Statistics Fundamentals, Part 2

Ivan Corneillet

Data Scientist

GENERAL ASSEMBLY

Learning Objectives

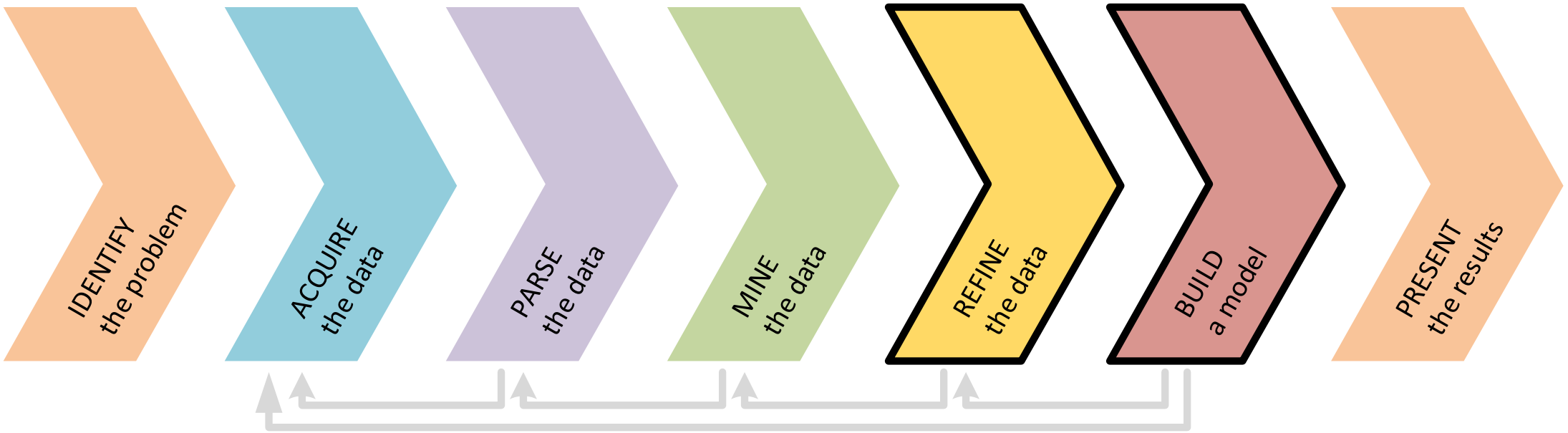
After this lesson, you should be able to:

- Explain the difference between causation and correlation
- Identify a normal distribution within a dataset using summary statistics and visualization
- Test a hypothesis within a sample case study
- Validate your findings using statistical analysis (t-tests, p-values, t-values, confidence intervals)

Outline

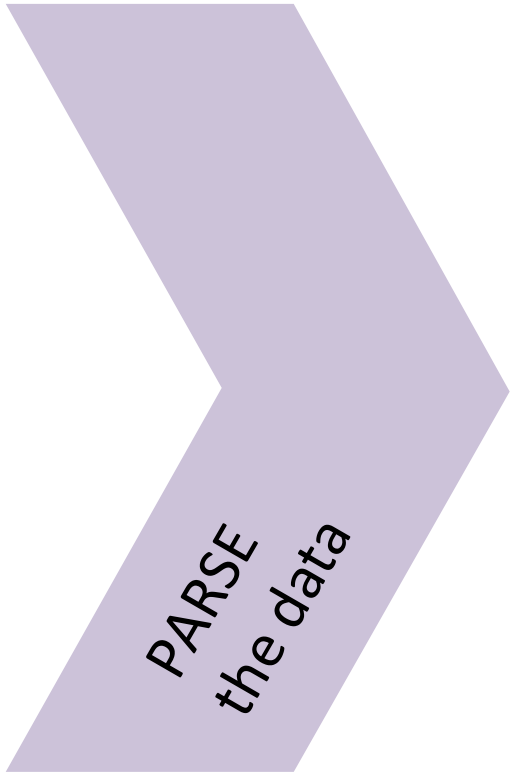
- Review
- ❺ Refine the Data and ❻ Build a Model
  - Causation and Correlation
  - If correlation doesn't imply causation, then what does?
  - Cofounding
  - Hill's Criteria for Causation
  - Do you really need causality or is correlation enough?
  - Data Mining and Spurious Correlations
  - Motivating Example: Codealong
  - The Normal Distribution
  - The 68 – 90 – 95 – 99.7 Rule
  - Hypothesis Testing
  - Two-Tail Hypothesis Test
  - t-value
  - p-value
  - Confidence Intervals
- Lab
- Review
- In-flight
  - **Unit Project 2 (due next session on 3/10)**
  - Final Project 1 (due in 2 weeks)

Today's Topic and the Data Science Workflow



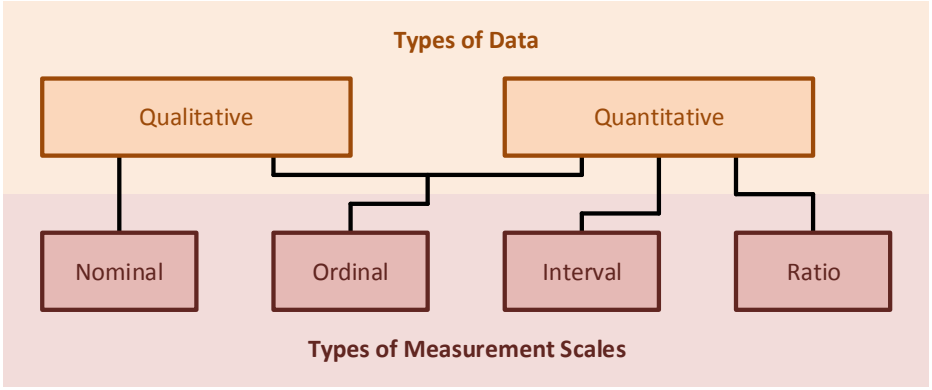
Review
❸ Parse the Data (Statistics)

❸ Parse the Data



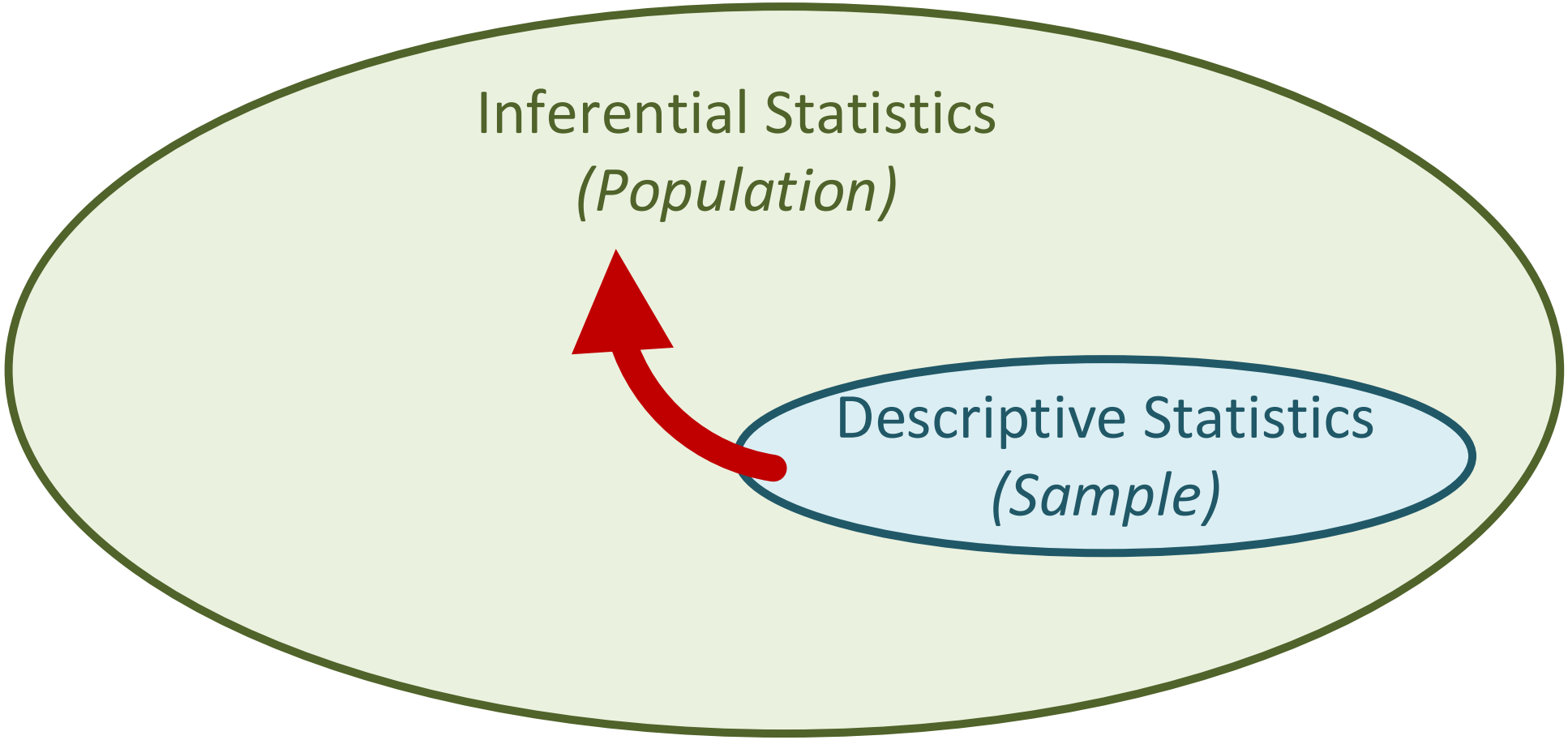
- Parse the Data
  - Read any documentation provided with the data *(session 2)*
  - Perform exploratory data analysis *(session 3)*
  - Verify the quality of the data *(sessions 2/3)*

Types of Data and Types of Measurement Scales

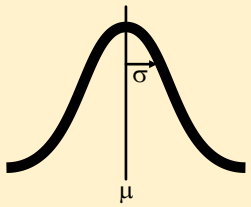
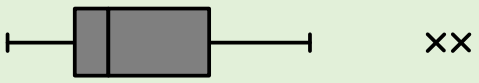
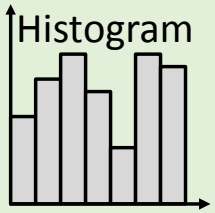


	Nominal	Ordinal	Interval	Ratio
Categorize?	✓	✓	✓	✓
Rank-order?	✗	✓	✓	✓
+; -?	✗	✗	✓	✓
*; /?	✗	✗	✗	✓

Descriptive and Inferential Statistics

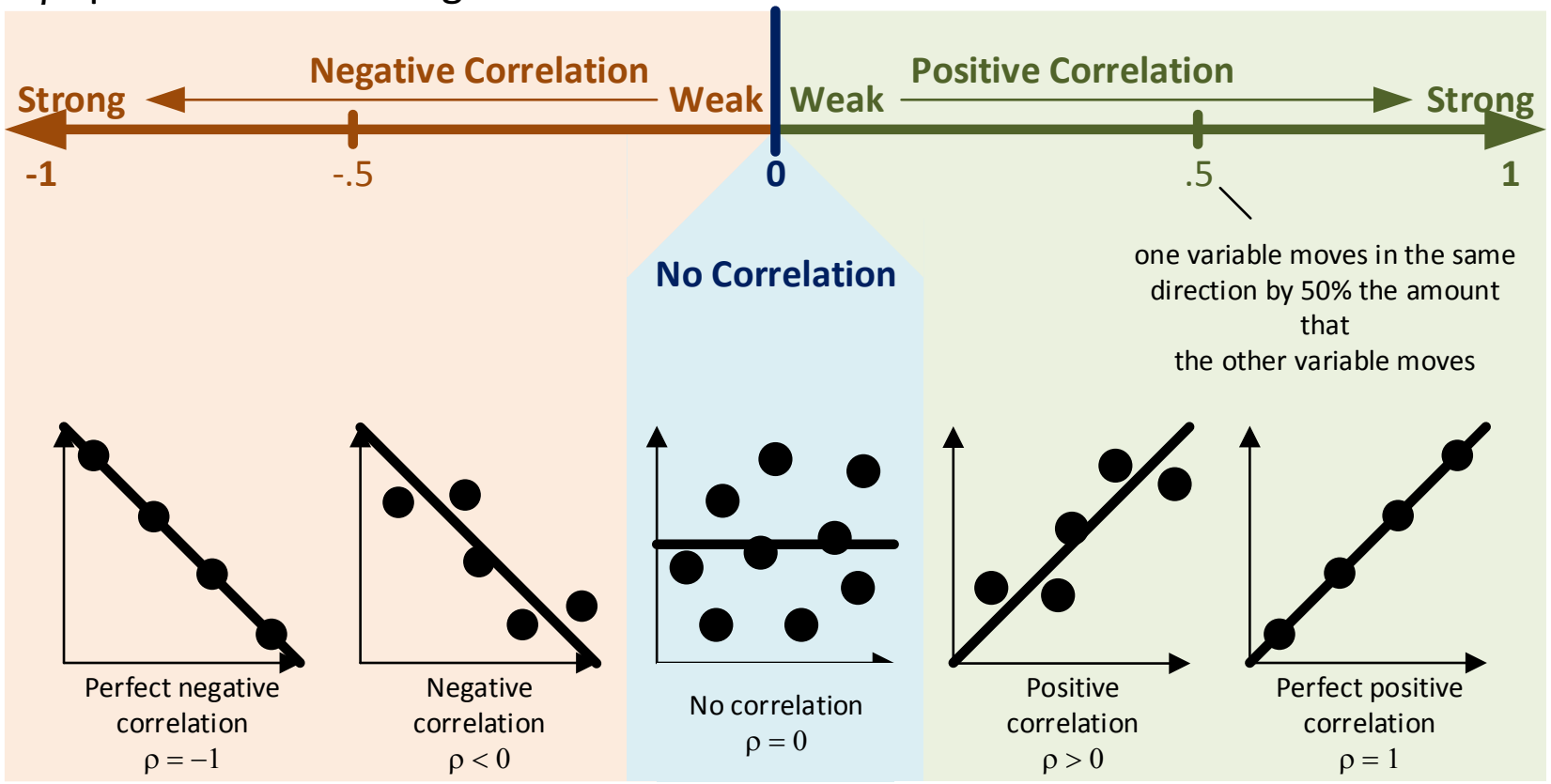


Descriptive Statistics

Measure of Centrality	Mean	Median	Mode
Measurement Scales	Interval - Ratio	Interval - Ratio	Nominal - Ratio
• In the dataset?	☹	😐	☺
• Easy of compute	☺	😐	☹
• Resistant to outliers?	☹	☺	☺
Measure of Dispersion	☺ (Variance, Standard Deviation)	☺ (Interquartile Range)	☹
Extensive used in mathematical models?	☺	☹	☹
Graphical Methods		Boxplot 	Histogram 

Correlation

ρ quantifies the strength and direction of movements of two random variables



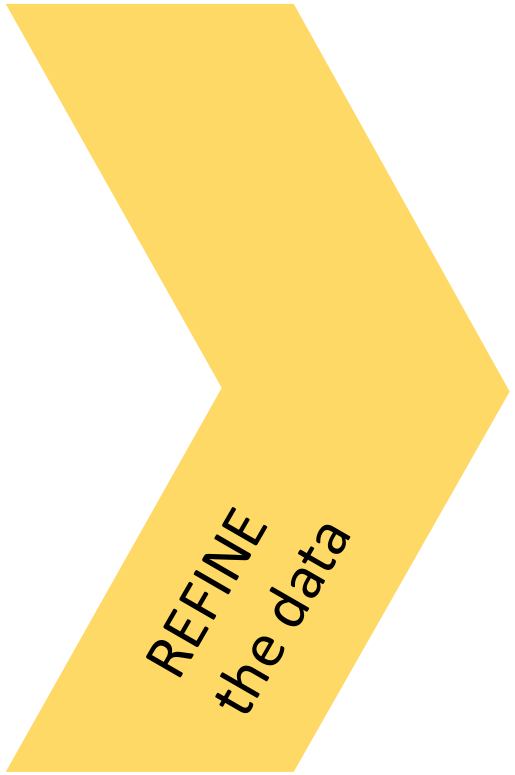
Descriptive Statistics, Correlation, and Python

Measure of Centrality	`.mean()`	`.median()`	`.mode()`
Measure of Dispersion	`.var(), .std()`	`.min(), .max()` `.quantile()`	
Summary	`.describe()`		
Graphical Methods		`.plot(kind = 'box')`	`.plot(kind = 'hist')`
Others	`.count(), sum(), .unique()` `.isnull(), dropna()`		

Correlation Matrix	`.corr()`
Scatter plot	*DataFrame*`.plot(kind = 'scatter', x = `*Series*`, y = `*Series*`)`
Scatter matrix	`pd.tools.plotting.scatter_matrix(`*DataFrame*`)`

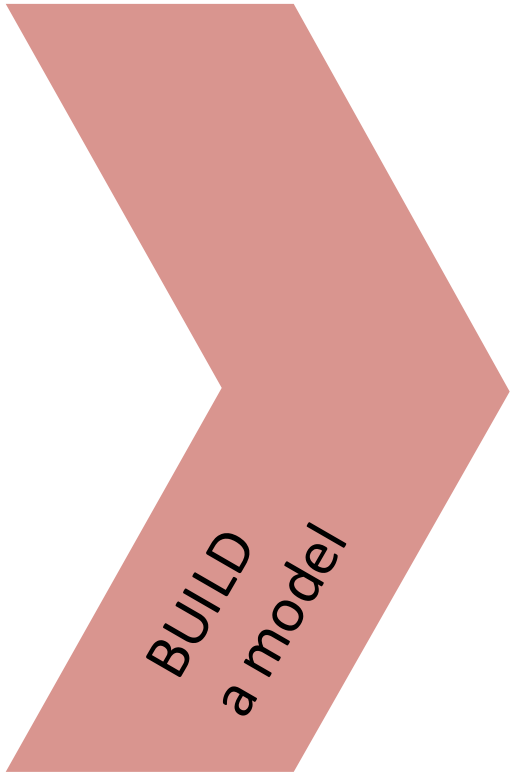
❺ Refine the Data
❻ Build a Model (Statistics)

❺ Refine the Data



- Refine the Data
  - Identify trends and outliers *(session 3)*
  - Apply descriptive *(session 3)* and **inferential statistics** *(session 4)*
  - Document *(session 2)* and transform data *(units 2-3)*

❻ Build a Model



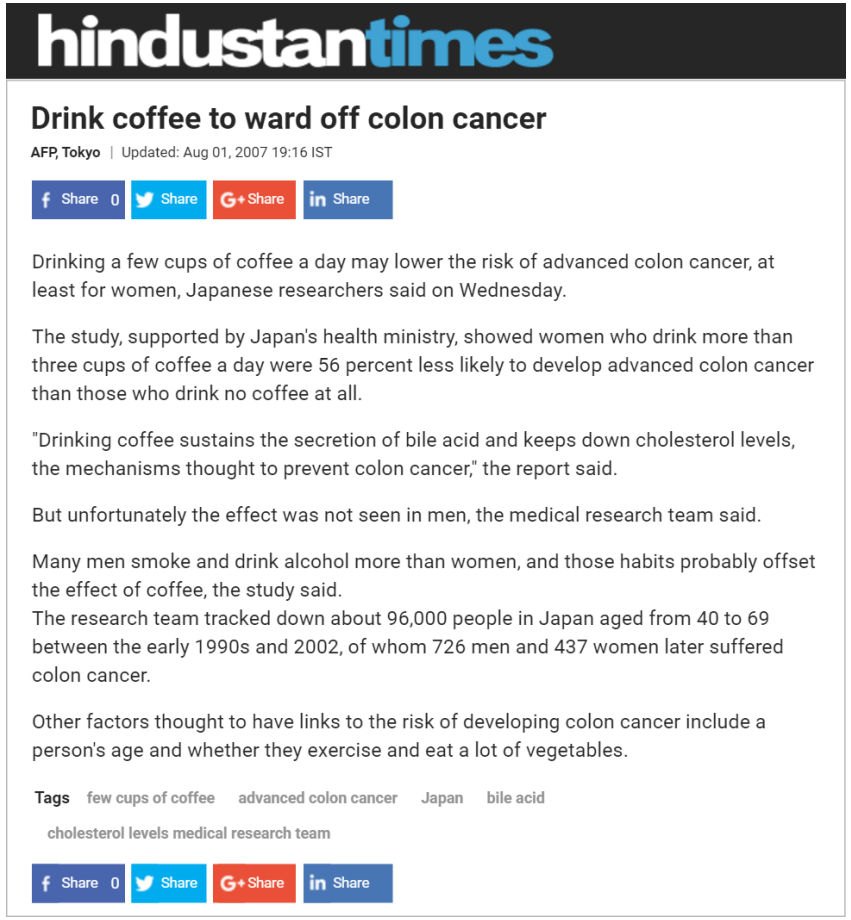
- Build a Model
  - Select appropriate model *(units 2-3)*
  - Build model *(units 2-3)*
  - **Evaluate** *(session 4)* and refine model *(units 2-3)*

Causation and Correlation

Causation and Correlation

- If an association is observed,
  - the first question to ask should always be…
    - **is it real?**

E.g., Coffee and Colon Cancer



hindustantimes

Drink coffee to ward off colon cancer

AFP, Tokyo | Updated: Aug 01, 2007 19:16 IST

Share 0 Share Share Share

Drinking a few cups of coffee a day may lower the risk of advanced colon cancer, at least for women, Japanese researchers said on Wednesday.

