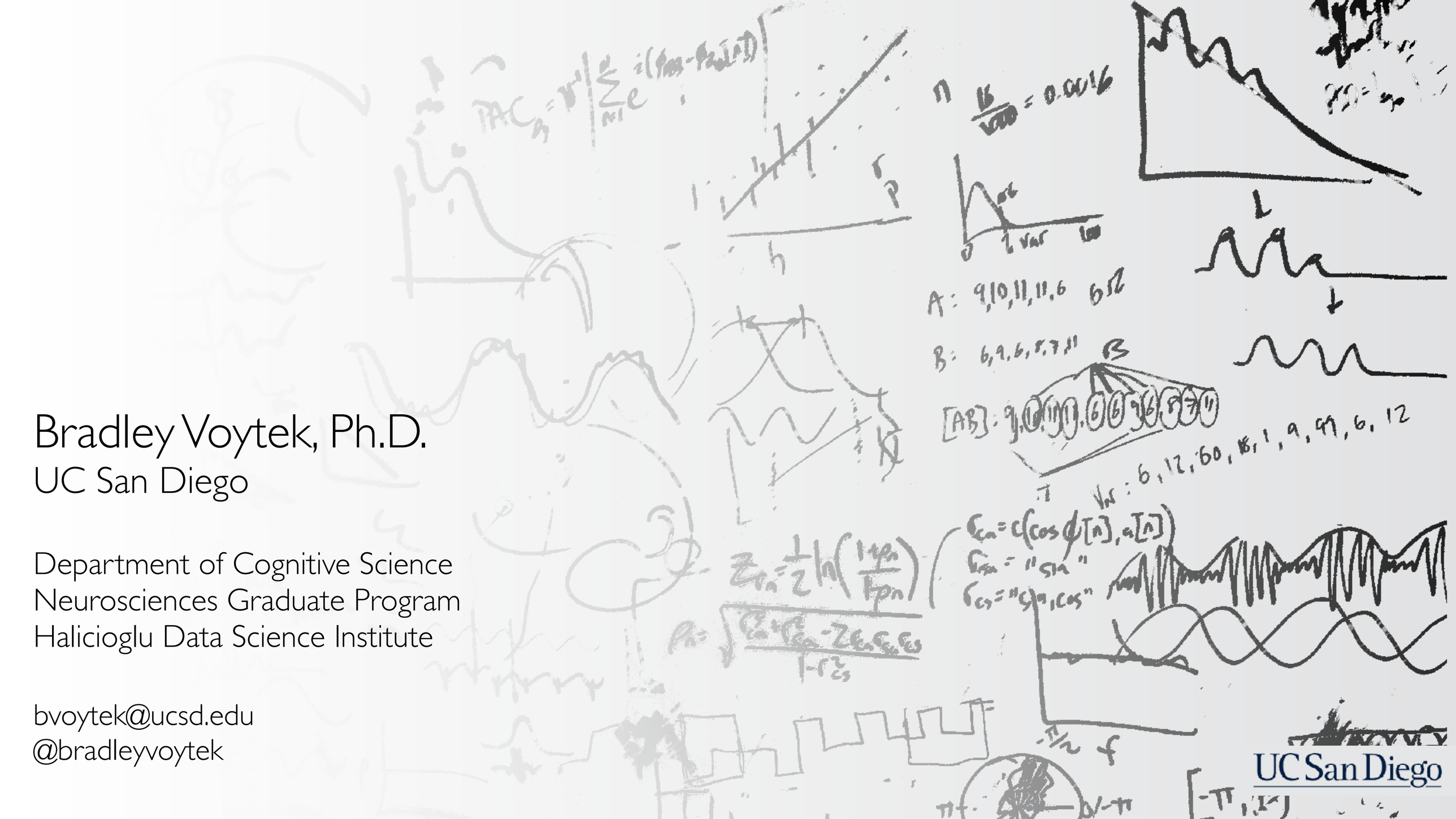


Bradley Voytek, Ph.D.

UC San Diego

Department of Cognitive Science
Neurosciences Graduate Program
Halicioglu Data Science Institute

bvoytek@ucsd.edu
@bradleyvoytek

UC San Diego

Fermi Estimation

What is the average height of your group (cm)?

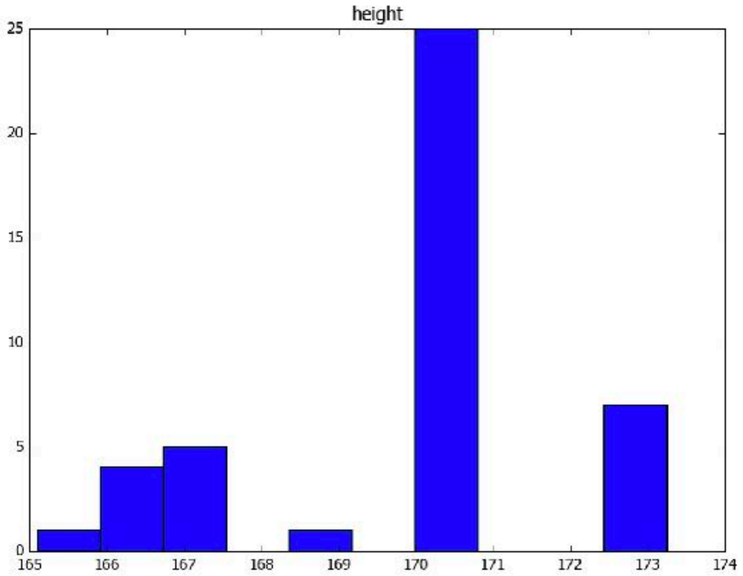
Short answer text

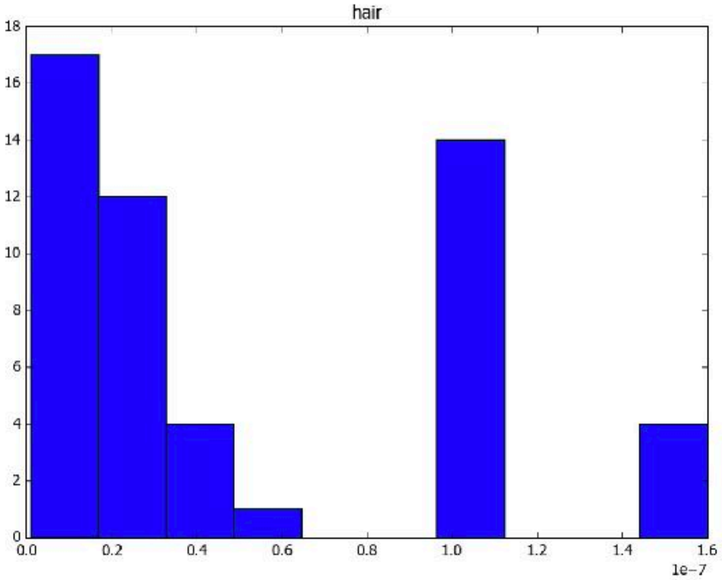
How fast does human hair grow (km/h)?

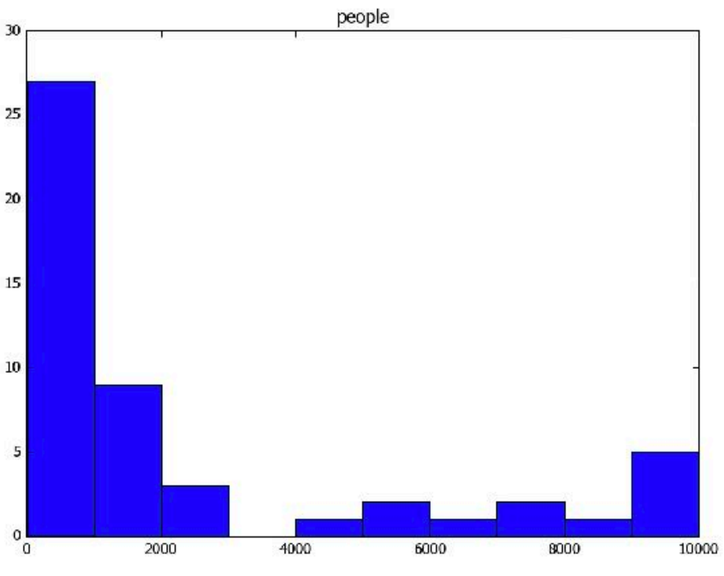
Short answer text

If every living person stood crammed together side-by-side, how large of an area would they occupy (km^2)?

Fermi Estimation







Fermi Estimation

What is the average height of your group (cm)?

Short answer text

How fast does human hair grow (km/h)?

Short answer text

If every living person stood crammed together side-by-side, how large of an area would they occupy (km^2)?

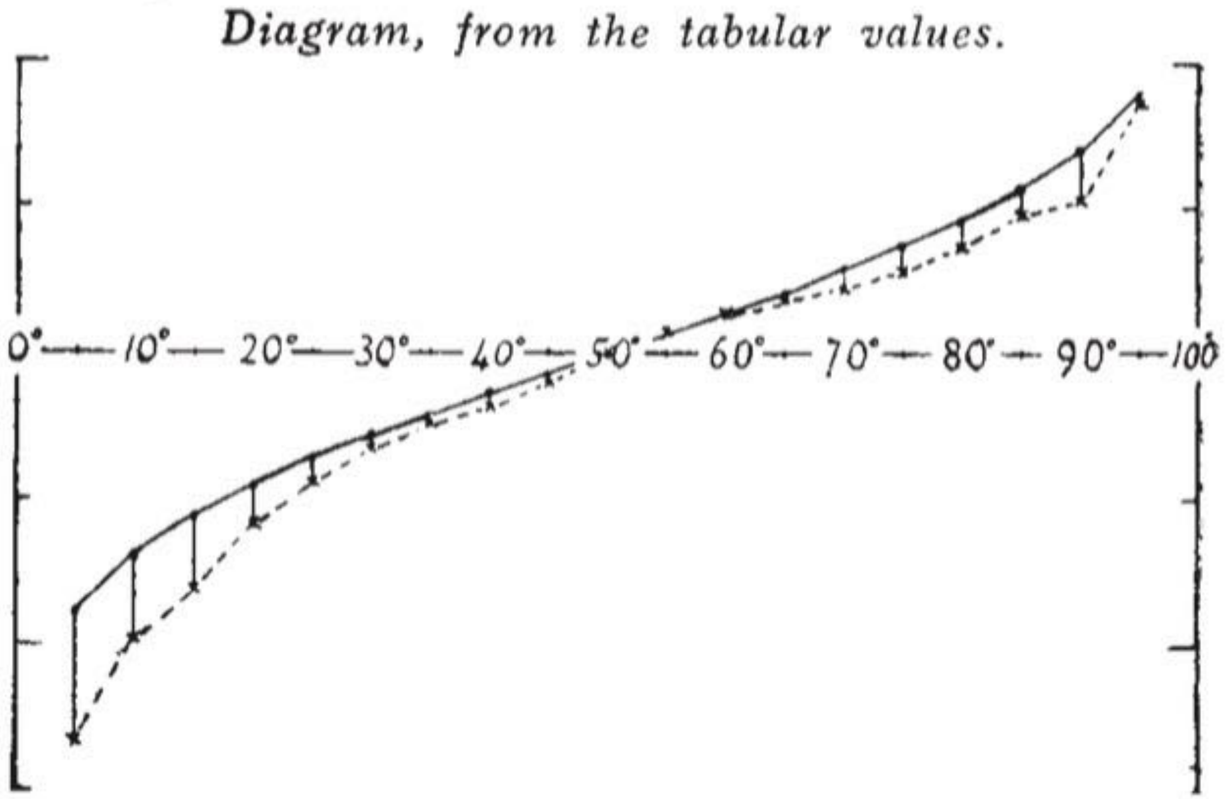
```
np.median(height), np.median(hair), np.median(people)
```

```
(170.0, 2.4899999999999998e-08, 1000.0)
```

hair: 1.7e-08

humans: 1000 km^2

Vox Populi



Middlemost estimate: 1207 lbs
True weight: 1198 lbs (0.8% under)

787 estimates

Source: Galton, *Nature* 1904

The Wisdom of Crowds

- Diversity of opinion: Each person should have private information even if it's just an eccentric interpretation of the known facts.

- Independence: People's opinions aren't determined by the opinions of those around them.

- Decentralization: People are able to specialize and draw on local knowledge.

- Aggregation: Some mechanism exists for turning private judgments into a collective decision.

