INFO 2950:
Intro to Data Science

Lecture 27
2023-12-04

Admin

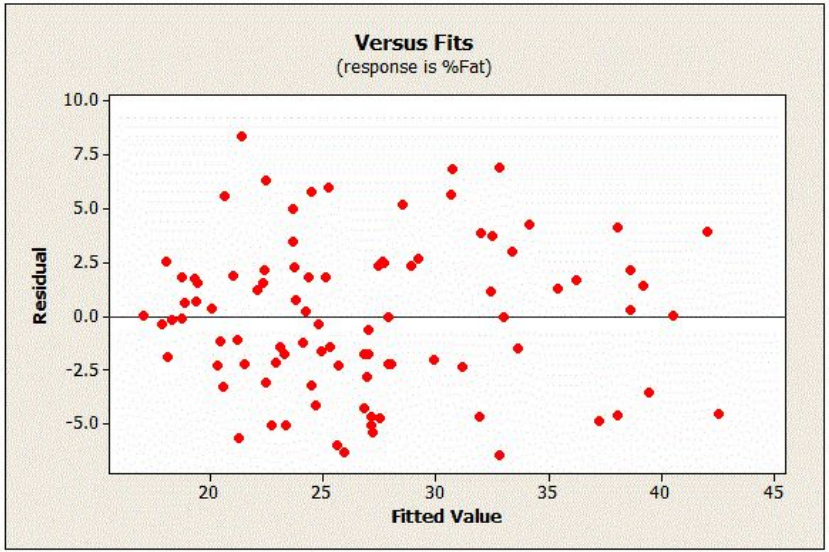
- Final project due at midnight **tonight**
- Weekly OH ending today (but TAs will still be staffing Ed Discussion this week)
- Final exam on Dec 10th, 2pm in Barton Hall (West entrance)
  - Students with SDS accommodations will be in a different room (and should have received an email from me)

INFO 2950 in a nutshell

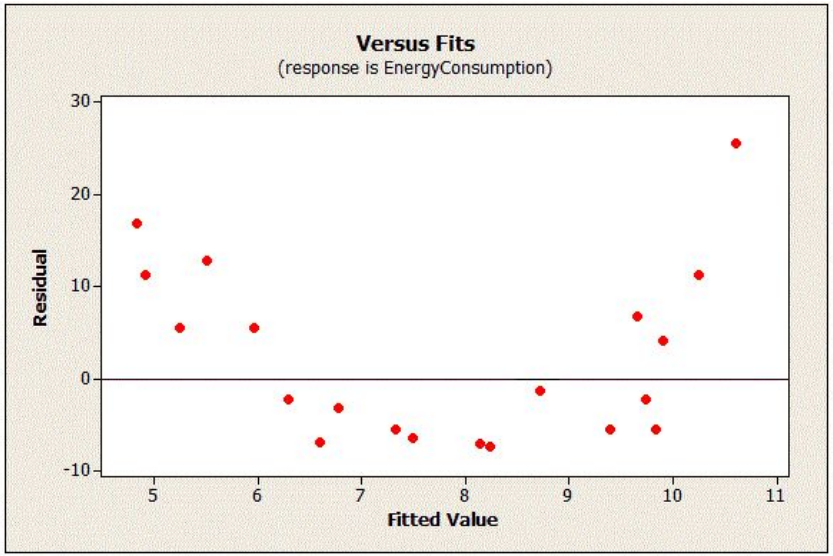
1. Programming with data
2. Describing one variable
3. Describing relationships between two variables
4. Predicting one variable from others
5. Distinguishing pattern from randomness

Which residual plot is concerning?

A



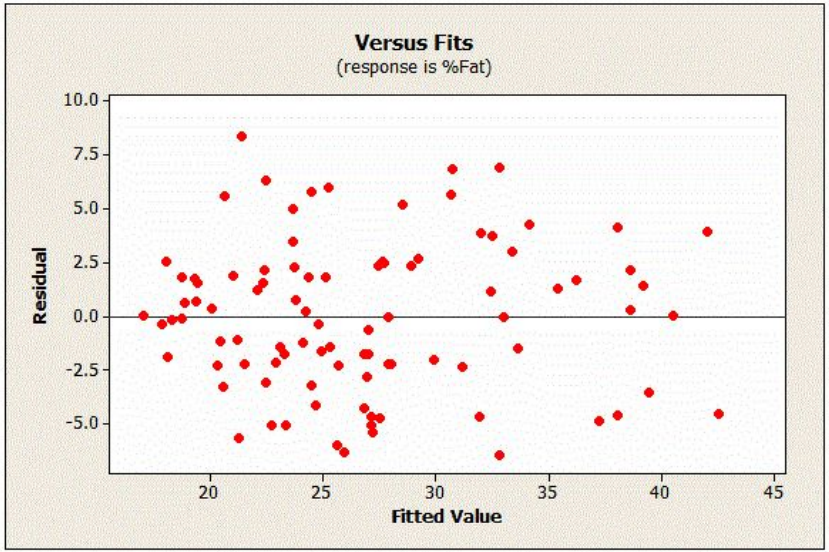
B



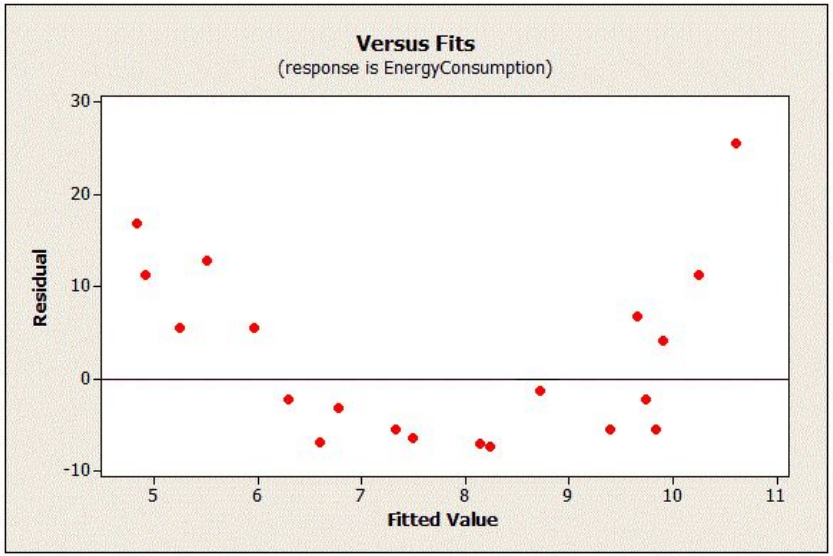
https://statisticsbyjim.com/regression/check-residual-plots-regression-analysis/

Residual Plots: good vs bad

Random Residuals ✅

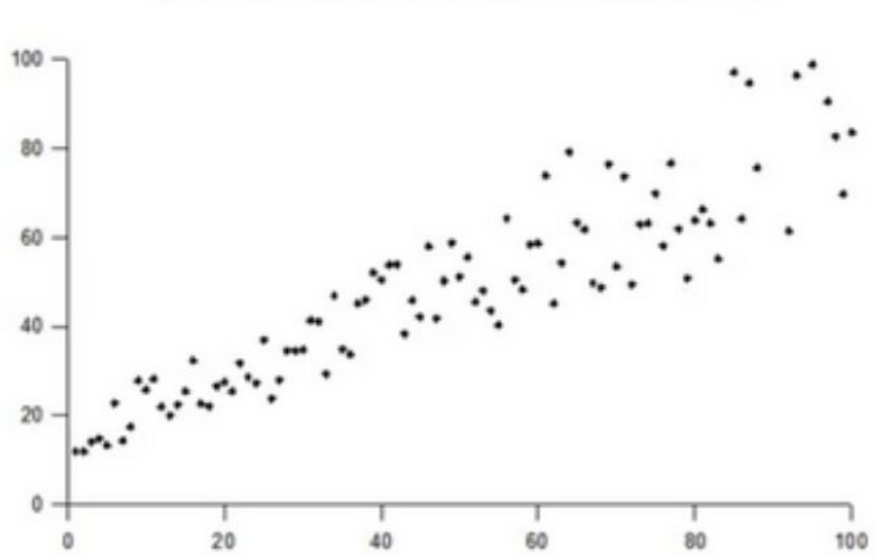


Non-Random Residuals ❌

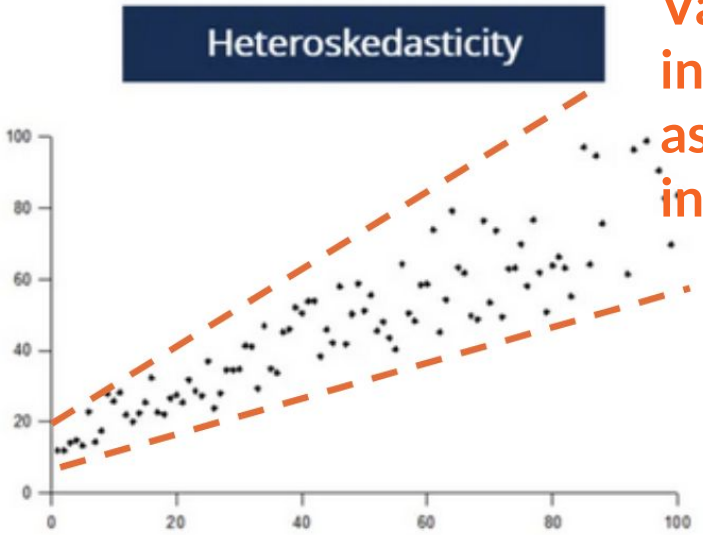


https://statisticsbyjim.com/regression/check-residual-plots-regression-analysis/

What's happening in this residual plot?



What's happening in this residual plot?



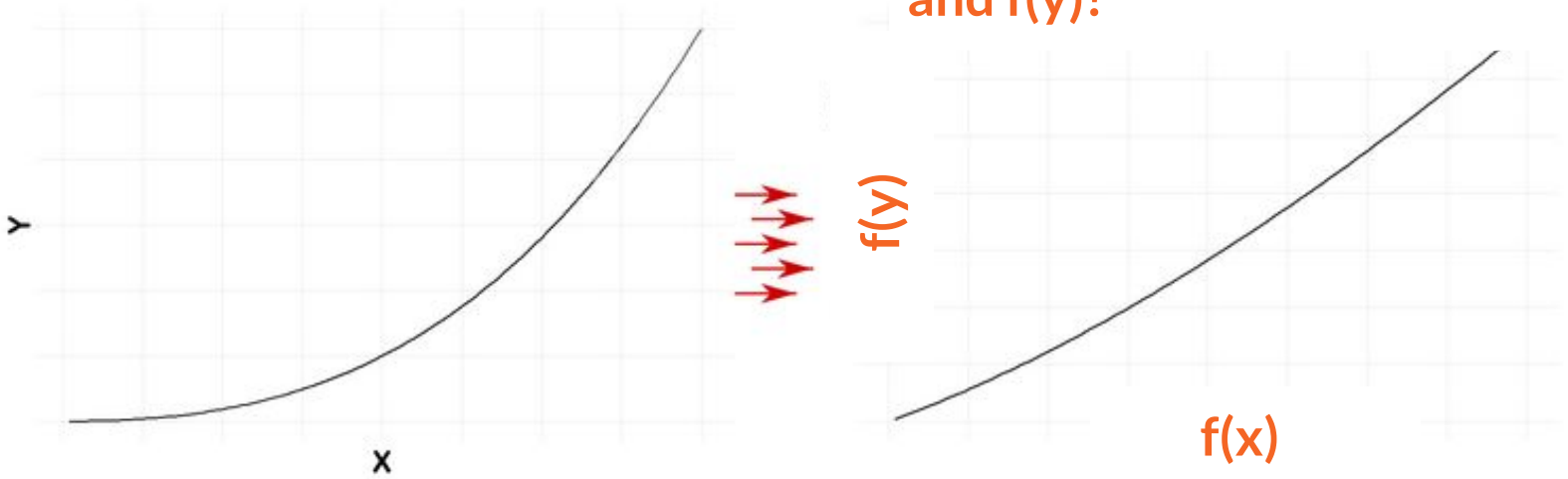
The classic tell: cone/fan shape in residual plot

What to do if residual plot bad

- Try using transformations
- Include missing variables (in multivariable regression)

Use transformations to get better linear fit

What would you use for f(x) and f(y)?

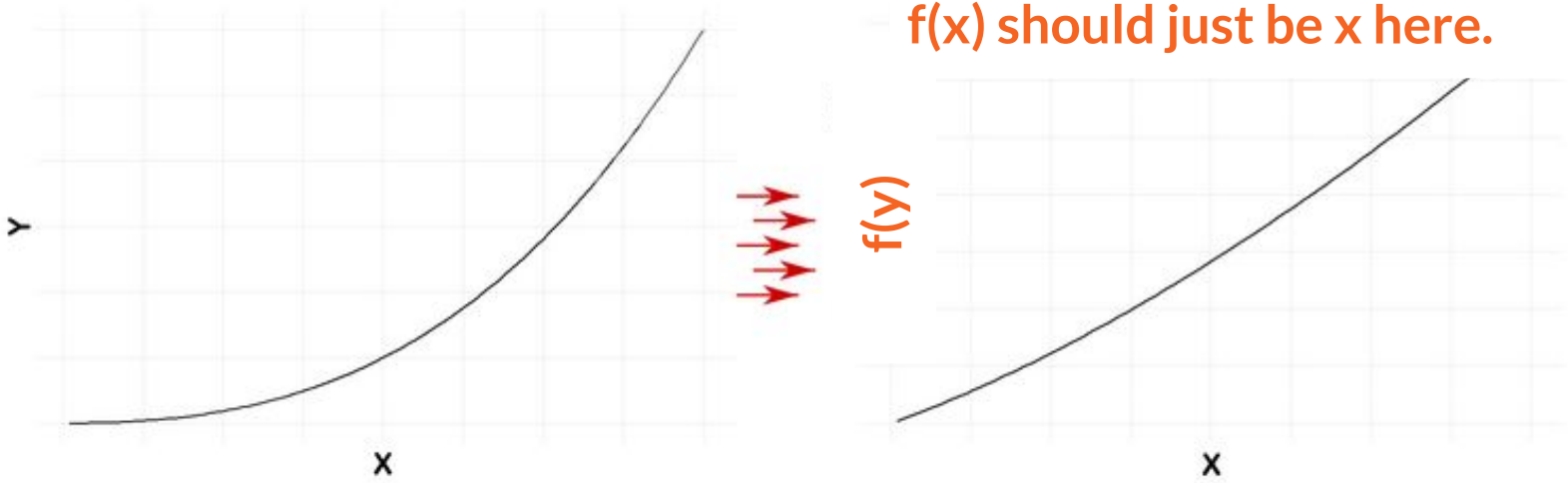


After the transformation, the relationship looks linear enough to run a linear regression

https://quantifyinghealth.com/square-root-transformation/

Use transformations to get better linear fit

f(y) could be sqrt, log, …
f(x) should just be x here.



After the transformation, the relationship looks linear enough to run a linear regression

https://quantifyinghealth.com/square-root-transformation/

Can we run the regression Temp ~ Pressure + Season?

y

Temp (F)	Pressure	Season
80	81	Summer
50	63	Fall
70	75	Spring
…	…	…

Can we run the regression Temp ~ Pressure + Season?

y	x_1			x_2	x_3	x_4
Temp (F)	**Pressure**	**Season**	**Spring**	**Summer**	**Fall**	**Winter**
80	81	Summer	0	1	0	0
50	63	Fall	0	0	1	0
70	75	Spring	1	0	0	0
…	…	…	…	…	…	…

No: we need to convert categorical variable (Season) to dummy variables!

Why shouldn't we include Spring in the regression (and what is it called)?

y	x_1			x_2	x_3	x_4
Temp (F)	**Pressure**	**Season**	**Spring**	**Summer**	**Fall**	**Winter**
80	81	Summer	0	1	0	0
50	63	Fall	0	0	1	0
70	75	Spring	1	0	0	0
…	…	…	…	…	…	…

Why shouldn't we include Spring in the regression (and what is it called)?

y	x_1			x_2	x_3	x_4
Temp (F)	**Pressure**	**Season**	**Spring**	**Summer**	**Fall**	**Winter**
80	81	Summer	0	1	0	0
50	63	Fall	0	0	1	0
70	75	Spring	1	0	0	0
…	…	…	…	…	…	…

Adding in the *reference level* would lead to multicollinearity

What should we do if if we have 10,000 potential input variables?

y	x_1			x_2	x_3	
Temp (F)	**Pressure**	**Season**	**Spring**	**Summer**	**Fall**	**…**
80	81	Summer	0	1	0	…
50	63	Fall	0	0	1	…
70	75	Spring	1	0	0	…
…	…	…	…	…	…	…

Feature selection (domain expertise, check for collinearity, SVD for dimension reduction, etc.)

y	x_1			x_2	x_3	
Temp (F)	**Pressure**	**Season**	**Spring**	**Summer**	**Fall**	**…**
80	81	Summer	0	1	0	…
50	63	Fall	0	0	1	…
70	75	Spring	1	0	0	…
…	…	…	…	…	…	…

Interactions: when to include them?

- When you think two variables' combinations have an additional effect on the output
    - E.g., intersectionality

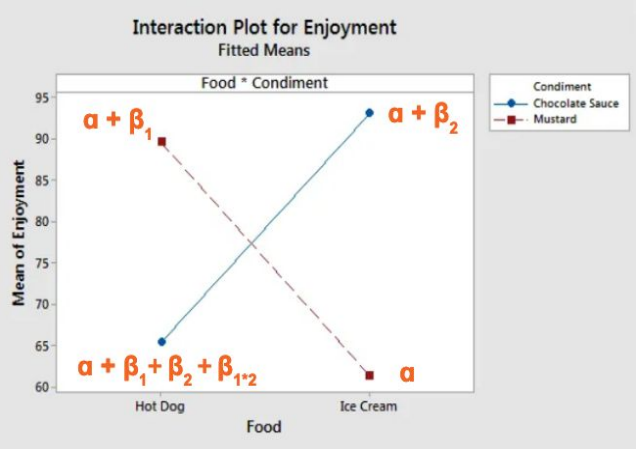
Interactions: when to include them?

- When you think two variables' combinations have an additional effect on the output
    - E.g., intersectionality

- **How to check for whether you have an interaction effect?**

Do you need regression interactions?

