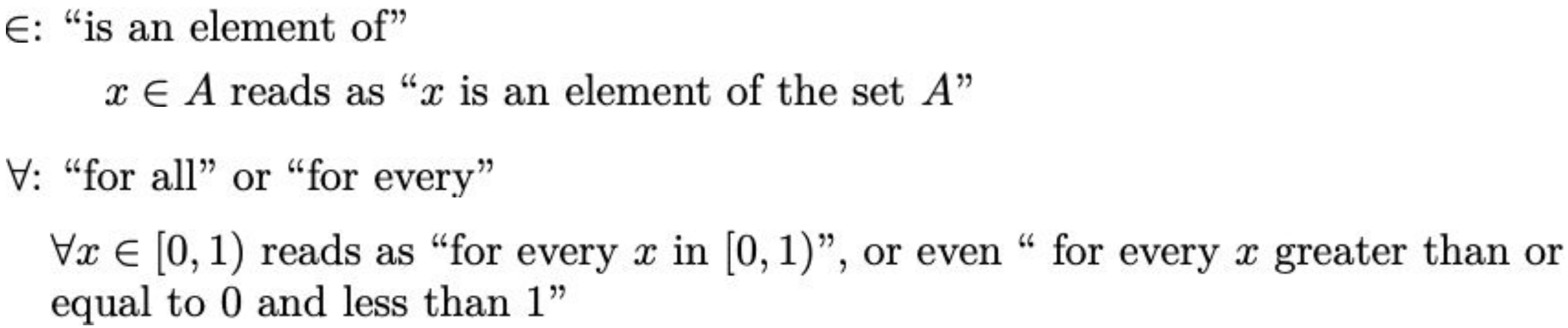
Recap (probability)

- Counting
  - Factorials
  - Binomial coefficients
- Definition of probability (experimental and theoretical)
- Definition of conditional probability

Dictionary (we’ll build on this as we go)



$\in$: “is an element of”

$x \in A$ reads as “x is an element of the set A”

$\forall$: “for all” or “for every”

$\forall x \in [0, 1)$ reads as “for every x in $[0, 1)$”, or even “ for every x greater than or equal to 0 and less than 1”

Sets

A set A is a collection of *things*, often numbers.

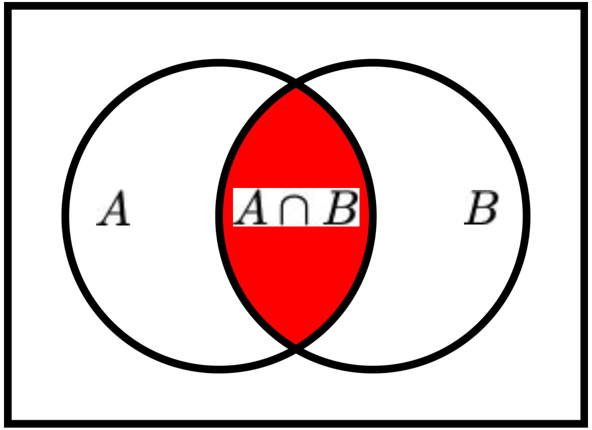
ex: $A = \{2, \pi, \text{unicorn}\}$

$|A|$: the *cardinality* of $A =$ the *size* of $A =$ the number of elements of A

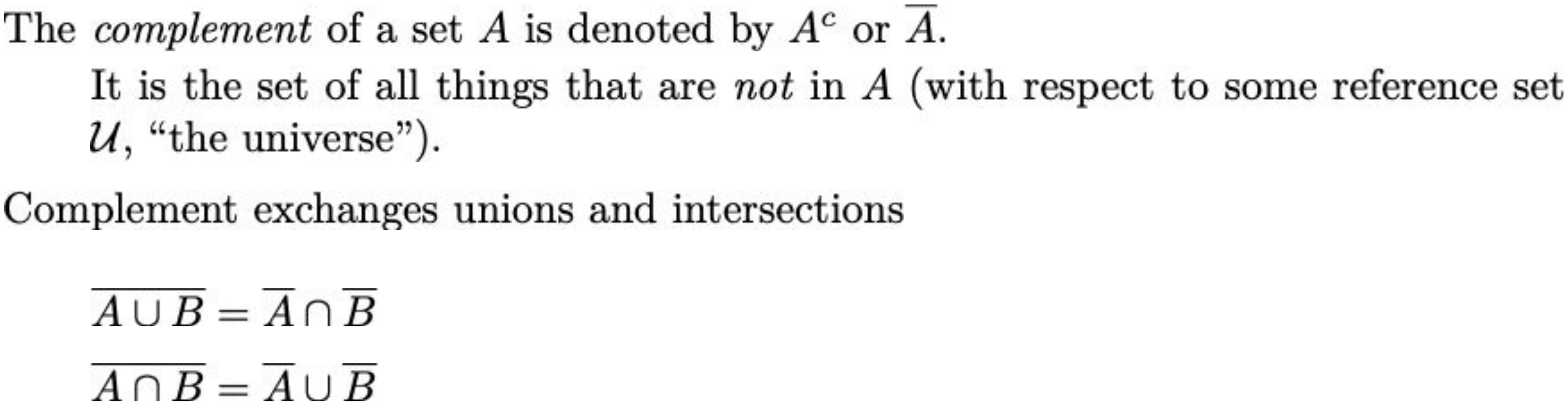
$A \cup B =$ the *union* of A and $B =$ the set of elements that are in both A and B

$A \cap B =$ the *intersection* of A and $B =$ the set of elements that are in either A or B

$|A \cup B| = |A| + |B| - |A \cap B|$



DeMorgan’s Laws



The *complement* of a set A is denoted by A^c or $\overline{A}$.

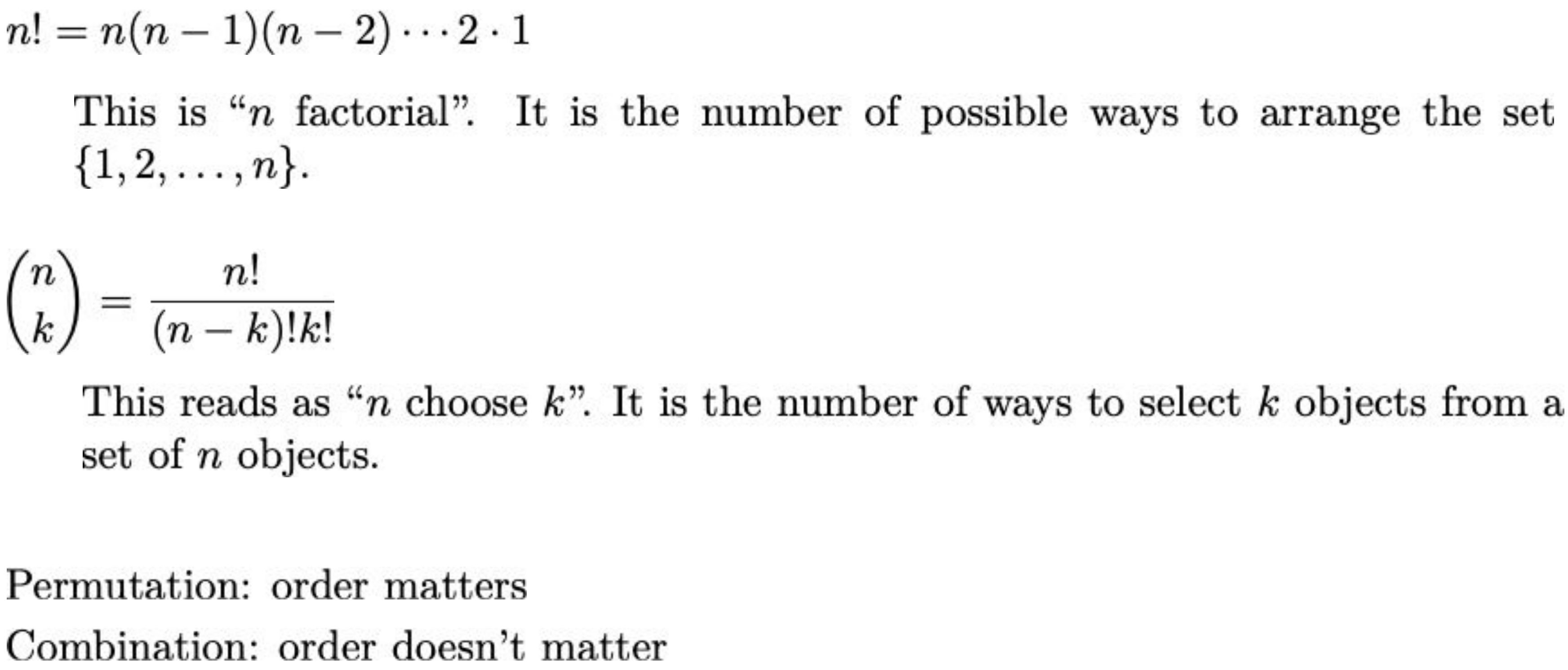
It is the set of all things that are *not* in A (with respect to some reference set $\mathcal{U}$, “the universe”).