Source: Surowiecki, *The Wisdom of Crowds* 2004

COGS 108
Data Science in Practice

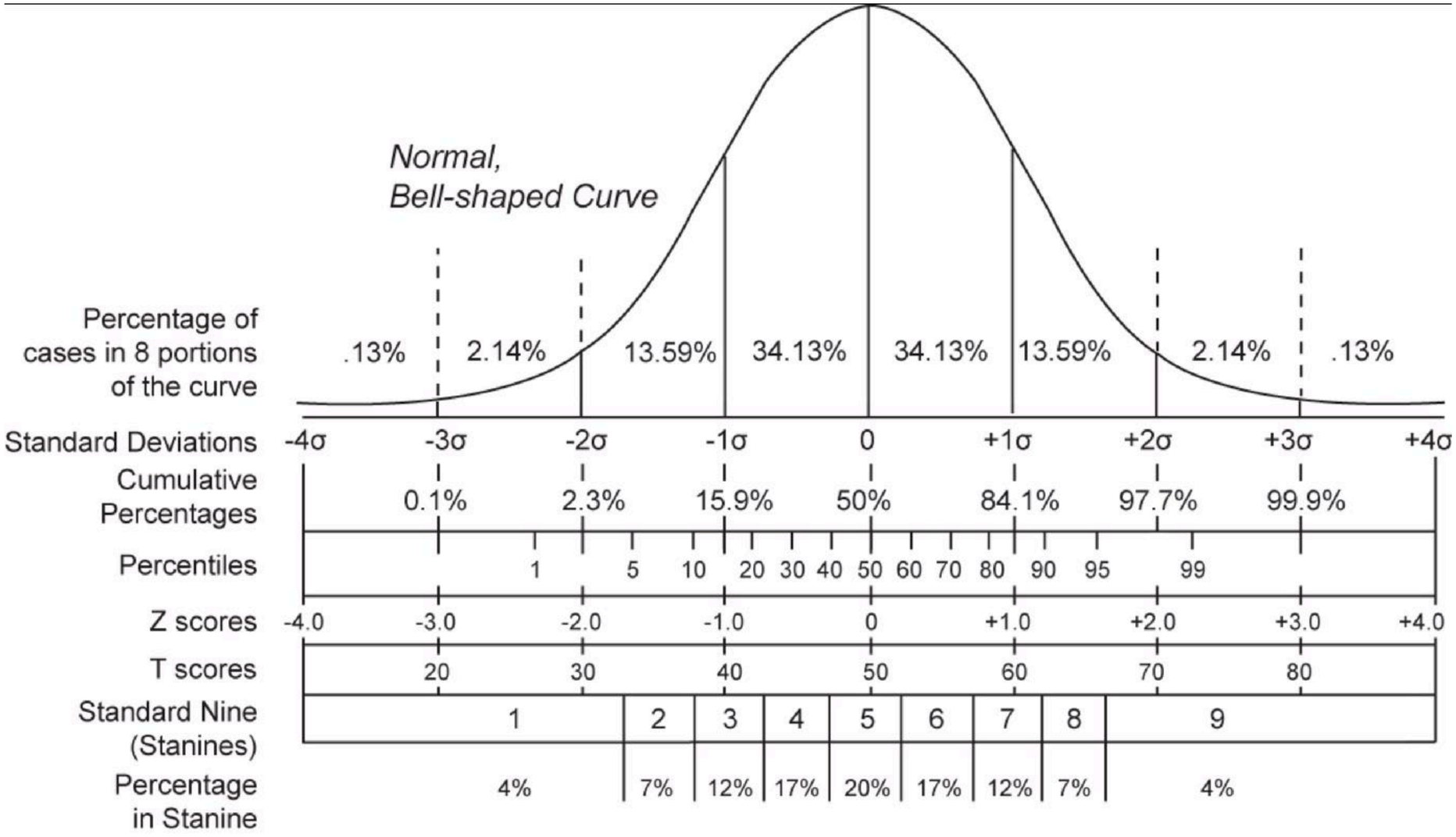
Linear modeling

UC San Diego

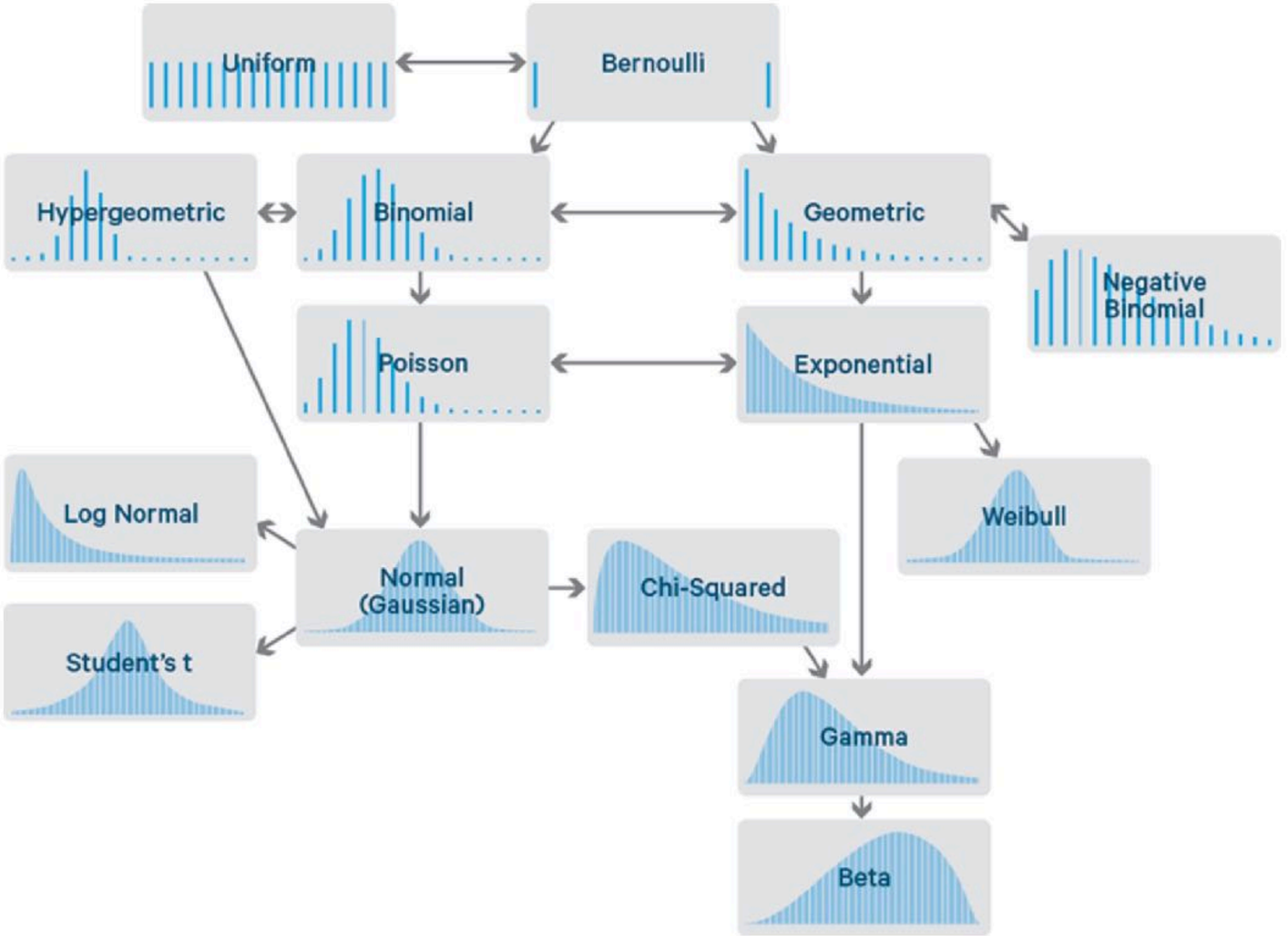
COGS 108
Data Science in Practice

~~*Linear modeling*~~
Intro to statistics

Distributions

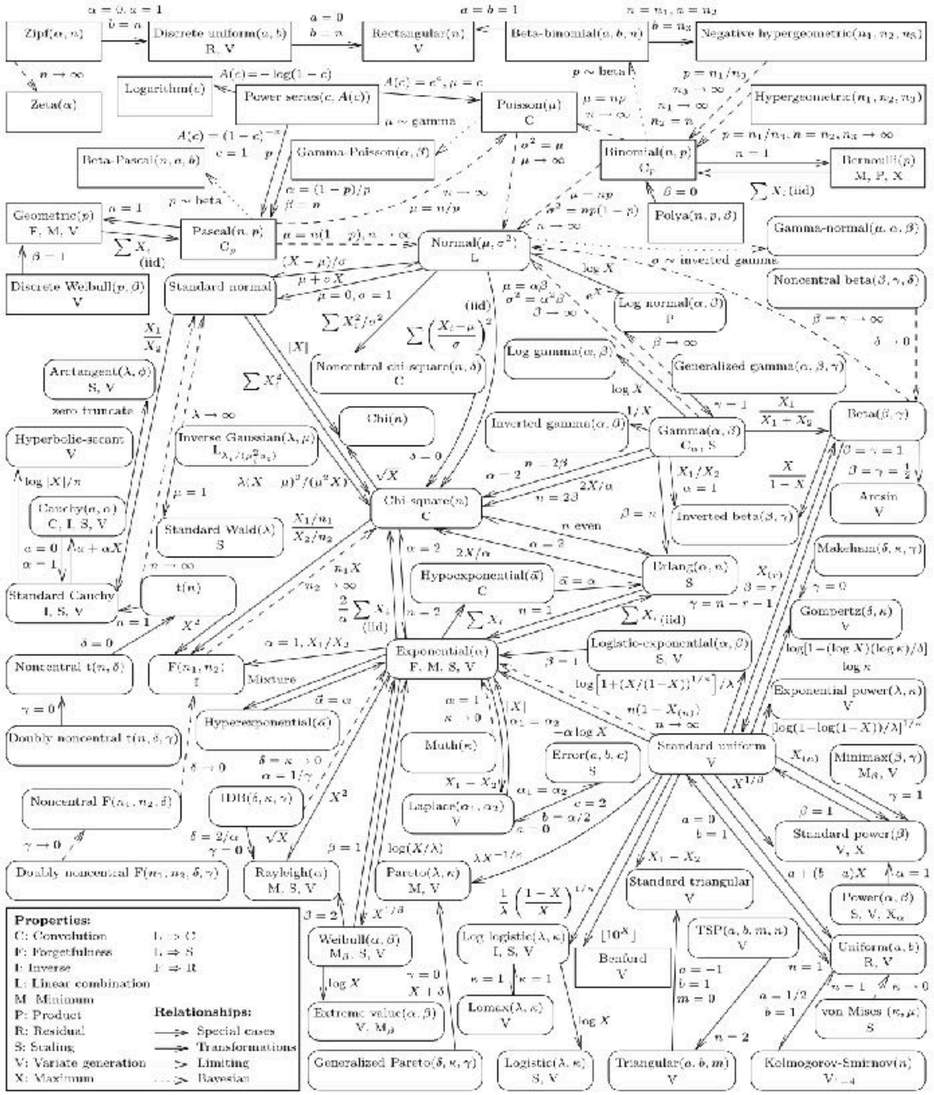


Distributions



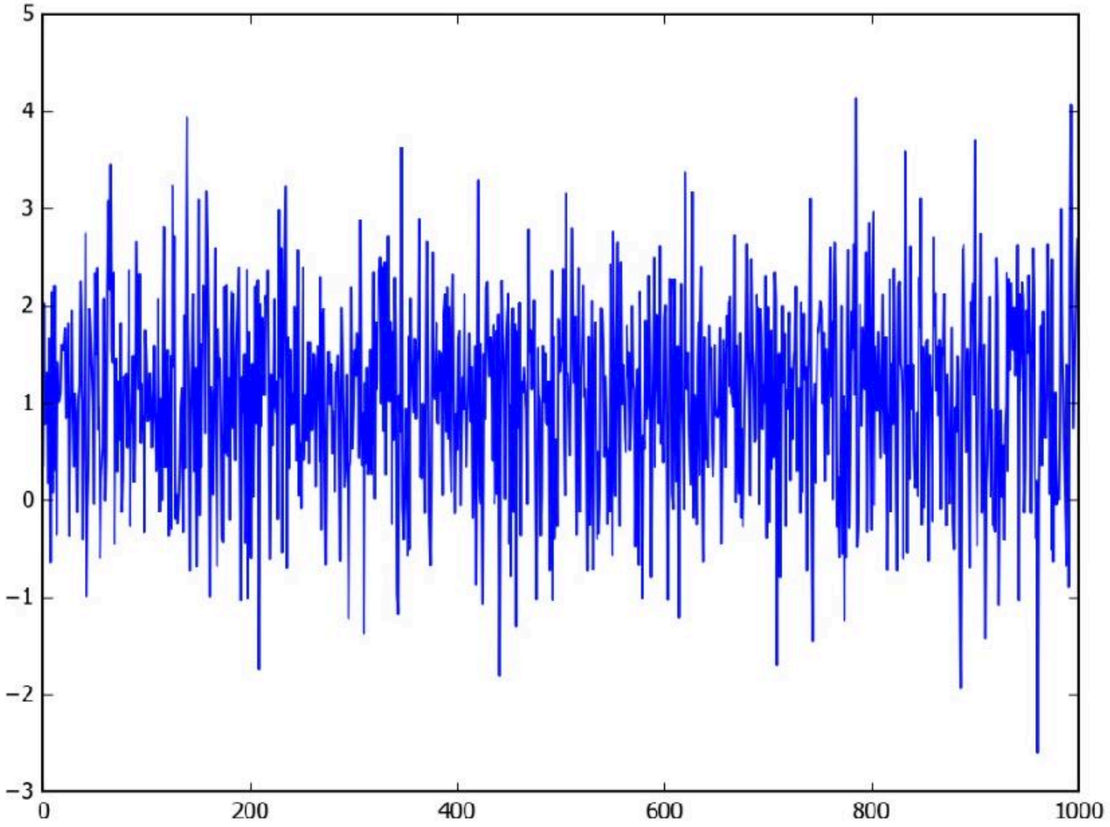
Source: *Cloudera blog*: http://blog.cloudera.com/blog/2015/12/common-probability-distributions-the-data-scientists-crib-sheet/

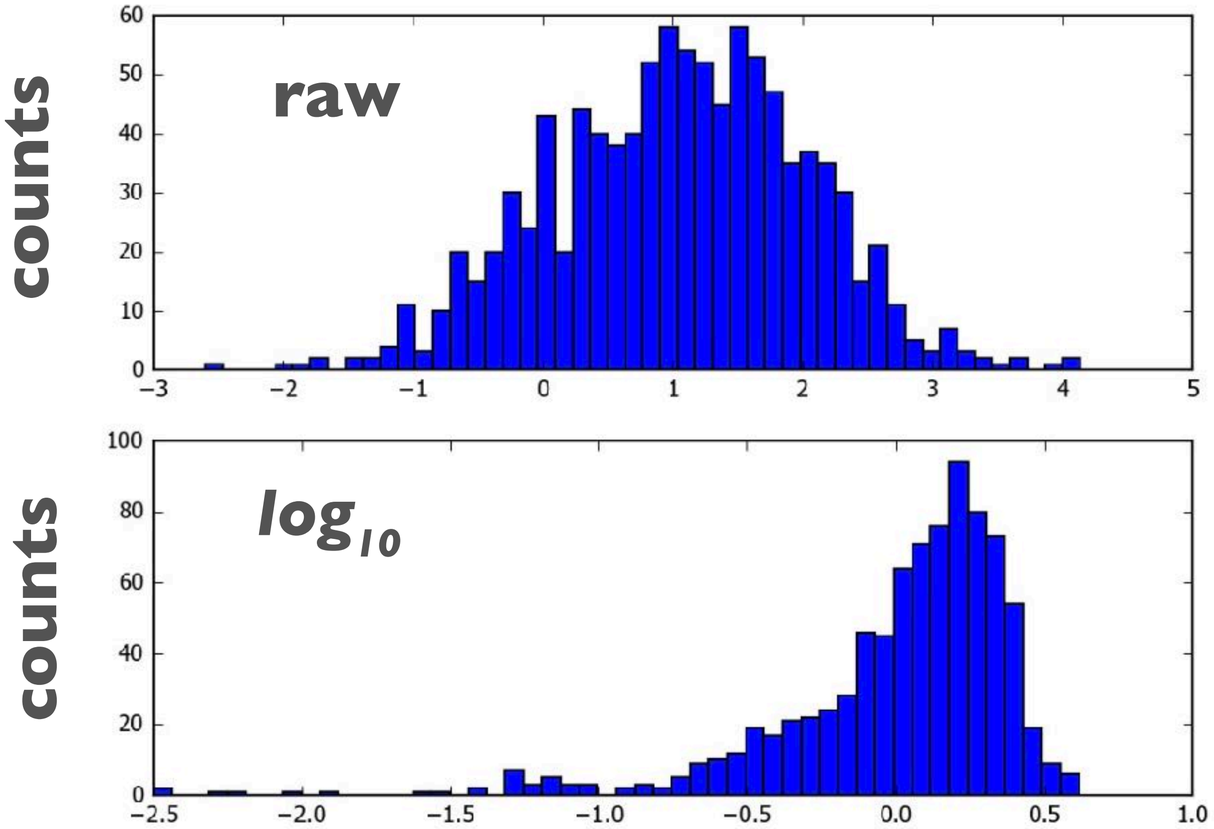
Distributions



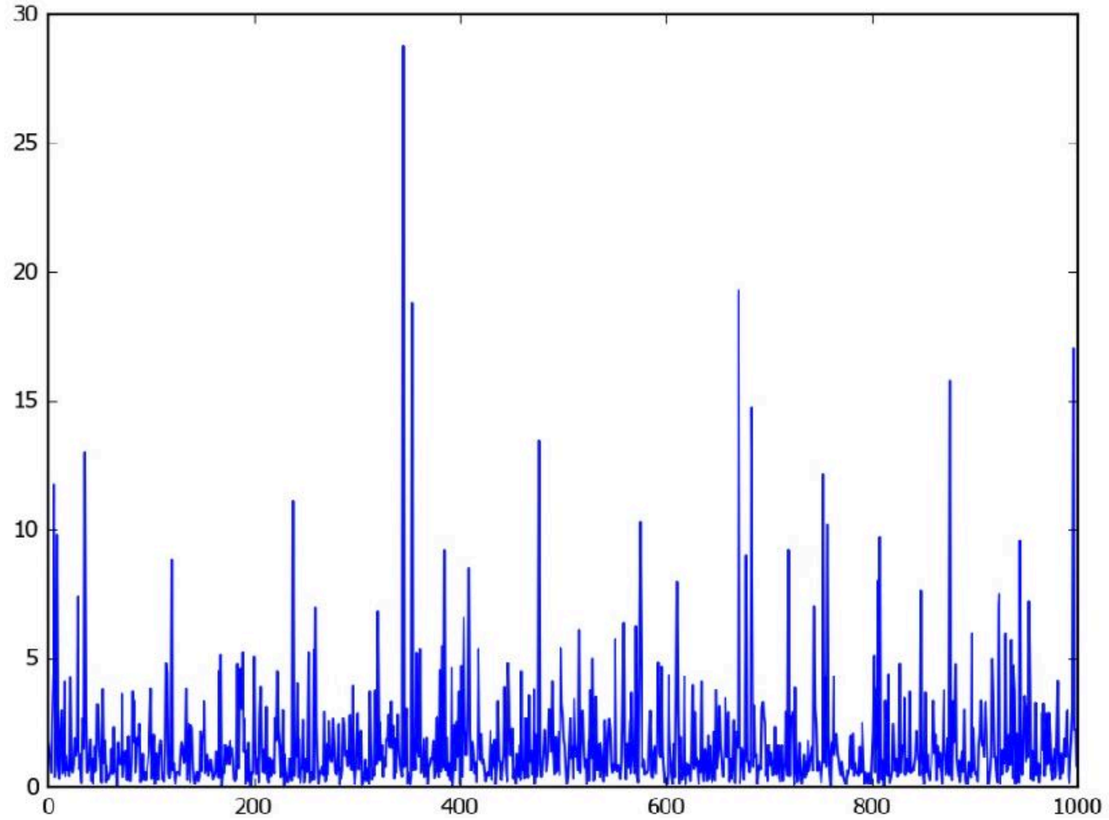
Source: Leemis & McQueston, *American Statistician* 2008

