Finding Possibly Related Entities

Elon Musk's Tesla Powerwalls Have Landed in Puerto Rico

The solar batteries have reportedly been spotted in San Juan's airport.

By John Patrick Pullen October 16, 2017

Exactly one week after Tesla CEO Elon Musk suggested his company could help with Puerto Rico's electricity crisis in the aftermath of Hurricane Maria, more of the company's Powerwall battery packs have arrived on the island, according to a photo snapped at San Juan airport Friday, Oct. 13.

Source: http://fortune.com/2017/10/16/elon-musks-tesla-powerwalls-have-landed-in-puerto-rico/

Elon Musk's Tesla Powerwalls Have Landed in Puerto Rico

How to automatically figure out Elon Musk and Tesla are related?

The solar batteries have reportedly been spotted in San Juan's airport.

By John Patrick Pullen October 16, 2017

Exactly one week after Tesla CEO Elon Musk suggested his company could help with Puerto Rico's electricity crisis in the aftermath of Hurricane Maria, more of the company's Powerwall battery packs have arrived on the island, according to a photo snapped at San Juan airport Friday, Oct. 13.

Source: http://fortune.com/2017/10/16/elon-musks-tesla-powerwalls-have-landed-in-puerto-rico/

Elon Musk's Tesla Powerwalls Have Landed in Puerto Rico

How to automatically figure out Elon Musk and Tesla are related?

The solar batteries have reportedly been spotted in San Juan's airport.

By John Patrick Pullen October 16, 2017

Exactly one week after Tesla CEO Elon Musk suggested his company could help with Puerto Rico's electricity crisis in the aftermath of Hurricane Maria, more of the company's Powerwall battery packs have arrived on the island, according to a photo snapped at San Juan airport Friday, Oct. 13.

Source: http://fortune.com/2017/10/16/elon-musks-tesla-powerwalls-have-landed-in-puerto-rico/

Co-Occurrences

Co-Occurrences

For example: count # news articles that have different named entities co-occur

Co-Occurrences

For example: count # news articles that have different named entities co-occur

	Elon Musk	Tesla	Apple	Tim Cook
Elon Musk		300	1	4
Tesla	300		5	1
Apple	1	5		195
Tim Cook	4	1	195	

Co-Occurrences

For example: count # news articles that have different named entities co-occur

	Elon Musk	Tesla	Apple	Tim Cook
Elon Musk		300	1	4
Tesla	300		5	1
Apple	1	5		195
Tim Cook	4	1	195	

What does it mean for a named entity to co-occur with itself?

Co-Occurrences

For example: count # news articles that have different named entities co-occur

	Elon Musk	Tesla	Apple	Tim Cook
Elon Musk		300	1	4
Tesla	300		5	1
Apple	1	5		195
Tim Cook	4	1	195	

What does it mean for a named entity to co-occur with itself?

Example: could count # articles in which word appears ≥ 2 times

Co-Occurrences

For example: count # news articles that have different named entities co-occur

	Elon Musk	Tesla	Apple	Tim Cook
Elon Musk		300	1	4
Tesla	300		5	1
Apple	1	5		195
Tim Cook	4	1	195	

Large values => possible related items

What does it mean for a named entity to co-occur with itself?

Example: could count # articles in which word appears ≥ 2 times

Different Ways to Count

Different Ways to Count

- Just saw: for all doc's, count # of doc's in which two named entities co-occur

Different Ways to Count

- Just saw: for all doc's, count # of doc's in which two named entities co-occur
    - This approach ignores # of co-occurrences *within a specific document* (e.g., if 1 doc has "Elon Musk" and "Tesla" appear 10 times, we count this as 1)

Different Ways to Count

- Just saw: for all doc's, count # of doc's in which two named entities co-occur
  - This approach ignores # of co-occurrences *within a specific document* (e.g., if 1 doc has "Elon Musk" and "Tesla" appear 10 times, we count this as 1)
  - Could instead add # co-occurrences, not just whether it happened in a doc