Complement exchanges unions and intersections

$$\overline{A \cup B} = \overline{A} \cap \overline{B}$$

$$\overline{A \cap B} = \overline{A} \cup \overline{B}$$

Counting



$$n! = n(n-1)(n-2)\cdots 2 \cdot 1$$

This is "n factorial". It is the number of possible ways to arrange the set $\{1, 2, \ldots, n\}$.

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

This reads as "n choose k". It is the number of ways to select k objects from a set of n objects.

Permutation: order matters

Combination: order doesn't matter

Probability

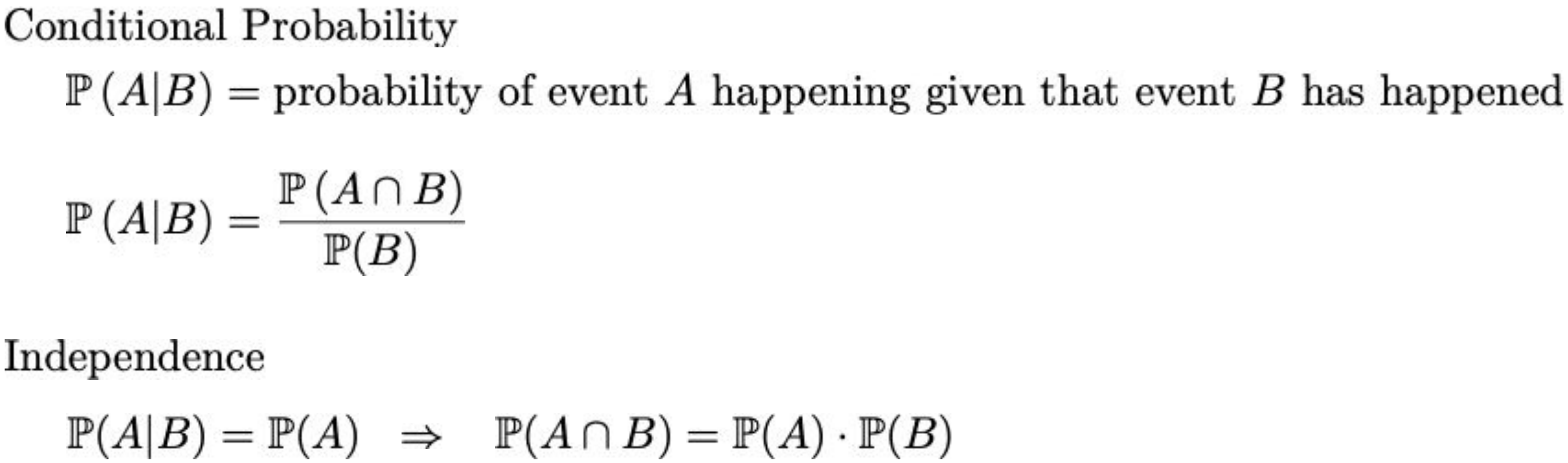
$S:$ the sample space of possible events

$\mathbb{P}(A) =$ probability that event A happens

$$\text{experimental probability:} \quad \mathbb{P}(A) = \frac{\#\text{ of times } A \text{ is observed}}{\#\text{ of observations}}$$

$$\text{theoretical probability:} \quad \mathbb{P}(A) = \frac{\#\text{ of ways } A \text{ can happen}}{\#\text{ of things that can happen}} = \frac{|A|}{|S|}$$

Probability



Conditional Probability

$\mathbb{P}\left(A|B\right) =$ probability of event A happening given that event B has happened

$$\mathbb{P}\left(A|B\right) = \frac{\mathbb{P}\left(A \cap B\right)}{\mathbb{P}(B)}$$

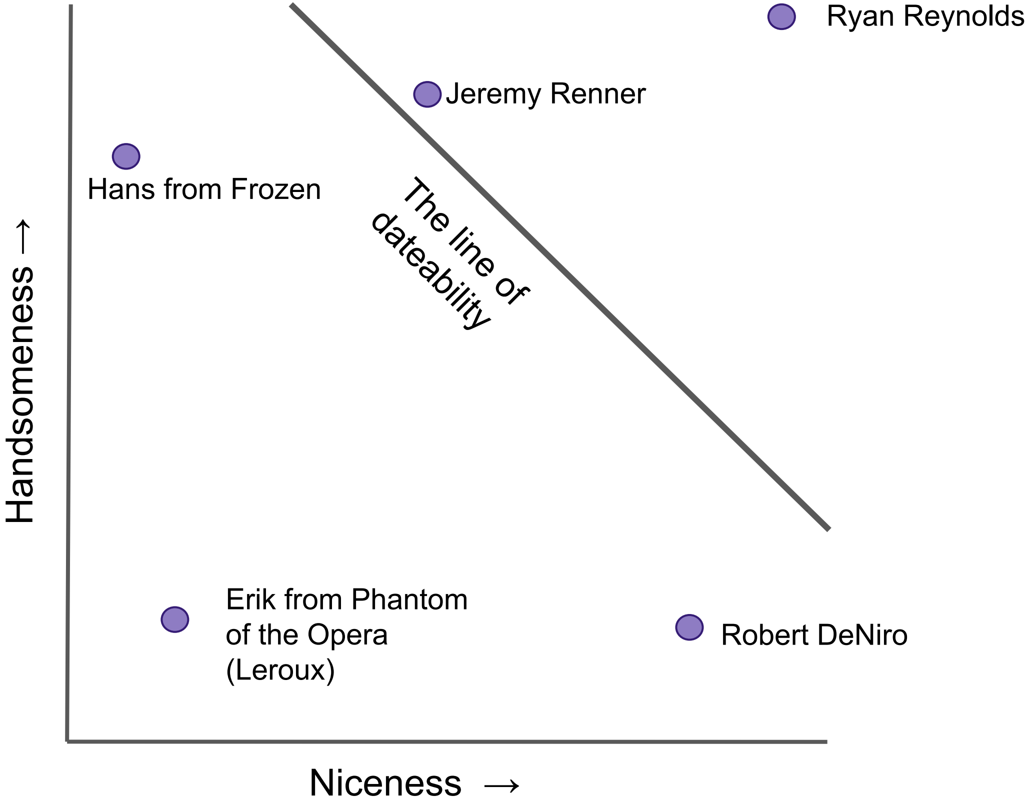
Independence

$$\mathbb{P}(A|B) = \mathbb{P}(A) \quad \Rightarrow \quad \mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

Outline (probability)

- A card trick
- The problem with intuition
  - The great square of men and Berkson’s paradox
  - Assuming independence and Sally Clark
- Bayes’ Theorem
  - Intuition and explanation
  - The prosecutor’s fallacy and the OJ Simpson trial
- Expected value
  - Definition
  - Intuition
  - Indicator variables

