Hypothesis testing: two samples

10-2: Inference for a Difference in Means of Two Normal Distributions, Variances Known

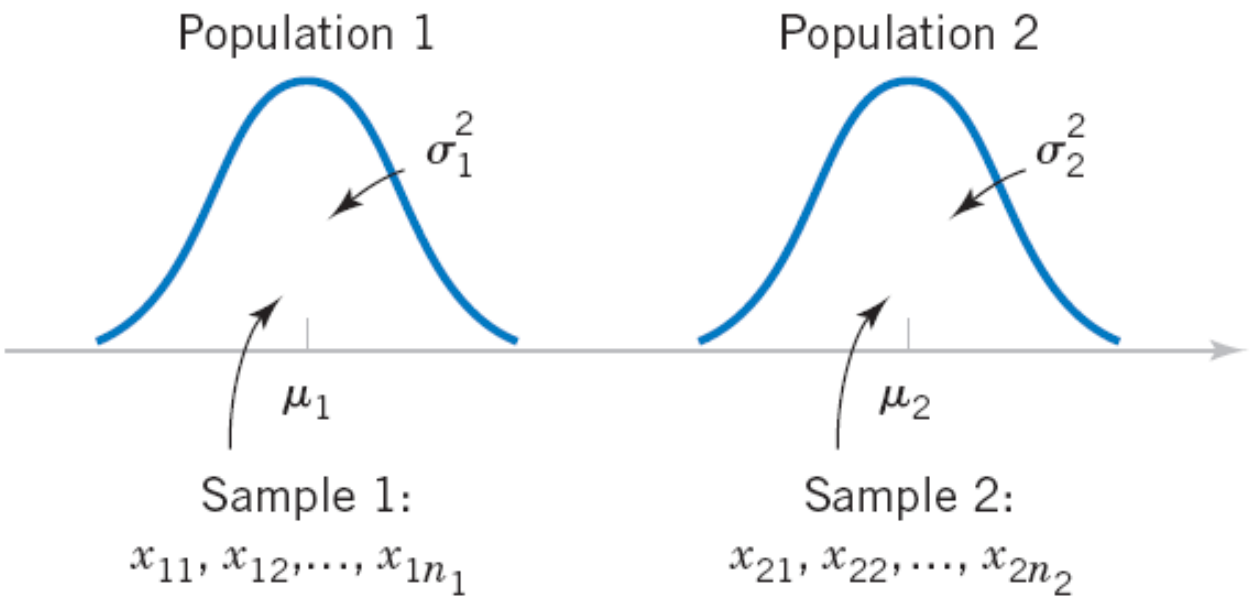


Figure 10-1 Two independent populations.

Figure 10-1 Two independent populations.

10-2: Inference for a Difference in Means of Two Normal Distributions, Variances Known

Assumptions

1. $X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample from population 1.
2. $X_{21}, X_{22}, \ldots, X_{2n_2}$ is a random sample from population 2.
3. The two populations represented by X_1 and X_2 are independent.
4. Both populations are normal.

$$E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

$$V(\bar{X}_1 - \bar{X}_2) = V(\bar{X}_1) + V(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

10-2: Inference for a Difference in Means of Two Normal Distributions, Variances Known

The quantity

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (10\text{-}1)$$

has a $N(0, 1)$ distribution.

10-2: Inference for a Difference in Means of Two Normal Distributions, Variances Known

10-2.1 Hypothesis Tests for a Difference in Means, Variances Known

usually $\Delta_0 = 0$

Null hypothesis: $H_0: \mu_1 - \mu_2 = \Delta_0$

Test statistic:
$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (10\text{-}2)$$

Alternative Hypotheses	*P*-Value	Rejection Criterion For for Fixed-Level Tests
$H_1: \mu_1 - \mu_2 \neq \Delta_0$	Probability above $\lvert z_0 \rvert$ and probability below $-\lvert z_0 \rvert$, $P = 2[1 - \Phi(\lvert z_0 \rvert)]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1: \mu_1 - \mu_2 > \Delta_0$	Probability above z_0, $P = 1 - \Phi(z_0)$	$z_0 > z_\alpha$
$H_1: \mu_1 - \mu_2 < \Delta_0$	Probability below z_0, $P = \Phi(z_0)$	$z_0 < -z_\alpha$

10-2.1 Hypotheses Tests on the Difference in Means, Variances Unknown

Case 2: $\sigma_1^2 \neq \sigma_2^2$

If H_0: $\mu_1 - \mu_2 = \Delta_0$ is true, the statistic

$$T_0^* = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \qquad (10\text{-}15)$$

is distributed as t-distribution with degrees of freedom given by

$$v = n_1 + n_2 - 2,$$

or more generally

Multiple null hypotheses: Bonferroni correction

- What if you have **m** independent null hypotheses? Say you have **m=25,000 genes** in a genome?
- What is the probability that at least one of the null-hypotheses will be shown to be false at significance threshold α_1?
- Answer: Family-Wise Error Rate or $FWER=1-(1-\alpha_1)^m \approx m\alpha_1$
- If m=20 and $\alpha_1=0.05$, $FWER= 0.6415$
- If you want to get $FWER< \alpha$, use $\alpha_1= \alpha/m$

Carlo Emilio Bonferroni
(1892 –1960)
Italian mathematician who worked on probability theory.

Example 10-7 Chocolate and Cardiovascular Health An article in *Nature* (2003, Vol. 424, p. 1013) described an

chocolate
In the expe
late per day
consisted o
average bo

Is there ev
plasma ant

Plasma antioxidants from chocolate

Dark chocolate may offer its consumers health benefits the milk variety cannot match.

There is some speculation that dietary flavonoids from chocolate, in particular (−)epicatechin, may promote cardiovascular health as a result of direct antioxidant effects or through antithrombotic mechanisms[1–3]. Here we show that consumption of plain, dark chocolate (Fig. 1) results in an increase in both the total antioxidant capacity and the (−)epicatechin content of blood plasma, but that these effects are markedly reduced when the chocolate is consumed with milk or if milk is incorporated as milk chocolate. Our findings indicate that milk may interfere with the absorption of antioxidants from chocolate *in vivo* and may therefore negate the potential health benefits that can be derived from eating moderate amounts of dark chocolate.

To determine the antioxidant content of different chocolate varieties, we took dark chocolate and milk chocolate prepared from the same batch of cocoa beans and defatted them twice with *n*-hexane before extracting them with a mixture of water, acetone and acetic acid (70.0:29.8:0.2 by volume). We measured their *in vitro* total antioxidant capacities using the ferric-reducing antioxidant potential (FRAP) assay[4]; FRAP reduced iron per 100 g for dark and milk chocolate, respectively. Volunteers must therefore consume twice as much milk chocolate as dark chocolate to receive a similar intake of antioxidants.

We recruited 12 healthy volunteers (7 women and 5 men with an average age of 32.2 ± 1.0 years (range, 25–35 years). Subjects were non-smokers, had normal blood lipid levels, were taking no drugs or vitamin supplements, and had an average weight of 65.8 ± 3.1 kg (range, 46.0–86.0 kg) and body-mass index of 21.9 ± 0.4 kg m^{-2} (range, 18.6–23.6 kg m^{-2}). On different days, following a crossover experimental design, subjects consumed 100 g dark chocolate, 100 g dark chocolate with 200 ml full-fat milk, or 200 g milk chocolate (containing the equivalent of up to 40 ml milk).

One hour after subjects had ingested the chocolate, or chocolate and milk, we measured the total antioxidant capacity of their plasma by FRAP assay. Plasma antioxidant levels increased significantly after consumption of dark chocolate alone, from $100 \pm 3.5\%$ to $118.4 \pm 3.5\%$ (*t*-test, $P < 0.001$), returning to baseline values ($95.4 \pm 3.6\%$) after 4 h (Fig. 2a). There was

TION.COM/ALAMY

Mauro Serafini*, Rossana Bugianesi*, Giuseppe Maiani*, Silvia Valtuena*, Simone De Santis*, Alan Crozier†

**Antioxidant Research Laboratory, Unit of Human Nutrition, National Institute for Food and Nutrition Research, Via Ardeatina 546, 00178 Rome, Italy*
e-mail: serafini@inran.it
†Plant Products and Human Nutrition Group, Graham Kerr Building, Division of Biochemistry and Molecular Biology, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK

Figure 1 Stack of benefits? Unlike its milky counterpart, dark chocolate may provide more than just a treat for the tastebuds.

could be due to the formation of secondary bonds between chocolate flavonoids and milk proteins[6,7], which would reduce the biological accessibility of the flavonoids and therefore the chocolate's potential antioxidant properties *in vivo*.

Our findings highlight the possibility

Vol. 424

Sweet matlab exercise #1

- **Download dark_vs_milk_chocolate_analysis_template.m at the course website. Correct all ?? In the file**
- **dark=[118.8 122.6 115.6 113.6 119.5 115.9 115.8 115.1 116.9 115.4 115.6 107.9];**
- **milk=[102.1 105.8 99.6 102.7 98.8 100.9 102.8 98.7 94.7 97.8 99.7 98.6]**
- **Use Z-statistics to calculate P-value of the null hypothesis H_0 that milk = dark against H_1 that dark > milk. P_value_z=2*[1-normcdf(|Z|)]**
- **Repeat using T-statistics. # of degrees of freedom is dof=2*(n-1)**
  P_value_t=2*tcdf(|T|, dof)

Sweet matlab exercise #1

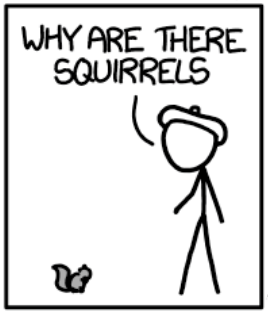
- dark=[118.8 122.6 115.6 113.6 119.5 115.9 115.8 115.1 116.9 115.4 115.6 107.9];
- milk=[102.1 105.8 99.6 102.7 98.8 100.9 102.8 98.7 94.7 97.8 99.7 98.6]
- x_dark=mean(dark) % sample mean dark chocolate
- x_milk=mean(milk) % sample mean milk chocolate
- s_dark=std(dark) % sample std dark chocolate
- s_milk=std(milk) % sample std milk chocolate
- n=12 % sample size of both dark and milk
- std_xdiff=sqrt(s_dark.^2./2+s_milk.^2./n) % std diff x
- z_stat=(x_dark-x_milk)./std_xdiff % z-statistic
- P_value_z=erfc(z_stat./sqrt(2))./2 % P-value of null true
- % P_value_z=9.9629e-34
- dof=(n-1)+(n-1) % # of degrees of freedom
- P_value_t=tcdf(z_stat,dof,'upper') % P-value of null true
- %P_value_t= 1.8417e-11

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 DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE
WHY IS LYING GOOD

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 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 SEX SO IMPORTANT

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 AREN'T MY ARMS GROWING

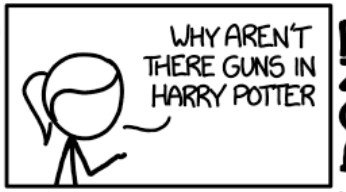
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

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 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 FEMALE MR MIMES

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 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 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 AREN'T THERE GUNS IN HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

Regression analysis
Two variables

(Montgomery and Runger: ch 11
Brani Vidakovic: ch 14)

Reminder

Covariance Defined

Covariance is a number qunatifying average dependence between two random variables.

The covariance between the random variables X and Y, denoted as $\text{cov}(X,Y)$ or σ_{XY} is

$$\sigma_{XY} = E\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right] = E\left(XY\right)-\mu_X\mu_Y \qquad (5\text{-}14)$$

The units of σ_{XY} are units of X times units of Y.

Unlike the range of variance, $-\infty < \sigma_{XY} < \infty$.

Correlation is “normalized covariance”

- Also called:
  Pearson correlation coefficient

$\rho_{XY}=\sigma_{XY}/\sigma_X\sigma_Y$
is the covariance normalized to be $-1 \leq \rho_{XY} \leq 1$

Karl Pearson (1852– 1936)
English mathematician and biostatistician

Covariance and Scatter Patterns

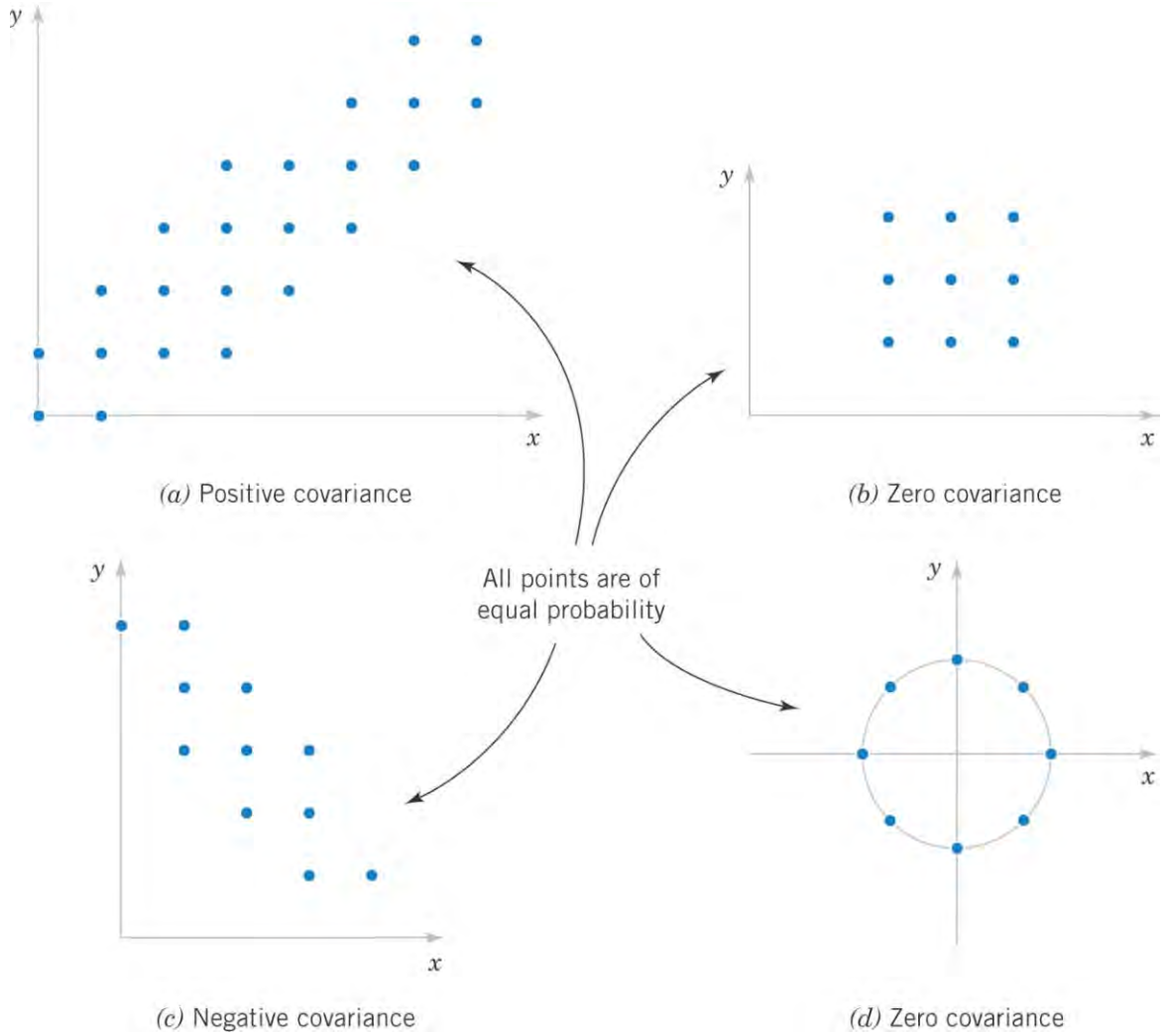


Figure 5-13 Joint probability distributions and the sign of cov(X, Y). Note that covariance is a measure of linear relationship. Variables with non-zero covariance are correlated.

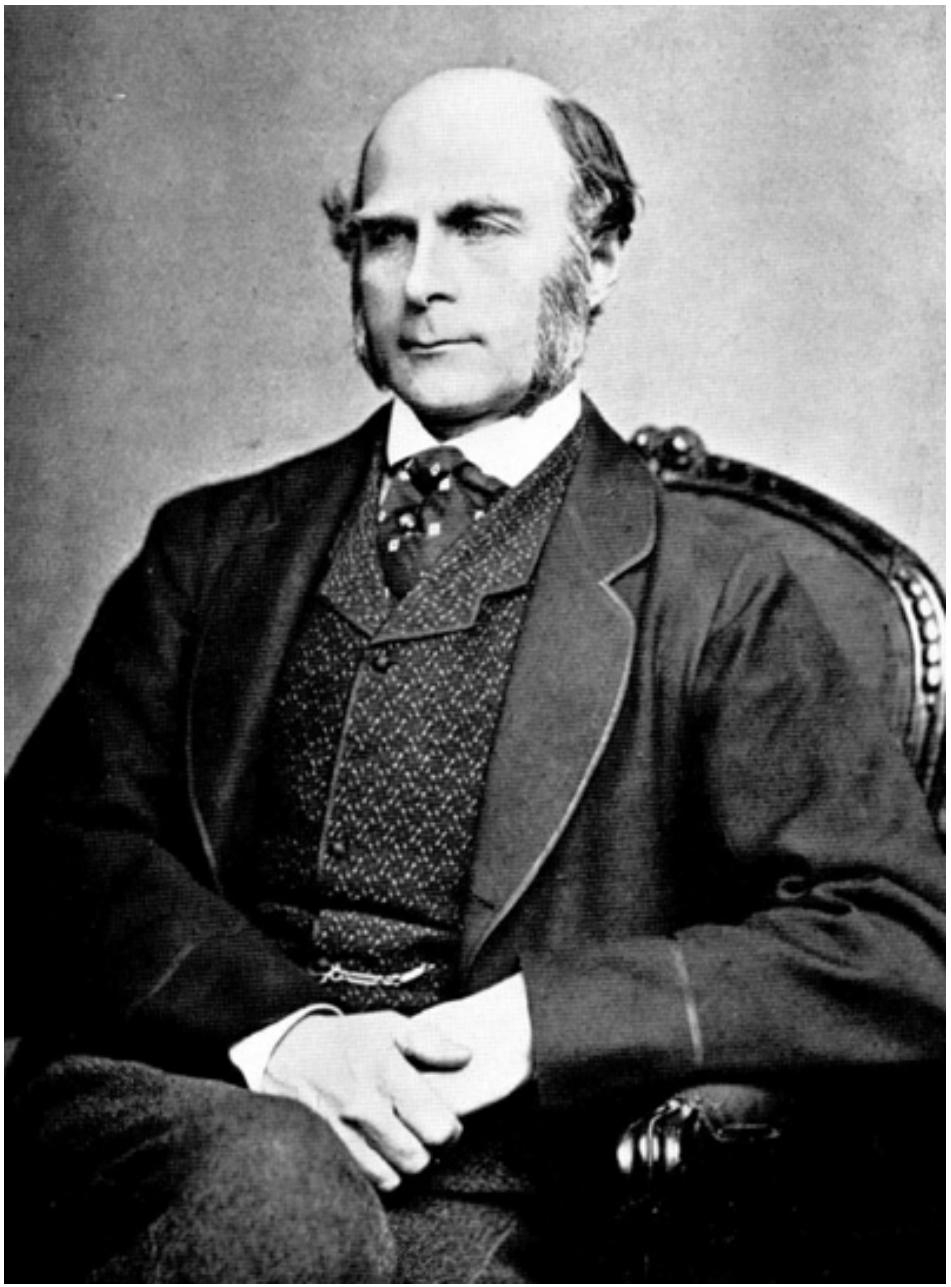
Regression analysis

• Many problems in engineering and science involve sample in which two or more variables were measured. They may not be independent from each other and one (or several) of them can be used to predict another

• Everyday example: in most samples height and weight of people are related to each other

• Biological example: in a cell sorting experiment the copy number of a protein may be measured alongside its volume

• Regression analysis uses a sample to build a model to predict protein copy number given a cell volume

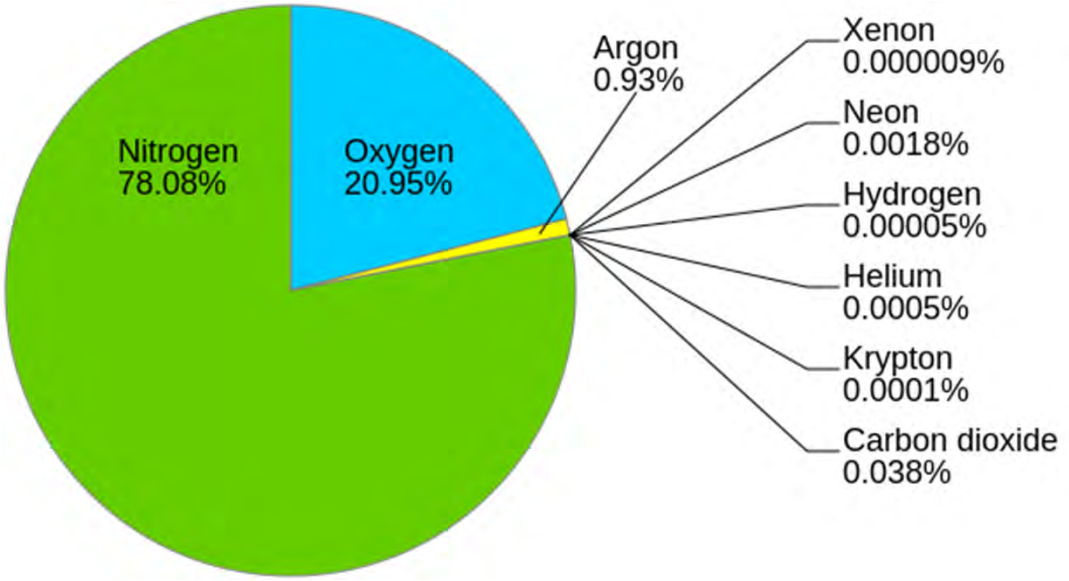


Sir Francis Galton,
(1822 -1911) was an English statistician, anthropologist, proto-geneticist, psychometrician, eugenicist, ("Nature vs Nurture", inheritance of intelligence), tropical explorer, geographer, inventor (Galton Whistle to test hearing), meteorologist (weather map, anticyclone).

Invented both correlation and regression analysis when studied heights of fathers and sons

Found that fathers with height above average tend to have sons with height also above average but closer to the average.
Hence "regression" to the mean

Two variable samples



- Oxygen can be distilled from the air
- Hydrocarbons need to be filtered out or the whole thing would go kaboom!!!
- When more hydrocarbons were removed, the remaining oxygen stays cleaner
- Except we don't know how dirty was the air to begin with

Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

$$Y = \beta_0 + \beta_1 X + \epsilon$$

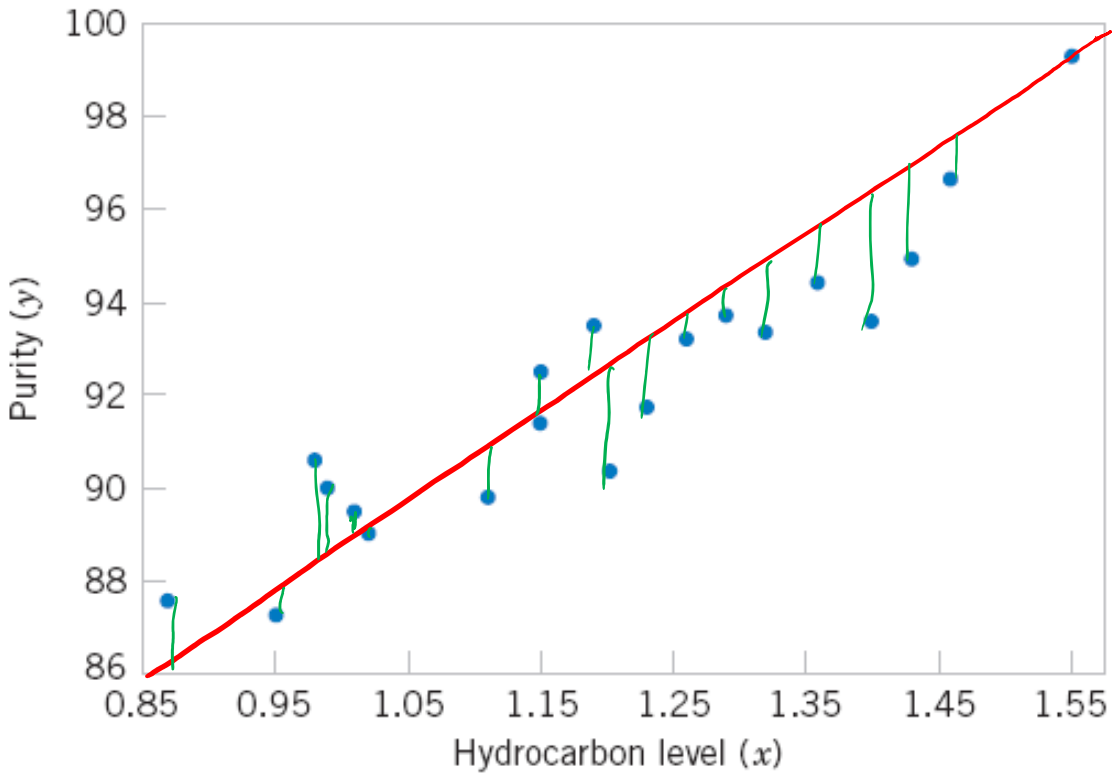


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

$$Y = 75 + 15 \cdot X + \epsilon$$

Linear regression

The **simple linear regression model** is given by

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

ε is the **random error term**

slope β_1 and intercept β_0 of the line are called **regression coefficients**

Note: Y , X and ε are random variables

The minimal assumption: $E(\varepsilon \mid x)=0$ →

$E(Y \mid x) = \beta_0 + \beta_1 x + E(\varepsilon \mid x) = \beta_0 + \beta_1 x$

$Y = \beta_0 + \beta_1 X + \epsilon$; $E(\epsilon | x) = 0 \quad \forall x$

How does one find β_0 & β_1?

$$Cov(Y, X) = Cov((\beta_0 + \beta_1 X + \epsilon), X) =$$

$$= Cov(\beta_0, X) + \beta_1 Cov(X, X) + Cov(\epsilon, X)$$

$Cov(\beta_0, X) = 0$ since β_0 is constant

$Cov(X, X) = E(X^2) - E(X)^2 = Var(X)$

$Cov(\epsilon, X) = E(\epsilon \cdot X) - E(\epsilon) \cdot E(X) =$

$= E(\epsilon \cdot X) = \sum_{all\ x} x \cdot E(\epsilon | x) = 0$

Thus $\boxed{\beta_1 = \frac{Cov(X, Y)}{Var(X)}}$ $\boxed{\beta_0 = E(Y) - \beta_1 E(X)}$

Method of least squares

- The **method of least squares** is used to estimate the parameters, β_0 and β_1 by minimizing the sum of the squares of the vertical deviations in Figure 11-3.

Figure 11-3 Deviations of the data from the estimated regression model.

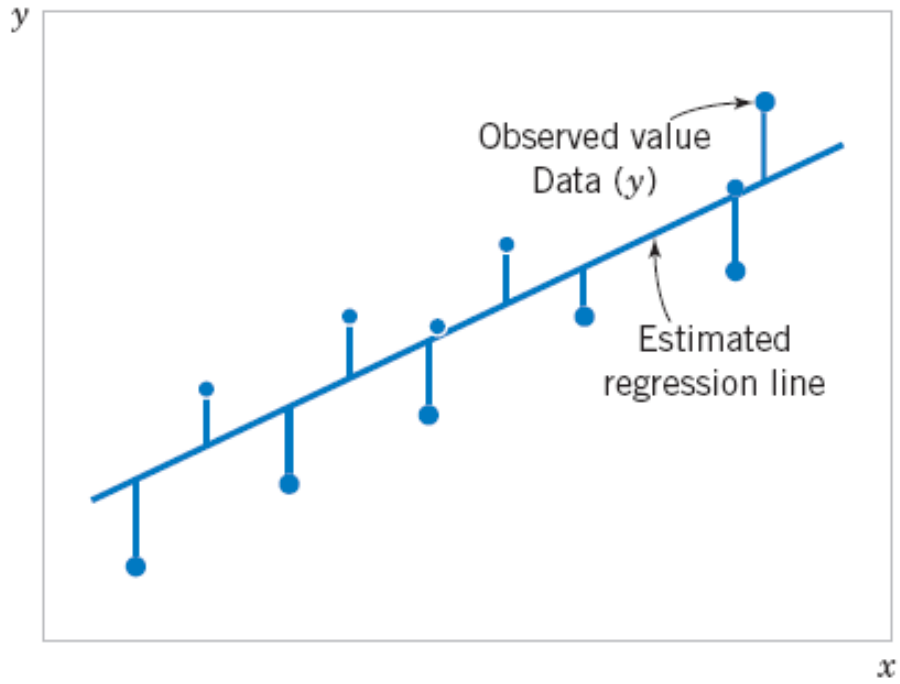


Figure 11-3 Deviations of the data from the estimated regression model.

Traditional notation

Definition

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{11-7}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{S_{xy}}{S_{xx}} \tag{11-8}$$

where $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ and $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$.

11-2: Simple Linear Regression

Definition

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{11-7}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n^2}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n^2}} = \frac{Cov(X,Y)}{Var(X)} \tag{11-8}$$

where $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ and $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$.

11-4: Hypothesis Tests in Simple Linear Regression

11-4.2 Analysis of Variance Approach to Test Significance of Regression

The analysis of variance identity is

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (11\text{-}24)$$

Symbolically,

$$SS_T = SS_R + SS_E \qquad (11\text{-}25)$$

11-7: Adequacy of the Regression Model

11-7.2 Coefficient of Determination (R^2)

VERY COMMONLY USED

- The quantity

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

is called the **coefficient of determination** and is often used to judge the adequacy of a regression model.

- $0 \le R^2 \le 1$;
- We often refer (loosely) to R^2 as the amount of variability in the data explained or accounted for by the regression model.

11-7: Adequacy of the Regression Model

11-7.2 Coefficient of Determination (R^2)

- For the oxygen purity regression model,

$$R^2 = SS_R/SS_T$$
$$= 152.13/173.38$$
$$= 0.877$$

- Thus, the model accounts for 87.7% of the variability in the data.

