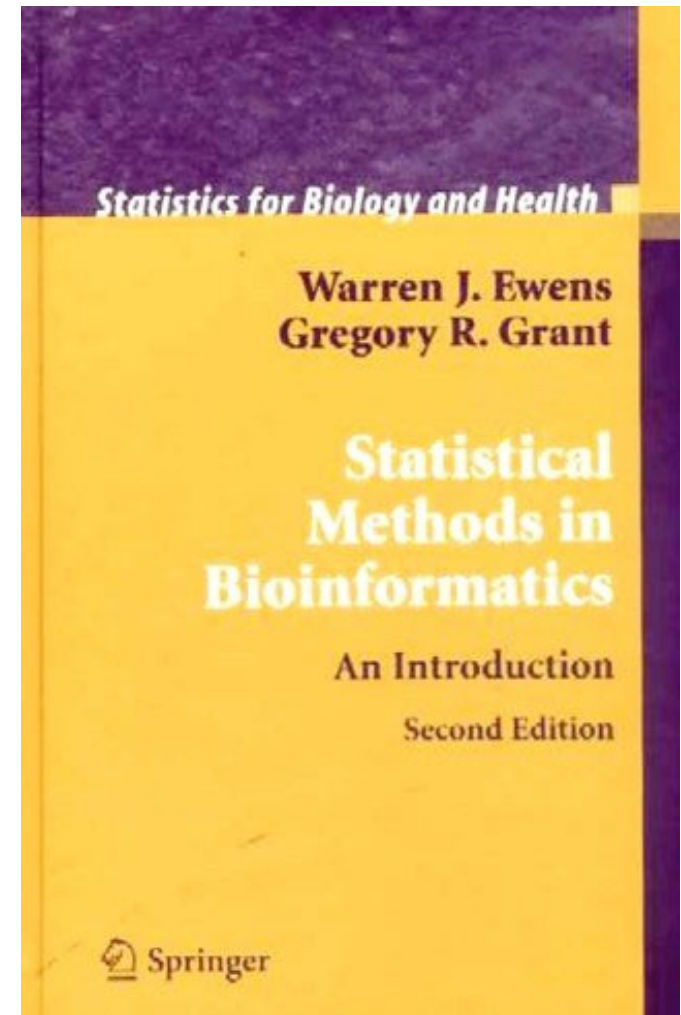
Genomics/Systems Biology Textbook

- 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*



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 binding strengths

Will not be covered:

- Imaging data such as e.g. fMRI brain scans
- Ecosystem population dynamics ☹️

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 Lots of Data

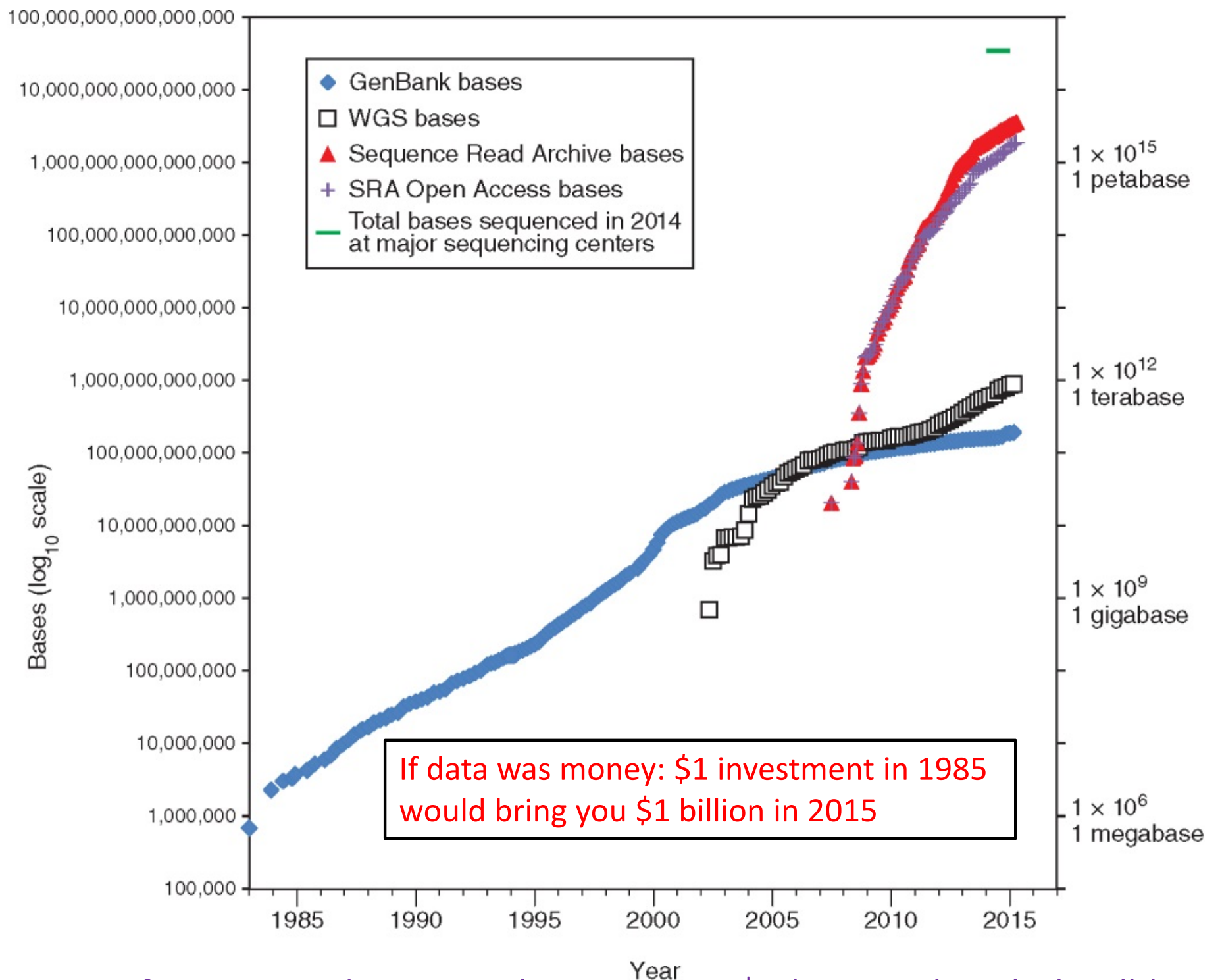
Base pairs	Unit	Abbreviation	Example
1	1 base pair	1 bp	A, C, G, T = 2 bits = 0.25 bytes
1000	1 kilobase pair	1 kb	
1,000,000	1 megabase pair	1 Mb	
10 ⁹	1 gigabase pair	1 Gb	
10 ¹²	1 terabase pair	1 Tb	
10 ¹⁵	1 petabase pair	1 Pb	