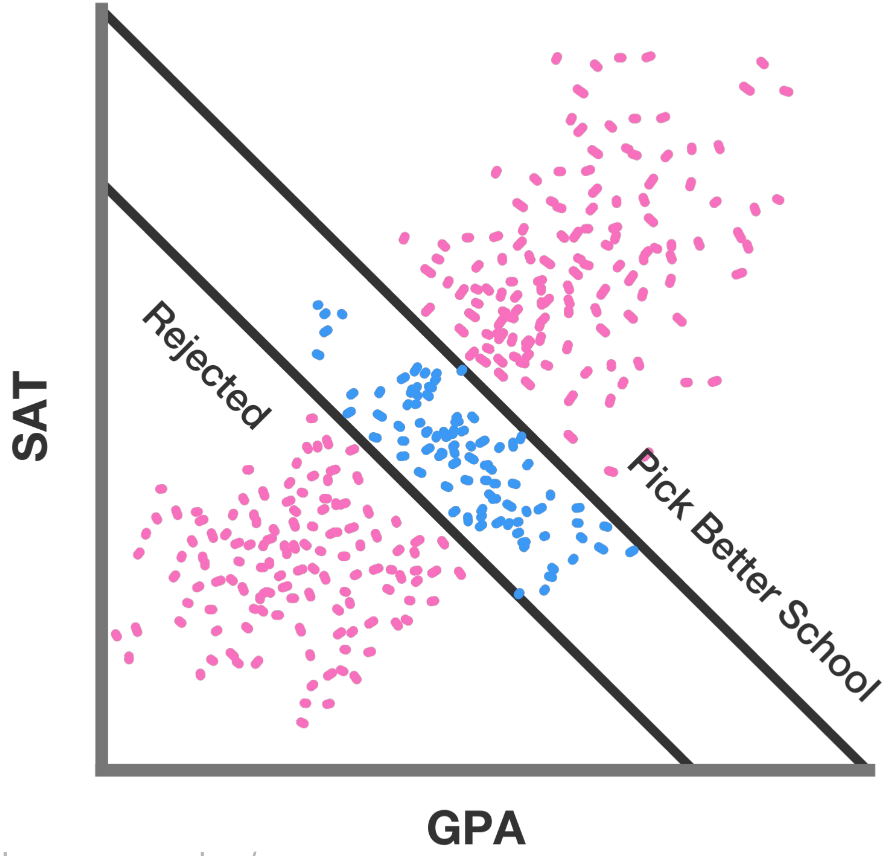
The great square of men



Stolen with love from Jordan Ellenberg:
https://slate.com/human-interest/2014/06/berksons-fallacy-why-are-handsome-men-such-jerks.html

An alternate explanation of the same example: http://corysimon.github.io/articles/berksons-paradox-are-handsome-men-really-jerks/

Berkson’s Paradox



https://brilliant.org/wiki/berksons-paradox/

Berkson’s paradox

- A false correlation induced through a sampling bias
- First documented in a hospital setting
  - Berkson, Joseph (June 1946). “Limitations of the Application of Fourfold Table Analysis to Hospital Data”. Biometrics Bulletin. 2 (3): 47–53
  - Leads to spurious statements like “Diabetes reduces cancer risk”
  - You’re in the hospital for a reason, so you are likely to have either cancer or diabetes and less likely to have both
- Spurious correlations: https://www.tylervigen.com/spurious-correlations
  - You can test and see that “amount of margarine bought in New Hampshire” correlates with the divorce rate
  - Also “number of people who died by drowning in a pool” correlates with “number of films Nicholas Cage has appeared in”

Assuming Independence

We’re going to be looking at a few examples that contain discussion of sensitive issues. If you would like to step out, now is the time.

Assuming Independence

- Sally Clark
  - Two children died from SIDS
  - A witness testified that the likelihood of this happening is 1 in 73 million
  - Sally Clark was wrongly convicted and sent to prison, later exonerated
  - https://www.theguardian.com/society/2007/mar/17/childrensservices.uknews
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC539414/

Bayes’ Theorem

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

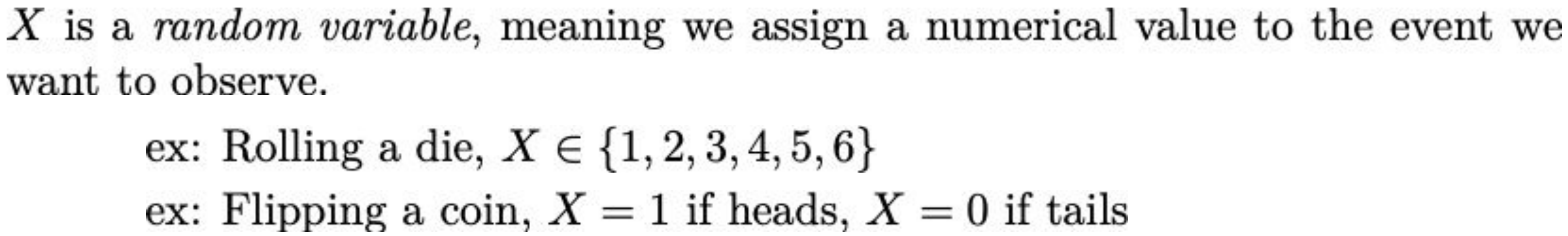
We also discussed the Monty Hall problem in the context of Bayes’ theorem. Check the lecture capture to see the discussion. If you want to see one of (a surprising number of) the papers about birds outperforming humans on the Monty Hall game, you can look here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3086893/

Julia Galef has a nice youtube video explaining Bayes’ theorem using visual thinking, which you can find here: https://www.youtube.com/watch?v=BrK7X_XlGB8

Prosecutor’s Fallacy

- Prosecutors try to confuse which conditional you want to look at
- Famous example:
  - Discussion about domestic abuse during the OJ Simpson trial
  - See a rundown by Steven Strogatz here: https://opinionator.blogs.nytimes.com/2010/04/25/chances-are/

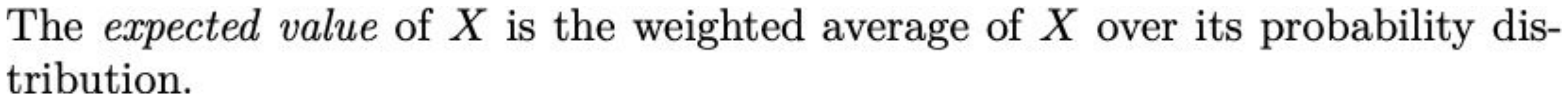
Expected Value



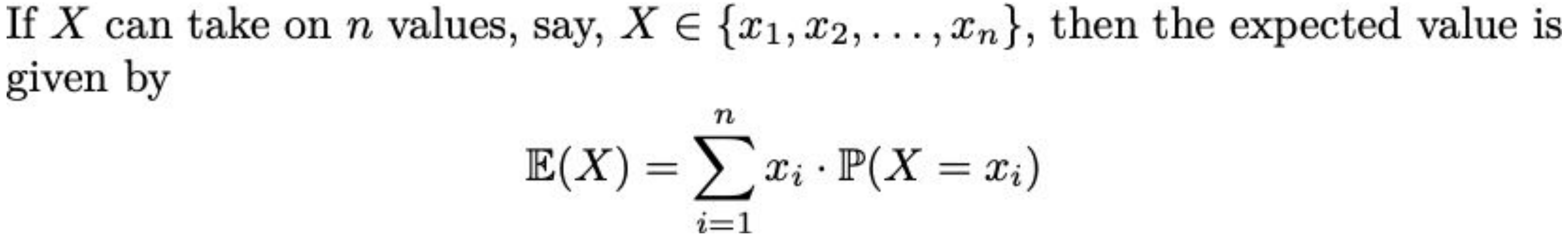
X is a *random variable*, meaning we assign a numerical value to the event we want to observe.

ex: Rolling a die, $X \in \{1, 2, 3, 4, 5, 6\}$

ex: Flipping a coin, $X = 1$ if heads, $X = 0$ if tails



The *expected value* of X is the weighted average of X over its probability distribution.