11-2: Simple Linear Regression

Estimating σ_ε^2

An **unbiased estimator** of σ_ε^2 is

$$\hat{\sigma}_\varepsilon^2 = \frac{SS_E}{n-2} \qquad (11\text{-}13)$$

where SS_E can be easily computed using

$$SS_E = SS_T - \hat{\beta}_1 S_{xy} \qquad (11\text{-}14)$$

11-3: Properties of the Least Squares Estimators

- Slope Properties

$$E(\hat{\beta}_1) = \beta_1 \qquad V(\hat{\beta}_1) = \frac{\hat{\sigma}_\varepsilon^2}{S_{xx}} = \frac{\hat{\sigma}_\epsilon^2}{n\,\sigma_x^2}$$

Large n → small variance of β_1

- Intercept Properties

$$E(\hat{\beta}_0) = \beta_0 \quad \text{and} \quad V(\hat{\beta}_0) = \hat{\sigma}_\varepsilon^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right] =$$

$$= \hat{\sigma}_\varepsilon^2 \left[1 + \frac{\mu_x^2}{\sigma_x^2}\right] \times \frac{1}{n}$$

11-4: Hypothesis Tests in Simple Linear Regression

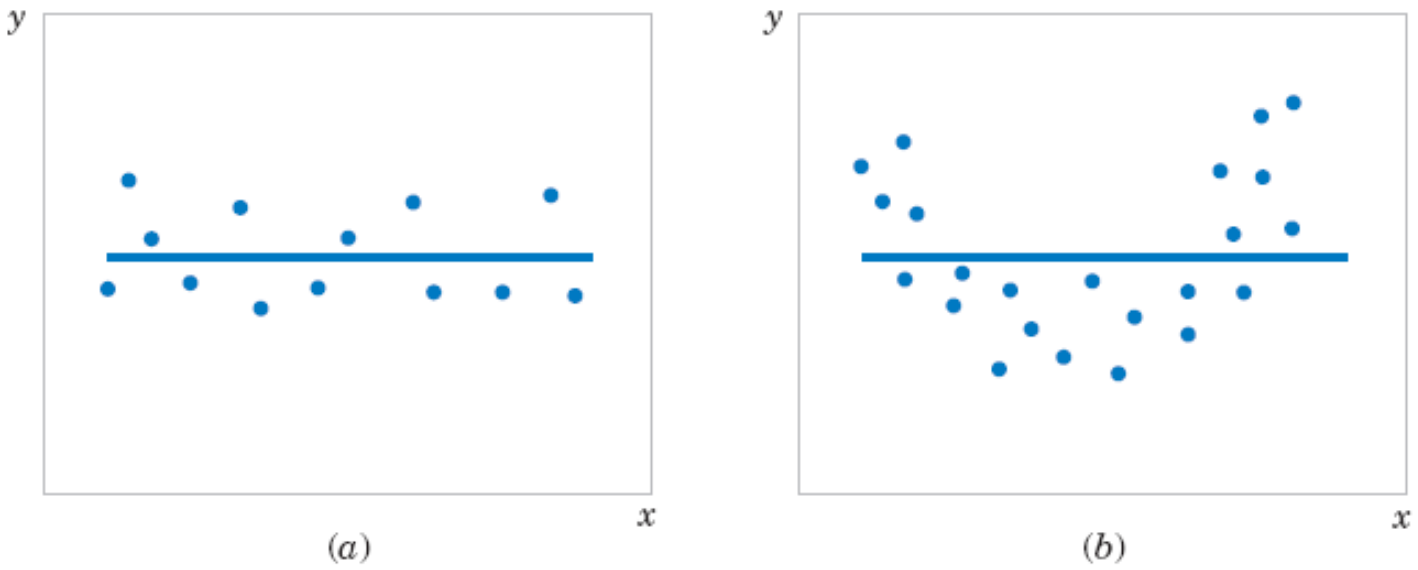


Figure 11-5 The hypothesis H_0: $\beta_1 = 0$ is not rejected.

Figure 11-5 The null hypothesis H_0: $\beta_1 = 0$ is accepted.

11-4: Hypothesis Tests in Simple Linear Regression

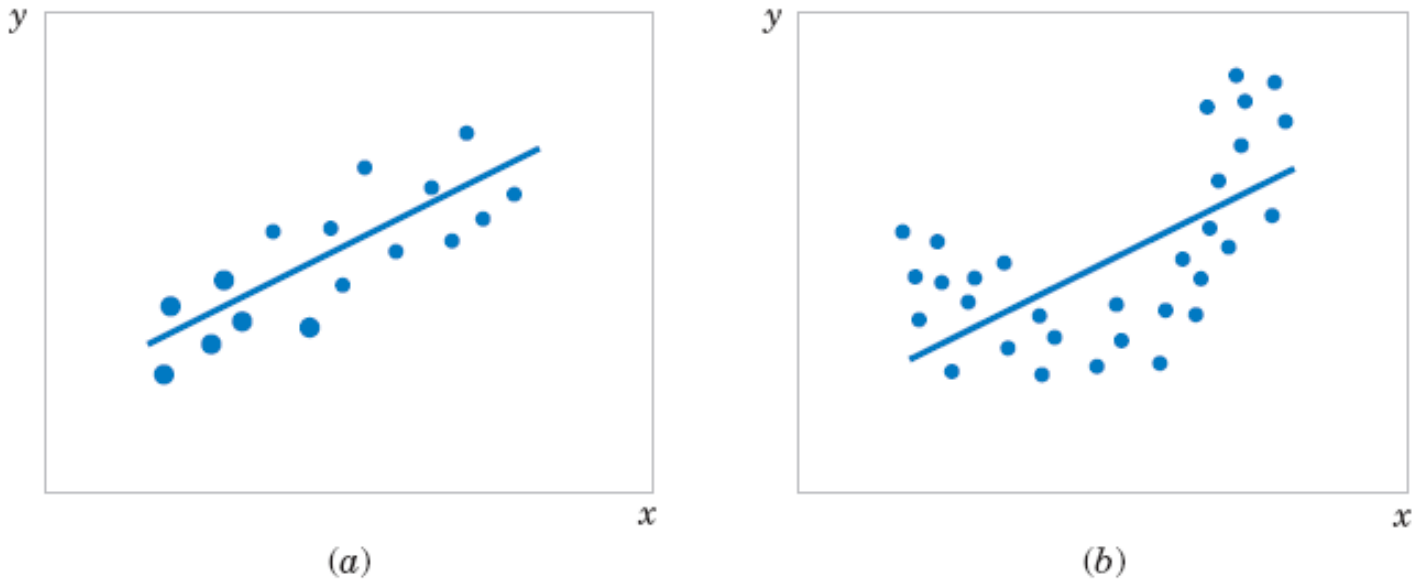


Figure 11-6 The hypothesis H_0: $\beta_1 = 0$ is rejected.

Figure 11-6 The **null hypothesis H_0: $\beta_1 = 0$ is rejected**.

11-4: Hypothesis Tests in Simple Linear Regression

11-4.1 Use of *Z*-tests for large n

An important special case of the hypotheses of Equation 11-18 is

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

These hypotheses relate to the **significance of regression**. *Failure* to reject H_0 is equivalent to concluding that there is no linear relationship between *X* and *Y*.

11-4: Hypothesis Tests in Simple Linear Regression

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

Choose α (e.g. $\alpha = 5\%$ for 95% confidence in rejecting H_0)

$$Z = \frac{\hat{\beta}_1 - 0}{\frac{\sigma_\epsilon}{\sigma_x} \cdot \frac{1}{\sqrt{n}}}$$

$$\sqrt{Var(\hat{\beta}_1)} = \frac{\sigma_\epsilon}{\sigma_x^2 \sqrt{n}}$$

for $\alpha = 5\%$

Reject H_0 if $|Z| > Z_{\alpha/2} = 1.96$

11-4: Hypothesis Tests in Simple Linear Regression

11-4.1 Use of *t*-tests for smaller n.

The number of degrees of freedom in **n-2**

One can always fit a straight line through two points so one needs n>=3

11-4: Hypothesis Tests in Simple Linear Regression

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

Choose α (e.g. $\alpha = 5\%$ for 95% confidence in rejecting H_0)

$$T = \frac{\hat{\beta}_1}{\frac{\hat{\sigma}_\epsilon}{\sigma_x} \cdot \frac{1}{\sqrt{n}}}$$

$t_{\alpha/2, n-2}$ is such $1 - \frac{\alpha}{2} = \text{tcdf}(t_{\alpha/2, n-2}, n-2)$

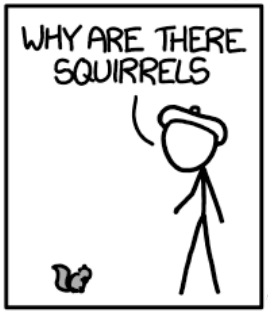
Reject H_0 if $|T| > t_{\alpha/2, n-2}$

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 DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE
WHY IS LYING GOOD

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 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 SEX SO IMPORTANT

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 AREN'T MY ARMS GROWING

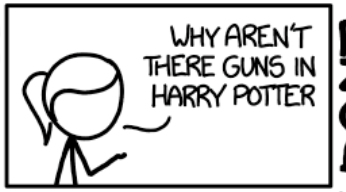
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

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 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 FEMALE MR MIMES

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 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 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 AREN'T THERE GUNS IN HARRY POTTER

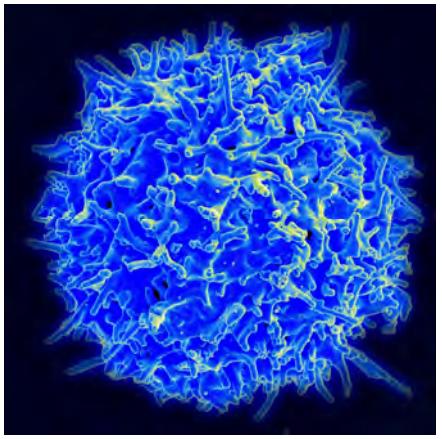


WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

Human T cell expression data

- The matrix contains 47 expression samples from Lukk et al, Nature Biotechnology 2010
- All samples are from T cells in different individuals
- Only the top 3000 genes with the largest variability were used
- The value is log2 of gene's expression level in a given sample as measured by the microarray technology

a T cell



A global map of human gene expression

Margus Lukk, Misha Kapushesky, Janne Nikkilä, Helen Parkinson, Angela Goncalves, Wolfgang Huber, Esko Ukkonen & Alvis Brazma

Affiliations | Corresponding author

Nature Biotechnology **28**, 322–324 (2010) | doi:10.1038/nbt0410-322

Although there is only one human genome sequence, different genes are expressed in many different cell types and tissues, as well as in different developmental stages or diseases. The structure of this 'expression space' is still largely unknown, as most transcriptomics experiments focus on sampling small regions. We have constructed a global gene expression map by integrating microarray data from 5,372 human samples representing 369 different cell and tissue types, disease states and cell lines. These have been compiled in an online resource (http://www.ebi.ac.uk/gxa/array/U133A) that allows the user to search for a gene of interest and

“Let’s Make a Deal” show with Monty Hall aired on NBC/ABC 1963-1986

WHEEL OF FORTUNE

Gene Expression "Wheel of Fortune"

- Each group gets a pair of genes that are known to be correlated.
- Each group also gets a random pair of genes selected by the "Wheel of Fortune". They may or may not be correlated
- Download (log-transformed) expression_table.mat
- Run command fitlm(x,y) on assigned and random pairs
- Record β_0, β_1, R^2, P-value of the slope β_1 and write them on the blackboard
- Validate Matlab result for R^2 using your own calculations
- Look up gene names (see gene_description in your workspace) and write down a brief description of biological functions of genes. Does their correlation make biological sense?

Correlated pairs

plausible biological connection based

on short description

g1=1994; g2=188; group 1

g1=2872; g2=1269; group 2

g1=1321; g2=10; group 3

g1= 886; g2=819; group 4

g1=2138; g2=1364; group 5

no obvious biological common function

g1=1+floor(rand.*3000); g2=1+floor(rand.*3000);
disp([g1, g2])

Confidence interval for population variance σ^2

- Up until now we were calculating the confidence interval on the **population average μ**
- What if one wants to put **confidence interval on population variance $\boldsymbol{\sigma^2}$**?
- We know an unbiased estimator of σ^2:

$$s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$$

- How to determine confidence interval?

$\vec{X} = (x_1, x_2 \dots x_n)$ $\quad x_i \to x_i - \bar{x}$

$y = |\vec{x}|^2 = \sum x_i^2 = (n-1)s^2$ $\quad \sum_{i=1}^{n} x_i = 0$

$$P(|\vec{x}|)\, d|\vec{x}| \sim \prod_{i=1}^{n-1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2}\right) dx_i$$

(left the last one since $x_n = -\sum_{i=1}^{n-1} x_i$)

Sphere area $\sim |\vec{x}|^{n-2}$

$(n-1)$-dim Space

$|\vec{x}| = \sqrt{y}$

$d|\vec{x}| = \frac{1}{\sqrt{y}}\, dy$

$\prod dx_i \sim |\vec{x}|^{n-2}\, d|\vec{x}|$

$$P(y)\, dy = y^{\frac{n-1}{2}-1} \exp\left(-\frac{y}{2}\right) dy$$

8-4 Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

$X=(n-1)S^2/\sigma^2$

We know n, S^2

want to estimate σ^2

$f(x,n) \sim x^{(n-1)/2-1}\exp(-x/2)$

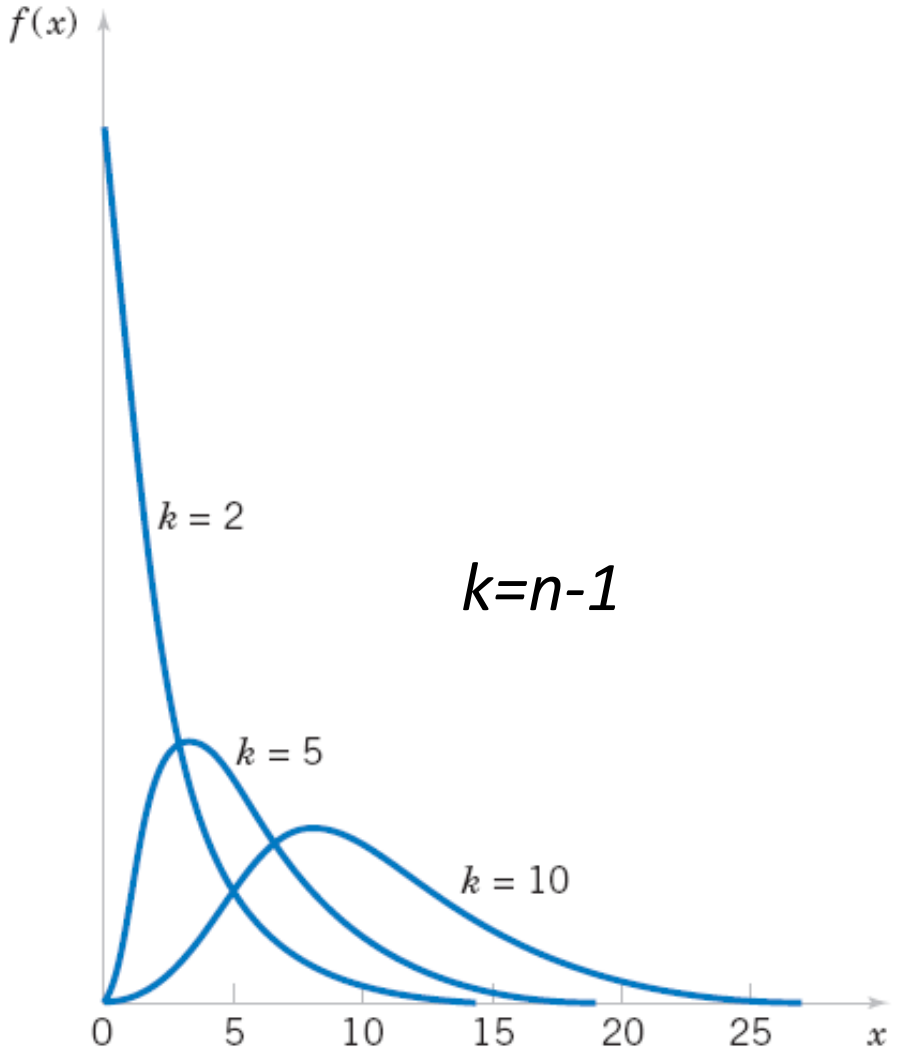
It is just Gamma PDF with $r=(n-1)/2$, and $\lambda=1/2$

Mean value:

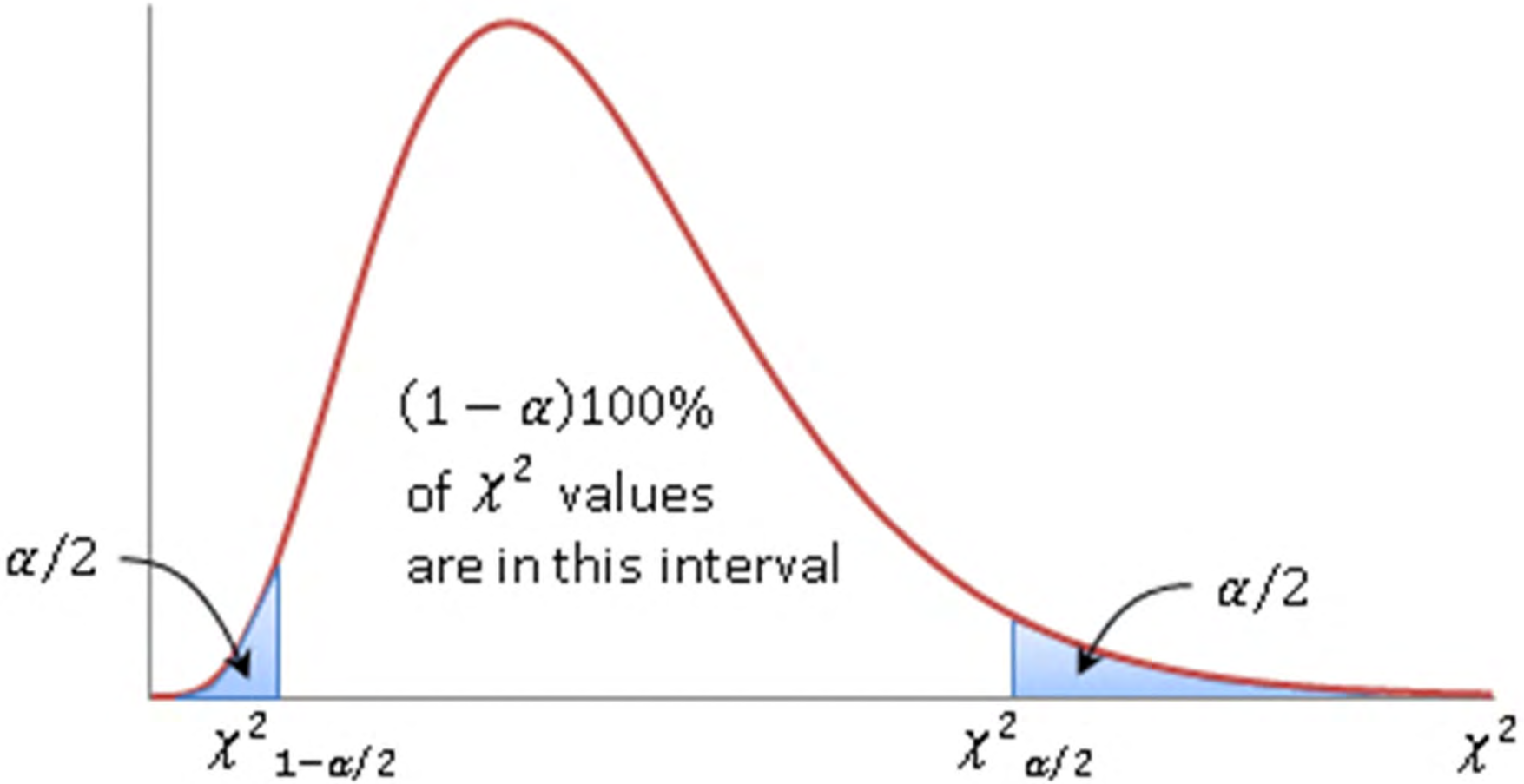
$n-1$

Standard deviation:

$$\sqrt{2(n-1)}$$



$k=n-1$



$$\chi^2_{1-\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\alpha/2}$$

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

Person’s chi-squared Goodness of fit test

Did you know that M&M's® Milk Chocolate Candies are supposed to come in the following percentages: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown?

http://www.scientificameriken.com/candy5.asp

"To our surprise M&Ms met our demand to review their procedures in determining candy ratios. It is, however, noted that the figures presented in their email differ from the information provided from their website (http://us.mms.com/us/about/products/milkchocolate/). An email was sent back informing them of this fact. To which M&Ms corrected themselves with one last email:

In response to your email regarding M&M'S CHOCOLATE CANDIES

Thank you for your email.
On average, our new mix of colors for M&M'S® Chocolate Candies is:

M&M'S® Milk Chocolate: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown.

M&M'S® Peanut: 23% blue, 23% orange, 15% green, 15% yellow, 12% red, 12% brown.

M&M'S® Kids MINIS®: 25% blue, 25% orange, 12% green, 13% yellow, 12% red, 13% brown.

M&M'S® Crispy: 17% blue, 16% orange, 16% green, 17% yellow, 17% red, 17% brown.

M&M'S® Peanut Butter and Almond: 20% blue, 20% orange, 20% green, 20% yellow, 10% red, 10% brown.

Have a great day!

Your Friends at Masterfoods USA
A Division of Mars, Incorporated

How to accept or reject the null hypothesis that these probabilities are correct from a finite sample?

Pearson chi^2 Goodness of Fit Test

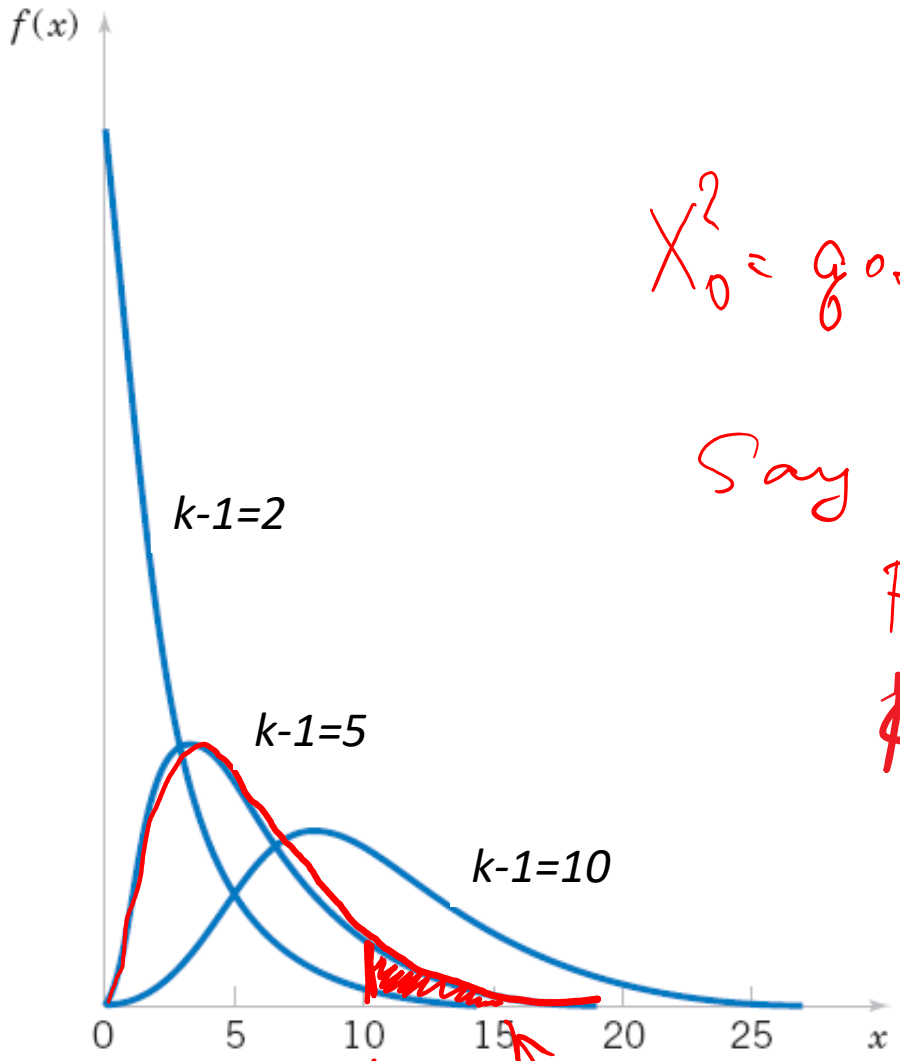
- Assume there is a sample of size n from a population with k classes (e.g. 6 M&M colors)
- Null hypothesis H_0: class i has frequency f_i in the population
- Alternative hypothesis H_1: some population frequencies are inconsistent with f_i
- Let O_i be the observed number of sample elements in the ith class and $E_i = n f_i$ be the expected number of sample elements in the ith class.
- Group any bin with E_i <3 with
  a) if numerical value of i is important, group it with its neighbor (k=i-1 or k=i+1) which has the smallest E_k until E_{group} >=3;
  b) If numerical value of i is irrelevant, group together all E_i <3 bins until E_{group} >=3
- The test statistic is

$$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad (9\text{-}47)$$

P-value is calculated based on the chi-square distribution with k-1 degrees of freedom:

P-value = Prob(H_0 is correct) =1-CDF_chi-squared(X_0^2, k-1)

chi^2 Goodness of Fit Test is a *one-sided* hypothesis



$$X_0^2 = \text{gof} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Say $X_0^2 = 10$

For M&M

$k = 6 \rightarrow k-1 = 5$

X_0^2

p-value that null hypothesis is correct

M&M group exercise

- DO NOT EAT CANDY BEFORE COUNTING IS FINISHED! THEN, **PLEASE, DO**.
- We will be testing three null hypotheses one after another:
  - M&M official data: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown
  - Website (fan collected) data from http://joshmadison.com/2007/12/02/mms-color-distribution-analysis: 18.36% blue, 20.76% orange, 18.44% green, 14.08% yellow, 14.20% red, 14.16% brown
  - Uniform distribution: 1/6~16.67% of each candy color
- You will estimate P-values for each one of these null hypotheses
- Hints: O_i – is the observed # of candies of color i; calculate the expected # E_i=(# candies in your sample)*f_i

Use **1-chi2cdf(X0squared, 5)** for P-value

$$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Statistical tests of independence

- Did I mix M&M candy well?

	blue	orange	green	yellow	red	brown
group 1	55	33	39	61	69	32
group 2	59	34	31	84	52	28
group 3	27	15	46	6	40	4
group 4	33	28	57	22	34	20

How to **test the hypothesis** if multiple sample are drawn from the same population?

- Table: samples (Student groups) – rows, classes (M&M colors) – columns
- Test if color fractions are **independent** from group
- ***P(Group 1 and Color = green) = P(Group 1)*P(Color green)***
- Compute for all groups/colors 6*4=24 in our case

$E_{green}(group\ 1) = n_{tot} * (group1/n_{tot}) * (green/n_{tot})$

- $\chi^2 = \sum_{groups\ \&\ colors}^{n_{tot}} \frac{(O_{color}\ (group) - E_{color}(group))^2}{E_{color}(group)}$
- # degrees of freedom=(colors-1)*(groups-1)

- Was the M&M box from Costco well mixed?
  Let's compare the first two groups' data

	blue	orange	green	yellow	red	brown
group 1	56	62	36	36	37	35
group 2	59	67	29	39	32	25
group 3	58	63	29	28	33	24
group 4	58	60	36	22	37	36

- Using $\chi^2 = \sum_{groups\ \&\ colors} \frac{\left(O_{color}\,(group) - E_{color}(group)\right)^2}{E_{color}(group)}$

with # degrees of freedom (colors-1)*(groups-1)

Find P-value of null hypothesis H_0 that
samples are independent from each other

Batch effect

Does color composition vary between Costco and Schnucks

- Costco: 114 67 70 145 121 60
- Schnucks: 60 43 103 28 74 24
- Test if they are significantly different from each other:
- Same test expect ngroups=2; ncolors=6;
- Results:

  Goodness of Fit =73.4774

  P-value = 1.9318e-14
- Batch effect is highly statistically significant!

Goodness of fit with a PDF defined by m parameters

- As before: k classes (e.g. M&M colors)
- Use parameter estimators to find the best parameters for the fit
  - Method of moments
  - MLE: method of maximum likelihood
- Use chi-squared distribution with k-1-m degrees of freedom
- As before: if $E_i < 3$, group it together with another group and reduce k by 1

$$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad (9\text{-}47)$$

9-7 Testing for Goodness of Fit

Example 9-12

EXAMPLE 9-12 Printed Circuit Board Defects

Poisson Distribution

The number of defects in printed circuit boards is hypothesized to follow a Poisson distribution. A random sample of $n = 60$ printed boards has been collected, and the following number of defects observed.

Number of Defects	Observed Frequency
0	32
1	15
2	9
3	4

9-7 Testing for Goodness of Fit

Example 9-12

The mean of the assumed Poisson distribution in this example is unknown and must be estimated from the sample data. The estimate of the mean number of defects per board is the sample average, that is, $(32 \cdot 0 + 15 \cdot 1 + 9 \cdot 2 + 4 \cdot 3)/60 = 0.75$. From the Poisson distribution with parameter 0.75, we may compute p_i, the theoretical, hypothesized probability associated with the ith class interval. Since each class interval corresponds to a particular number of defects, we may find the p_i as follows:

$$p_1 = P(X = 0) = \frac{e^{-0.75}(0.75)^0}{0!} = 0.472$$

$$p_2 = P(X = 1) = \frac{e^{-0.75}(0.75)^1}{1!} = 0.354$$

$$p_3 = P(X = 2) = \frac{e^{-0.75}(0.75)^2}{2!} = 0.133$$

$$p_4 = P(X \geq 3) = 1 - (p_1 + p_2 + p_3) = 0.041$$

9-7 Testing for Goodness of Fit

Example 9-12

The expected frequencies are computed by multiplying the sample size $n = 60$ times the probabilities p_i. That is, $E_i = np_i$. The expected frequencies follow:

Number of Defects	Probability	Expected Frequency
0	0.472	28.32
1	0.354	21.24
2	0.133	7.98
3 (or more)	0.041	2.46

9-7 Testing for Goodness of Fit

Example 9-12

Since the expected frequency in the last cell is less than 3, we combine the last two cells:

Number of Defects	Observed Frequency	Expected Frequency
0	32	28.32
1	15	21.24
2 (or more)	13	10.44

The chi-square test statistic in Equation 9-47 will have $k - p - 1 = 3 - 1 - 1 = 1$ degree of freedom, because the mean of the Poisson distribution was estimated from the data.

9-7 Testing for Goodness of Fit

Example 9-12

The seven-step hypothesis-testing procedure may now be applied, using $\alpha = 0.05$, as follows:

1. **Parameter of interest:** The variable of interest is the form of the distribution of defects in printed circuit boards.
2. **Null hypothesis:** H_0: The form of the distribution of defects is Poisson.
3. **Alternative hypothesis:** H_1: The form of the distribution of defects is not Poisson.
4. **Test statistic:** The test statistic is

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(o_i - E_i)^2}{E_i}$$

9-7 Testing for Goodness of Fit

Example 9-12

5. **Reject H_0 if:** Reject H_0 if the P-value is less than 0.05.