3

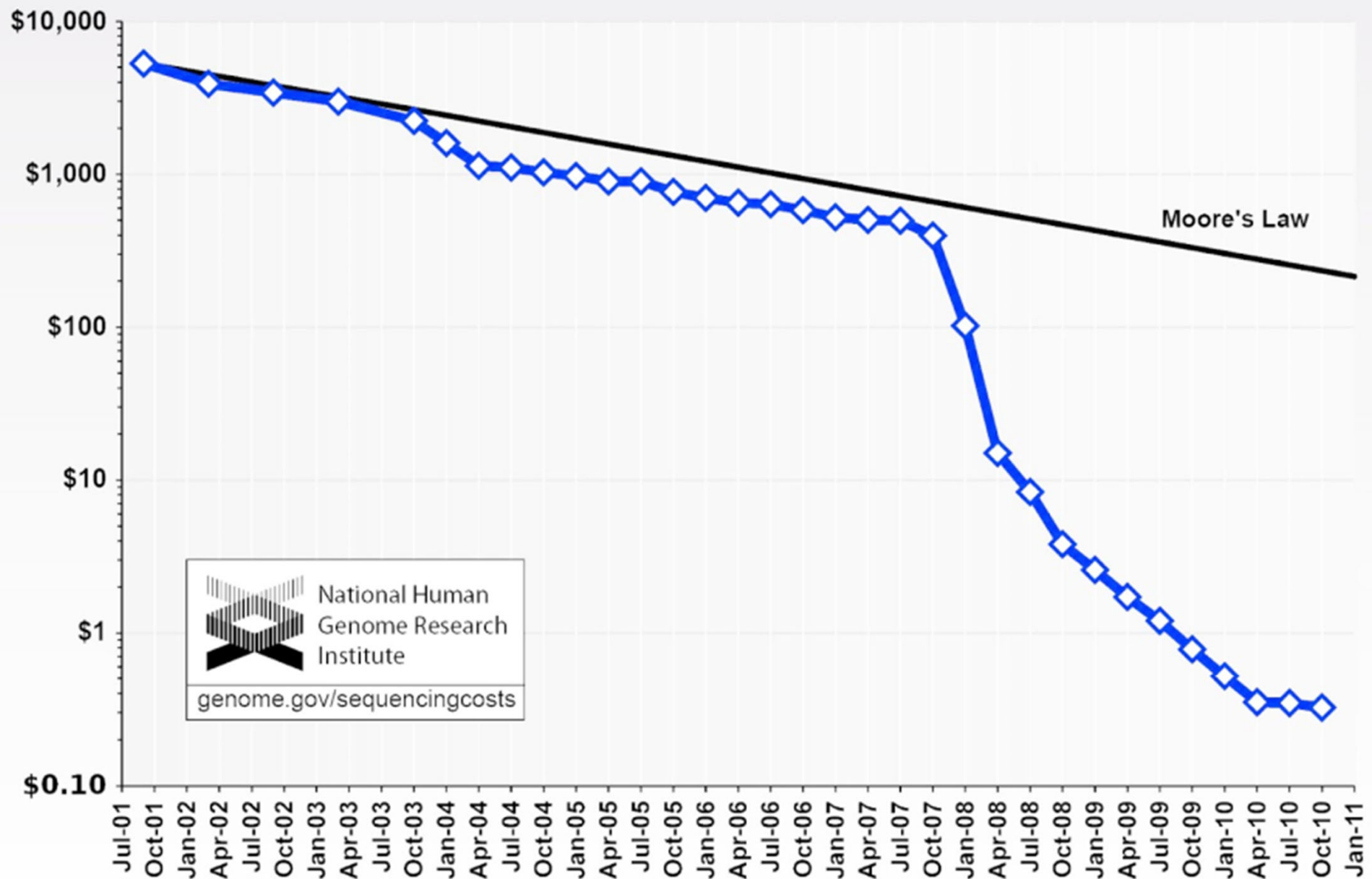
Size	Abbreviation	No. bytes	Examples
Bytes	–	1	1 byte is typically 8 bits, used to encode a single character of text
Kilobytes	1 kb	10 ³	Size of a text file with up to 1000 characters
Megabytes	1 MB	10 ⁶	Size of a text file with 1 million characters
Gigabytes	1 GB	10 ⁹	600 GB: size of GenBank (uncompressed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

Bacterial genome = "War & Peace"

Human Genome

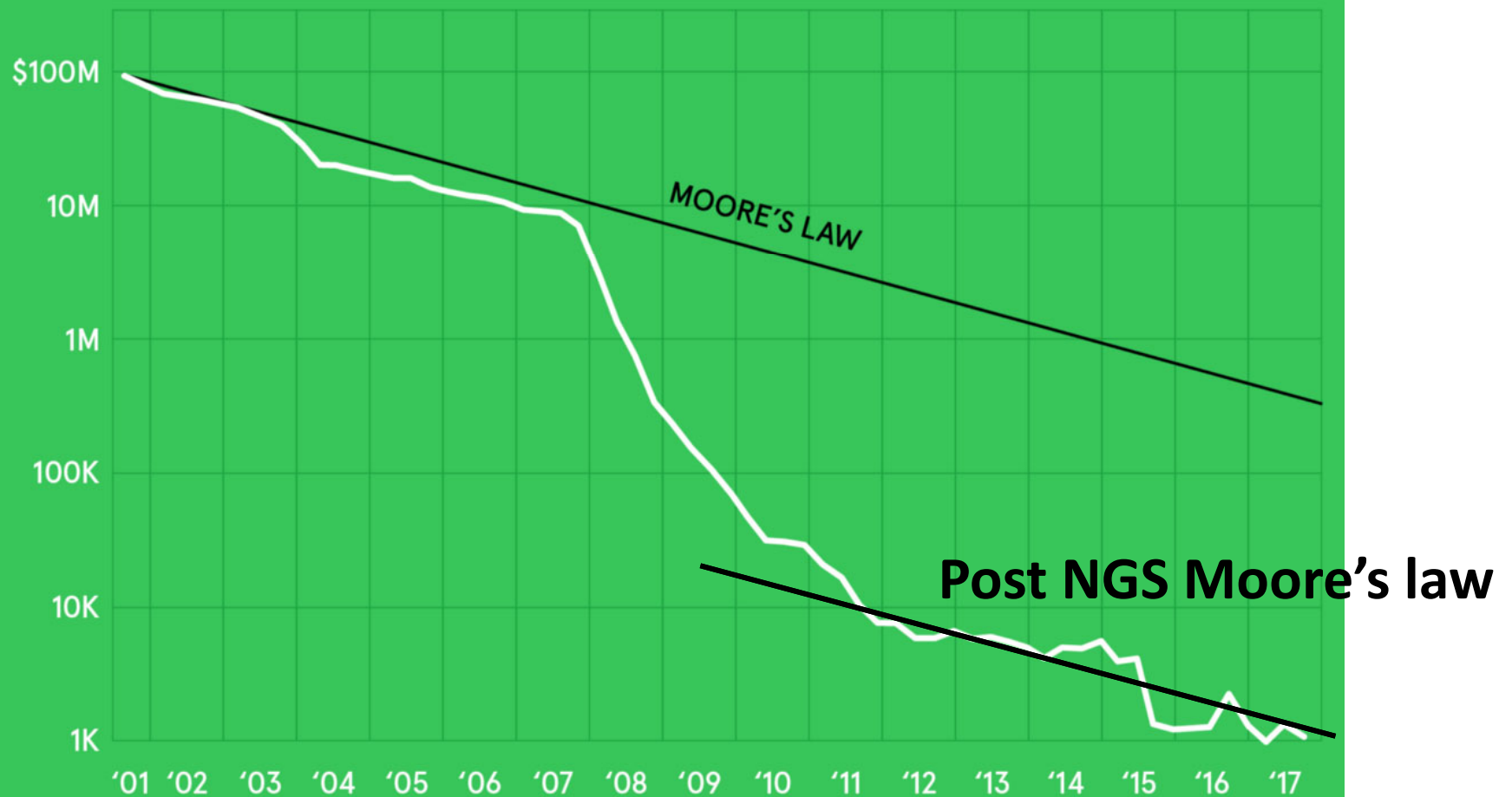


Cost per Megabase of DNA Sequence



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?

<u>Data Phase</u>	<u>Astronomy</u>	<u>Twitter</u>
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}		
<u>YouTube</u>	<u>Genomics</u>	
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.