If X can take on n values, say, $X \in \{x_1, x_2, \ldots, x_n\}$, then the expected value is given by

$$\mathbb{E}(X) = \sum_{i=1}^{n} x_i \cdot \mathbb{P}(X = x_i)$$

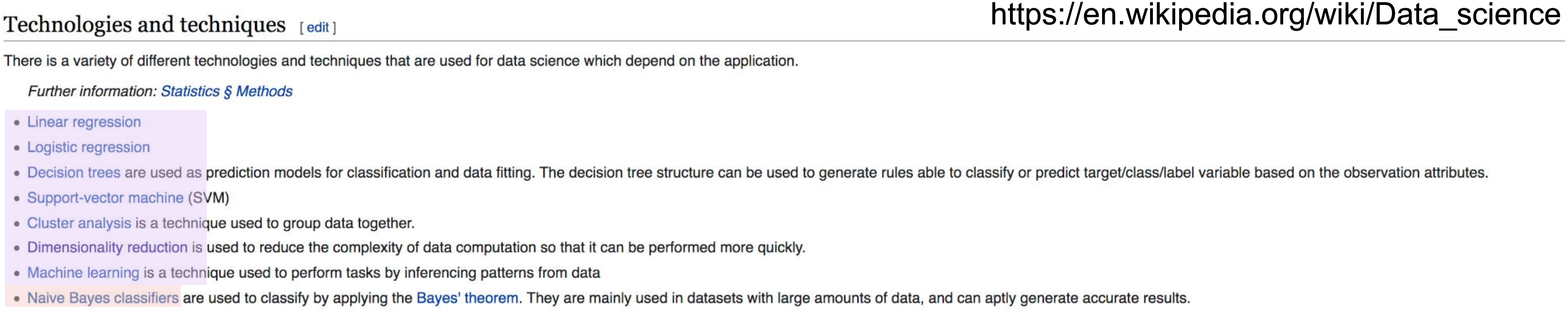
Linear Algebra

Arc of the course

Why do we need math again?

Linear Algebra | Probability | LA & Pr

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

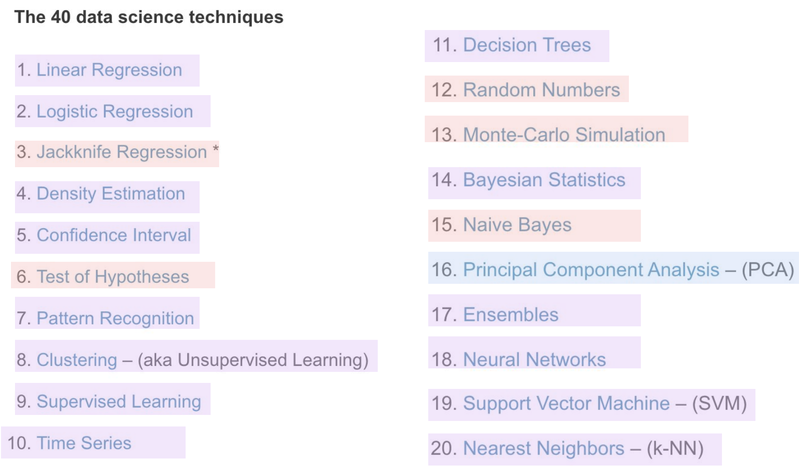


Technologies and techniques [edit]

There is a variety of different technologies and techniques that are used for data science which depend on the application.

Further information: Statistics § Methods

- Linear regression
- Logistic regression
- Decision trees are used as prediction models for classification and data fitting. The decision tree structure can be used to generate rules able to classify or predict target/class/label variable based on the observation attributes.
- Support-vector machine (SVM)
- Cluster analysis is a technique used to group data together.
- Dimensionality reduction is used to reduce the complexity of data computation so that it can be performed more quickly.
- Machine learning is a technique used to perform tasks by inferencing patterns from data
- Naive Bayes classifiers are used to classify by applying the Bayes' theorem. They are mainly used in datasets with large amounts of data, and can aptly generate accurate results.



The 40 data science techniques

1. Linear Regression
2. Logistic Regression
3. Jackknife Regression *
4. Density Estimation
5. Confidence Interval
6. Test of Hypotheses
7. Pattern Recognition
8. Clustering – (aka Unsupervised Learning)
9. Supervised Learning
10. Time Series
11. Decision Trees
12. Random Numbers
13. Monte-Carlo Simulation
14. Bayesian Statistics
15. Naive Bayes
16. Principal Component Analysis – (PCA)
17. Ensembles
18. Neural Networks
19. Support Vector Machine – (SVM)
20. Nearest Neighbors – (k-NN)

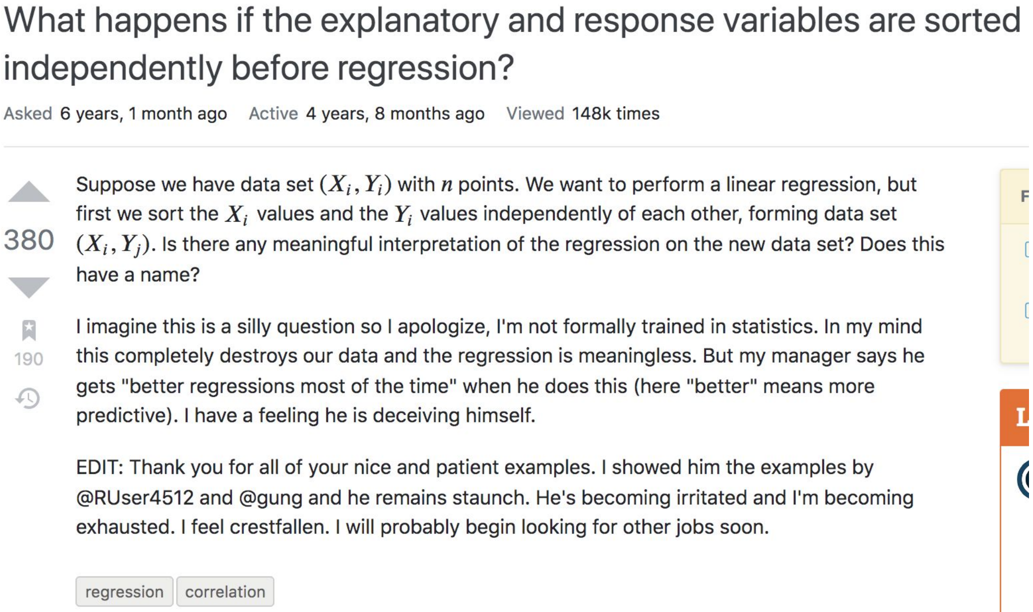
They all use probability, linear algebra, or both!

Why do we need math again?

- Examples: HIV, Ebola, cholera

Why do we need math again?

- Examples: HIV, Ebola, cholera
- Stack exchange yikes example



What happens if the explanatory and response variables are sorted independently before regression?

Asked 6 years, 1 month ago Active 4 years, 8 months ago Viewed 148k times

Suppose we have data set (X_i, Y_i) with n points. We want to perform a linear regression, but first we sort the X_i values and the Y_i values independently of each other, forming data set (X_i, Y_j). Is there any meaningful interpretation of the regression on the new data set? Does this have a name?