- Use interaction plots
  (harder to interpret with logit)



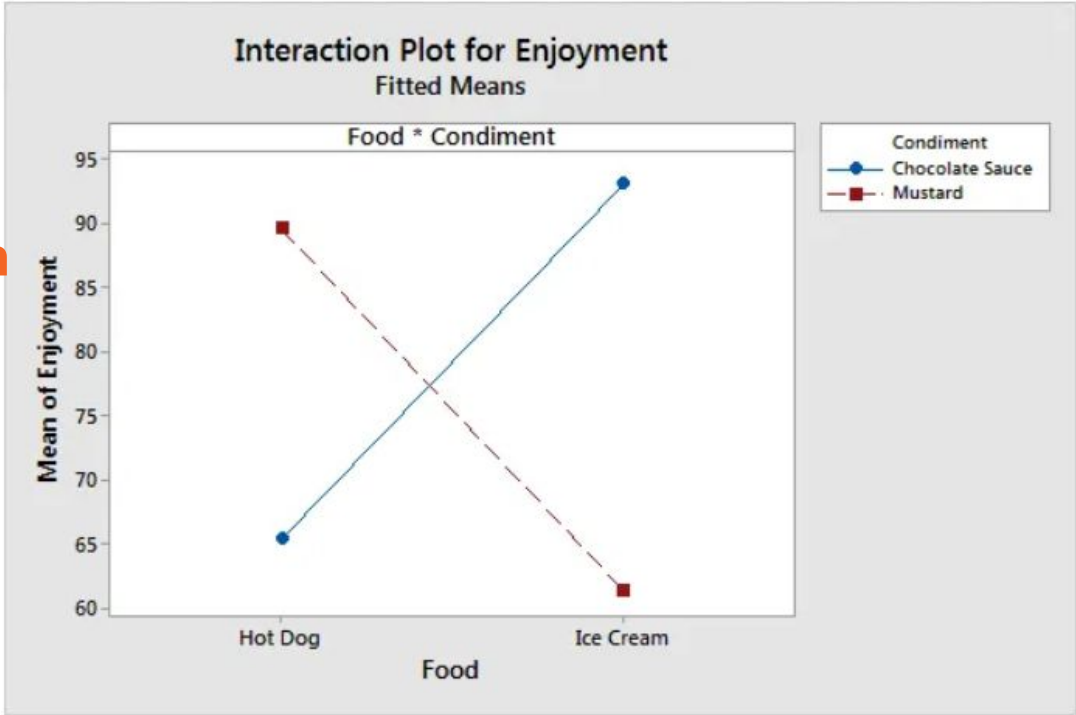
- Use probabilistic thinking

P(Democrat)	Men	Women
No college degree	$\sigma(\alpha + 0\beta_F + 0\beta_C) = 0.40$	$\sigma(\alpha + 1\beta_F + 0\beta_C) = 0.51$
College degree	$\sigma(\alpha + 0\beta_F + 1\beta_C) = 0.48$	$\boldsymbol{\sigma(\alpha + 1\beta_F + 1\beta_C)} = 0.59 \neq 0.65$

Visual aid: interaction plot

Interaction

$$\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_{1*2} x_1 * x_2$$



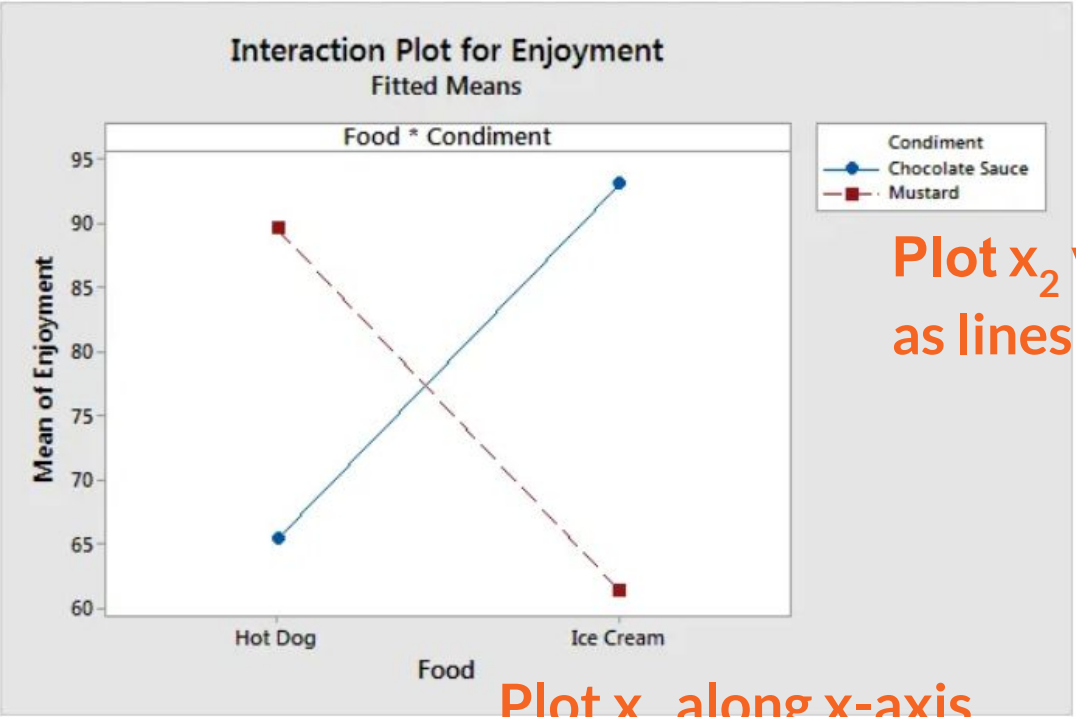
https://statisticsbyjim.com/regression/interaction-effects/

Visual aid: interaction plot

Plot predicted outcome

$\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_{1*2} x_1 * x_2$

along y-axis

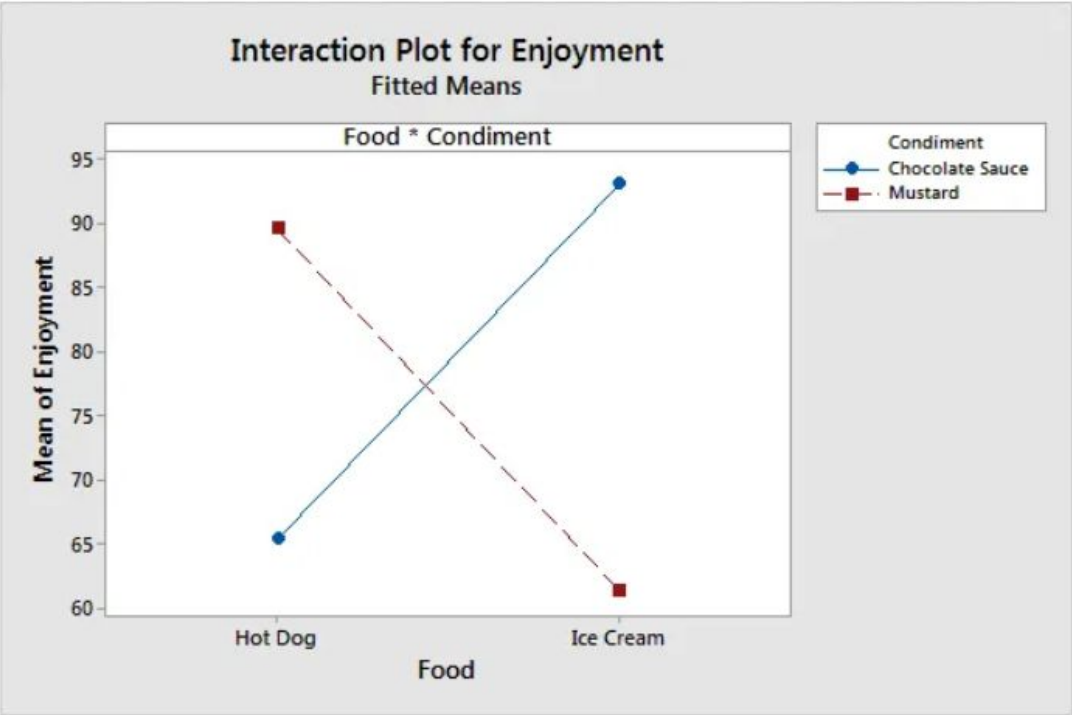


Plot x_2 values as lines

Plot x_1 along x-axis

https://statisticsbyjim.com/regression/interaction-effects/

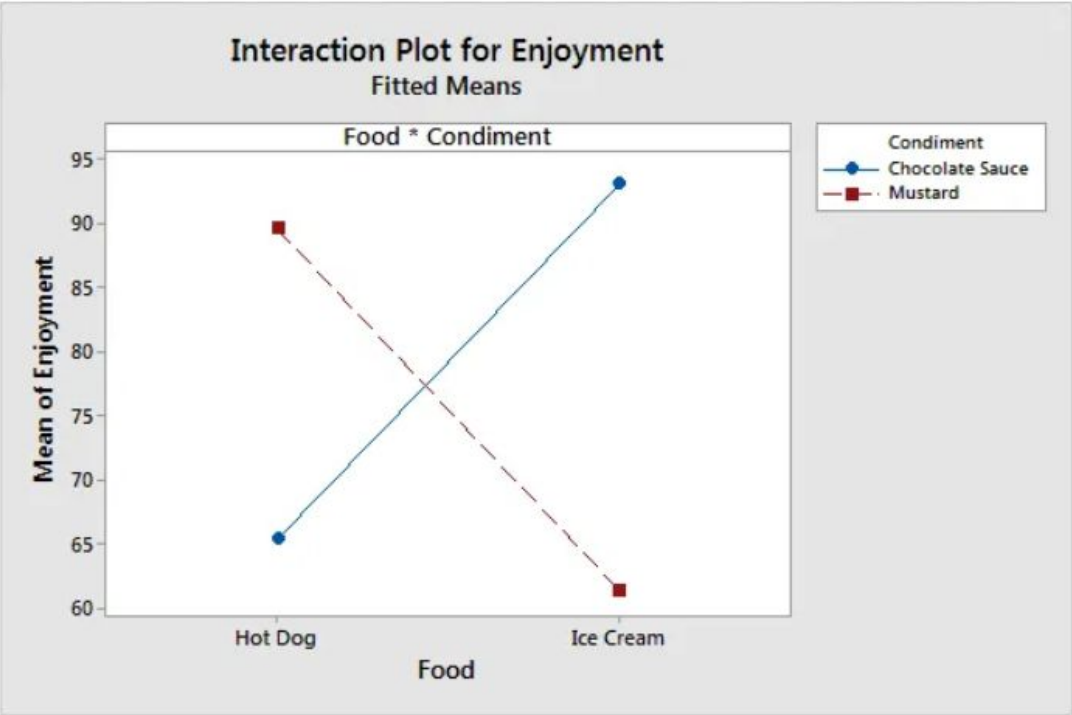
Should you include an interaction term here?



https://statisticsbyjim.com/regression/interaction-effects/

Should you include an interaction term here?

Yes, because these lines cross!



https://statisticsbyjim.com/regression/interaction-effects/

Which of these should you do to interpret regressions with interaction effects?

- Predict
- Summarize
- Oddities/outliers

Which of these should you do to interpret regressions with interaction effects?

- **Predict**
- ~~Summarize~~ Easy to be misleading if you're not careful: a negative coefficient on interactions doesn't necessarily mean the overall effect of having both variables is negative
- **Oddities/outliers**

Interpreting Regressions

- If **no interactions**, interpret each different x_i separately
  - Summarize Relationship | Predict Outcome | Outliers & Oddities
- If **have interactions**, **interpret by plugging in values for different combinations of x_i**
  - Predict Outcome | Outliers & Oddities

How to predict with interaction effects?

- $y = 4.5 - 0.4x_1 - 0.5x_2 - 0.1x_1{}^{*}x_2$

Demographic	x_1	x_2	Example	Value	Equation
White male	0	0	$y^{hat} = 4.5 - 0.4{*}0 - 0.5{*}0 - 0.1{*}0{*}0$	4.5	$y^{hat} = \alpha$
Non-white male	1	0	$y^{hat} = 4.5 - 0.4{*}1 - 0.5{*}0 - 0.1{*}1{*}0$	4.1	$y^{hat} = \alpha + \beta_1$
White non-male	0	1	$y^{hat} = 4.5 - 0.4{*}0 - 0.5{*}1 - 0.1{*}0{*}1$	4.0	$y^{hat} = \alpha + \beta_2$
Non-white non-male	1	1	$y^{hat} = 4.5 - 0.4{*}1 - 0.5{*}1 - 0.1{*}1{*}1$	__	$y^{hat} =$ _____

Multivar Regression: Interactions

- Interpretation: our model predicts that non-white non-male instructors be rated lowest by student evaluations at 3.5, which is a full point lower than white male instructors, who we predict to have the highest student evaluations at 4.5.

Demographic	x_1	x_2	Example	Value	Equation
White male	0	0	$y^{hat} = 4.5 - 0.4*0 - 0.5*0 - 0.1*0*0$	4.5	$y^{hat} = \alpha$
Non-white male	1	0	$y^{hat} = 4.5 - 0.4*1 - 0.5*0 - 0.1*1*0$	4.1	$y^{hat} = \alpha + \beta_1$
White non-male	0	1	$y^{hat} = 4.5 - 0.4*0 - 0.5*1 - 0.1*0*1$	4.0	$y^{hat} = \alpha + \beta_2$
Non-white non-male	1	1	$y^{hat} = 4.5 - 0.4*1 - 0.5*1 - 0.1*1*1$	3.5	$y^{hat} = \alpha + \beta_1 + \beta_2 + \beta_{1*2}$

Interpreting regressions: prediction (first line), summary (next lines)

Canvas > Modules > 2. Regression and Linear Models

Review how to apply in examples & derive the things in this table!

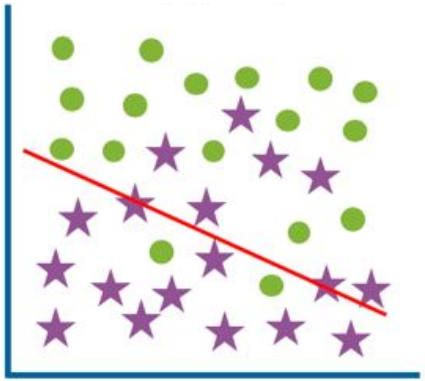
Midterm Fall 2023 - Review Topics.pdf

Download Midterm Fall 2023 - Review Topics.pdf (76.7 KB) | Alternative formats

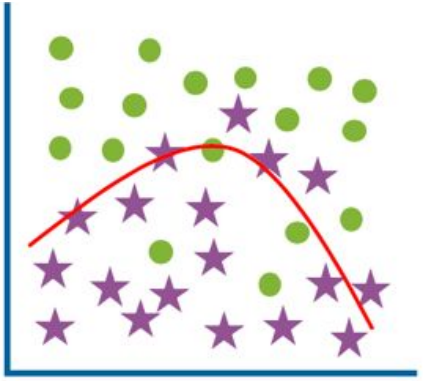
Model	Regression Interpretation
Linear $y = \alpha + \beta x$	If x=0, y = α 1 unit change in x is associated with a β unit change in y
Linear-log $y = \alpha + \beta \ln(x)$	If x=1, y = α If x is multiplied by *e*, we expect a β unit change in y 1% change in x is associated with a 0.01*β unit change in y
Log-linear $\ln(y) = \alpha + \beta x$	If x=0, $y = e^{\alpha}$ For a 1 unit change in x, we expect y to be multiplied by e^{β} 1 unit change in x is associated with a 100*(exp(β)-1)% change in y
Log-log	If x=1, $y = e^{\alpha}$

Which is the ideal model?

A



B

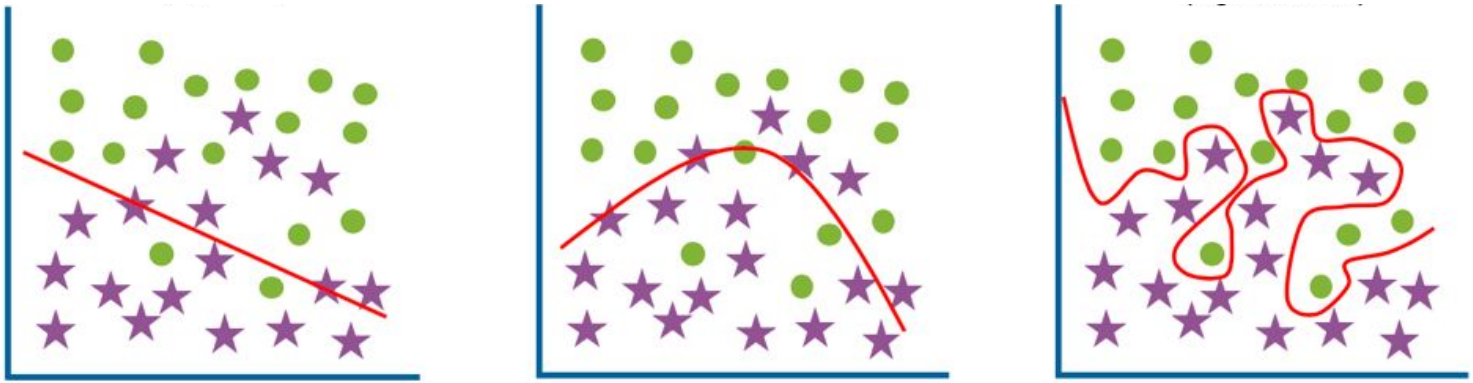


C

https://www.ibm.com/cloud/learn/overfitting

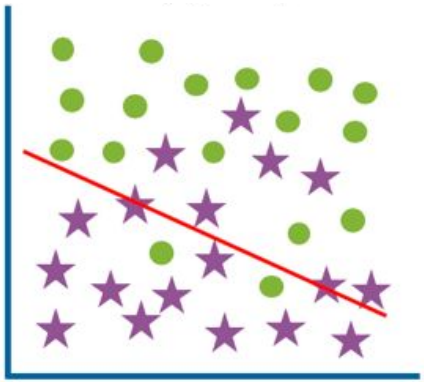
Which is the ideal model?

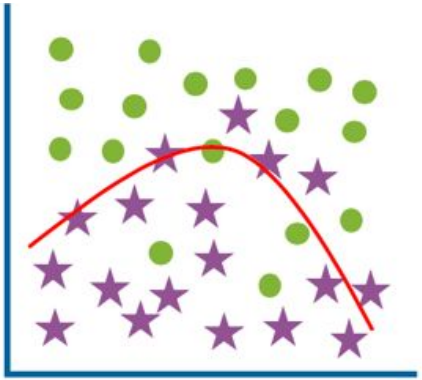
Ideal



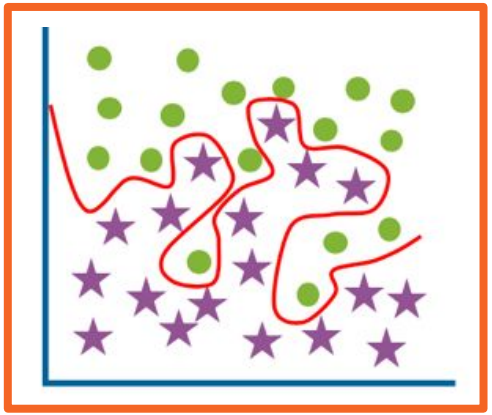
https://www.ibm.com/cloud/learn/overfitting

Which is the ideal model?



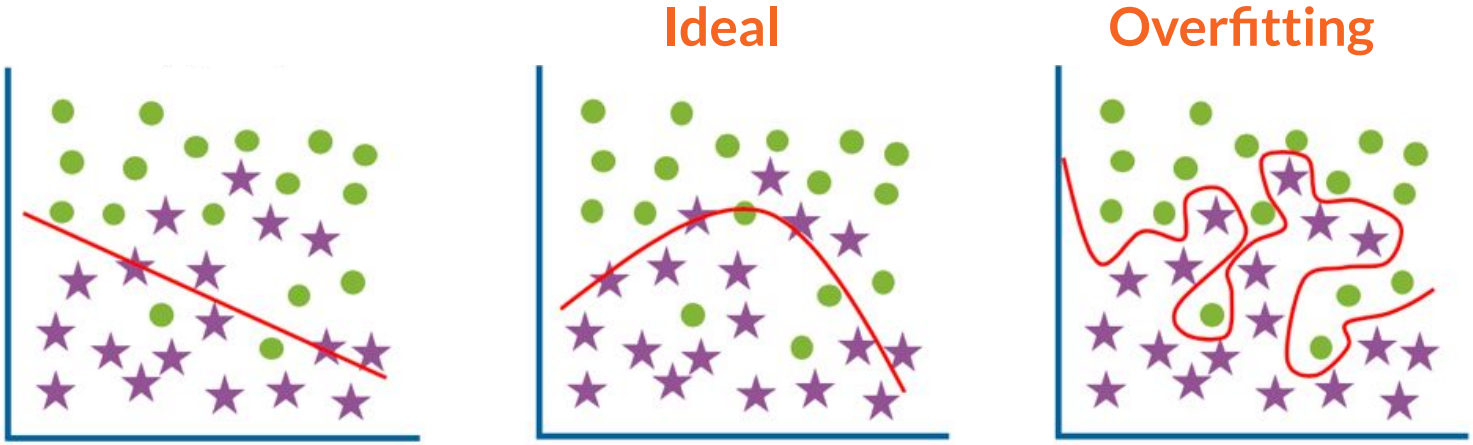


What's this called?



https://www.ibm.com/cloud/learn/overfitting

Which is the ideal model?



https://www.ibm.com/cloud/learn/overfitting

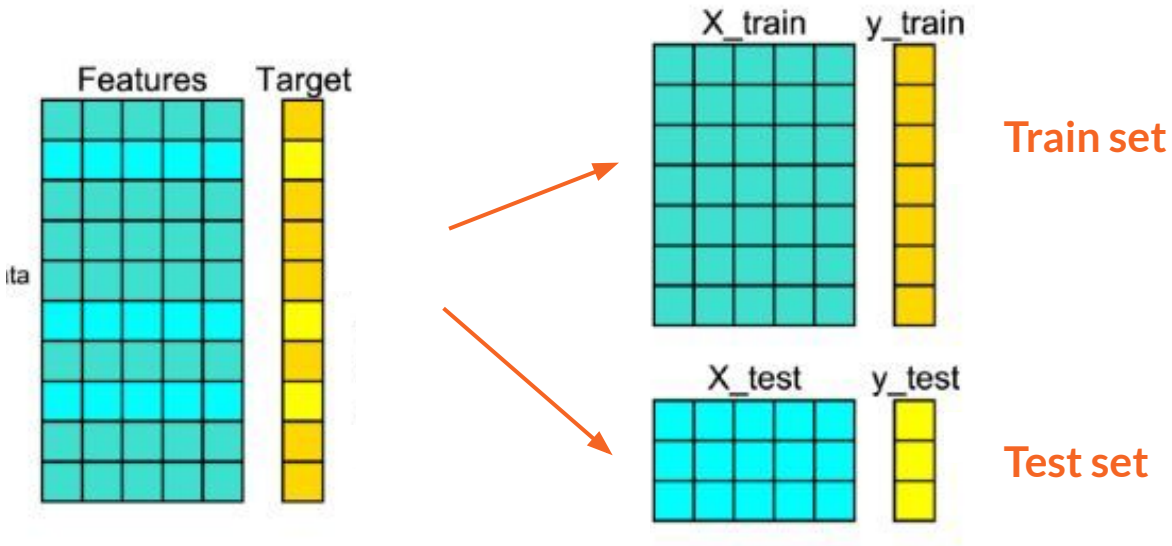
https://datasciencedojo.com/blog/data-science-memes/

How to overcome overfitting?