The study, supported by Japan's health ministry, showed women who drink more than three cups of coffee a day were 56 percent less likely to develop advanced colon cancer than those who drink no coffee at all.

"Drinking coffee sustains the secretion of bile acid and keeps down cholesterol levels, the mechanisms thought to prevent colon cancer," the report said.

But unfortunately the effect was not seen in men, the medical research team said.

Many men smoke and drink alcohol more than women, and those habits probably offset the effect of coffee, the study said.
The research team tracked down about 96,000 people in Japan aged from 40 to 69 between the early 1990s and 2002, of whom 726 men and 437 women later suffered colon cancer.

Other factors thought to have links to the risk of developing colon cancer include a person's age and whether they exercise and eat a lot of vegetables.

Tags few cups of coffee advanced colon cancer Japan bile acid cholesterol levels medical research team

Share 0 Share Share Share

CancerConnect.com
community content connection

Home » Coffee Does Not Decrease Risk of Colorectal Cancer
Categories: Colon Cancer, News, Rectal Cancer

Coffee Does Not Decrease Risk of Colorectal Cancer

Contrary to the results of several previous studies, coffee consumption does not appear to reduce the risk of colorectal cancer, according to the results of a study published in the *International Journal of Cancer*.[1]

Colorectal cancer is the second leading cause of cancer-related deaths in the United States. The disease develops in the large intestine, which includes the colon (the longest part of the large intestine) and the rectum (the last several inches).

Some studies have indicated that coffee may have a protective effect against colon cancer; however, researchers continue to evaluate this link in an effort to establish more direct evidence. In order to examine the relationship between coffee consumption and colorectal cancer, researchers from Harvard conducted a review of 12 studies that included 646,848 participants and 5,403 cases of colorectal cancer.

They evaluated high versus low coffee consumption and found no significant effect of coffee consumption on colorectal cancer risk. The review included four studies in the United States, five in Europe, and three in Japan. The data from each country was very similar. There were no significant differences by gender or site of cancer; however, there was a slight inverse relationship (reduction in risk) between coffee consumption and colon cancer for women, which was even more pronounced among Japanese women (21% for total study, 38% for Japanese women).

The researchers observed that inverse associations between coffee consumption and colorectal cancer "were slightly stronger in studies that controlled for smoking and alcohol and in studies with shorter follow-up times."

They concluded that coffee is "unlikely to have a strong protective effect on colorectal cancer risk"; however, they also note that it does not appear to increase the risk of colorectal cancer either.

Reference:

[1] Je Y, Liu W, Giovannucci E. Coffee consumption and risk of colorectal cancer: A systematic review and meta-analysis of prospective cohort studies. *International Journal of Cancer*. 2009; 124: 1662-1668.

E.g., Alcohol and Dementia Risk

BBC NEWS

Friday, 25 January, 2002, 12:13 GMT

Alcohol 'could reduce dementia risk'

Small amounts of alcohol could reduce the risk of dementia in older people regardless of the type of alcoholic drink consumed, research suggests.

It is known that light-to-moderate consumption lessens the risk of coronary heart disease and stroke, but Dutch scientists think it could be good for mental health.

The team at Erasmus University Medical School in Rotterdam compared the risk of developing dementia between individuals who regularly consumed alcohol with those who did not consume alcohol.

Light-to-moderate alcohol consumption (one to three drinks a day) was associated with a 42% risk reduction of all dementia and about a 70% reduction in risk of vascular dementia (dementia caused by a series of small strokes).

Out of 8,000 people who took part, 197 individuals developed dementia - of these, 146 had Alzheimer's disease, 29 developed vascular dementia and 22 got other types of dementia, it is reported in the Lancet medical journal.

"It is particularly impressive that just 1-3 drinks per day can reduce the risk of vascular dementia"

Dr Richard Harvey, Alzheimer's Society

The team suggests alcohol may have a direct effect on brain activity by stimulating the release of the chemical acetylcholine in the hippocampus area of the brain.

Acetylcholine is known to facilitate memory and learning processes, however high alcohol intake inhibits acetylcholine production.

Monique Breteler, who led the research, said: "In recent years, evidence has been accumulating that vascular factors may be involved in the cause of dementia, both vascular dementia and Alzheimer's disease.

"Our findings lend further support to the vascular hypothesis of dementia.

Limited intake

"We saw some indication for a stronger relation with alcohol in persons with a genetically determined susceptibility for Alzheimer's disease.

"Our findings can help focus research into the specific mechanisms that underlie the development of dementing illnesses."

Alzheimer's disease is the most common cause of dementia, accounting for 50% of all cases.

Vascular dementia accounts for about 20% of cases.

The Alzheimer's Society has welcomed the survey findings.

The society's research director Dr Richard Harvey said: "This interesting new study confirms the results of previous research which has suggested that light to moderate alcoholic consumption is actually good for our health.

"It is particularly impressive that just 1-3 drinks per day can reduce the risk of vascular dementia.

"Clearly, however, excessive alcohol consumption is not good for our long term health and increases the risk of serious diseases such as cirrhosis of the liver.

"It is very much the case of a little of what you fancy appears to do you good."

All those taking part in the research were aged 55+ and did not have dementia at the start of the study.

WebMD

Drinking and Dementia: Is There a Link?

Study Shows Drinkers With Genetic Predisposition to Alzheimer's Disease at Higher Risk

By Salynn Boyles
WebMD Health News

Sept. 2, 2004 -- Drinking alcohol in middle age may increase the risk of late-life dementia in people who are genetically predisposed to develop Alzheimer's disease, according to findings from a Scandinavian study.

Researchers from Stockholm's Karolinska Institute reported that infrequent drinkers have a twofold increase in the risk of dementia in old age among carriers of a gene that has been linked to Alzheimer's. Gene carriers who frequently drink had a threefold increase in risk.

But the findings also show a protective effect for infrequent drinkers who did not have the genetic risk factor. Low-risk teetotalers and frequent drinkers in the study were twice as likely to experience mild cognitive declines later in life as infrequent drinkers.

The findings are reported in the Sept. 4 issue of the *BMJ* (formerly the *British Medical Journal*).

"Earlier studies indicated that light to moderate drinking may be protective, but this study shows that the picture is much more complex," researcher Miia Kivipelto, MD, PhD, tells WebMD. "The more people with this susceptibility gene drank, the more their risk for dementia increased."

Apolipoprotein E

The study included just more than 1,000 men and women followed for an average of 23 years, who were between the ages of 65 and 79 at follow-up. At enrollment, the participants provided details about their alcohol consumption.

People were considered infrequent drinkers if they drank alcohol less than once a month and frequent drinkers if they drank several times a month.

The researchers also took blood samples to determine which study participants were carriers of the apolipoprotein E genotype. The genotype is an established risk factor for dementia in old age, and as many as one in four Americans are carriers, Kivipelto says.

The Karolinska researchers reported that dementia risk appeared to be directly related to drinking frequency among study participants who were carriers of the gene.

"Our current data indicate that frequent alcohol drinking has harmful effects on the brain, and this may be more pronounced if there is genetic susceptibility," the researchers wrote. "We therefore do not want to encourage people to drink more alcohol in the belief that they are protecting themselves against dementia."

1 | 2 NEXT PAGE >

WebMD

Drinking and Dementia: Is There a Link?

Study Shows Drinkers With Genetic Predisposition to Alzheimer's Disease at Higher Risk

Lifestyle Influences

Alzheimer's Association vice president for medical and scientific affairs Bill Thies, PhD, echoes the sentiment. Thies tells WebMD that even though the data do suggest a protective benefit for light to moderate drinking, the studies examining drinking and old-age dementia are far from conclusive.

"Nobody is suggesting that people who don't drink alcohol start doing so to improve their health," he says.

Thies says there are many other things people with family histories of Alzheimer's or other age-related dementias can do to reduce their risk, including keeping their blood pressure, blood sugar, and cholesterol under control, maintaining a healthy weight, getting plenty of exercise, and eating well. Other tips can be found in the "Maintain Your Brain" section of the Alzheimer's Association web site (www.alz.org).

"We have much better evidence that these lifestyle factors contribute to Alzheimer's," Thies says.

< PREVIOUS PAGE 1 | 2

Causation and Correlation (cont.)

- Why is this?
  - Sensational headlines
  - No robust data analysis
  - Lack of understanding of the difference between *causation* and *correlation*
    - **"caused" ≠ "measured" or "associated"**
    - ***Correlation does not imply causation***

- Understanding this difference is critical in the data science workflow, especially when **Identifying** the problem and **Acquiring** the data
  - We need to fully articulate our question and use the right data to answer it, including any *confounders*
- Additionally, this comes up when **Presenting** our results to stakeholders

If correlation doesn’t imply causation, then what does?

Activity

Simpson's Paradox (a.k.a., Yule–Simpson effect, reversal paradox, or amalgamation paradox): A trend appears in different groups of data but disappears or reverses when these groups are combined

	Democrats	Republicans
North	94% ✓	85%
South	7% ✓	0%
Overall	61%	80% ✓

Simpson's Paradox: The Arithmetic

	Democrats	Republicans
North	**145 / 154 = 94%**	**138 / 162 = 85%**
South	**7 / 94 = 7%**	**0 / 10 = 0%**
Overall	**152 / 248 = 61%**	**138 / 172 = 80%**

Simpson's Paradox: Takeaway

- Determining causality on the basis of correlations is tricky and can even lead to contradictory conclusions

- "Partial evidence may be worse than no evidence if it leads to an illusion of knowledge, and so to overconfidence and certainty where none is justified. It's better to know that you don't know" – Michael Nielsen

Activity – Take 2

You suffer from kidney stones, and your doctor offers you two choices: treatment A or treatment B. Which will you chose?

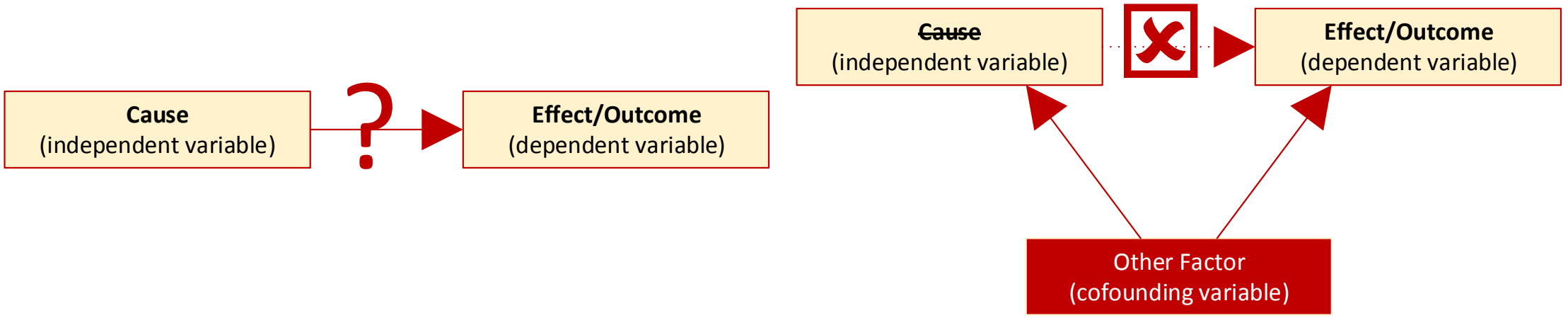
	Treatment A	Treatment B
Effectiveness	**61%**	**80%**

The gotcha… Still going with treatment B?

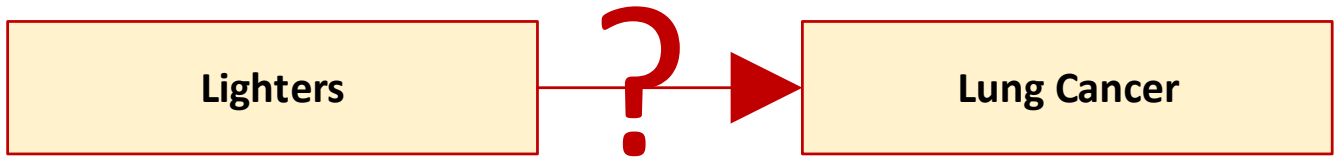
The gotcha…	Treatment A	Treatment B
Patients with large kidney stones	94%	85%
Those with small kidney stones	7%	0%
Overall	61%	80%

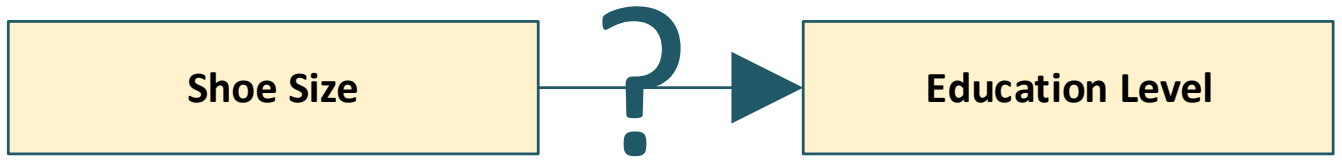
Cofounding

Cofounding

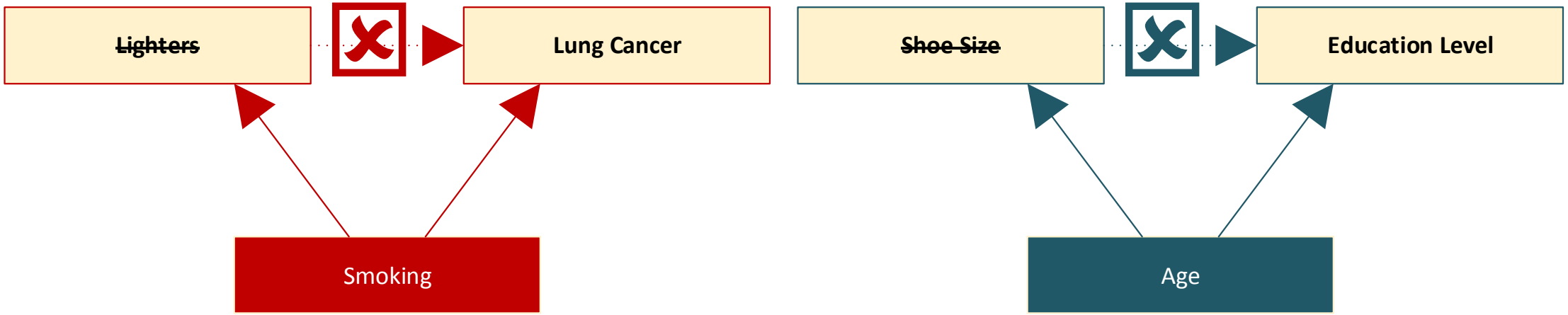

Cause
(independent variable)
?
Effect/Outcome
(dependent variable)
~~Cause~~
(independent variable)
Effect/Outcome
(dependent variable)
Other Factor
(cofounding variable)

Cofounding (cont.)





Cofounding (cont.)



If you really need to establish causality…

The Bradford Hill Criteria (a.k.a., Hill's Criteria for Causation): a group of minimal conditions necessary to establish causality (commonly used in the medical field)

Strength	(or effect size): a small association does not mean that there is not a causal effect, though the larger the association, the more likely that it is causal
Consistency	Consistent findings observed by different persons in different places with different samples strengthens the likelihood of an effect
Specificity	Causation is likely if there is a very specific population at a specific site and disease with no other likely explanation. The more specific an association between a factor and an effect is, the bigger the probability of a causal relationship
Temporality	**The effect has to occur after the cause** (and if there is an expected delay between the cause and expected effect, then the effect must occur after that delay)
Biological gradient	Greater exposure should generally lead to greater incidence of the effect. However, in some cases, the mere presence of the factor can trigger the effect. In other cases, an inverse proportion is observed: greater exposure leads to lower incidence
Plausibility	A plausible mechanism between cause and effect is helpful (but Hill noted that knowledge of the mechanism is limited by current knowledge)
Coherence	Coherence between epidemiological and laboratory findings increases the likelihood of an effect. However, Hill noted that "… lack of such [laboratory] evidence cannot nullify the epidemiological effect on associations"
Experiment	"Occasionally it is possible to appeal to experimental evidence"
Analogy	The effect of similar factors may be considered

Do you really need causality or is correlation enough?

Amazon

EXAMPLE

- "the Amazon Voice"
  - Hand-crafted reviews and title recommendations
  - Considered one of the company's crown jewels and a source of its competitive advantage
- What if Amazon could recommend specific books to customers based on their individual shopping preferences?
  - Comparing people with other people was cumbersome
  - All it needed to do was find associations among product themselves
    - "item-to-item" collaborative filtering patent
  - Data-generated material generated vastly more sales and the "Amazon Voice" group was disbanded
  - Knowing *what*, not *why*, is good enough

“Item-to-Item” Collaborative Filtering

Collaborative recommendations using item-to-item similarity mappings

US 6266649 B1

Publication number	US6266649 B1
Publication type	Grant
Application number	US 09/157,198
Publication date	Jul 24, 2001
Filing date	Sep 18, 1998
Priority date	Sep 18, 1998
Fee status	Paid
Also published as	EP1121658A1, EP1121658A4, WO2000017792A1
Inventors	Gregory D. Linden, Jennifer A. Jacobi, Eric A. Benson
Original Assignee	Amazon.Com, Inc.
Export Citation	BiBTeX, EndNote, RefMan

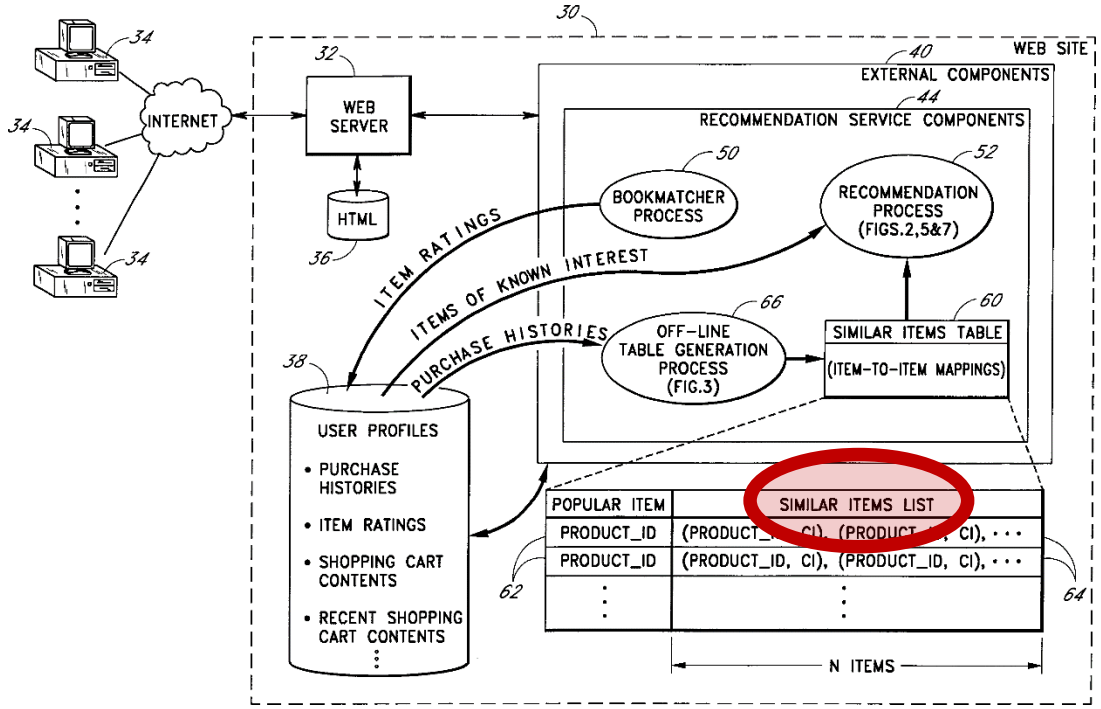
Patent Citations (22), Non-Patent Citations (39), Referenced by (1104), Classifications (23), Legal Events (9)