Plot used to make this prediction

Growth of DNA Sequencing

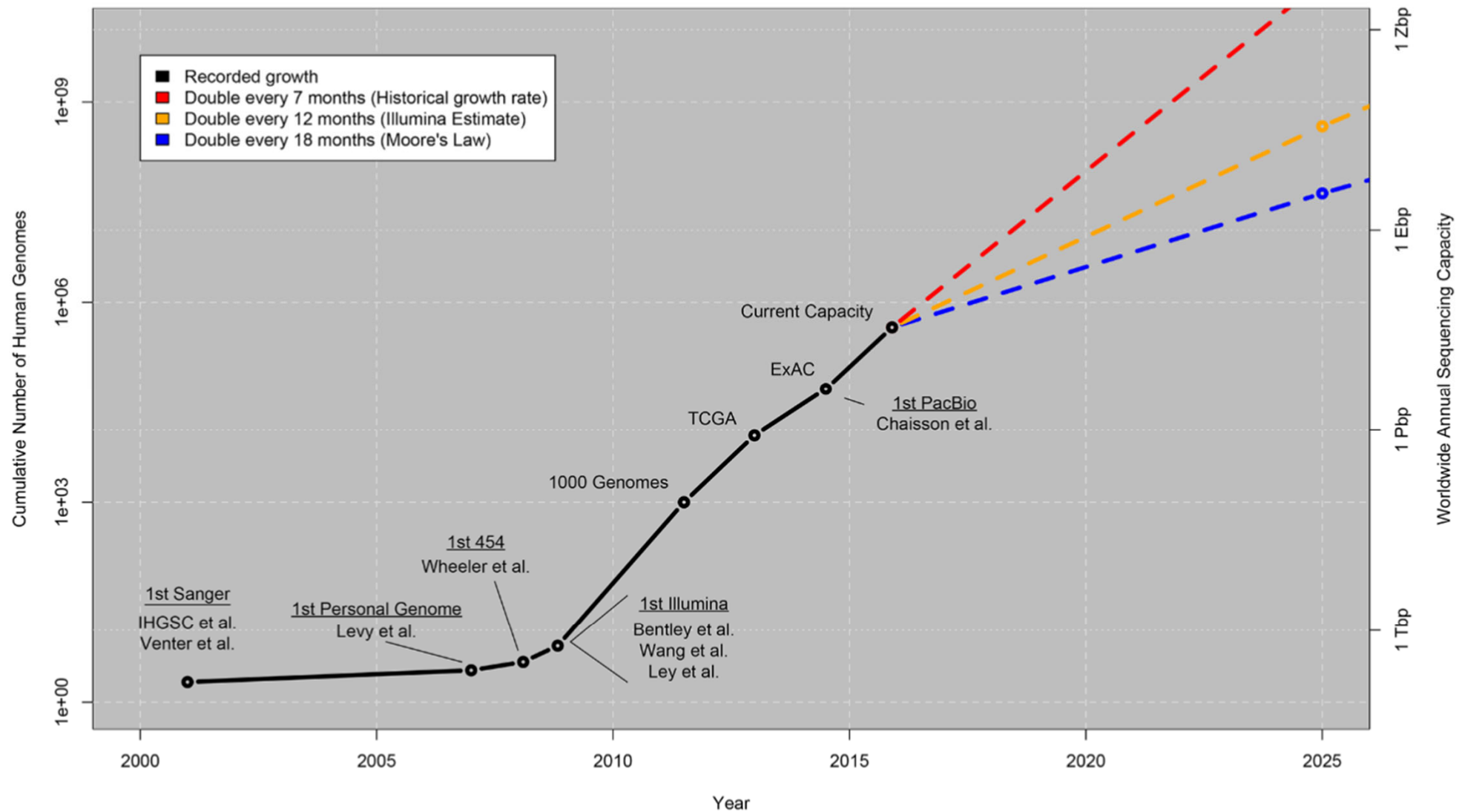
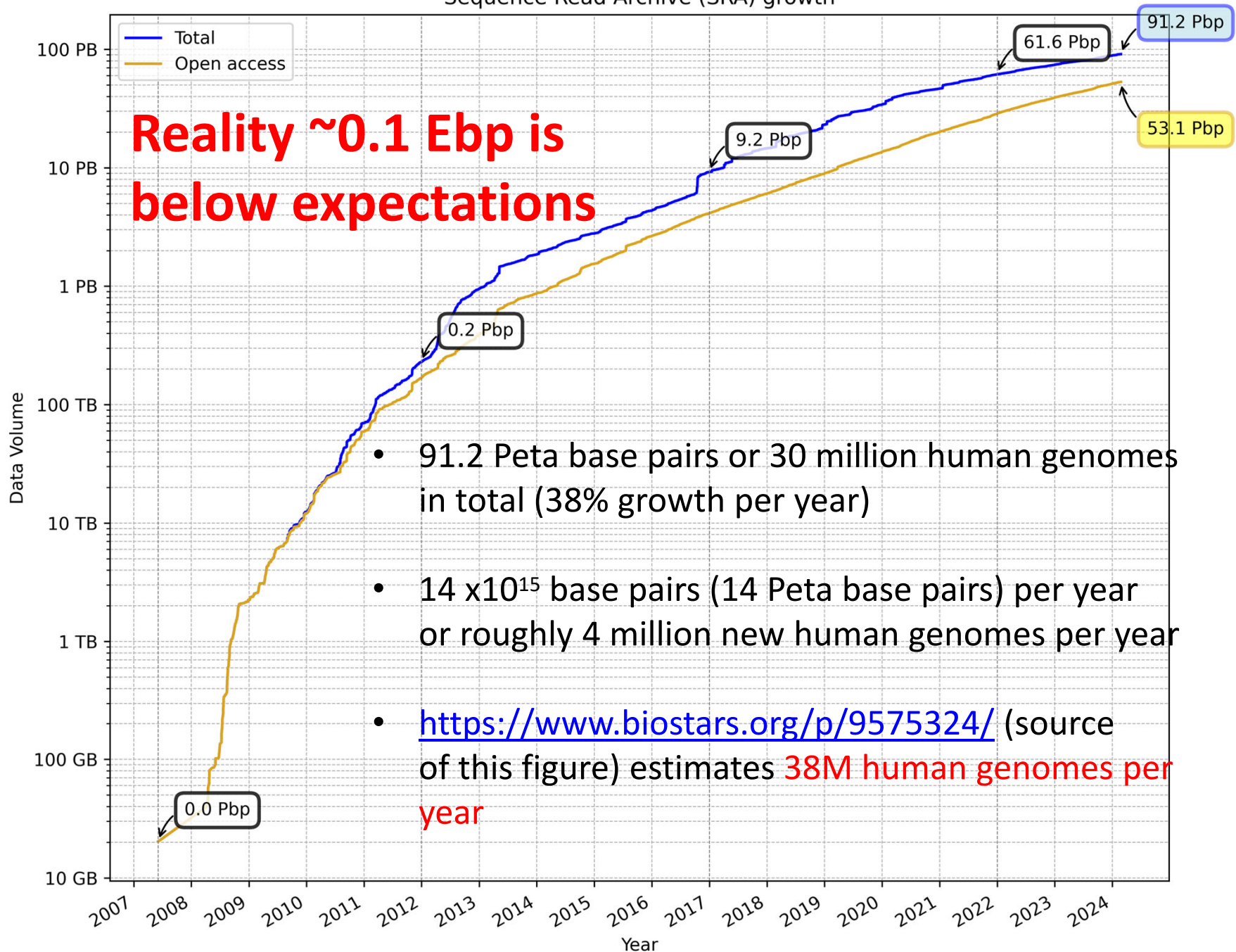


Fig 1. Growth of DNA sequencing. The plot shows the growth of DNA sequencing both in the total number of human genomes sequenced (left axis) as well as the worldwide annual sequencing capacity (right axis: Tera-basepairs (Tbp), Peta-basepairs (Pbp), Exa-basepairs (Ebp), Zetta-basepairs (Zbps)). The values through 2015 are based on the historical publication record, with selected milestones in sequencing (first Sanger through first PacBio human genome published) as well as three exemplar projects using large-scale sequencing: the 1000 Genomes Project, aggregating hundreds of human genomes by 2012 [3]; The Cancer Genome Atlas (TCGA), aggregating over several thousand tumor/normal genome pairs [4]; and the Exome Aggregation Consortium (ExAC), aggregating over 60,000 human exomes [5]. Many of the genomes sequenced to date have been whole exome rather than whole genome, but we expect the ratio to be increasingly favored towards whole genome in the future. The values beyond 2015 represent our projection under three possible growth curves as described in the main text.