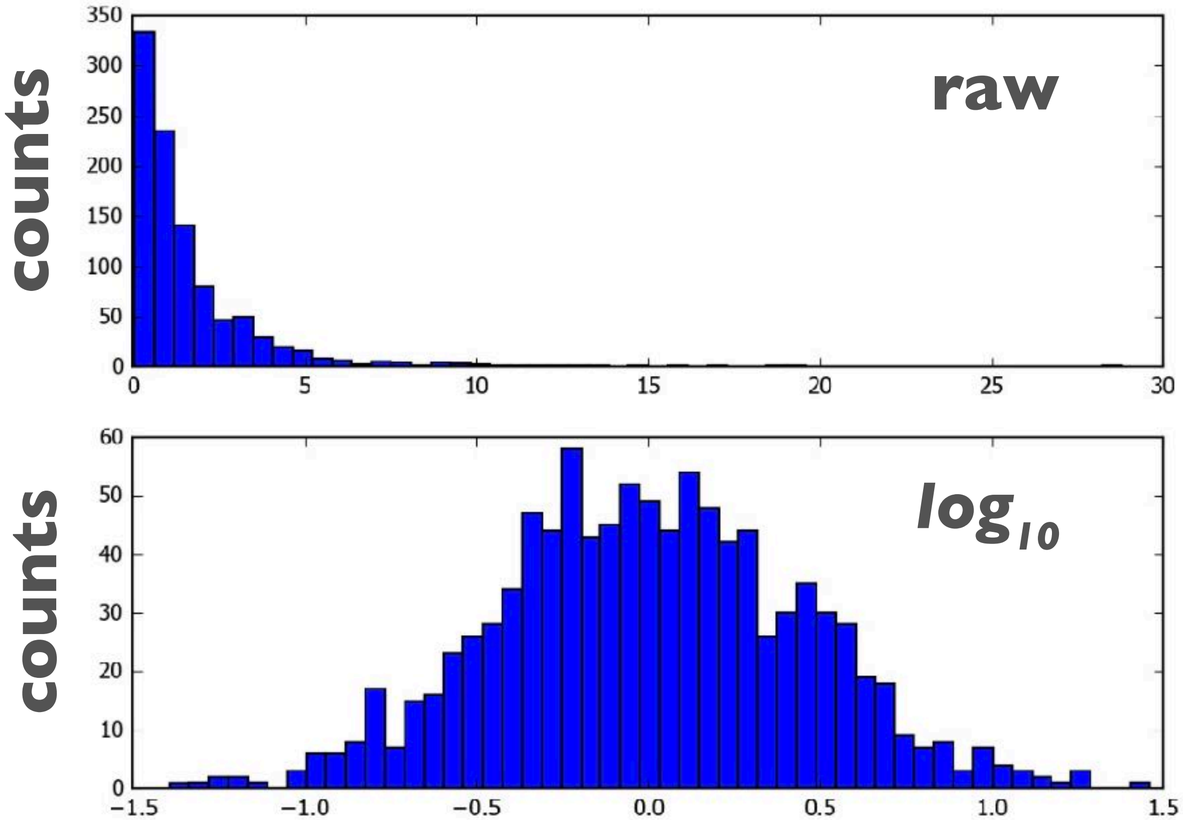
Normal


5
4
3
2
1
0
−1
−2
−3
0
200
400
600
800
1000

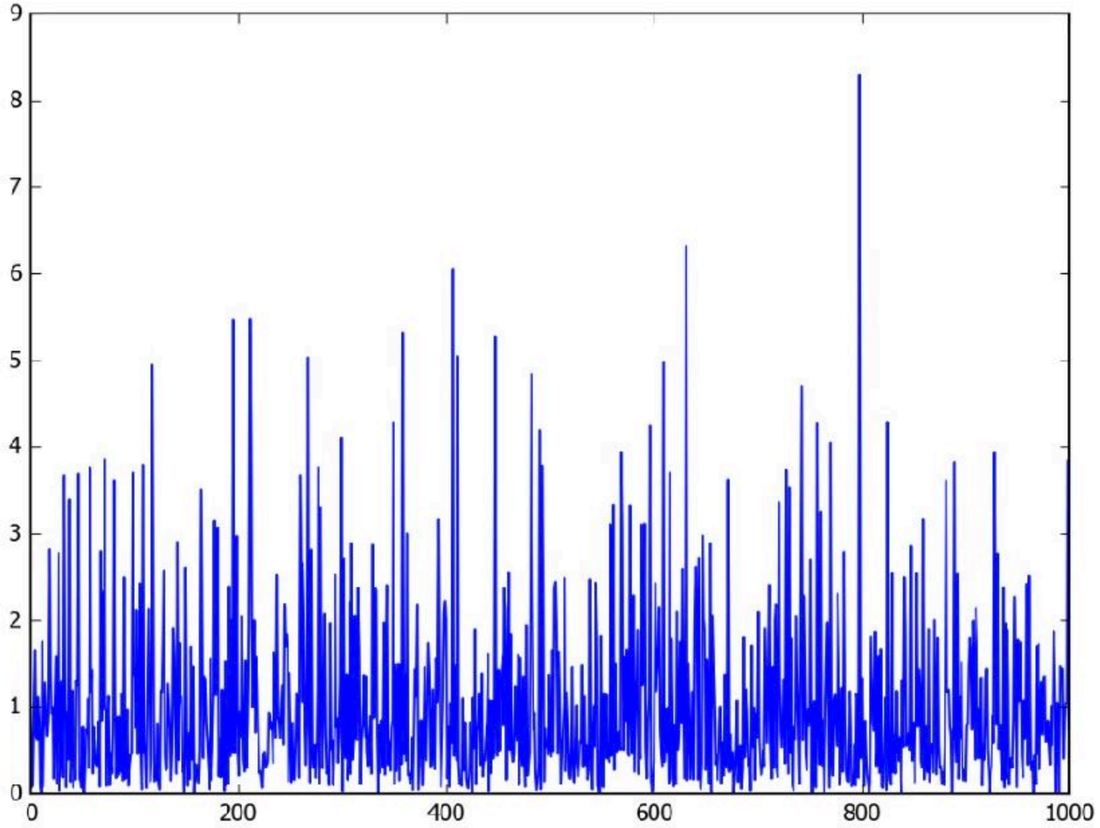

counts
raw
60
50
40
30
20
10
0
−3
−2
−1
0
1
2
3
4
5
counts
log_{10}
100
80
60
40
20
0
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0

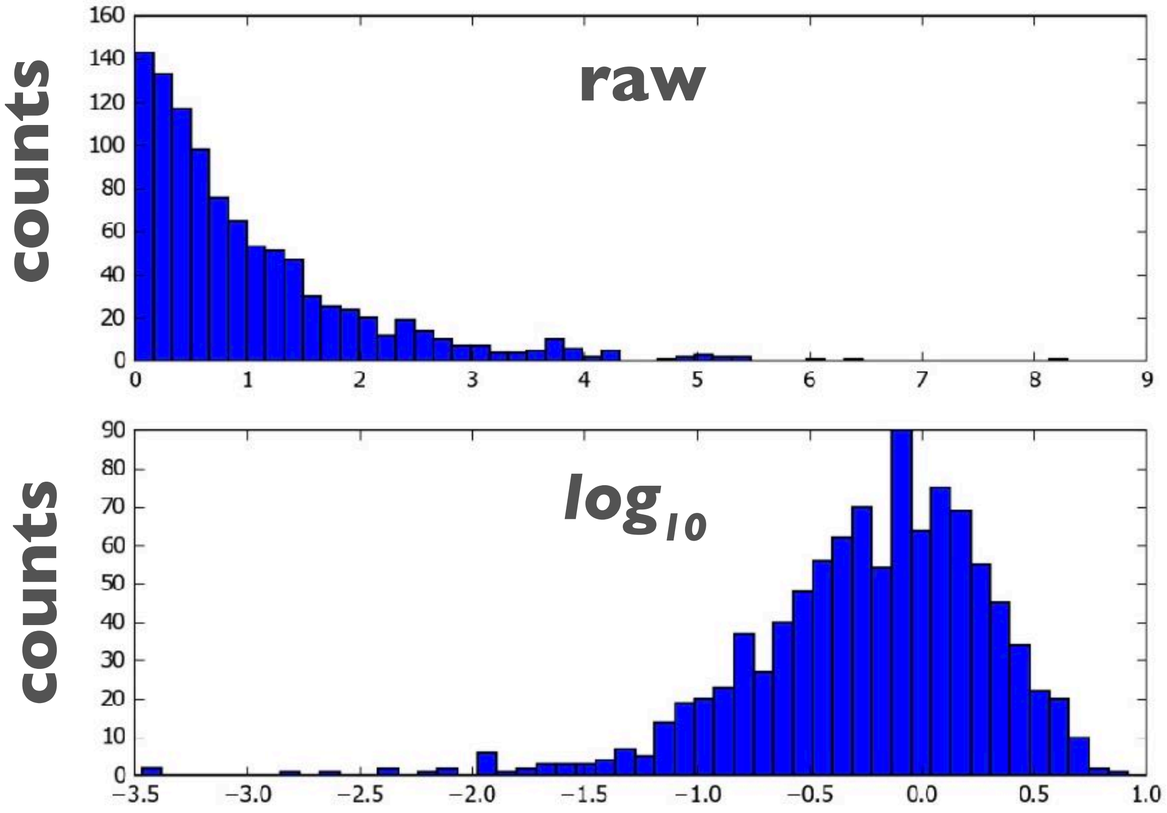
Lognormal


30
25
20
15
10
5
0
0
200
400
600
800
1000

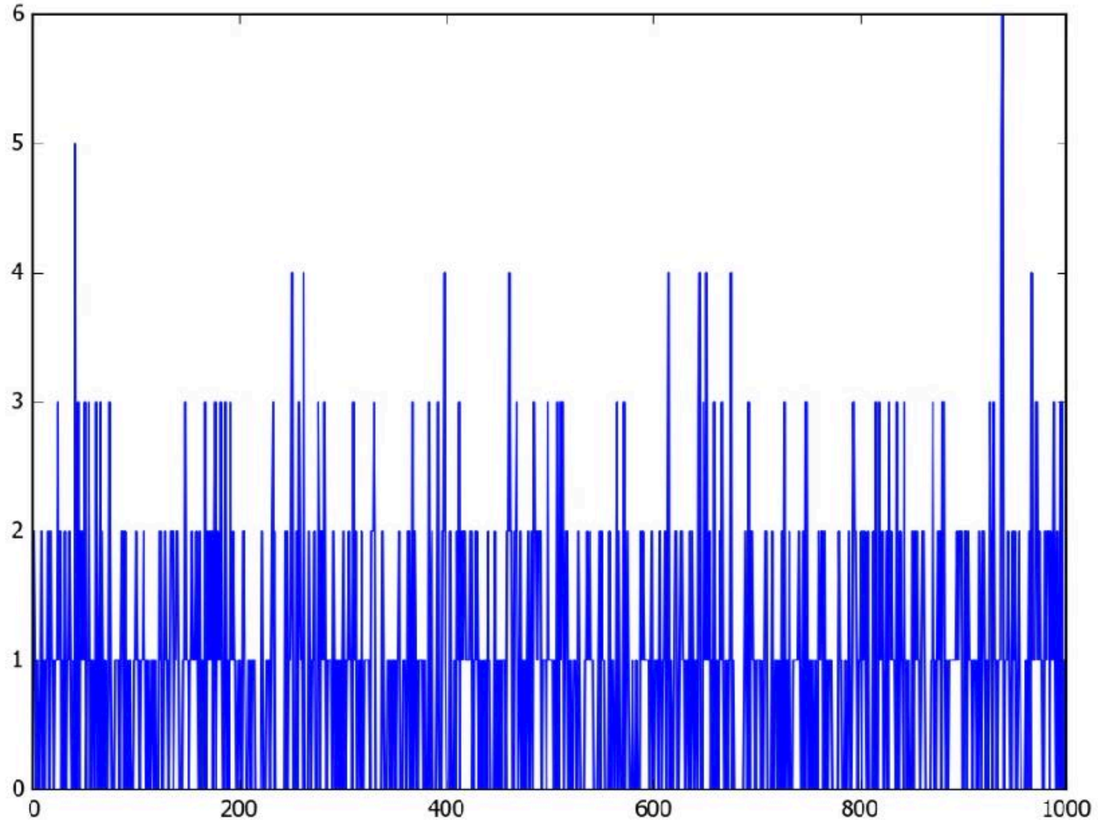

counts
raw
350
300
250
200
150
100
50
0
0
5
10
15
20
25
30
counts
log10
60
50
40
30
20
10
0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