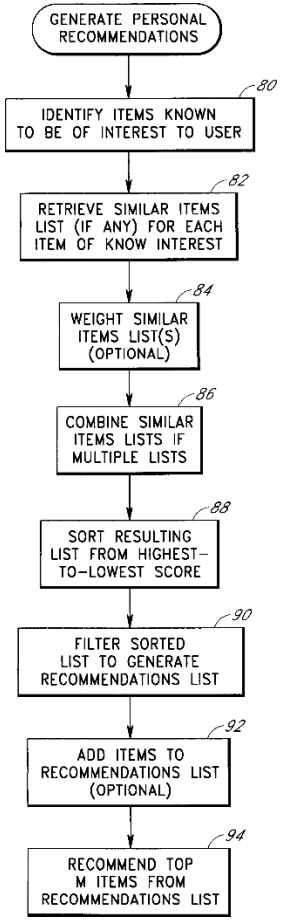
External Links: USPTO, USPTO Assignment, Espacenet

ABSTRACT

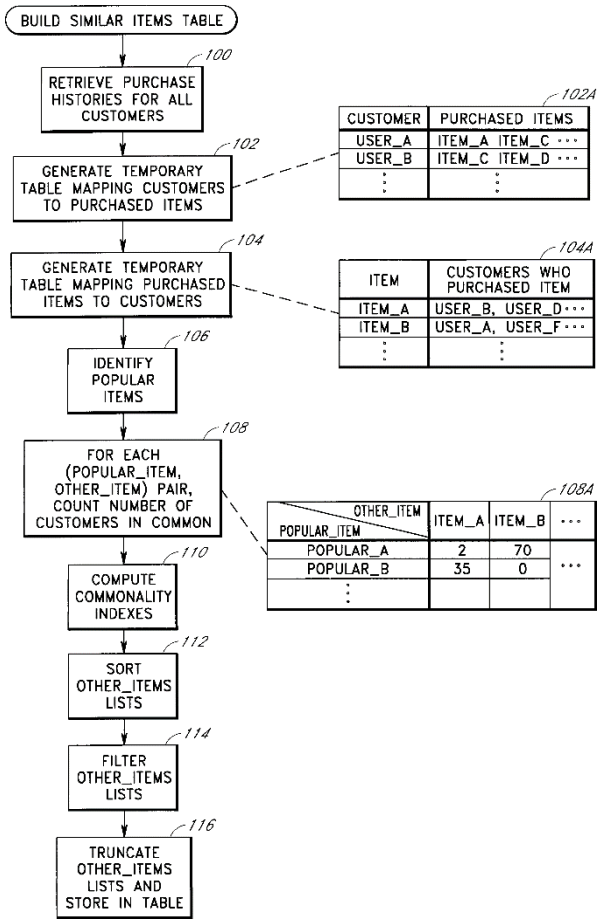
A recommendations service recommends items to individual users based on a set of items that are known to be of interest to the user, such as a set of items previously purchased by the user. In the disclosed embodiments, the service is used to recommend products to users of a merchant's Web site. The service generates the recommendations using a previously-generated table which maps items to lists of “similar” items. The similarities reflected by the table are based on the collective interests of the community of users. For example, in one embodiment, the similarities are based on correlations between the purchases of items by users (e.g., items A and B are similar because a relatively large portion of the users that purchased item A also bought item B). The table also includes scores which indicate degrees of similarity between individual items. To generate personal recommendations, the service retrieves from the table the similar items lists corresponding to the items known to be of interest to the user. These similar items lists are appropriately combined into a single list, which is then sorted (based on combined similarity scores) and filtered to generate a list of recommended items. Also disclosed are various methods for using the current and/or past contents of a user's electronic shopping cart to generate recommendations. In one embodiment, the user can create multiple shopping carts, and can use the recommendation service to obtain recommendations that are specific to a designated shopping cart. In another embodiment, the recommendations are generated based on the current contents of a user's shopping cart, so that the recommendations tend to correspond to the current shopping task being performed by the user.

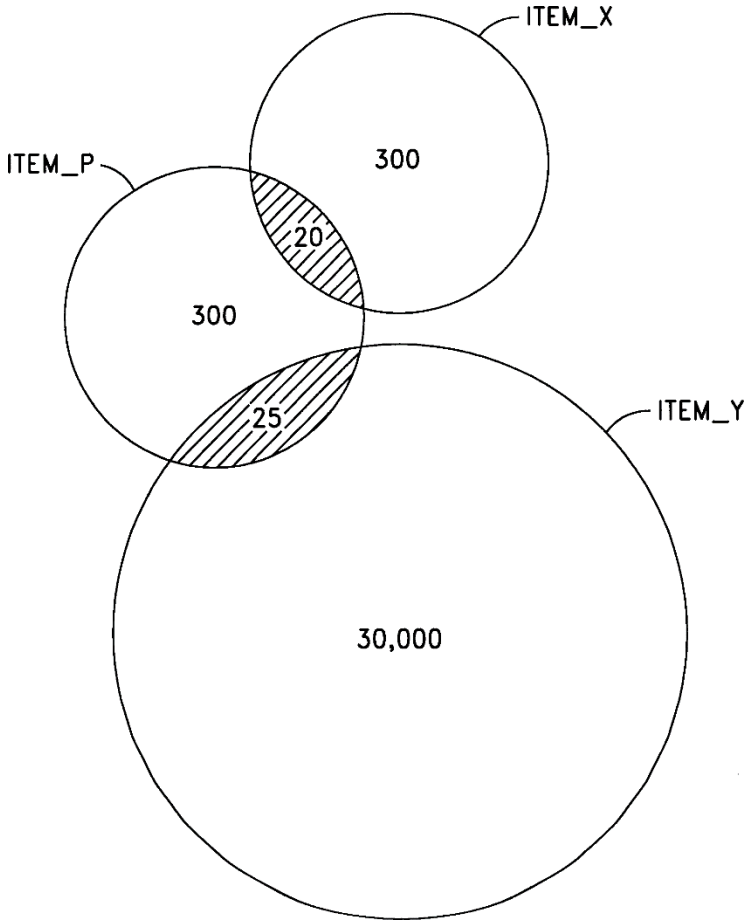
“Item-to-Item” Collaborative Filtering (cont.)



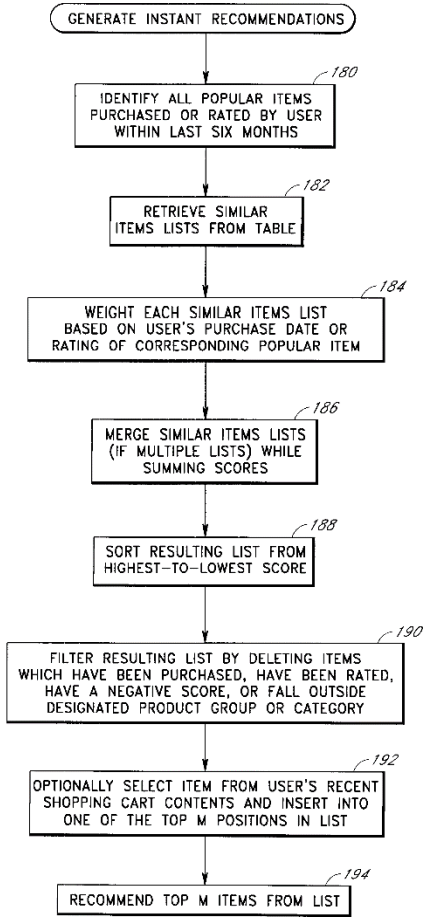


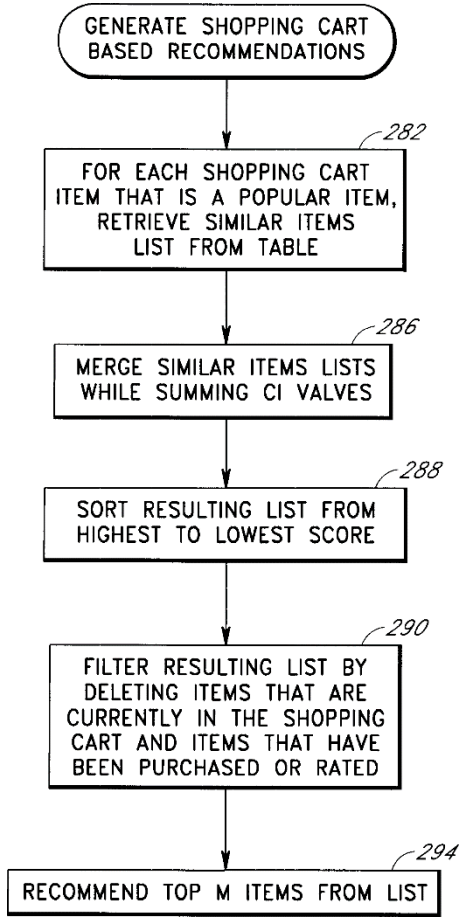
“Item-to-Item” Collaborative Filtering (cont.)





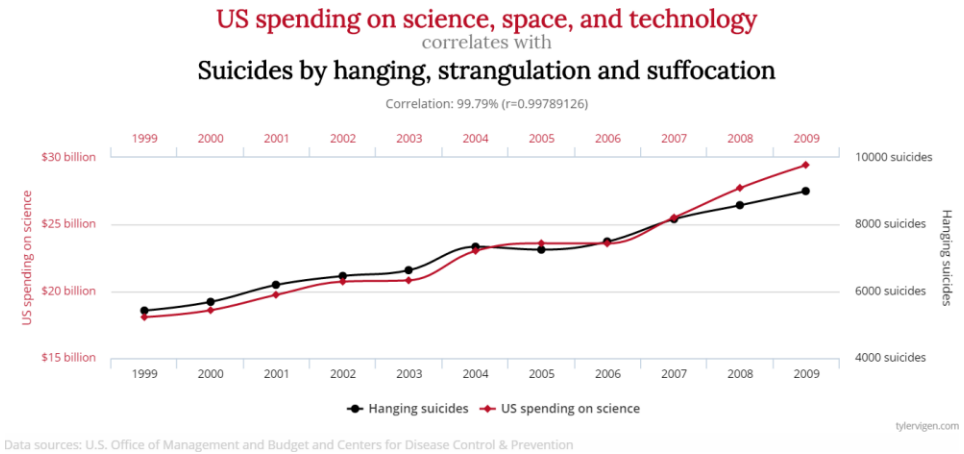
“Item-to-Item” Collaborative Filtering (cont.)

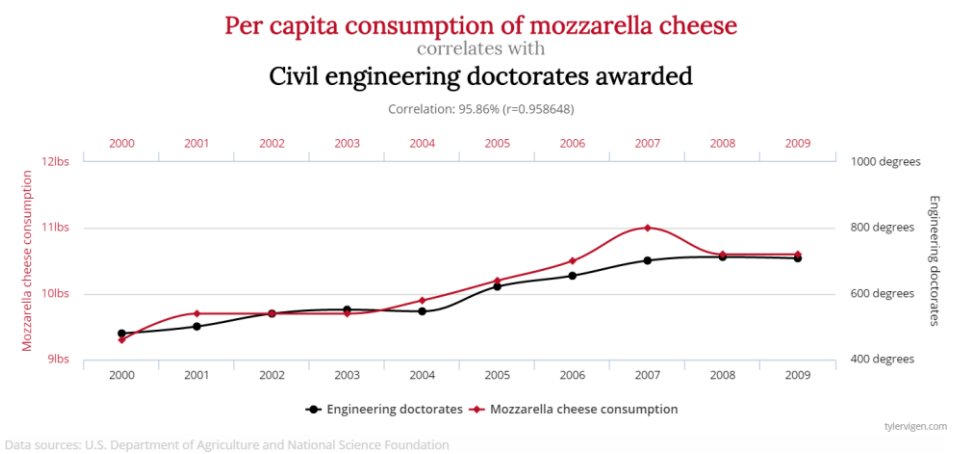


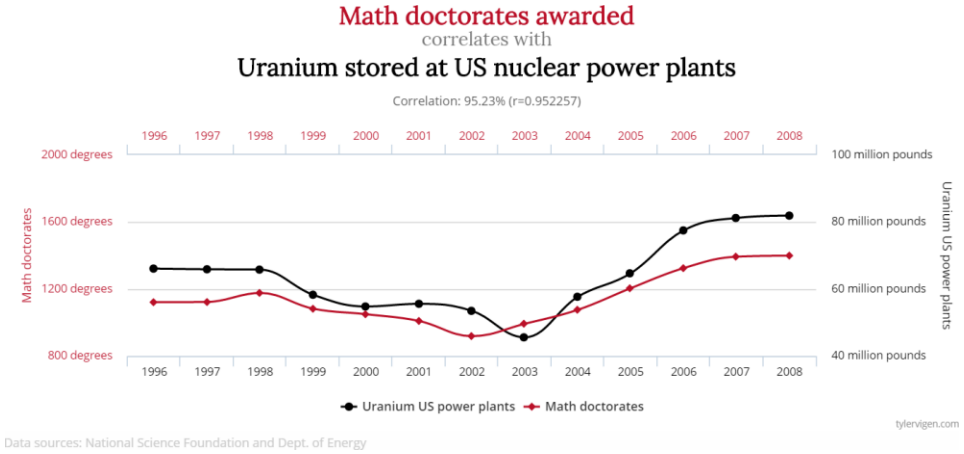


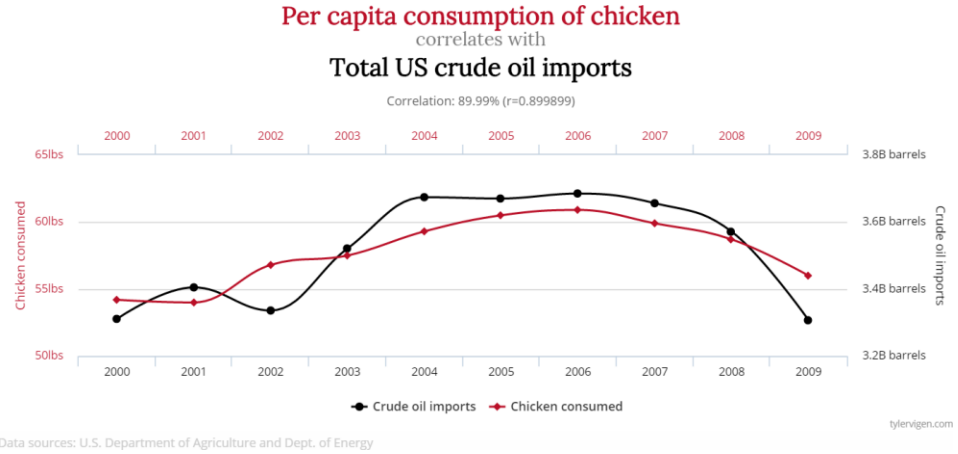
Data Mining, “Fooled by Randomness”, and Spurious Correlations