Different Ways to Count

- Just saw: for all doc's, count # of doc's in which two named entities co-occur
  - This approach ignores # of co-occurrences *within a specific document* (e.g., if 1 doc has "Elon Musk" and "Tesla" appear 10 times, we count this as 1)
  - Could instead add # co-occurrences, not just whether it happened in a doc
- Instead of looking at # doc's, look at co-occurrences within a *sentence*, or a *paragraph*, etc

Different Ways to Count

- Just saw: for all doc's, count # of doc's in which two named entities co-occur
  - This approach ignores # of co-occurrences *within a specific document* (e.g., if 1 doc has "Elon Musk" and "Tesla" appear 10 times, we count this as 1)
  - Could instead add # co-occurrences, not just whether it happened in a doc
- Instead of looking at # doc's, look at co-occurrences within a *sentence*, or a *paragraph*, etc

Bottom Line

Different Ways to Count

- Just saw: for all doc's, count # of doc's in which two named entities co-occur
  - This approach ignores # of co-occurrences *within a specific document* (e.g., if 1 doc has "Elon Musk" and "Tesla" appear 10 times, we count this as 1)
  - Could instead add # co-occurrences, not just whether it happened in a doc
- Instead of looking at # doc's, look at co-occurrences within a *sentence*, or a *paragraph*, etc

Bottom Line

- There are many ways to count co-occurrences

Different Ways to Count

- Just saw: for all doc's, count # of doc's in which two named entities co-occur
  - This approach ignores # of co-occurrences *within a specific document* (e.g., if 1 doc has "Elon Musk" and "Tesla" appear 10 times, we count this as 1)
  - Could instead add # co-occurrences, not just whether it happened in a doc
- Instead of looking at # doc's, look at co-occurrences within a *sentence*, or a *paragraph*, etc

Bottom Line

- There are many ways to count co-occurrences
- You should think about what makes the most sense/is reasonable for the problem you're looking at

We aim to find *interesting* relationships by looking at co-occurrences

Image source: http://www.awf.org/sites/default/files/media/gallery/wildlife/Plains%20Zebra/Z-Billy_Dodson_3.jpg?itok=rzMdZ7LM

Black and white frequently co-occur, but is this relationship interesting?

Image source: http://www.awf.org/sites/default/files/media/gallery/wildlife/Plains%20Zebra/Z-Billy_Dodson_3.jpg?itok=rzMdZ7LM

Black and white frequently co-occur, but is this relationship interesting?

	Green	White	Black
Green	1000	200	200
White	200	2000	350
Black	200	350	2000

How I'm counting: For each pixel, look at neighboring 4 pixels and compare their values (1 of "green green", "green white", "green black", "white white", "white black", "black black")

Image source: http://www.awf.org/sites/default/files/media/gallery/wildlife/Plains%20Zebra/Z-Billy_Dodson_3.jpg?itok=rzMdZ7LM