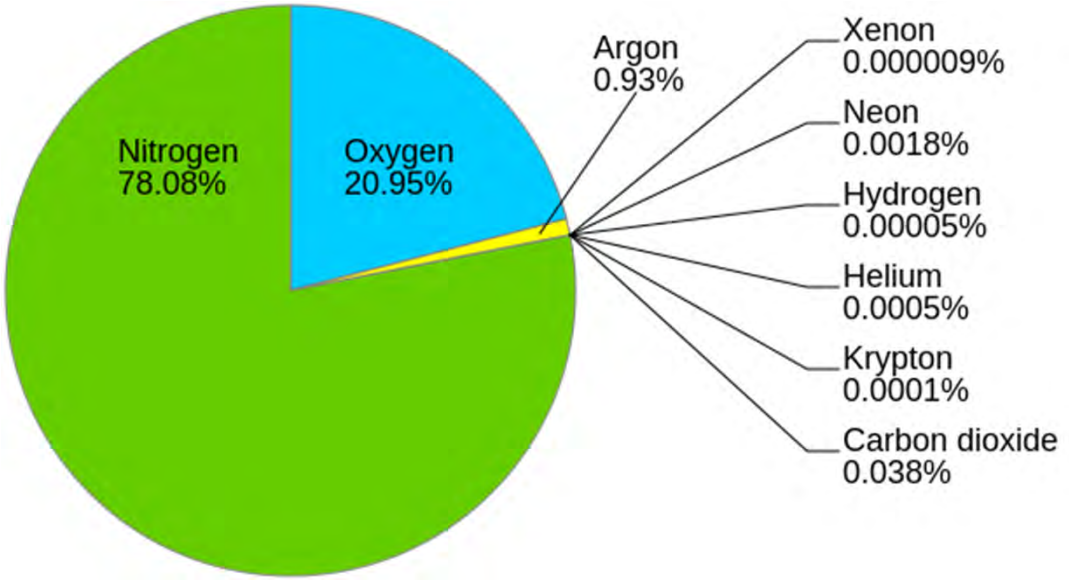
6. **Computations:**

$$\chi_0^2 = \frac{(32 - 28.32)^2}{28.32} + \frac{(15 - 21.24)^2}{21.24} + \frac{(13 - 10.44)^2}{10.44} = 2.94$$

7. **Conclusions:** We find from Appendix Table III that $\chi^2_{0.10,1} = 2.71$ and $\chi^2_{0.05,1} = 3.84$. Because $\chi_0^2 = 2.94$ lies between these values, we conclude that the P-value is between 0.05 and 0.10. Therefore, since the P-value exceeds 0.05 we are unable to reject the null hypothesis that the distribution of defects in printed circuit boards is Poisson. The exact P-value computed from Minitab is 0.0864.

Reminder

Two variable samples



- Oxygen can be distilled from the air
- Hydrocarbons need to be filtered out or the whole thing would go kaboom!!!
- When more hydrocarbons were removed, the remaining oxygen stays cleaner
- Except we don't know how dirty was the air to begin with

Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

Linear regression

The **simple linear regression model** is given by

$$Y = \beta_0 + \beta_1 X + \varepsilon = \hat{Y} + \varepsilon$$

ε is the **random error term**

slope β_1 and intercept β_0 of the line are called **regression coefficients**

Note: Y , $\hat{Y}$, X and ε are random variables

The minimal assumption: $E(\varepsilon \mid x)=0$ →

$E(Y \mid x) = \beta_0 + \beta_1 x + E(\varepsilon \mid x) = \beta_0 + \beta_1 x$

$$Y = \beta_0 + \beta_1 X + \epsilon$$

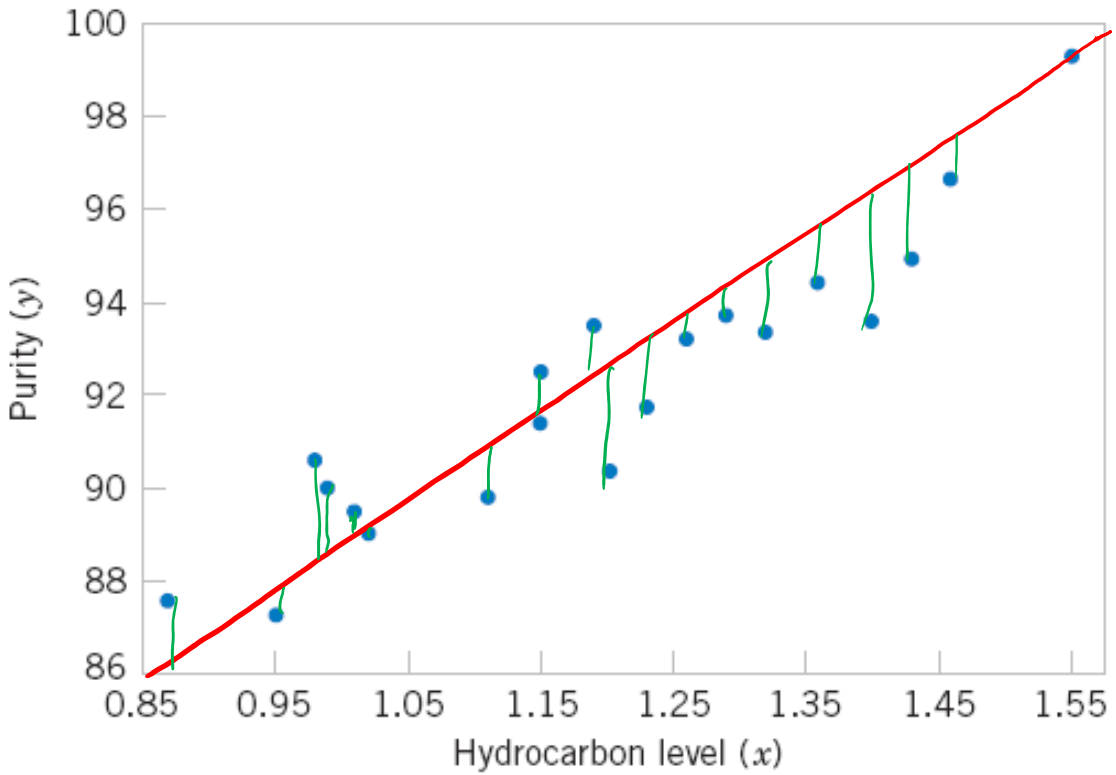


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

$$Y = 75 + 15 \cdot X + \epsilon$$

$$Y = \beta_0 + \beta_1 X + \epsilon \; ; \quad E(\epsilon \mid x) = 0 \;\; \forall x$$

How does one find β_0 & β_1?

$$Cov(Y, X) = Cov((\beta_0 + \beta_1 X + \epsilon), X) =$$

$$= Cov(\beta_0, X) + \beta_1 Cov(X, X) + Cov(\epsilon, X)$$

$Cov(\beta_0, X) = 0$ since β_0 is constant

$Cov(X, X) = E(X^2) - E(X)^2 = Var(X)$

$Cov(\epsilon, X) = E(\epsilon \cdot X) - E(\epsilon) \cdot E(X) =$

$= E(\epsilon \cdot X) = \sum_{\text{all } x} x \cdot E(\epsilon \mid x) = 0$

Thus $\boxed{\beta_1 = \frac{Cov(X, Y)}{Var(X)}}$ $\boxed{\beta_0 = E(Y) - \beta_1 E(X)}$

Method of least squares

- The **method of least squares** is used to estimate the parameters, β_0 and β_1 by minimizing the sum of the squares of the vertical deviations in Figure 11-3.

Figure 11-3 Deviations of the data from the estimated regression model.

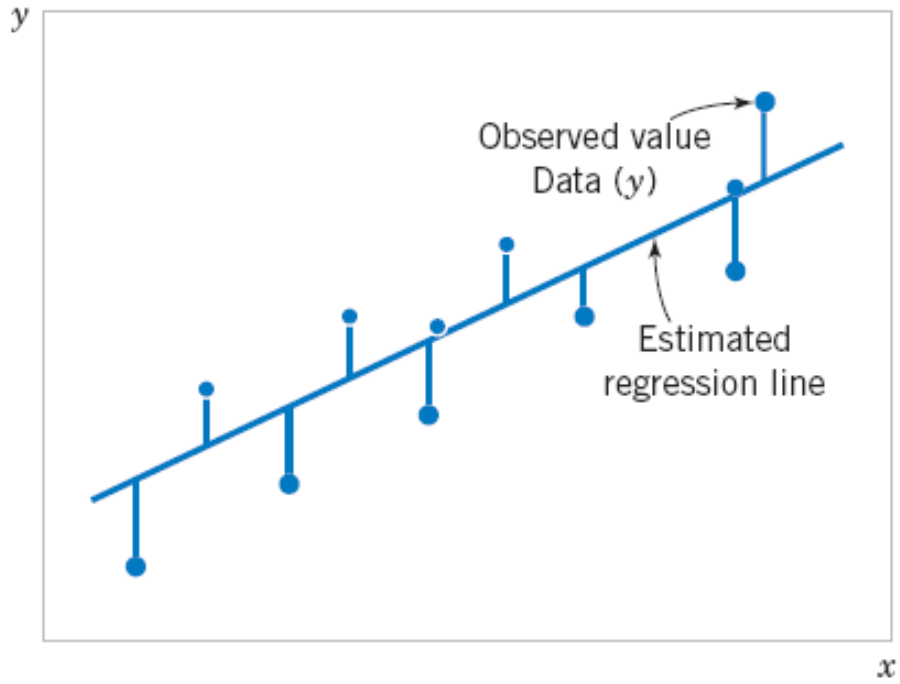


Figure 11-3 Deviations of the data from the estimated regression model.

Traditional notation

Definition

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{11-7}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{S_{xy}}{S_{xx}} \tag{11-8}$$

where $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ and $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$.

Connection to Cov(X,Y)/Var(X) result

Definition

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{11-7}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n^2}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n^2}} = \frac{Cov(X,Y)}{Var(X)} \tag{11-8}$$

where $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ and $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$.

Different types of y

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{11-7}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n^2}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n^2}} = \frac{Cov(X,Y)}{Var(X)} \tag{11-8}$$

where $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ and $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$.

$\bar{y} = \sum y_i / n$

$\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0$

$\varepsilon_i = y_i - \hat{y}_i$

The analysis of variance identity is

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (11\text{-}24)$$

Symbolically,

$$SS_T = SS_R + SS_E \qquad (11\text{-}25)$$

11-7: Adequacy of the Regression Model

11-7.2 Coefficient of Determination (R^2)

VERY COMMONLY USED

- The quantity

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

is called the **coefficient of determination** and is often used to judge the adequacy of a regression model.

- $0 \le R^2 \le 1$;
- We often refer (loosely) to R^2 as the amount of variability in the data explained or accounted for by the regression model.

11-2: Simple Linear Regression

Estimating σ_ε^2

An **unbiased estimator** of σ_ε^2 is

$$\hat{\sigma}_\varepsilon^2 = \frac{SS_E}{n-2} \qquad (11\text{-}13)$$

where SS_E can be easily computed using

$$SS_E = SS_T - \hat{\beta}_1 S_{xy} \qquad (11\text{-}14)$$

Multiple Linear Regression

(Chapters 12-13 in
Montgomery, Runger)

12-1: Multiple Linear Regression Model

12-1.1 Introduction

- Many applications of regression analysis involve situations in which there are more than one regressor variable X_k used to predict Y.

- A regression model then is called a **multiple regression model**.

Multiple Linear Regression Model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots \beta_k x_k + \varepsilon$$

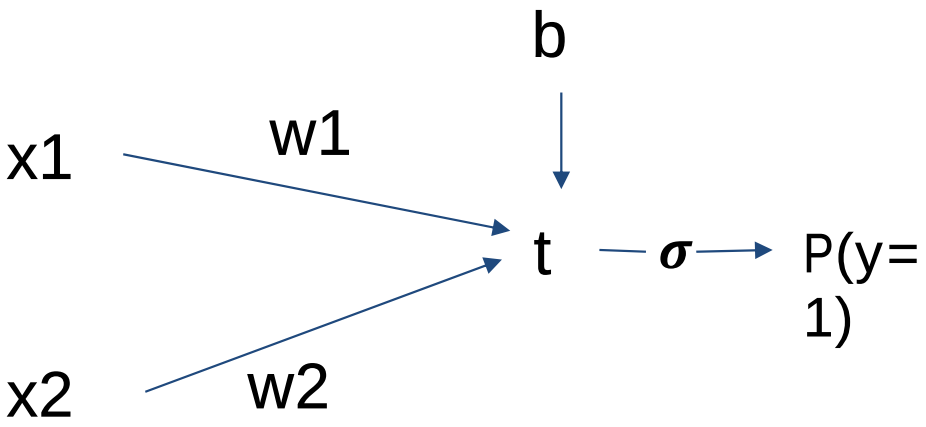
One can also use powers and products of other variables or even non-linear functions like $\exp(x_i)$ or $\log(x_i)$ instead of $x_3, \ldots x_k$.

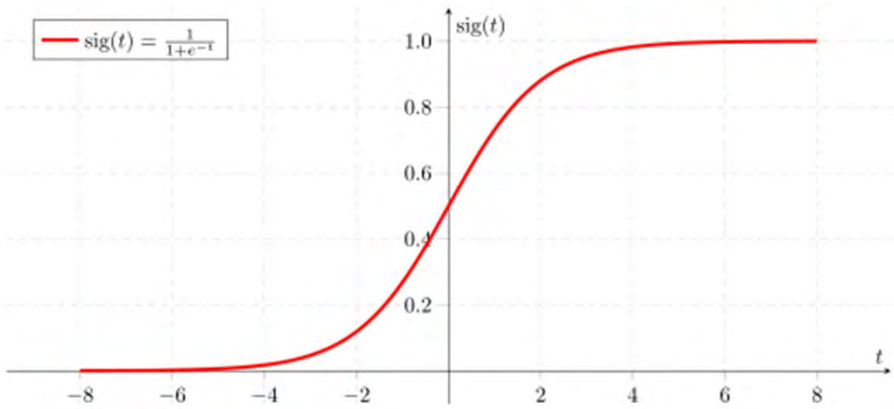
Example: the general two-variable quadratic regression has 6 constants:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1)^2 + \beta_4 (x_2)^2 + \beta_5 (x_1 x_2) + \varepsilon$$

Logistic Regression

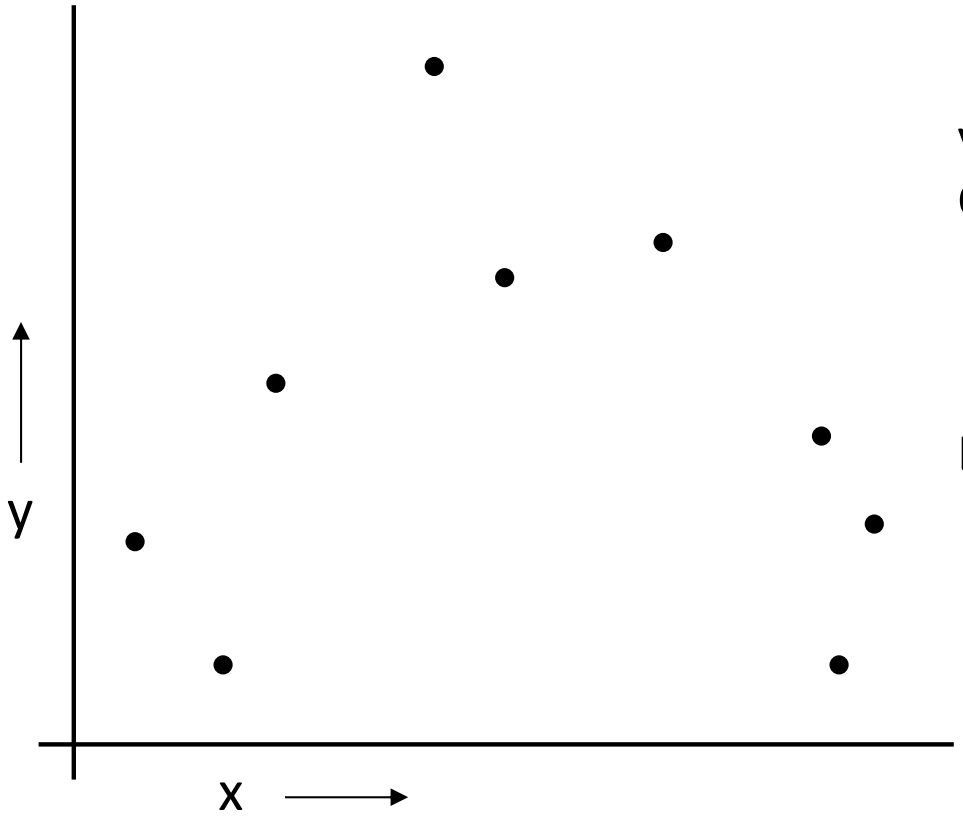
$$P(y=1) = \sigma(x1*w1 + x2*w2 + b)$$





How to know where to stop adding new variables or powers of old variables?

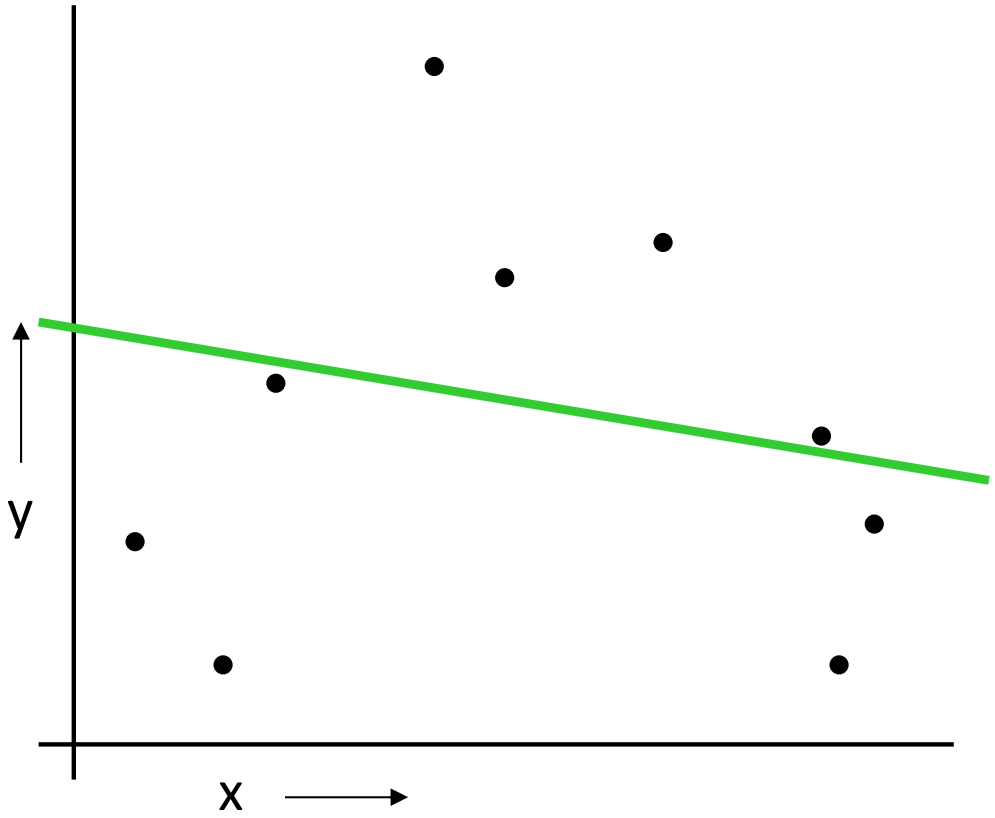
A Regression Problem



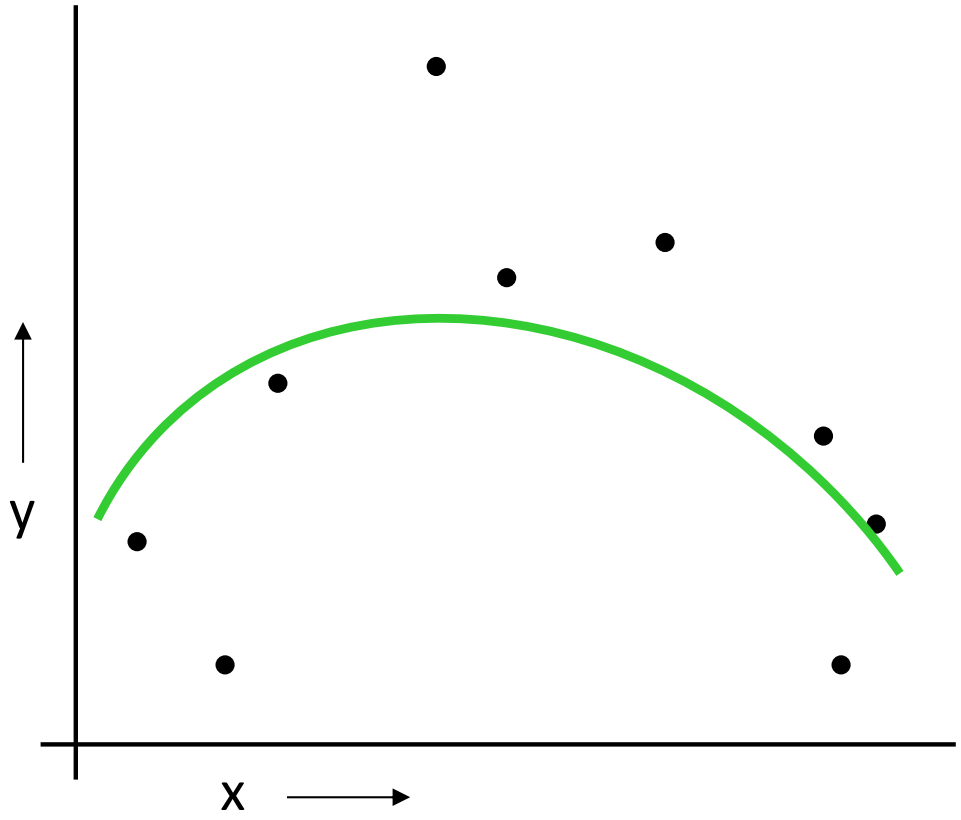
y = f(x) + noise
Can we learn f from this data?

Let's consider three methods…

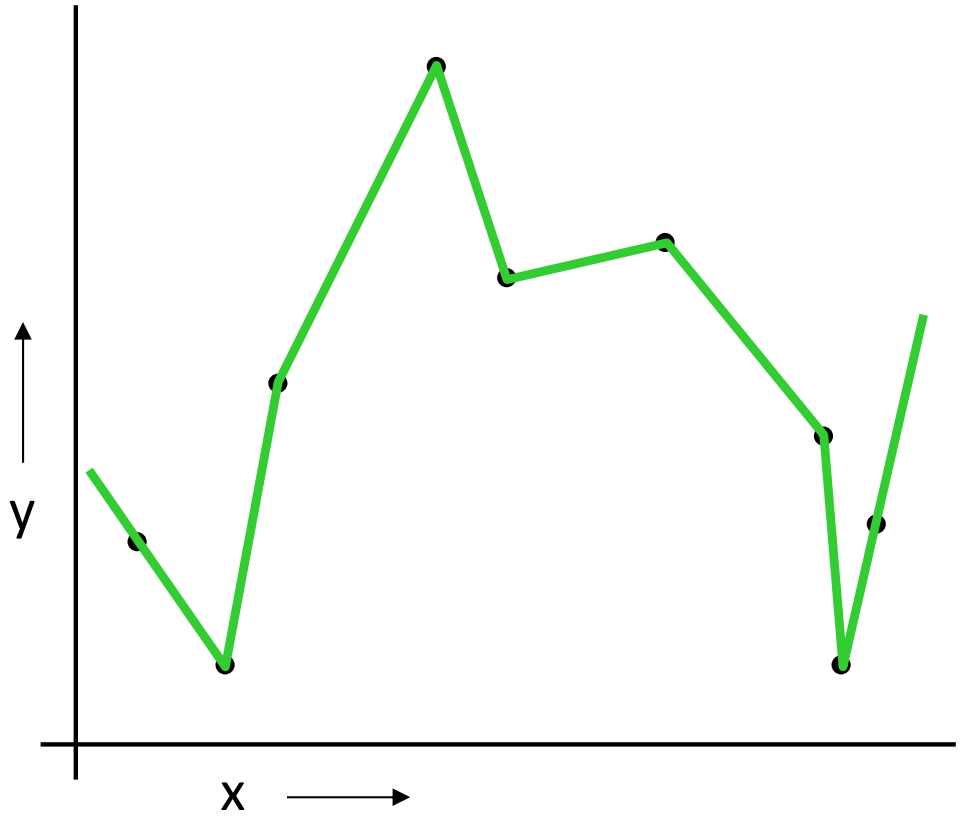
Linear Regression



Quadratic Regression

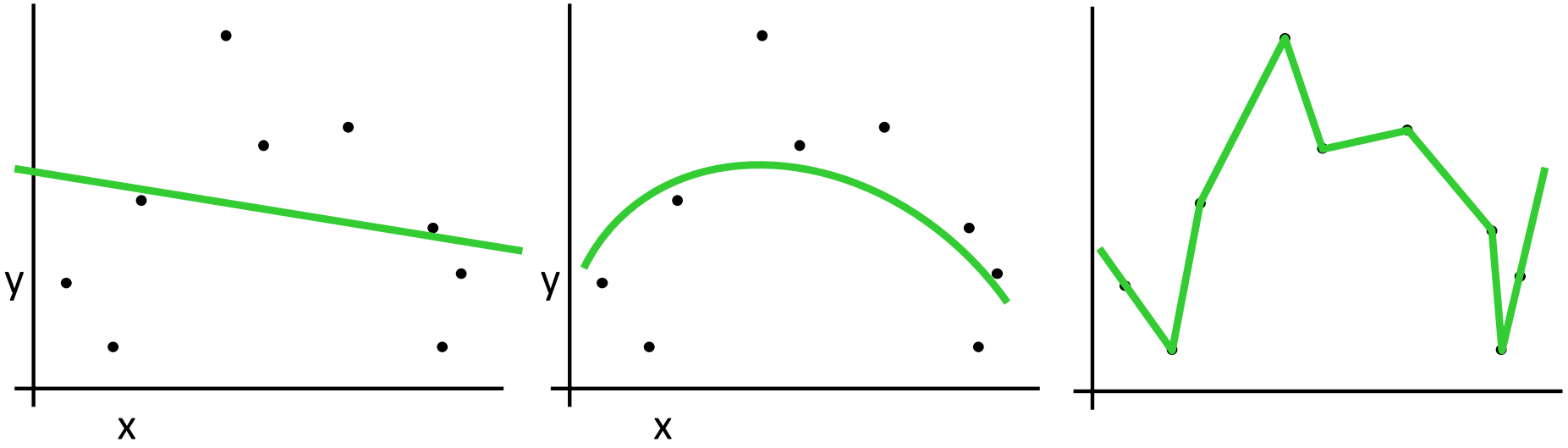


Join-the-dots



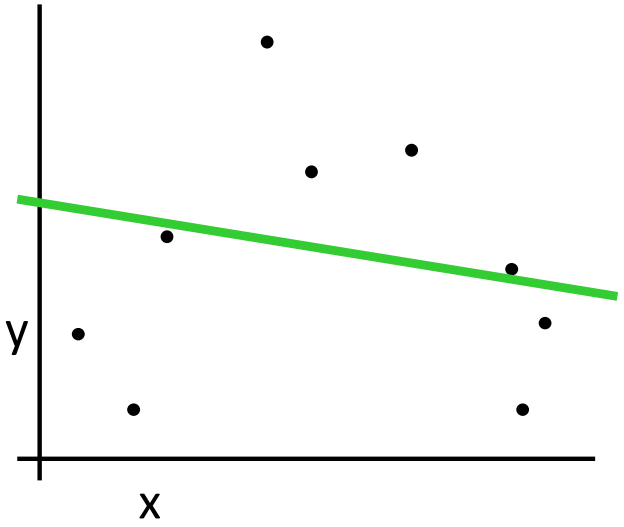
Also known as piecewise linear nonparametric regression if that makes you feel better

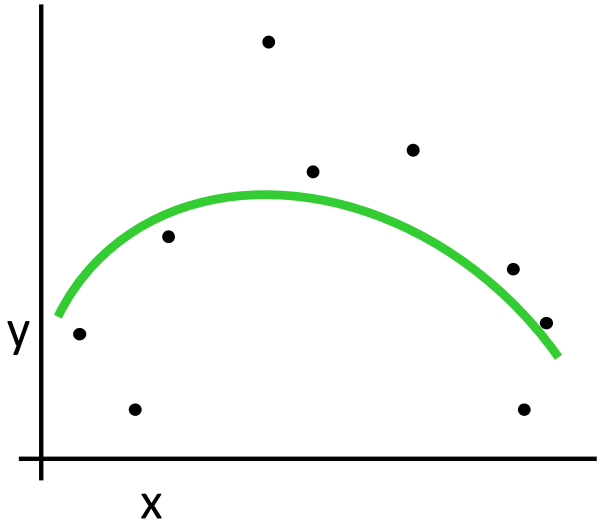
Which is best?

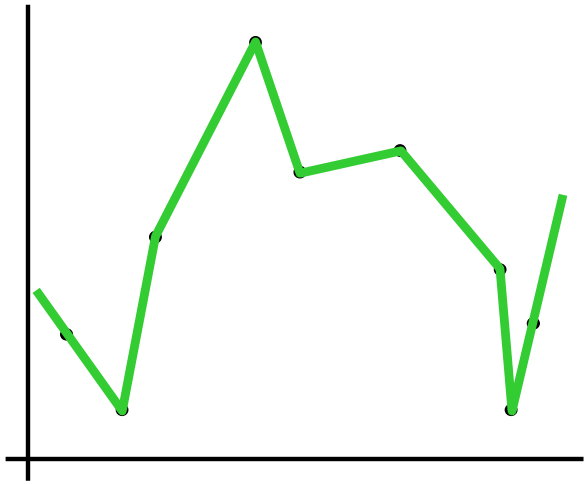


Why not choose the method with the best fit to the data?

What do we really want?



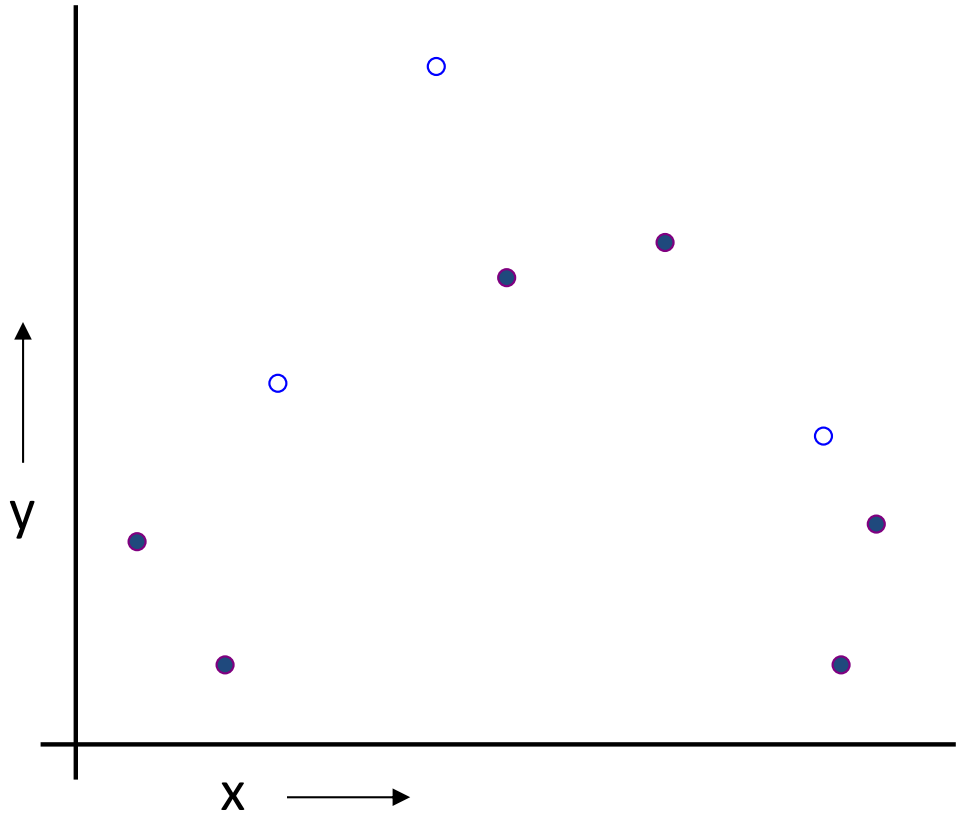




Why not choose the method with the best fit to the data?