Sequence Read Archive (SRA) growth



<https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/>

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



How DNA could store all the world's data

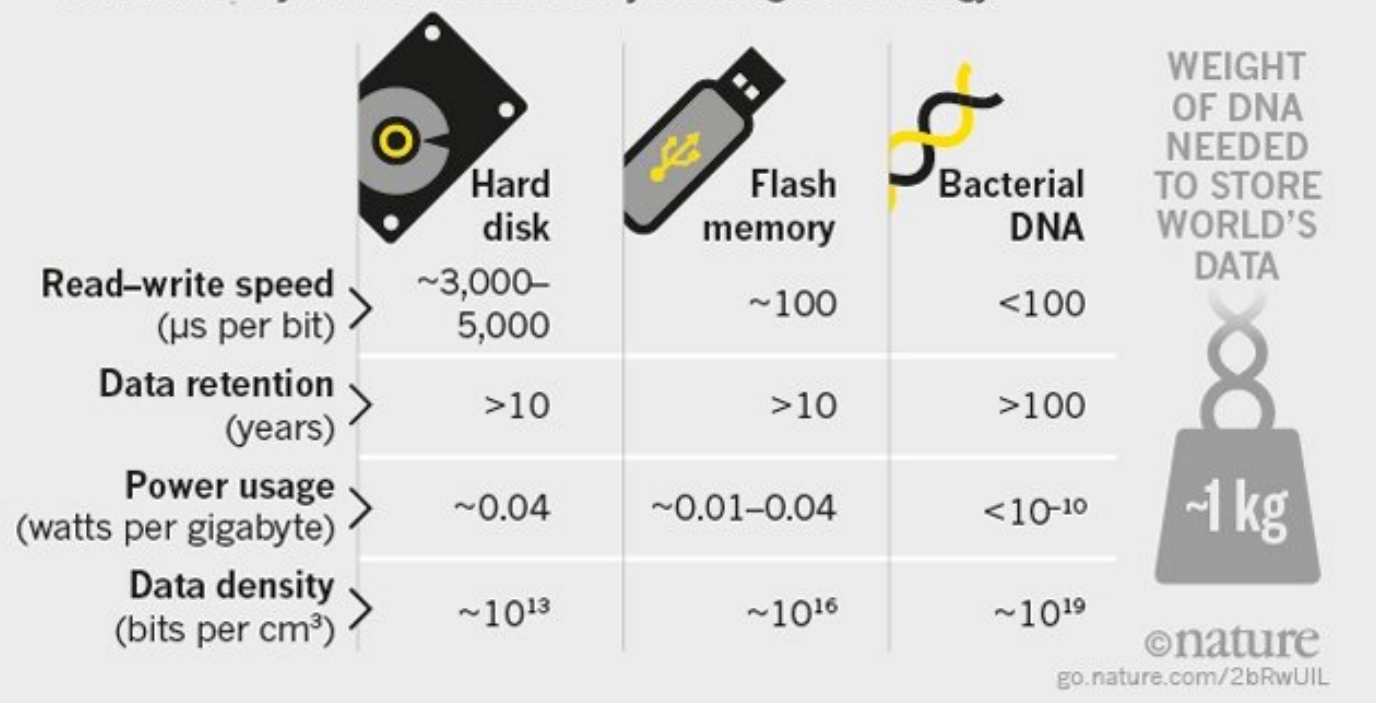
Modern archiving technology may hold an answer to that problem

Andy Extnance

31 August 2016

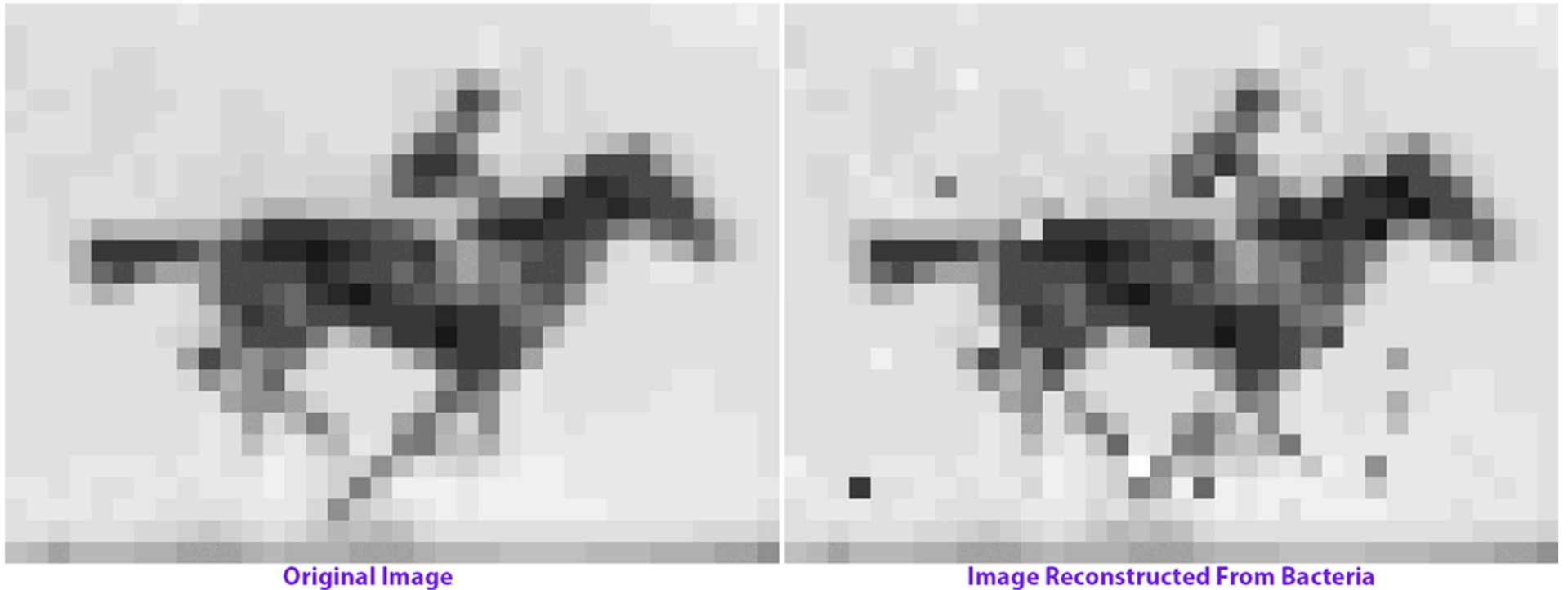
STORAGE LIMITS

Estimates based on bacterial genetics suggest that digital DNA could one day rival or exceed today's storage technology.



- 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
- <https://www.youtube.com/watch?v=7Hk9jct2ozY>
- 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?

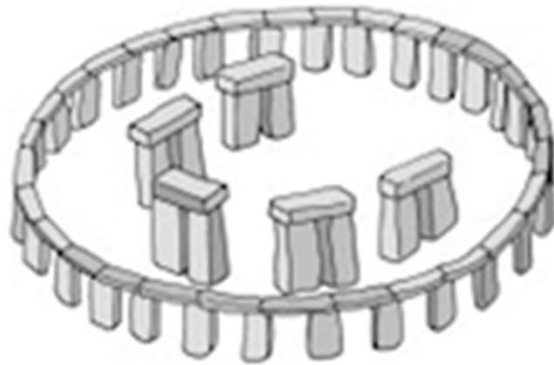
Primer on complex system

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

HËNJ



80x



30x



30x



10x



5x



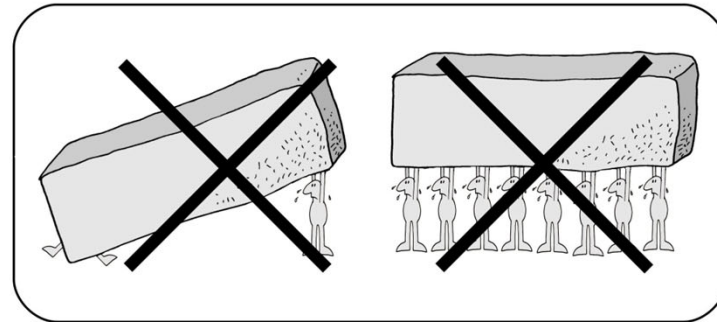
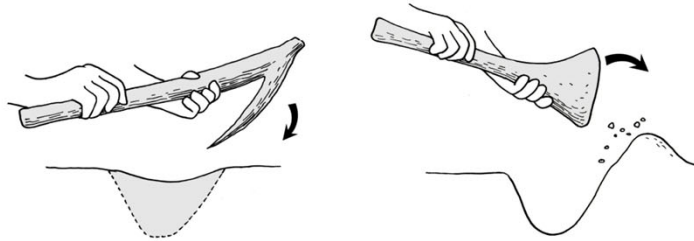
1x