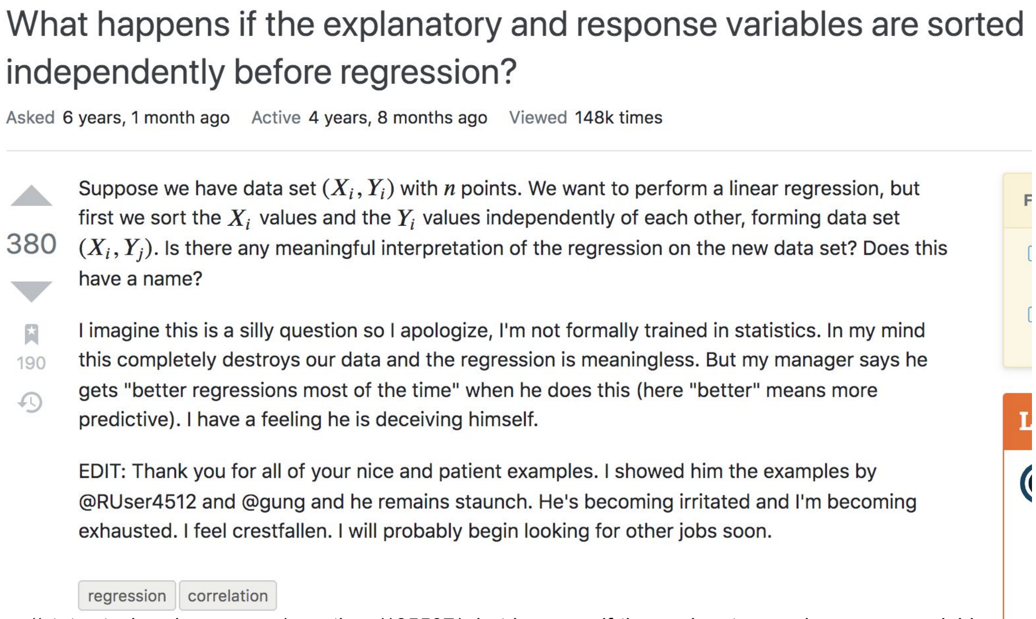
I imagine this is a silly question so I apologize, I'm not formally trained in statistics. In my mind this completely destroys our data and the regression is meaningless. But my manager says he gets "better regressions most of the time" when he does this (here "better" means more predictive). I have a feeling he is deceiving himself.

EDIT: Thank you for all of your nice and patient examples. I showed him the examples by @RUser4512 and @gung and he remains staunch. He's becoming irritated and I'm becoming exhausted. I feel crestfallen. I will probably begin looking for other jobs soon.

https://stats.stackexchange.com/questions/185507/what-happens-if-the-explanatory-and-response-variables-are-sorted-independently

Why do we need math again?

- Examples: HIV, Ebola, cholera
- Stack exchange yikes example



What happens if the explanatory and response variables are sorted independently before regression?

Asked 6 years, 1 month ago Active 4 years, 8 months ago Viewed 148k times

380

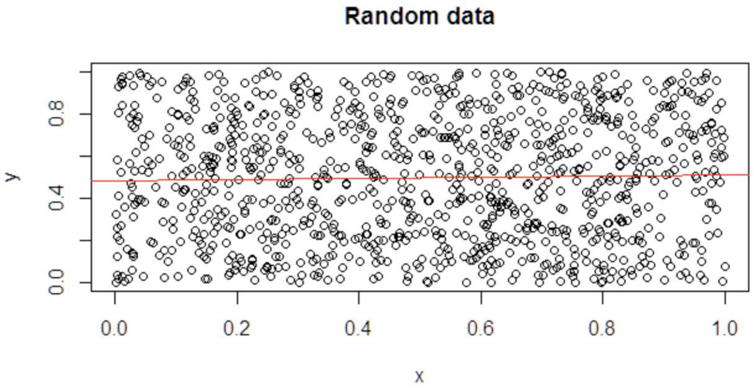
190

Suppose we have data set (X_i, Y_i) with n points. We want to perform a linear regression, but first we sort the X_i values and the Y_i values independently of each other, forming data set (X_i, Y_j). Is there any meaningful interpretation of the regression on the new data set? Does this have a name?

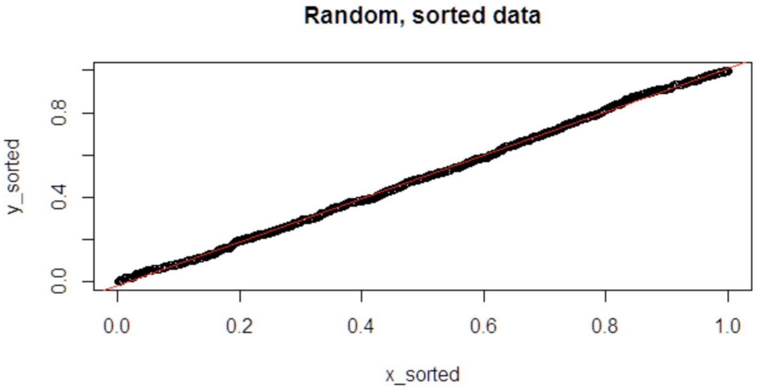
I imagine this is a silly question so I apologize, I'm not formally trained in statistics. In my mind this completely destroys our data and the regression is meaningless. But my manager says he gets "better regressions most of the time" when he does this (here "better" means more predictive). I have a feeling he is deceiving himself.

EDIT: Thank you for all of your nice and patient examples. I showed him the examples by @RUser4512 and @gung and he remains staunch. He's becoming irritated and I'm becoming exhausted. I feel crestfallen. I will probably begin looking for other jobs soon.

regression correlation



Random data



Random, sorted data

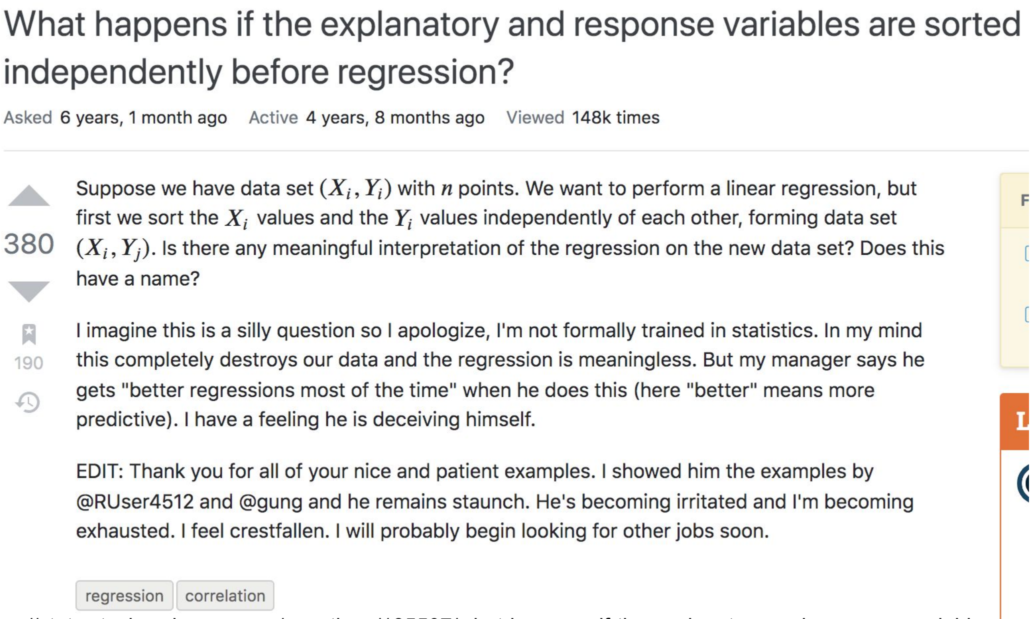
https://stats.stackexchange.com/questions/185507/what-happens-if-the-explanatory-and-response-variables-are-sorted-independently

Why do we need math again?

- Examples: HIV, Ebola, cholera
- Stack exchange yikes example

X	Y
1	4
2	2
4	3
3	1

X	Y
1	1
2	2
3	3
4	4



What happens if the explanatory and response variables are sorted independently before regression?