- Feature selection (choosing fewer inputs)
- Using train/validation/test split

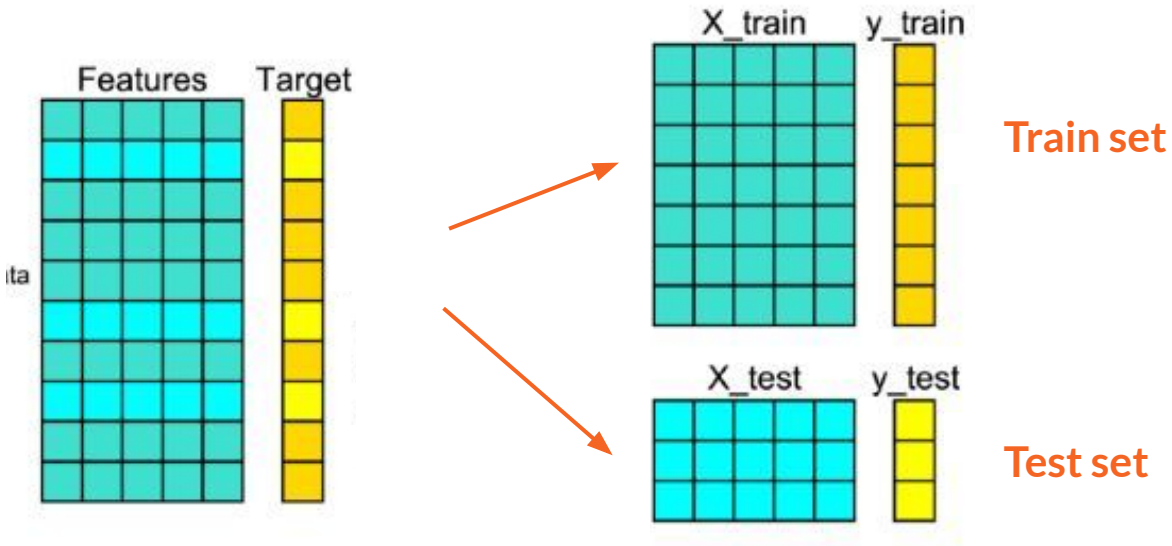
https://www.ibm.com/cloud/learn/overfitting

What's the train/test split %?



https://builtin.com/data-science/train-test-split

Train/test: 70%/30%



https://builtin.com/data-science/train-test-split

Word bank: X_train, y_train, X_test, y_test

```
model = LinearRegression().fit(_____,_____)

y_hat_train = model.predict(_____)

y_hat_test = model.predict(_____)
```

Word bank: X_train, y_train, X_test, y_test

```
model = LinearRegression().fit(X_train,y_train)

y_hat_train = model.predict(X_train)

y_hat_test = model.predict(X_test)
```

Word bank: X_train, y_train, X_test, y_test

```
model = LinearRegression().fit(X_train,y_train)

y_hat_train = model.predict(X_train)

y_hat_test = model.predict(X_test)
```

To get evaluation metrics:

- What do you compare y_hat_train to?
- What do you compare y_hat_test to?

Word bank: X_train, y_train, X_test, y_test

```
model = LinearRegression().fit(X_train,y_train)

y_hat_train = model.predict(X_train)

y_hat_test = model.predict(X_test)
```

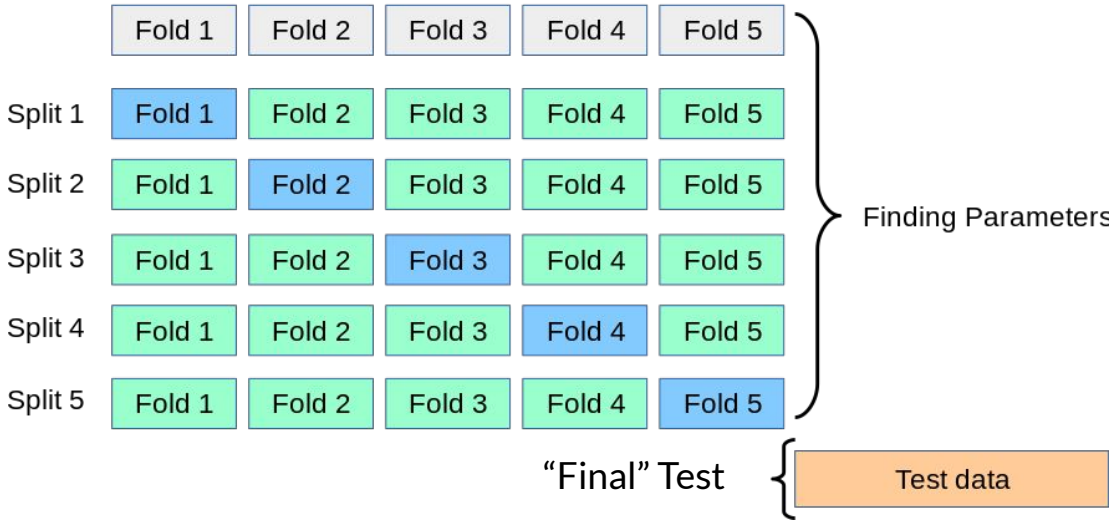
To get evaluation metrics:

- What do you compare y_hat_train to? y_train
- What do you compare y_hat_test to? y_test

What if you have concerns about your test set (too big, too small, or too similar/dissimilar to train set)?

What if you have concerns about your test set (too big, too small, or too similar/dissimilar to train set)?

Cross Validation



https://scikit-learn.org/stable/modules/cross_validation.html

Which of these evaluation metrics would you use for numeric (non-binary) outputs y?

A

- MSE
- RMSE
- MAE
- MAPE

B

- Precision
- Recall
- F1 score
- AUC-ROC

Which of these evaluation metrics would you use for numeric (non-binary) outputs y?

Prediction (numeric)

- MSE
- RMSE
- MAE
- MAPE

Classification (binary)

- Precision
- Recall
- F1 score
- AUC-ROC

Why isn't accuracy on the binary evaluation metric list?

Classification (binary)

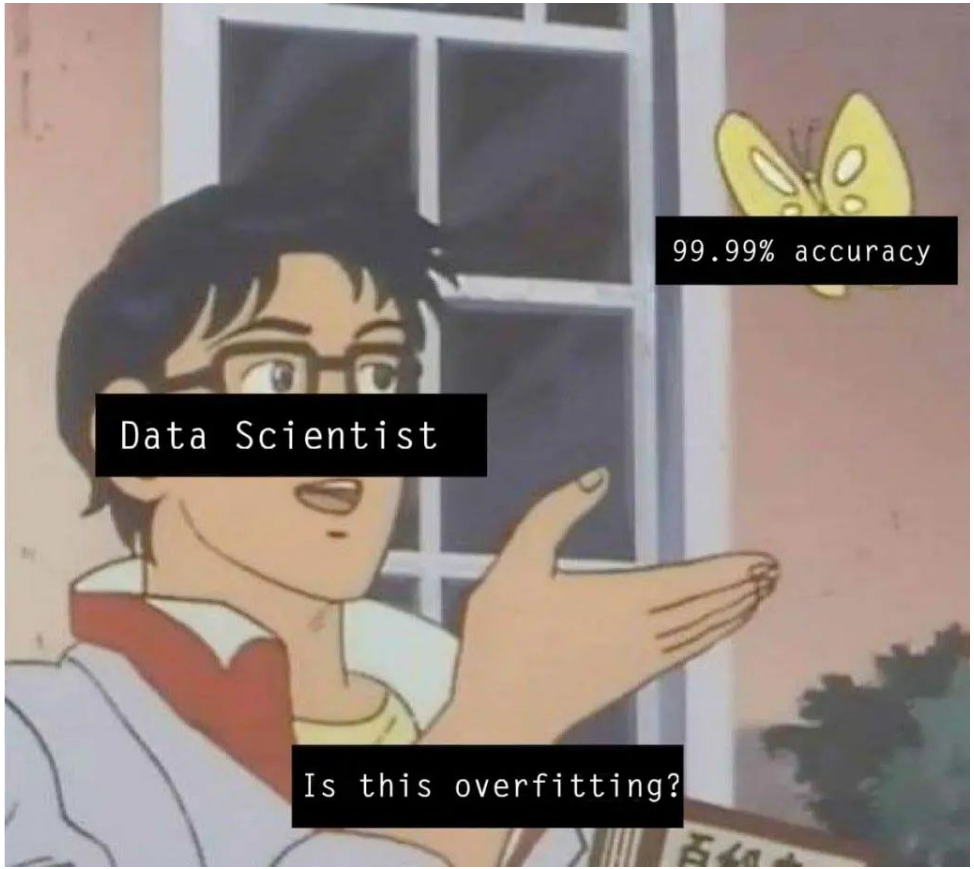
- Precision
- Recall
- F1 score
- AUC-ROC

Why isn't accuracy on the binary evaluation metric list?

Accuracy isn't good for datasets with imbalanced classes (e.g. if only 1% of the population is 1 and 99% is 0, always returning 0 gives you 99% accuracy but that's not a good model)

Classification (binary)

- Precision
- Recall
- F1 score
- AUC-ROC



https://datasciencedojo.com/blog/data-science-memes/

Evaluation metrics suggestions

- Understand the differences between TN, TP, FN, FP and when to use them
- Know whether metrics being “good” vs. “bad” correspond to increases or decreases in the metric

Which is indicative of overfitting (not generalizing well to out-of-sample data)?

A

Train set metrics “good”, test set metrics “bad”

B

Test set metrics “good”, train set metrics “bad”

Which is indicative of overfitting (not generalizing well to out-of-sample data)?

A

Train set metrics “good”, test set metrics “bad”

B

Test set metrics “good”, train set metrics “bad”

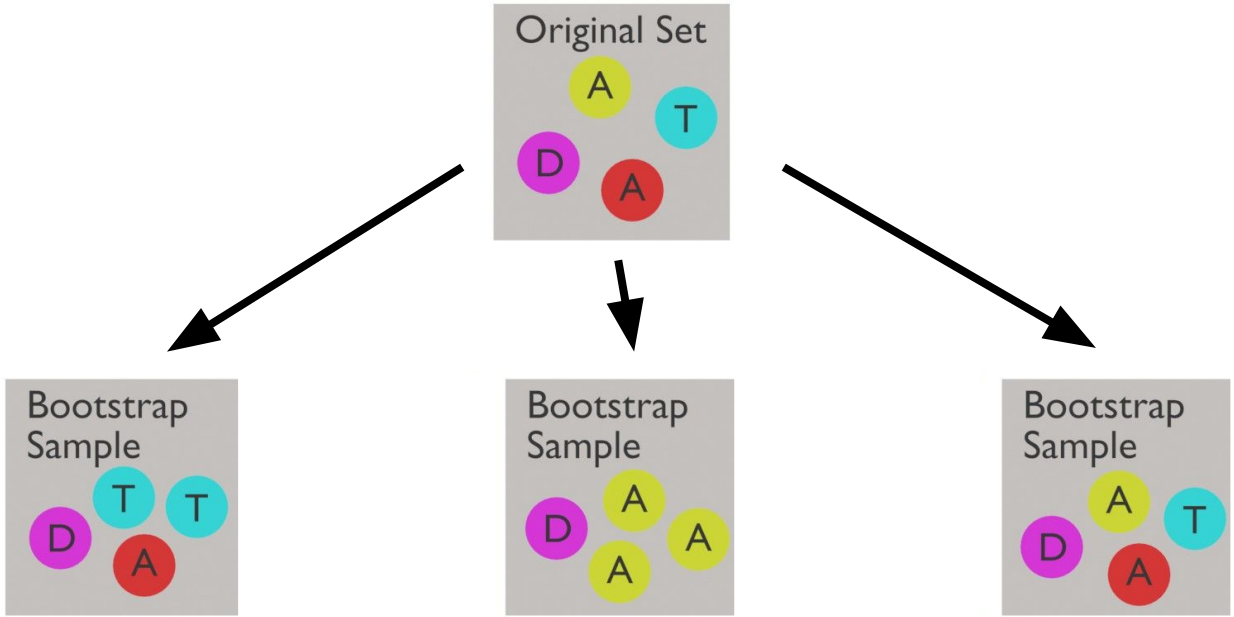
Resampling methods

Without Replacement	With Replacement
Cross Validation	Bootstrap

Calculate statistic (evaluation metric; mean and variance of multiple evaluation metrics) based on the left-out fold (using a model trained on the left-in samples)

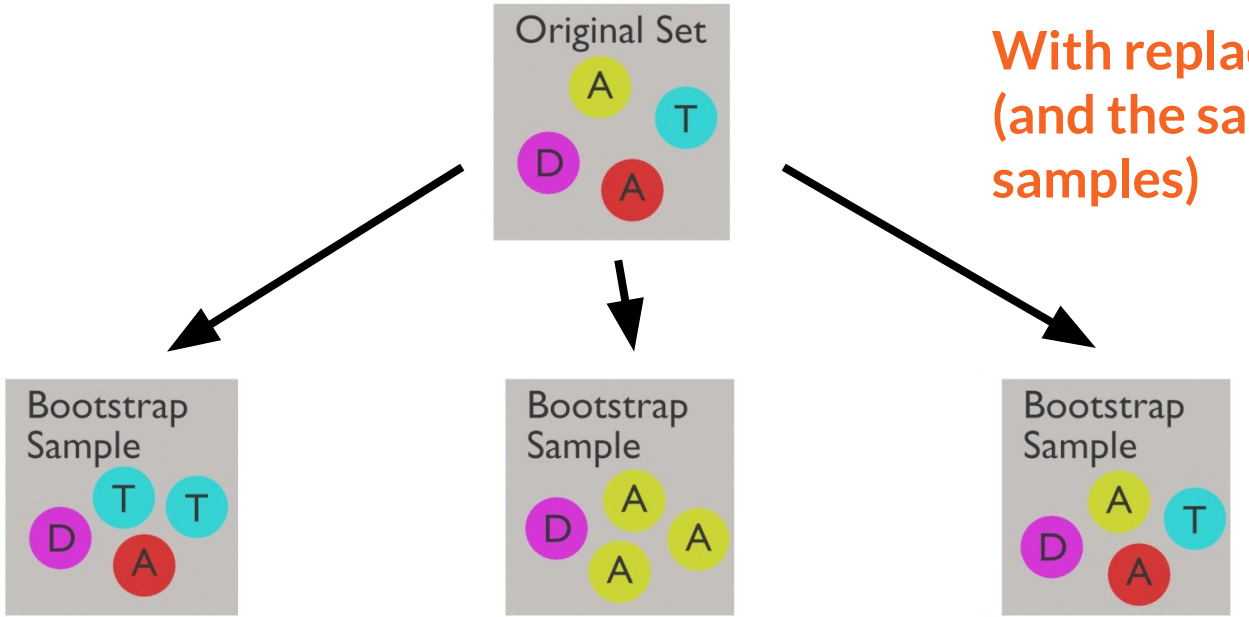
Calculate statistic (predicted values; mean and variance of multiple predicted values) based on the left-in samples

Bootstrapping: is this drawing with or without replacement?



https://commons.wikimedia.org/wiki/File:Sampling_with_replacement_and_out-of-bag_dataset.jpg

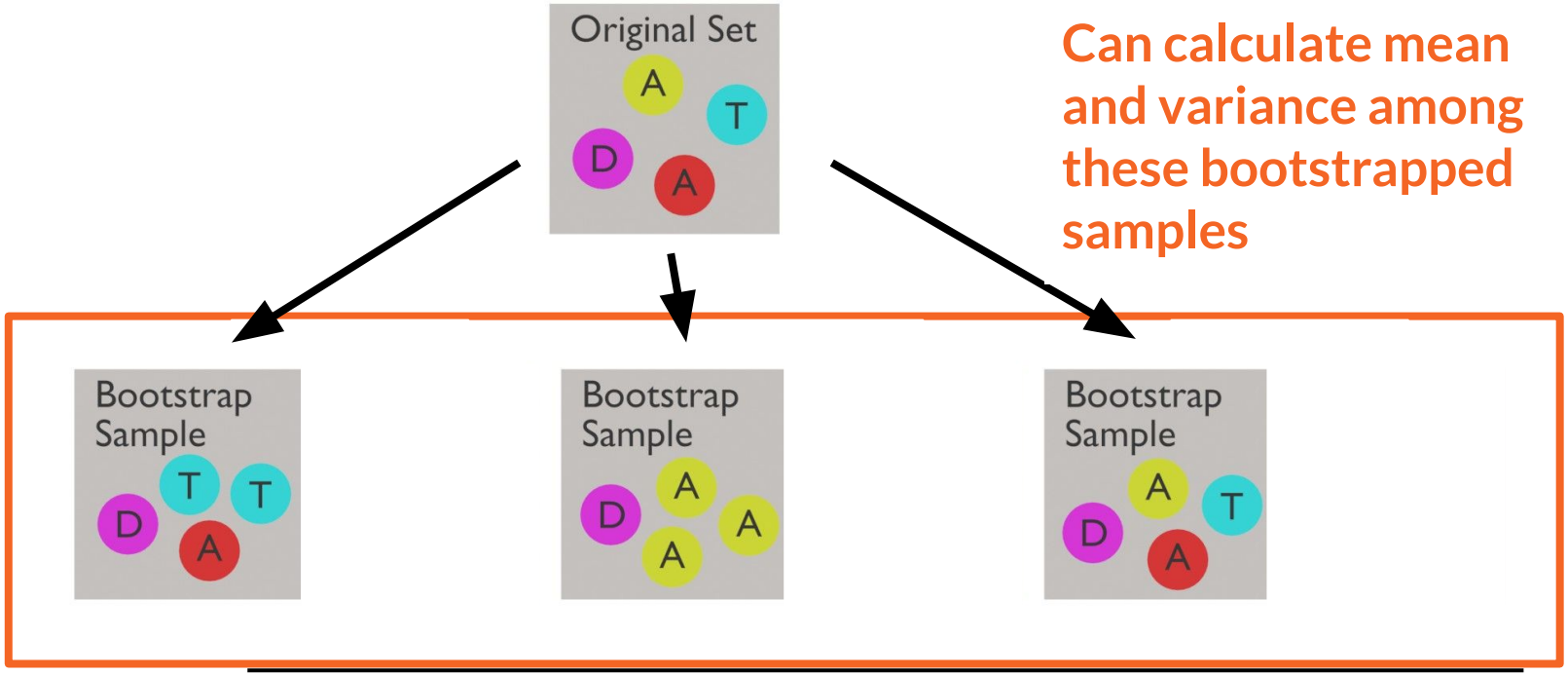
Bootstrapping: is this drawing with or without replacement?