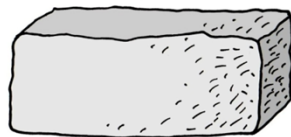
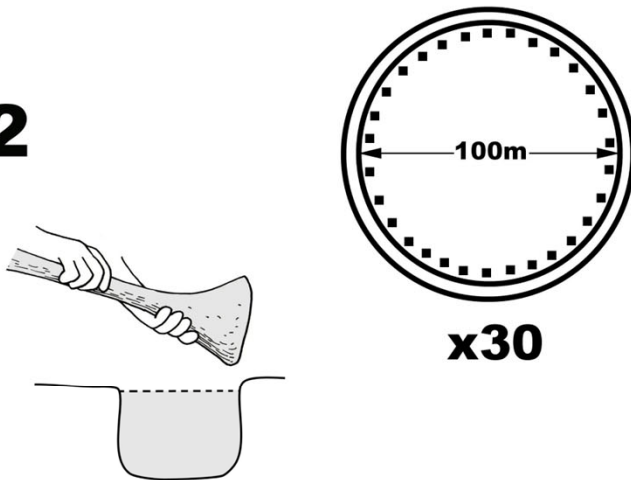
3x

Parts interact → they need to be assembled to work

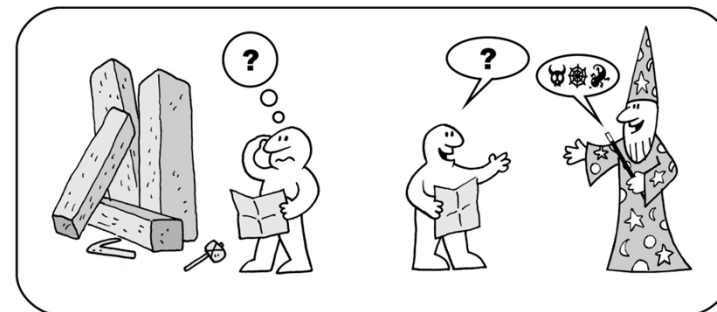
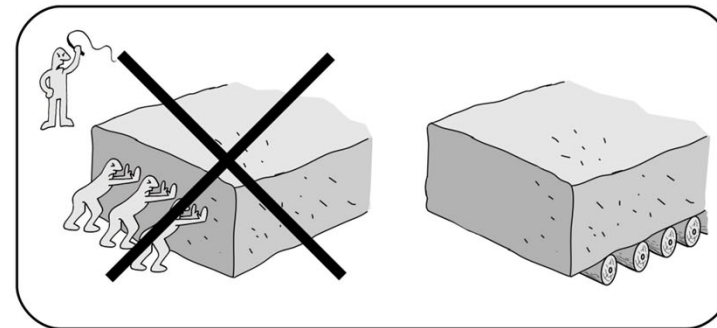
1