Asked 6 years, 1 month ago Active 4 years, 8 months ago Viewed 148k times

380

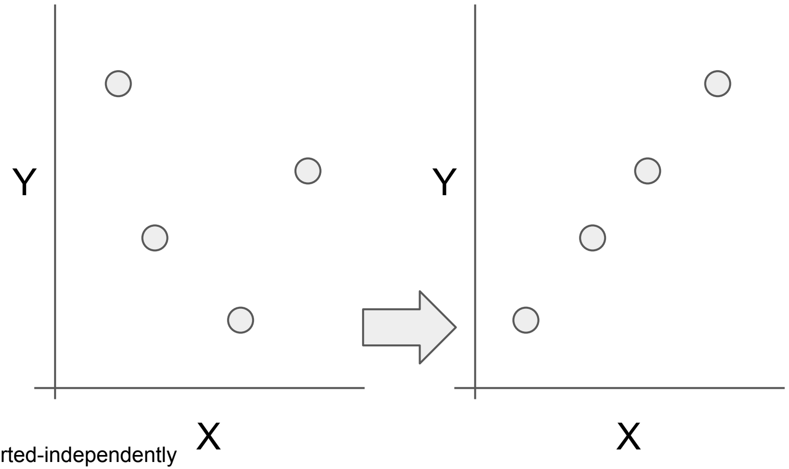
190

Suppose we have data set (X_i, Y_i) with n points. We want to perform a linear regression, but first we sort the X_i values and the Y_i values independently of each other, forming data set (X_i, Y_j). Is there any meaningful interpretation of the regression on the new data set? Does this have a name?

I imagine this is a silly question so I apologize, I'm not formally trained in statistics. In my mind this completely destroys our data and the regression is meaningless. But my manager says he gets "better regressions most of the time" when he does this (here "better" means more predictive). I have a feeling he is deceiving himself.

EDIT: Thank you for all of your nice and patient examples. I showed him the examples by @RUser4512 and @gung and he remains staunch. He's becoming irritated and I'm becoming exhausted. I feel crestfallen. I will probably begin looking for other jobs soon.

regression correlation



https://stats.stackexchange.com/questions/185507/what-happens-if-the-explanatory-and-response-variables-are-sorted-independently

Recap

Review matrices as a transformation

Columns as readout of where it sends the standard basis

Visualizing - what kinds of things can matrices do

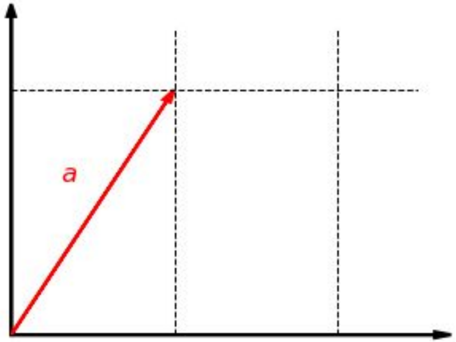
Multiplication as composition

Linearity

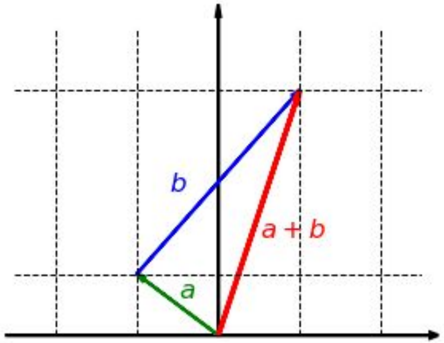
Define linearity

Matrices are linear transformations - they preserve lines/planes/etc

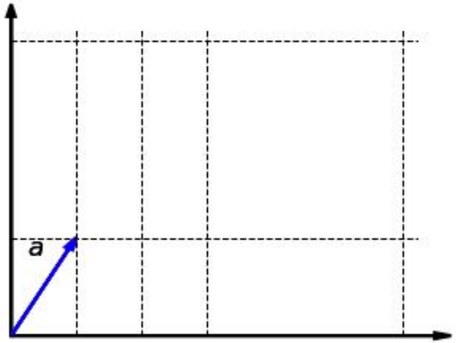
Example: f(x,y) = (2x, y), i.e. the matrix $\begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$



What does f do?



How does f affect vector addition?



How does f affect scalar multiplication?

Inverse of a matrix

- If a matrix is a transformation, what about undo-ing that transformation?
    - E.g. rotations, shears, etc
- This is also a matrix!
- We write it A^{-1}
- $A A^{-1}$ = the identity matrix = I
- $(A^{-1})^{-1} = A$

Matrix inverses

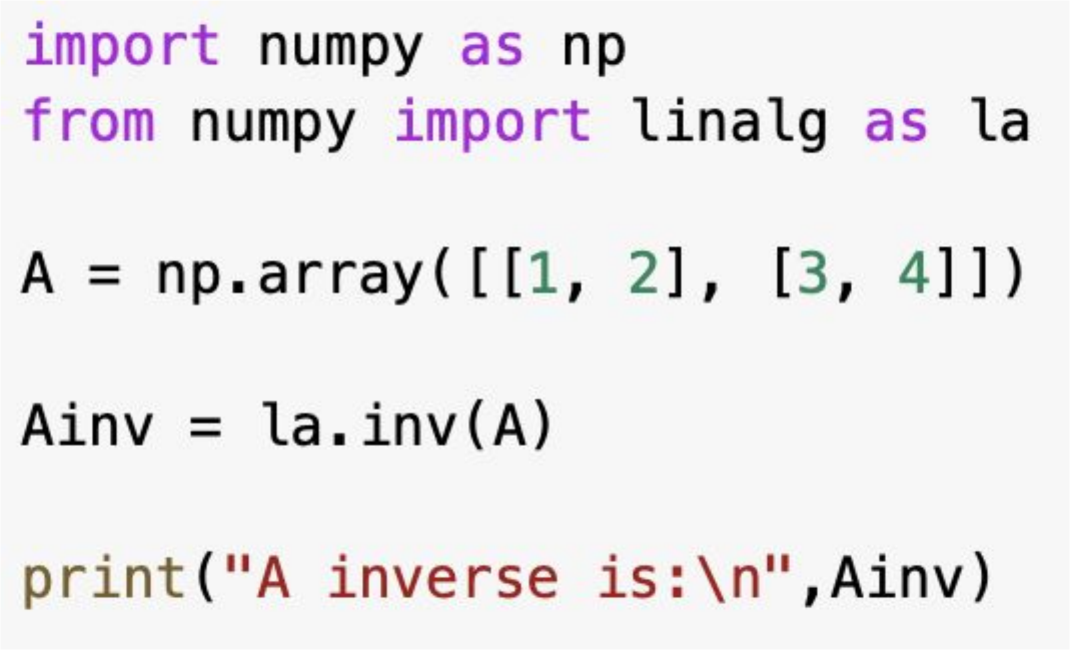
How to calculate the inverse of a matrix? For a 2x2:

$$\mathbf{A}^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

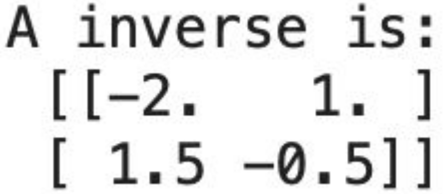
(We haven’t introduced determinants yet, but for those of you have seen it)

Matrix Inverses

What about bigger matrices? There are methods (and often you can figure it out from the matrix structure), but we can also use code:



```
import numpy as np
from numpy import linalg as la

A = np.array([[1, 2], [3, 4]])

Ainv = la.inv(A)

print("A inverse is:\n",Ainv)
```



```
A inverse is:
 [[-2.   1. ]
 [ 1.5 -0.5]]
```

Does every matrix have an inverse?