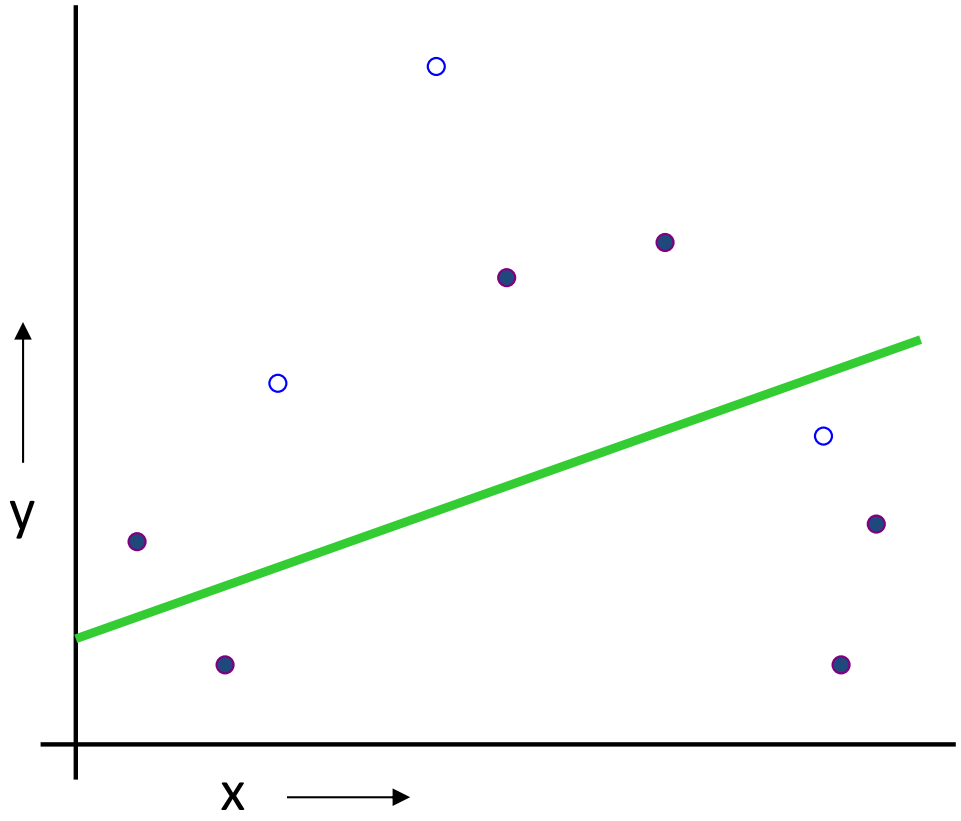
"How well are you going to predict future data drawn from the same distribution?"

The test set method



1. Randomly choose 30% of the data to be in a test set
2. The remainder is a training set

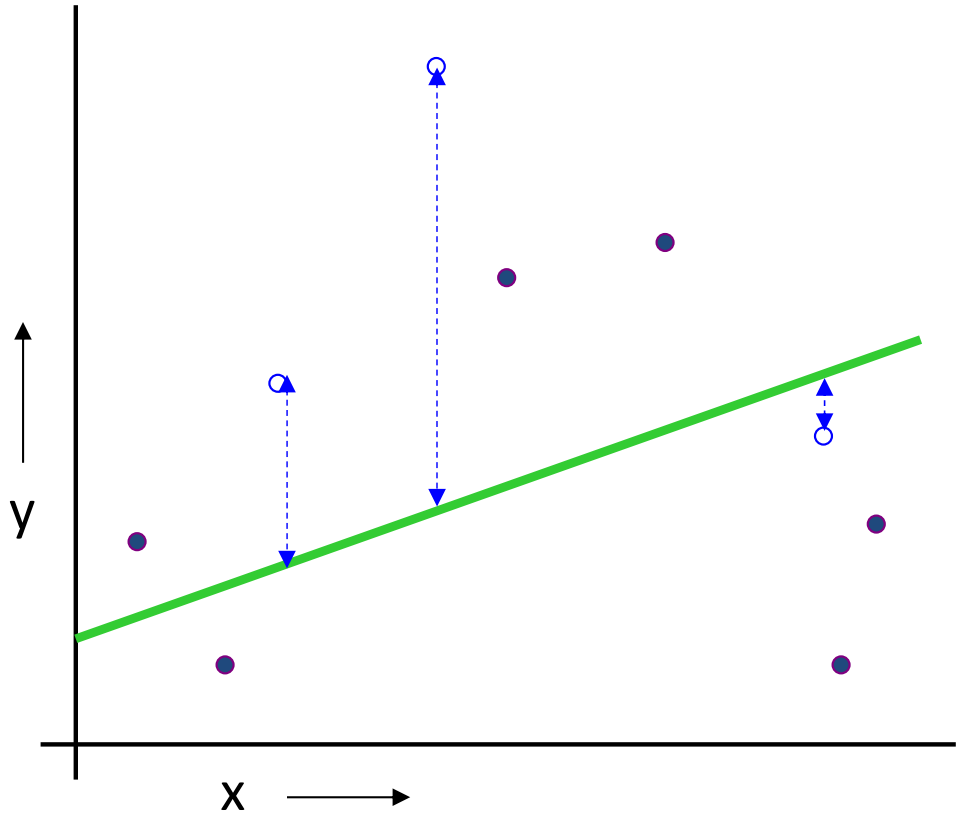
The test set method



(Linear regression example)

1. Randomly choose 30% of the data to be in a test set
2. The remainder is a training set
3. Perform your regression on the training set

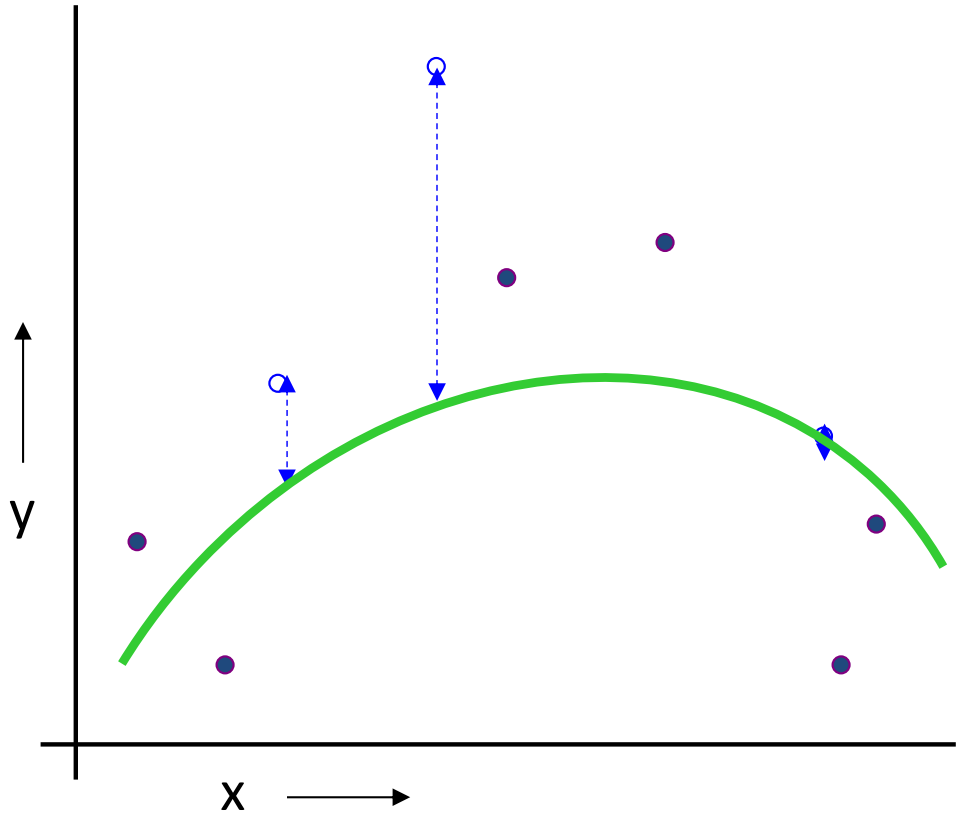
The test set method



(Linear regression example)
Mean Squared Error = 2.4

1. Randomly choose 30% of the data to be in a test set
2. The remainder is a training set
3. Perform your regression on the training set
4. Estimate your future performance with the test set

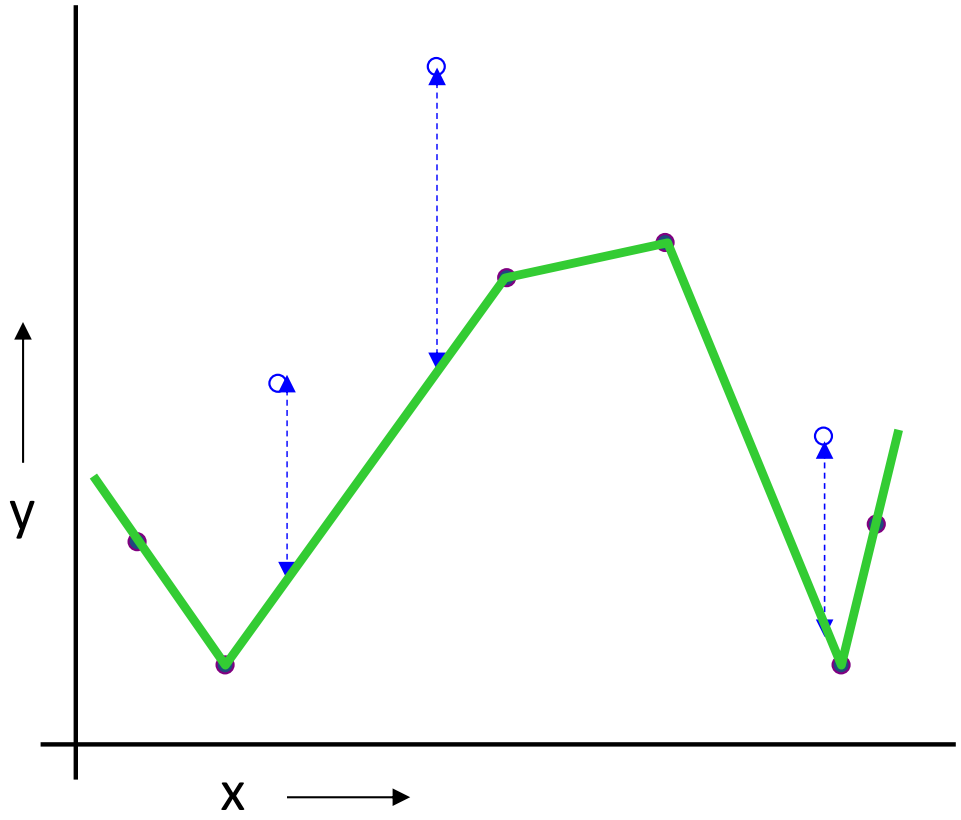
The test set method



(Quadratic regression example)
Mean Squared Error = 0.9

1. Randomly choose 30% of the data to be in a test set
2. The remainder is a training set
3. Perform your regression on the training set
4. Estimate your future performance with the test set

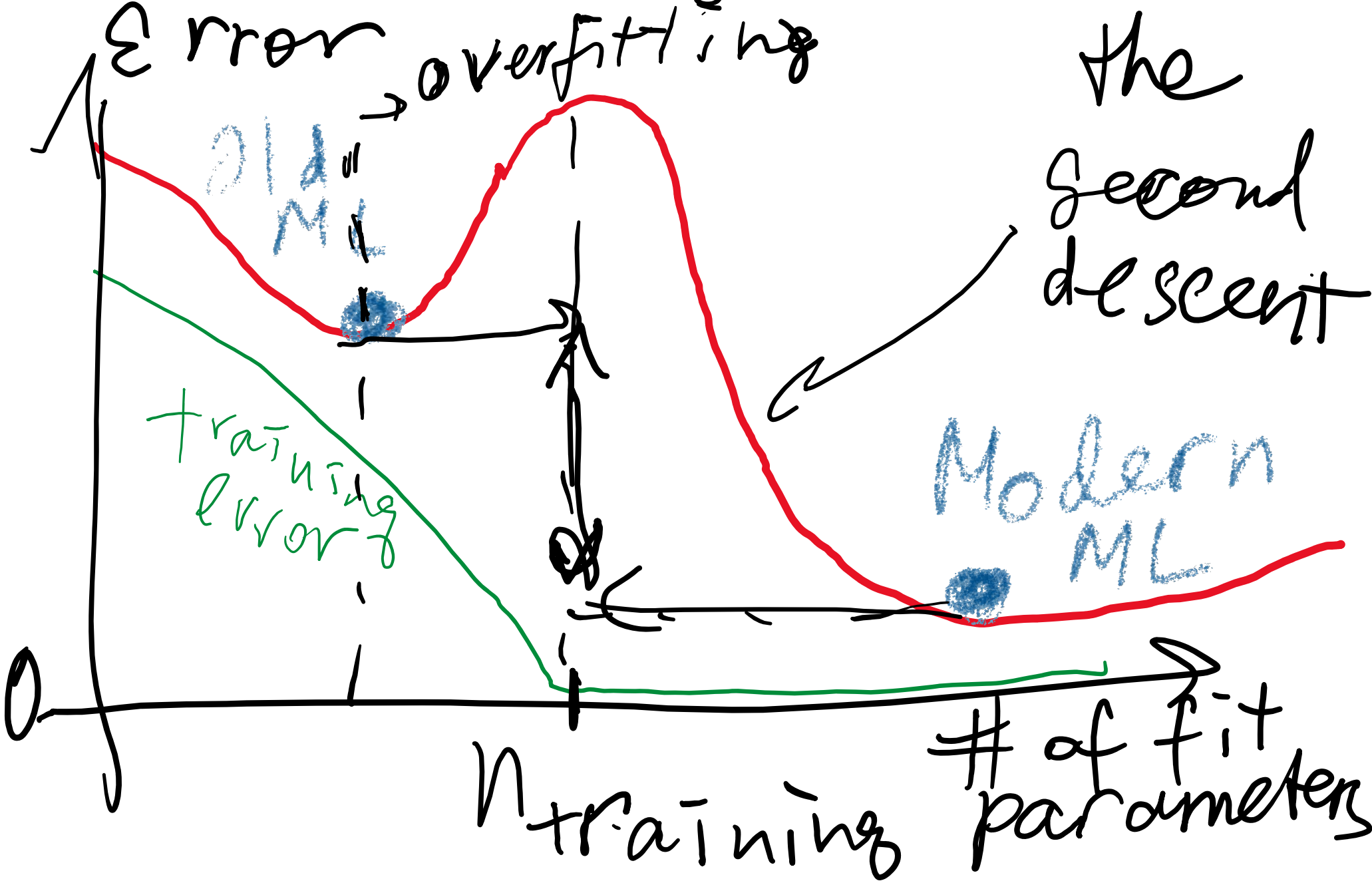
The test set method



(Join the dots example)
Mean Squared Error = 2.2

1. Randomly choose 30% of the data to be in a test set
2. The remainder is a training set
3. Perform your regression on the training set
4. Estimate your future performance with the test set

Double descend- the main reason modern Machine Learning works so well



12-1: Multiple Linear Regression Model

12-1.3 Matrix Approach to Multiple Linear Regression

Suppose the model relating the regressors to the response is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

In matrix notation this model can be written as

$$\mathbf{y = X\beta + \varepsilon} \qquad (12\text{-}6)$$

12-1: Multiple Linear Regression Model

12-1.3 Matrix Approach to Multiple Linear Regression

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \text{ and } \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

12-1.3 Matrix Approach to Multiple Linear Regression

We wish to find the vector $\hat{\beta}$ that minimizes the sum of squares of error terms:

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (y - X\beta)'(y - X\beta)$$

$$0 = \frac{\partial L}{2\partial \beta} = -X'(y - X\beta) = -X'y + (X'X)\beta$$

The resulting least squares estimate is

$$\hat{\beta} = (X'X)^{-1} X'y \qquad (12\text{-}7)$$

Analog of $\frac{1}{\text{Var}(X)}$

Analog of $\text{Cov}(X,Y)$

Multiple Linear Regression Model

$$\hat{\beta} = (X'X)^{-1} X'y$$

H is an idempotent matrix

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y,$$

$$\hat{y} = Hy, \quad \text{and} \quad e = (I - H)y.$$

$H = H^1$; $H^2 = X(X'X)^{-1}X'X(X'X)^{-1}X' = X(X'X)^{-1}X' = H$

($X'X$ cancels to I)

Vectors $\hat{y}$ & e are orthogonal since

$\hat{y}'e = y'H(I-H)y = 0$ since

$$H(I - H) = H - H^2 = H - H = 0.$$

12-1: Multiple Linear Regression Models

12-1.4 Properties of the Least Squares Estimators

Unbiased estimators:

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}] \\ &= \boldsymbol{\beta} \end{aligned}$$

Covariance Matrix of Estimators:

$$\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} C_{00} & C_{01} & C_{02} \\ C_{10} & C_{11} & C_{12} \\ C_{20} & C_{21} & C_{22} \end{bmatrix}$$

12-1: Multiple Linear Regression Models

12-1.4 Properties of the Least Squares Estimators

Individual variances and covariances:

$$V(\hat{\beta}_j) = \sigma^2 C_{jj}, \qquad j = 0, 1, 2$$

$$\text{cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}, \qquad i \neq j$$

In general,

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2\mathbf{C}$$

12-1: Multiple Linear Regression Models

Estimating error variance σ_ε^2

An unbiased estimator of error variance σ_ε^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SS_E}{n-p} \qquad (12\text{-}16)$$

Here p=k+1 for k-variable multiple linear regression

R^2 and Adjusted R^2

The **coefficient of multiple determination R^2**

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

The **adjusted R^2** is

$$R^2_{\text{adj}} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} \approx 1 - \frac{\hat{\sigma}^2}{S_y^2} \qquad (12\text{-}23)$$

- The adjusted R^2 statistic penalizes adding terms to the MLR model.
- It can help guard against overfitting (including regressors that are not really useful)

How to know where to stop adding variables?

- Adding new variables x_i to MLR
  watch the adjusted R^2

- Once the adjusted R^2
  no longer increases = stop.
  Now you did the best you can.

Matlab exercise

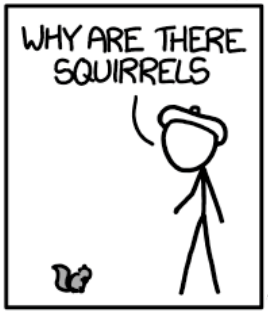
- Every group works with
  g0=2907; g1=1527; g2=2629; g3=2881;
  g4=1144; g5=1066;
- Compute Multiple Linear Regression (MLR):
  where
  y=exp_t (g0); x1= exp_t (g1); x2= exp_t (g2);
- How much better the MLR did compared to the Single Linear Regression (SLR)?
- Continue increasing the number of genes in x until R_adj starts to decrease

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 DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE
WHY IS LYING GOOD

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 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 SEX SO IMPORTANT

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 AREN'T MY ARMS GROWING

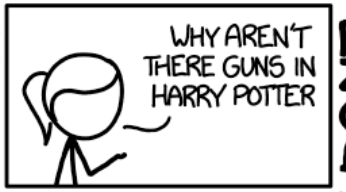
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

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 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 FEMALE MR MIMES

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 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 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 AREN'T THERE GUNS IN HARRY POTTER



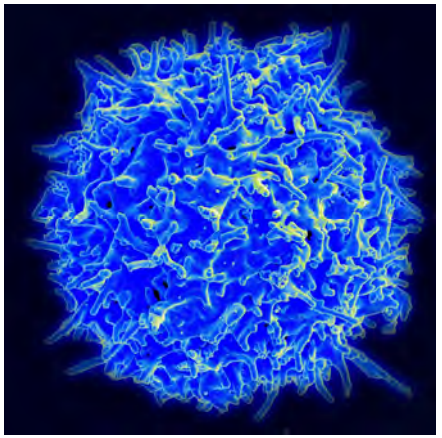
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

Clustering analysis of gene expression data

Chapter 11 in
Jonathan Pevsner,
Bioinformatics and Functional Genomics,
3rd edition
(Chapter 9 in 2nd edition)

Human T cell expression data

- The matrix contains 47 expression samples from Lukk et al, Nature Biotechnology 2010
- All samples are from T cells in different individuals
- Only the top 3000 genes with the largest variability were used
- The value is log2 of gene's expression level in a given sample as measured by the microarray technology



A global map of human gene expression

Margus Lukk, Misha Kapushesky, Janne Nikkilä, Helen Parkinson, Angela Goncalves, Wolfgang Huber, Esko Ukkonen & Alvis Brazma

Affiliations | Corresponding author

Nature Biotechnology **28**, 322–324 (2010) | doi:10.1038/nbt0410-322

Although there is only one human genome sequence, different genes are expressed in many different cell types and tissues, as well as in different developmental stages or diseases. The structure of this 'expression space' is still largely unknown, as most transcriptomics experiments focus on sampling small regions. We have constructed a global gene expression map by integrating microarray data from 5,372 human samples representing 369 different cell and tissue types, disease states and cell lines. These have been compiled in an online resource (http://www.ebi.ac.uk/gxa/array/U133A) that allows the user to search for a gene of interest and

WHEEL OF FORTUNE

Correlated pairs

plausible biological connection based

on short description

g1=1994; g2=188; group 1

g1=2872; g2=1269; group 2

g1=1321; g2=10; group 3

g1= 886; g2=819; group 4

g1=2138; g2=1364; group 5

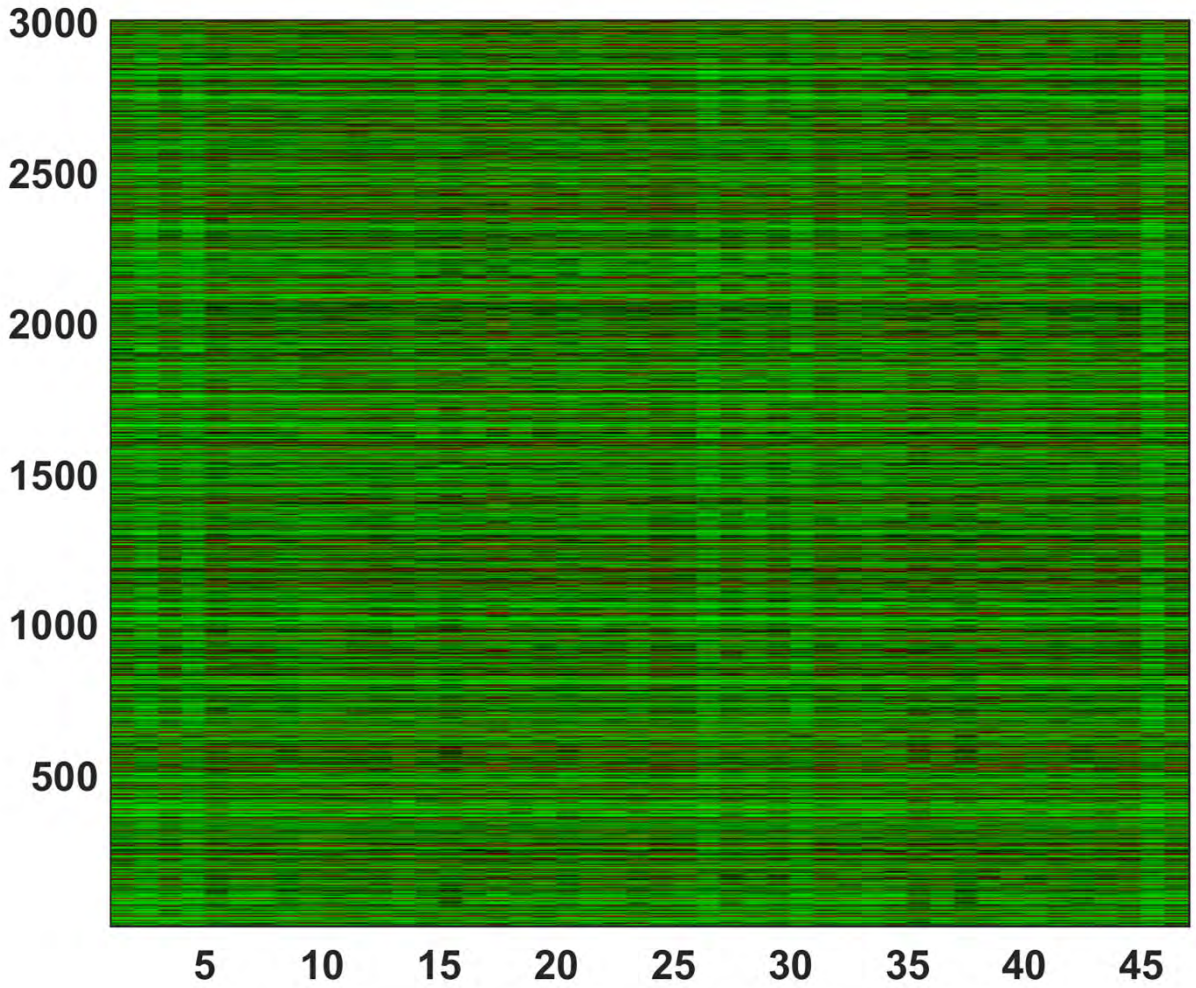
no obvious biological common function

```
g1=1+floor(rand.*3000); g2=1+floor(rand.*3000);
disp([g1, g2])
```

Matlab exercise

- Every group works with
  g0=2907; g1=1527; g2=2629; g3=2881;
  g4=1144; g5=1066;
- Compute Multiple Linear Regression (MLR),
  where y=exp_t (g0);
  x1= exp_t (g1); x2= exp_t (g2);
- How much better the MLR did compared to the Single Linear Regression (SLR)?
- Continue increasing the number of genes in x until R_adj starts to decrease

How to find the entire groups of mutually correlated genes if you have many genes and many samples?



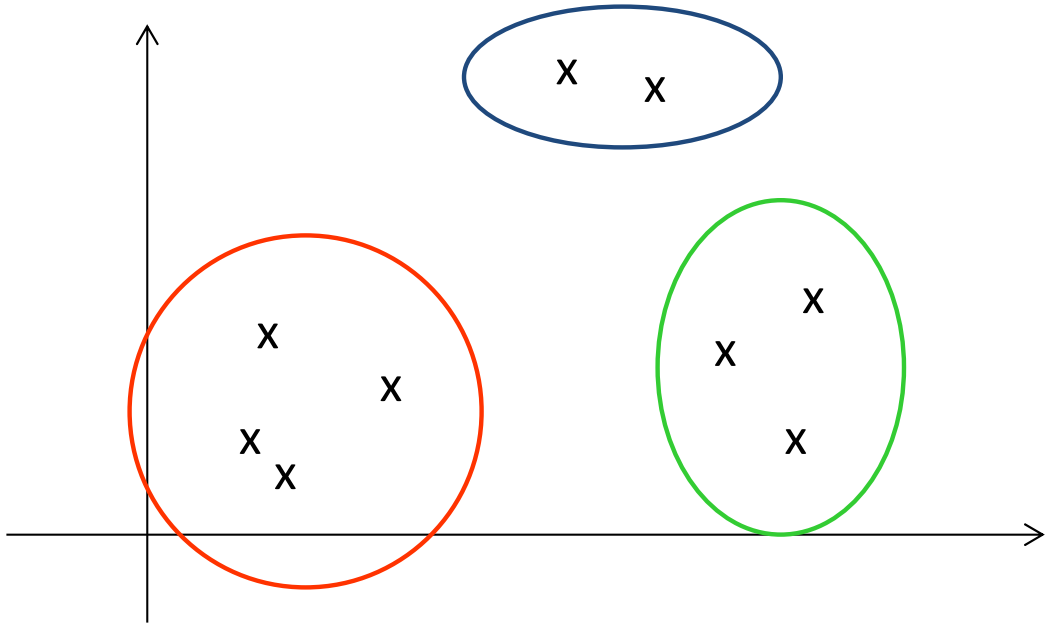
Clustering to the rescue!

Clustering is a part of Machine Learning

- ***Supervised Learning:***
  A machine learning technique whereby a system uses a set of human-labelled training examples to learn how to correctly perform a task
  Example: a sample of cancer expression profiles each **annotated** with cancer type
  Goal: predict cancer type based on expression pattern
- ***Unsupervised Learning (including clustering):***
  In machine learning, unsupervised learning is a class of problems in which one seeks to determine how the data are organized. One only has unlabeled examples.
  Example: a sample of breast cancer expression profiles.
  Goal: Identify several different (yet unknown) subtypes with potentially different treatments

What is clustering?

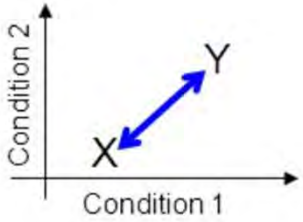
- The goal of clustering is to
  - group data points that are close (or **similar**) to each other
  - Usually, one needs to identify such groups (or clusters) in an **unsupervised** manner
  - Sometimes one takes into account **prior information** (Bayesian methods)
- Need to define some distance d_{ij} between objects i and j
- Clustering is easy in 2 dimensions but hard in 3000 dimensions -> need to somehow reduce dimensionality



How to define the distance?