With replacement (and the same # samples)

https://commons.wikimedia.org/wiki/File:Sampling_with_replacement_and_out-of-bag_dataset.jpg

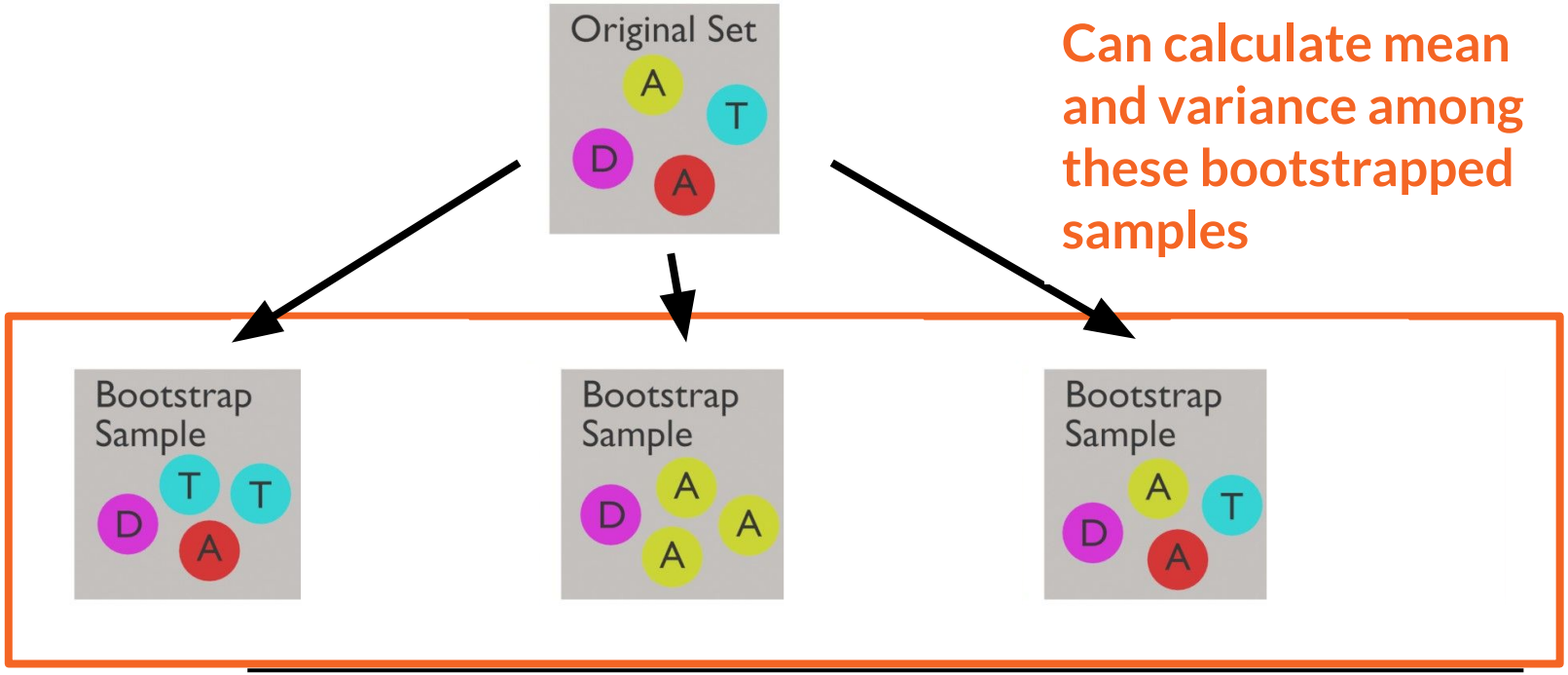
Bootstrapping: is this drawing with or without replacement?



Can calculate mean and variance among these bootstrapped samples

https://commons.wikimedia.org/wiki/File:Sampling_with_replacement_and_out-of-bag_dataset.jpg

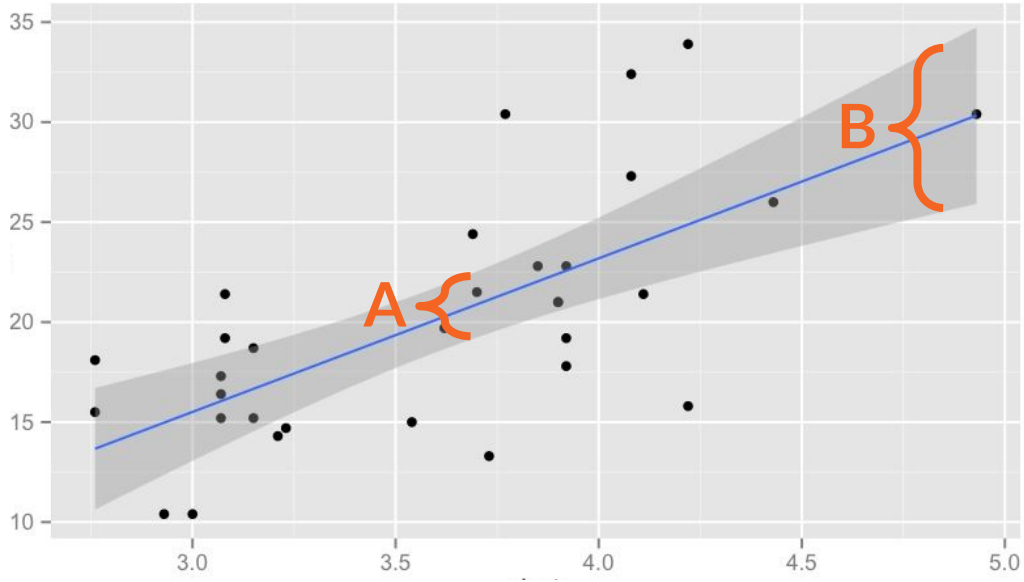
We can run a regression on each bootstrapped population sample



Can calculate mean and variance among these bootstrapped samples

https://commons.wikimedia.org/wiki/File:Sampling_with_replacement_and_out-of-bag_dataset.jpg

Are we more confident about our predictions at A or B?

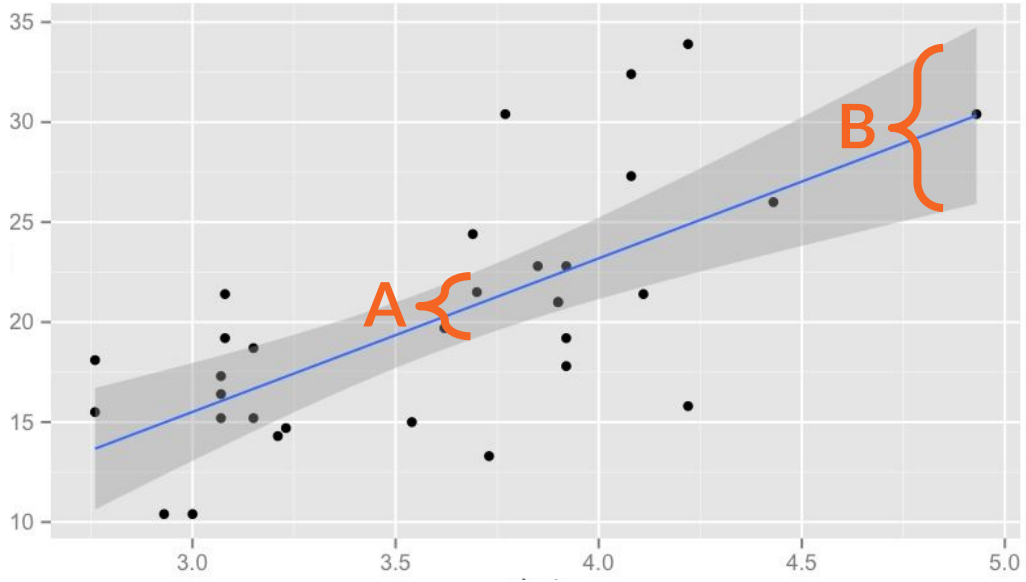


https://stats.stackexchange.com/questions/85560/shape-of-confidence-interval-for-predicted-values-in-linear-regression

Are we more confident about our predictions at A or B?

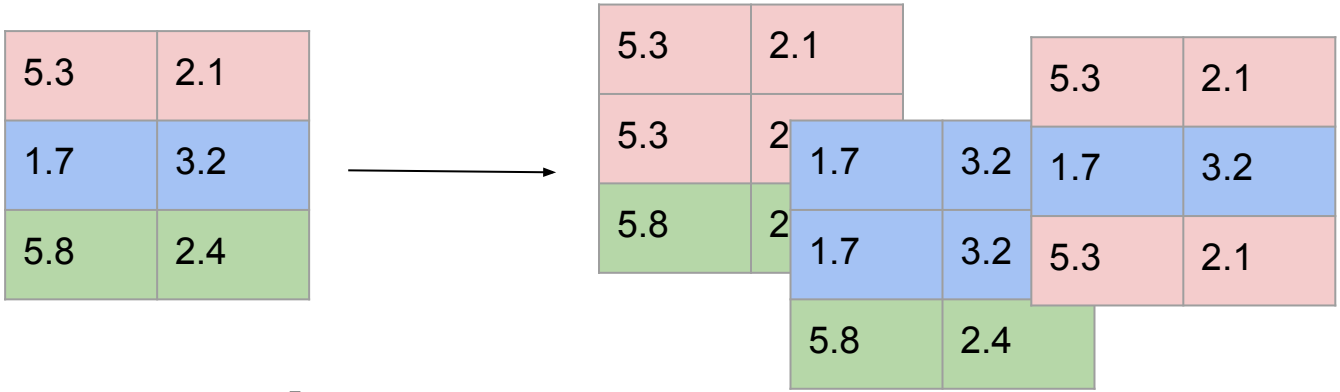
Narrower confidence interval → more confident (A)

Bootstrap gives us a "margin of error"

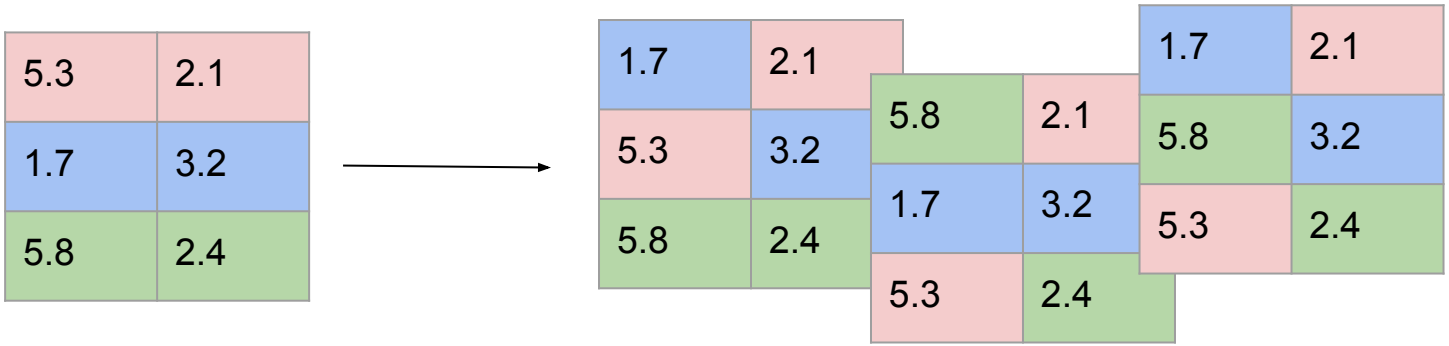


https://stats.stackexchange.com/questions/85560/shape-of-confidence-interval-for-predicted-values-in-linear-regression

Bootstrap

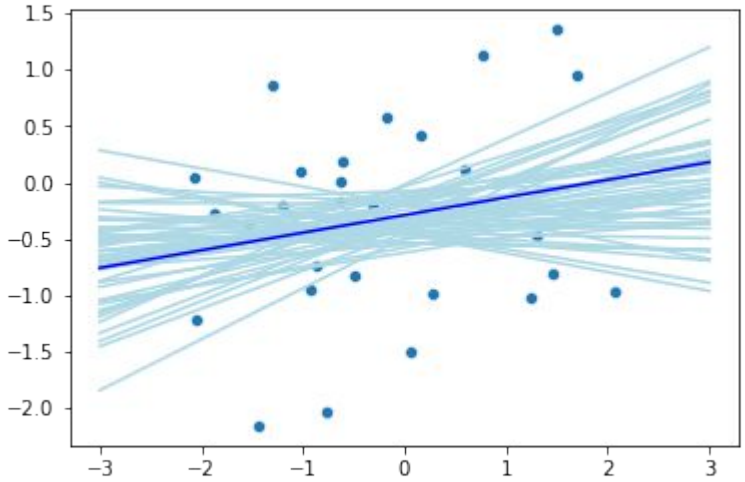


Permutation



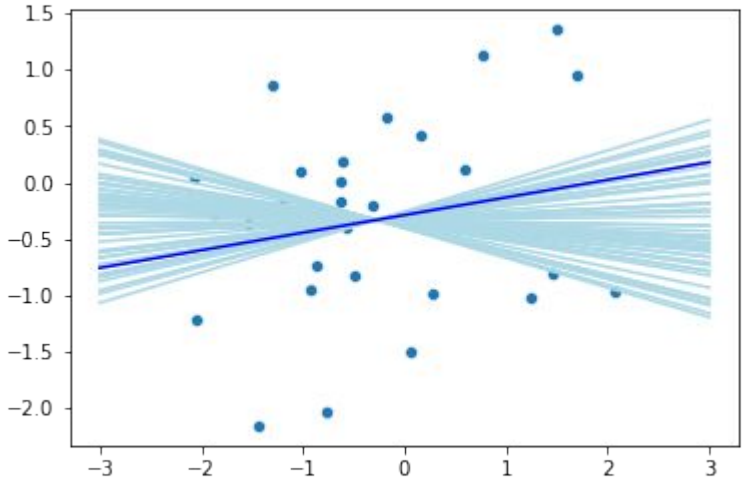
Bootstrap test Resampling may change (mean X, mean Y), so lines don't all pass through the same point.

Confidence region is centered around original slope



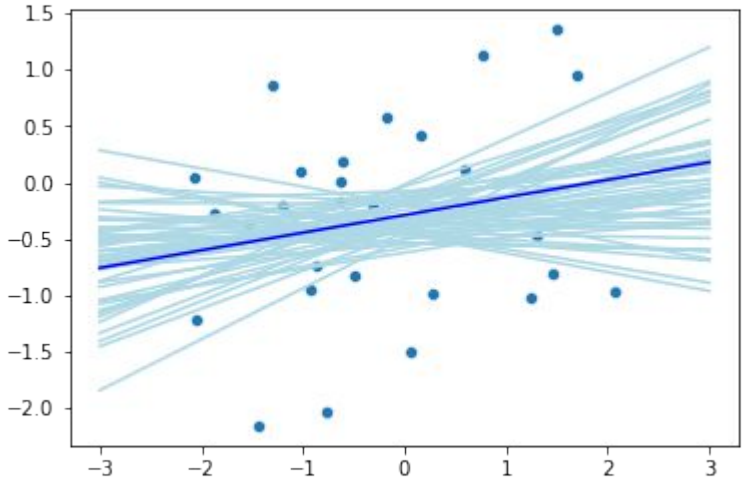
Permutation test All regression lines go through the point at (mean X, mean Y). Permutation doesn't change these means.

Confidence region is centered around 0 slope (not Y=0!)



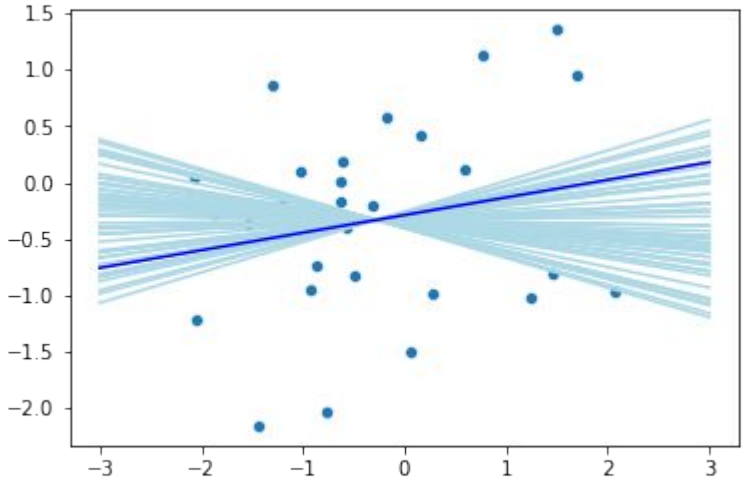
Bootstrap test

Generally: use for confidence intervals



Permutation test

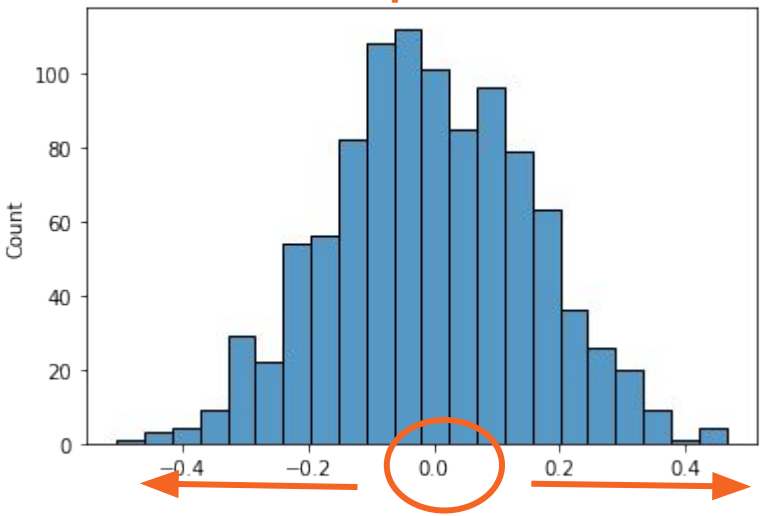
Generally: use for hypothesis tests



Compare your slope to permutation slopes

```
seaborn.histplot(permutation_slopes, bins=30)
```

Permutations of β centered at zero



Instead of bootstrapping/permutations, can use probability distributions to get margin of error, significance

Often need to assume i.i.d.
(i_____ & i_____ d_____)

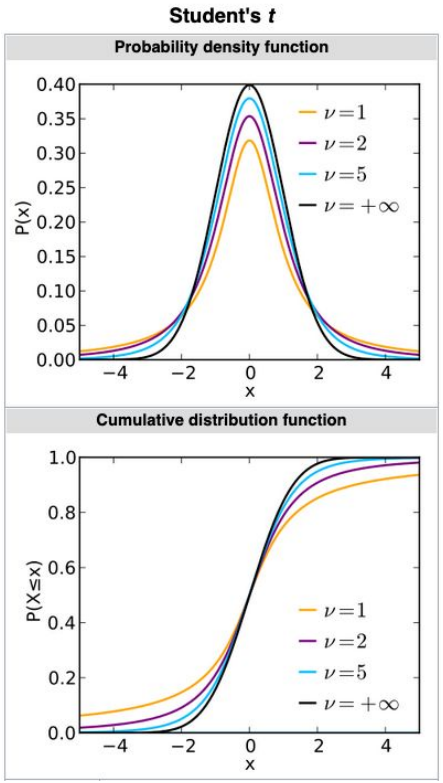
Instead of bootstrapping/permutations, can use probability distributions to get margin of error, significance

Often need to assume i.i.d.
(independent & identically distributed)

Instead of bootstrapping/permutations, can use probability distributions to get margin of error, significance

- Often need to assume i.i.d. (independent & identically distributed)
- Determine whether you should use pdf or cdf for distributions (= vs. >)

Hypothesis Testing: Regression

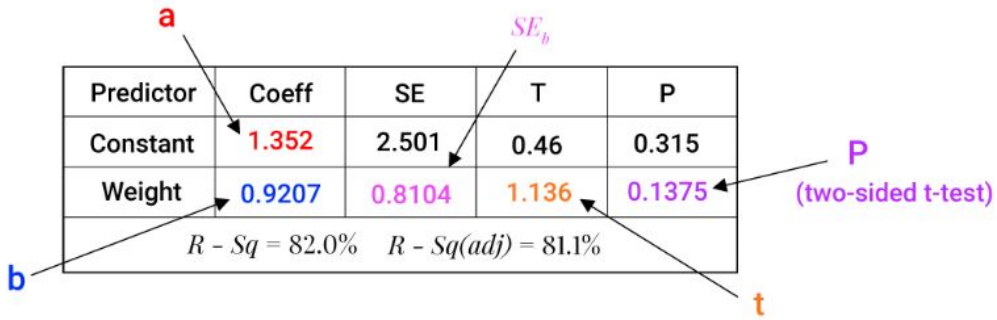


- t-statistic: $\mathbf{t = (b - \beta) / SE_b}$ Just set $\beta = 0$
  - Calculate t-statistic (free in Python)
    - β = the coefficient under the null
  - Compare to t-distribution
  - Decide if t spooky enough to reject null

https://en.wikipedia.org/wiki/Student%27s_t-distribution

Hypothesis Testing: Regression

Is the weight coefficient significant at the 5% level?



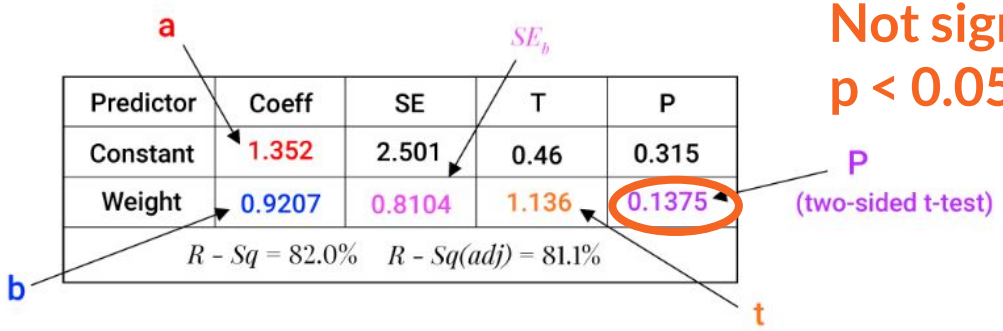
Predictor	Coeff	SE	T	P
Constant	1.352	2.501	0.46	0.315
Weight	0.9207	0.8104	1.136	0.1375
$R - Sq = 82.0\%$		$R - Sq(adj) = 81.1\%$		

If $H_o: \beta = 0$ and $H_a: \beta \neq 0$, then

Test Statistic: $t = \frac{b-\beta}{SE_b} \rightarrow t = \frac{b-\beta}{SE_b} \rightarrow t = \frac{0.9207 - 0}{0.8104} = 1.136$

https://calcworkshop.com/linear-regression/t-test/

Hypothesis Testing: Regression



Not significant, $p < 0.05$

If $H_o: \beta = 0$ *and* $H_a: \beta \neq 0$, then

Test Statistic: $t = \frac{b - \beta}{SE_b} \rightarrow t = \frac{b - \beta}{SE_b} \rightarrow t = \frac{0.9207 - 0}{0.8104} = 1.136$

https://calcworkshop.com/linear-regression/t-test/

Now we can also think about the CI for each individual slope (instead of regression output)!