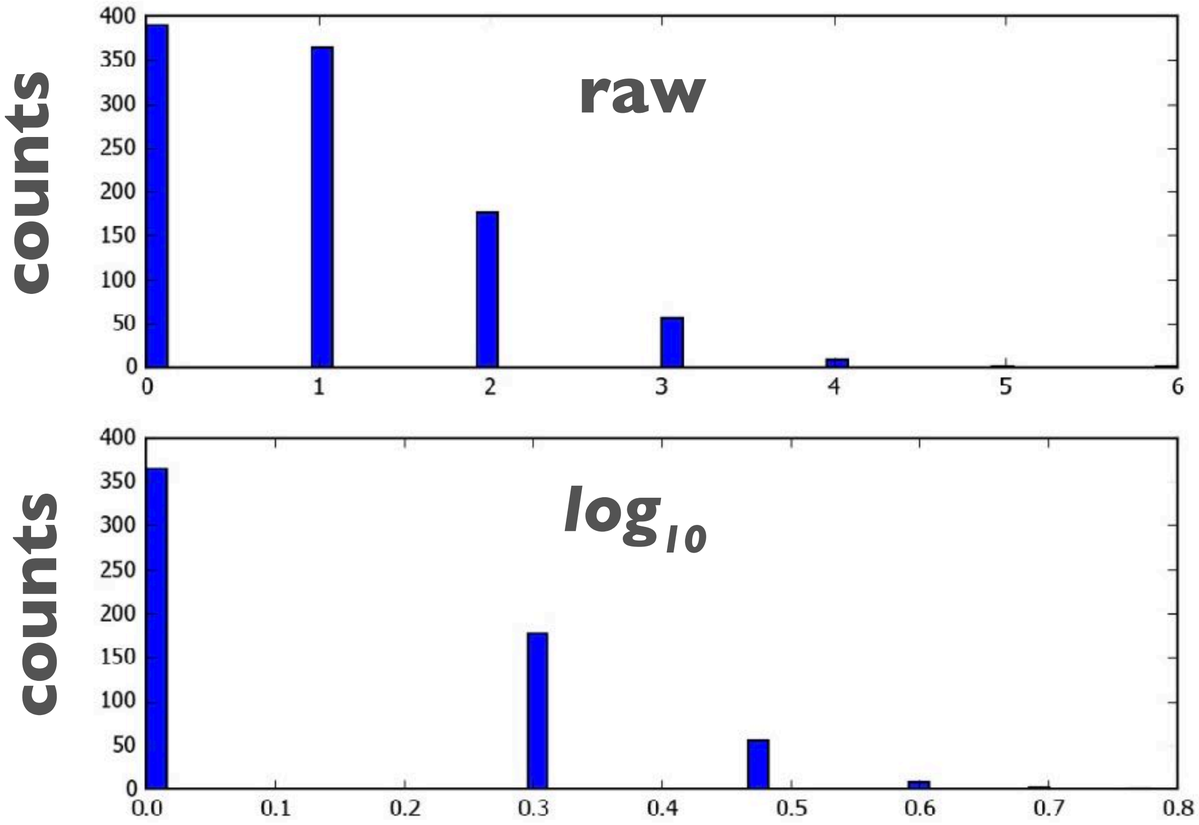
Exponential



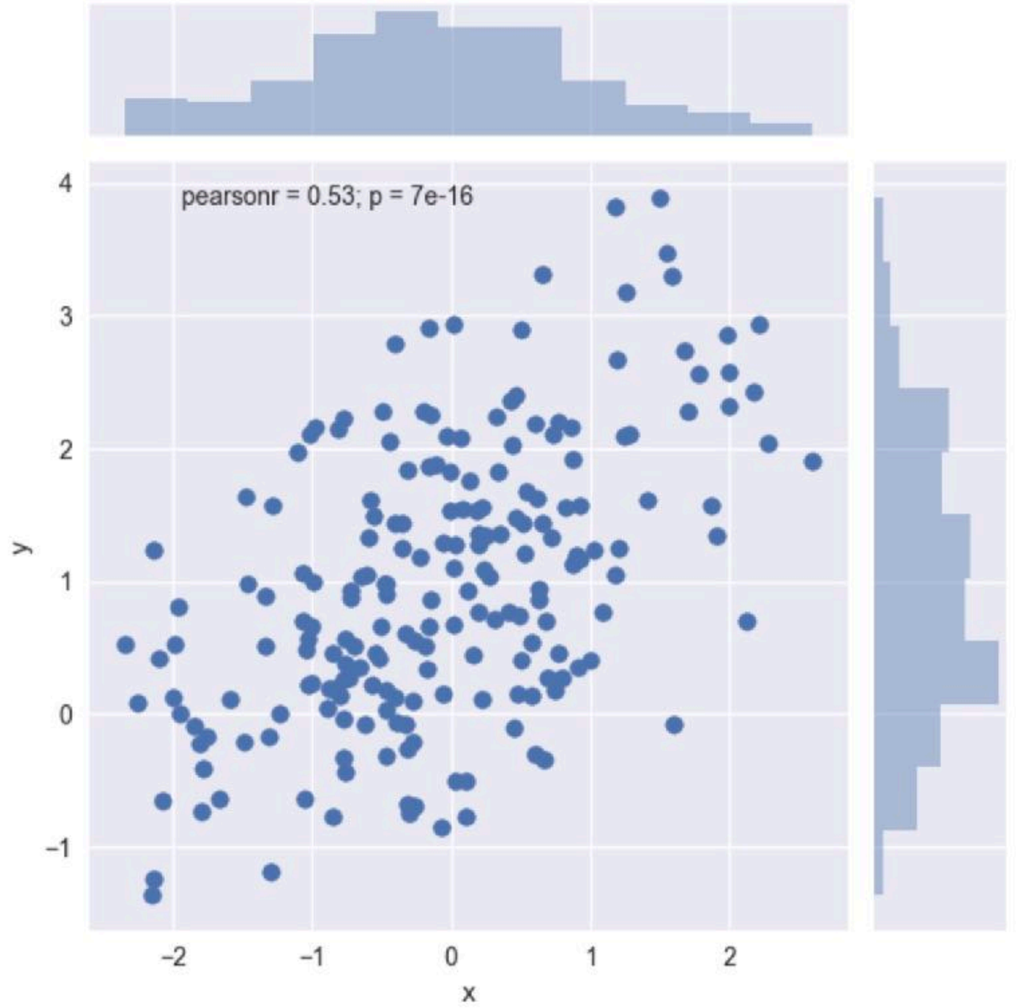


Poisson

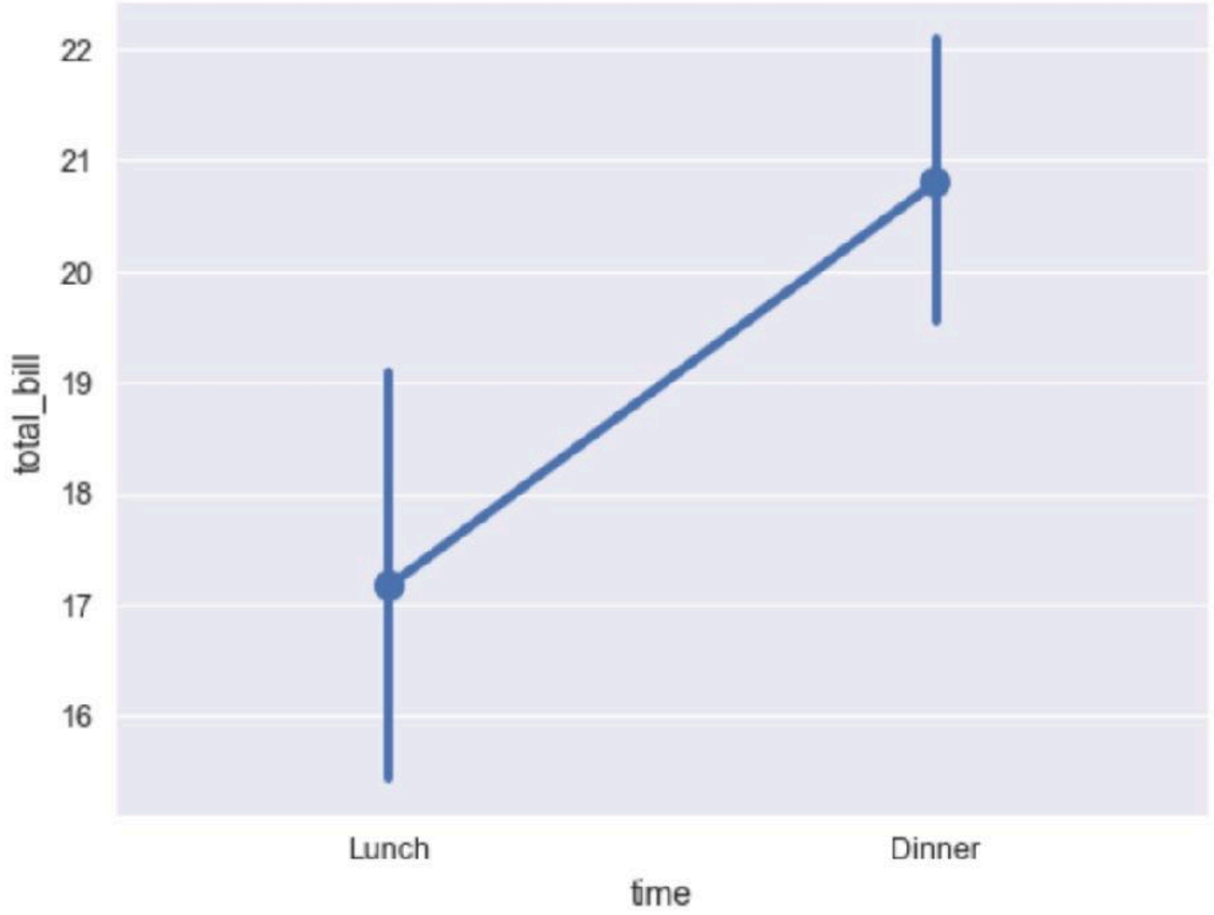

0
1
2
3
4
5
6
0
200
400
600
800
1000


counts
raw
0
50
100
150
200
250
300
350
400
0
1
2
3
4
5
6
counts
log10
0
50
100
150
200
250
300
350
400
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

Data distributions



Data distributions



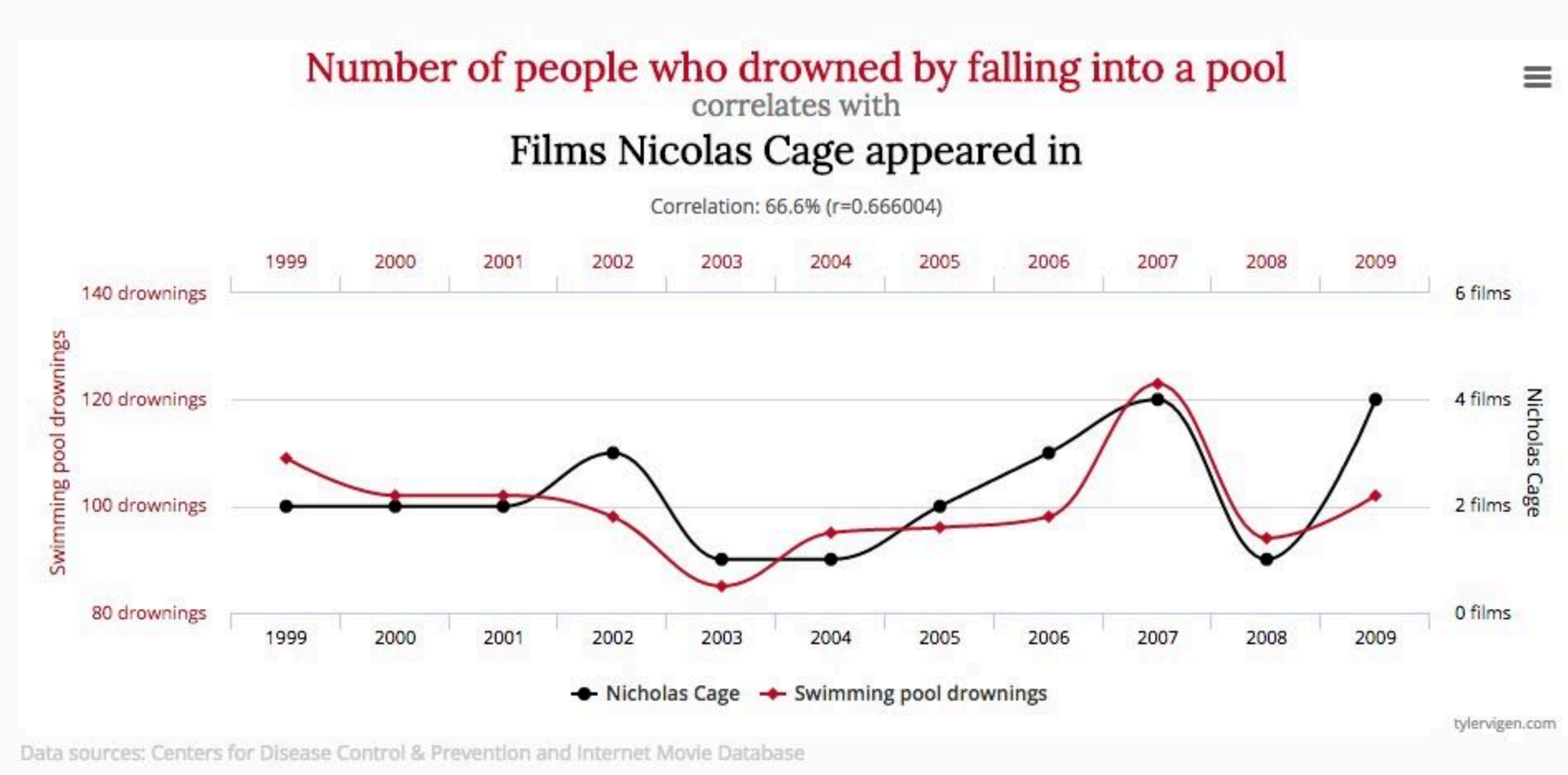
Student's t-test

Student's t-test

…brewing a batch of beer is a time consuming and expensive process, so in order to draw conclusions from experimental methods applied to just a few batches, Gosset had to figure out the role of chance in determining how a batch of beer had turned out. Guinness frowned upon academic publications, so Gosset had to publish his results under the modest pseudonym, "Student."

Source: Stanton, *An Introduction to Data Science*

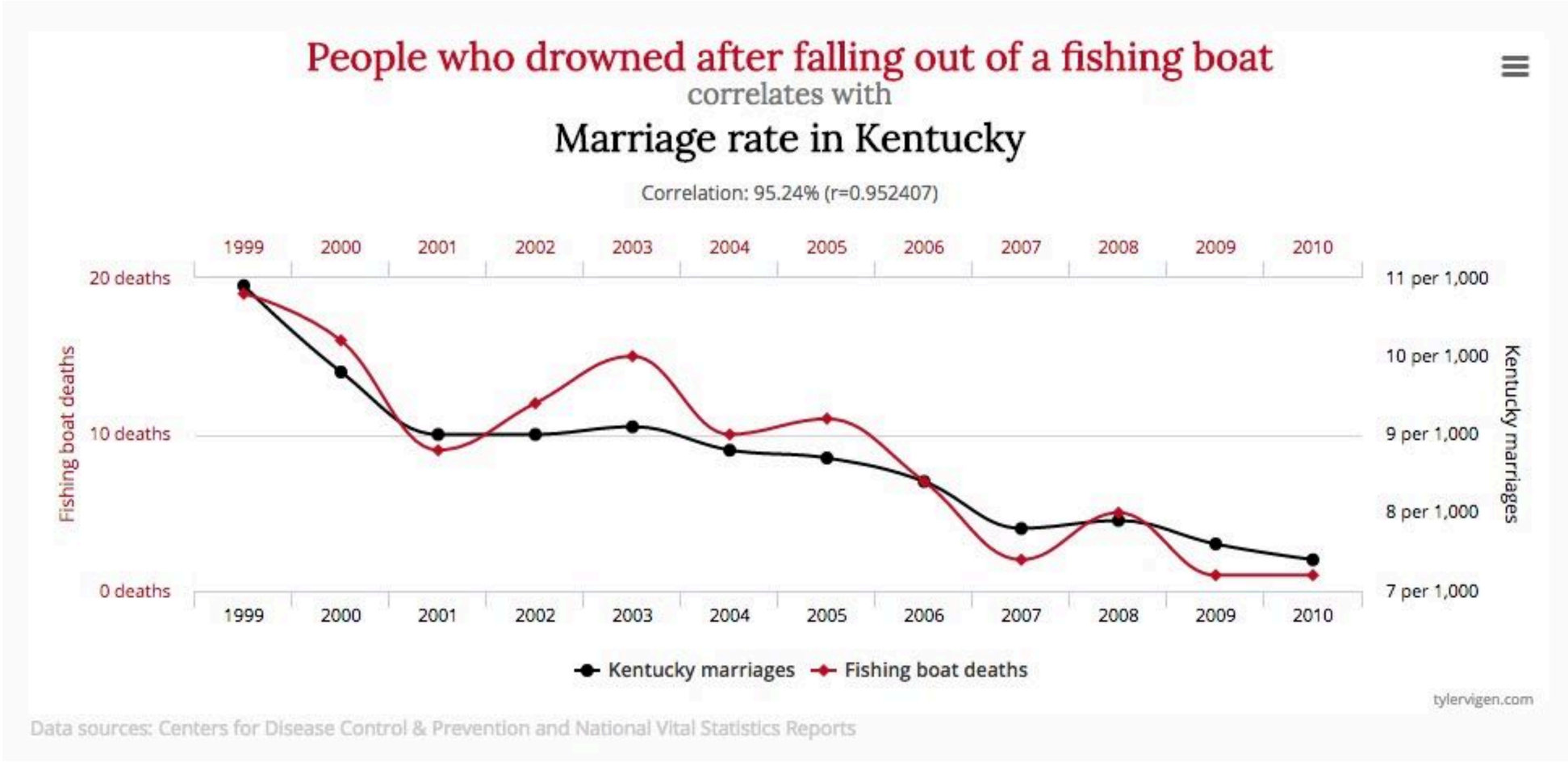
Don’t trust your data (or narratives)