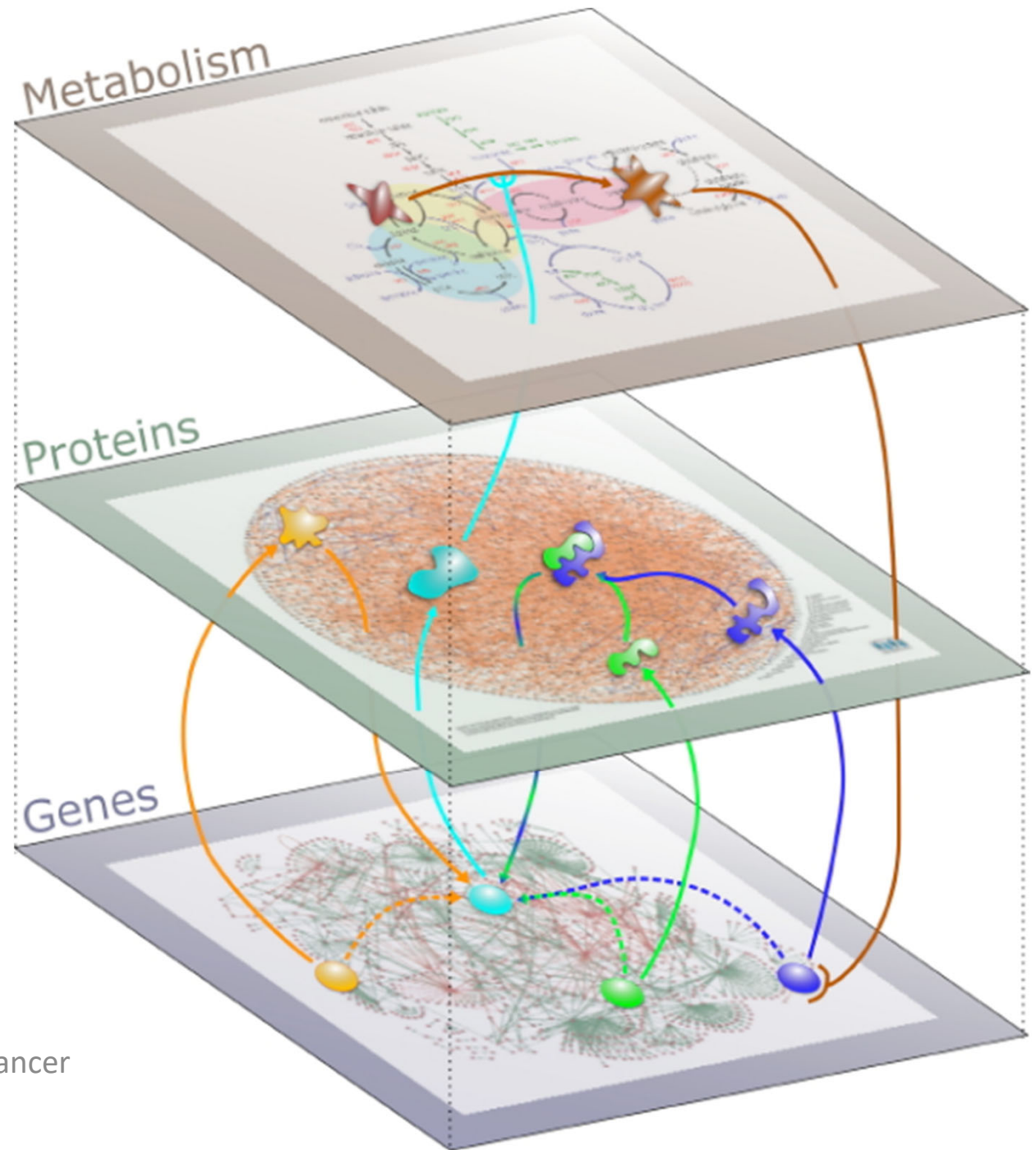
2



30x



Intra-cellular Networks operate on multiple levels

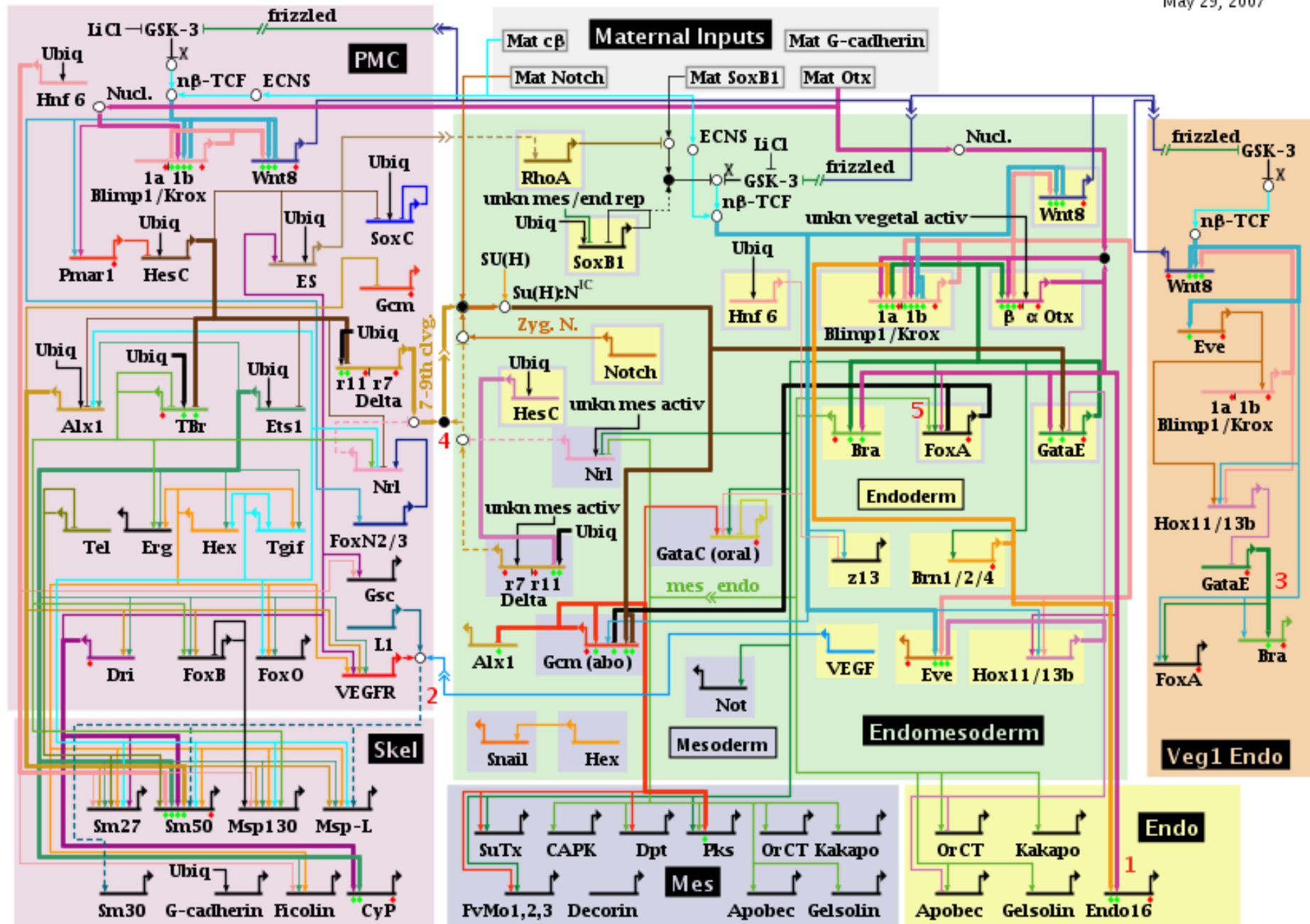


Slides by Amitabh Sharma, PhD

Northeastern University & Dana Farber Cancer
Institute

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

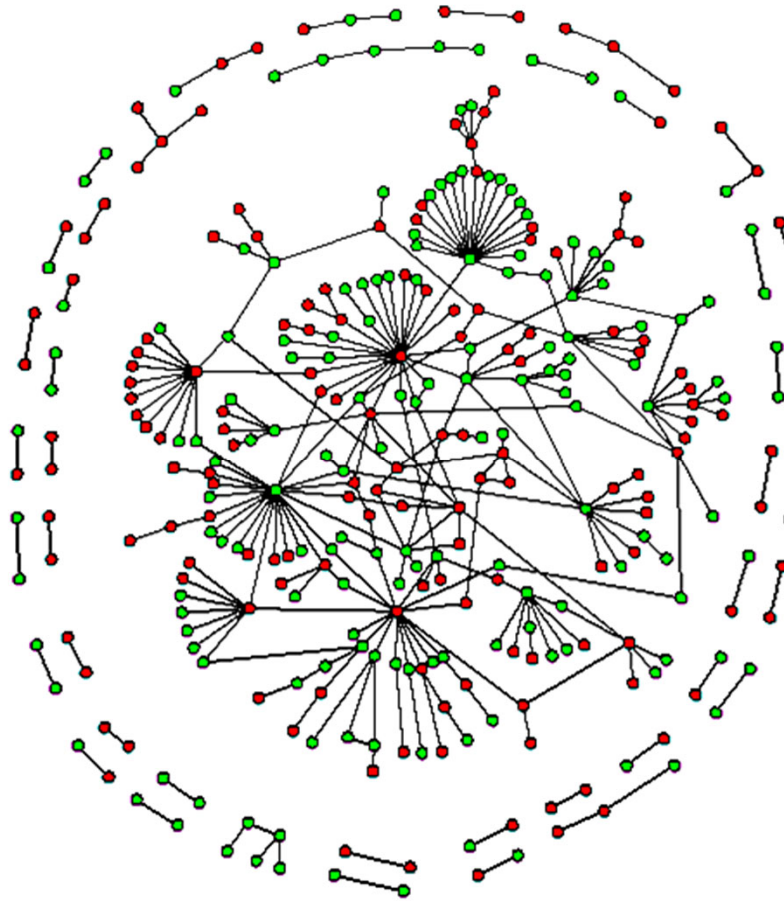
May 29, 2007



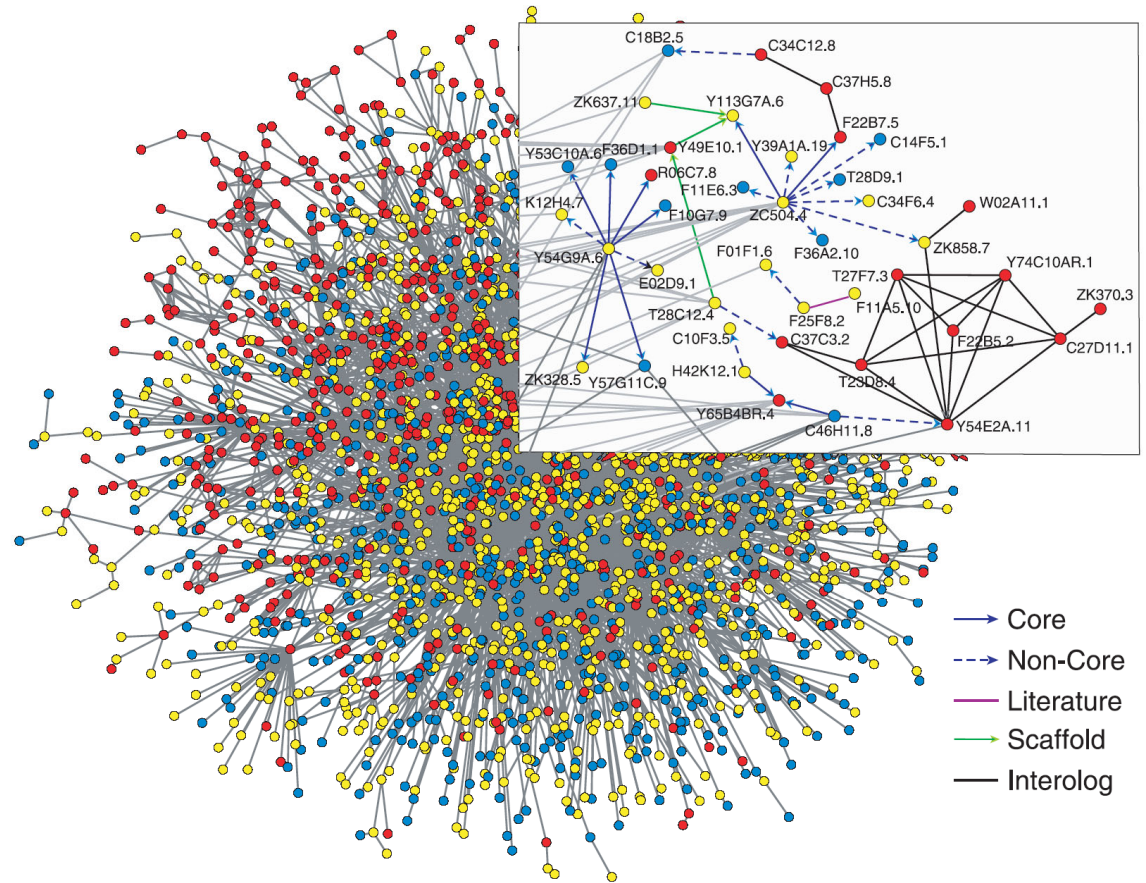
Ubiq=ubiquitous; Mat = maternal; activ = activator; rep = repressor;
 unkn = unknown; Nucl. = nuclearization; x = β-catenin source;
 nβ-TCF = nuclearized β-catenin-Tcf1; ES = early signal;
 ECNS = early cytoplasmic nuclearization system; Zyg. N. = zygotic Notch