$100(1-\alpha)\%$ **Confidence Interval for Slope**

Point Estimate $\pm$ **Margin of Error**

$$b \pm t^{*}\left(SE_{b}\right)$$

https://en.wikipedia.org/wiki/Confidence_interval

Hypothesis testing for regressions

Hypothesis 1: The cost of ice cream changes with distance from Ithaca, NY ($\beta_{distance} \neq \_\_\_$)

Hypothesis 2: Chocolate-based flavors are more popular on the East Coast than in the Midwest ($\beta_{eastcoast} \_\_ \_\_$)

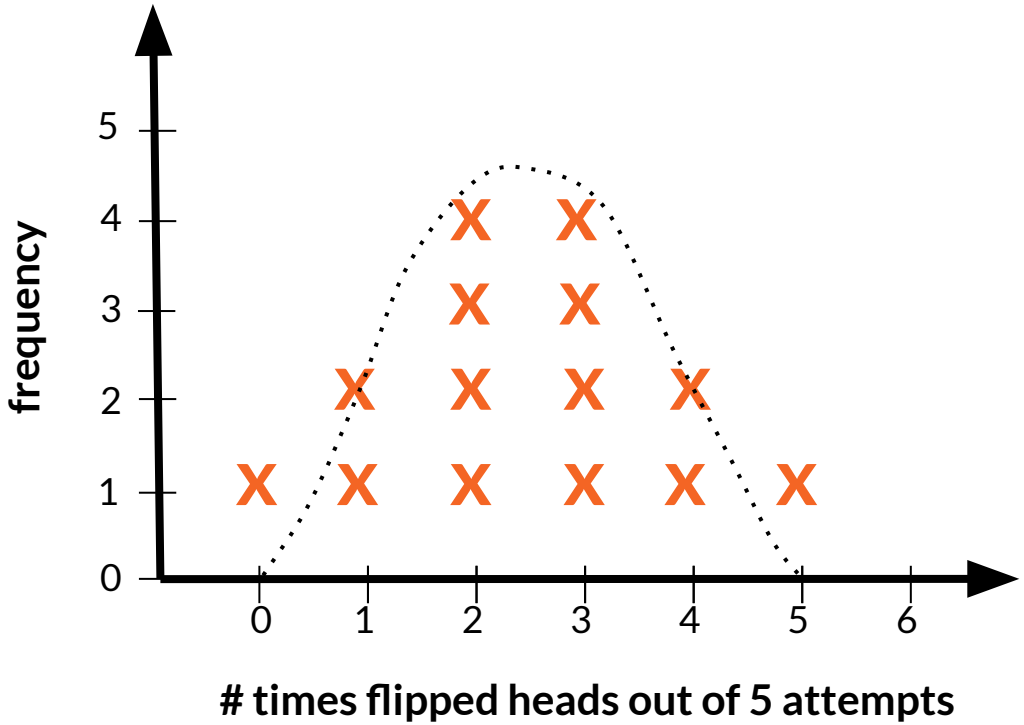
(reference dummy)

Hypothesis testing for regressions

Hypothesis 1: The cost of ice cream changes with distance from Ithaca, NY ($\beta_{distance} \neq 0$)

Hypothesis 2: Chocolate-based flavors are more popular on the East Coast than in the Midwest ($\beta_{eastcoast} > 0$)

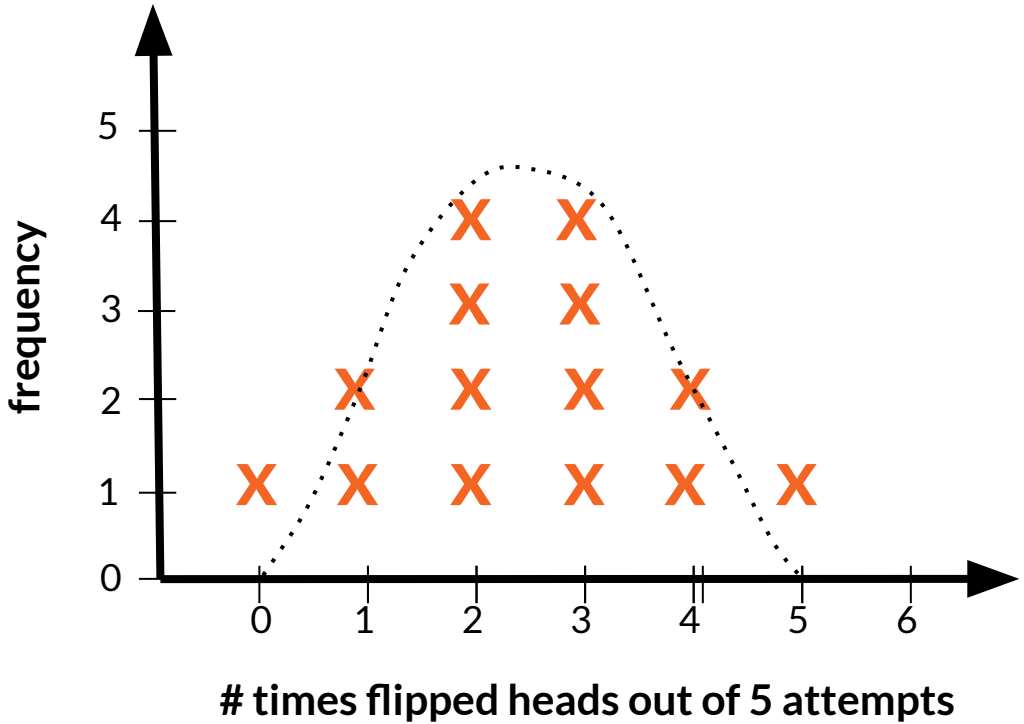
What distribution is this based on?



What distribution is this based on?

Binomial distribution:
Counting the number of **positive events X out of total events N** where each event has **probability *p*** to be positive

(at large N → normal distribution)



Binomial distribution μ and σ

$$\mathbb{E}[X] = ?$$

$$Var[X] = N\ p(1-p)$$

$$Std[X] = \sqrt{N\ p(1-p)}$$

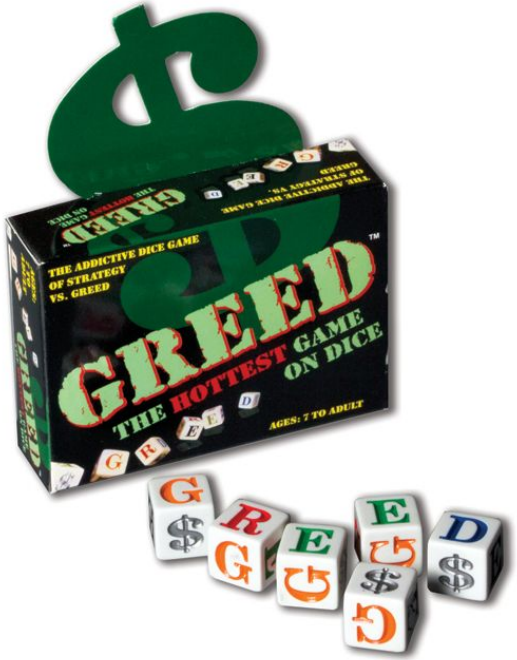
Binomial distribution μ and σ

$$\mathbb{E}[X] = N\ p$$

$$Var[X] = N\ p(1-p)$$

$$Std[X] = \sqrt{N\ p(1-p)}$$

The Greed Game: which distribution to model when someone wins with *k* points?



- If I roll a 1-5, you get +1 point
- If I roll a 6, you reset to 0 points and you're out of the game
- You can sit down at any time (if a 6 has not yet been rolled) and keep the points you've accumulated

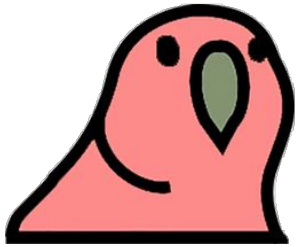
https://woodexpressions.com/wp-content/uploads/2017/03/Greed.jpg https://www.statsmedic.com/apstats-chapter6-day9

Greed game: Geometric distribution

Shifted Geometric Distribution Cumulative Distribution Function	$P(X \le k) = 1-(1-p)^k$ $P(X \ge k) = (1-p)^{k-1}$ $P(X > k) = 1-P(X \le k) = (1-p)^k$

- What is the probability that someone won with *n* points?
  - **The probability that the first "success" (i.e., the game ending) takes at least *n+1* trials = $P(X \ge n+1) = (1-p)^n = (1-⅙)^n$ =?**

https://calcworkshop.com/discrete-probability-distribution/geometric-distribution/



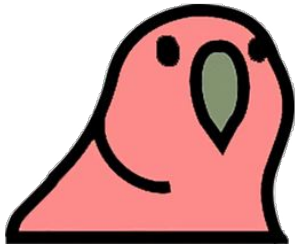
What distribution to model this?

We survey students, and we know that on average 3 out of every 75 students love parrots. What is the probability it takes asking **6 students** to find the first **3** parrot-loving students?

X = number of students to survey
k = 6, r = 3, p = 3/75, want to find P(X = k)

https://www.producthunt.com/products/party-parrots

https://online.stat.psu.edu/stat414/lesson/11/11.5



Negative Binomial

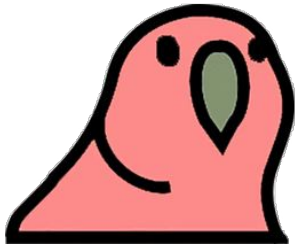
We survey students, and we know that on average 3 out of every 75 students love parrots. What is the probability it takes asking **6 students** to find the first **3** parrot-loving students?

X = number of students to survey (i.e. trials until 3 successes)

$P(X = k) = (5 \text{ choose } 2) * (1-(3/75))^{6-3} * (3/75))^3 =$
$10*0.96^3*0.04^3 = 0.000566$
Unlikely it'd take only 6 trials to find 3 parrot lovers!

https://www.producthunt.com/products/party-parrots

https://online.stat.psu.edu/stat414/lesson/11/11.5

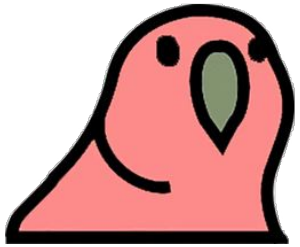


What distribution to model this?

There is a mean of 3 parrot-lovers per discussion section. What is the probability that a randomly selected discussion section has one parrot-lover?

X = number of parrot lovers in a discussion section
λ = mean = 3

https://www.producthunt.com/products/party-parrots



Poisson Example

There is a mean of 3 parrot-lovers per discussion section. What is the probability that a randomly selected discussion section has one parrot-lover?

$$\Pr(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$P(X=1) = 3^1 * e^{-3} / 1! = 3e^{-3} = 0.15$

https://www.producthunt.com/products/party-parrots

What distribution to model this?

- Our hypothesis test's *test statistic* will be a sum of a squared value. Instead of a "z-score" we now have:

$$\sum_{(i \cdot j)} \frac{(O - E)^2}{E}$$

- O = observed values
- E = expected values
- i = the number of rows in the table
- j = the number of columns in the table

https://courses.lumenlearning.com/introstats1/chapter/test-of-independence/

χ^2 distribution for independence tests

- Our hypothesis test's *test statistic* will be a sum of a squared value. Instead of a "z-score" we now have:

$$\sum_{(i \cdot j)} \frac{(O - E)^2}{E}$$

- O = observed values
- E = expected values
- i = the number of rows in the table
- j = the number of columns in the table

https://courses.lumenlearning.com/introstats1/chapter/test-of-independence/

Is the null hypothesis spooky or boring?

What was that noise?

Boring hypothesis

It was just the wind. Old houses make strange noises. Go back to sleep.

Spooky hypothesis

It's a ghost! Don't go in the basement!!!

65

https://www.housebeautiful.com/lifestyle/a38046654/sk-pierce-haunted-victorian-mansion/

Null

Alternative

What was that noise?

Boring hypothesis

It was just the wind. Old houses make strange noises. Go back to sleep.

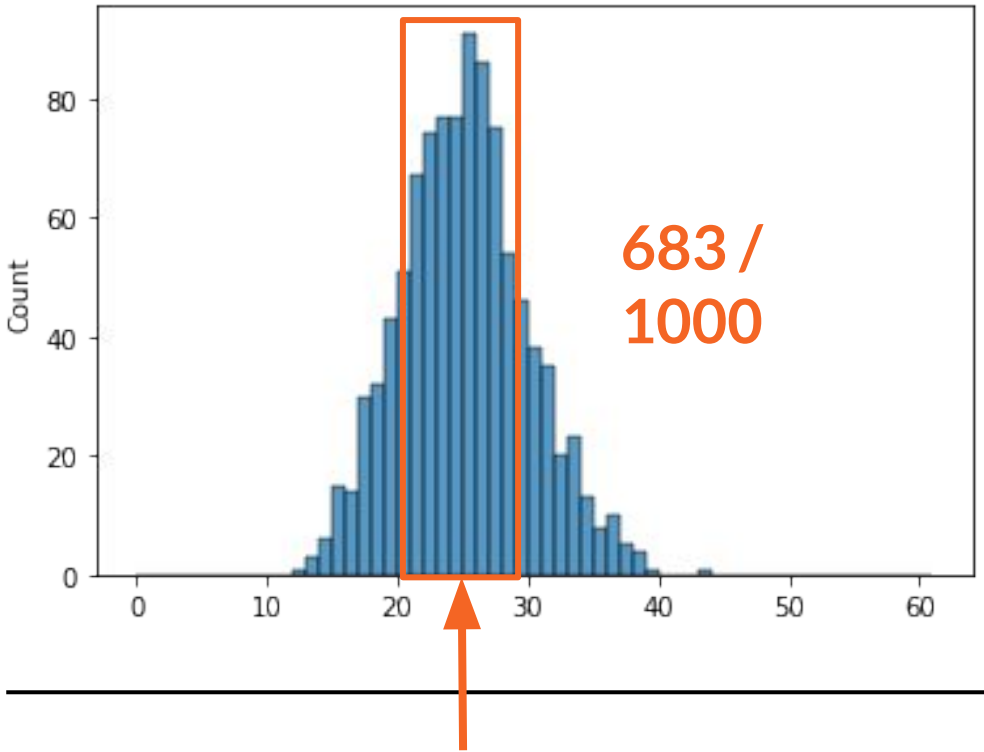
Spooky hypothesis

It's a ghost! Don't go in the basement!!!

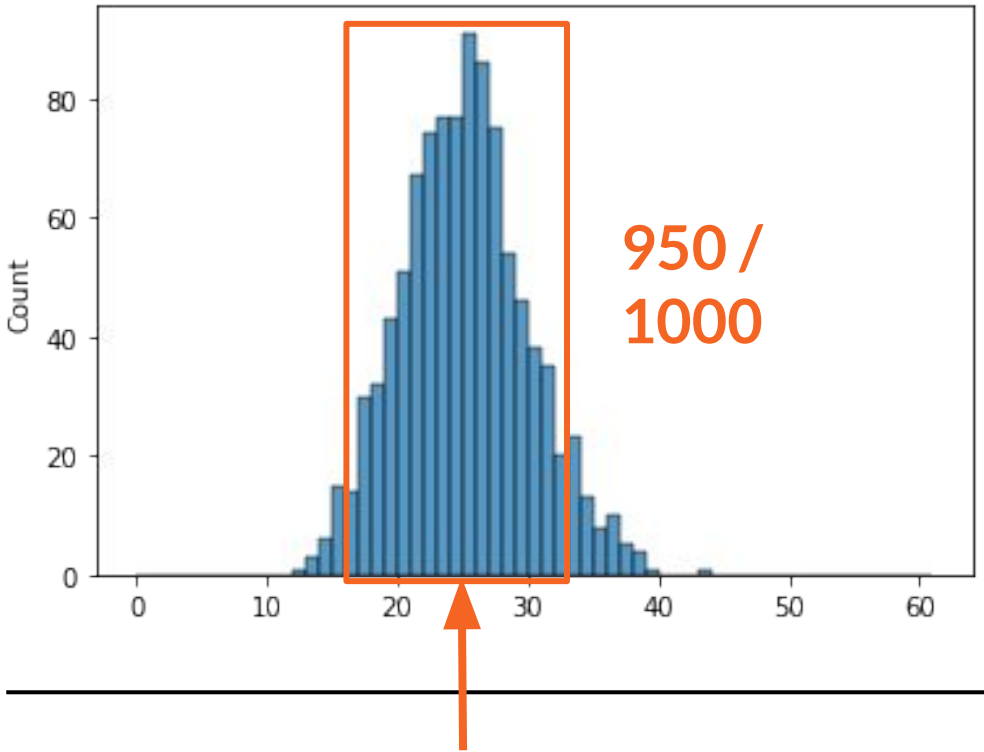
65

https://www.housebeautiful.com/lifestyle/a38046654/sk-pierce-haunted-victorian-mansion/

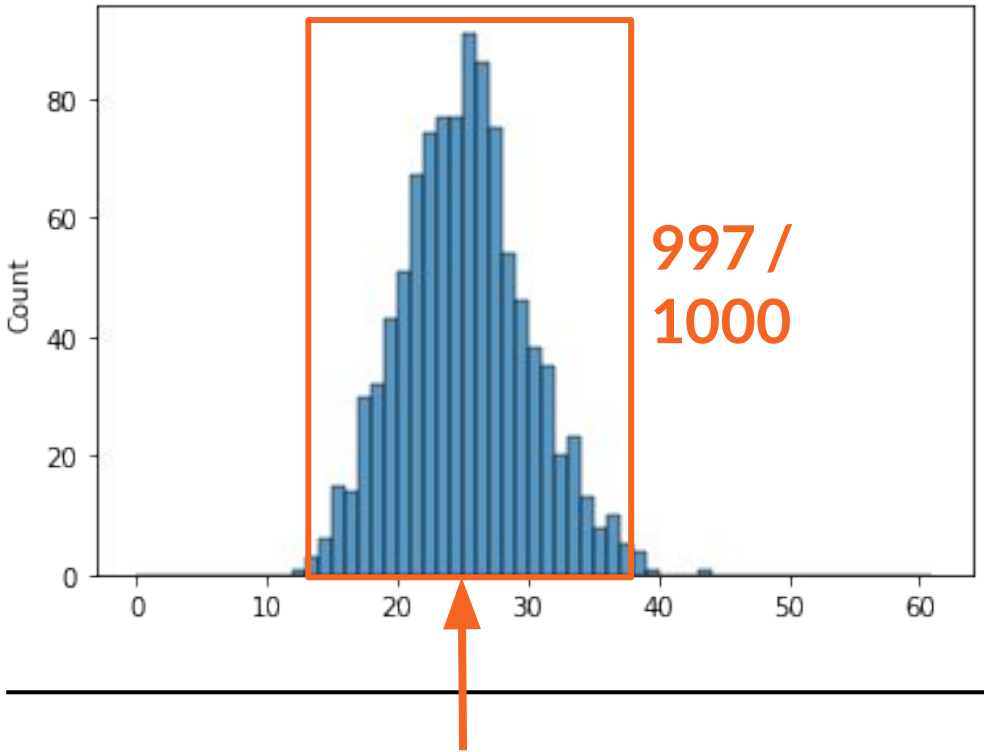
One standard deviation: boring



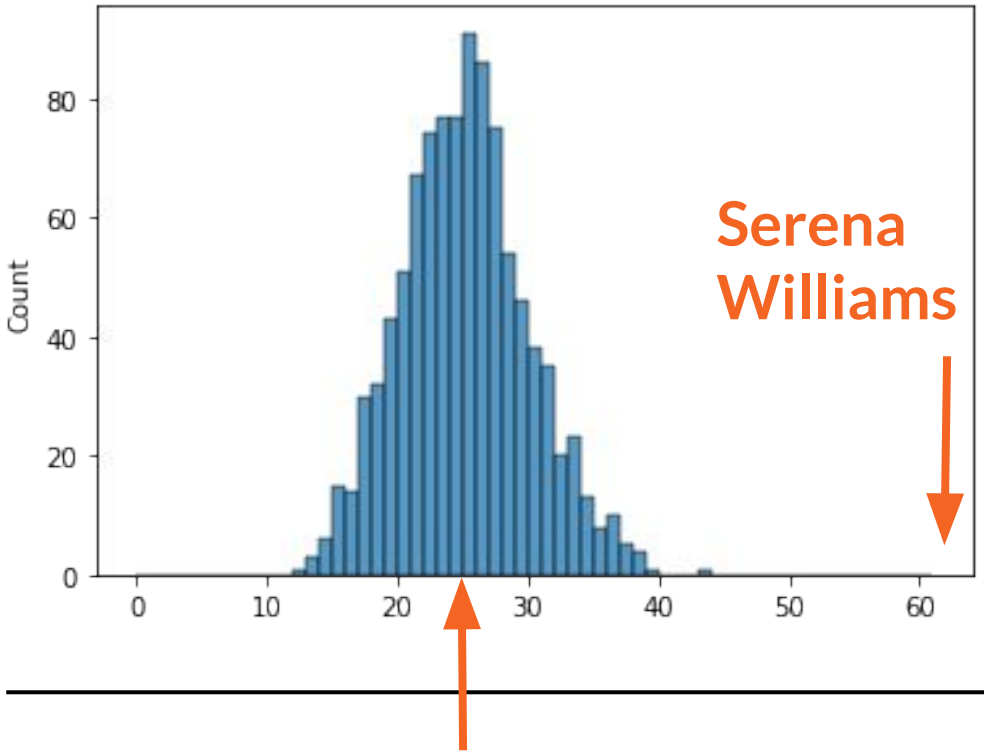
Two standard deviations: boring?