Spurious Correlations

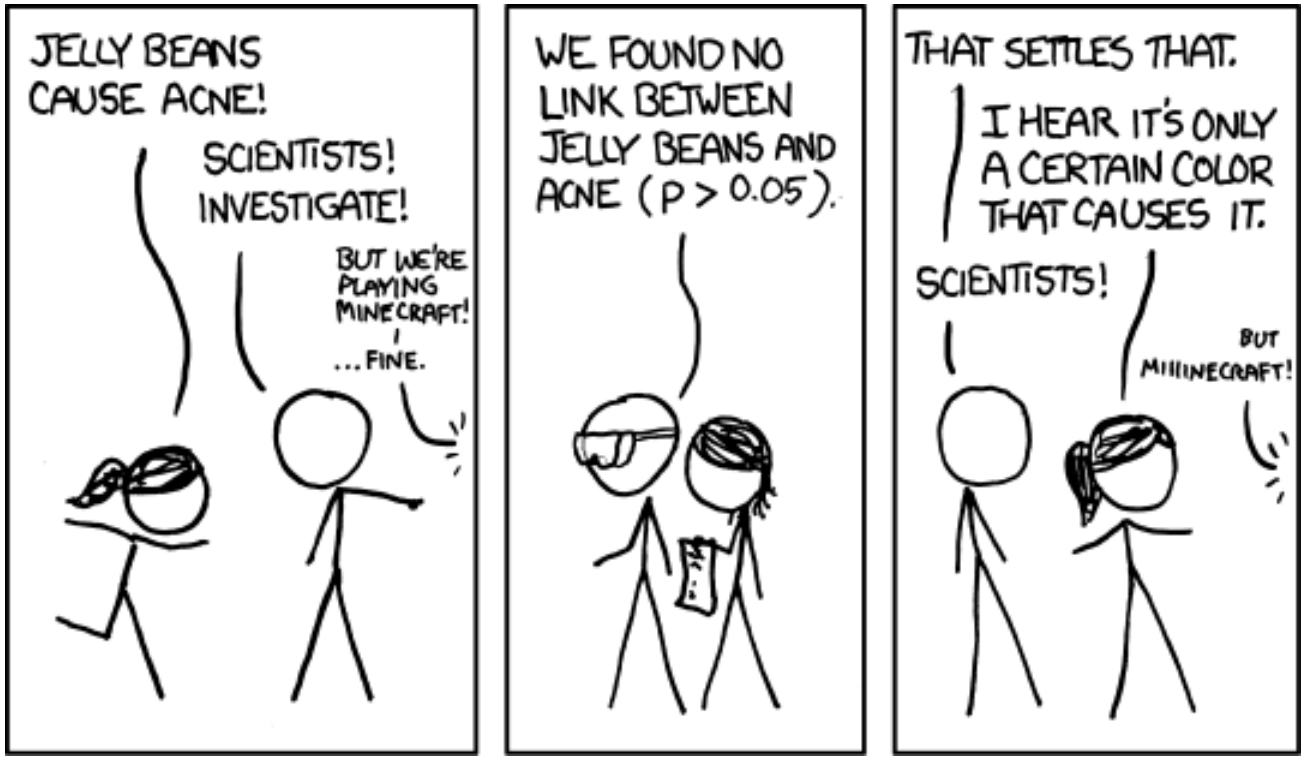




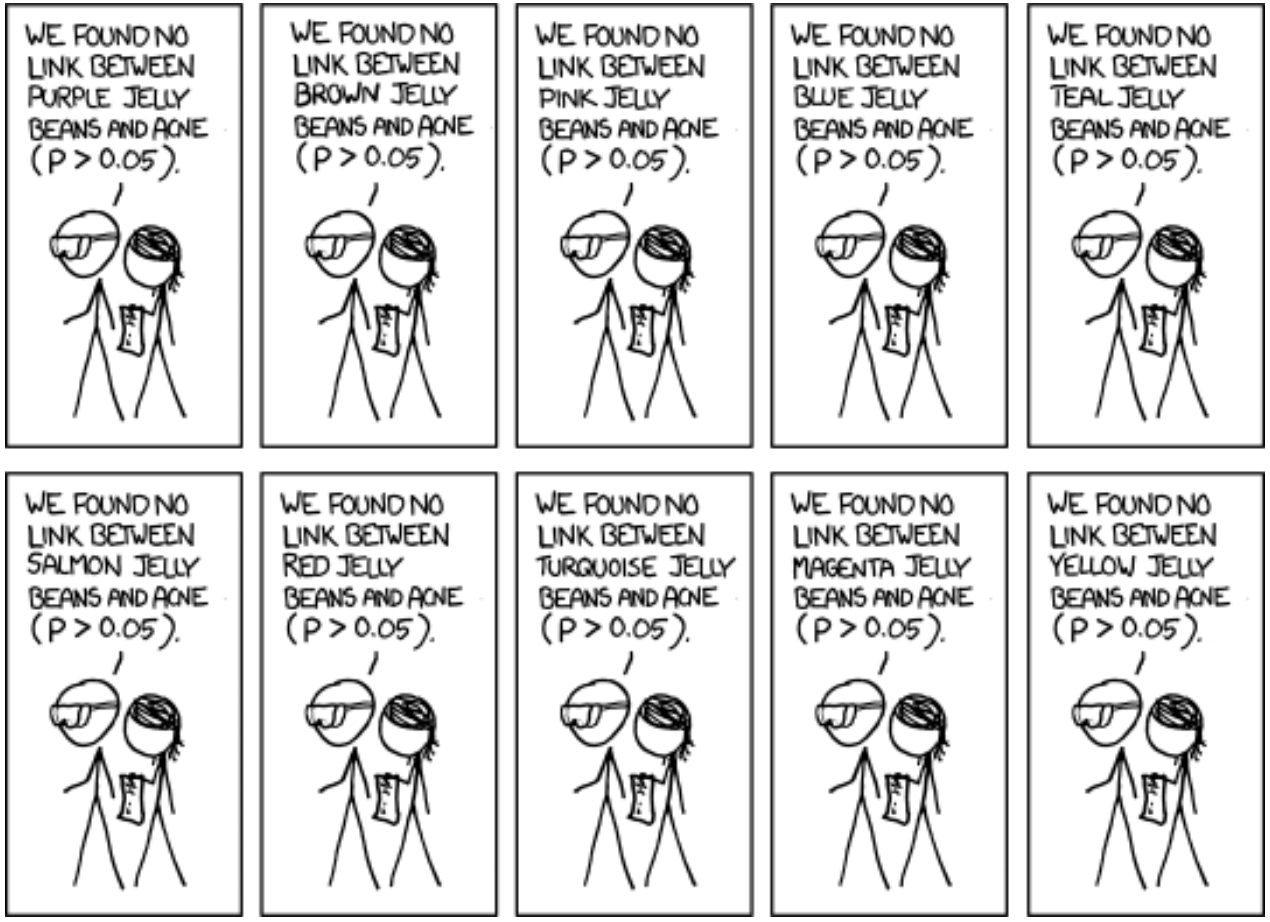




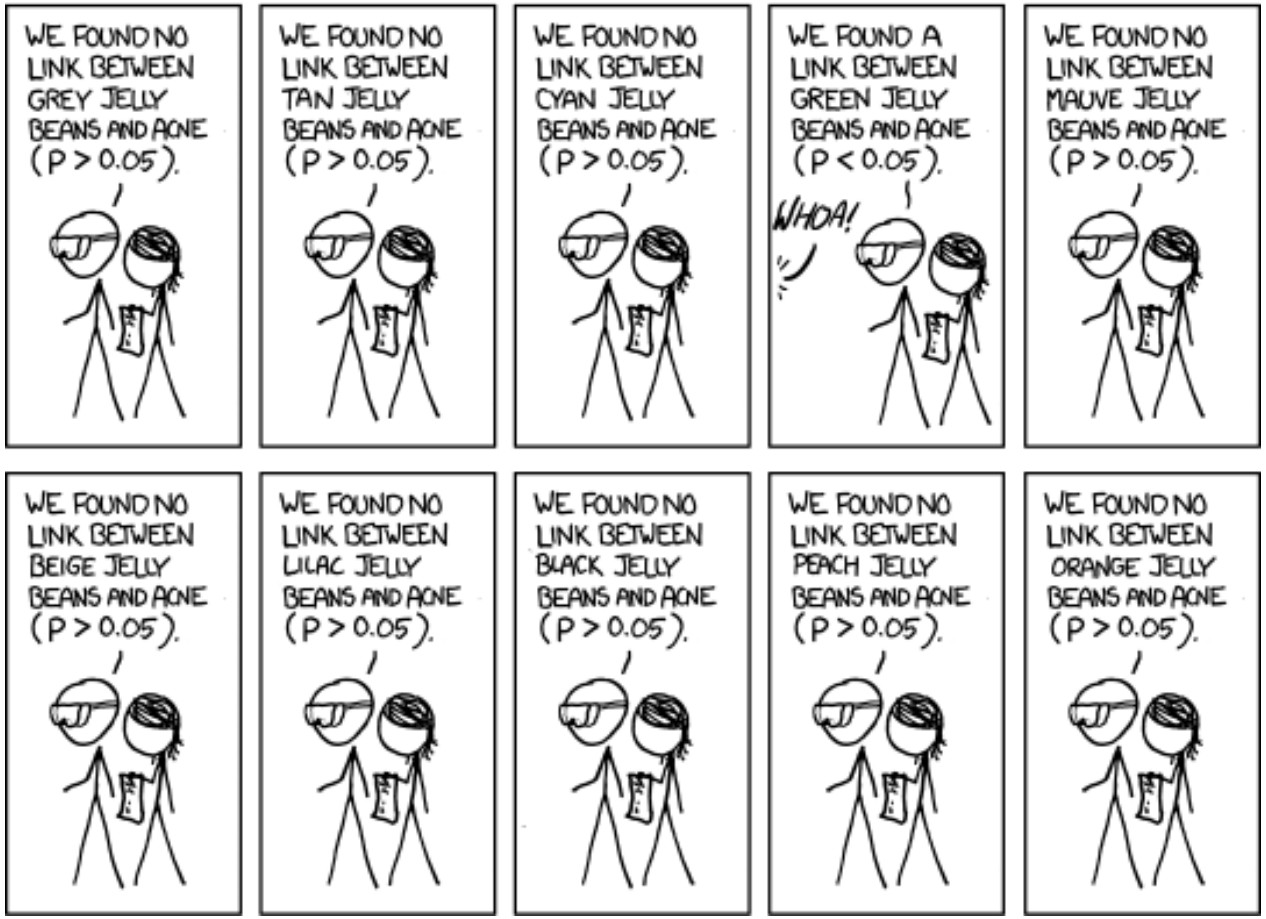
Data Mining


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!

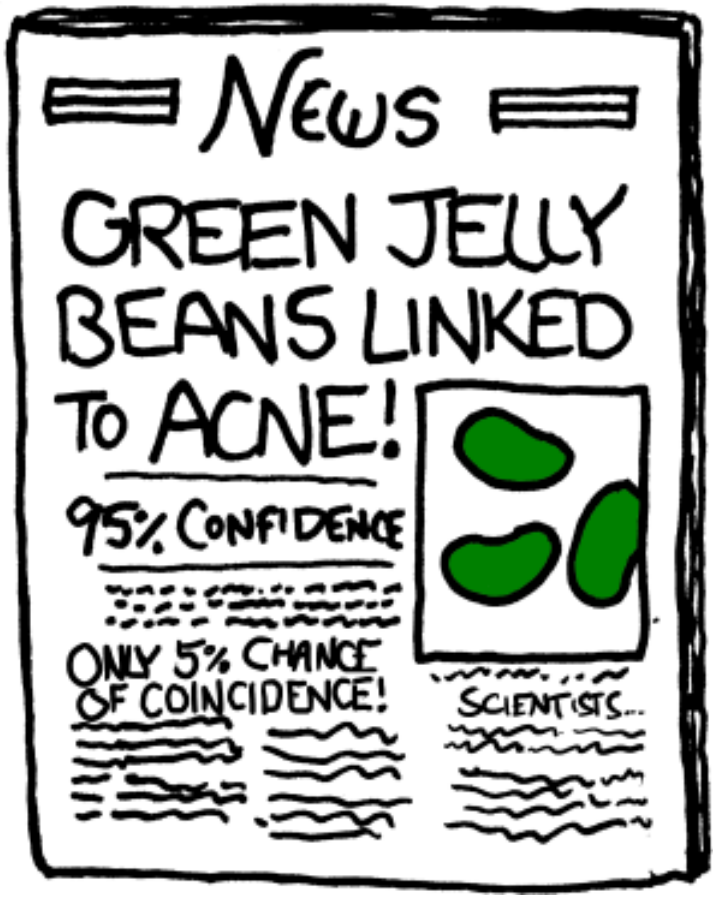
Data Mining (cont.)



Data Mining (cont.)



Data Mining (cont.)



Motivating Example: Codealong

We are using our usual Zillow dataset…

SF-DAT-21 | Codealong 4

Motivating Example

In [1]:
```
import os
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.formula.api as sm

pd.set_option('display.max_rows', 10)
pd.set_option('display.notebook_repr_html', True)
pd.set_option('display.max_columns', 10)

%matplotlib inline
plt.style.use('ggplot')
```

In [2]:
```
df = pd.read_csv(os.path.join('..', 'datasets', 'zillow-04-starter.csv'), index_col = 'ID')
```

Two new variables **M1** and **M2** have been added to the dataset

```
In [3]: df
```

Out[3]:

	Address	DateOfSale	SalePrice	IsAStudio	BedCount	...	Size	LotSize	BuiltInYear	M1	M2
ID											
15063471	55 Vandewater St APT 9, San Francisco, CA	12/4/15	710000	0	1	...	550	NaN	1980	1.099658	0.097627
15063505	740 Francisco St, San Francisco, CA	11/30/15	2150000	0	NaN	...	1430	2435	1948	3.687657	0.430379
15063609	819 Francisco St, San Francisco, CA	11/12/15	5600000	0	2	...	2040	3920	1976	8.975475	0.205527
15064044	199 Chestnut St APT 5, San Francisco, CA	12/11/15	1500000	0	1	...	1060	NaN	1930	2.317325	0.089766
15064257	111 Chestnut St APT 403, San Francisco, CA	1/15/16	970000	0	2	...	1299	NaN	1993	1.380945	-0.152690
...	...	...	...	...	...	...	...	...	...	...	...
2124214951	412 Green St APT A, San Francisco, CA	1/15/16	390000	1	NaN	...	264	NaN	2012	0.428094	-0.804647
2126960082	355 1st St UNIT 1905, San Francisco, CA	11/20/15	860000	0	1	...	691	NaN	2004	1.302833	0.029844
2128308939	33 Santa Cruz Ave, San Francisco, CA	12/10/15	830000	0	3	...	1738	2299	1976	1.608882	0.876824
2131957929	1821 Grant Ave, San Francisco, CA	12/15/15	835000	0	2	...	1048	NaN	1975	1.025920	-0.542707
2136213970	1200 Gough St, San Francisco, CA	1/10/16	825000	0	1	...	900	NaN	1966	1.383641	0.354282

1000 rows × 11 columns

Activity: Knowledge Check

EXERCISE

ANSWER THE FOLLOWING QUESTIONS (10 minutes)

1. Perform some Exploratory Analysis on the these two “mystery” variables `M1` and `M2` and how they relate to `SalePrice`

2. When finished, share your answers with your table

DELIVERABLE

Answers to the above questions

Exploratory Analysis – Correlation

Exploratory Analysis on `M1` and `M2` and how they relate to `SalePrice`

Correlation

In [4]:
```
df[ ['M1', 'M2', 'SalePrice'] ].corr()
```

Out[4]:

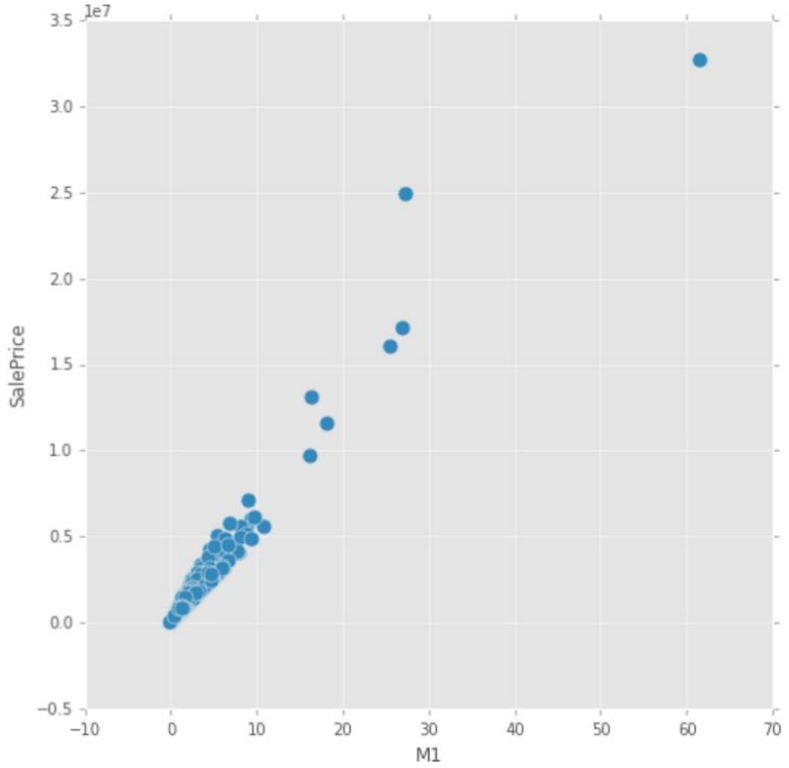
	M1	M2	SalePrice
M1	1.000000	0.166624	0.970612
M2	0.166624	1.000000	0.022003
SalePrice	0.970612	0.022003	1.000000

Exploratory Analysis – Scatter plots

Scatter plots

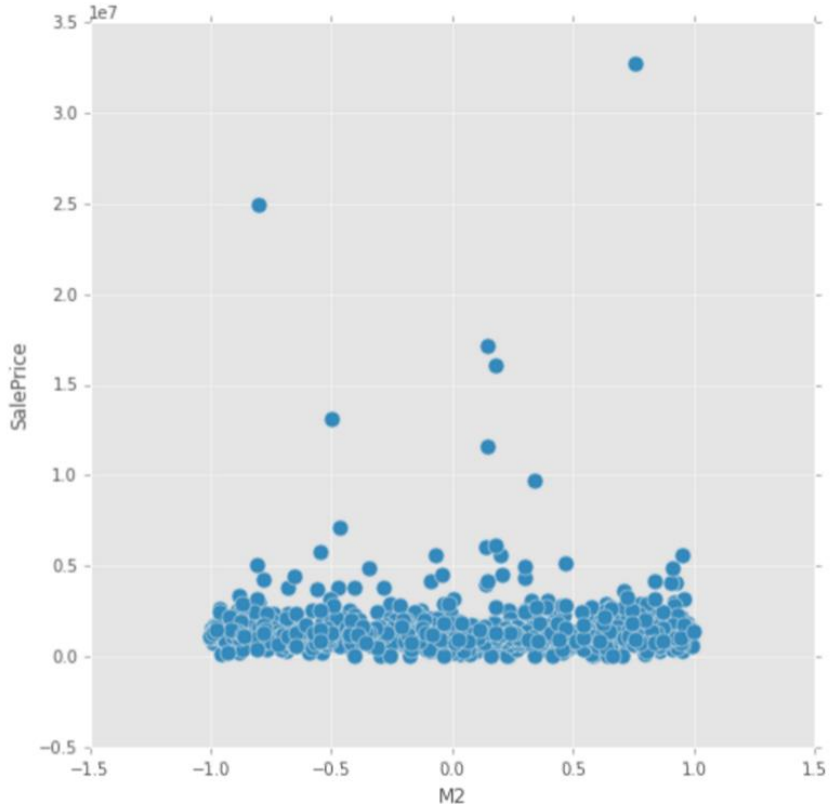
In [5]:
```
df.plot(kind = 'scatter', x = 'M1', y = 'SalePrice', s = 100, figsize = (8, 8))
```

Out[5]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x187e92b0>
```



Exploratory Analysis – Scatter plots (cont.)

```
In [6]: df.plot(kind = 'scatter', x = 'M2', y = 'SalePrice', s = 100, figsize = (8, 8))
Out[6]: <matplotlib.axes._subplots.AxesSubplot at 0x18af3630>
```



Out first Machine Learning Model – `SalePrice` as a function of `M1`

Your first Machine Learning Models!

SalePrice as a function of M1

In [7]:
```
X = df[ ['M1'] ]
y = df['SalePrice']

fit = sm.OLS( y, X ).fit()
```

How do we interpret these results?

```
In [8]: fit.summary()
```

Out[8]: OLS Regression Results

Dep. Variable:	SalePrice	**R-squared:**	0.963
Model:	OLS	**Adj. R-squared:**	0.963
Method:	Least Squares	**F-statistic:**	2.567e+04
Date:	Mon, 07 Mar 2016	**Prob (F-statistic):**	0.00
Time:	13:29:46	**Log-Likelihood:**	-14393.
No. Observations:	1000	**AIC:**	2.879e+04
Df Residuals:	999	**BIC:**	2.879e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
M1	6.241e+05	3894.990	160.228	0.000	6.16e+05 6.32e+05

Omnibus:	1044.296	**Durbin-Watson:**	1.921
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	901486.247
Skew:	3.948	**Prob(JB):**	0.00
Kurtosis:	149.879	**Cond. No.**	1.00

How do we interpret these results? (cont.)

```
In [8]: fit.summary()
```

Out[8]: OLS Regression Results

Dep. Variable:	SalePrice	**R-squared:**	0.963
Model:	OLS	**Adj. R-squared:**	0.963
Method:	Least Squares	**F-statistic:**	2.567e+04
Date:	Mon, 07 Mar 2016	**Prob (F-statistic):**	0.00
Time:	13:29:46	**Log-Likelihood:**	-14393.
No. Observations:	1000	**AIC:**	2.879e+04
Df Residuals:	999	**BIC:**	2.879e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
M1	6.241e+05	3894.990	160.228	0.000	6.16e+05 6.32e+05

Omnibus:	1044.296	**Durbin-Watson:**	1.921
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	901486.247
Skew:	3.948	**Prob(JB):**	0.00
Kurtosis:	149.879	**Cond. No.**	1.00

`SalePrice` as a function of `M1`. But how good is it?

$$SalePrice = 6.241 \times 10^5 \times M1$$

How do we interpret these results? (cont.)