Source: http://www.tylervigen.com/spurious-correlations

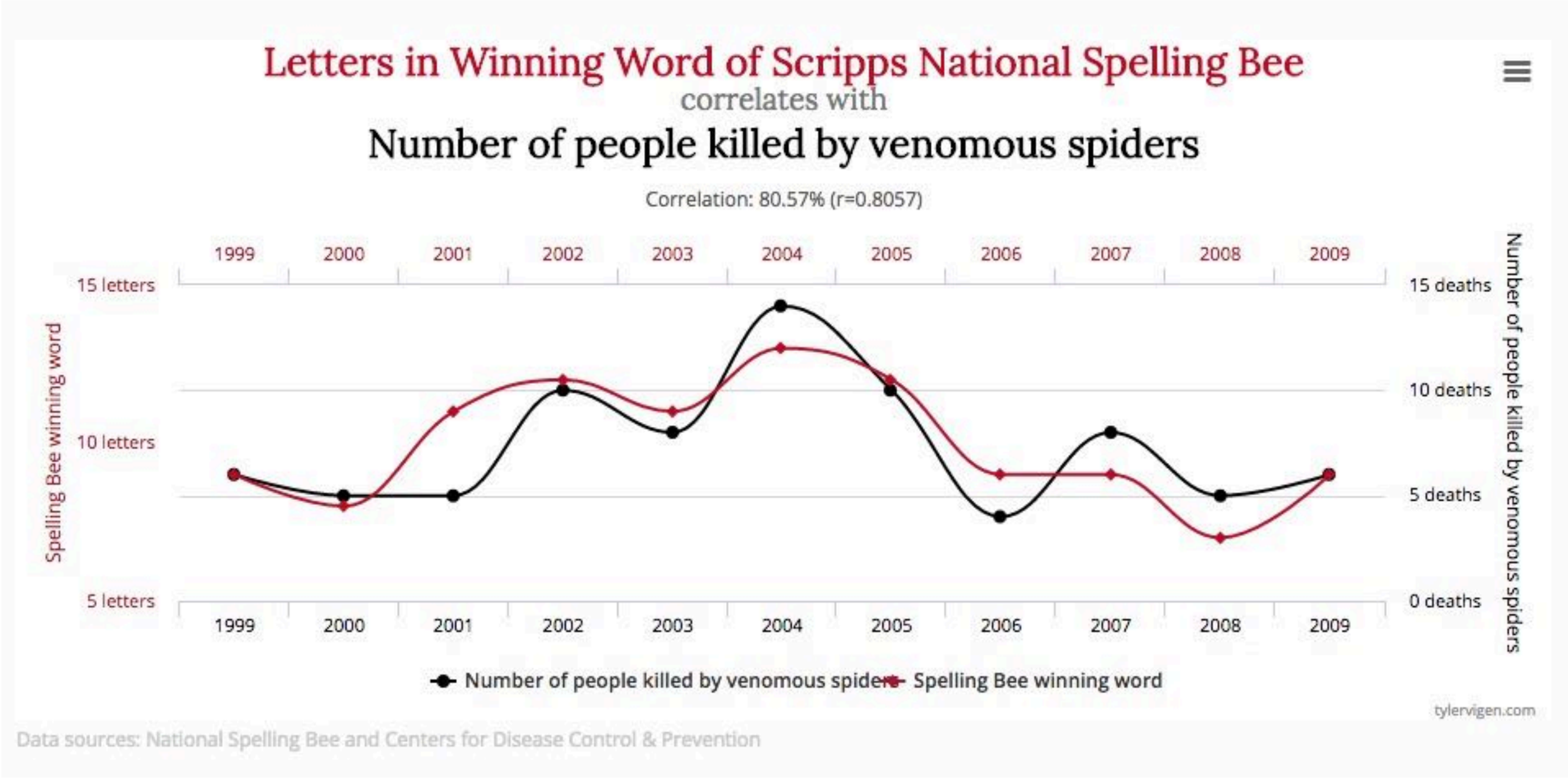
UC San Diego

Don’t trust your data (or narratives)



Source: http://www.tylervigen.com/spurious-correlations

Don’t trust your data (or narratives)



Source: http://www.tylervigen.com/spurious-correlations

A line

slope

$$y = mx + b$$

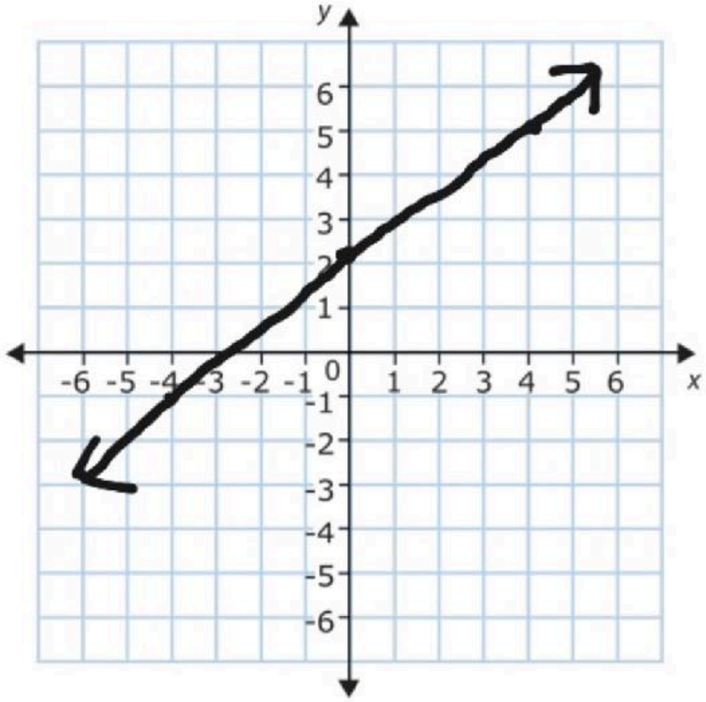
intercept

A line

slope

$$y = mx + b$$

intercept



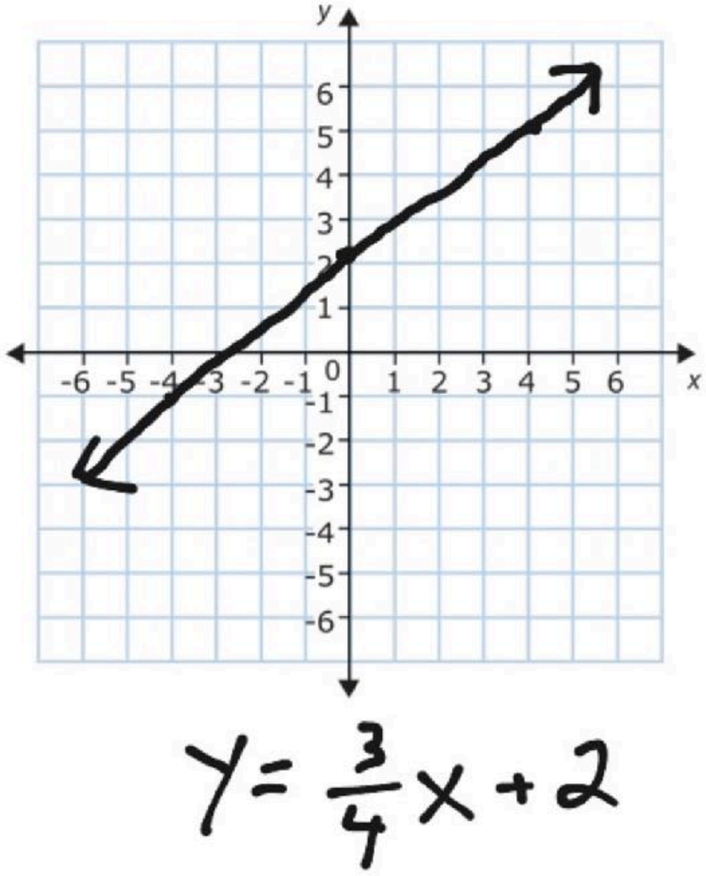
$$y = \frac{3}{4}x + 2$$

Ordinary Least Squares (OLS)

slope

$$y = mx + b$$

intercept



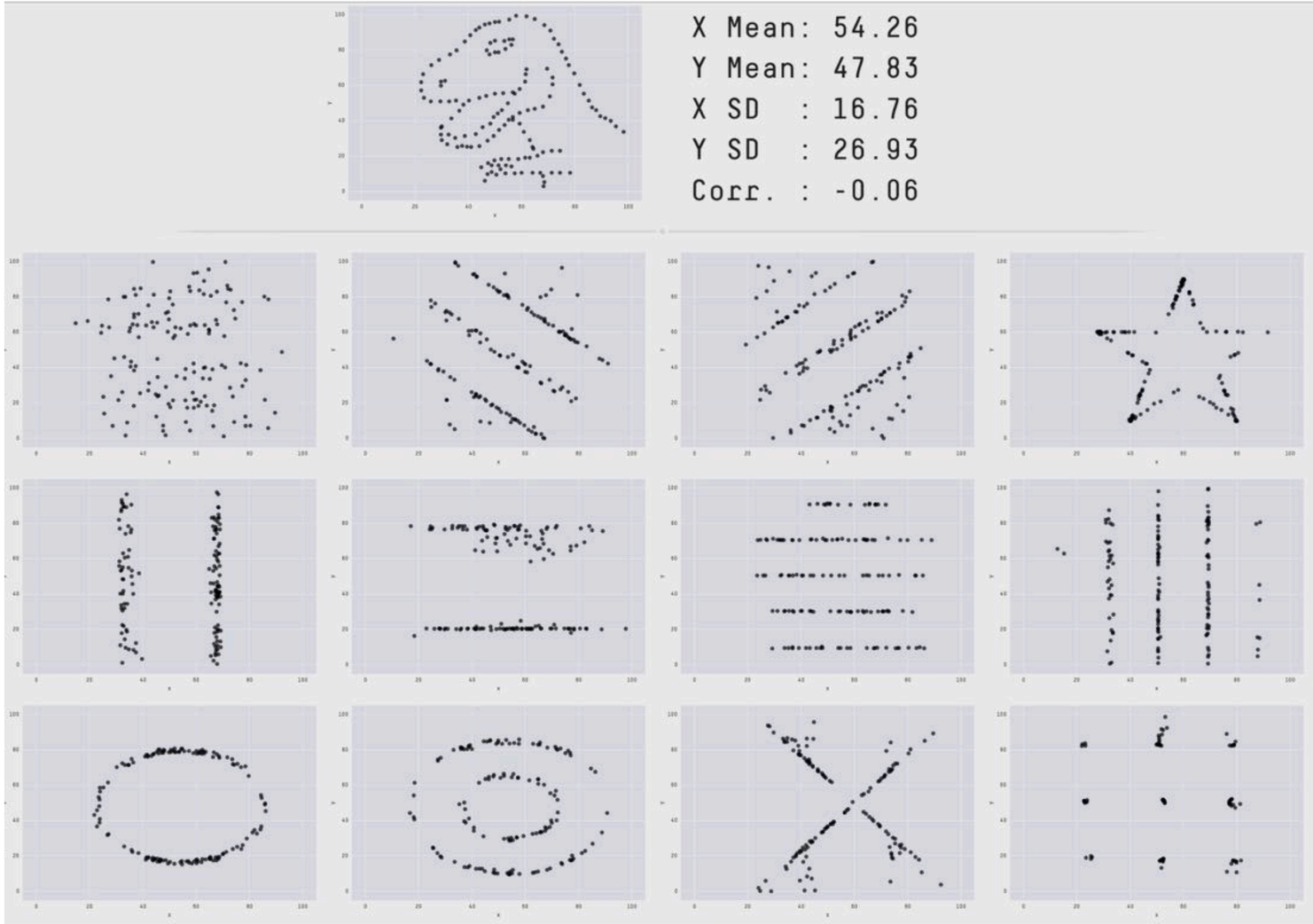
intercept

error

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

slope

Visualize your data!



Source: Matejka & Fitzmaurice, *ACM SIGCHI* 2017

Visualize your data!