3 standard deviations: spooky?



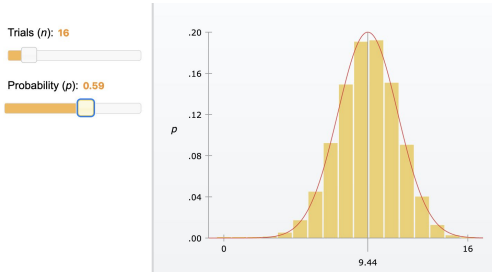
Many standard devs: very spooky



The 68/95/99.7 rule

If a distribution is approximately normal:

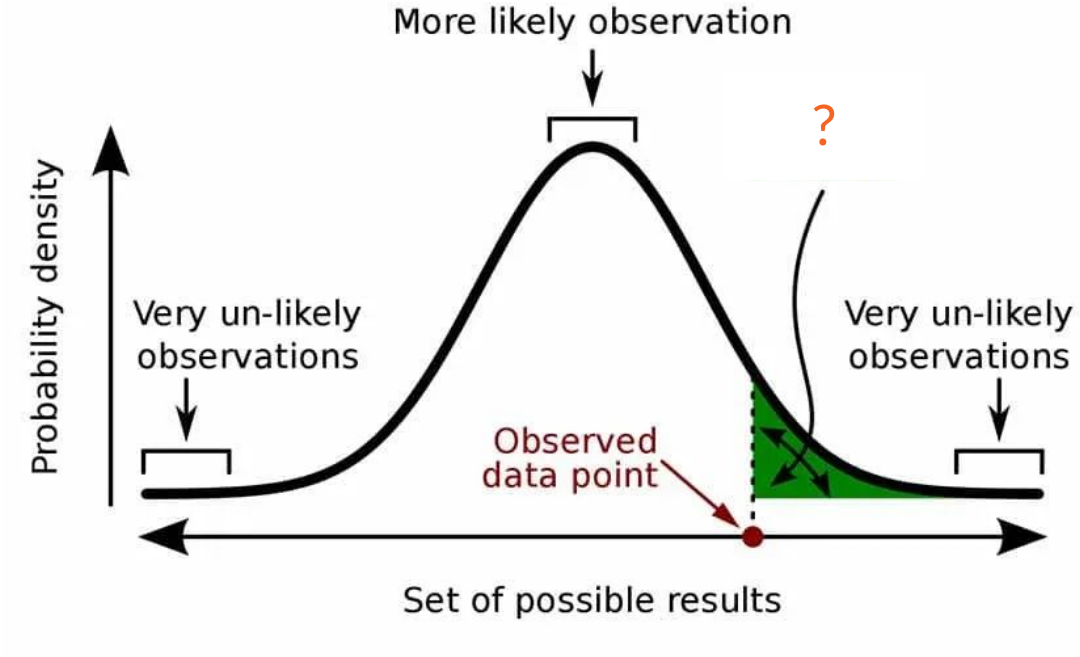
- 68% within ONE standard deviation
- 95% within TWO standard deviations
- 99.7% within THREE standard deviations
- Almost nothing outside 3 sd



When N is large and p is not close to 0 or 1, binomial is approximately normal

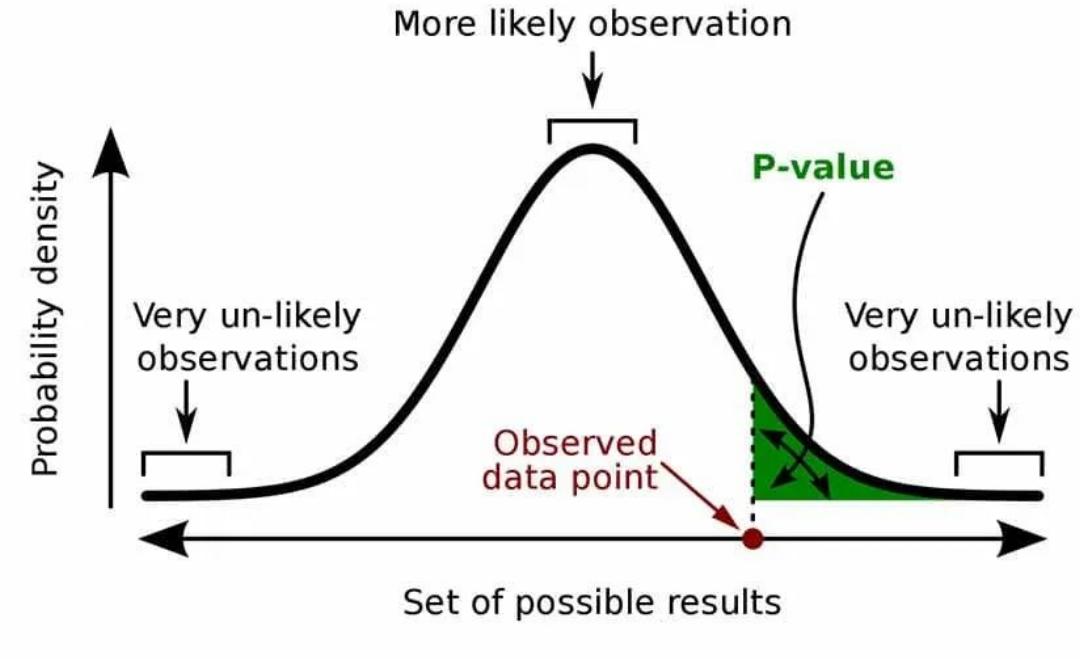
https://datasciencedojo.com/blog/data-science-memes/

What is the thing in green?



https://www.simplypsychology.org/p-value.html

P-value: The probability that something occurred, given the null hypothesis is true



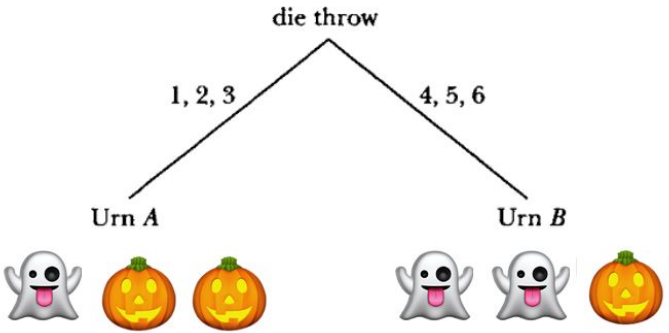
https://www.simplypsychology.org/p-value.html

Hypothesis test suggestions

- Know how to formulate a hypothesis (both a null and alternative)
- Know how to choose what distributions (and their mean/var formulas) to use
- Know how to work with contingency tables
  - Independence tests
  - Joint, marginal, conditional probabilities

Bayes' rule

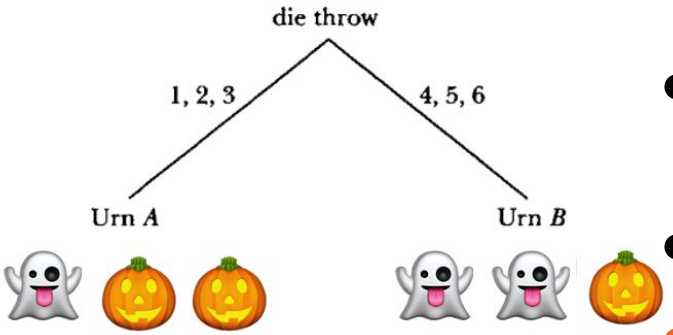
$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$



https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.10.2.179

Bayes' rule

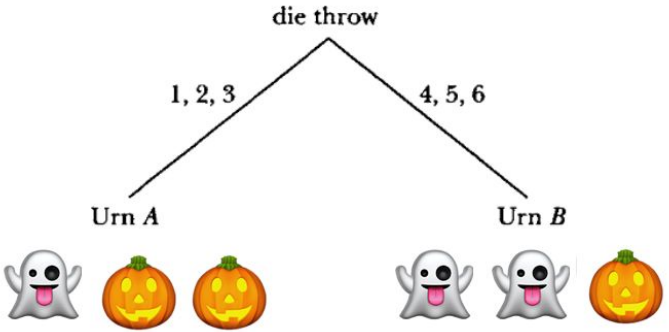
$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$



- First draw from urns was a pumpkin. It is returned to original urn **with replacement**
- And now, we take a second draw from the **same urn**, and get a pumpkin
- What is the probability we drew from Urn A?
- **How do we define A and B (in English) if we use Bayes' rule?**

https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.10.2.179

Bayes' rule



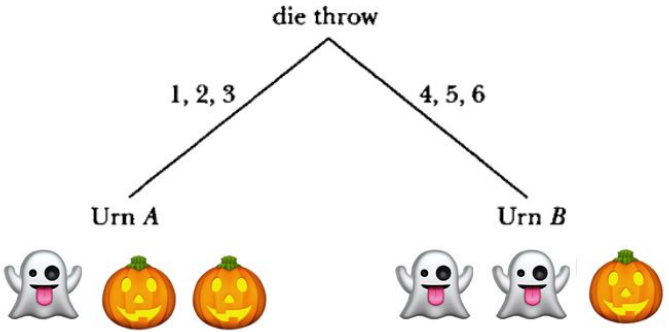
$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

A = drawing from urn A

B = 1st draw 🎃 and 2nd draw 🎃

https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.10.2.179

Bayes' rule: solve!



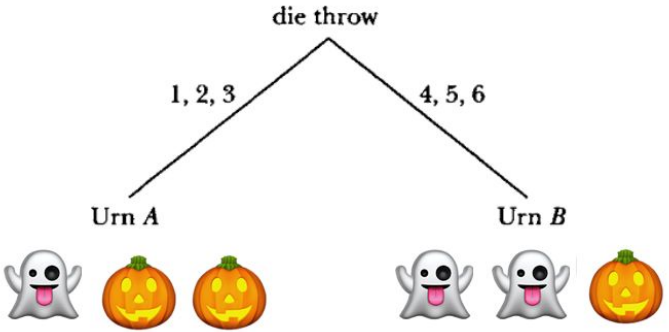
$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

A = drawing from urn A

B = 1st draw 🎃 and 2nd draw 🎃

https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.10.2.179

Bayes' rule



$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

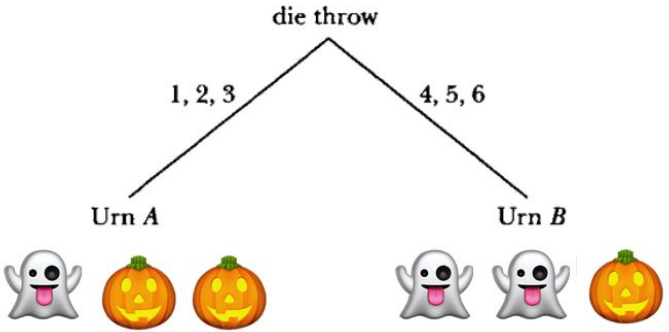
P(A&B) = ½ * ⅔ * ⅔ =

A = drawing from urn A

B = 1st draw 🎃 and 2nd draw 🎃

https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.10.2.179

Bayes' rule



$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

P(A&B) = ½ * ⅔ * ⅔ =

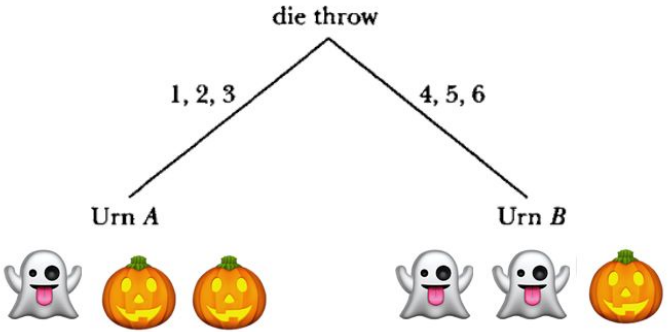
= ½ * ⅔ * ⅔ + ½ * ⅓ * ⅓

A = drawing from urn A

B = 1st draw 🎃 and 2nd draw 🎃

https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.10.2.179

Bayes' rule



$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

P(A&B) = ½ * ⅔ * ⅔ =

= 4/18 / 5/18

= 4/5

= ½ * ⅔ * ⅔ + ½ * ⅓ * ⅓

A = drawing from urn A

B = 1st draw 🎃 and 2nd draw 🎃

https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.10.2.179

What method would you use for classifying text: sports or not?

Training data

Want to classify new data

Text	Tag
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports
"A very close game"	

https://monkeylearn.com/blog/practical-explanation-naive-bayes-classifier/

Naive Bayes!

Training data

Want to classify new data

Text	Tag
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports
"A very close game"	

https://monkeylearn.com/blog/practical-explanation-naive-bayes-classifier/

Naive Bayes!

In math: P(Sports | “A very close game”)

Notice: we aren’t outputting a binary estimate of Sports or Not Sports – we’re outputting an estimate of the probability that our tag is Sports

Want to classify new data

Text	Tag
"A very close game"	Sports / Not Sports?

https://monkeylearn.com/blog/practical-explanation-naive-bayes-classifier/

Naive Bayes Assumptions

?

$$P(y|x_1,...,x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

$$P(y|x_1,...,x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y)$$

$$y = argmax_y P(y)\prod_{i=1}^{n} P(x_i|y)$$

https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html

Naive Bayes Assumptions

Assume all words are independent!!

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

$$P(y|x_1, ..., x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html

Naive Bayes Assumptions

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

We can ignore comparing the denominator since it's constant w.r.t. y

$$P(y|x_1, ..., x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

The goal of calculating $P(y|x_1, ..., x_n)$ is to find the argmax of it!

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html

Naive Bayes!

What are two things we should make sure to do to avoid computation issues?

Naive Bayes!

What are two things we should make sure to do to avoid computation issues?

1. Add log probabilities instead of multiplying raw probabilities
2. Laplace correction so everything doesn't get 0'd out

Log prob quiz!

What probability has $\log_e$ **-2.3**?

One in _____

What is the $\log_e$ of **One in 100,000**?

_____

Log prob quiz!

What probability has $\log_e$ **-2.3**?

One in ___10____

What is the $\log_e$ of **One in 100,000**?

___-11.5_____

Log intuition quiz!!!

If the odds are **99,999 to 1** (large), the log odds ratio is _____

Log intuition quiz!!!

If the odds are **99,999 to 1** (large), the log odds ratio is **11.5**

Log probabilities

One in	Probability	Log_{10}	Log_e
10	0.1	-1	-2.3
100	0.01	-2	-4.6
1,000	0.001	-3	-6.9
10,000	0.0001	-4	-9.2
100,000	0.00001	-5	-11.5

Machine learning methods

Which of these can I use to classify *known* binary y's?

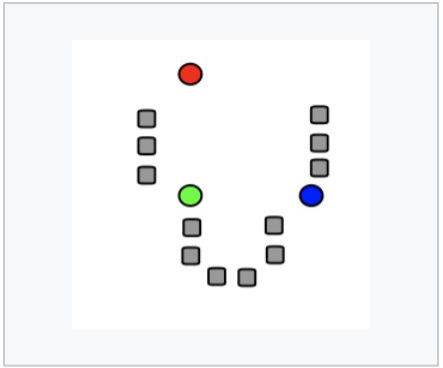
- Logistic regression
- Naive Bayes
- Clustering
- Neural nets

Machine learning methods

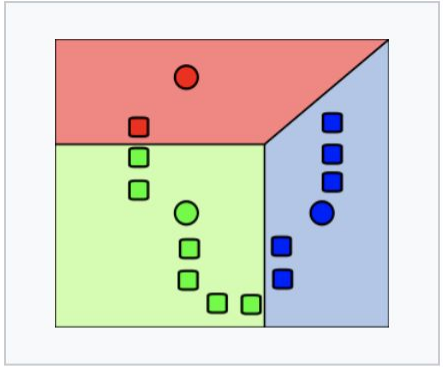
Which of these can I use to classify *known* binary y's?

- Logistic regression
- Naive Bayes
- ~~Clustering~~(no output needed for unsupervised learning algorithm!)
- Neural nets

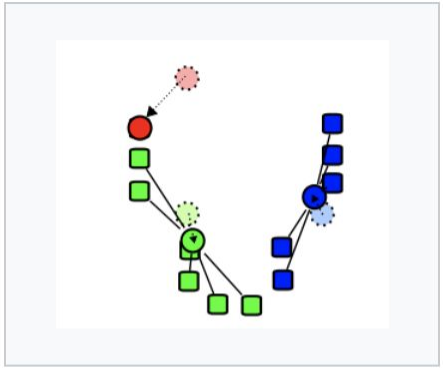
K-Means clustering: what is k here?



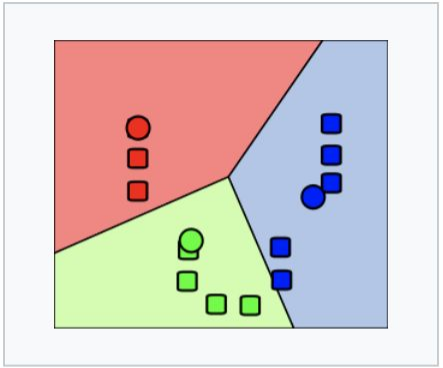
1. *k* initial "means" (in this case *k*=[redacted]) are randomly generated within the data domain (shown in color).



2. *k* clusters are created by associating every observation with the nearest mean.



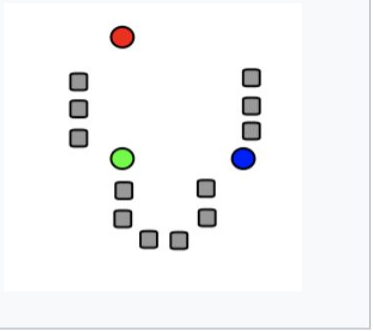
3. The centroid of each of the *k* clusters becomes the new mean.



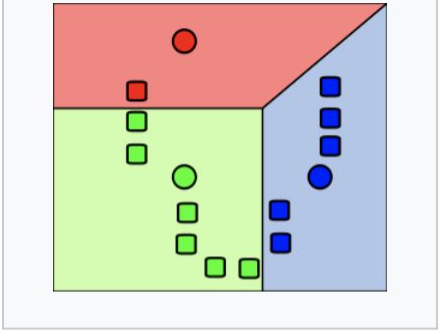
4. Steps 2 and 3 are repeated until convergence has been reached.

https://en.wikipedia.org/wiki/K-means_clustering

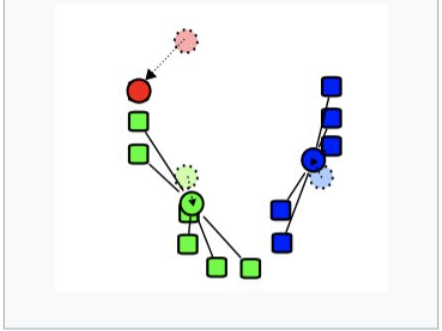
K= 3



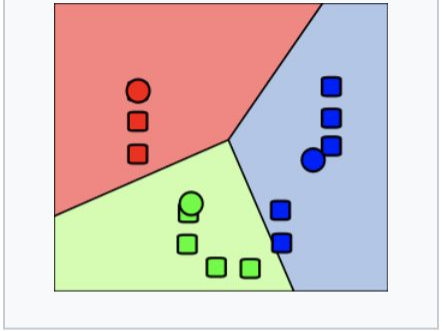
1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).