```
In [8]: fit.summary()
```

Out[8]: OLS Regression Results

Dep. Variable:	SalePrice	**R-squared:**	0.963
Model:	OLS	**Adj. R-squared:**	0.963
Method:	Least Squares	**F-statistic:**	2.567e+04
Date:	Mon, 07 Mar 2016	**Prob (F-statistic):**	0.00
Time:	13:29:46	**Log-Likelihood:**	-14393.
No. Observations:	1000	**AIC:**	2.879e+04
Df Residuals:	999	**BIC:**	2.879e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std er	t	P>|t|	[95.0% Conf. Int.]
M1	6.241e+05	3894.9	160.228	0.000	6.16e+05 6.32e+05

Omnibus:	1044.296	**Durbin-Watson:**	1.921
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	901486.247
Skew:	3.948	**Prob(JB):**	0.00
Kurtosis:	149.879	**Cond. No.**	1.00

SalePrice as a function of M2

SalePrice as a function of M2

In [9]:
```
X = df[ ['M2'] ]
y = df['SalePrice']

fit = sm.OLS( y, X ).fit()
```

In [10]:
```
fit.summary()
```

Out[10]:

OLS Regression Results

Dep. Variable:	SalePrice	**R-squared:**	0.000
Model:	OLS	**Adj. R-squared:**	-0.001
Method:	Least Squares	**F-statistic:**	0.06941
Date:	Mon, 07 Mar 2016	**Prob (F-statistic):**	0.792
Time:	13:29:46	**Log-Likelihood:**	-16036.
No. Observations:	1000	**AIC:**	3.207e+04
Df Residuals:	999	**BIC:**	3.208e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std e	t	P>|t|	[95.0% Conf. Int.]
M2	3.195e+04	1.216e+05	0.263	0.792	-2.06e+05 2.7e+05

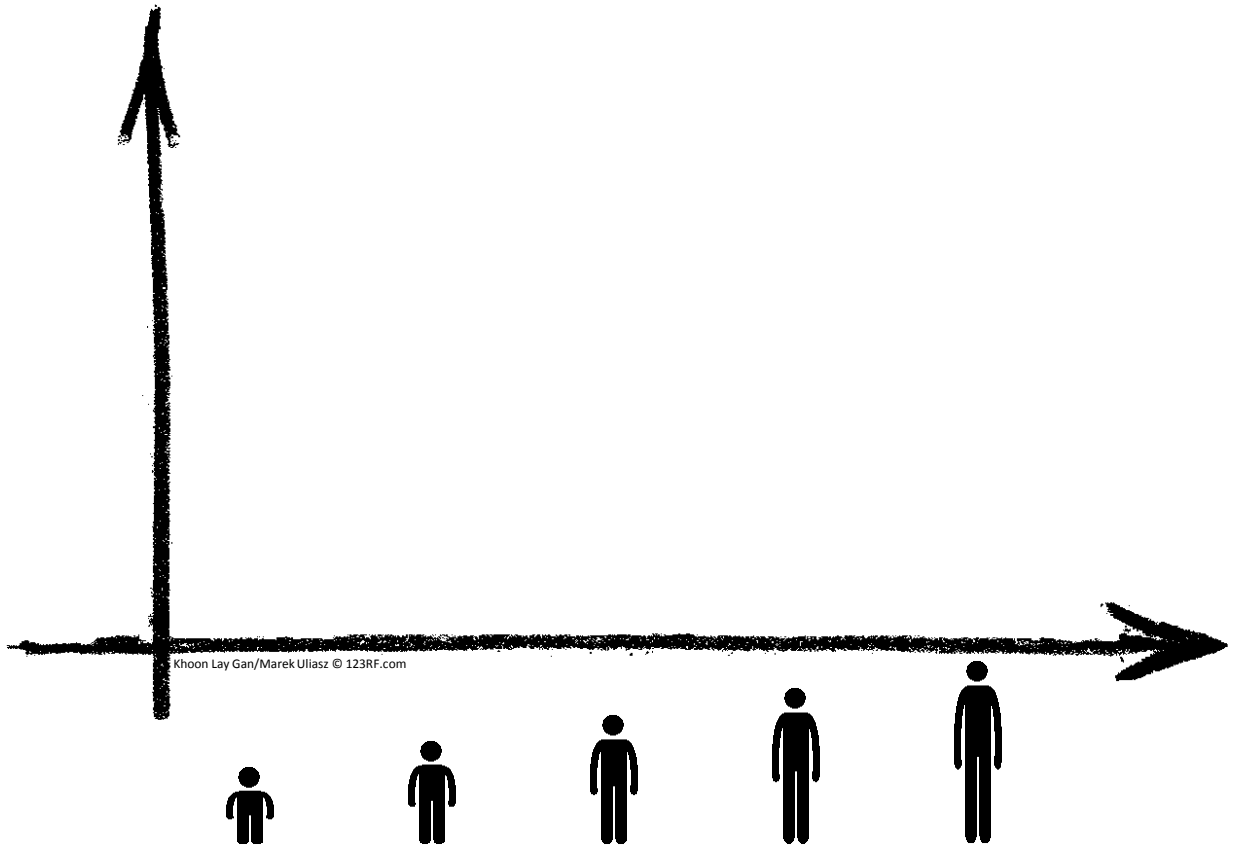
Omnibus:	1664.600	**Durbin-Watson:**	0.971
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	986904.813
Skew:	10.532	**Prob(JB):**	0.00
Kurtosis:	155.453	**Cond. No.**	1.00

Do these coefficients make sense? From a statistical standpoint, are they “significant”?

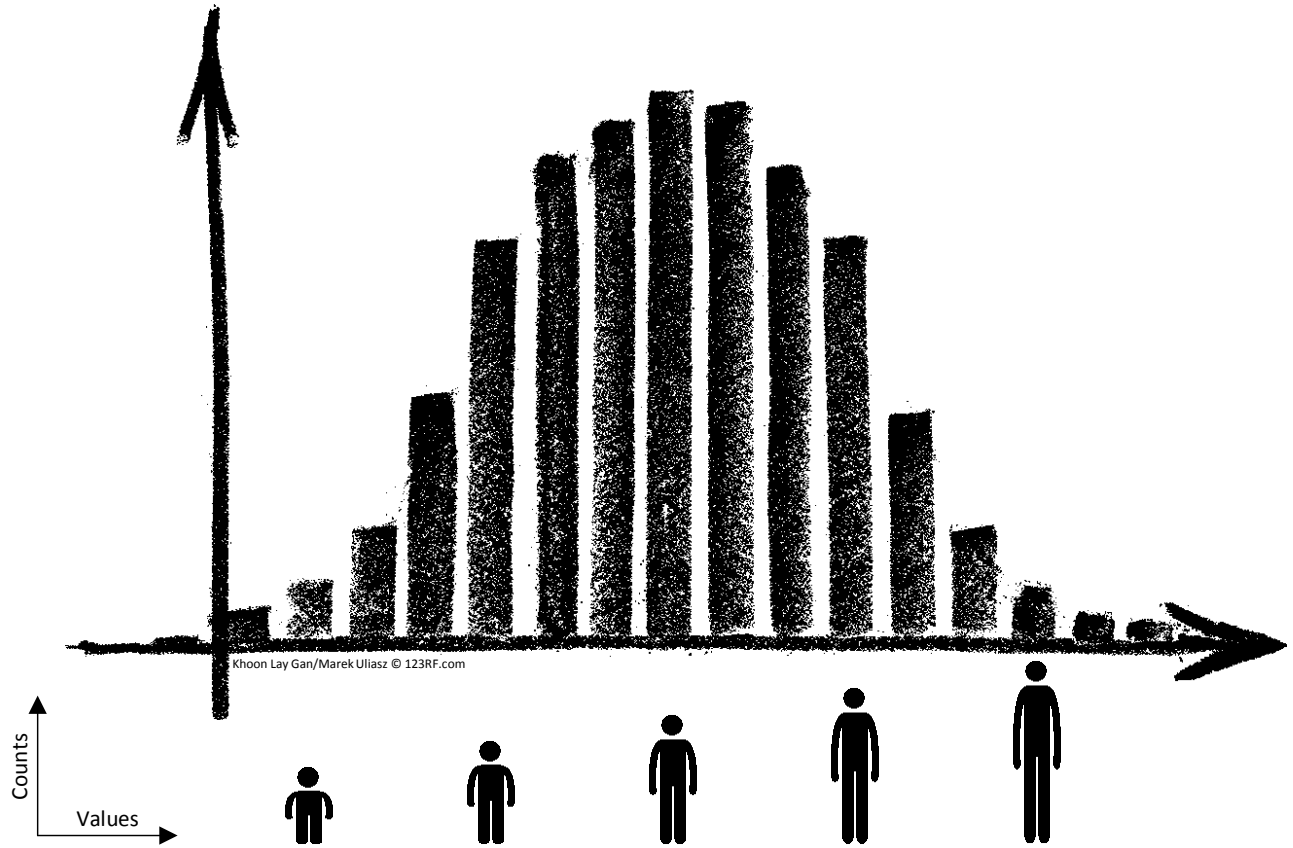
	coef	std err	t	P>|t|	[95.0% Conf. Int.]
M1	6.241e+05	3894.990	160.228	0.000	6.16e+05 6.32e+05
M2	3.195e+04	1.21e+05	0.263	0.792	-2.06e+05 2.7e+05

The Normal Distribution

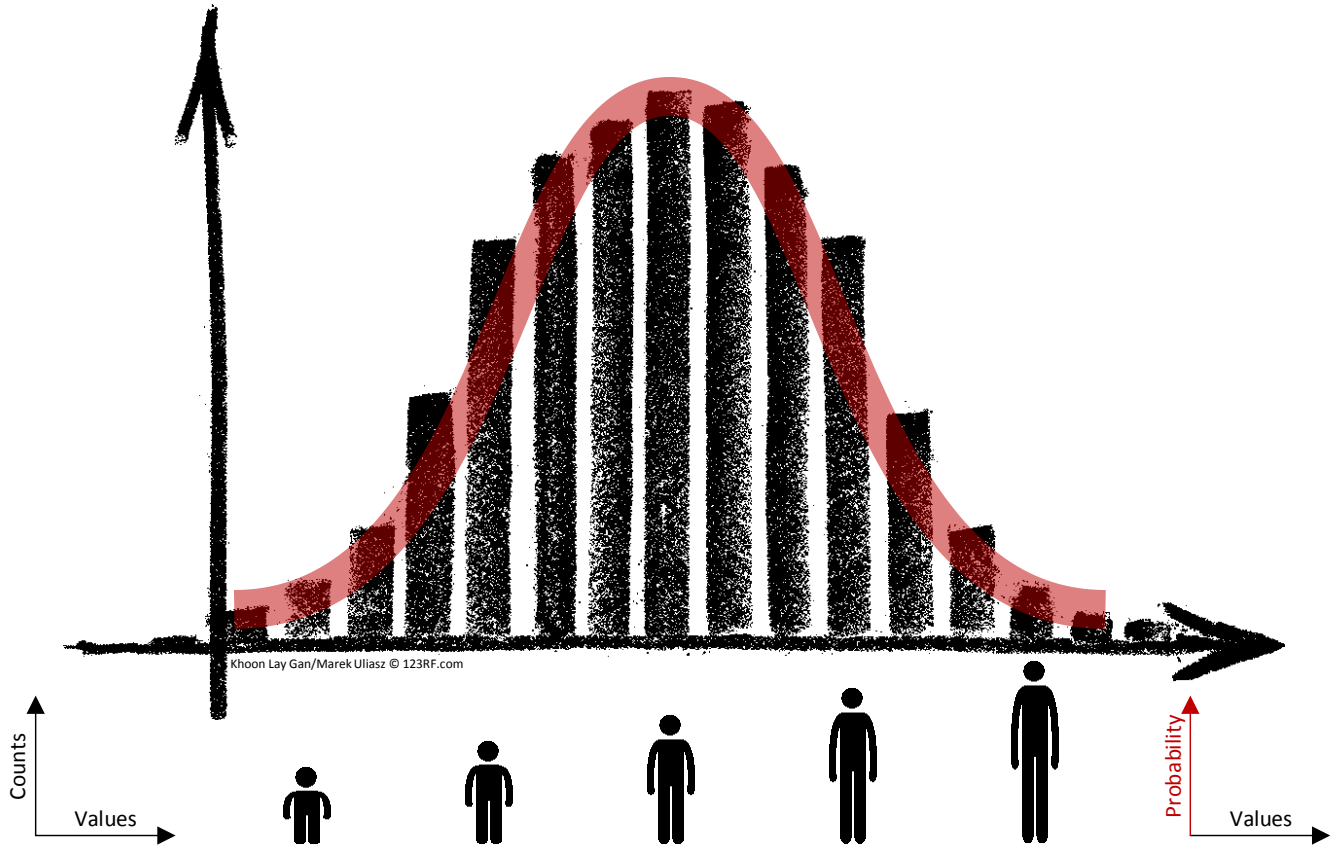
How is people’s height distributed?



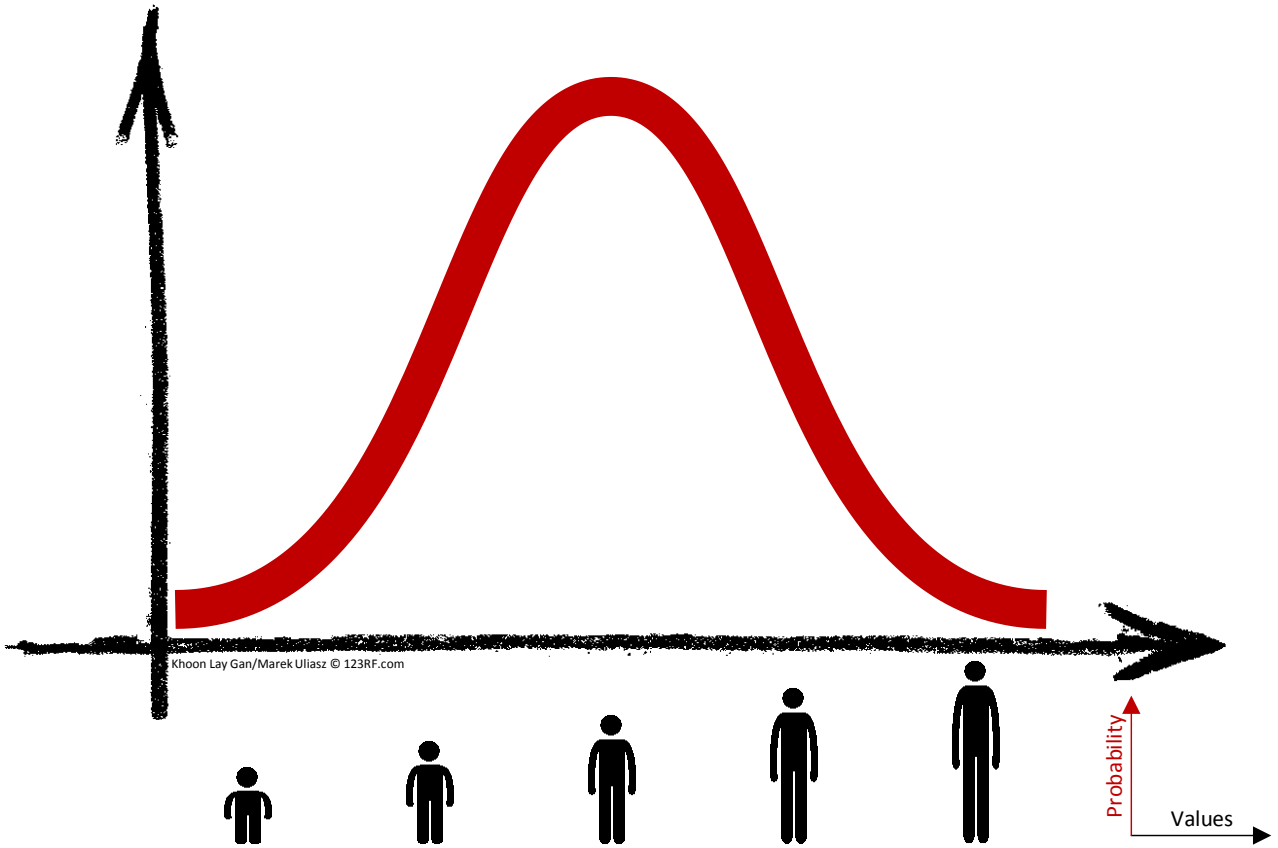
People’s height follows a bell shape distribution. (For men, the average height is around 70 inches (5-10), with few people shorter than 67 inches, and few as tall as 73 inches)



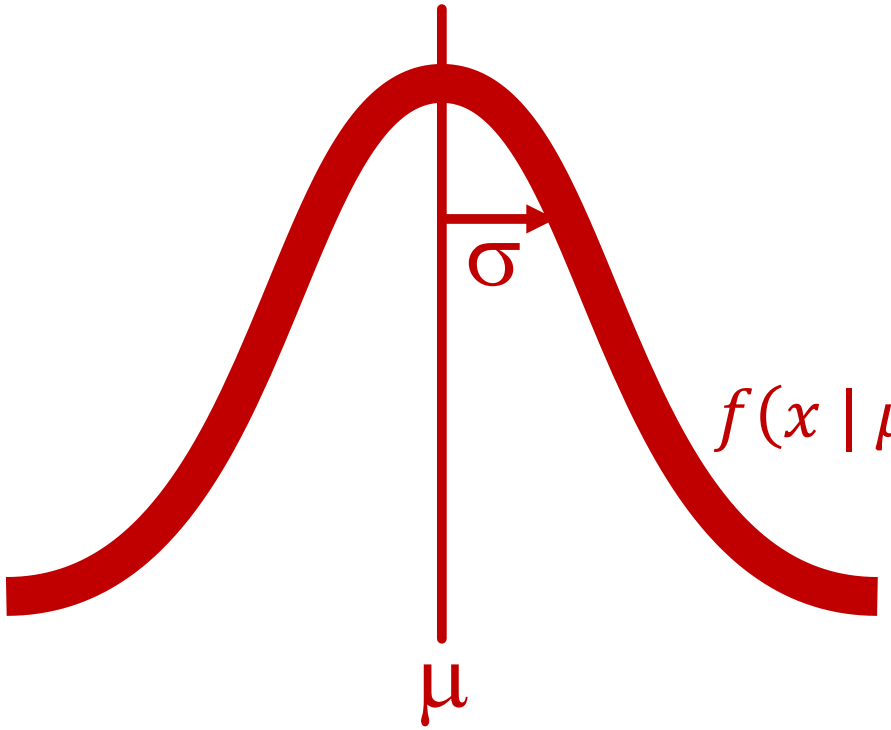
People’s height follows a bell shape distribution (cont.)



The Normal Distribution

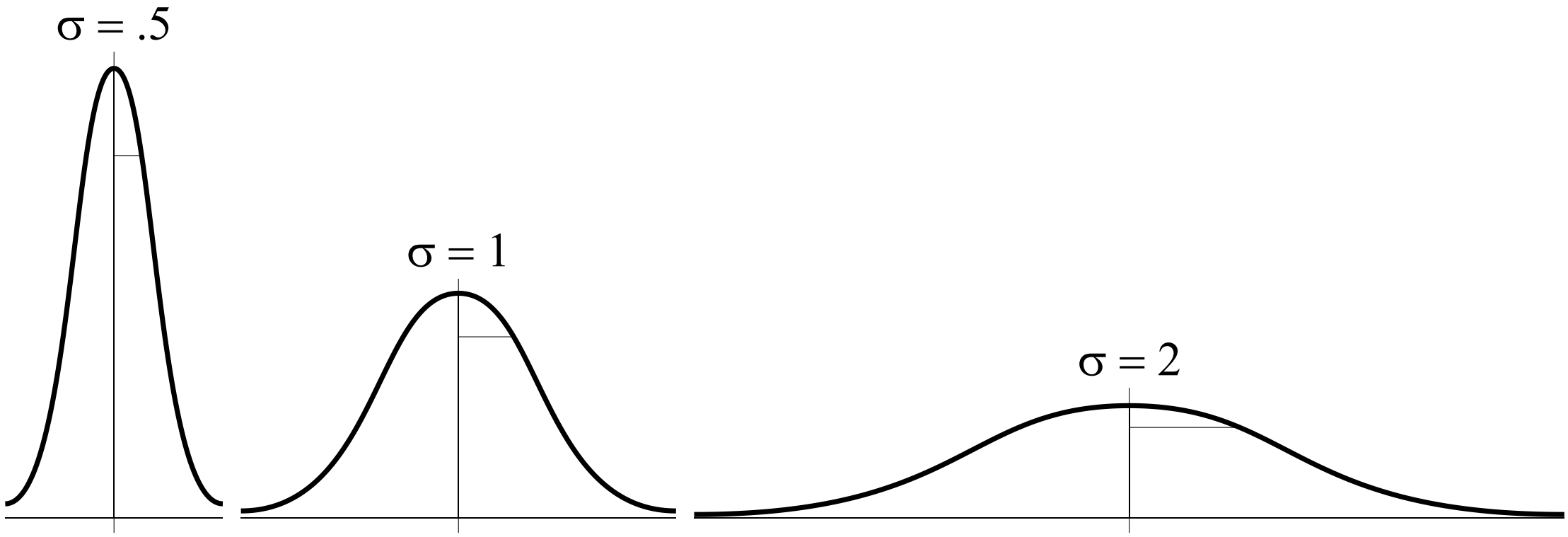


The Normal Distribution (cont.)

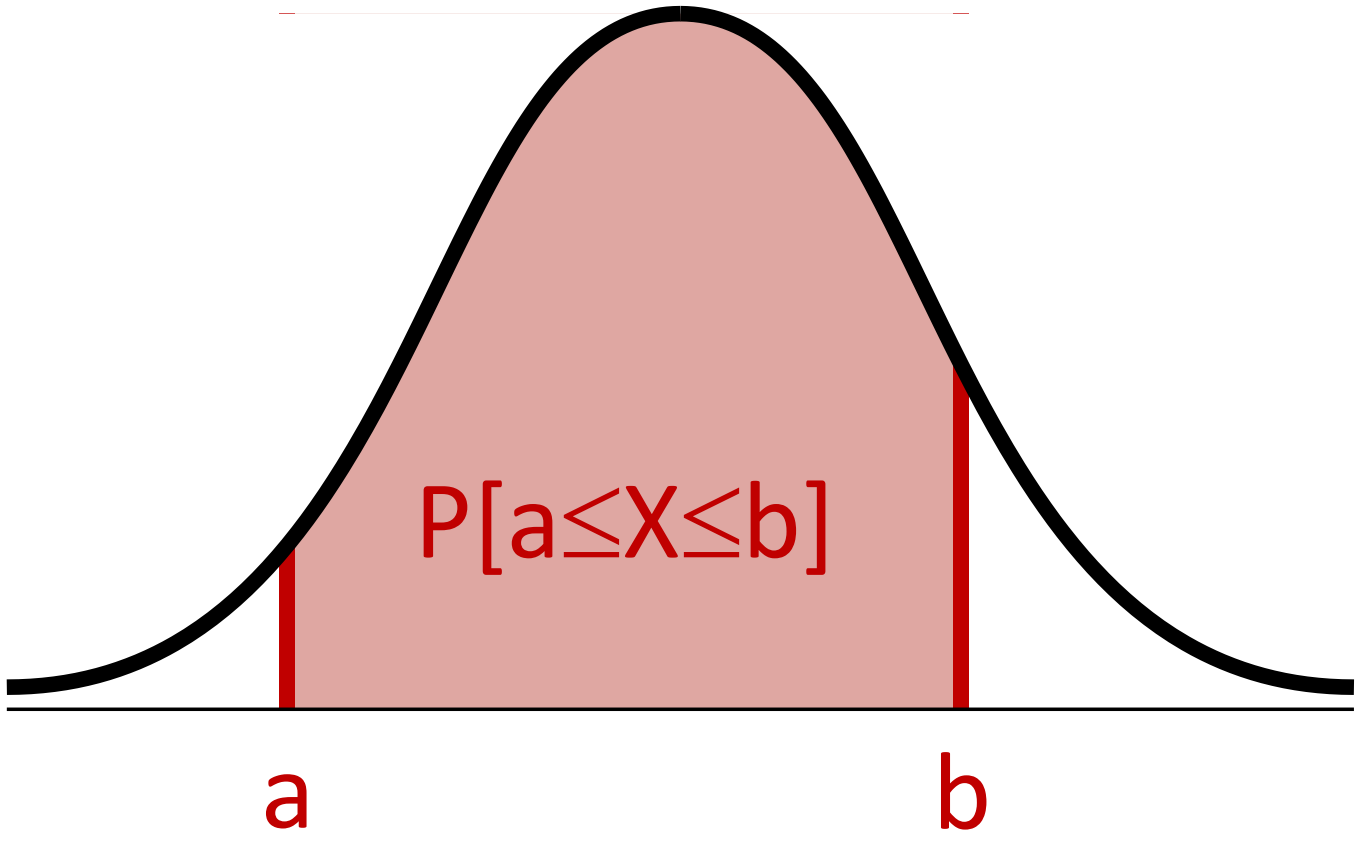


$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

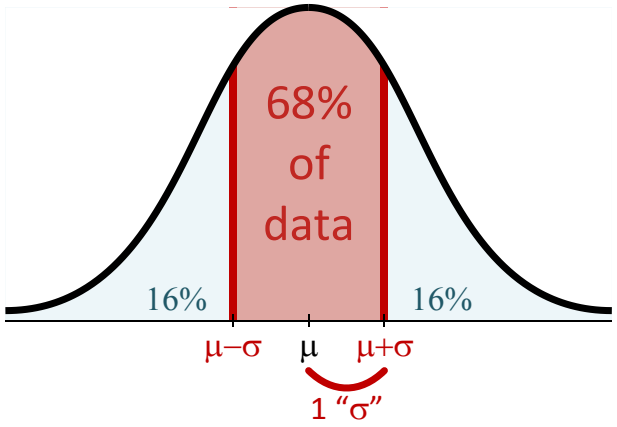
This is a probability density function: The area under the curve is always 1 (for any σ)

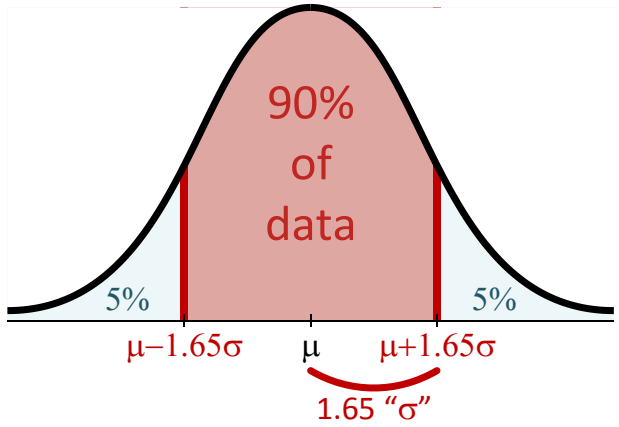


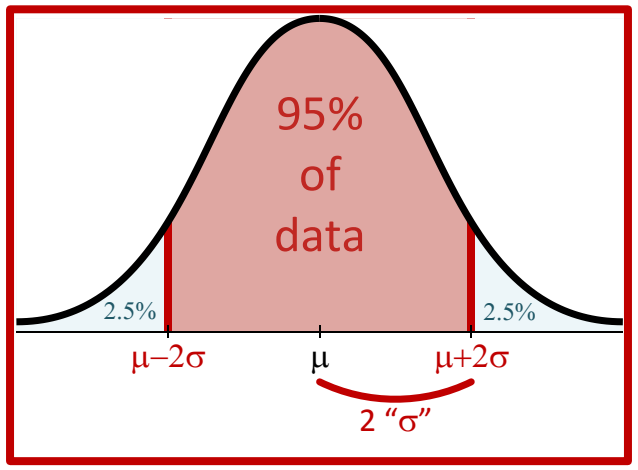
How to read a probability density function?

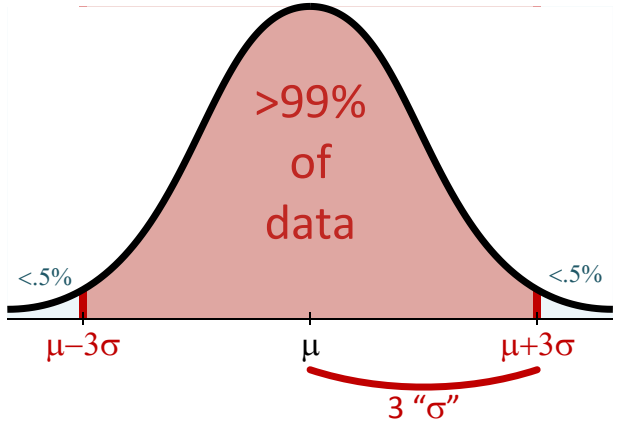


The 68 – 90 – 95 – 99.7 Rule









Hypothesis Testing

- A hypothesis is an assumption about the a population parameter. E.g.,
  - M1's coefficient is 6.241×10^5
  - M2's coefficient is 3.195×10^4
- In both cases, we made a statement about a population parameter that may or may not be true
- The purpose of hypothesis testing is to make a statistical conclusion about **rejecting** or **failing to reject** such statement

Two-Tail Hypothesis Test

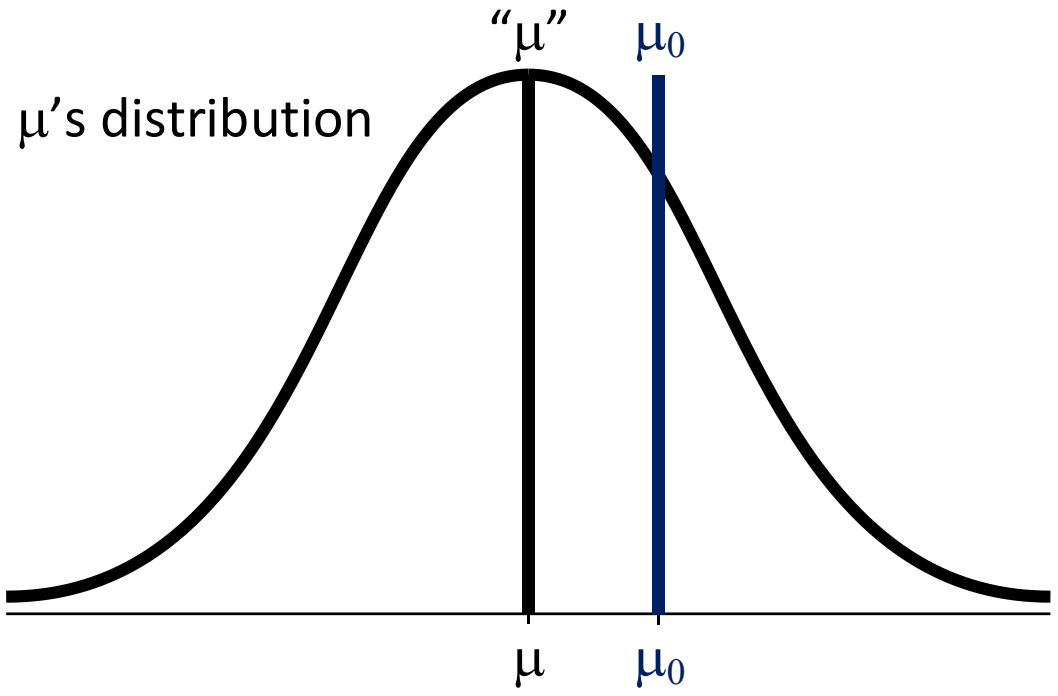