Source: Matejka & Fitzmaurice, *ACM SIGCHI* 2017

COGS 108
Data Science in Practice

Distributions and outliers

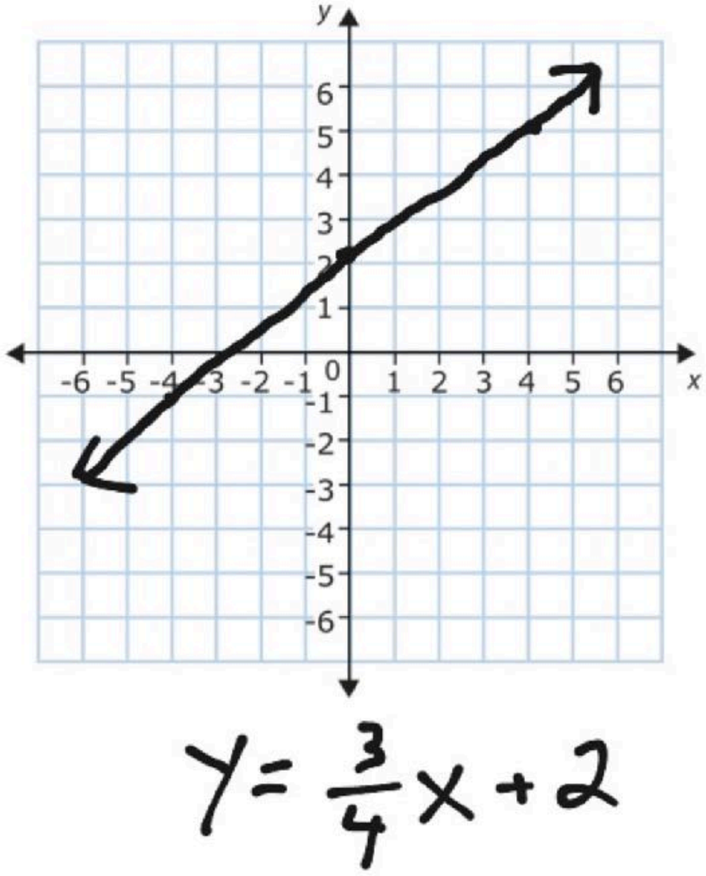
UC San Diego

Ordinary Least Squares (OLS)

slope

$$y = mx + b$$

intercept



intercept

error

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

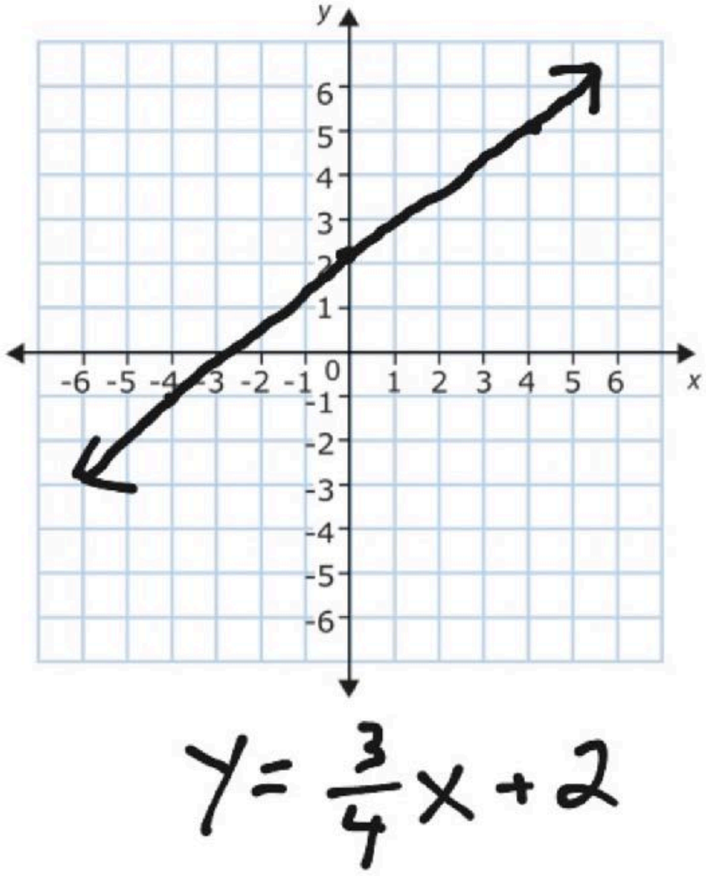
slope

Ordinary Least Squares (OLS)

slope

$$y = mx + b$$

intercept



intercept

error

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

slope

Ordinary Least Squares (OLS)

intercept *error*

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

slope

Ordinary Least Squares (OLS)

parameter
intercept

residual
error

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

slope
parameter

Ordinary Least Squares (OLS)

parameter
intercept

residual
error

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

slope
parameter

Goal: find values for the parameters to provide a "best" fit for the data

Ordinary Least Squares (OLS)

parameter
intercept

residual
error

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

slope
parameter

Goal: find values for the parameters to provide a "best" fit for the data

how?

Ordinary Least Squares (OLS)

parameter
intercept

residual
error

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

slope
parameter

minimize this!

Ordinary Least Squares (OLS)

parameter
intercept

residual
error

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

slope
parameter

minimize this!

how?

Ordinary Least Squares (OLS)

parameter
intercept

residual
error

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

slope
parameter

$$\min \sum_{i=1}^{n} \varepsilon_i^2$$

Ordinary Least Squares (OLS)

parameter
intercept

residual
error

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

slope
parameter

$$\min \sum_{i=1}^{n} \varepsilon_i^2$$

why squared?

UC San Diego

Ordinary Least Squares (OLS)

parameter
intercept

residual
error

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

slope
parameter

$$\min \sum_{i=1}^{n} \varepsilon_i^2$$

$$y_i - \beta_0 - \beta_1 x_i = \varepsilon_i$$

Ordinary Least Squares (OLS)

parameter
intercept

residual
error

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

slope
parameter

$$\min \sum_{i=1}^{n} \varepsilon_i^2$$

$$y_i - \beta_0 - \beta_1 x_i = \varepsilon_i$$

$$\min \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Ordinary Least Squares (OLS)

parameter
intercept

residual
error

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

slope
parameter

$$\beta_0 = \mu_y - \beta_1 \mu_x$$

$$\beta_1 = r_{xy} \frac{\sigma_y}{\sigma_x}$$

Ordinary Least Squares (OLS)

parameter
intercept

residual
error

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

slope
parameter

$$\beta_0 = \mu_y - \beta_1 \mu_x$$

WHAT DOES THIS MEAN?!?

$$\beta_1 = r_{xy} \frac{\sigma_y}{\sigma_x}$$

Ordinary Least Squares (OLS)

parameter
intercept

residual
error

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

slope
parameter

$$\beta_0 = \mu_y - \beta_1 \mu_x$$

Ordinary Least Squares (OLS)

parameter
intercept

residual
error

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

slope
parameter

$$\beta_0 = \mu_y - \beta_1 \mu_x$$

when we see the average value of *x*, we predict the average value of *y*

Ordinary Least Squares (OLS)

parameter
intercept

residual
error

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

slope
parameter

$$\beta_0 = \mu_y - \beta_1 \mu_x$$

when we see the average value of *x*, we predict the average value of *y*

$$\beta_1 = r_{xy} \frac{\sigma_y}{\sigma_x}$$

Ordinary Least Squares (OLS)

parameter
intercept

residual
error

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

slope
parameter

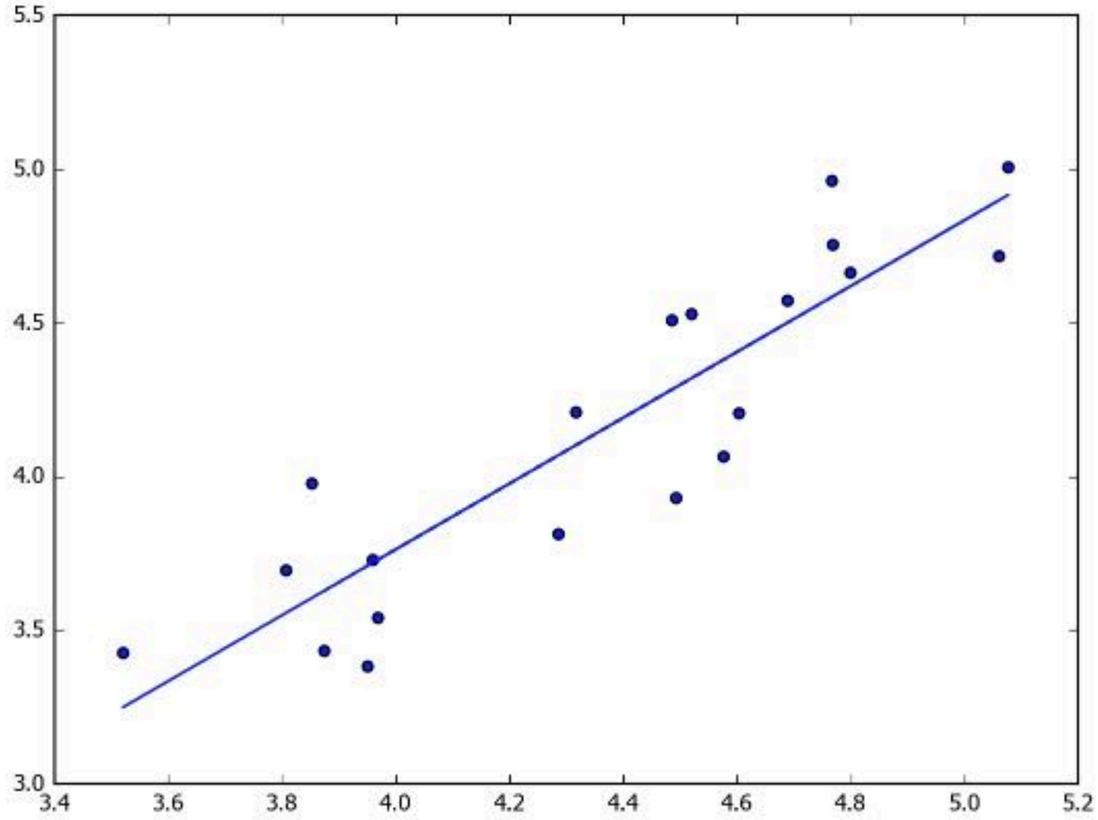
$$\beta_0 = \mu_y - \beta_1 \mu_x$$

when we see the average value of *x*, we predict the average value of *y*

$$\beta_1 = r_{xy} \frac{\sigma_y}{\sigma_x}$$

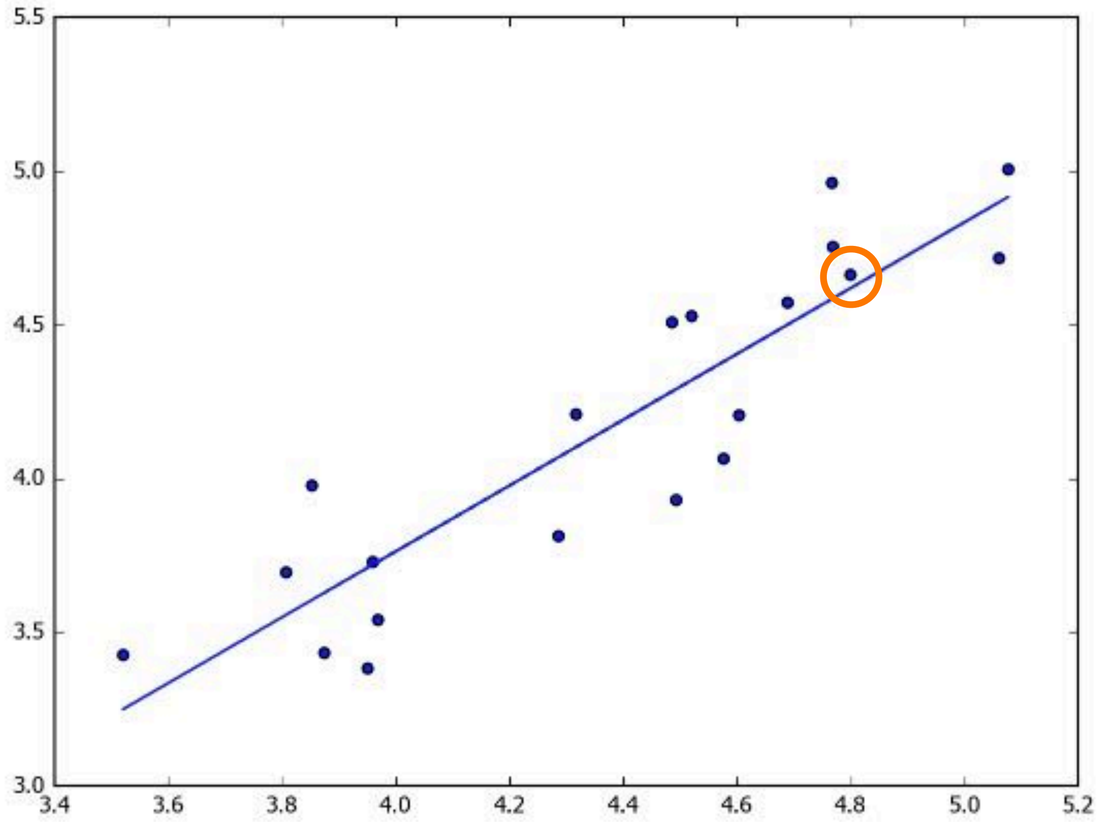
when the input increases by σ_x, the prediction increases by the *x*,*y* correlation

Outliers!



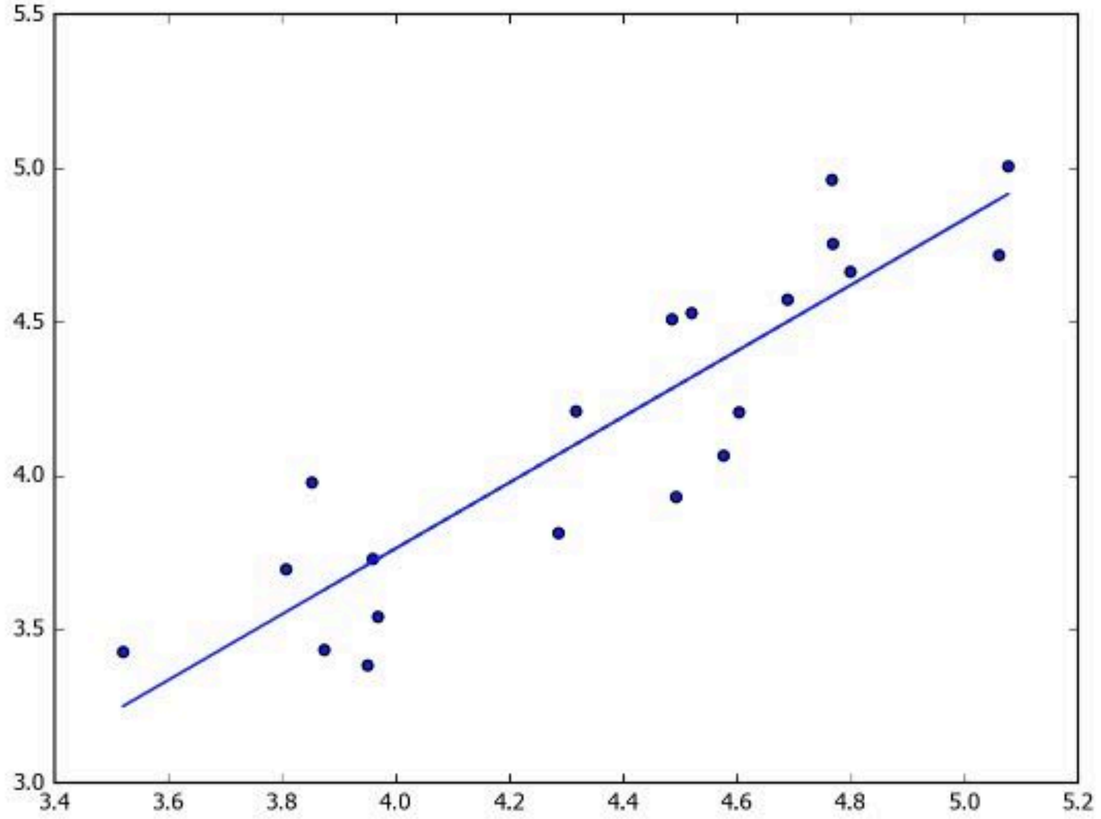

```
r = 0.899192316859 ; p = 7.04649901141e-08
```

Outliers!