Copyright © 2001-2007 Hamid Bolouri and Eric Davidson

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

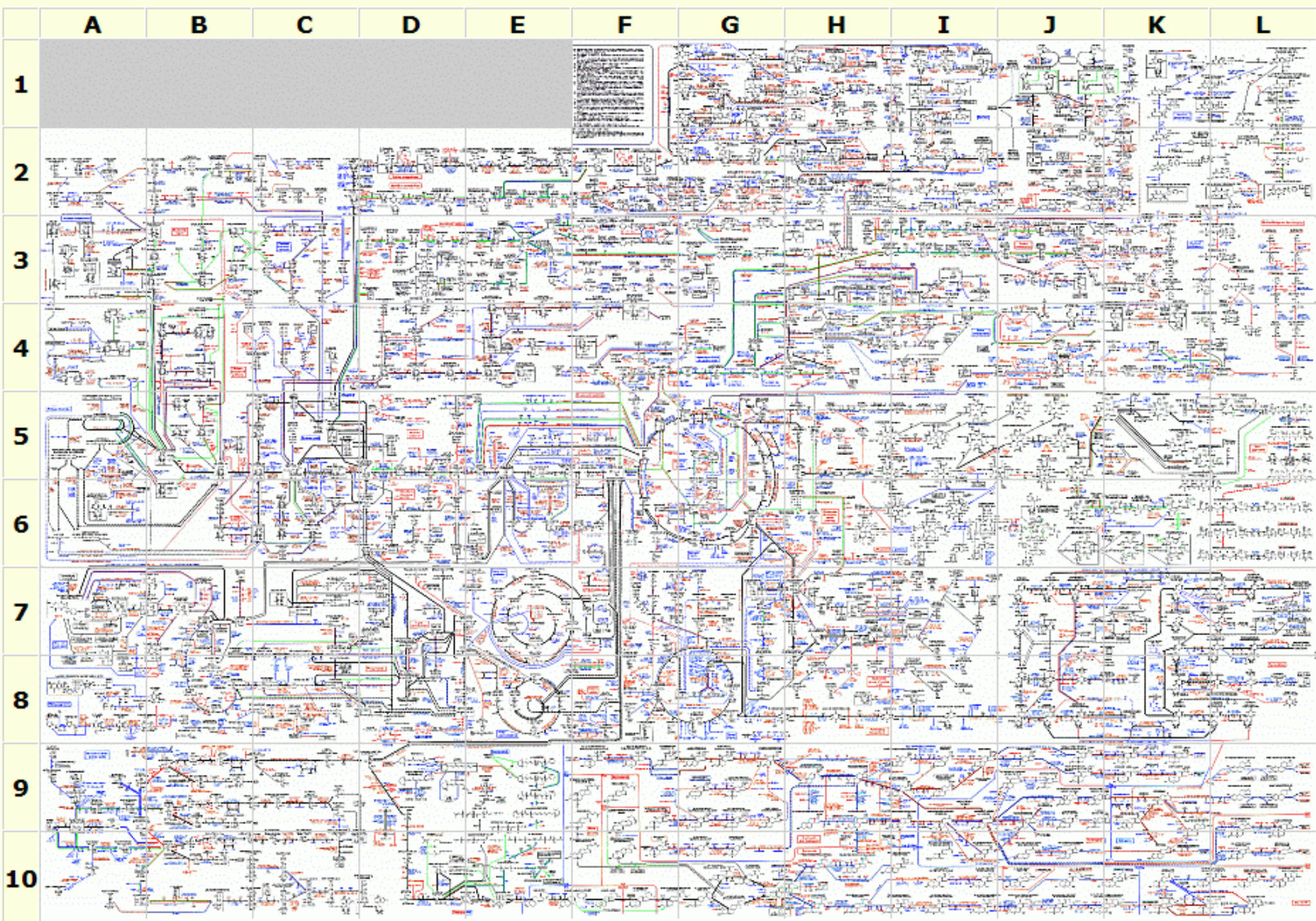


Baker's yeast *S. cerevisiae* (only nuclear proteins shown)
From S. Maslov, K. Sneppen, Science 2002



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

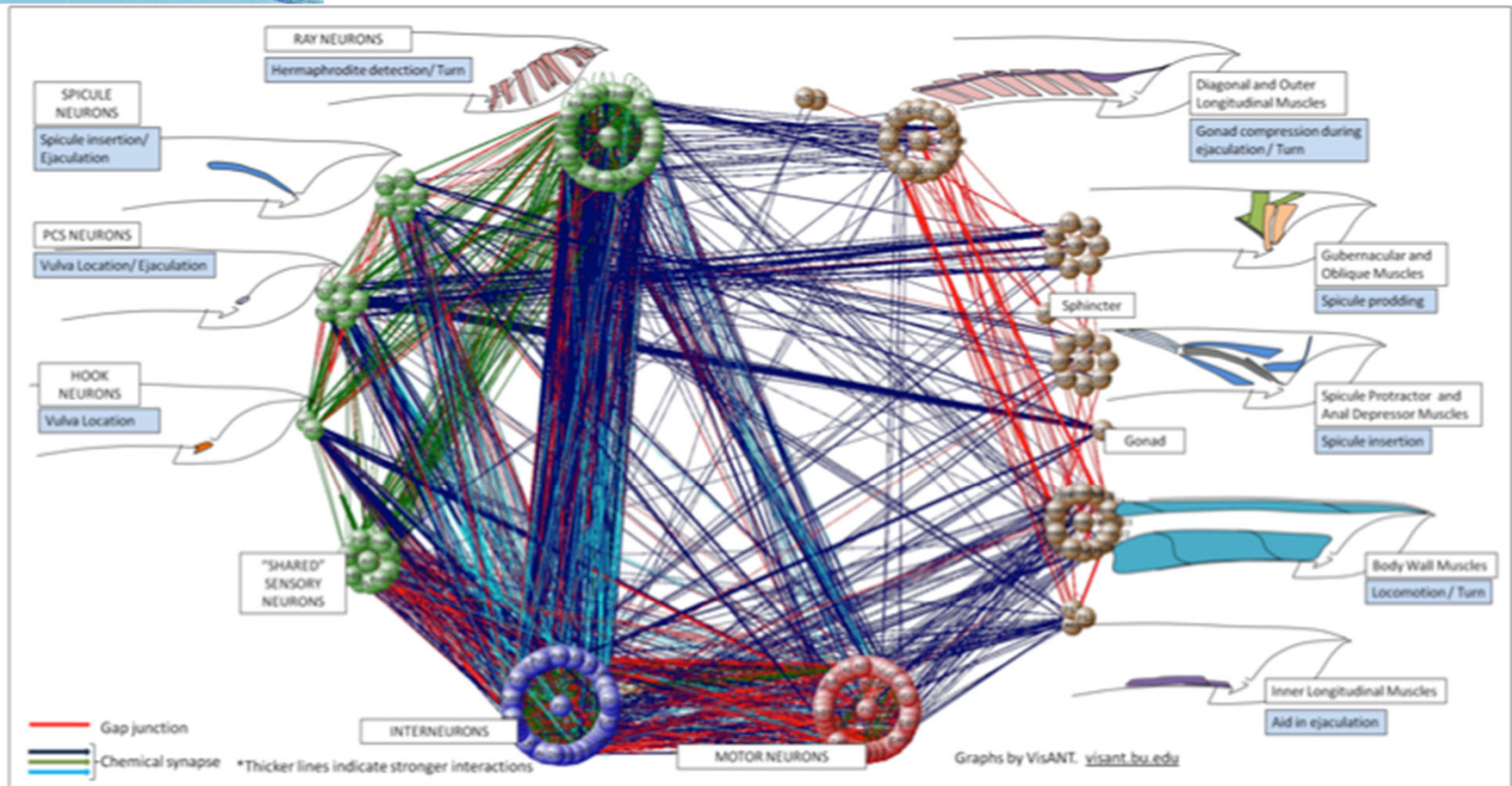
Metabolic pathway chart by ExPASy: 5702 reactions as of December 2015