- The *null hypothesis* (H_0) represents the status quo; that the mean of the population is equal to a specific value:

$$H_0: \mu = \mu_0$$

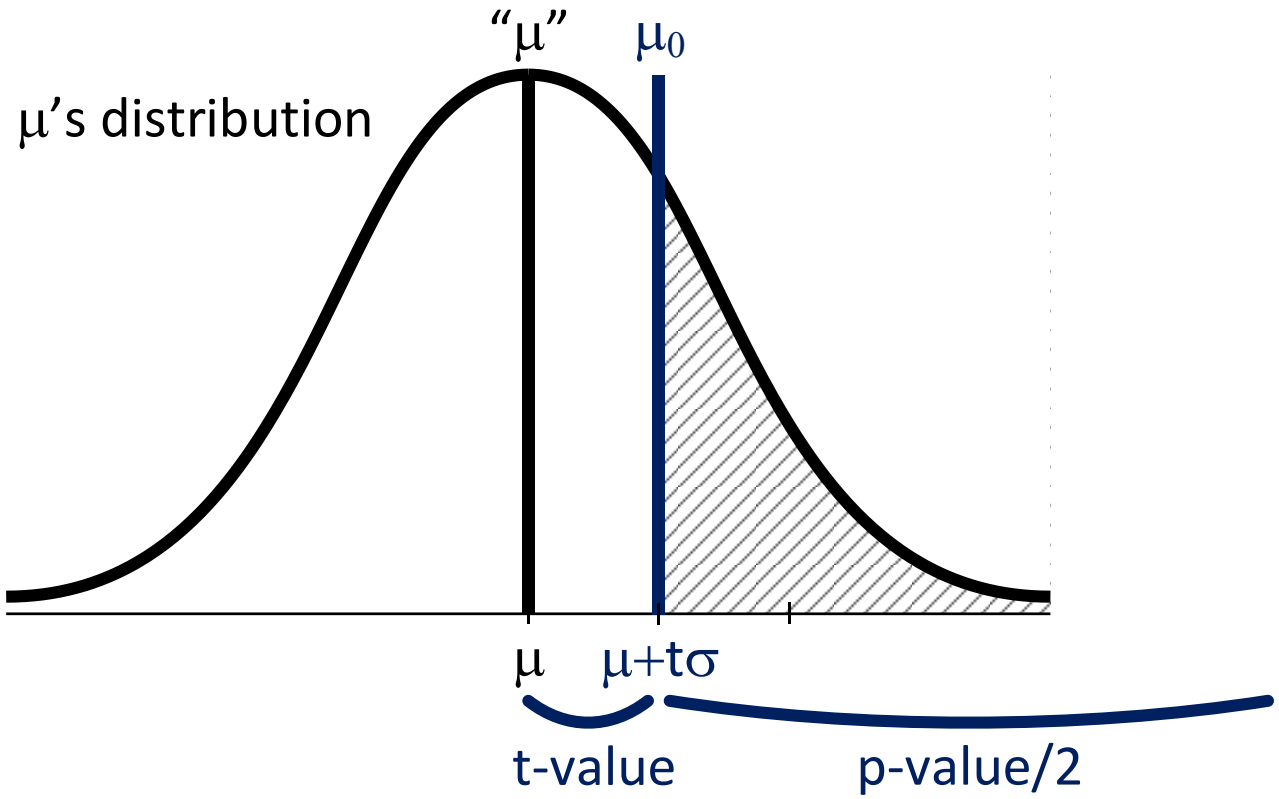
- The *alternate hypothesis* (H_a) represents the opposite of the null hypothesis and holds true if the *null hypothesis* is found to be false:

$$H_a: \mu \neq \mu_0$$

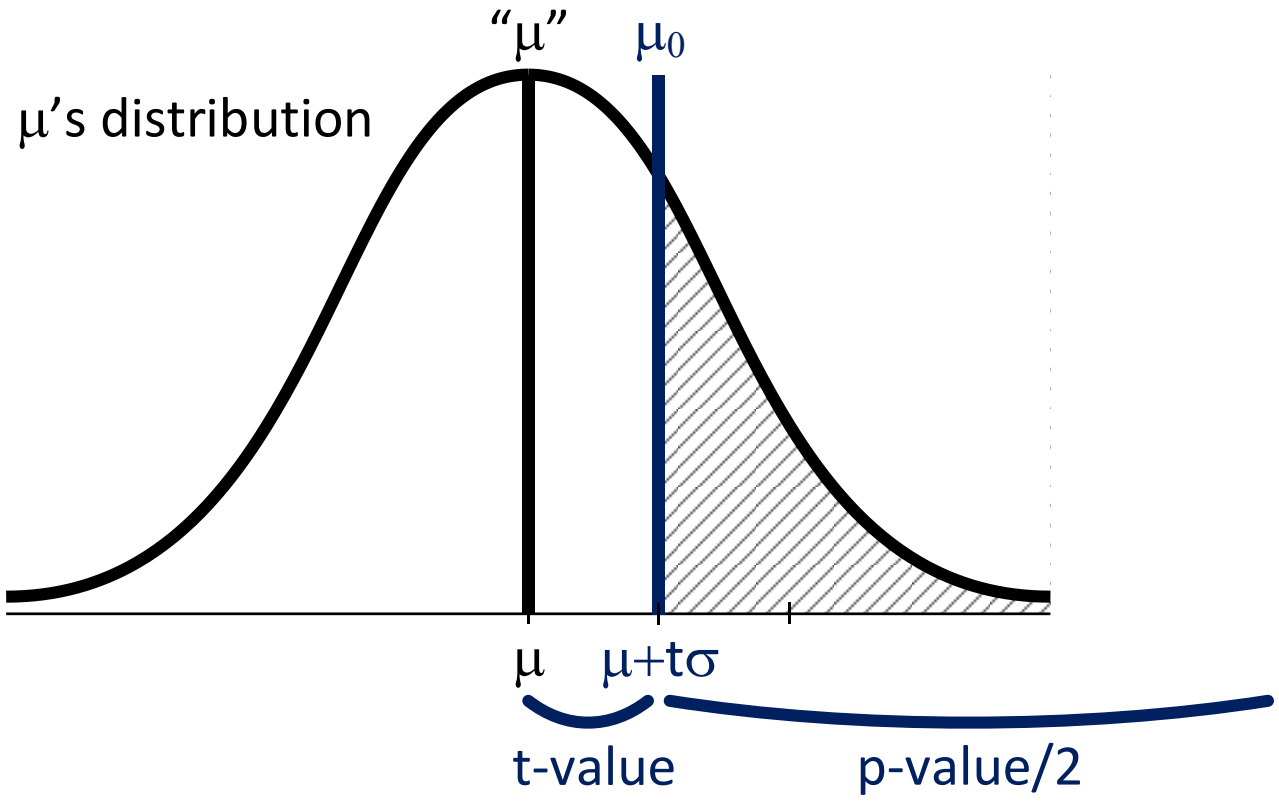
Two-Tail Hypothesis Test (cont.)



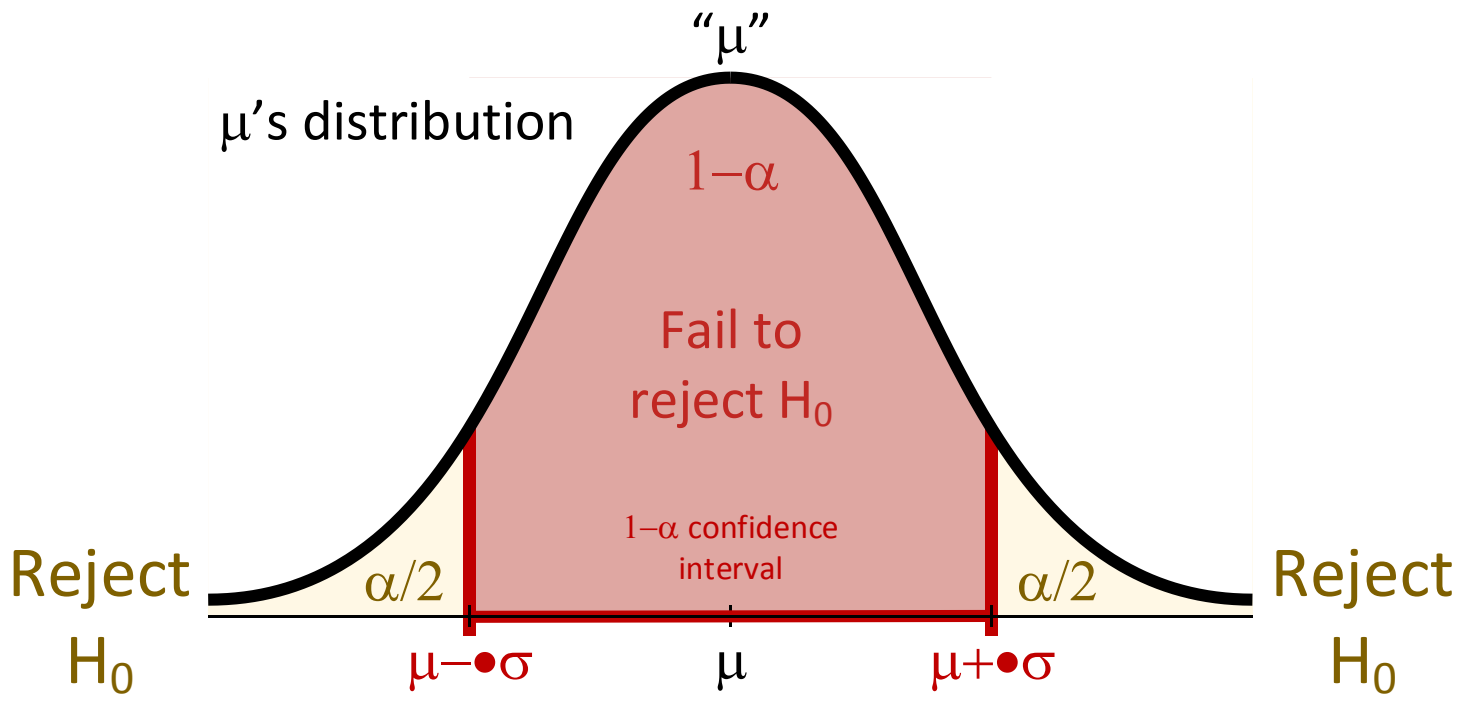
t-value measures the difference to μ_0 in σ. t-values of large magnitudes (either negative or positive) are less likely. The far left and right "tails" of the distribution curve represent instances of obtaining extreme values of t, far from μ



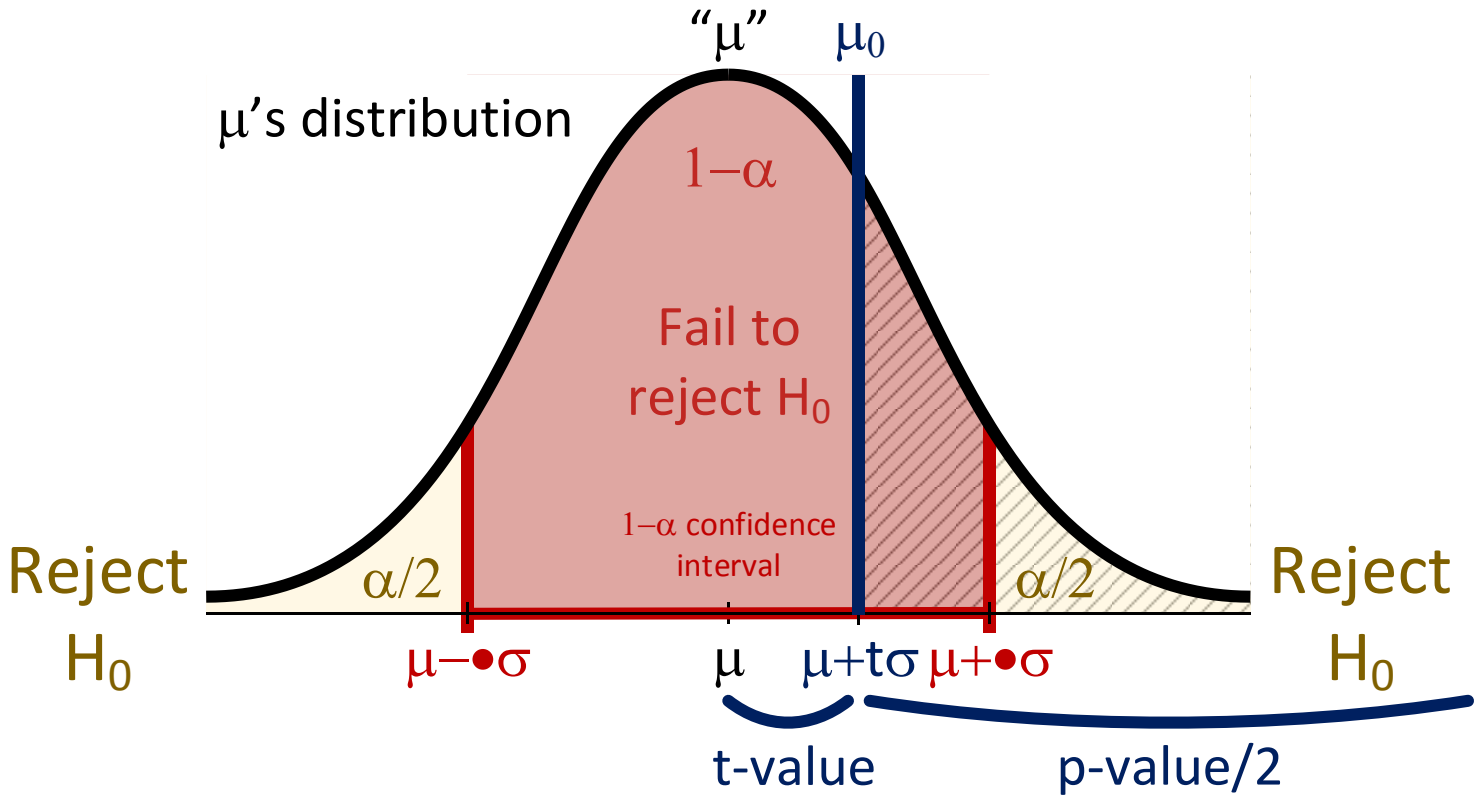
p-value determines the probability (assuming the H_0 is true) of observing a more extreme test statistic in the direction of H_a than the one observed



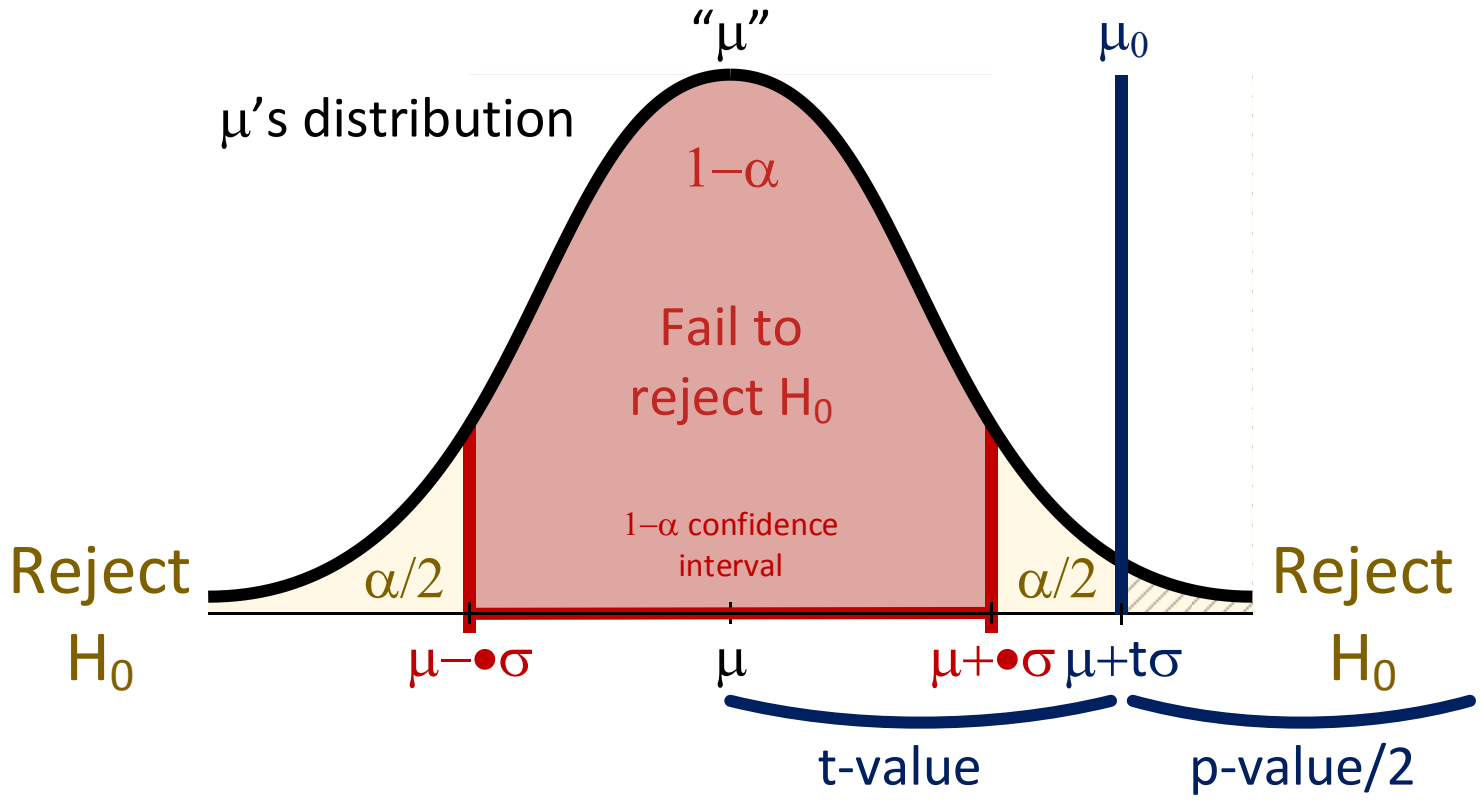
Two-Tail Hypothesis Test (simplified) (cont.)



Two-Tail Hypothesis Test (cont.)



Two-Tail Hypothesis Test (cont.)



Two-Tail Hypothesis Test (cont.)

|t-value|	p-value	$1-\alpha$ Confidence Interval ($[\mu_0 -\cdot \sigma, \mu_0 +\cdot \sigma]$)	H_0 / H_a	Conclusion
$\geq \cdot$	$\leq \alpha$	μ_0 is outside	Found evidence that $\mu \neq \mu_0$: Reject H_0	$\mu \neq \mu_0$
$< \cdot$	$> \alpha$	μ_0 is inside	Did not find that $\mu \neq \mu_0$: Fail to reject H_0	$\mu = \mu_0$

Two-Tail Hypothesis Test ($\alpha = .05\%$) (cont.)

|t-value|	p-value	95% Confidence Interval ($[\mu_0 - 2\sigma, \mu_0 +$	H_0 / H_a	Conclusion
$\geq$ "~ 2"(*) (*) (check the t-table slide)	$\leq .025$	μ_0 is outside	Found evidence that $\mu \neq \mu_0$: Reject H_0	$\mu \neq \mu_0$
$<$ "~ 2"(*)	$> .025$	μ_0 is inside	Did not find that $\mu \neq \mu_0$: Fail to reject H_0	$\mu = \mu_0$ (assume)

Activity

EXERCISE

ANSWER THE FOLLOWING QUESTIONS (5 minutes)

1. What are the *null* and *alternate hypothesis* for the `M1` and `M2` coefficients? (Hint: What makes these coefficients "statistically" significant?)

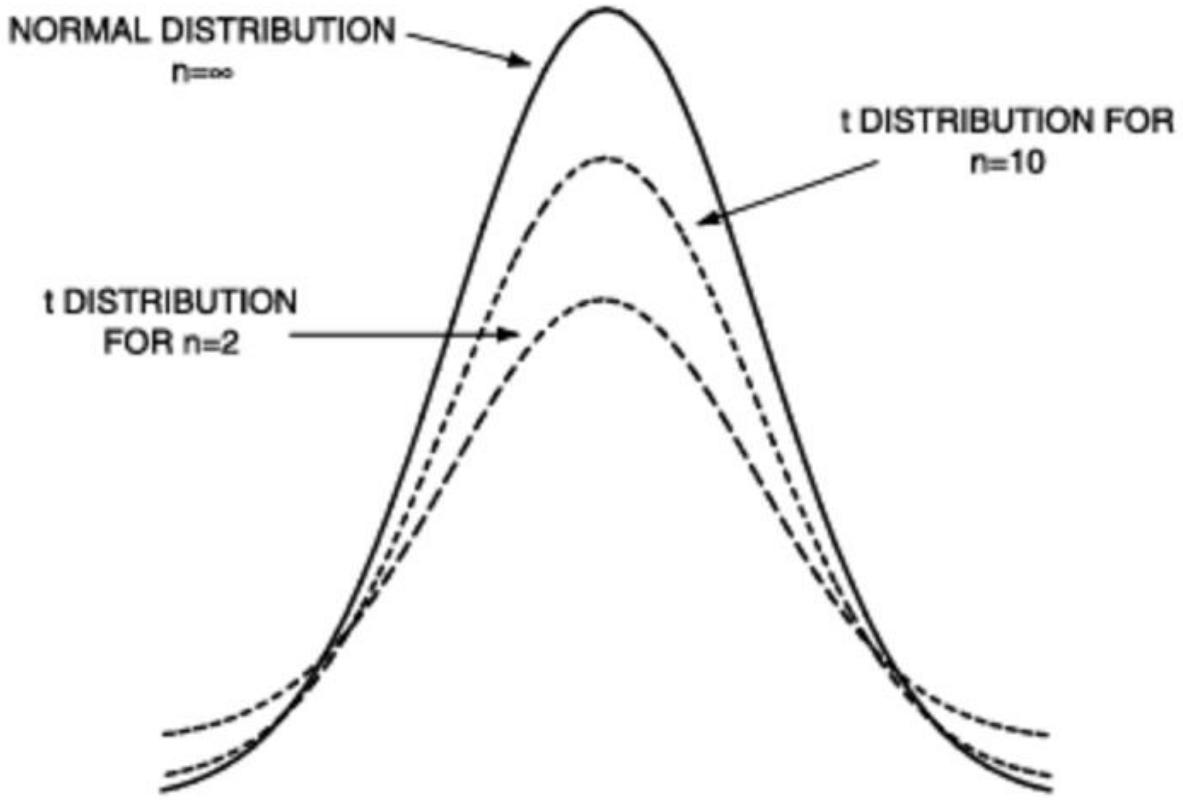
	coef	std err	t	P>|t|	[95.0% Conf. Int.]
M1	6.241e+05	3894.990	160.228	0.000	6.16e+05 6.32e+05
M2	3.195e+04	1.21e+05	0.263	0.792	-2.06e+05 2.7e+05

2. When finished, share your answers with your table

DELIVERABLE

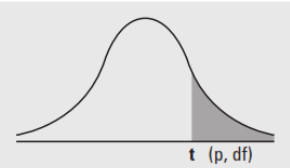
Answers to the above questions

FYI: t-tests use the Student's t-distribution, not the normal distribution…



Student's t-distribution table (cont.)

Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	———	———	80%	90%	95%	98%	99%	99.9%

Lab

Today's Closing Thought

ScienceNews
MAGAZINE OF THE SOCIETY FOR SCIENCE & THE PUBLIC

Context

SCIENCE PAST AND PRESENT
TOM SIEGFRIED

CONTEXT NUMBERS

P value ban: small step for a journal, giant leap for science

Editors reject flawed system of null hypothesis testing

BY TOM SIEGFRIED 3:18PM, MARCH 17, 2015

Imagine, if you dare, a world without P values.

Perhaps you're already among the lucky participants in the human race who don't know what a P value is. Trust me, you don't want to. P stands for pernicious, and P values are at the root of all (well, most) scientific evil.