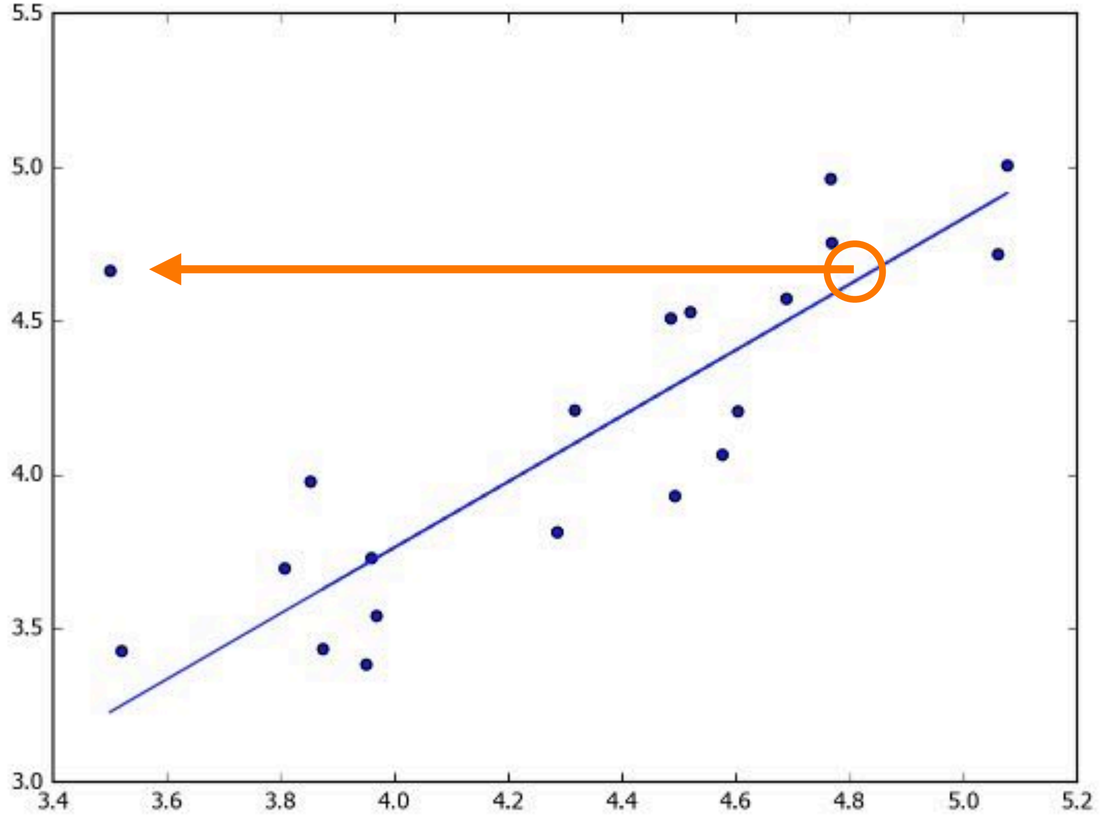



```
r = 0.899192316859 ; p = 7.04649901141e-08
```

Outliers!

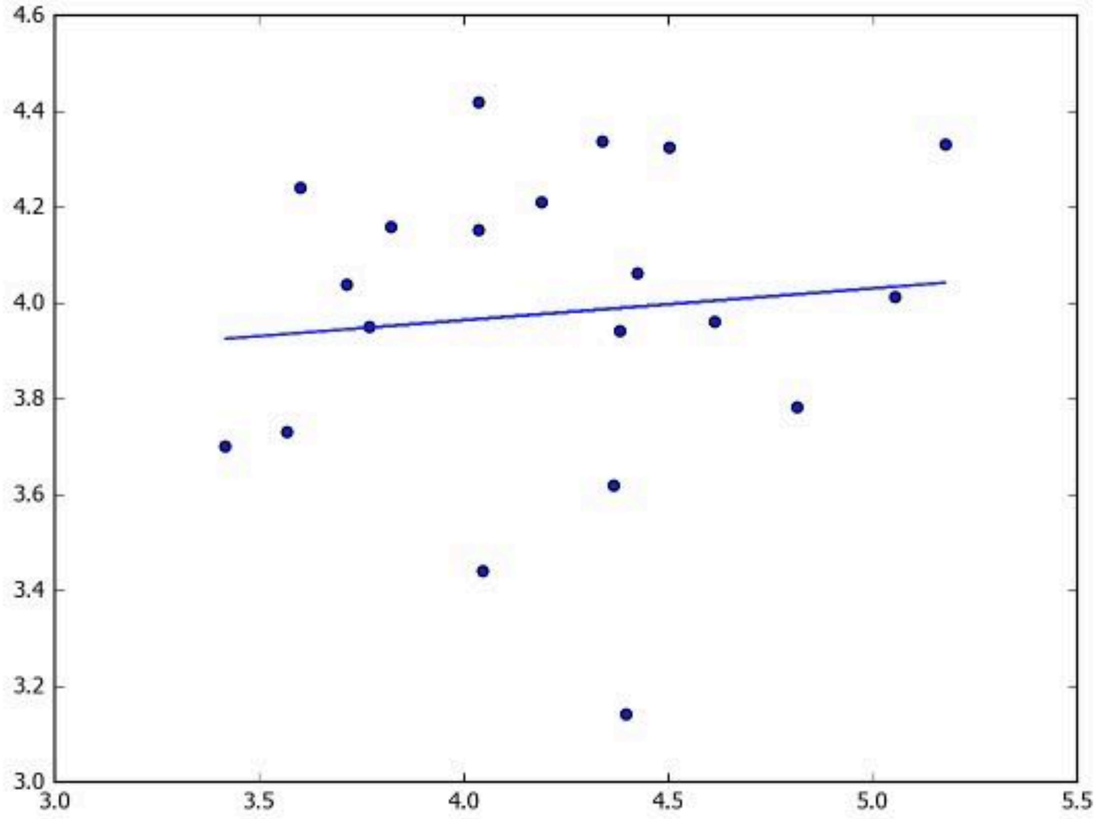


r = 0.899192316859 ; p = 7.04649901141e-08



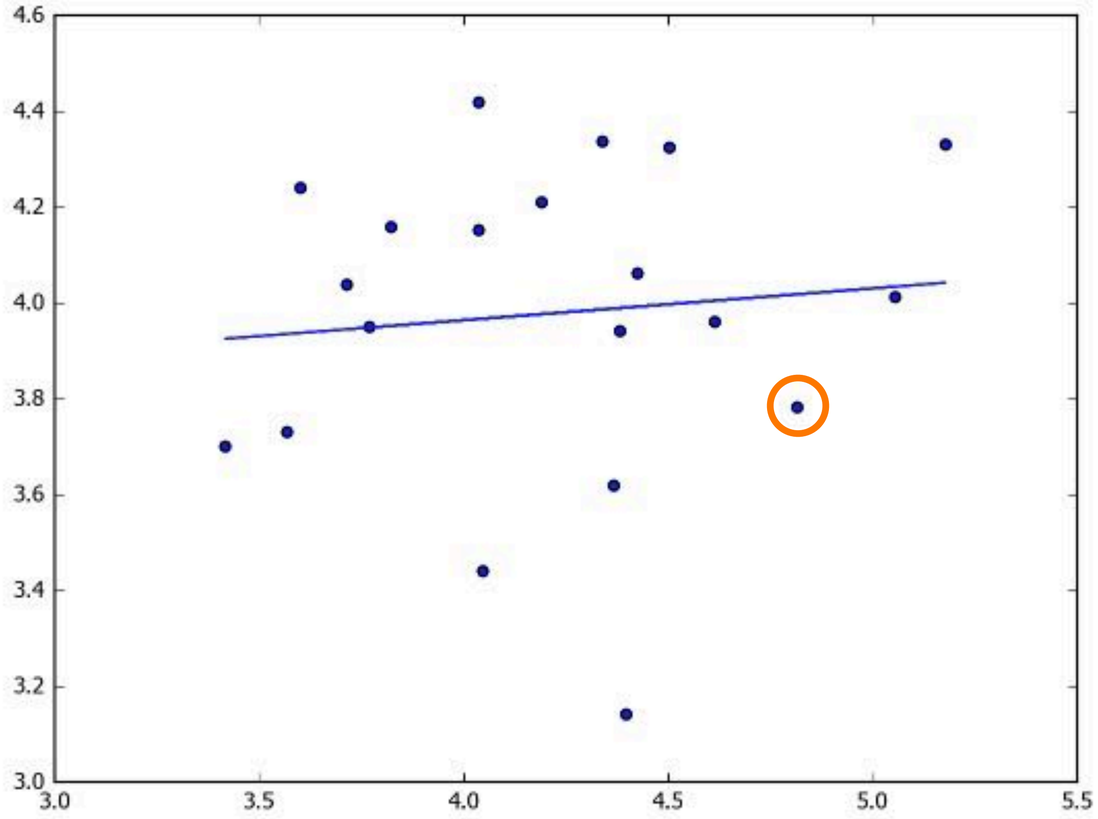
0.709892176011 0.000454396182138

Outliers!



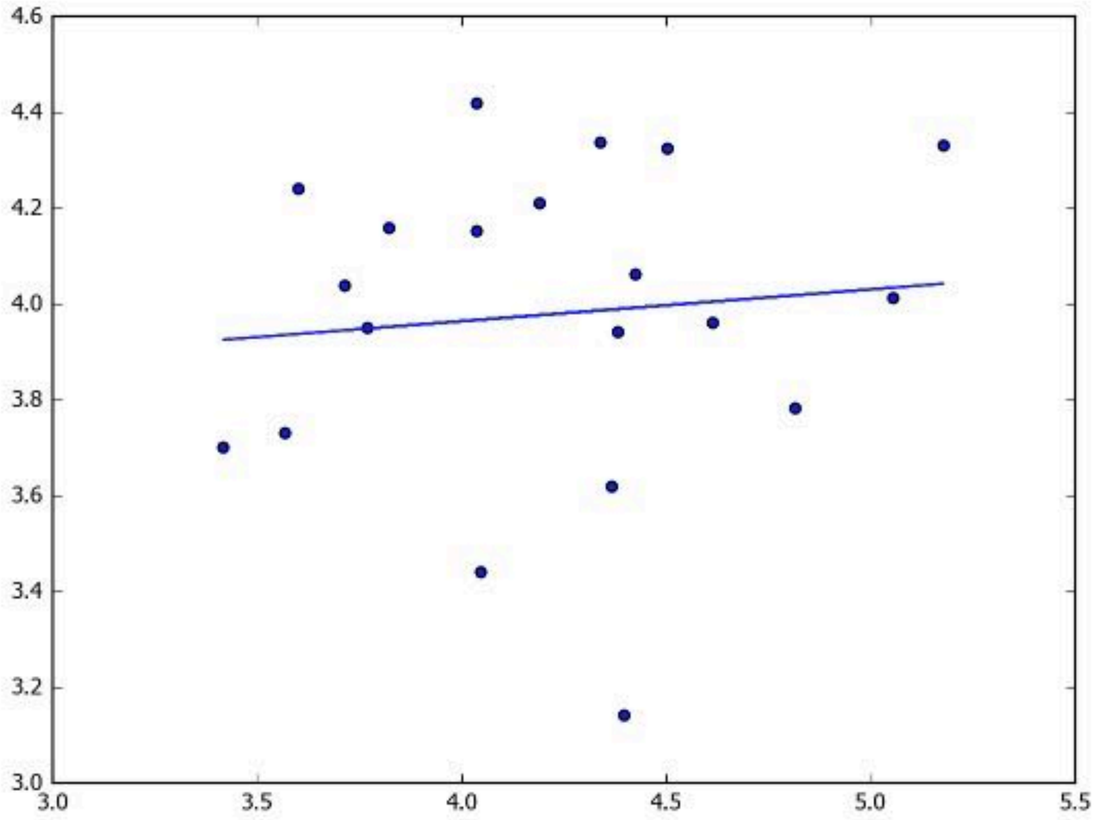

```
r = 0.0982029696446 ; p = 0.680415082529
```

Outliers!

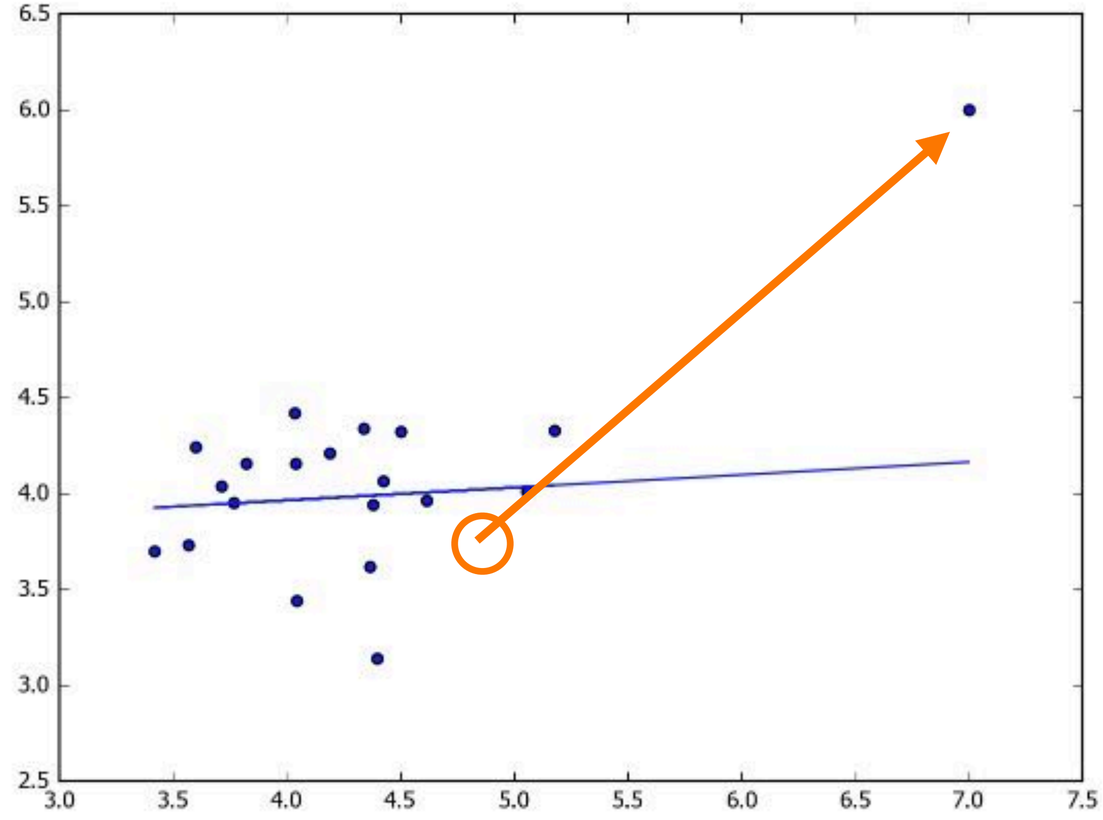



```
r = 0.0982029696446 ; p = 0.680415082529
```

Outliers!