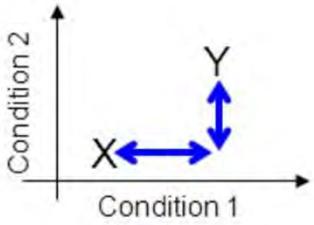
- Euclidean distance:
  - Most commonly used distance
  - Sphere shaped cluster
  - Corresponds to the geometric distance into the multidimensional space

$$d(X,Y) = \sqrt{\sum_i (x_i - y_i)^2}$$



- City Block (Manhattan) distance:
  - Sum of differences across dimensions
  - Less sensitive to outliers
  - Diamond shaped clusters

$$d(X,Y) = \sum_i |x_i - y_i|$$



The Canberra distance metric is calculated in R by

$$\sum \left(\frac{|x_i - y_i|}{|x_i + y_i|} \right).$$

Correlation coefficient distance

$$d(X,Y) = 1 - \rho(X,Y) = 1 - \frac{Cov(X,Y)}{\sqrt{(Var(X) \cdot Var(Y)}}$$

Common types of clustering algorithms

- Hierarchical if one doesn't know in advance the # of clusters
  - Agglomerative: start with N clusters and gradually merge them into 1 cluster
  - Divisive: start with 1 cluster and gradually break it up into N clusters
- Non-hierarchical algorithms
  - K-means clustering:
    - Iteratively apply the following two steps:
    - Calculate the centroid (center of mass) of each cluster
    - Assign each to the cluster to the nearest centroid
  - Principal Component Analysis (PCA)
    - plot pairs of top eigenvectors of the covariance matrix $Cov(X_i, X_j)$ and uses visual information to group

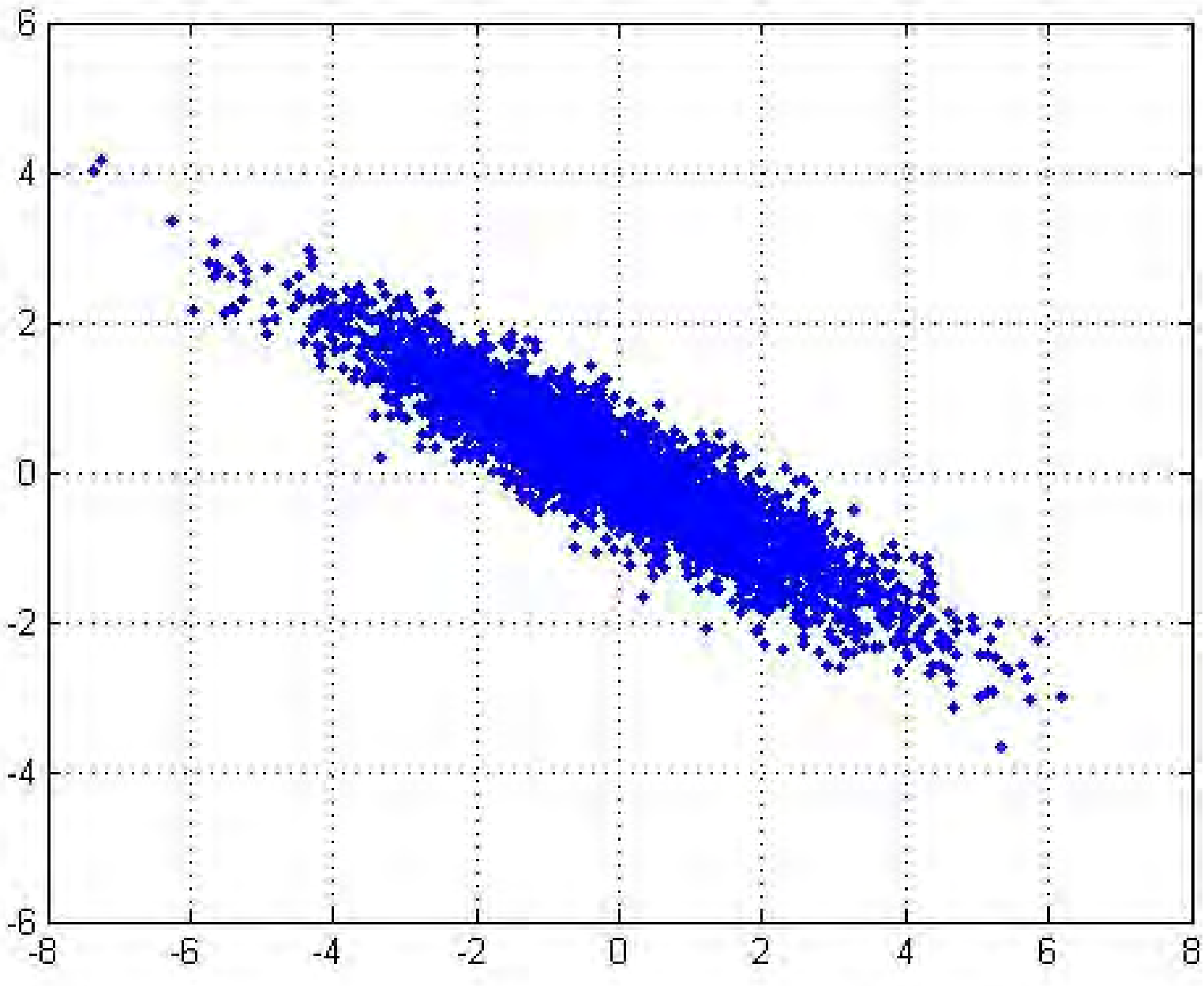
The Principal Components

- **Vectors** originating from the center of mass

- Principal component #1 points
  in the direction of the **largest variance**.

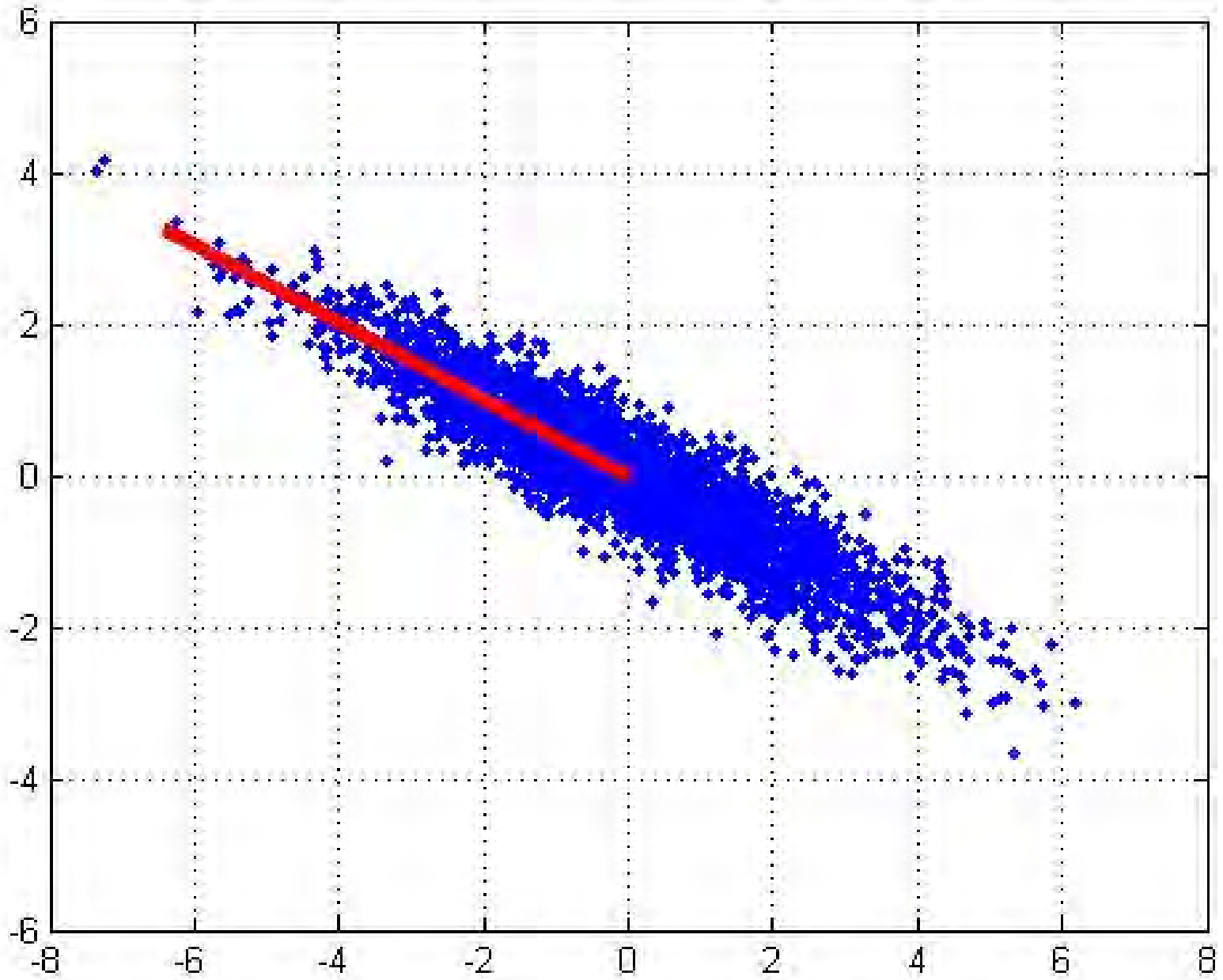
- Each subsequent principal component…
  - is **orthogonal** to the previous ones, and
  - points in the directions of the **largest variance of the residual subspace**

2D Gaussian dataset



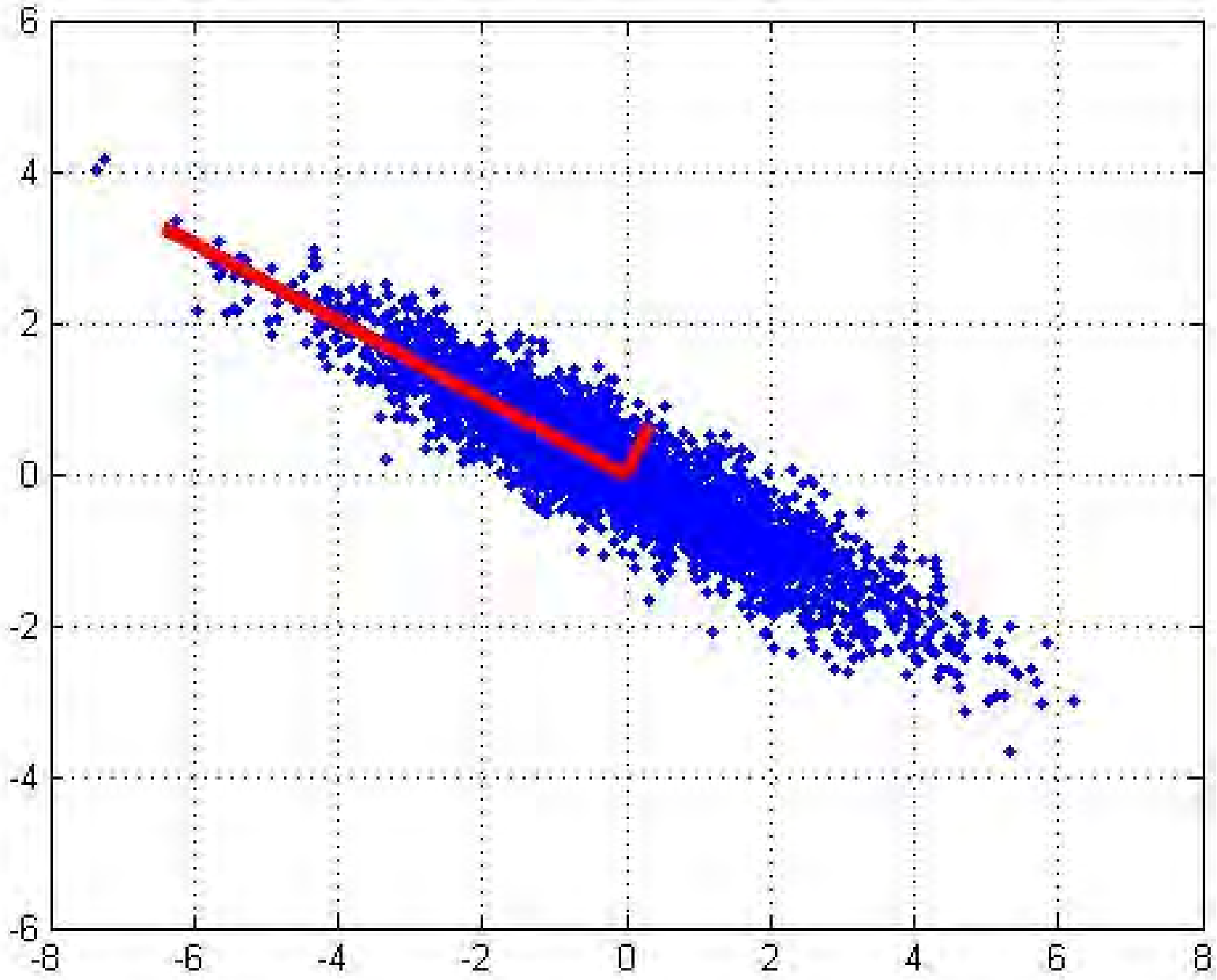
Adapted from lectures Prof. Pat Virtue at CMU based on original slide from Barnabas Poczos

1st PCA axis



Adapted from lectures Prof. Pat Virtue at CMU based on original slide from Barnabas Poczos

2nd PCA axis



Adapted from lectures Prof. Pat Virtue at CMU based on original slide from Barnabas Poczos

Data for PCA

$$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N} \qquad \mathbf{X} = \begin{bmatrix} (\mathbf{x}^{(1)})^T \\ (\mathbf{x}^{(2)})^T \\ \vdots \\ (\mathbf{x}^{(N)})^T \end{bmatrix}$$

We assume the data is **centered**

$$\mu = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^{(i)} = \mathbf{0}$$

Q: What if your data is **not** centered?

A: Subtract off the sample mean

Adapted from lectures Prof. Pat Virtue at CMU based on the original slides from Matt Gormley

Sample Covariance Matrix

The sample covariance matrix is given by:

$$\Sigma_{jk} = \frac{1}{N}\sum_{i=1}^{N}(x_j^{(i)} - \mu_j)(x_k^{(i)} - \mu_k)$$

Since the data matrix is centered, we rewrite as:

$$\mathbf{\Sigma} = \frac{1}{N}\mathbf{X}^T\mathbf{X}$$

$$\mathbf{X} = \begin{bmatrix} (\mathbf{x}^{(1)})^T \\ (\mathbf{x}^{(2)})^T \\ \vdots \\ (\mathbf{x}^{(N)})^T \end{bmatrix}$$

Adapted from lectures Prof. Pat Virtue at CMU based on the original slides from Matt Gormley

PCA algorithm

PCA algorithm($\mathbf{X}$, k): top k eigenvalues/eigenvectors

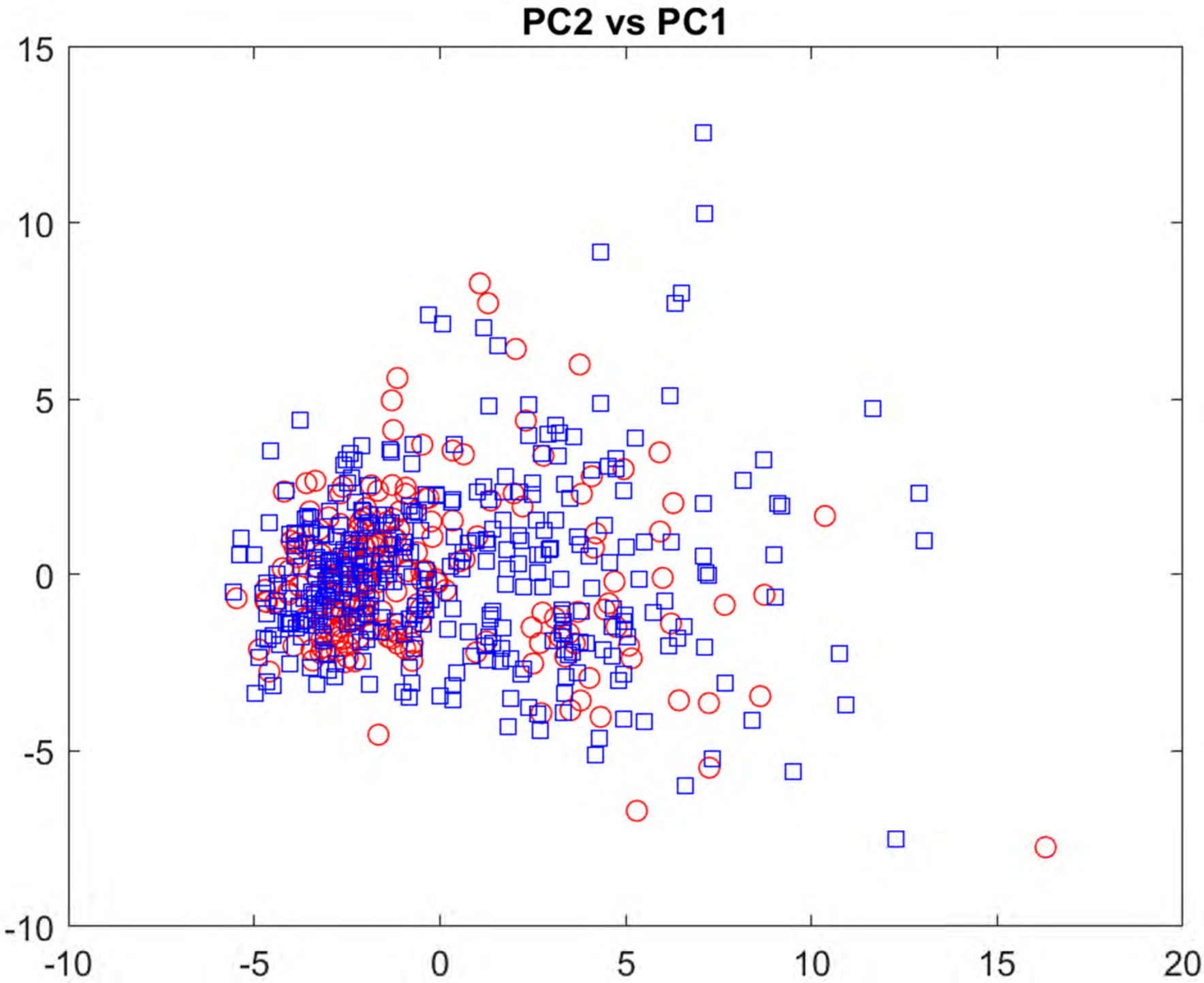
- $\{ \lambda_i, \mathbf{u}_i \}_{i=1:m}$ = eigenvectors/eigenvalues of Σ
  $\ldots \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$
- **PCA** basis vectors = the eigenvectors of Σ
- Larger eigenvalue $\Rightarrow$ more important eigenvectors

PCA and units

- When different variables have different units (like temperature and mass), the meaning of principal components is a somewhat arbitrary
- One way of making the PCA less arbitrary is to use variables scaled so as to have unit variance, by standardizing the data
- Before making PCA of X transform it using Z=zscore(X);

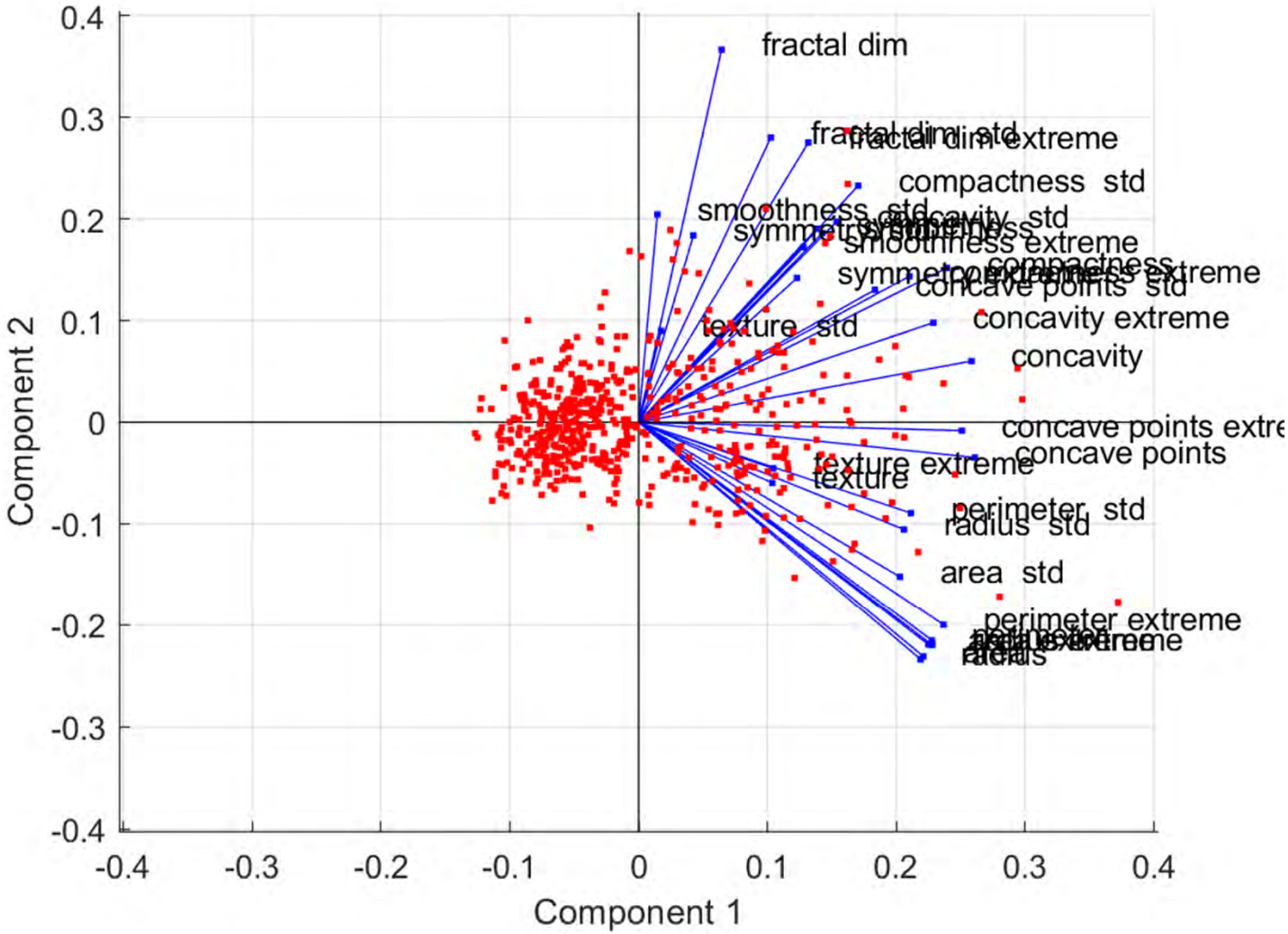
Group project 4

- load cancer_wdbc.mat
- Z=zscore(cancerwdbc);
- [coeff_z, score_z, latent_z] = pca(Z);
- ic=find(cancer_yn==1); whos ic;
  inc=find(cancer_yn==0); whos inc;
- figure; plot(score_z(ic,1), score_z(ic,2),'ro'); hold on; plot(score_z(inc,1), score_z(inc,2),'bs');
  title('PC2 vs PC1');
- Plot pairs of score_z components
  - 1st principal component vs 2nd principal component.
  - 1st principal component vs 3rd principal component
  - 3rd principal component vs 2nd principal component

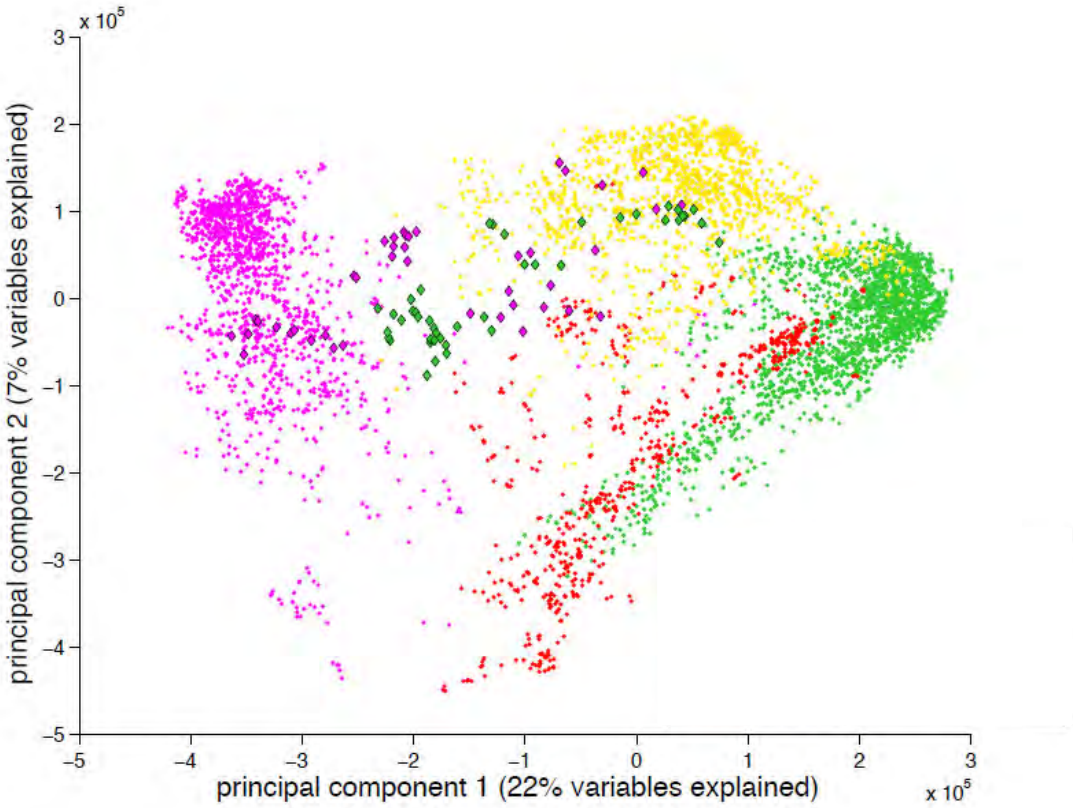

PC2 vs PC1
15
10
5
0
-5
-10
-10
-5
0
5
10
15
20

Which variables contribute to which PC? Add loadings (coeff eigenvectors)

- figure; biplot(coeff_z(:,1:2),'scores',score_z(:,1:2), 'VarLabels', feature_names);



Example of Principal Component Analysis (PCA) clustering



7000 gene expression samples of model plant *Arabidopsis thaliana*

- root
- seedling
- vegetative(leaf+shoot+stem)
- reproductive(flower+seed+silique)
- cultured root
- cultured aerial

Plant J. 2016 Mar 25. doi: 10.1111/tpj.13175. [Epub ahead of print]

Large-scale atlas of microarray data reveals the distinct expression landscape of different tissues in Arabidopsis.

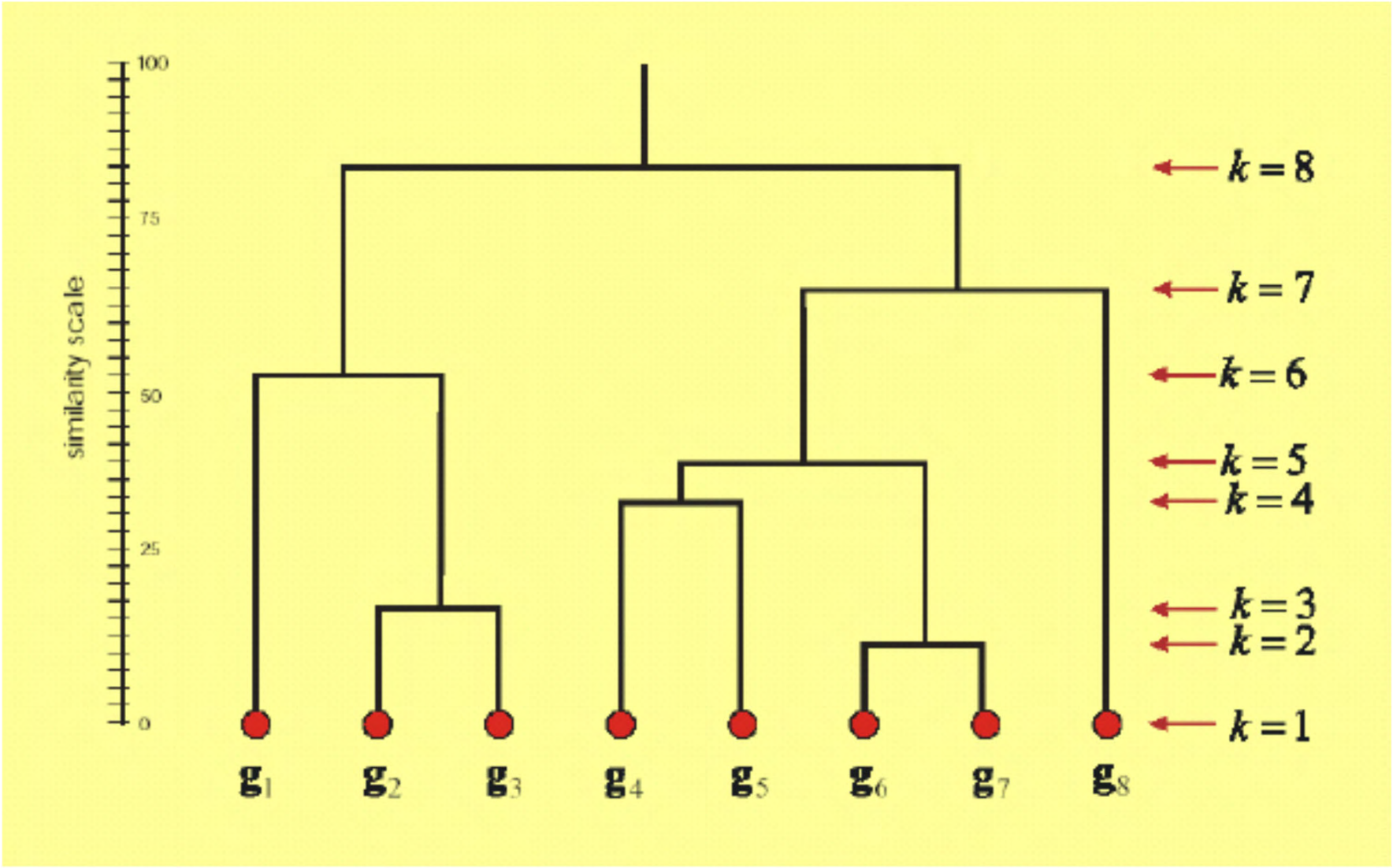
He F[1], Yoo S[2,3], Wang D[4], Kumari S[5], Gerstein M[4], Ware D[5,6], Maslov S[1,7].