2. *k* clusters are created by associating every observation with the nearest mean.

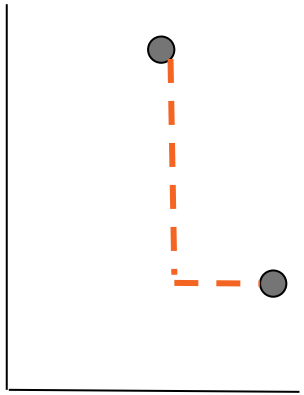


3. The centroid of each of the *k* clusters becomes the new mean.



4. Steps 2 and 3 are repeated until convergence has been reached.

https://en.wikipedia.org/wiki/K-means_clustering



Which distance metric is this pic?

Euclidean / ℓ_2 / "as the crow flies"

Use Pythagorean theorem: square root of sum of squares

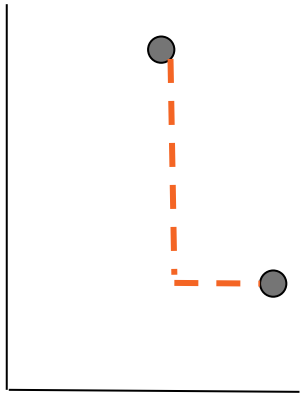
Absolute / ℓ_1 / Manhattan / "city block"

Sum of absolute values for each variable

Cosine / inner product

Ignore magnitude, compare angle between vectors

Which distance metric is this pic?



Euclidean / ℓ_2 / "as the crow flies"

Use Pythagorean theorem: square root of sum of squares

Absolute / ℓ_1 / Manhattan / "city block"

Sum of absolute values for each variable

Cosine / inner product

Ignore magnitude, compare angle between vectors

How might we recommend movies?

	User 1	User 2	User 3	User 4	···	User 13435
Airplane!	9	6		7		
Akira		4	7	8		8
Aladdin	6			7		
Alexander Nevsky				6		
...						
Zoolander			9	5		7

How might we recommend movies?

- **Collaborative Filtering**
- **SVD**

	User 1	User 2	User 3	User 4	···	User 13435
Airplane!	9	6		7		
Akira		4	7	8		8
Aladdin	6			7		
Alexander Nevsky				6		
...						
Zoolander			9	5		7

Collaborative Filtering

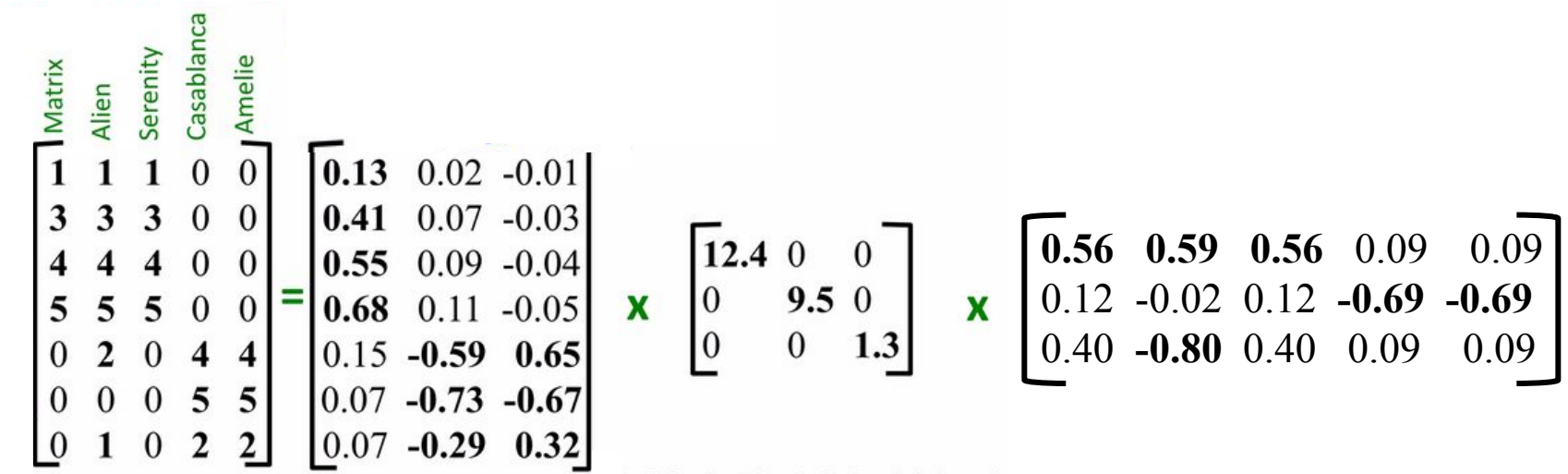
User-User

$$r_{xi} = \frac{\sum_{y \in N} s_{xy} \cdot r_{yi}}{\sum_{y \in N} s_{xy}}$$

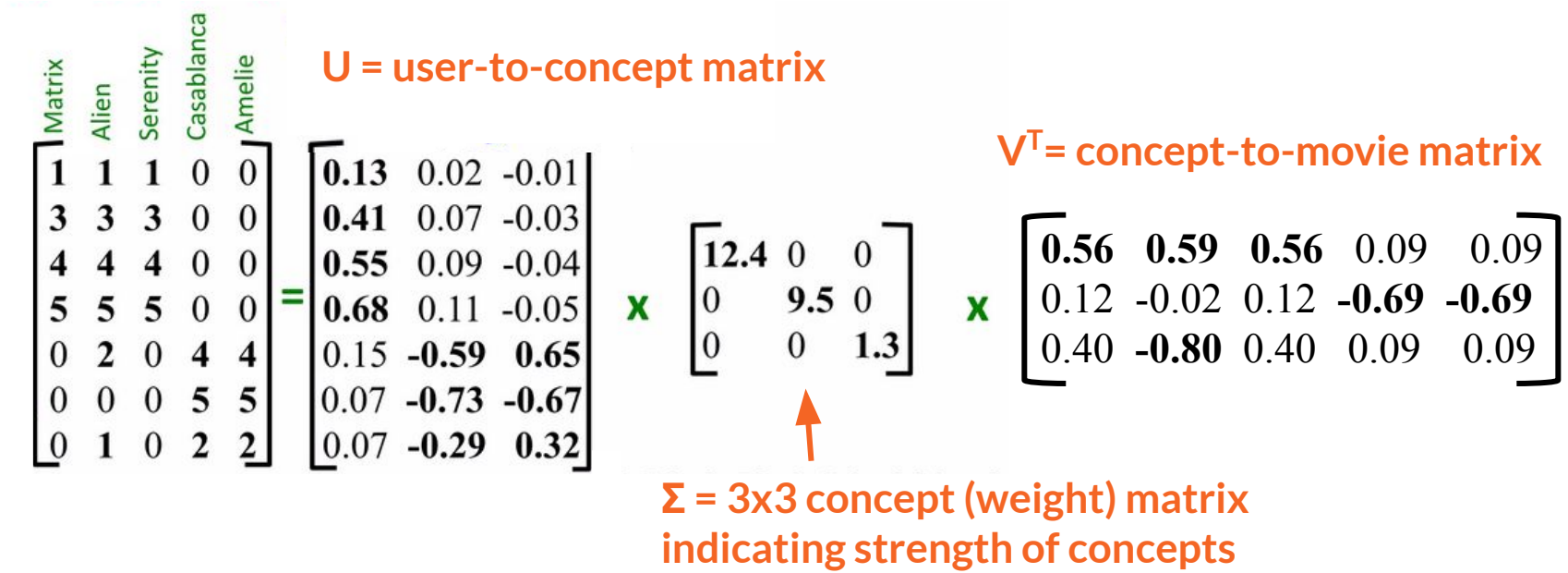
Item-Item

$$r_{xi} = \frac{\sum_{j \in N(i;x)} s_{ij} \cdot r_{xj}}{\sum_{j \in N(i;x)} s_{ij}}$$

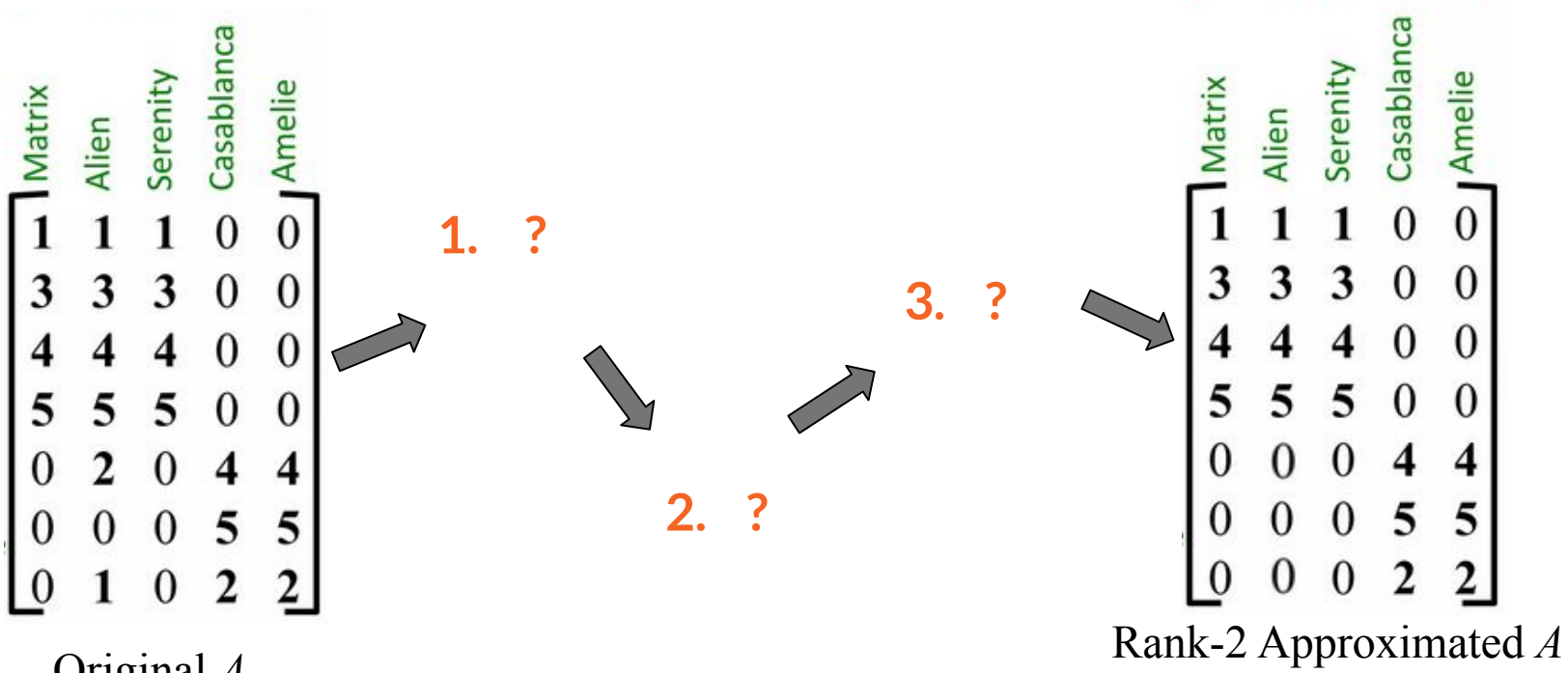
SVD: what are these 3 matrices called?



SVD: what are these 3 matrices called?



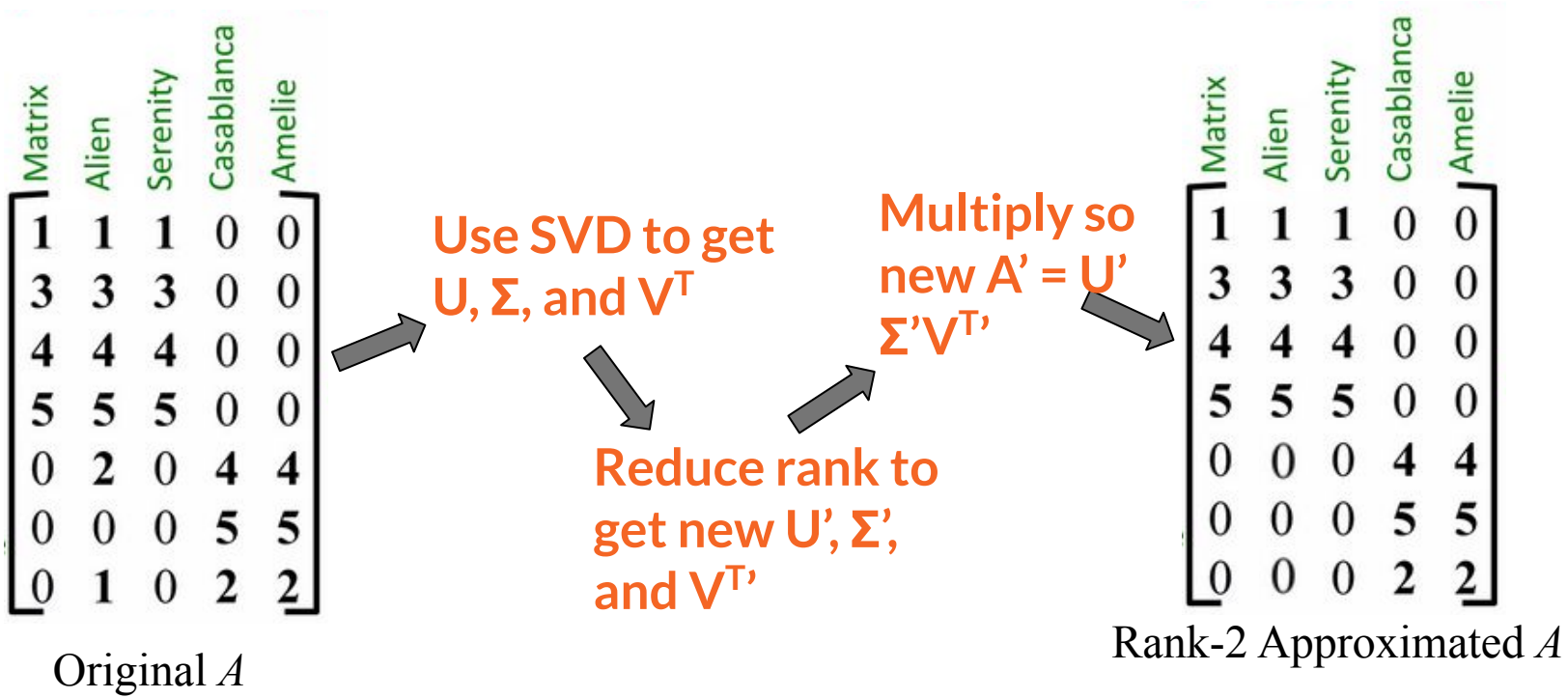
SVD: what are the steps?



Original A

Rank-2 Approximated A

SVD: original vs. approximation



Original *A*

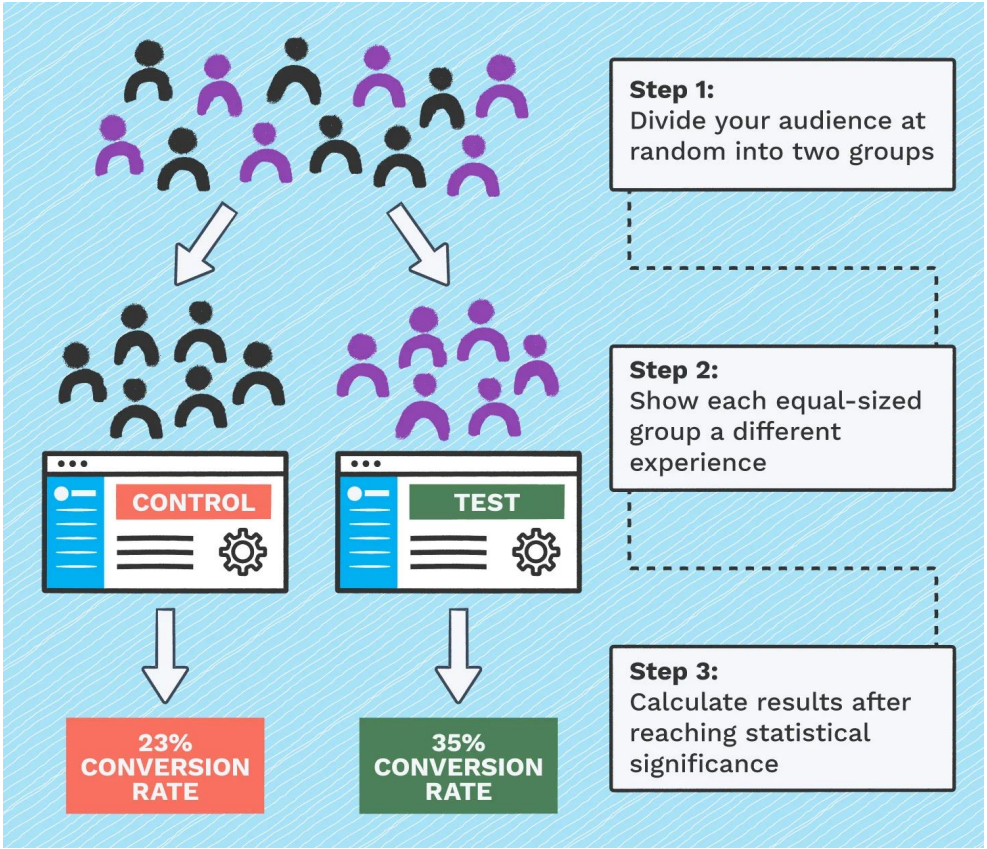
Rank-2 Approximated *A*

Why SVD?

- Allows you to **reduce dimensions** on really big (high dimensional) data
- Allows you to **interpret important concepts**
- Allows you to **approximate** any matrix
- Can be used to **fill in missing elements** if df is missing values at random

https://twitter.com/WomenInStat/status/1285610321747611653

A/B experimentation



https://www.adroll.com/blog/a-b-testing-an-introduction-infographic

Distill question into A and B

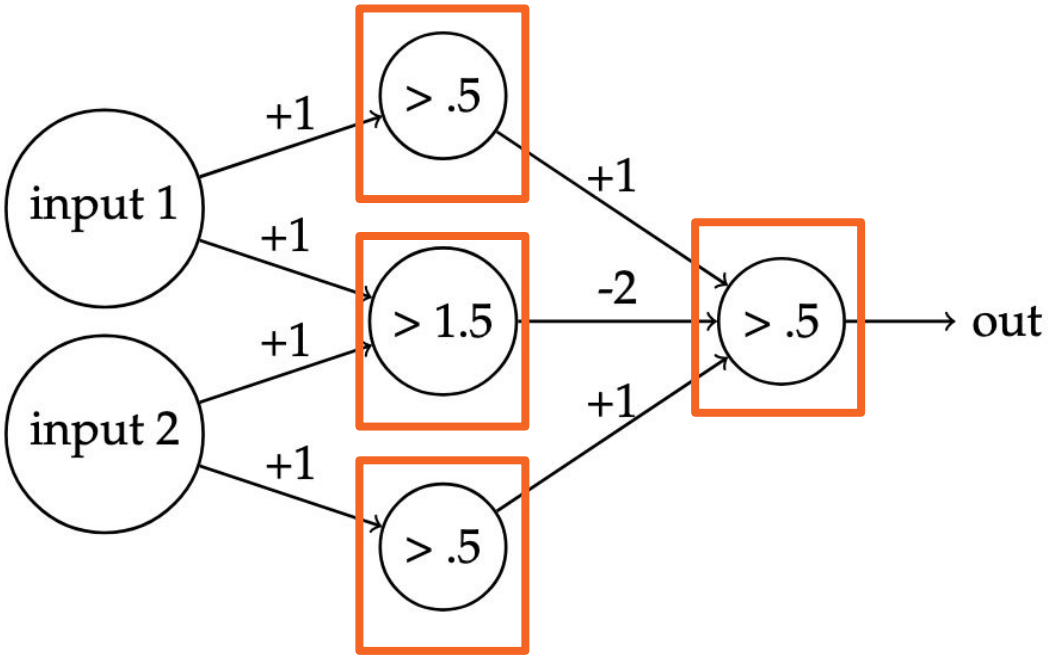
Question	A	B	Metric
Should we make “Tweet” button size bigger?	?	?	?