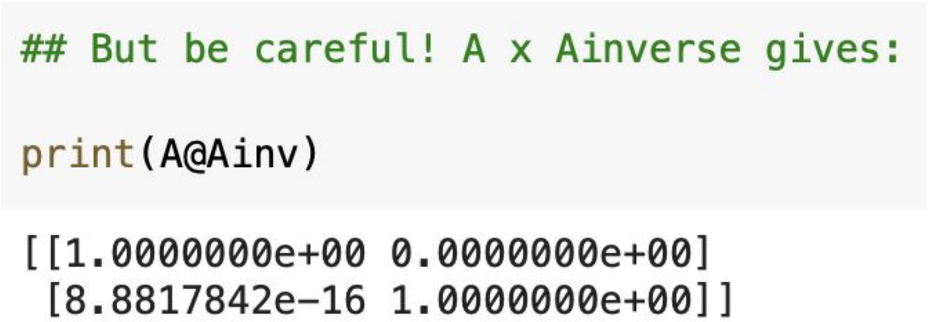
Discuss—is every transformation undo-able?

Examples

- Matrices that aren’t invertible—what goes wrong? What happens to the calculation if we try?

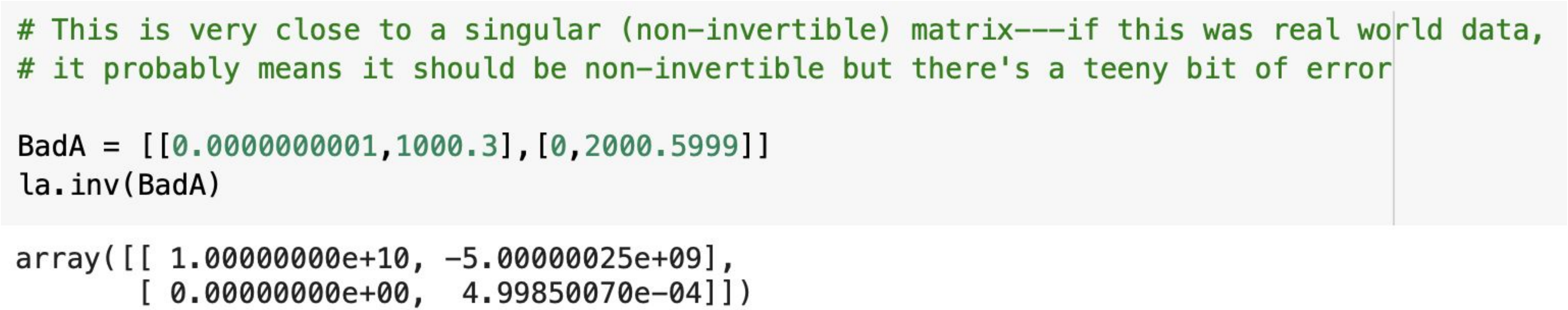
$$\mathbf{A}^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

But be careful…



```
## But be careful! A x Ainverse gives:

print(A@Ainv)
```

```
[[1.0000000e+00 0.0000000e+00]
 [8.8817842e-16 1.0000000e+00]]
```



```
# This is very close to a singular (non-invertible) matrix---if this was real world data,
# it probably means it should be non-invertible but there's a teeny bit of error

BadA = [[0.0000000001,1000.3],[0,2000.5999]]
la.inv(BadA)
```

```
array([[ 1.00000000e+10, -5.00000025e+09],
       [ 0.00000000e+00,  4.99850070e-04]])
```

Eigenvalues and eigenvectors

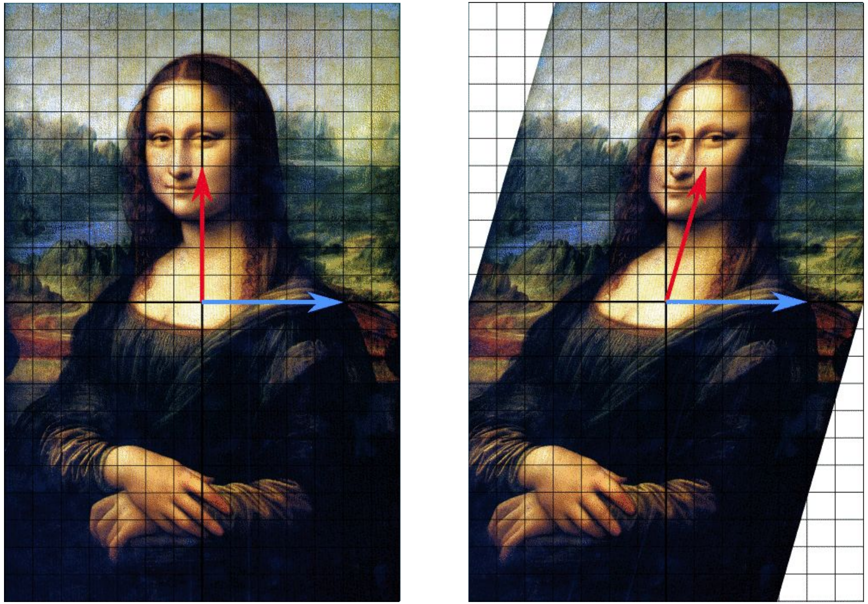
- How do we get a sense of what a matrix is doing?
- One of the most important concepts in linear algebra is eigenvalues/eigenvectors

- These are the basis of so many techniques: PCA, tons of network analysis techniques including Google pagerank, and even how we calculate the basic reproduction number (R0) for infectious diseases!

Eigenvalues and eigenvectors

We call a vector $\vec{v}$ an eigenvector of a matrix A, with *eigenvalue* λ if

$$A\vec{v} = \lambda\vec{v}$$



Eigenvalues and eigenvectors

For example:

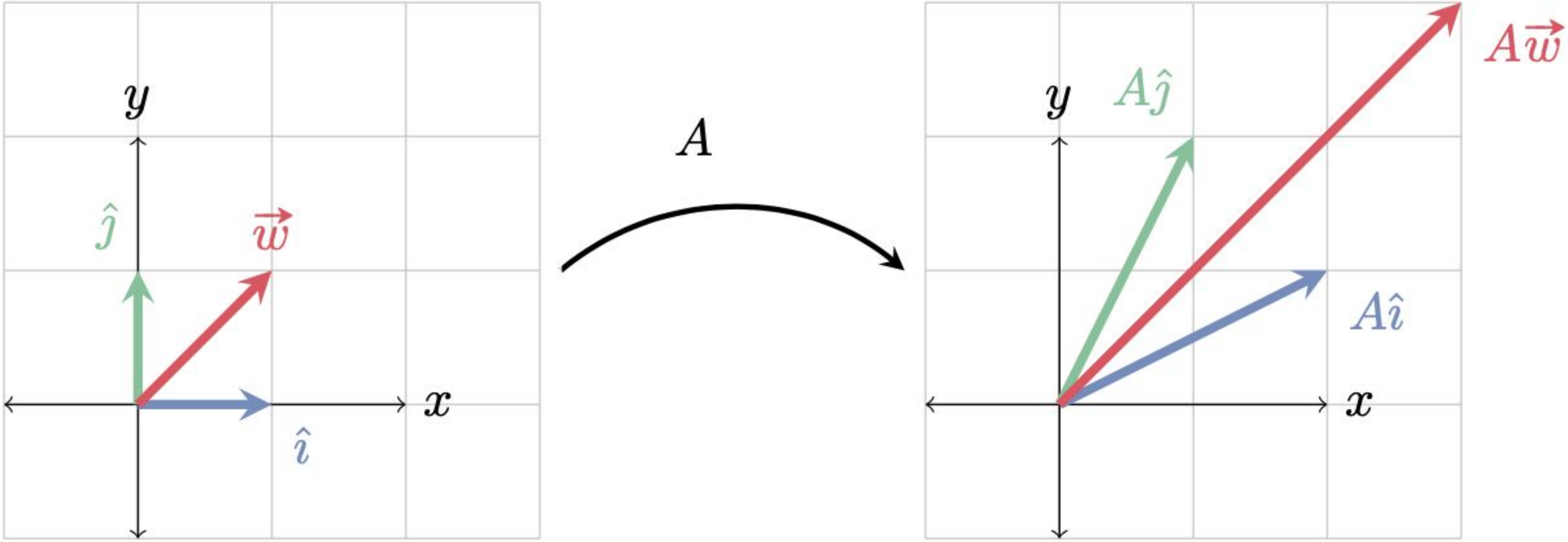
$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \qquad v_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad v_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