Hierarchical clustering

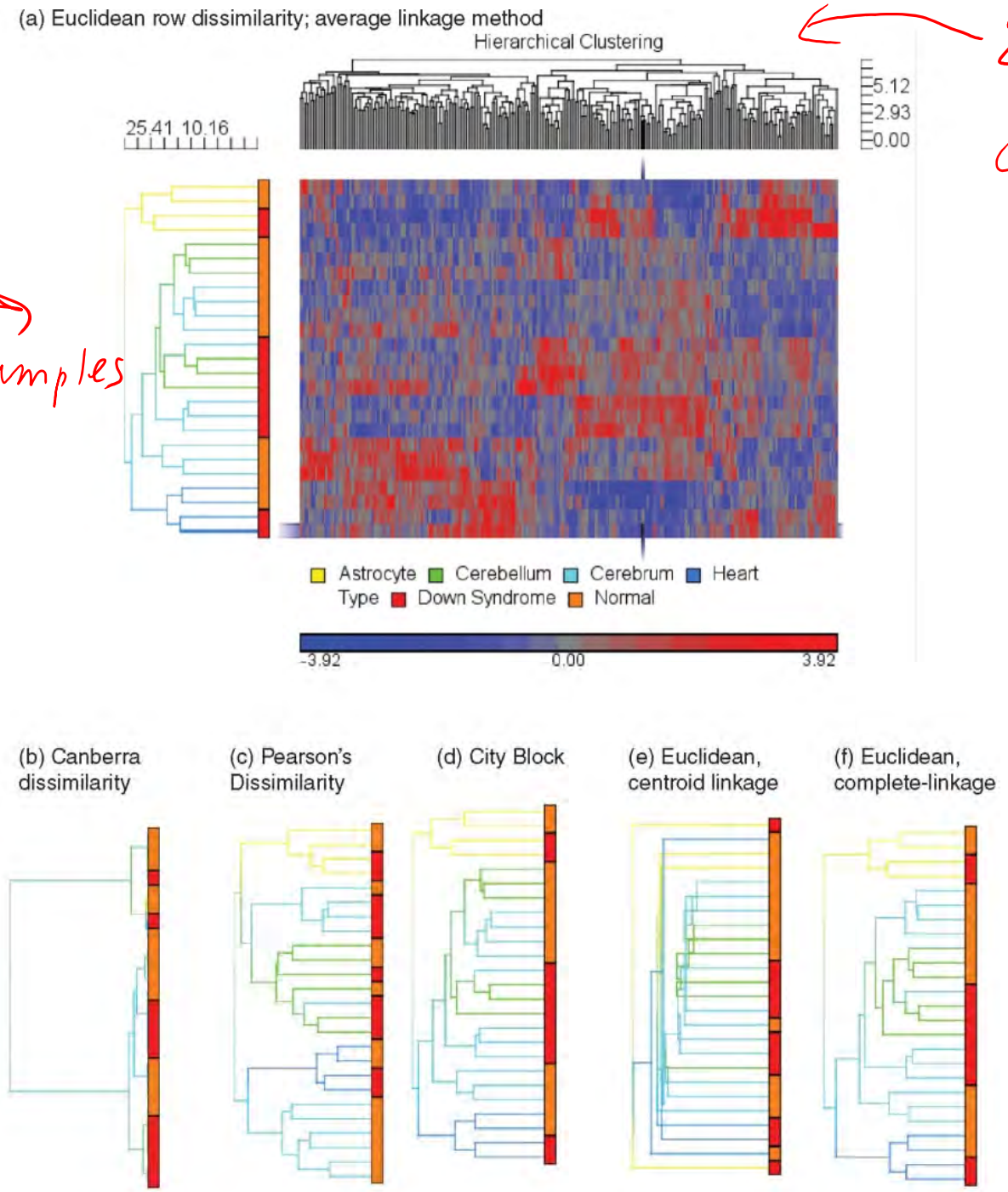
UPGMA algorithm

- Hierarchical agglomerative clustering algorithm
- **UPGMA** = Unweighted Pair Group Method with Arithmetic mean
- Iterative algorithm:
- Start with a pair with the smallest d(X,Y)
- Cluster these two together and replace it with their arithmetic mean (X+Y)/2
- Recalculate all distances to this new “cluster node”
- Repeat until all nodes are merged

Output of UPGMA algorithm



UPGMA algorithm



250 genes on chromosome 21

25 samples

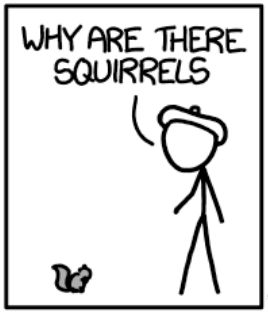
FIGURE 11.16 Hierarchical clustering of 250 chromosome 21 transcripts in 25 samples using Partek software. (a) Hierarchical clustering of microarray data using the default settings of Euclidean dissimilarity for rows (samples) and columns (transcripts). Colors correspond to expression intensity values.

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 DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE
WHY IS LYING GOOD

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 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 SEX SO IMPORTANT

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 AREN'T MY ARMS GROWING

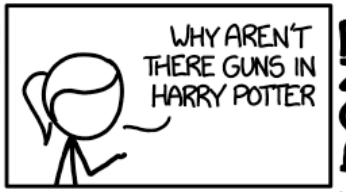
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

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 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 FEMALE MR MIMES

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 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 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 AREN'T THERE GUNS IN HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

Clustering
Matlab demo

Choices of distance metrics in clustergram(… 'RowPDistValue' …, 'ColumnPDistValue' …,)

Metric	Description
'euclidean'	Euclidean distance (default).
'seuclidean'	Standardized Euclidean distance. Each coordinate difference between rows in X is scaled by dividing by the corresponding element of the standard deviation S=nanstd(X). To specify another value for S, use D=pdist(X,'seuclidean',S).
'cityblock'	City block metric.
'minkowski'	Minkowski distance. The default exponent is 2. To specify a different exponent, use D = pdist(X,'minkowski',P), where P is a scalar positive value of the exponent.
'chebychev'	Chebychev distance (maximum coordinate difference).
'mahalanobis'	Mahalanobis distance, using the sample covariance of X as computed by nancov. To compute the distance with a different covariance, use D = pdist(X,'mahalanobis',C), where the matrix C is symmetric and positive definite.
'cosine'	One minus the cosine of the included angle between points (treated as vectors).
'correlation'	One minus the sample correlation between points (treated as sequences of values).
'spearman'	One minus the sample Spearman's rank correlation between observations (treated as sequences of values).
'hamming'	Hamming distance, which is the percentage of coordinates that differ.
'jaccard'	One minus the Jaccard coefficient, which is the percentage of nonzero coordinates that differ.
custom distance function	A distance function specified using @: D = pdist(X,@distfun) A distance function must be of form d2 = distfun(XI,XJ) taking as arguments a 1-by-*n* vector XI, corresponding to a single row of X, and an *m*2-by-*n* matrix XJ, corresponding to multiple rows of X. distfun must accept a matrix XJ with an arbitrary number of rows. distfun must return an *m*2-by-1 vector of distances d2, whose *k*th element is the distance between XI and XJ(k,:).

Choices of hierarchical clustering algorithm in clustergram(…'linkage',…)

`X`	Matrix with two or more rows. The rows represent observations, the columns represent categories or dimensions.
`method`	Algorithm for computing distance between clusters.

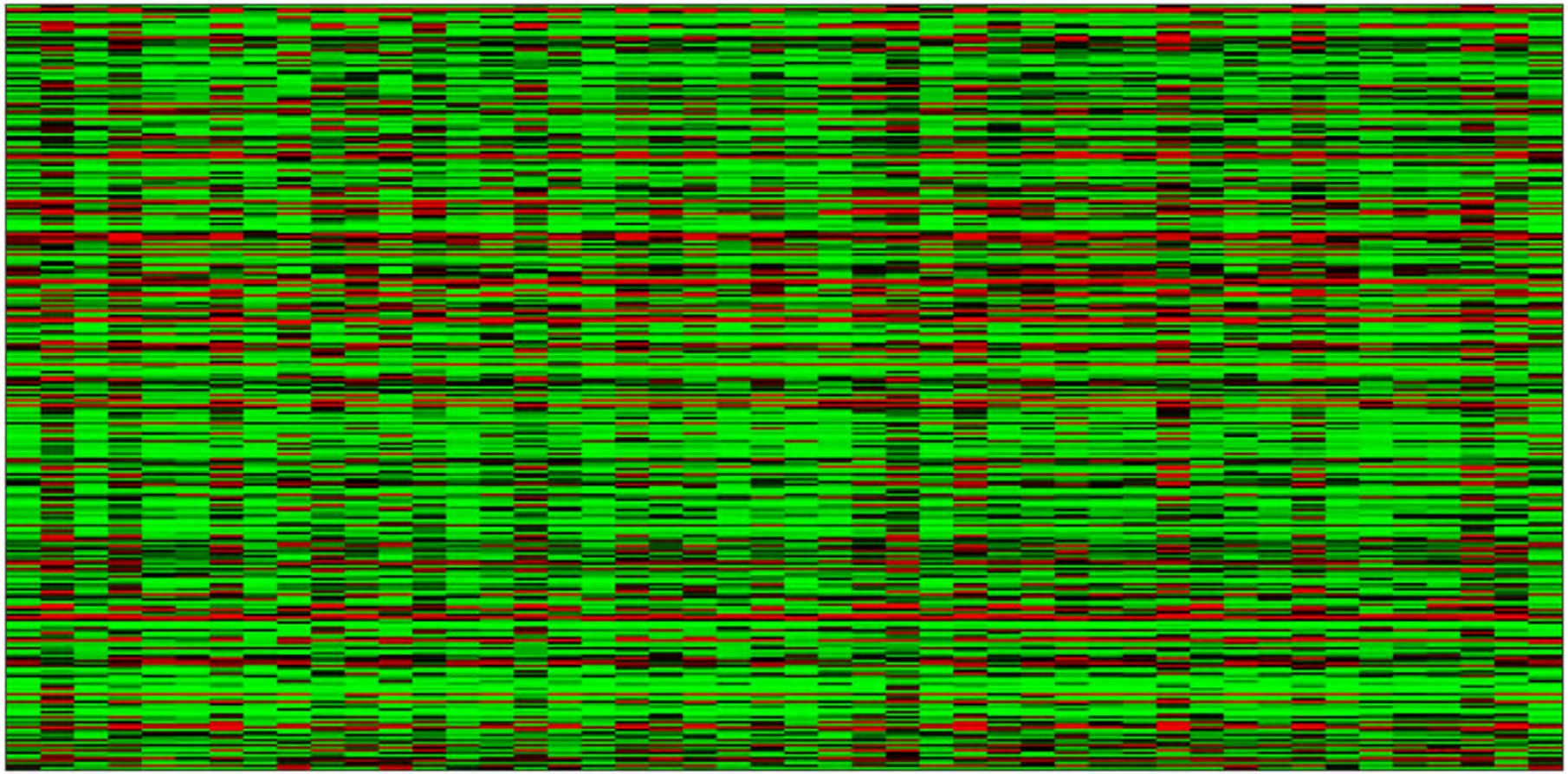
Method	Description
`'average'`	Unweighted average distance (UPGMA)
`'centroid'`	Centroid distance (UPGMC), appropriate for Euclidean distances only
`'complete'`	Furthest distance
`'median'`	Weighted center of mass distance (WPGMC), appropriate for Euclidean distances only
`'single'`	Shortest distance
`'ward'`	Inner squared distance (minimum variance algorithm), appropriate for Euclidean distances only
`'weighted'`	Weighted average distance (WPGMA)

Default: `'single'`

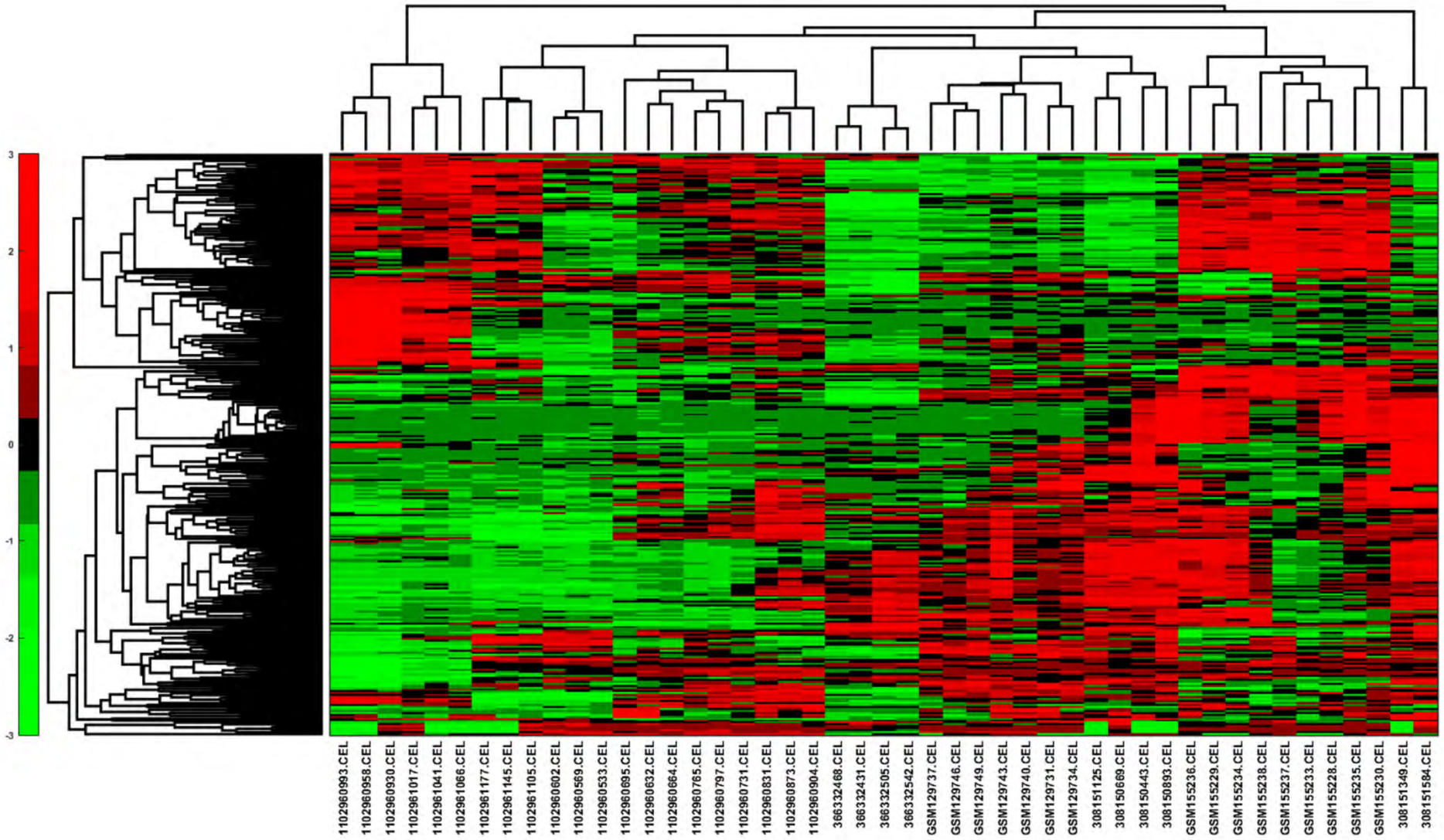
Clustering group exercise

- Each group will analyze a cluster of genes identified in the T cell expression table
- Analyze the table of top 100 genes **by variance** in 47 samples
- Cluster them using:
  - Group 1: UPGMA = 'linkage', 'average', 'RowPDistValue', 'euclidean',
  - Group 2: 'linkage', 'single', 'RowPDistValue', 'cityblock',
  - Group 3: 'linkage', 'average', 'RowPDistValue', 'correlation',
  - Group 4: UPGMA = 'linkage', 'single', 'RowPDistValue', 'euclidean',
  - Group 5: UPGMA = 'linkage', 'weighted', 'RowPDistValue', 'correlation',
- Use clustergram(…, 'Standardize','Row', 'linkage', ***as specified for your group,*** 'RowPDistValue' ***as specified for your group,*** 'RowLabels',gene_names1,'ColumnLabels', array_names)

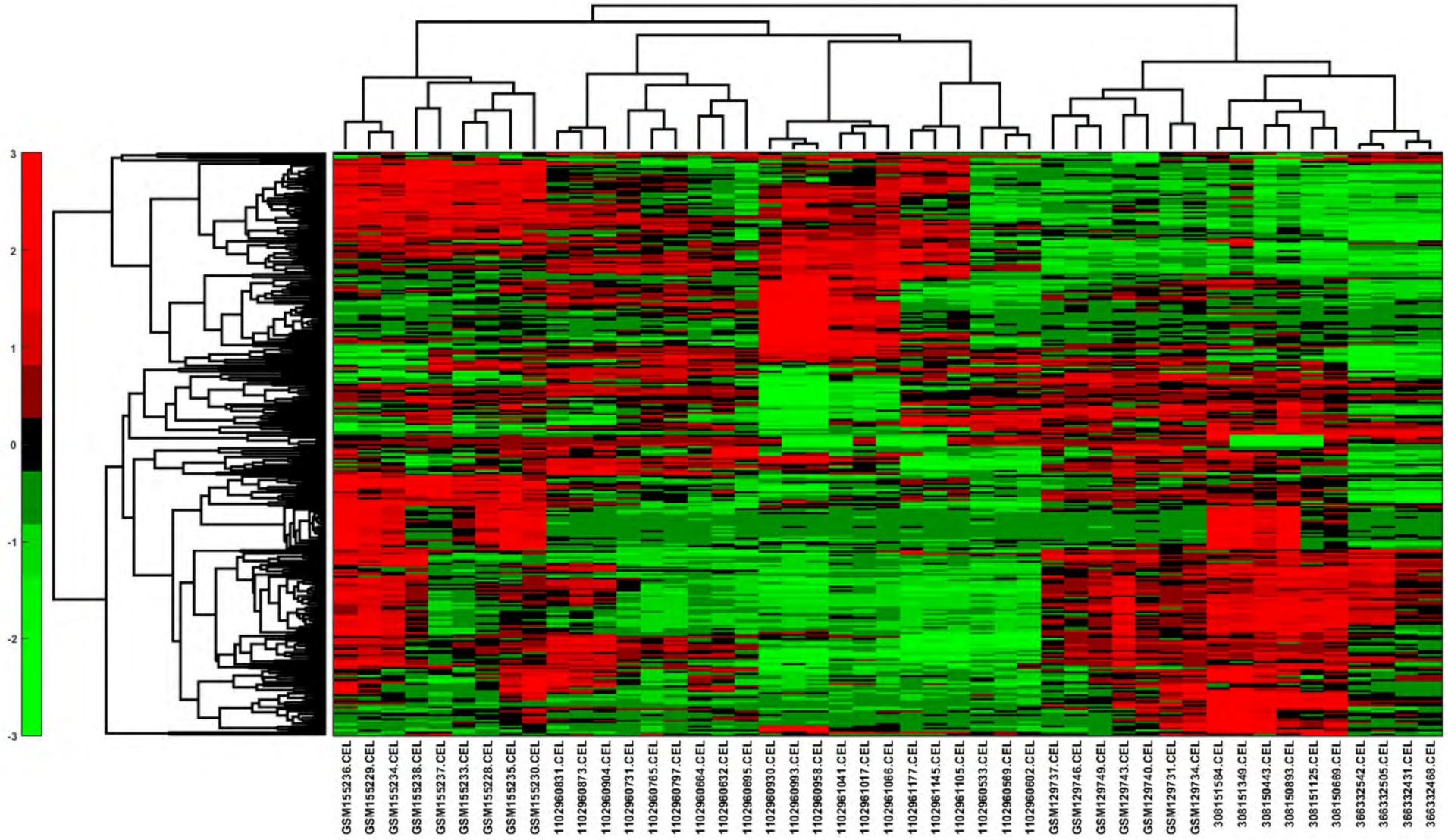
Before clustering



UPGMA hierarchical clustering, Euclidian distance



UPGMA hierarchical clustering, correlation distance



Clustering group exercise

- Each group will analyze a cluster of genes identified in the T cell expression table
- Analyze the table of top 100 genes **by variance** in 47 samples
- Cluster them using:
  - Group 1: UPGMA = 'linkage', 'average', 'RowPDistValue', 'euclidean',
  - Group 2: 'linkage', 'single', 'RowPDistValue', 'cityblock',
  - Group 3: 'linkage', 'average', 'RowPDistValue', 'correlation',
  - Group 4: UPGMA = 'linkage', 'single', 'RowPDistValue', 'euclidean',
  - Group 5: UPGMA = 'linkage', 'weighted', 'RowPDistValue', 'correlation',
- Use clustergram(…, 'Standardize','Row', 'linkage', ***as specified for your group,*** 'RowPDistValue' ***as specified for your group,*** 'RowLabels',gene_names1,'ColumnLabels', array_names)

Cluster analysis group exercise

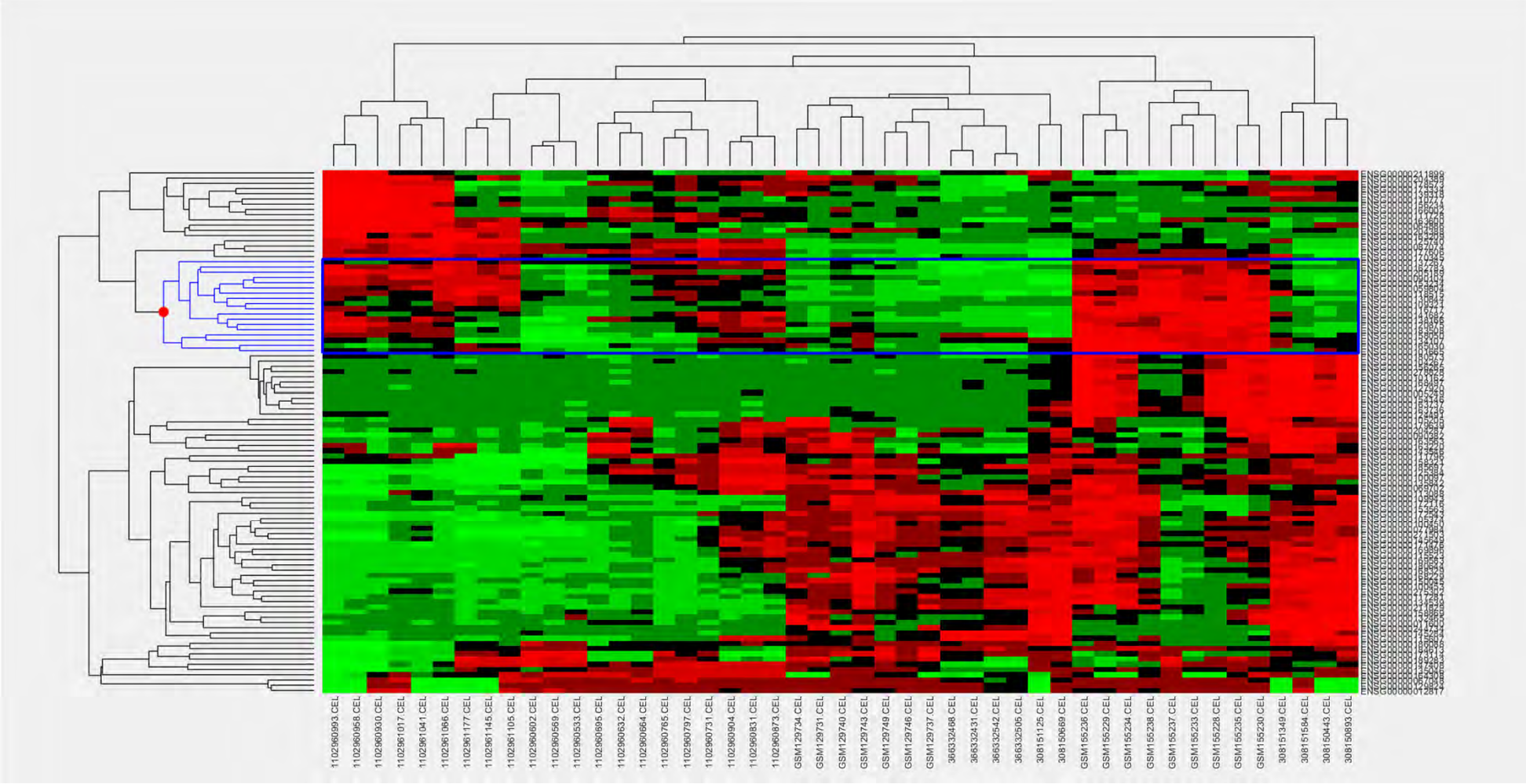
- Which biological functions are overrepresented in different clusters?
- Pick a cluster:
  - Select a node on the tree of rows,
  - Right click
  - Choose "export group info" into the workspace
  - Name it gene_list
- Run the following two Matlab commands to display genes
  - g1=gene_list.RowNodeNames;
  - for m=1:length(g1); disp(g1{m}); end;

Search for shared biological functions

- copy the list of displayed genes
- go to "Start Analysis" on https://david.ncifcrf.gov/tools.jsp
- Paste genes from gene list displayed by Matlab into the box in the left panel of the website
- select ENSEMBL_GENE_ID and “gene list” radio button
- Click "Functional Annotation Clustering“
- Select groups in “Annotation Summary Results” which have many genes from your list. Definitely select “PUBMED_ID” and interaction databases like “Biogrid”
- First look at "Functional Annotation Chart" rectangular button below to display all overrepresented terms. Sort by “Benjamini” correction for multiple hypotheses tetsting
- Select "Functional Annotation Clustering" rectangular button below to display annotation results for gene list broken into multiple groups (clusters) each with related biological functions
- Write down the # of genes in the cluster and the top functions in two most interesting clusters

Using Group 1 options:

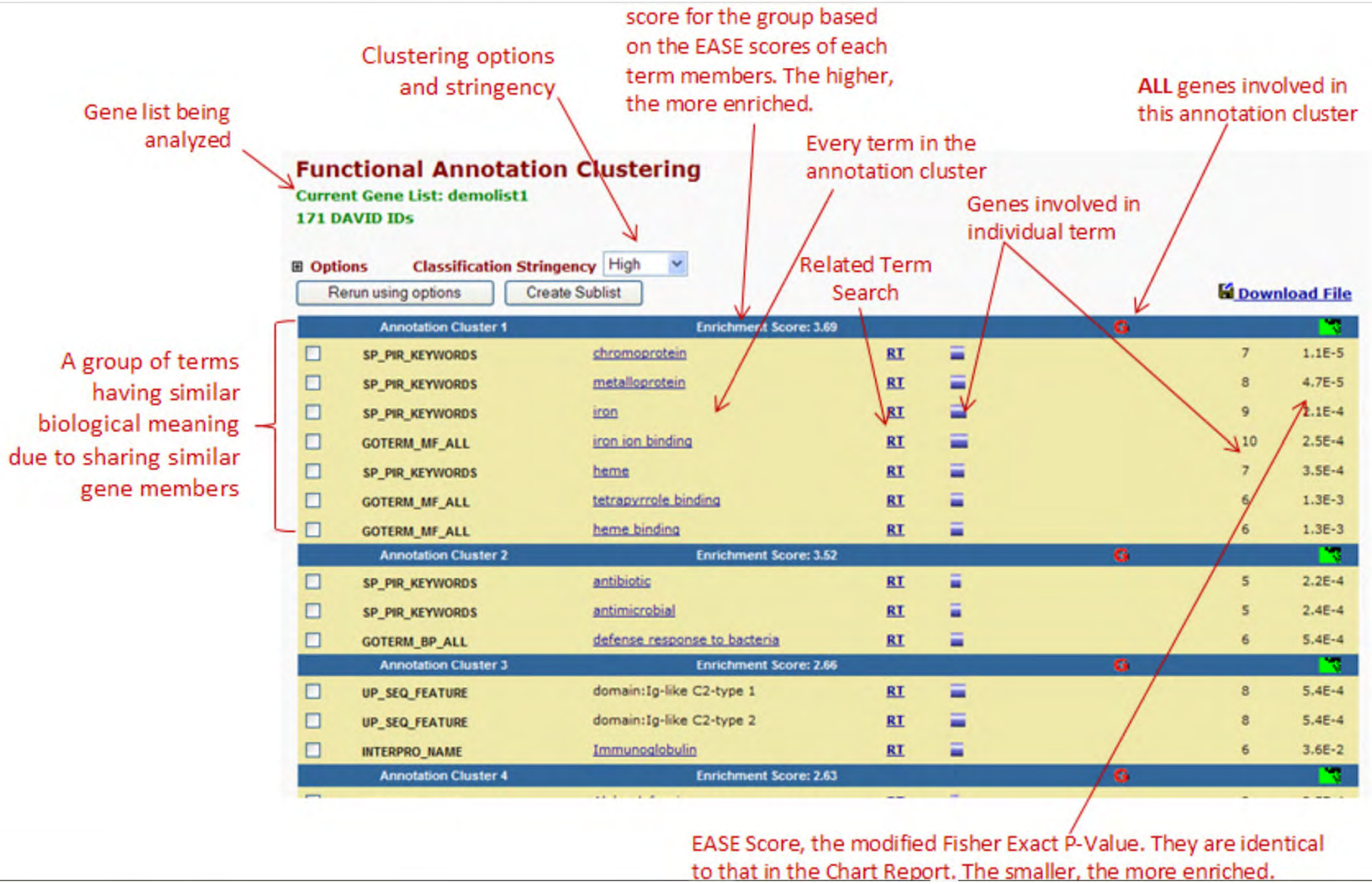
‘linkage’, ‘average’, ‘RowPDistValue’, ’euclidean’,



54 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_CC_DIRECT	nucleus	RT		16	88.9	8.1E-7	3.7E-5
	PIR_SUPERFAMILY	dual specificity protein phosphatase (MAP kinase phosphatase)	RT		3	16.7	4.0E-5	8.0E-5
	GOTERM_MF_DIRECT	protein tyrosine/threonine phosphatase activity	RT		3	16.7	3.4E-5	1.3E-3
	GOTERM_MF_DIRECT	MAP kinase tyrosine phosphatase activity	RT		3	16.7	3.4E-5	1.3E-3
	GOTERM_MF_DIRECT	MAP kinase tyrosine/serine/threonine phosphatase activity	RT		3	16.7	5.9E-5	1.5E-3
	INTERPRO	Mitogen-activated protein (MAP) kinase phosphatase	RT		3	16.7	3.3E-5	1.9E-3
	SMART	RHOD	RT		3	16.7	2.5E-4	4.8E-3
	INTERPRO	Rhodanese-like domain	RT		3	16.7	2.2E-4	6.2E-3
	SMART	DSPc	RT		3	16.7	8.4E-4	8.0E-3
	INTERPRO	Dual specificity phosphatase, catalytic domain	RT		3	16.7	6.0E-4	9.2E-3
	INTERPRO	Dual specificity phosphatase, subgroup, catalytic domain	RT		3	16.7	6.6E-4	9.2E-3
	GOTERM_BP_DIRECT	endoderm formation	RT		3	16.7	5.6E-5	1.1E-2
	UP_KW_CELLULAR_COMPONENT	Nucleus	RT		13	72.2	1.5E-3	1.3E-2
	SMART	PTPc_motif	RT		3	16.7	2.3E-3	1.5E-2
	GOTERM_MF_DIRECT	phosphoprotein phosphatase activity	RT		3	16.7	8.0E-4	1.5E-2
	INTERPRO	Protein-tyrosine phosphatase, catalytic	RT		3	16.7	1.4E-3	1.6E-2
	UP_KW_PTM	Ubl conjugation	RT		7	38.9	4.5E-3	1.9E-2
	UP_KW_PTM	Isopeptide bond	RT		6	33.3	5.4E-3	1.9E-2
	INTERPRO	Protein-tyrosine phosphatase, active site	RT		3	16.7	2.1E-3	2.0E-2
	INTERPRO	Protein-tyrosine/Dual specificity phosphatase	RT		3	16.7	2.8E-3	2.3E-2
	UP_SEQ_FEATURE	DOMAIN:Rhodanese	RT		3	16.7	1.9E-4	2.4E-2
	KEGG_PATHWAY	MAPK signaling pathway	RT		5	27.8	5.9E-4	2.8E-2
	GOTERM_MF_DIRECT	myosin phosphatase activity	RT		3	16.7	2.4E-3	3.6E-2
	GOTERM_MF_DIRECT	protein tyrosine phosphatase activity	RT		3	16.7	4.2E-3	5.3E-2
	GOTERM_CC_DIRECT	nucleoplasm	RT		10	55.6	2.3E-3	5.4E-2
	GOTERM_BP_DIRECT	negative regulation of MAPK cascade	RT		3	16.7	7.0E-4	6.8E-2


Gene list being analyzed
Clustering options and stringency
score for the group based on the EASE scores of each term members. The higher, the more enriched.
ALL genes involved in this annotation cluster
Every term in the annotation cluster
Genes involved in individual term
Related Term Search
A group of terms having similar biological meaning due to sharing similar gene members
Functional Annotation Clustering
Current Gene List: demolist1
171 DAVID IDs
Options
Classification Stringency
High
Rerun using options
Create Sublist
Download File
Annotation Cluster 1 Enrichment Score: 3.69
SP_PIR_KEYWORDS chromoprotein RT 7 1.1E-5
SP_PIR_KEYWORDS metalloprotein RT 8 4.7E-5
SP_PIR_KEYWORDS iron RT 9 1.1E-4
GOTERM_MF_ALL iron ion binding RT 10 2.5E-4
SP_PIR_KEYWORDS heme RT 7 3.5E-4
GOTERM_MF_ALL tetrapyrrole binding RT 6 1.3E-3
GOTERM_MF_ALL heme binding RT 6 1.3E-3
Annotation Cluster 2 Enrichment Score: 3.52
SP_PIR_KEYWORDS antibiotic RT 5 2.2E-4
SP_PIR_KEYWORDS antimicrobial RT 5 2.4E-4
GOTERM_BP_ALL defense response to bacteria RT 6 5.4E-4
Annotation Cluster 3 Enrichment Score: 2.66
UP_SEQ_FEATURE domain:Ig-like C2-type 1 RT 8 5.4E-4
UP_SEQ_FEATURE domain:Ig-like C2-type 2 RT 8 5.4E-4
INTERPRO_NAME Immunoglobulin RT 6 3.6E-2
Annotation Cluster 4 Enrichment Score: 2.63
EASE Score, the modified Fisher Exact P-Value. They are identical to that in the Chart Report. The smaller, the more enriched.