Of course, I don't mean evil in the sense of James Bond's villains. It's an unintentional evil, but nevertheless a diabolical conspiracy of ignorance that litters the scientific literature with erroneous results. P values are supposed to help scientists decide whether an apparently meaningful experimental result is really just a fluke. But in fact, P values confuse more than they clarify. They are misused, misunderstood and misrepresented.

But now somebody is finally trying to do something about it.

Last month a scientific journal — *Basic and Applied Social Psychology* — announced that it won't publish papers that mention the unmentionable P value. No longer will the journal permit published papers to report the P value's use in the process of "null hypothesis testing," which psychologists and scientists in many other fields routinely rely on. Anyone embarking on a research career soon gets infected with this method. When you want to test to see whether a food additive causes cancer, or a medicine cures a disease, you assume that it doesn't — the null hypothesis — and then do an experiment comparing the drug or medicine with a placebo, or another drug, or whatever. If more people survive with the medicine than with the placebo, maybe the medicine works. Or maybe that result was a fluke — the luck of the draw. P values supposedly tell you whether the difference you saw was luck or reality.

Except that they don't. P value calculations tell you only the probability of seeing a result at least as big as what you saw *if there is no real effect*. (In other words, the P value calculation assumes the null hypothesis is true.) A small P value — low probability of the data you measured — might mean the null hypothesis is wrong, or it might mean that you just saw some unusual data. You don't know which. And if there *is* a real effect, your calculation of a P value is rendered meaningless, because that calculation assumed that there *wasn't* a real effect.

> "The use of P values and null hypothesis testing is 'surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students.'"
>
> – *William Rozeboom*

Nevertheless, the scientific establishment — the peer-reviewed journals that supposedly police scientific standards and decide what research gets published — has largely insisted on P values as a measure of publication worthiness. But now the editors of *Basic and Applied Social Psychology* have gone rogue.

"The [P value] fails to provide the probability of the null hypothesis, which is needed to provide a strong case for rejecting it," David Trafimow and Michael Marks of New Mexico State University write in the journal's editorial announcing the P value ban.

It's no great shock that some of the world's statistical organizations have reacted a bit negatively. In a statement, the American Statistical Association expressed concern that the P value–ban "may have its own negative consequences." More than two dozen "distinguished statistical professionals" are developing a statement for the association "to appear later this year" that will "highlight the issues and competing viewpoints." Composing such a statement was a very good idea — 50 years ago.