$$Av_1 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \end{pmatrix} = 3 \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 3 \cdot v_1$$

$$Av_2 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix} = 1 \cdot v_2$$

Eigenvalues and eigenvectors

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$



Other examples

- Rotation matrix

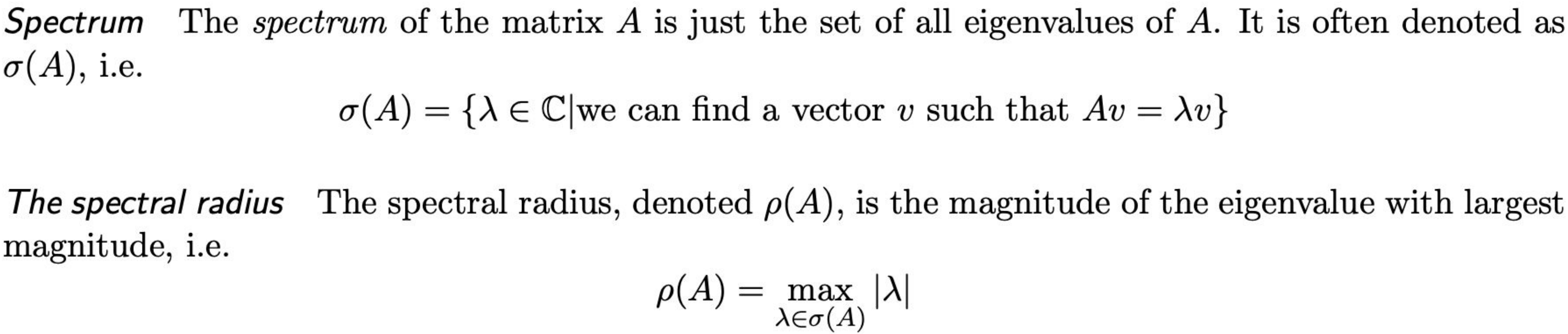
$$A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

- Only one eigenvalue/eigenvector

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

$$Av = v \quad \Rightarrow \quad \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} \quad \Rightarrow \quad \begin{aligned} x &= x \\ x + y &= y \end{aligned} \quad \Rightarrow \quad \begin{aligned} x &= 0 \\ y &= 1 \end{aligned} \quad \Rightarrow \quad v = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Spectrum & spectral radius



Spectrum The *spectrum* of the matrix A is just the set of all eigenvalues of A. It is often denoted as $\sigma(A)$, i.e.

$$\sigma(A) = \{\lambda \in \mathbb{C} | \text{we can find a vector } v \text{ such that } Av = \lambda v\}$$

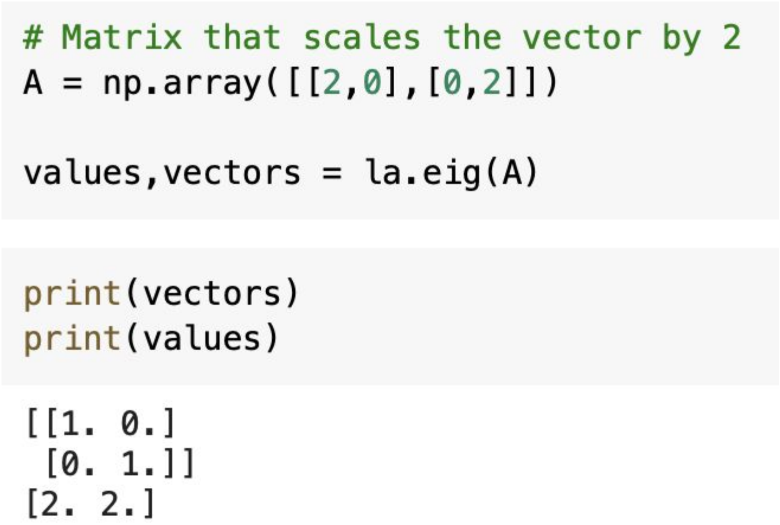
The spectral radius The spectral radius, denoted $\rho(A)$, is the magnitude of the eigenvalue with largest magnitude, i.e.

$$\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda|$$

How do we calculate the eigenvalues and eigenvectors of a matrix?

More on this next time when we do determinants!

We can also do with python:



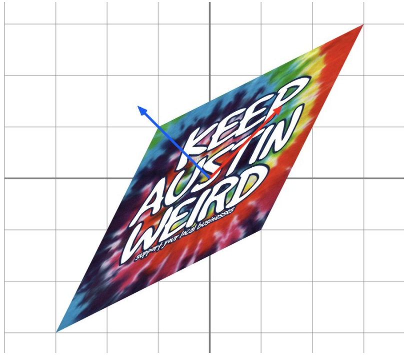
```
# Matrix that scales the vector by 2
A = np.array([[2,0],[0,2]])

values,vectors = la.eig(A)
```

```
print(vectors)
print(values)
```

```
[[1. 0.]
 [0. 1.]]
[2. 2.]
```

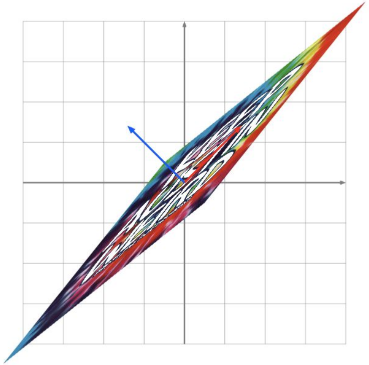
Why are eigenvalues and eigenvectors a good marker of how a matrix transforms data/space?

$$A = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



Eigenvalues: 1.5, 0.5.

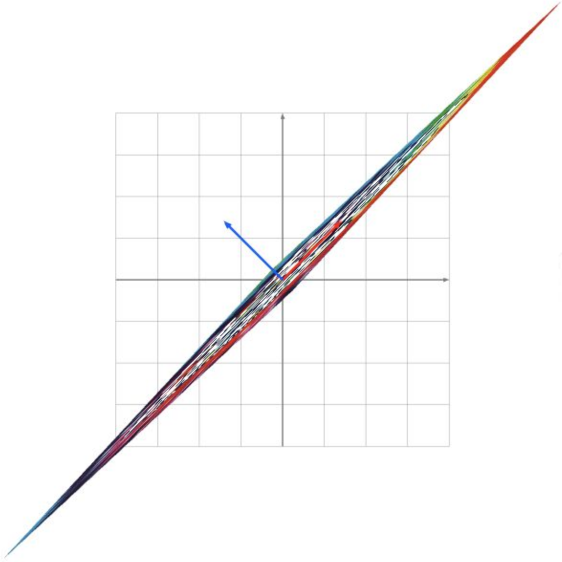
$$A^2 = \begin{bmatrix} 1.2 & 1 \\ 1 & 1.2 \end{bmatrix}$$



Eigenvalues: 2.25, 0.25.

As you apply the matrix repeatedly, it will stretch the data along the eigenvector with the largest eigenvalue

$$A^3 = \begin{bmatrix} 1.7 & 1.6 \\ 1.6 & 1.7 \end{bmatrix}$$



Eigenvalues: $3.375, 0.125$.

Vector Spaces

Addition, scalar multiplication

Examples - include some weird ones

Fourier decomposition

What is a vector space?

A vector space is a set of vectors that can be added and can be multiplied by constants called scalars.

$$\{(a,b) \mid a,b \in \mathbb{R}\}$$

and $c = \mathbb{R}$

(Vector spaces are what we’ve been working with this whole time!)

How do matrices fit in to this?

What is a vector space?

Matrices move vectors around the vector space, or even take vectors from one vector space to another!