Functional Annotation Clustering

Help and Manual

Current Gene List: List_3
Current Background: Homo sapiens
18 DAVID IDs

Options **Classification Stringency** Medium

Rerun using options | Create Sublist

25 Cluster(s)

Download File

Annotation Cluster 1	Enrichment Score: 5.2	G		Count	P_Value	Benjamini
DISGENET	Juvenile arthritis	RT		7	1.5E-8	4.7E-7
DISGENET	Juvenile psoriatic arthritis	RT		7	1.5E-8	4.7E-7
DISGENET	Polyarthritis, Juvenile, Rheumatoid Factor Negative	RT		7	1.5E-8	4.7E-7
DISGENET	Polyarthritis, Juvenile, Rheumatoid Factor Positive	RT		7	1.5E-8	4.7E-7
DISGENET	Juvenile-Onset Still Disease	RT		7	1.8E-8	4.7E-7
KEGG_PATHWAY	MAPK signaling pathway	RT		5	5.9E-4	2.8E-2
BIOGRID_INTERACTION	mitogen-activated protein kinase 1(MAPK1)	RT		4	3.8E-3	1.0E0
WIKIPATHWAYS	MAPK signaling pathway	RT		3	5.8E-2	6.9E-1
GAD_DISEASE_CLASS	UNKNOWN	RT		5	1.5E-1	9.9E-1
Annotation Cluster 2	**Enrichment Score: 2.83**	**G**		**Count**	**P_Value**	**Benjamini**
INTERPRO	Mitogen-activated protein (MAP) kinase phosphatase	RT		3	3.3E-5	1.9E-3
GOTERM_MF_DIRECT	protein tyrosine/threonine phosphatase activity	RT		3	3.4E-5	1.3E-3
GOTERM_MF_DIRECT	MAP kinase tyrosine phosphatase activity	RT		3	3.4E-5	1.3E-3
PIR_SUPERFAMILY	dual specificity protein phosphatase (MAP kinase phosphatase)	RT		3	4.0E-5	8.0E-5
GOTERM_BP_DIRECT	endoderm formation	RT		3	5.6E-5	1.1E-2
GOTERM_MF_DIRECT	MAP kinase tyrosine/serine/threonine phosphatase activity	RT		3	5.9E-5	1.5E-3
PUBMED_ID	27880917	RT		4	1.7E-4	2.5E-2
UP_SEQ_FEATURE	DOMAIN:Rhodanese	RT		3	1.9E-4	2.4E-2
INTERPRO	Rhodanese-like domain	RT		3	2.2E-4	6.2E-3
SMART	RHOD	RT		3	2.5E-4	4.8E-3

Annotation Cluster 3	Enrichment Score: 2.43	G		Count	P_Value	Benjamini
DISGENET	Arsenic Poisoning, Inorganic	RT		3	3.5E-3	4.6E-2
DISGENET	Nervous System, Organic Arsenic Poisoning	RT		3	3.5E-3	4.6E-2
DISGENET	Arsenic Poisoning	RT		3	3.5E-3	4.6E-2
DISGENET	Arsenic Encephalopathy	RT		3	3.5E-3	4.6E-2
DISGENET	Arsenic Induced Polyneuropathy	RT		3	3.5E-3	4.6E-2
DISGENET	Dermatologic disorders	RT		3	5.1E-3	5.6E-2
Annotation Cluster 4	**Enrichment Score: 2.26**	G		**Count**	**P_Value**	**Benjamini**
PUBMED_ID	19322201	RT		7	1.3E-8	5.9E-6
BIOGRID_INTERACTION	ELAV like RNA binding protein 1(ELAVL1)	RT		7	4.4E-3	1.0E0
UCSC_TFBS	CEBPA	RT		7	1.8E-1	1.0E0
UCSC_TFBS	CDPCR3HD	RT		7	6.5E-1	1.0E0
UCSC_TFBS	FOXD3	RT		5	7.4E-1	1.0E0
Annotation Cluster 5	**Enrichment Score: 2.14**	G		**Count**	**P_Value**	**Benjamini**
GOTERM_BP_DIRECT	negative regulation of transcription from RNA polymerase II promoter	RT		6	1.4E-3	9.1E-2
BIOGRID_INTERACTION	retinoid X receptor alpha(RXRA)	RT		3	6.1E-3	1.0E0
GOTERM_MF_DIRECT	protein heterodimerization activity	RT		3	4.5E-2	3.7E-1
Annotation Cluster 6	**Enrichment Score: 1.95**	G		**Count**	**P_Value**	**Benjamini**
REACTOME_PATHWAY	Generic Transcription Pathway	RT		7	2.8E-3	1.7E-1
REACTOME_PATHWAY	RNA Polymerase II Transcription	RT		7	4.6E-3	1.7E-1
REACTOME_PATHWAY	Gene expression (Transcription)	RT		7	8.2E-3	2.0E-1
GAD_DISEASE_CLASS	UNKNOWN	RT		5	1.5E-1	9.9E-1
Annotation Cluster 7	**Enrichment Score: 1.76**	G		**Count**	**P_Value**	**Benjamini**
PUBMED_ID	18029348	RT		6	1.8E-5	3.4E-3
UP_KW_PTM	Isopeptide bond	RT		6	5.4E-3	1.9E-2
PUBMED_ID	15342556	RT		3	7.9E-3	4.8E-1
PUBMED_ID	26496610	RT		3	1.0E-1	1.0E0
GOTERM_MF_DIRECT	metal ion binding	RT		4	4.5E-1	1.0E0
UCSC_TFBS	TAL1ALPHAE47	RT		3	7.9E-1	1.0E0

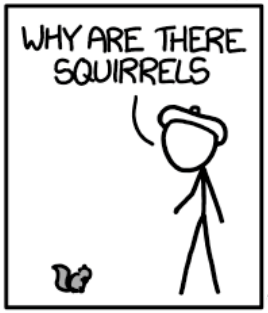
Annotation Cluster 3	Enrichment Score: 2.43	G		Count	P_Value	Benjamini
DISGENET	Arsenic Poisoning, Inorganic	RT		3	3.5E-3	4.6E-2
DISGENET	Nervous System, Organic Arsenic Poisoning	RT		3	3.5E-3	4.6E-2
DISGENET	Arsenic Poisoning	RT		3	3.5E-3	4.6E-2
DISGENET	Arsenic Encephalopathy	RT		3	3.5E-3	4.6E-2
DISGENET	Arsenic Induced Polyneuropathy	RT		3	3.5E-3	4.6E-2
DISGENET	Dermatologic disorders	RT		3	5.1E-3	5.6E-2
Annotation Cluster 4	**Enrichment Score: 2.26**	G		**Count**	**P_Value**	**Benjamini**
PUBMED_ID	19322201	RT		7	1.3E-8	5.9E-6
BIOGRID_INTERACTION	ELAV like RNA binding protein 1(ELAVL1)	RT		7	4.4E-3	1.0E0
UCSC_TFBS	CEBPA	RT		7	1.8E-1	1.0E0
UCSC_TFBS	CDPCR3HD	RT		7	6.5E-1	1.0E0
UCSC_TFBS	FOXD3	RT		5	7.4E-1	1.0E0
Annotation Cluster 5	**Enrichment Score: 2.14**	G		**Count**	**P_Value**	**Benjamini**
GOTERM_BP_DIRECT	negative regulation of transcription from RNA polymerase II promoter	RT		6	1.4E-3	9.1E-2
BIOGRID_INTERACTION	retinoid X receptor alpha(RXRA)	RT		3	6.1E-3	1.0E0
GOTERM_MF_DIRECT	protein heterodimerization activity	RT		3	4.5E-2	3.7E-1
Annotation Cluster 6	**Enrichment Score: 1.95**	G		**Count**	**P_Value**	**Benjamini**
REACTOME_PATHWAY	Generic Transcription Pathway	RT		7	2.8E-3	1.7E-1
REACTOME_PATHWAY	RNA Polymerase II Transcription	RT		7	4.6E-3	1.7E-1
REACTOME_PATHWAY	Gene expression (Transcription)	RT		7	8.2E-3	2.0E-1
GAD_DISEASE_CLASS	UNKNOWN	RT		5	1.5E-1	9.9E-1
Annotation Cluster 7	**Enrichment Score: 1.76**	G		**Count**	**P_Value**	**Benjamini**
PUBMED_ID	18029348	RT		6	1.8E-5	3.4E-3
UP_KW_PTM	Isopeptide bond	RT		6	5.4E-3	1.9E-2
PUBMED_ID	15342556	RT		3	7.9E-3	4.8E-1
PUBMED_ID	26496610	RT		3	1.0E-1	1.0E0
GOTERM_MF_DIRECT	metal ion binding	RT		4	4.5E-1	1.0E0
UCSC_TFBS	TAL1ALPHAE47	RT		3	7.9E-1	1.0E0

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 DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE
WHY IS LYING GOOD

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 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 SEX SO IMPORTANT

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 AREN'T MY ARMS GROWING

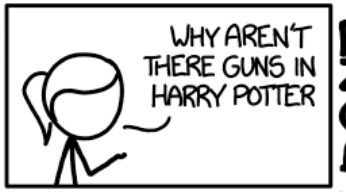
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

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 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 FEMALE MR MIMES

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 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 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

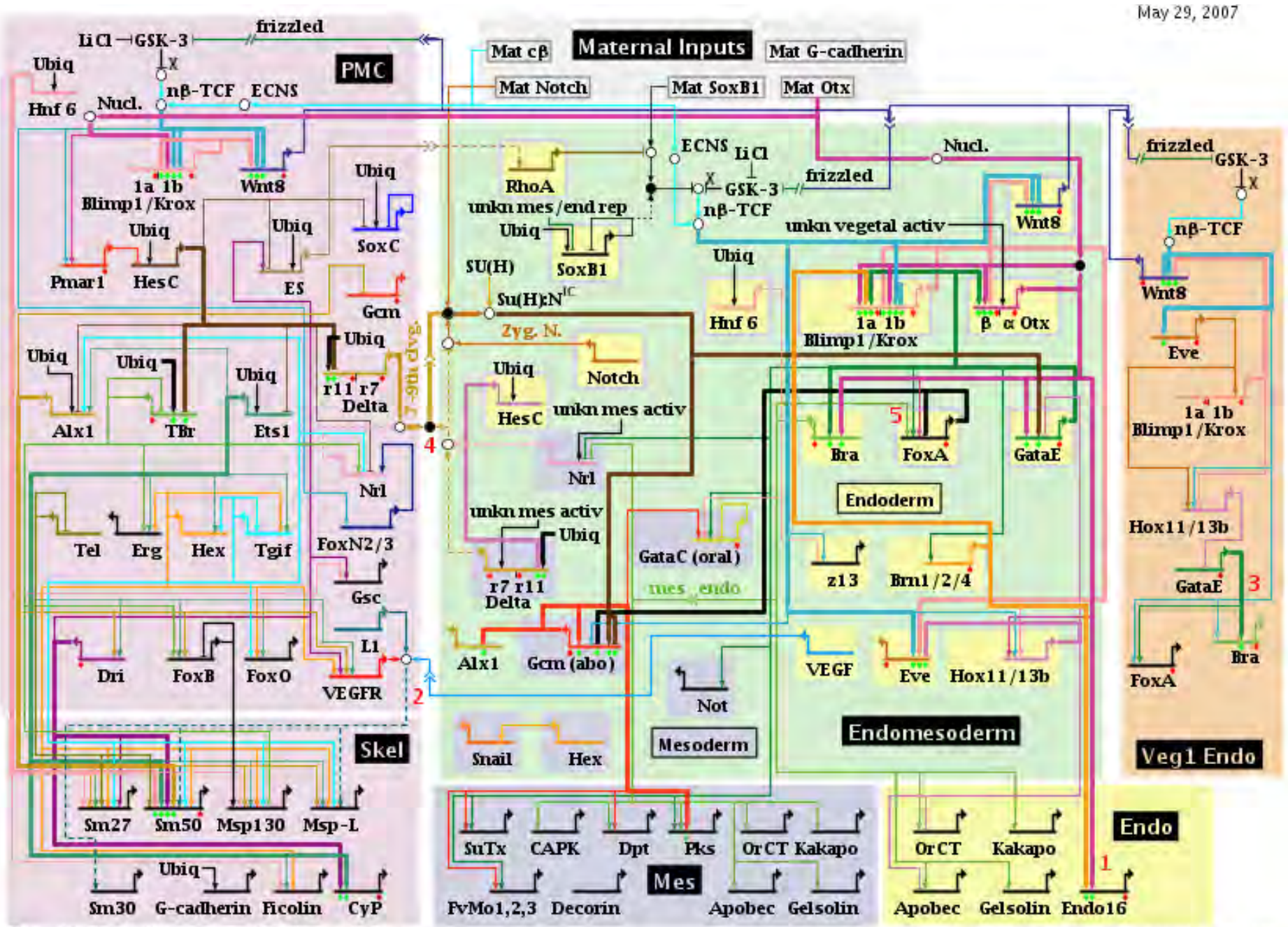
WHY AREN'T THERE GUNS IN HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

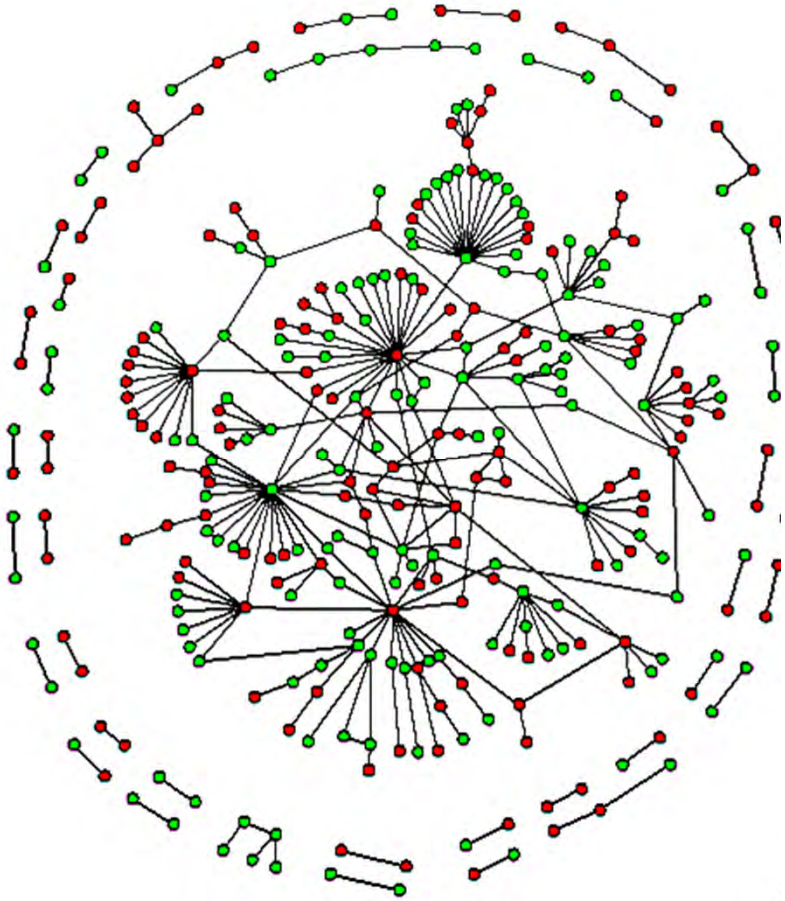
Reminder from the first lecture

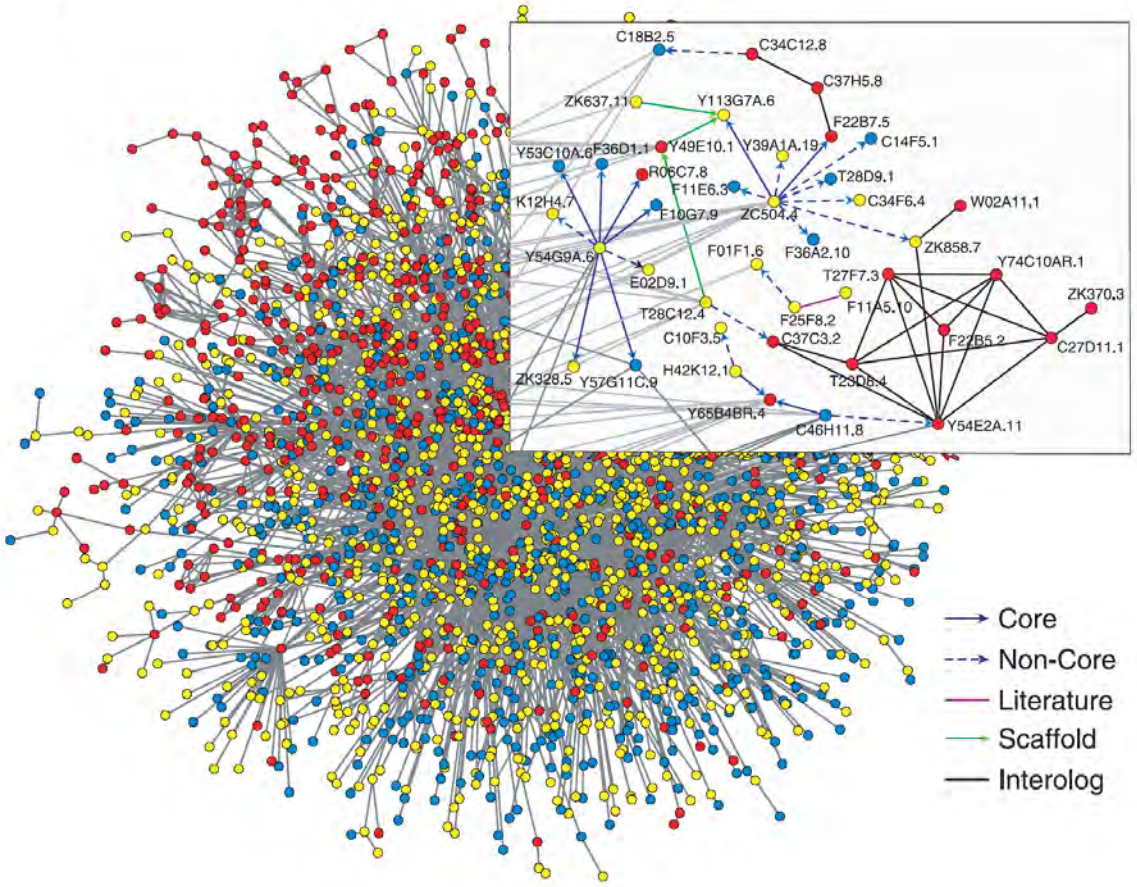
Sea urchin embryonic development (from endomesoderm up to 30 hours) by Davidson's lab



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

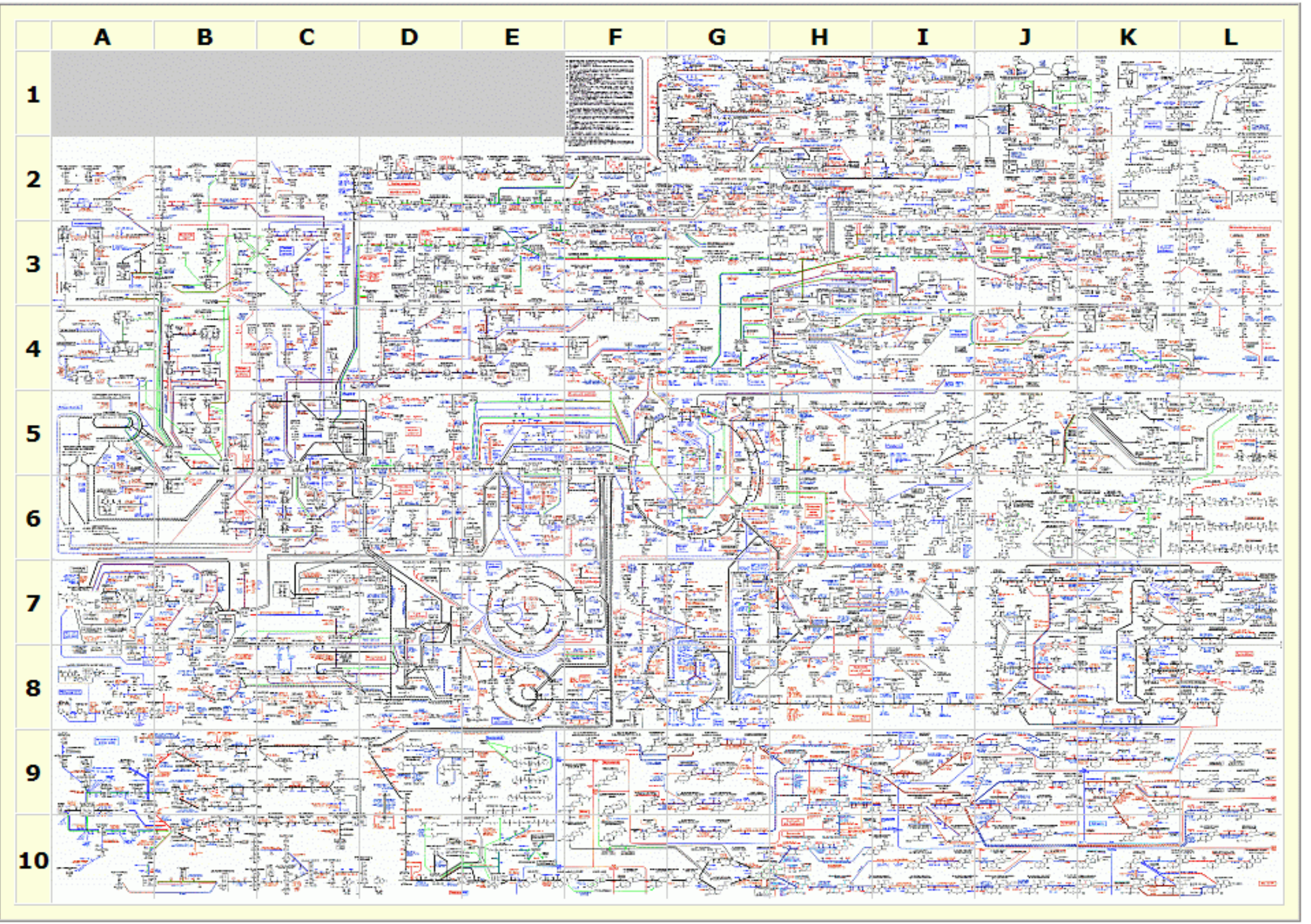




Baker's yeast *S. cerevisiae* (only nuclear proteins shown)
From S. Maslov, K. Sneppen, Science 2002

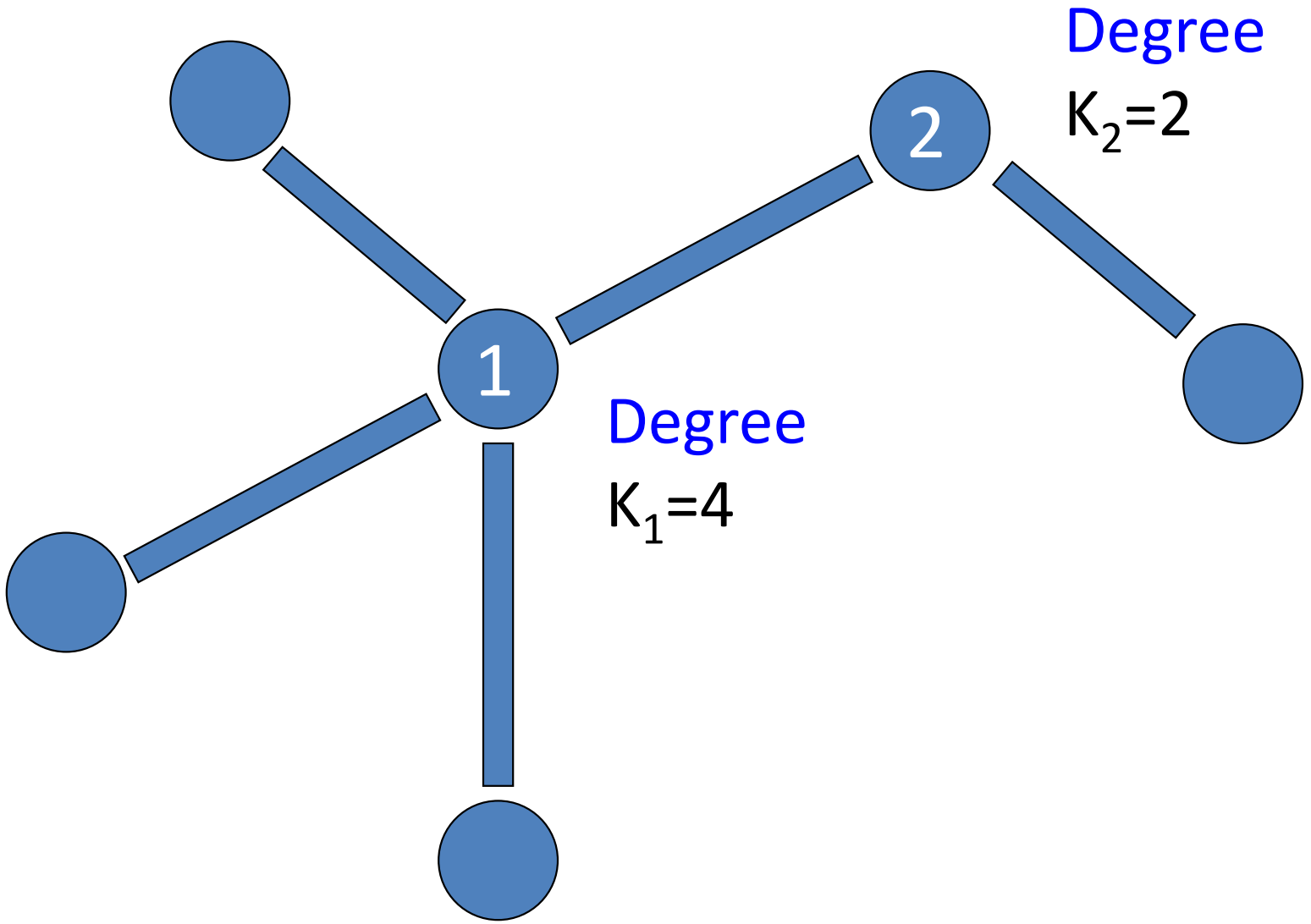
Worm *C. elegans*
From S. Lee et al , Science 2004

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

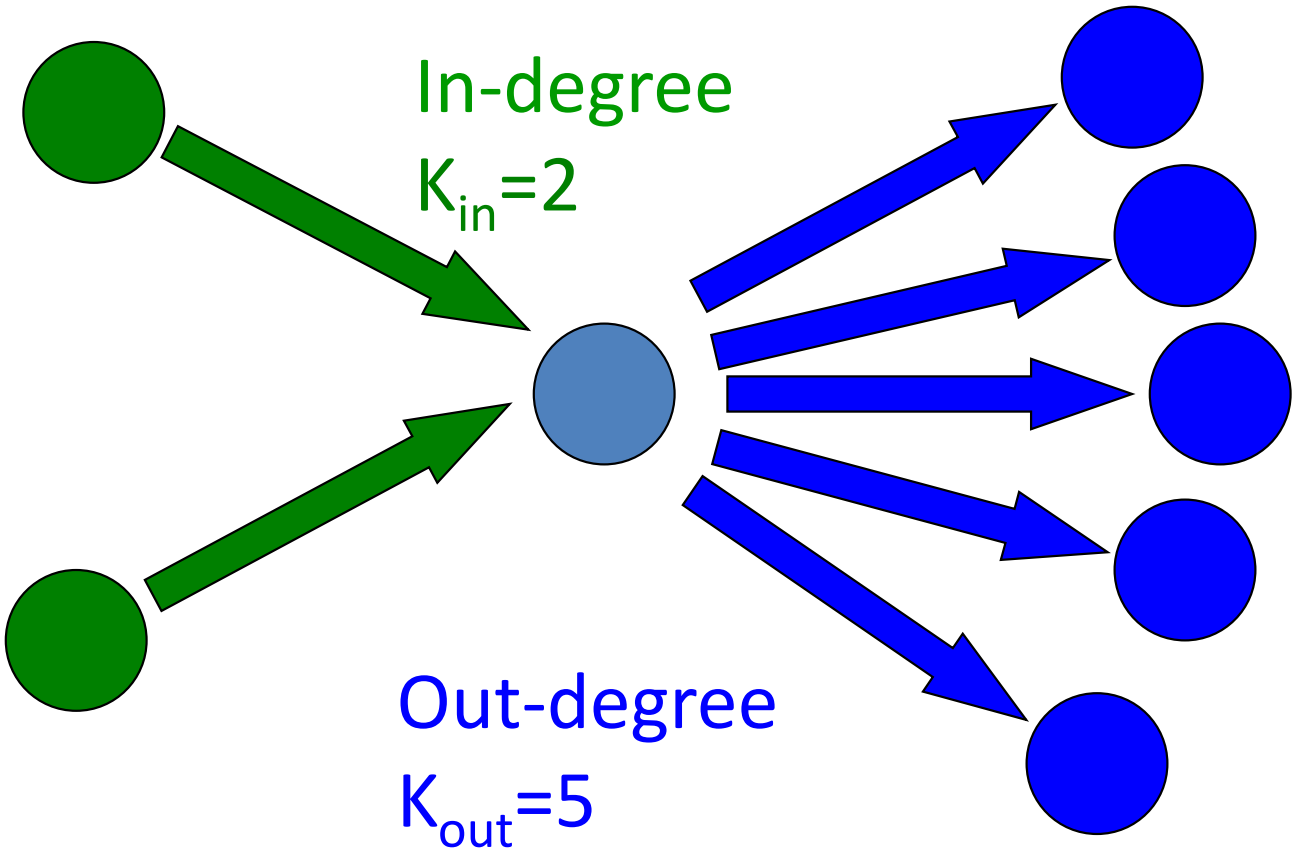


Basic concepts of network analysis

Degree of a node – its # of neighbors

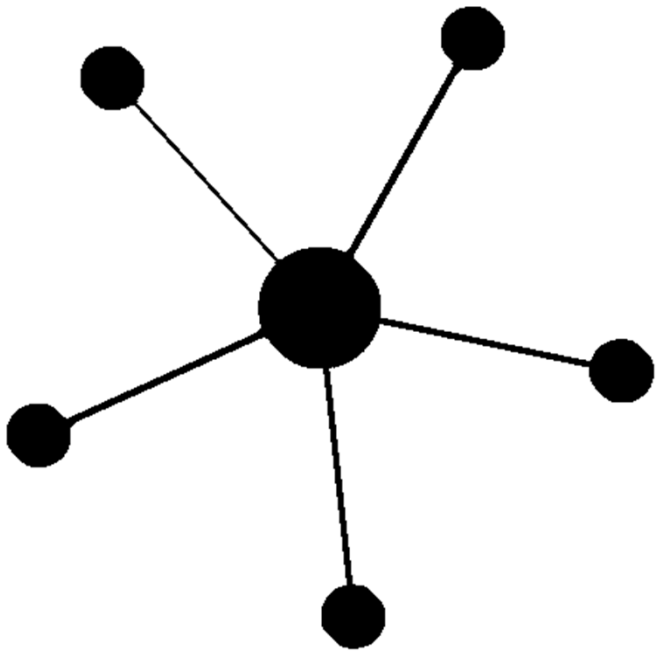


Directed networks have in- and out- degrees



How to find “important” nodes?