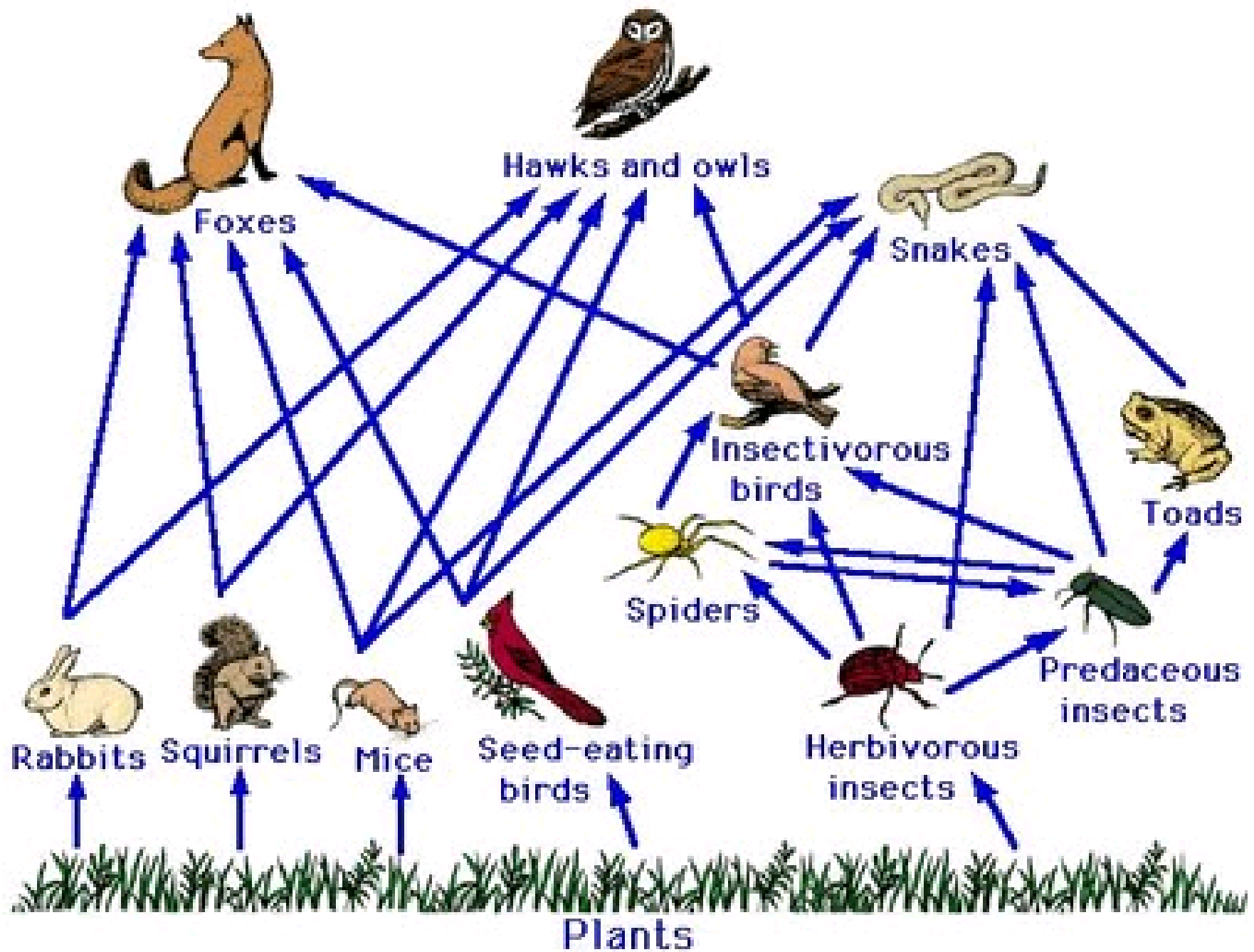


Brain and nerves of a worm



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



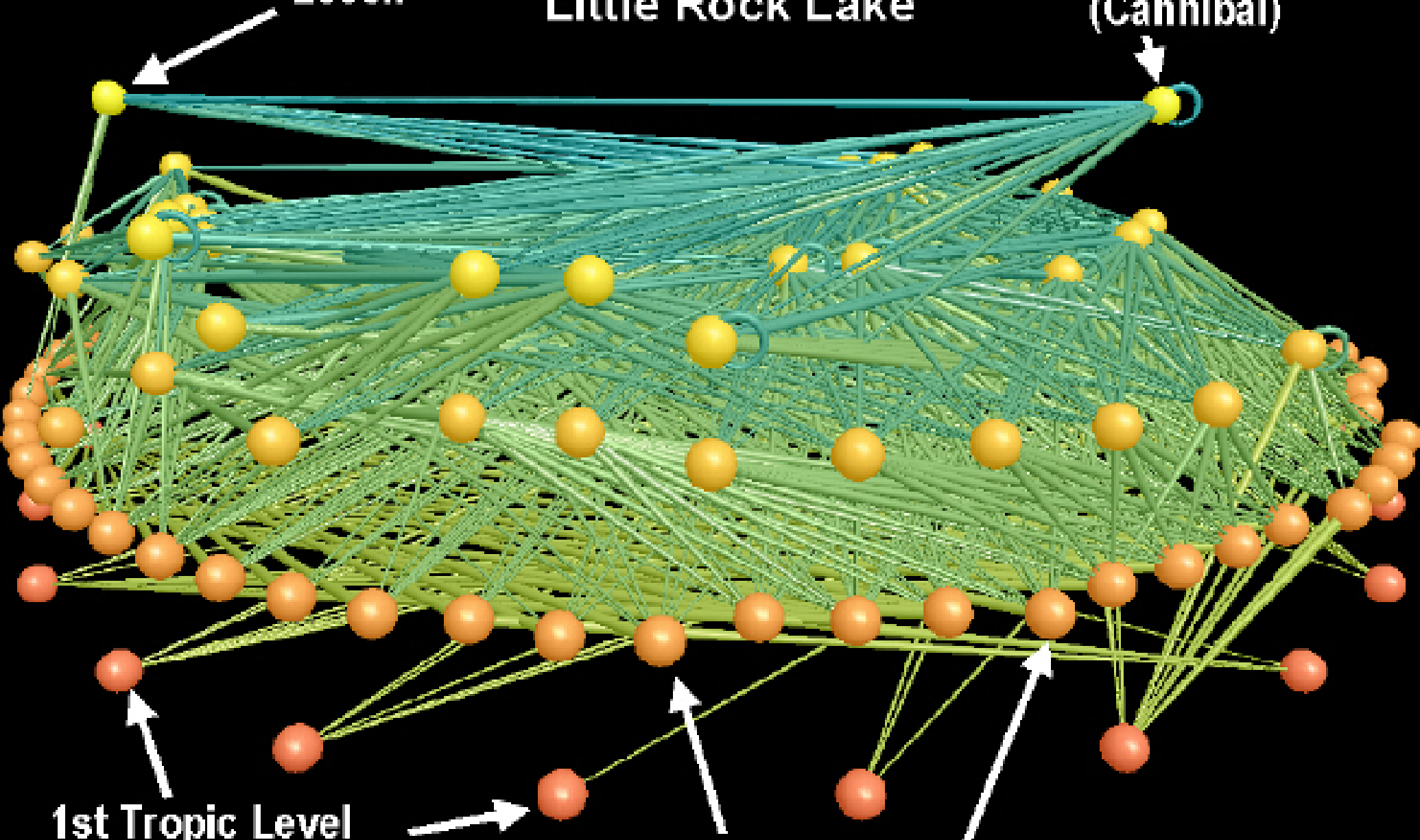


by Neo Martinez and Richard Williams

Food Web of Little Rock Lake

Smallmouth Bass
(Cannibal)

Leech



1st Trophic Level
Mostly Phytoplankton

2nd Trophic Level
Many Zooplankton

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

Int. ARCD

WHY AREN'T MY QU
WHY AREN'T MY QU
WHY AREN'T THERE

EN DYING
 AIL LAYING EGGS
 IL EGGS HATCHING
 ANY FOREIGN MILITARY BASES IN AMERICA

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

SPRING
 LAND
 WORKS

二、

Credit: XKCD
comics