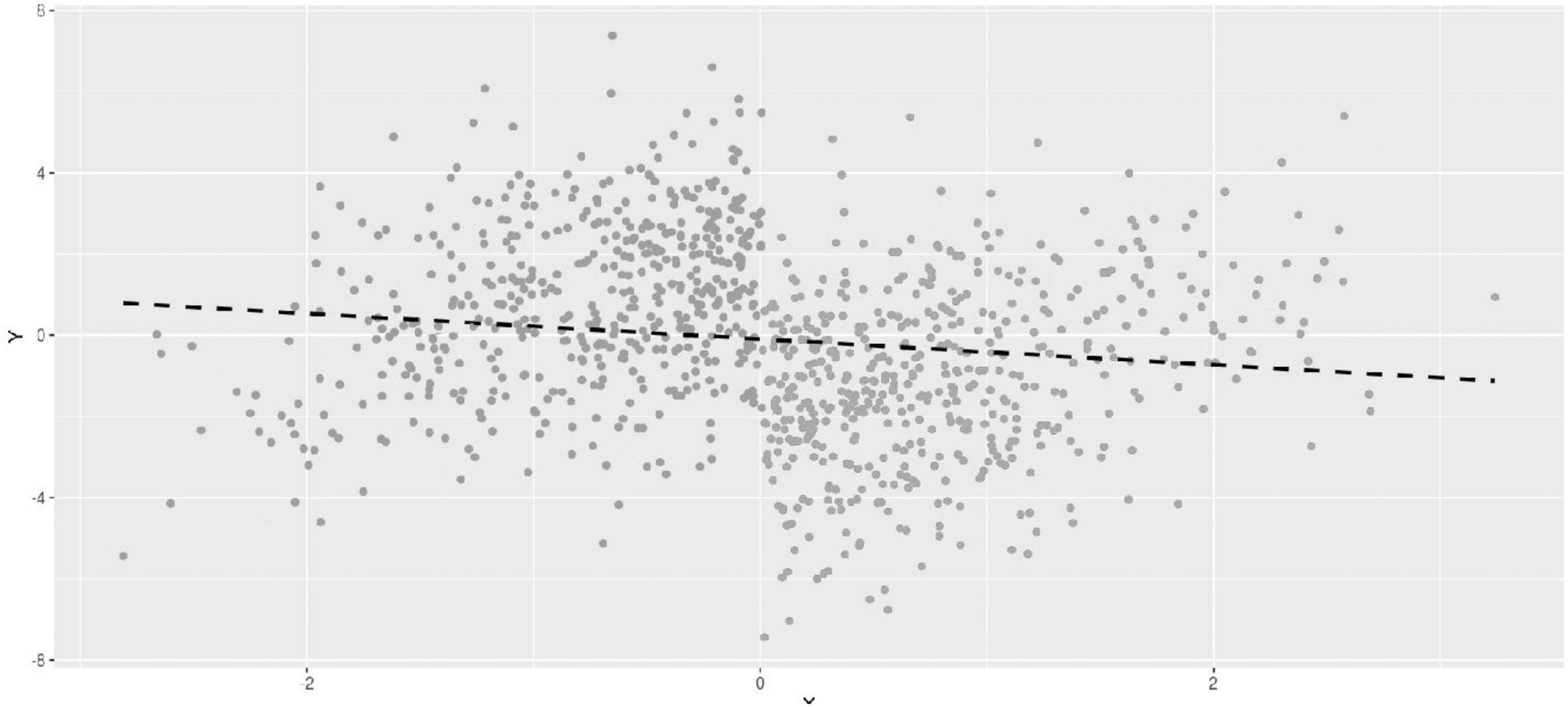



```
r = 0.0982029696446 ; p = 0.680415082529
```



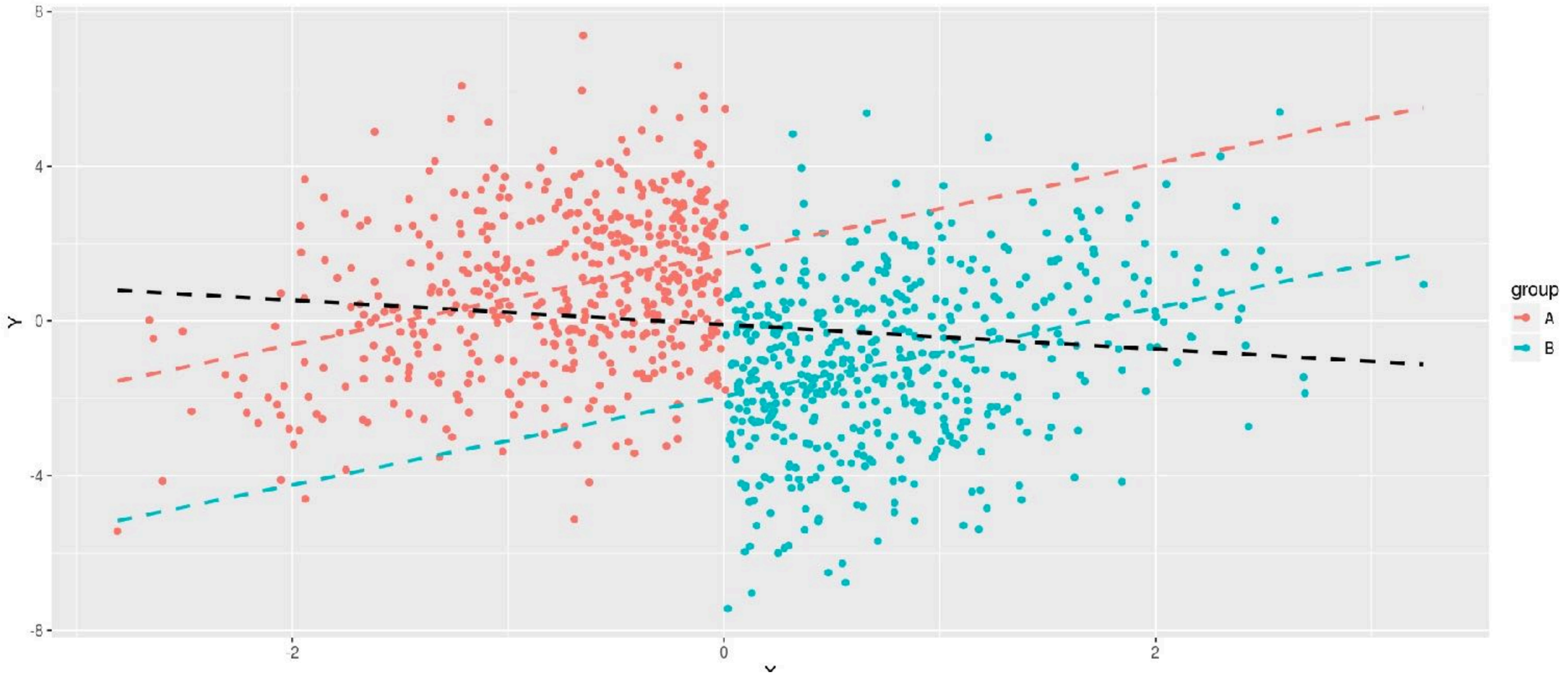

```
r = 0.702033103659 ; p = 0.000559678419262
```

Simpson's Paradox



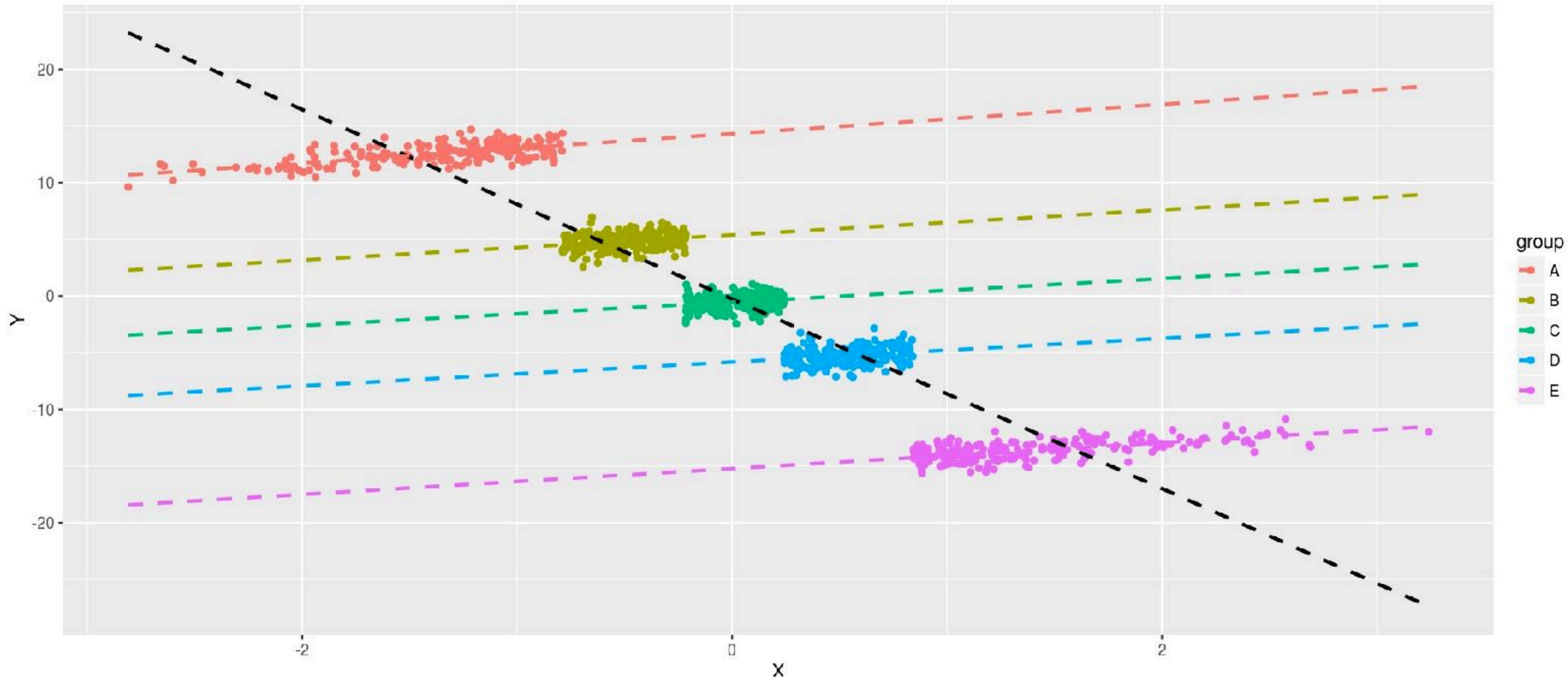
Source: Emil O. W. Kirkegaard

Simpson's Paradox



Source: Emil O. W. Kirkegaard

Simpson's Paradox



Source: Emil O. W. Kirkegaard

Simpson's Paradox - Real world example

Example 2—Airlines on-time data: Here are numbers of flights on time and delayed for two airlines at five airports in June 1991. The table shows that Alaska Airlines outperforms America West at all 5 cities. If you collapse the table over city, it appears that America West outperforms Alaska.

	Alaska Airlines			America West Airlines		
	On time	**Delayed**	**Delay%**	**On time**	**Delayed**	**Delay%**
LA	497	62	11.1%	694	117	14.4%
Phoenix	221	12	5.4%	4840	415	7.9%
San Diego	212	20	8.6%	383	65	14.5%
San Fran.	503	102	16.9%	320	129	28.7%
Seattle	1841	305	14.2%	201	61	23.3%
Total	**3274**	**501**	**13.3%**	**6438**	**787**	**10.9%**

Simpson's Paradox - Intuition

- **Totally contrived example:**
  - You and a friend are both avid gamers

Simpson's Paradox - Intuition

- **Totally contrived example:**
  - You and a friend are both avid gamers
  - You are also both kind of competitive

Simpson's Paradox - Intuition

- **Totally contrived example:**
  - You and a friend are both avid gamers
  - You are also both kind of competitive
  - You make a friendly wager: you'll both play a question-answering game twice, and whomever has the best record, wins

Simpson's Paradox - Intuition

- **Totally contrived example:**
  - You and a friend are both avid gamers
  - You are also both kind of competitive
  - You make a friendly wager: you'll both play a question-answering game twice, and whomever has the best record, wins

- **Day 1:**
  - you: 98/99 (98.99%)

- **Day 1:**
  - them: 1/1 (**100%**)

Simpson's Paradox - Intuition

- **Totally contrived example:**
  - You and a friend are both avid gamers
  - You are also both kind of competitive
  - You make a friendly wager: you'll both play a question-answering game twice, and whomever has the best record, wins

- **Day1:**
  - you: 98/99 (98.99%)
- **Day2:**
  - you: 0/1 (0%)

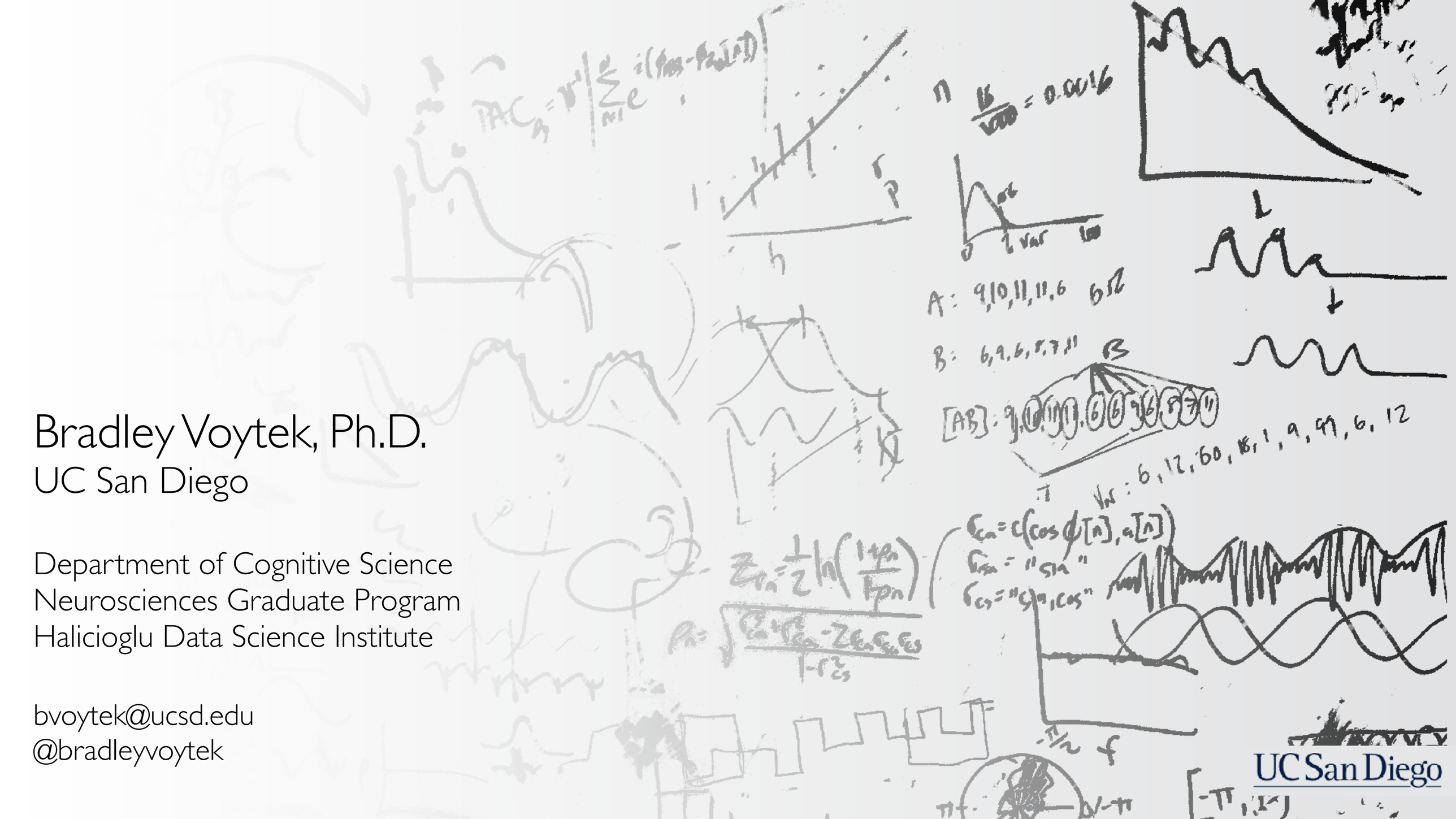
- **Day1:**
  - them: 1/1 (**100%**)
- **Day2:**
  - them: 1/99 (**1.01%**)

Simpson's Paradox - Intuition

- **Totally contrived example:**
  - You and a friend are both avid gamers
  - You are also both kind of competitive
  - You make a friendly wager: you'll both play a question-answering game twice, and whomever has the best record, wins

- **Day1:**
  - you: 98/99 (98.99%)
- **Day2:**
  - you: 0/1 (0%)
- **Total:**
  - you: 98/100 (**98%**)

- **Day1:**
  - them: 1/1 (**100%**)
- **Day2:**
  - them: 1/99 (**1.01%**)
- **Total:**
  - them: 2/100 (2%)



Bradley Voytek, Ph.D.

UC San Diego

Department of Cognitive Science
Neurosciences Graduate Program
Halicioglu Data Science Institute

bvoytek@ucsd.edu
@bradleyvoytek

UC San Diego