- By their degree
- Hubs = important
- Example: Google’s PageRank



How Google PageRank algorithm works?

- Google was solving the following problem in mid-1990s: too many websites match a typical search query: need to rank websites.
- Other popular search engines (e.g. Altavista) count the # of times a query word appears in website's text. Websites respond by putting lots of invisible words
- One could rank the importance of webpages by number of hyperlinks pointing to it (in-degree K_{in}) but:
  - Too democratic: It doesn't take into account the importance of webpages sending hyperlinks
  - it's easy to trick and artificially boost the rank
- Google's solution: simulate the behavior of many "random surfers" and then count the number of times they visited each webpage = it's PageRank
  - Popular pages send more surfers your way → the PageRank weight is proportional to K_{in} but weighted by popularity

PageRank algorithm is Google's $2.8T idea

- PageRank assigns to every webpage an importance score G_i
- The meaning of G_i – how often random surfers visit this website
- To determine solves a self-consistent Eq.:

  $G_i \sim \sum_j T_{ij} G_j$. Here
  $T_{ij} = A_{ij}/K_{out}$ (j) is the normalized adjacency matrix
- It finds the principal eigenvector (the one with the largest eigenvalue).

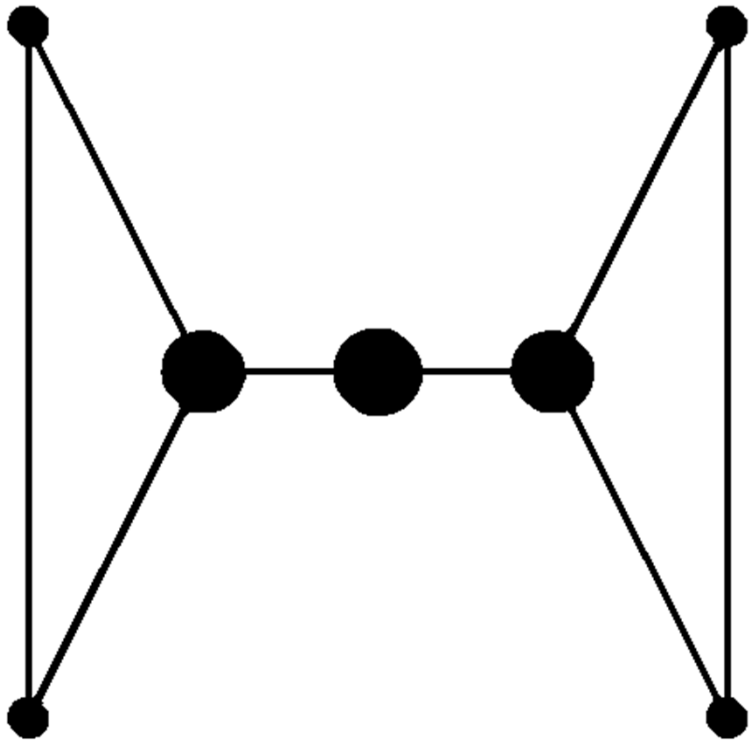
Problem with PageRank algorithm and how Google solved it

- Problem: surfers can be trapped in infinite loops with one or more entrances and no exits
- Model with random jumps mimicking surfers getting bored when following a chain of links

  $G_i \sim (1-\alpha)\sum_j T_{ij} G_j + \alpha \sum_j G_j$
- α=0.15 meaning that an average web surfer (circa 1995) on average jumped around $1/\alpha \approx 6$ webpages before going somewhere else

How to find “important” nodes?

- By their connectivity
- Connectors = important
- Betweenness-centrality



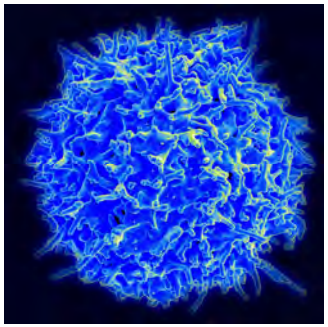
Betweenness centrality: definition

- Take a node i
- There are (N-1)*(N-2)/2 pairs of other nodes
- For each pair find the shortest path on the network
- If more than one shortest path, sample them equally
- Betweenness-centrality C(i) ~ the number of shortest paths going through node i

How is it connected to expression data analysis?

T-cell expression data

- The matrix contains 47 expression samples from Lukk et al, Nature Biotechnology 2011
- All samples are normal T-cells from different individuals
- Only the top 3000 genes with the largest variability were used
- The value is log2 of gene's expression level in a given sample as measured by microarray technology



A global map of human gene expression

Margus Lukk, Misha Kapushesky, Janne Nikkilä, Helen Parkinson, Angela Goncalves, Wolfgang Huber, Esko Ukkonen & Alvis Brazma

Affiliations | Corresponding author

Nature Biotechnology **28**, 322–324 (2010) | doi:10.1038/nbt0410-322

Although there is only one human genome sequence, different genes are expressed in many different cell types and tissues, as well as in different developmental stages or diseases. The structure of this 'expression space' is still largely unknown, as most transcriptomics experiments focus on sampling small regions. We have constructed a global gene expression map by integrating microarray data from 5,372 human samples representing 369 different cell and tissue types, disease states and cell lines. These have been compiled in an online resource (http://www.ebi.ac.uk/gxa/array/U133A) that allows the user to search for a gene of interest and

Correlated pairs

plausible biological connection based

on short description

g1=1994; g2=188; group 1

g1=2872; g2=1269; group 2

g1=1321; g2=10; group 3

g1= 886; g2=819; group 4

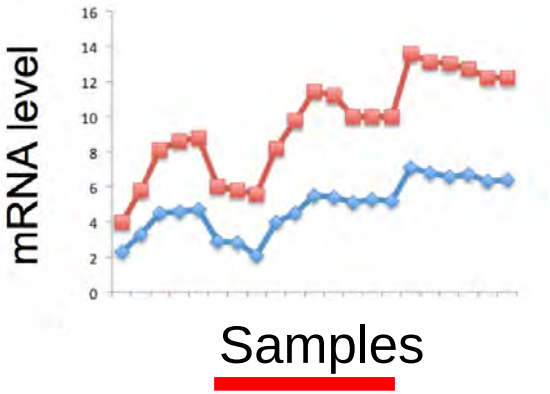
g1=2138; g2=1364; group 5

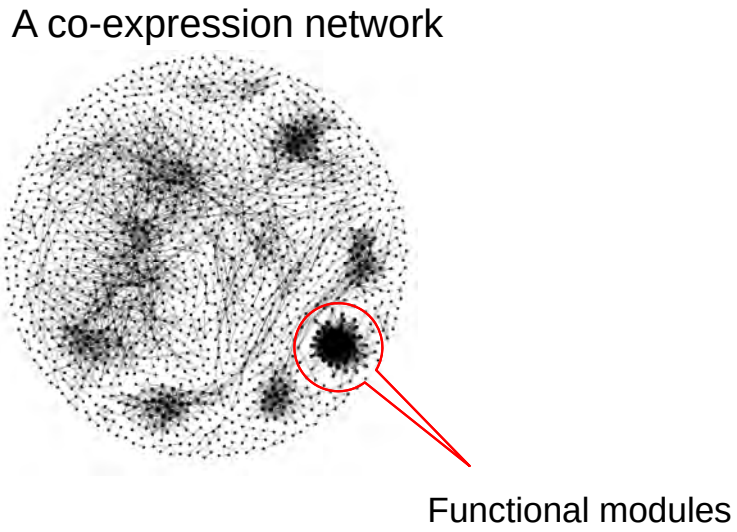
no obvious biological common function

```
g1=1+floor(rand.*3000); g2=1+floor(rand.*3000);
disp([g1, g2])
```

To analyze
correlations in expression
for all pairs of genes:
Co-expression networks

How to construct a co-expression network?





- Start with a matrix of log2 of expression levels of N genes in K samples (conditions): for our T-cell data N=3000, K=47
- For each of N(N-1)/2 pairs of genes i and j calculate the correlation coefficient $\rho_{ij}=\sigma_{ij}/\sigma_i\sigma_j$ of gene levels across K samples
- Put a threshold, e.g. $\rho_{ij}>0.85$, or otherwise select the most correlated pairs of genes (~4500 in our case). Now you have a weighted network.
- Identify densely interconnected functional modules in this network.
- Modules can be used to infer unknown functions of genes via "Guilt by Association" principle.

How to install Gephi software for network analysis?

- Install Gephi from: https://gephi.org/users/download/
- One of the common problems with installation is the version of Java on your computer. One possible solution is here: https://github.com/gephi/gephi/issues/1787.

Sometimes after installation Gephi may complain that it cannot find java version 1.8 or higher. In this case you need to go to C:\Program Files\Gephi-0.9.2\etc
Open file gephi.conf using notepad.exe (MS Word does not work!).
Add a line jdkhome="C:\Program Files (x86)\Java\jre1.8.0_231"
(the numbers in …jre1.8.0_231 may be changed to reflect the actual directory where Java is installed on your computer).
If JDK is not installed on your computer, you need to install itfirst from https://www.java.com/en/download/win10.jsp"

Co-expression network analysis exercise

- Start Gephi and open coexpression_network_random_start.gephi
- Run “Layout” → Fruchterman Reingold → Speed 10.0
- Run “Average degree”, “Network diameter”, “Modularity” in the Statistics tab in the right panel.
- Color nodes by “modularity class”:
  Appearance → Nodes → Partition → Palette Icon → Modularity class
- Size nodes first by “degree”.
  Appearance → Nodes → Ranking → Multiple Circles Icon → Degree
  - If the nodes are too small, select “Min size”: 10 and “Max size”:80
  - Nodes in large tightly connected clusters have large degree
- Then size nodes by “betweenness-centrality”
  Appearance → Nodes → Ranking → Multiple Circles Icon → Betweenness-centrality
  - Large circles are “coordinator” genes connecting different co-expressed clusters to each other. Potentially biologically interesting

Disease-disease similarity network

- Based on the table summarizing **all current medical knowledge** of genes implicated in diseases:
  - Rows: 516 common human diseases
  - Columns: 25,000 human genes
  - Matrix element $D_{i\alpha}$ =1 if the gene α is known to be involved in the disease i . 0 – otherwise
- Constructed disease-disease similarity network:
  - Weight of the edge - # of shared genes between two diseases
  - Easy to construct: the adjacency matrix A of the network is simply $A=D \bullet D^{+}$

Disease network analysis exercise

- Start Gephi and open disease_disease_random_start.gephi
- Run "Layout" → Fruchterman Reingold → Speed 10.0
  Observe how clusters emerge.
- Run "Average degree", "Network diameter", "Modularity" analysis tools in the right panel.
- Color nodes with medical term: "disorder class"
  Appearance → Nodes → Partition → Palette Icon → Disorder class
- Then color nodes by "modularity class". See how well it agrees with the previous color.
  Appearance → Nodes → Partition → Palette Icon → Modularity class
- Size nodes first by "degree".
  Appearance → Nodes → Ranking → Multiple Circles Icon → Degree
  - Which disease has the largest degree?
- Size nodes by "betweenness centrality"
  Appearance → Nodes → Ranking → Multiple Circles Icon → Degree
  - Which diseases have the largest betweenness-centrality?

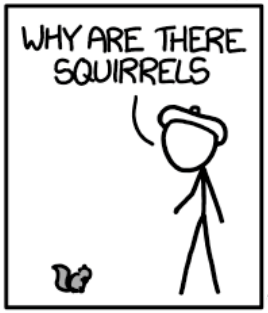
  These "connector" diseases linking different diseases clusters to each other. They highlight potentially interesting connections between diseases

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 DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE
WHY IS LYING GOOD

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 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 SEX SO IMPORTANT

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 AREN'T MY ARMS GROWING

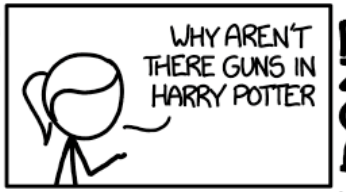
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

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 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 FEMALE MR MIMES

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 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 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 AREN'T THERE GUNS IN HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

Review for the Final Exam

Rules

- Closed book exam; no books, notes, laptops, smartphones, etc.
- However, calculators (not on a smartphone) can be used.
- You can prepare one cheat sheet (letter size, two-sided if needed)
- Printouts provided:
  - Distributions means/variances/pdfs
  - Standard normal distribution CDF table

Name	Probability Distribution	Mean	Variance	Section in Book
Discrete				
Uniform	$\frac{1}{n}, a \le b$	$\frac{(b+a)}{2}$	$\frac{(b-a+1)^2-1}{12}$	3-5
Binomial	$\binom{n}{x}p^x(1-p)^{n-x}$, $x = 0, 1, \ldots, n, 0 \le p \le 1$	np	$np(1-p)$	3-6
Geometric	$(1-p)^{x-1}p$, $x = 1, 2, \ldots, 0 \le p \le 1$	$1/p$	$(1-p)/p^2$	3-7.1
Negative binomial	$\binom{x-1}{r-1}(1-p)^{x-r}p^r$ $x = r, r+1, r+2, \ldots, 0 \le p \le 1$	r/p	$r(1-p)/p^2$	3-7.2

This will be provided

Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots, 0 < \lambda$	λ	λ	3-9
Continuous				
Uniform	$\frac{1}{b-a}, a \le x \le b$	$\frac{(b+a)}{2}$	$\frac{(b-a)^2}{12}$	4-5
Normal	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ $-\infty < x < \infty, -\infty < \mu < \infty, 0 < \sigma$	μ	σ^2	4-6
Exponential	$\lambda e^{-\lambda x}, 0 \le x, 0 < \lambda$	$1/\lambda$	$1/\lambda^2$	4-8
Erlang	$\frac{\lambda^r x^{r-1}e^{-\lambda x}}{(r-1)!}, 0 < x, r = 1, 2, \ldots$	r/λ	r/λ^2	4-9.1
Gamma	$\frac{\lambda^r x^{r-1}e^{-\lambda x}}{\Gamma(r)}, 0 < x, 0 < r, 0 < \lambda$	r/λ	r/λ^2	4-9.2

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511967	0.515953	0.519939	0.532922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555760	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621719	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.805106	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823815	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.878999	0.881000	0.882977
1.2	0.884930	0.886860	0.888767	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903199	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935744	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959071	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965621	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705
2.0	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613
2.5	0.993790	0.993963	0.994132	0.994297	0.994457	0.994614	0.994766	0.994915	0.995060	0.995201
2.6	0.995339	0.995473	0.995604	0.995731	0.995855	0.995975	0.996093	0.996207	0.996319	0.996427
2.7	0.996533	0.996636	0.996736	0.996833	0.996928	0.997020	0.997110	0.997197	0.997282	0.997365
2.8	0.997445	0.997523	0.997599	0.997673	0.997744	0.997814	0.997882	0.997948	0.998012	0.998074
2.9	0.998134	0.998193	0.998250	0.998305	0.998359	0.998411	0.998462	0.998511	0.998559	0.998605
3.0	0.998650	0.998694	0.998736	0.998777	0.998817	0.998856	0.998893	0.998930	0.998965	0.998999
3.1	0.999032	0.999065	0.999096	0.999126	0.999155	0.999184	0.999211	0.999238	0.999264	0.999289
3.2	0.999313	0.999336	0.999359	0.999381	0.999402	0.999423	0.999443	0.999462	0.999481	0.999499
3.3	0.999517	0.999533	0.999550	0.999566	0.999581	0.999596	0.999610	0.999624	0.999638	0.999650
3.4	0.999663	0.999675	0.999687	0.999698	0.999709	0.999720	0.999730	0.999740	0.999749	0.999758
3.5	0.999767	0.999776	0.999784	0.999792	0.999800	0.999807	0.999815	0.999821	0.999828	0.999835
3.6	0.999841	0.999847	0.999853	0.999858	0.999864	0.999869	0.999874	0.999879	0.999883	0.999888
3.7	0.999892	0.999896	0.999900	0.999904	0.999908	0.999912	0.999915	0.999918	0.999922	0.999925
3.8	0.999928	0.999931	0.999933	0.999936	0.999938	0.999941	0.999943	0.999946	0.999948	0.999950
3.9	0.999952	0.999954	0.999956	0.999958	0.999959	0.999961	0.999963	0.999964	0.999966	0.999967

What may be on the final exam?

- Probability Multiplication, Combinatorics
- Bayes Theorem
- Discrete & Continuous Random Variables
- Joint Probability Distributions, Covariation/Correlations
- Sampling distributions and parameter point estimation
- Confidence Intervals
- Hypothesis testing for one and two samples
- Other topics
- Look at Homework 1-5 for examples of problems

One-sample hypothesis testing

3. **(8 points)** The college bookstore tells prospective students that the average cost of its textbooks is $52 with a standard deviation of $4.50. A group of statistics students think that the average cost is **actually higher**. In order to test bookstore's claim against this alternative hypothesis, the students bought a random sample of 100 books. The mean price of this sample was $52.80. Perform the hypothesis test at the 5% level of significance and state your decision.

Two-sample hypothesis

Mating Calls. In a study of mating calls in the gray treefrogs *Hyla hrysoscelis* and *Hyla versicolor*, Gerhart (1994) reports that in a location in Lousiana the following data on the length of male advertisement calls have been collected:

	Sample size	Average duration	SD of duration	Duration range
Hyla chrysoscelis	43	0.65	0.18	0.36–1.27
Hyla versicolor	12	0.54	0.14	0.36–0.75

The two species cannot be distinguished by external morphology, but *H. chrysoscelis* are diploids while *H. versicolor* are tetraploids. The triploid crosses exhibit high mortality in larval stages, and if they attain sexual maturity, they are sterile. Females responding to the mating calls try to avoid mismatches.

Based on the data summaries provided, test whether the length of call is a discriminatory characteristic? Use $\alpha = 0.05$.

	Sample size	Average duration	SD of duration
Hyla chrysoscelis	43	0.65	0.18
Hyla versicolor	12	0.54	0.14

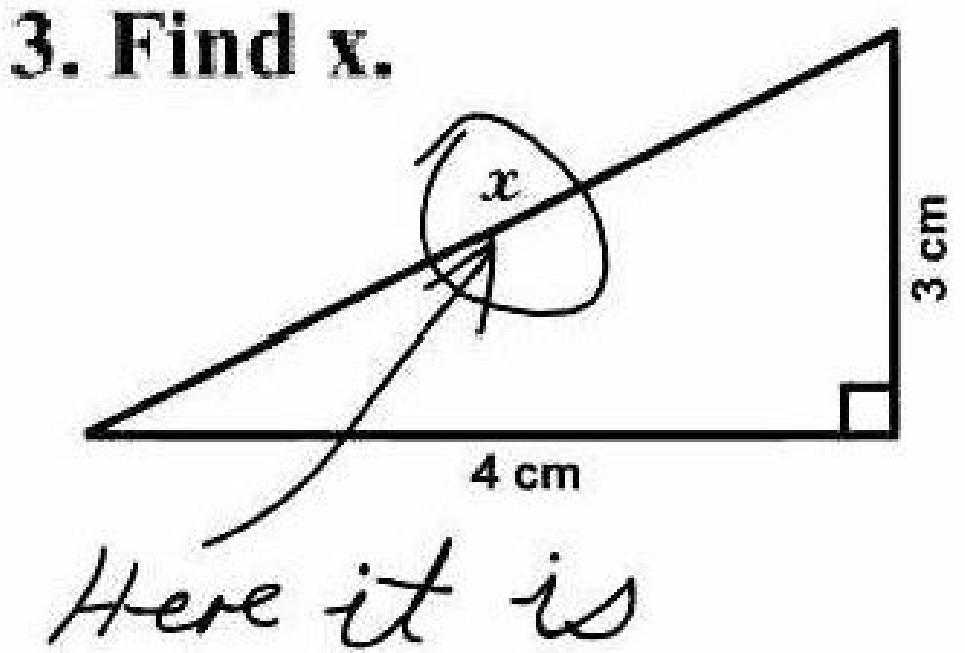
Based on the data summaries provided, test whether the length of call is a discriminatory characteristic? Use $\alpha = 0.05$.

Confidence intervals

2. **(6 points)** The operations manager of a large production plant would like to estimate the mean amount of time a worker takes to assemble a new electronic component. Assume that the standard deviation of this assembly time is 3.6 minutes. After observing a sample of 100 workers assembling similar devices, the manager noticed that their average time was 16.2 minutes. Construct a **90% confidence interval** for the population mean of the assembly time.

What is X in this problem?

- What is X? Look for keywords:
  - Find the probability that….
  - What is the mean (or variance) of…
- What are the parameters?
  Look for keywords:
  - Given that…
  - Assuming that…
- **Is X discrete or continuous?**



Discrete Probability Distributions

(8 points) You are doing a long series of experiments. Assume that each of your experiments has a probability of 0.02 of succeeding. Assume that your experiments are independent.

(A) (2 points) What is the probability that you first succeed on tenth experiment?

(B) (2 points) What is the probability that it requires more than five experiments for you to succeed?

(C) (2 points) What is the mean number of experiments needed to succeed once?

(D) (2 points) What is the probability that the second experiment that worked is the tenth one since you started?

Continuous Probability Distributions

(12 points) Time interval separating subsequent bus arrivals at a stop is an exponential random variable with mean 20 minutes. Steve and Andrew work at the same place and each will be late to work unless they board a bus on or before 8:40am. Steve comes to the bus stop exactly at 8am. Andrew also comes to the same bus stop but at a random time, uniformly distributed between 8am and 8:30am. Both of them take the first bus that arrives.

(a) (4 points) What is the probability that Steve will be late for work tomorrow?

(b) **(4 points)** What is the probability that Andrew will be late for work tomorrow?

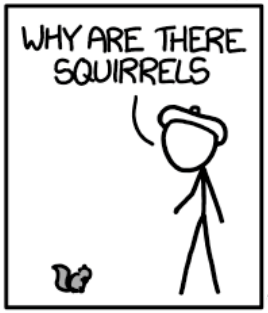
(c) **(4 points)** What is the probability that Steve and Andrew will ride the same bus

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 DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS
WHY IS THERE HELL IF GOD FORGIVES
WHY IS GPS FREE
WHY IS LYING GOOD

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 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 SEX SO IMPORTANT

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 AREN'T MY ARMS GROWING

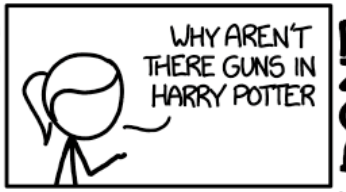
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS

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 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 FEMALE MR MIMES

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 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 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

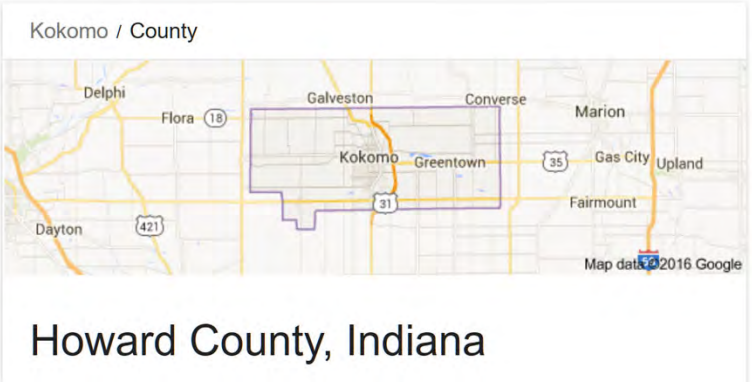
WHY AREN'T THERE GUNS IN HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

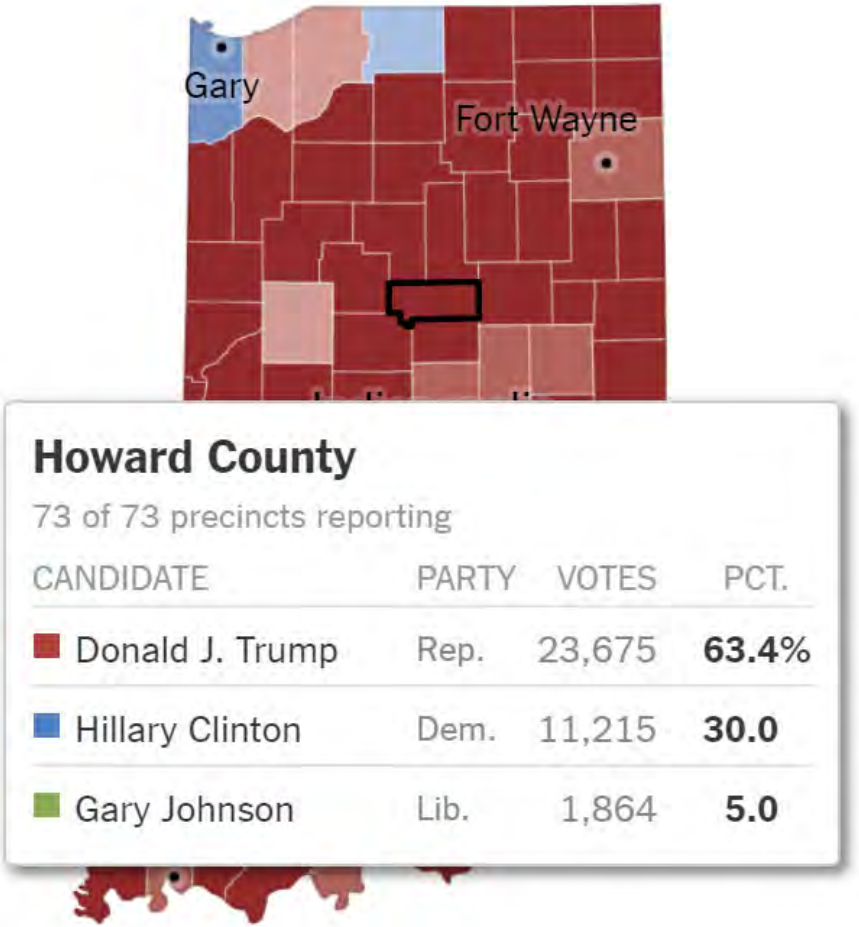
Bayes theorem

Kokomo, Indiana. In Kokomo, IN, 65% of the people are conservative, 20% are liberal, and 15% are independent. Records show that in a particular election, 82% of conservatives voted, 65% of liberals voted, and 50% of independents voted. If a person from the city is selected at random and it is learned that she did not vote, what is the probability that the person is liberal?



Howard County, Indiana

As of the 2010 census, the population was 82,752. The county seat is Kokomo, IN.



Kokomo, Indiana. In Kokomo, IN, 65% of the people are conservative, 20% are liberal, and 15% are independent. Records show that in a particular election, 82% of conservatives voted, 65% of liberals voted, and 50% of independents voted. If a person from the city is selected at random and it is learned that she did not vote, what is the probability that the person is liberal?

Joint Probability Distributions

1. **(20 points)** The joint probability mass function of discrete random variables X and Y taking values x = 1, 2, 3 and y = 1, 2, 3, respectively, is given by $f_{XY}(x, y) = c^*(x + y)$. Determine the following:

a) **(2 points)** Find c
b) **(2 points)** Find probability of the event, where X = 1 and Y < 3
c) **(2 points)** Find marginal probability $P_Y(Y = 2)$
d) **(2 points)** Find marginal probability distribution of the random variable X
e) **(2 points)** Find E(X), E(Y), V(X), and V(Y)
f) **(2 points)** Find conditional probability distribution of Y given that X = 1
g) **(2 points)** Conditional probability distribution of X given that Y = 2
h) **(2 points)** Are X and Y independent?
i) **(2 points)** What is the covariance for X and Y?
j) **(2 points)** What is the correlation for X and Y?

1. **(20 points)** The joint probability mass function of discrete random variables X and Y taking values x = 1, 2, 3 and y = 1, 2, 3, respectively, is given by $f_{XY}(x, y) = c*(x + y)$. Determine the following:

a) **(2 points)** Find c

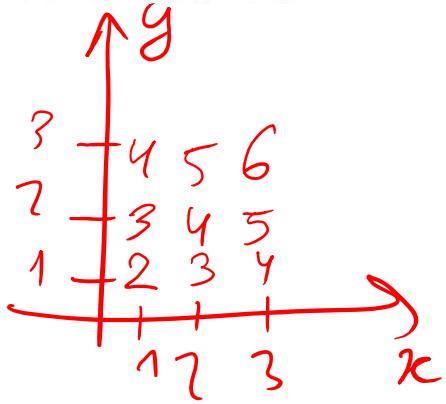
b) **(2 points)** Find probability of the event, where X = 1 and Y < 3

c) **(2 points)** Find marginal probability $P_Y(Y = 2)$

f) **(2 points)** Find conditional probability distribution of Y given that X = 1

(a) $1 = c \cdot (2 + 3 + 4 + 3 + 4 + 5 + 4 + 5 + 6)$

$c = 1/36$



$\times c$

(b) $P(X=1, Y<3) = \frac{2+3}{36} = \frac{5}{36}$

(c) $P_Y(Y=2) = \frac{3+4+5}{36} = \frac{12}{36} = \frac{1}{3}$

(f) $P(Y=2 \mid X=1) = \frac{P(Y=2, X=1)}{P_X(X=1)} = \frac{3/36}{(2+3+4)/36} = \frac{1}{3}$

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 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 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 SEX SO IMPORTANT

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 AREN'T MY ARMS GROWING

WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY IS YKK ON ALL ZIPPERS
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
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 ARE THERE SO MANY CROWS IN ROCHESTER, MN
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 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 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 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 AREN'T THERE GUNS IN HARRY POTTER

WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY IS LIFE SO BORING
WHY IS GPS FREE
WHY IS THERE FEMALE MR MIMES
WHY DO IGUANAS DIE
WHY IS THERE HELL IF GOD FORGIVES