Distill question into A and B

Question	A	B	Metric
Should we make “Tweet” button size bigger?	Current (small) button size	New (bigger) button size	• # button clicks

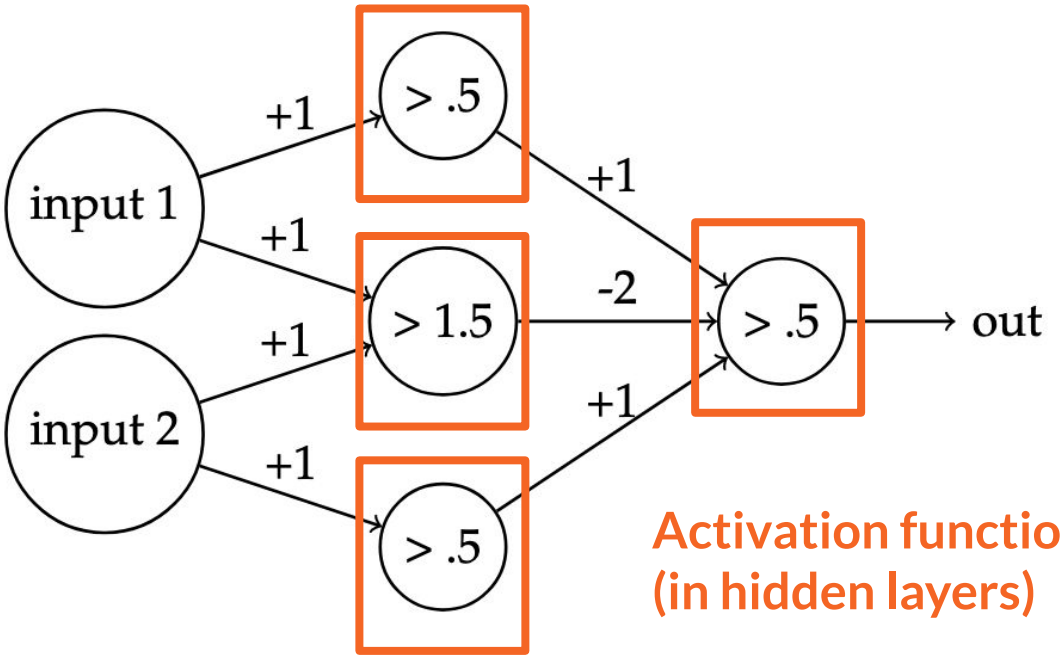
- Be able to describe some potential pitfalls of an experiment

What are these parts of a neural net called?



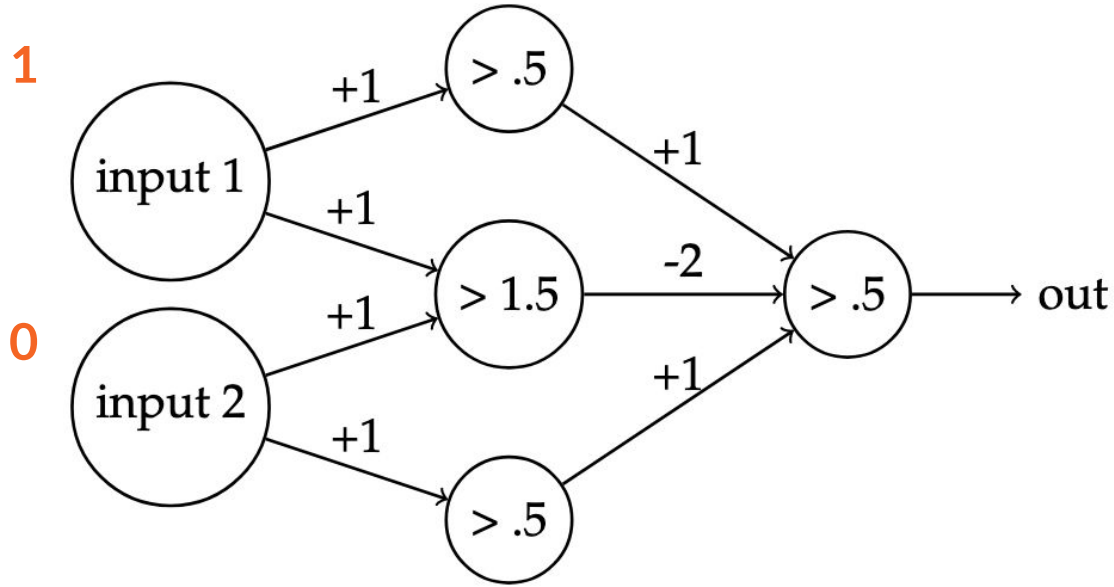
Rumelhart, Hinton & Williams. https://www.diva-portal.org/smash/get/diva2:1304961/FULLTEXT01.pdf

What are these parts of a neural net called?



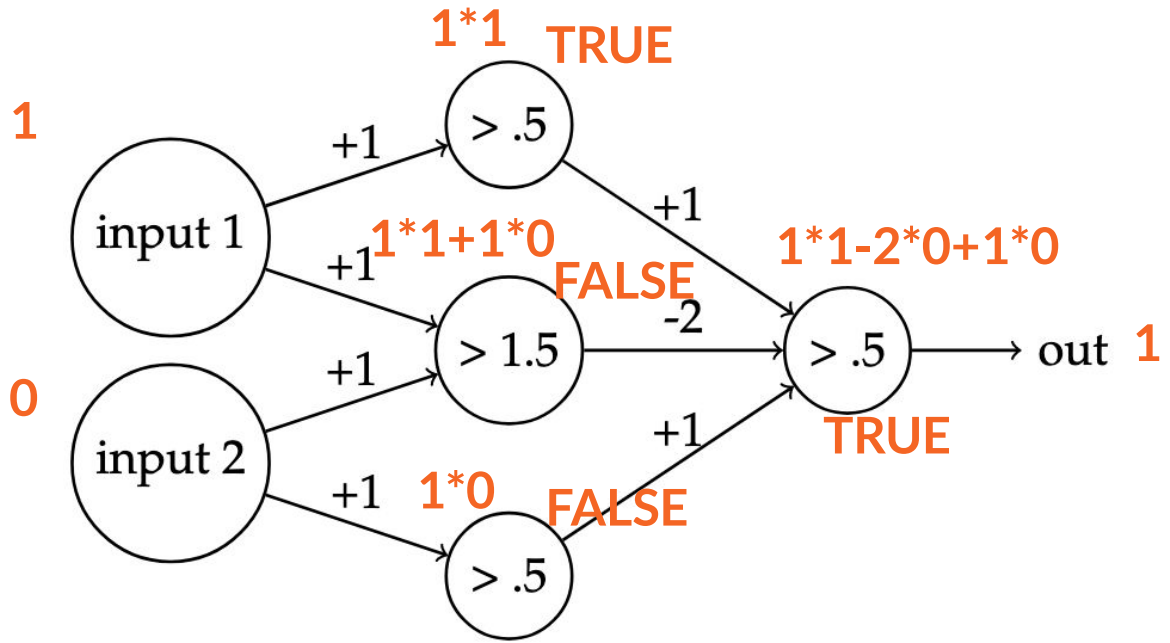
Rumelhart, Hinton & Williams. https://www.diva-portal.org/smash/get/diva2:1304961/FULLTEXT01.pdf

Multi-layer perceptron: XOR(1,0)?

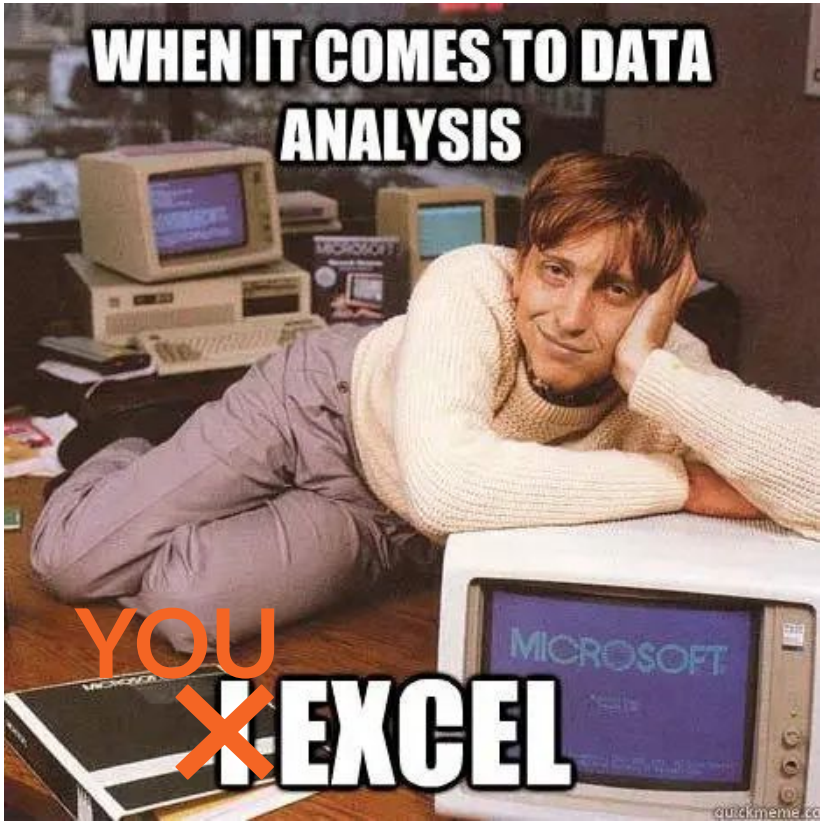


Rumelhart, Hinton & Williams. https://www.diva-portal.org/smash/get/diva2:1304961/FULLTEXT01.pdf

Multi-layer perceptron: XOR(1,0)?



Rumelhart, Hinton & Williams. https://www.diva-portal.org/smash/get/diva2:1304961/FULLTEXT01.pdf



https://datasciencedojo.com/blog/data-science-memes/

For each, name one tool

1. Programming with data
2. Describing one variable
3. Describing relationships between two variables
4. Predicting one variable from others
5. Distinguishing pattern from randomness

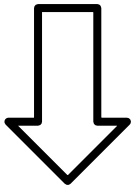
Examples

1. Pandas, SQL
2. Mean, median, variance
3. Covariance, correlation (Pearson, Spearman)
4. Regression, classification
5. Bootstrap, t-test, permutation test

From Lecture 1…

What is INFO 2950?

- Computational tools + real data + [*data science skills*]



- Use data ethically to create evidence to support an argument

What will you get out of 2950?

- The ability to think critically about & generate your own data analyses
- Foundational knowledge for future DS, ML, AI courses
- The skills to ace a basic DS job interview in industry

Course Goals

- Use statistical methods, data processing, and real-world knowledge to make arguments using data

Course Goals

- Use statistical methods, data processing, and real-world knowledge to make arguments using data
- Ability to execute each phase of a typical DS project:
  - Data collection
  - Exploration and summarization
  - Model fitting
  - Hypothesis testing
  - Communication of findings

Ethics, always

https://idatassist.com/are-ethics-in-data-science-really-that-important/

Choose the right research questions

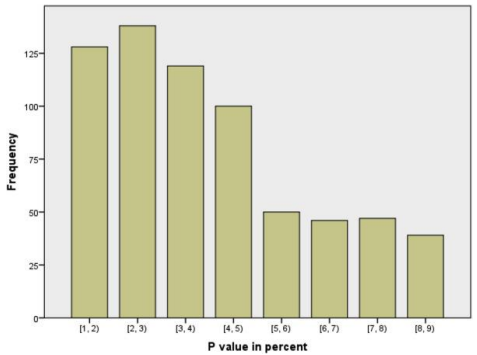
- Remember that your work has repercussions in the real world (e.g. eugenicists Fisher, Spearman, et al.)
- What data would you actually need to have in order to answer your question? Is it possible to get it?
- Are you grouping people or places together in ways that hide important distinctions?

Data paranoia

- Be careful with getting data (use reputable sources; respect private data; don't scrape personal data without consent)
- Be careful about data quality (null island)
- Be suspicious when reusing existing data. Is it really what you think it is?

https://www.bloomberg.com/news/newsletters/2022-05-04/maplab-inside-null-island-the-most-real-of-fictional-places

Model paranoia



- Don't "game the system" to make your results what you want, via: p-hacking, testing on your train set, stopping A/B experiments early, …
- Be suspicious of p=0.0499
- If you don't know how to interpret your results, be **very careful** in sharing your results and overclaiming

(Perneger & Combescure 2017)

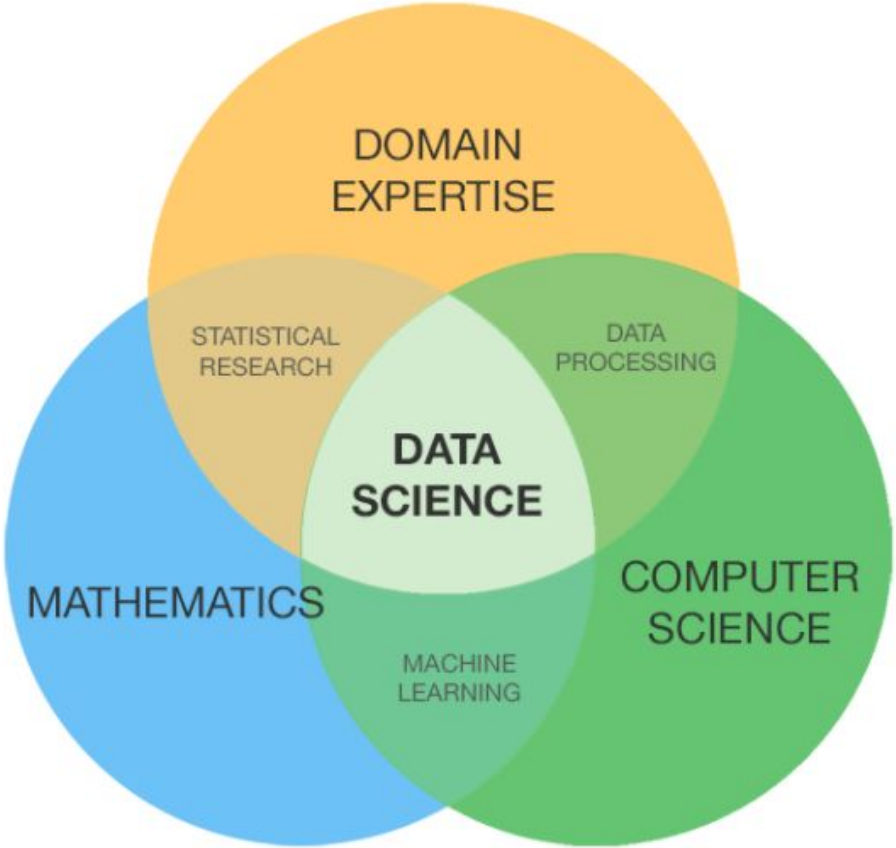
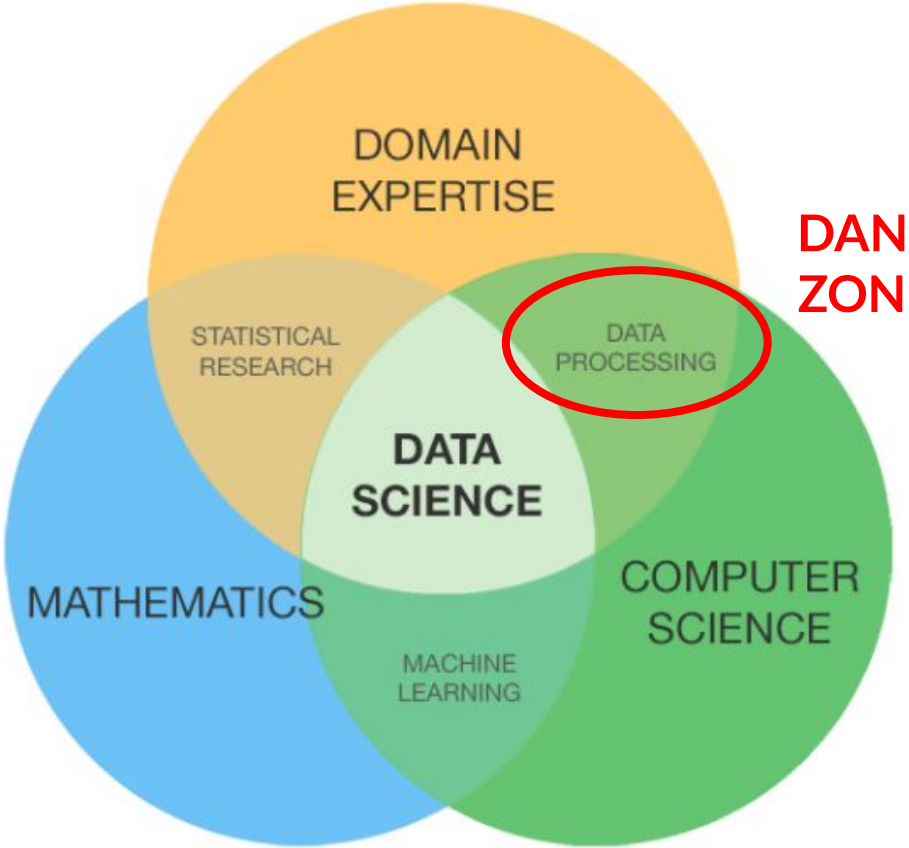
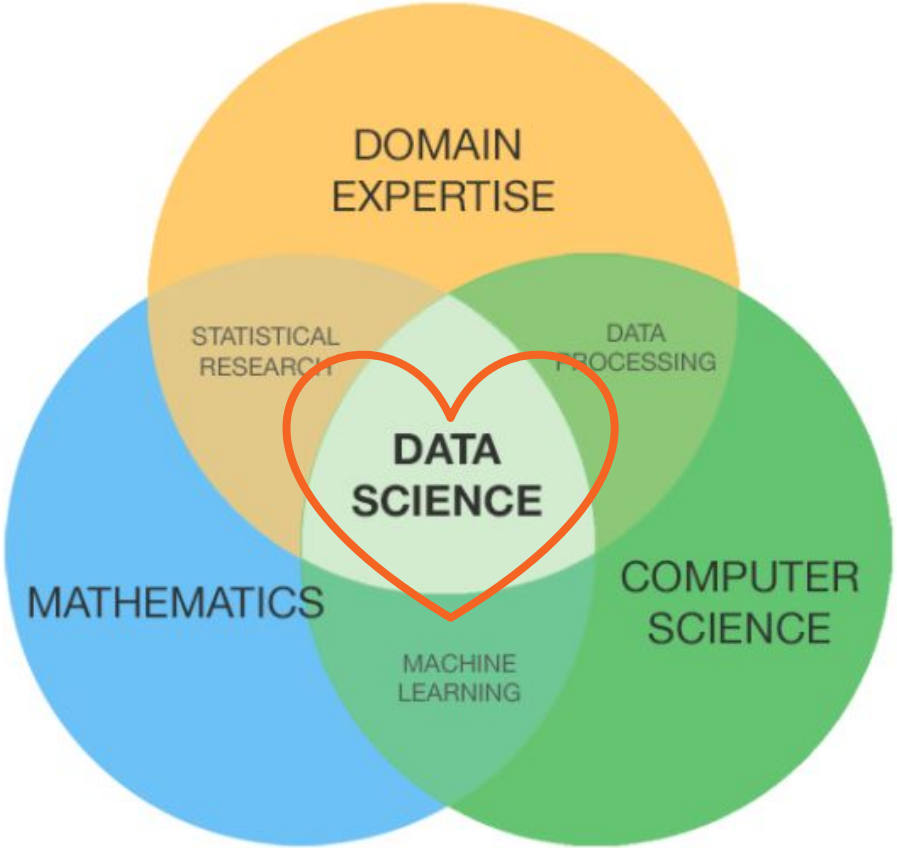


Image source: https://towardsdatascience.com/data-science-interview-guide-4ee9f5dc778



DANGER ZONE

Content source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram



After 2950!

Content source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Thank you, staff!

- **Graduate TAs**
  - Andrea Wang
  - Anna Choi
  - Rejoice Hu
  - Tangwuyou Su

- **Undergraduate TAs**
  - Bella, Sydney, Karla, Arunabh, Alexia, Chiara, Ryan, Hao, Elliot, Zack, Gaby, Samhita, Jonathan, Anya, Ethan, Annie, Ahmed, Julius, Cassandra, Kevin, Sarah, Charlie

Thank you, staff!

- **Graduate TAs**
  - Andrea Wang
  - Anna Choi
  - Rejoice Hu
  - Tangwuyou Su

Please remember to fill out your course evaluations!

- **Undergraduate TAs**
  - Bella, Sydney, Karla, Arunabh, Alexia, Chiara, Ryan, Hao, Elliot, Zack, Gaby, Samhita, Jonathan, Anya, Ethan, Annie, Ahmed, Julius, Cassandra, Kevin, Sarah, Charlie

Please keep in touch!

- We **LOVE** hearing about how you apply your 2950 skills in future classes / internships / jobs
- If you think of any topics we should add in future iterations of the class, let us know!
- koenecke@cornell.edu | ret85@cornell.edu

Thank you, students!