And in fact, for decades, many distinguished statistical professionals and others have been harping on the intellectual bankruptcy of P values and null hypothesis testing. "Despite the awesome pre-eminence this method has attained … it is based upon a fundamental misunderstanding of the nature of rational inference, and is seldom if ever appropriate to the aims of scientific research," the philosopher of science William Rozeboom wrote — in 1960. Later he called it "surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students."

Many others since Rozeboom have argued just as forcefully that P values are pathological. Their widespread use in scientific research renders many if not most scientific papers guilty of reporting a finding that will later turn out to be wrong. P values pose a serious problem that has plagued the scientific process for nearly a century.

Yet they remain persistently misunderstood. In an account of the *Basic and Applied Social Psychology* ban, a prestigious international scientific journal stated that "the closer to zero the P value gets, the greater the chance that the null hypothesis is false." That's utterly wrong, but it is often how P values get explained and understood. And perhaps that's the best reason to get rid of them.

Follow me on Twitter: @tom_siegfried

Review

Review

You should now be able to:

- Explain the difference between causation and correlation
- Identify a normal distribution within a dataset using summary statistics and visualization
- Test a hypothesis within a sample case study
- Validate your findings using statistical analysis (t-tests, p-values, t-values, confidence intervals)

Q & A

Before Next Class

For the next class, make sure you can commit code into your GitHub repository

- To commit code into GitHub, you need to be authenticated (using SSH keys): Open a terminal and check the output of the following command:

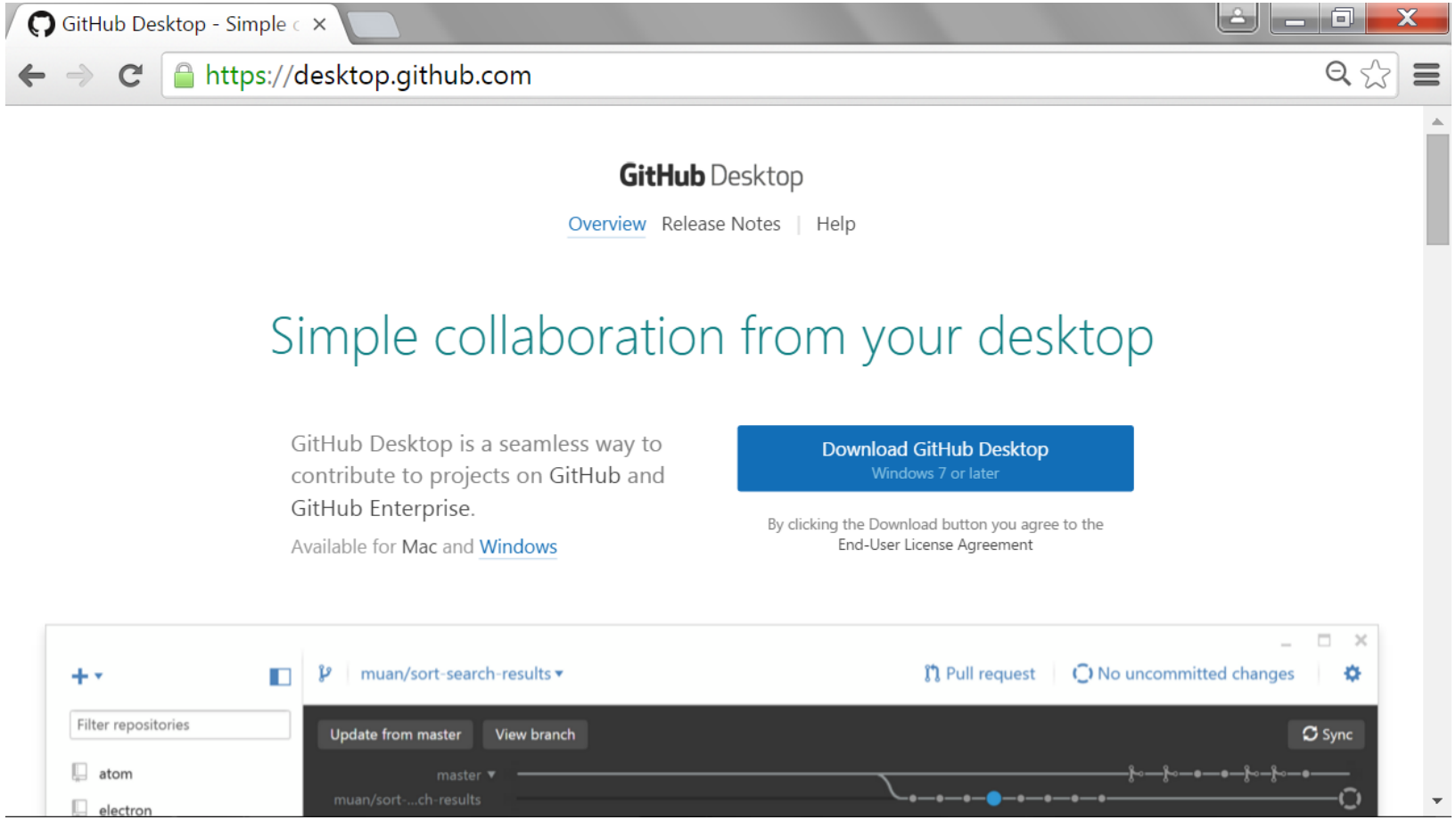
```
ssh -T git@github.com
```

- If you get the following message, you are good to go:

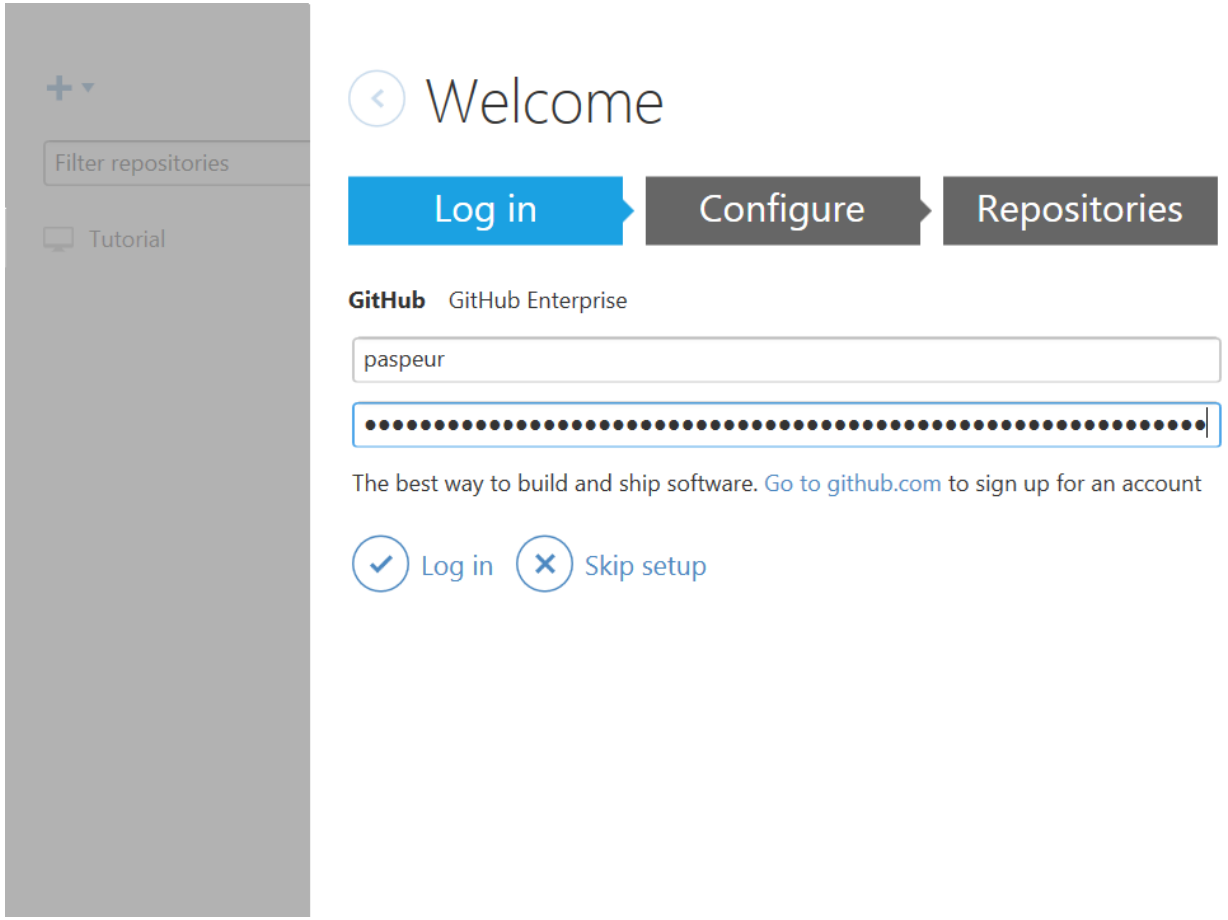
```
Hi □□□□□□□! You've successfully authenticated, but GitHub does not
provide shell access.
```

- If not, you can install the GitHub Desktop client (shown next) which will properly set you up

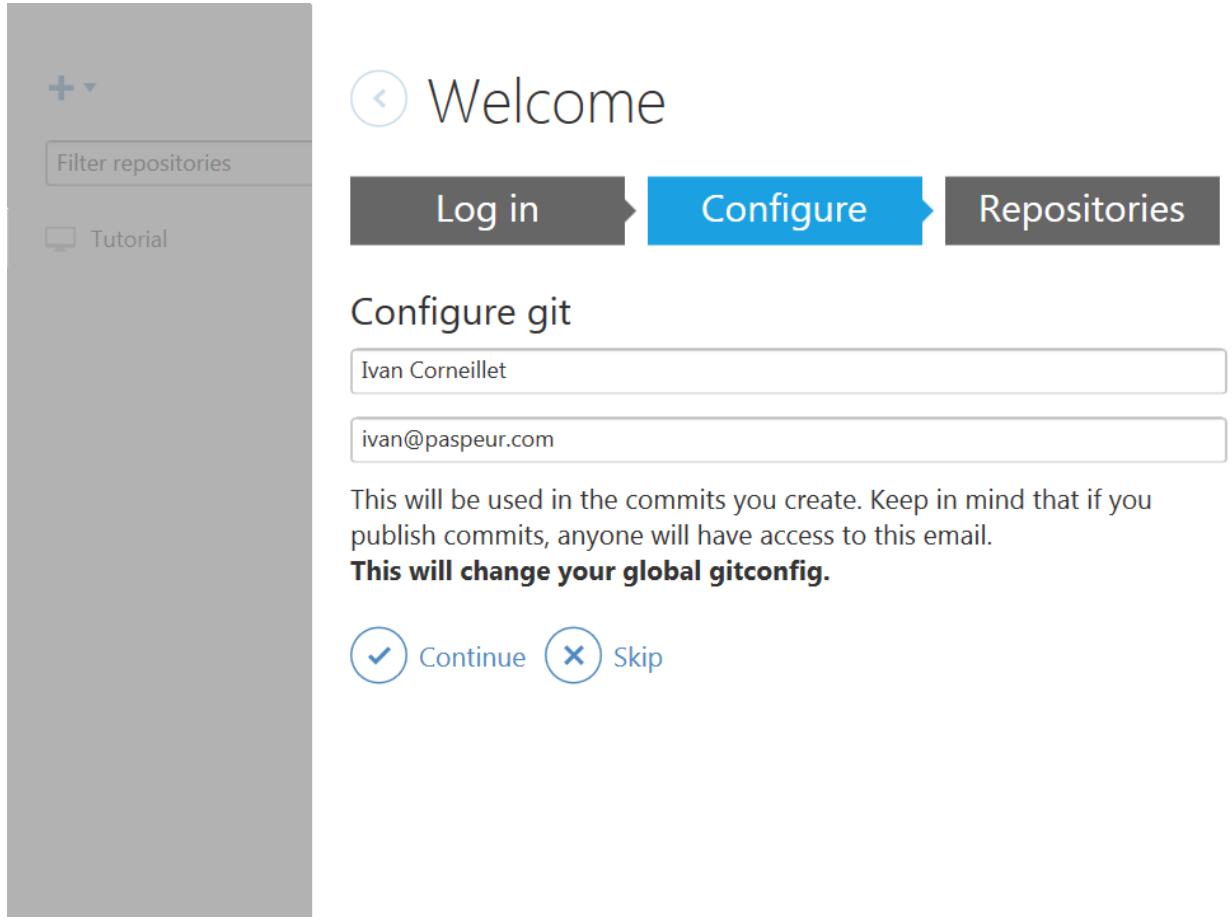
Install the GitHub Desktop client (https://desktop.github.com)



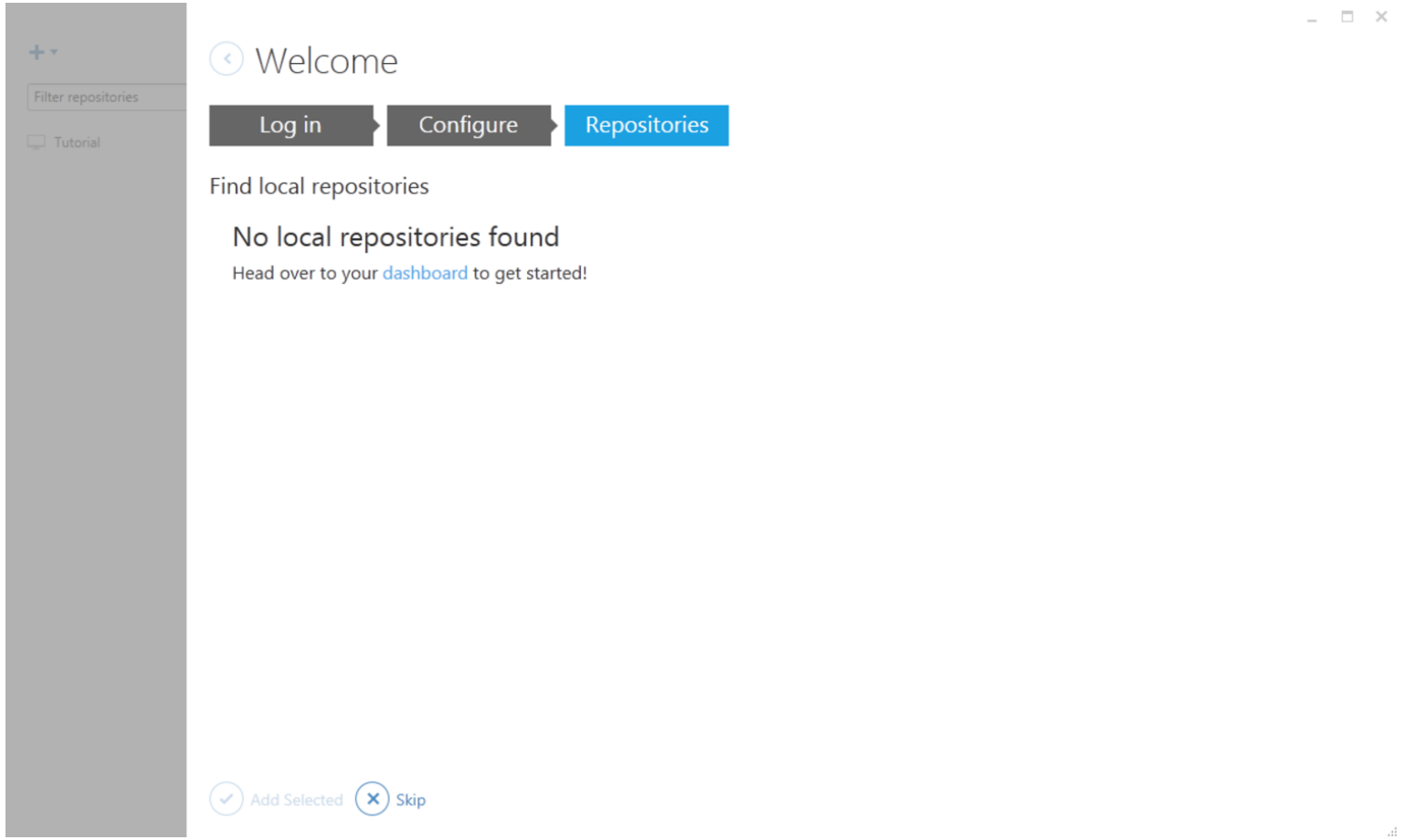
We will use the shell provided by the GitHub Desktop client so we don’t have to deal with authentication (the client will do it on your behalf)



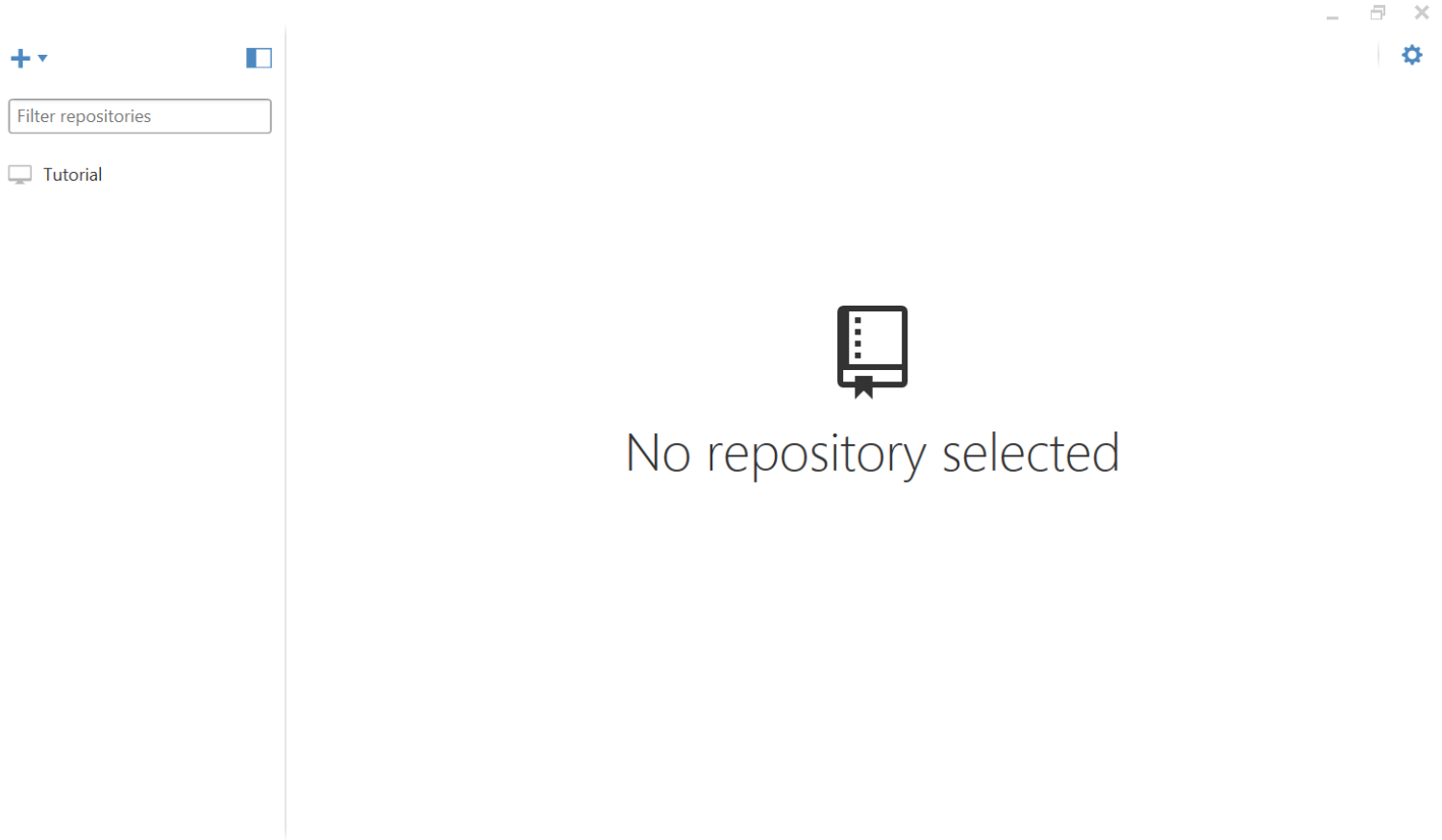
Don’t skip this step…



Feel free to select (or not) the repositories the client finds on your machine

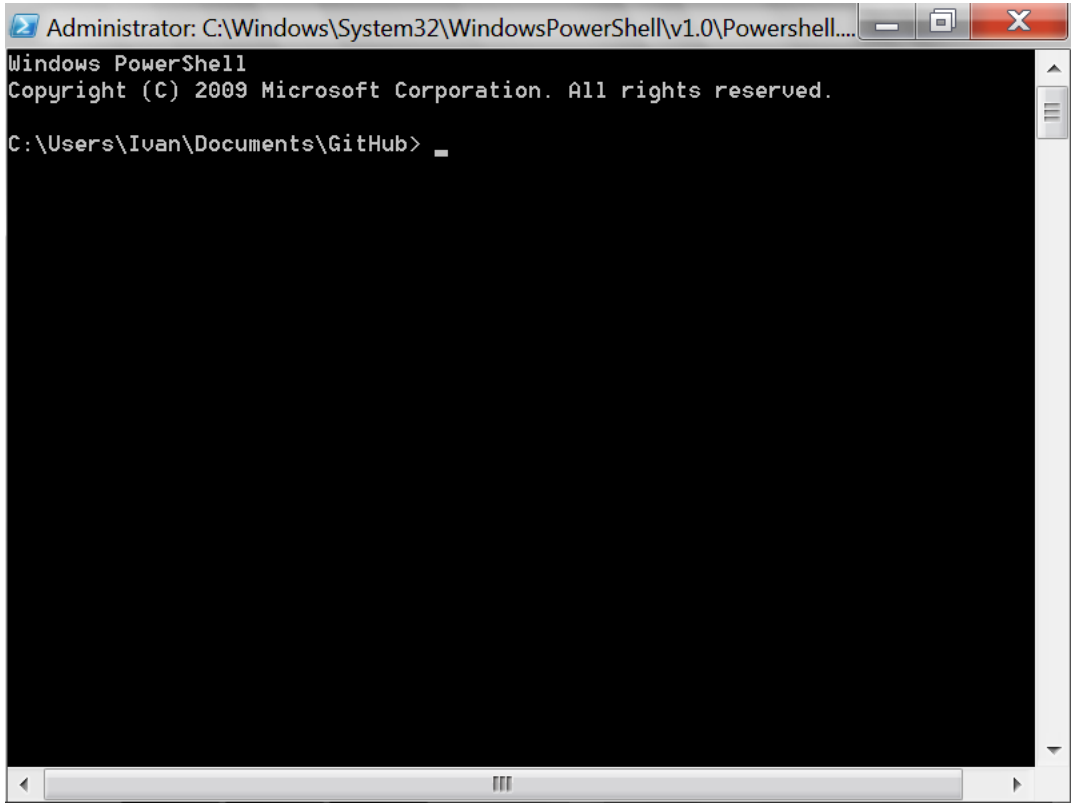


This is what I got after the installation completed (I asked the client not to manage any of my local Git repositories)

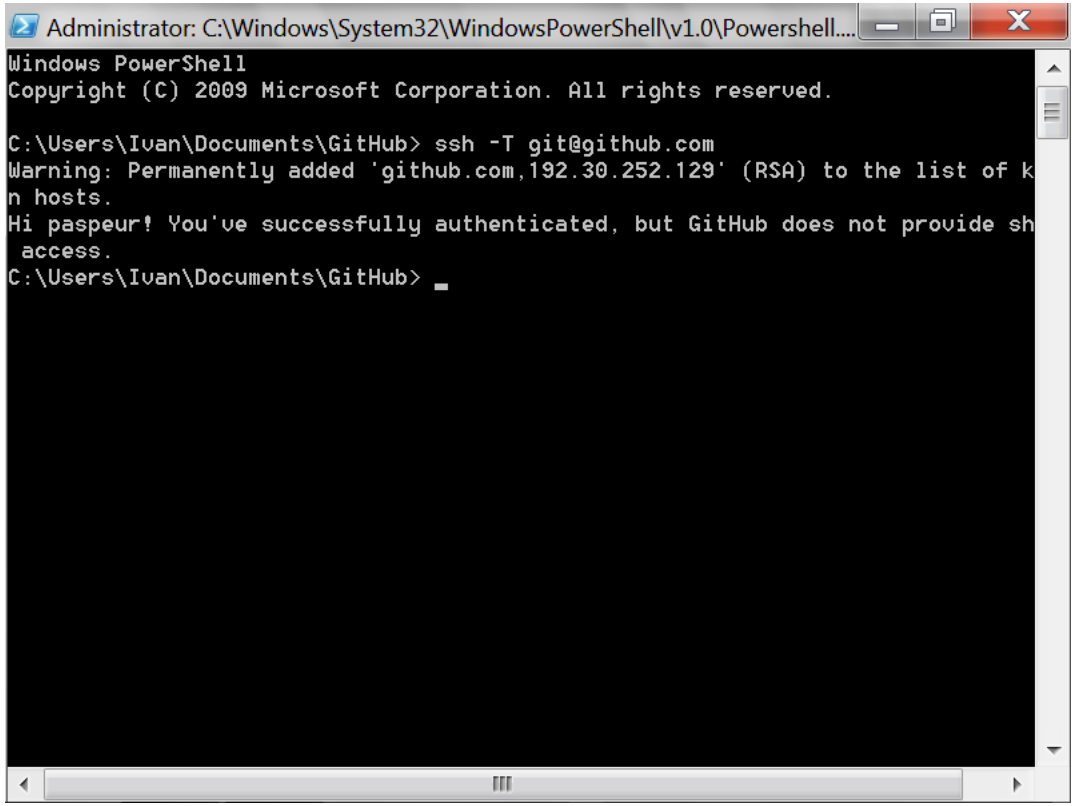


Exit the client then run “Git Shell” from “GitHub, Inc”

You'll then be greeted with a command line prompt



To check that you are correctly authenticated, type `ssh -T git@github.com` and check the response



Exit Ticket

Don’t forget to fill out your exit ticket here

Sources

- Section “If correlation doesn’t imply causation, then what does?” from Michael Nielsen
- Slide #41 – Big Data: A Revolution That Will Transform How We Live, Work, and Think
- 47 – tylervigen.com
- 48 – 51 – xkcd.com