	Green	White	Black
Green	1000	200	200
White	200	2000	350
Black	200	350	2000

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

1000 of these cards: Green, Green

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

1000 of these cards: Green, Green

200 of these cards: Green, White

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

1000 of these cards: Green, Green

200 of these cards: Green, White

200 of these cards: Green, Black

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

1000 of these cards: Green, Green

200 of these cards: Green, White

200 of these cards: Green, Black

2000 of these cards: White, White

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

1000 of these cards: Green, Green

200 of these cards: Green, White

200 of these cards: Green, Black

2000 of these cards: White, White

350 of these cards: White, Black

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

1000 of these cards: Green, Green

200 of these cards: Green, White

200 of these cards: Green, Black

2000 of these cards: White, White

350 of these cards: White, Black

2000 of these cards: Black, Black

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

1000 of these cards: Green, Green

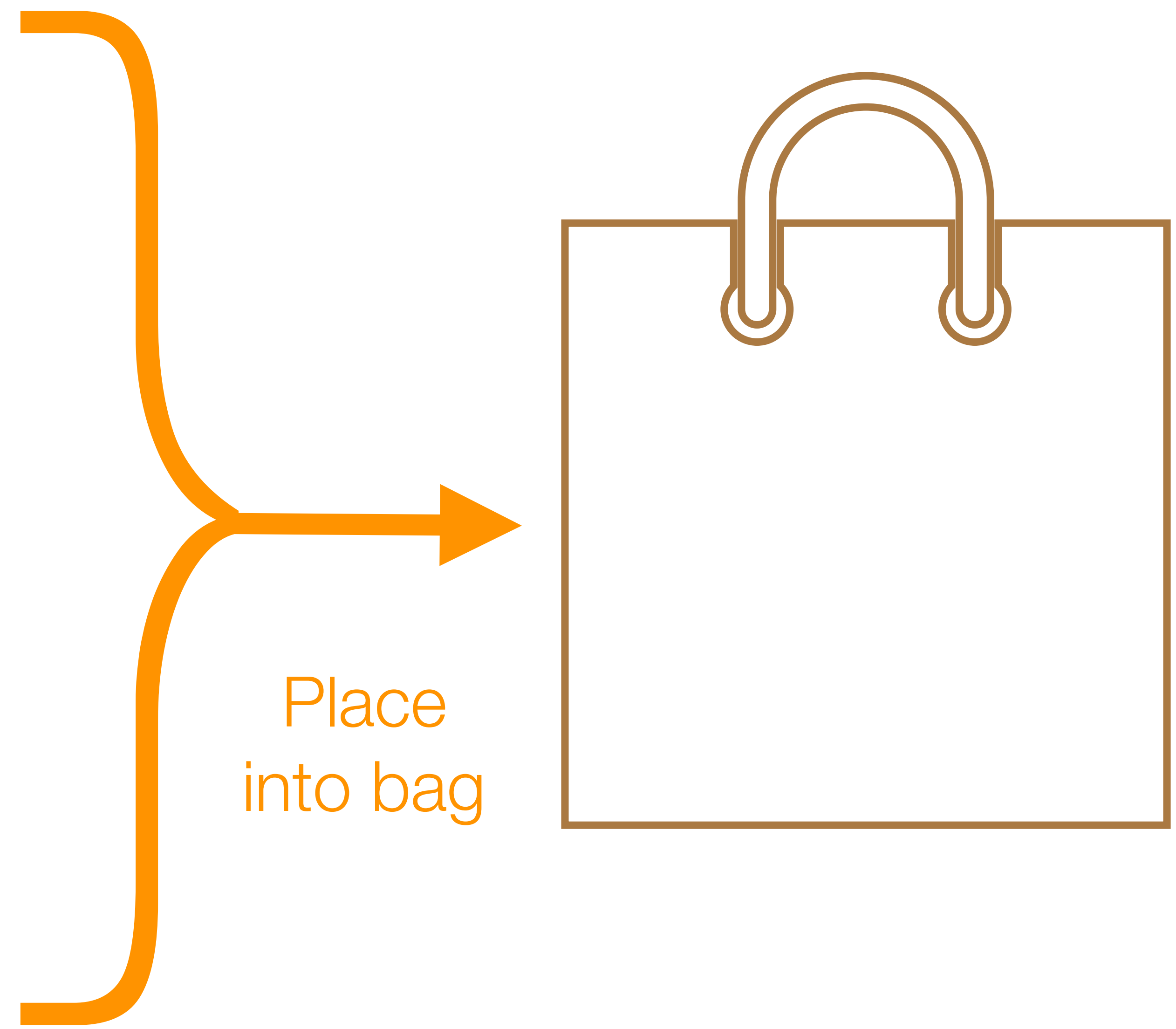
200 of these cards: Green, White

200 of these cards: Green, Black

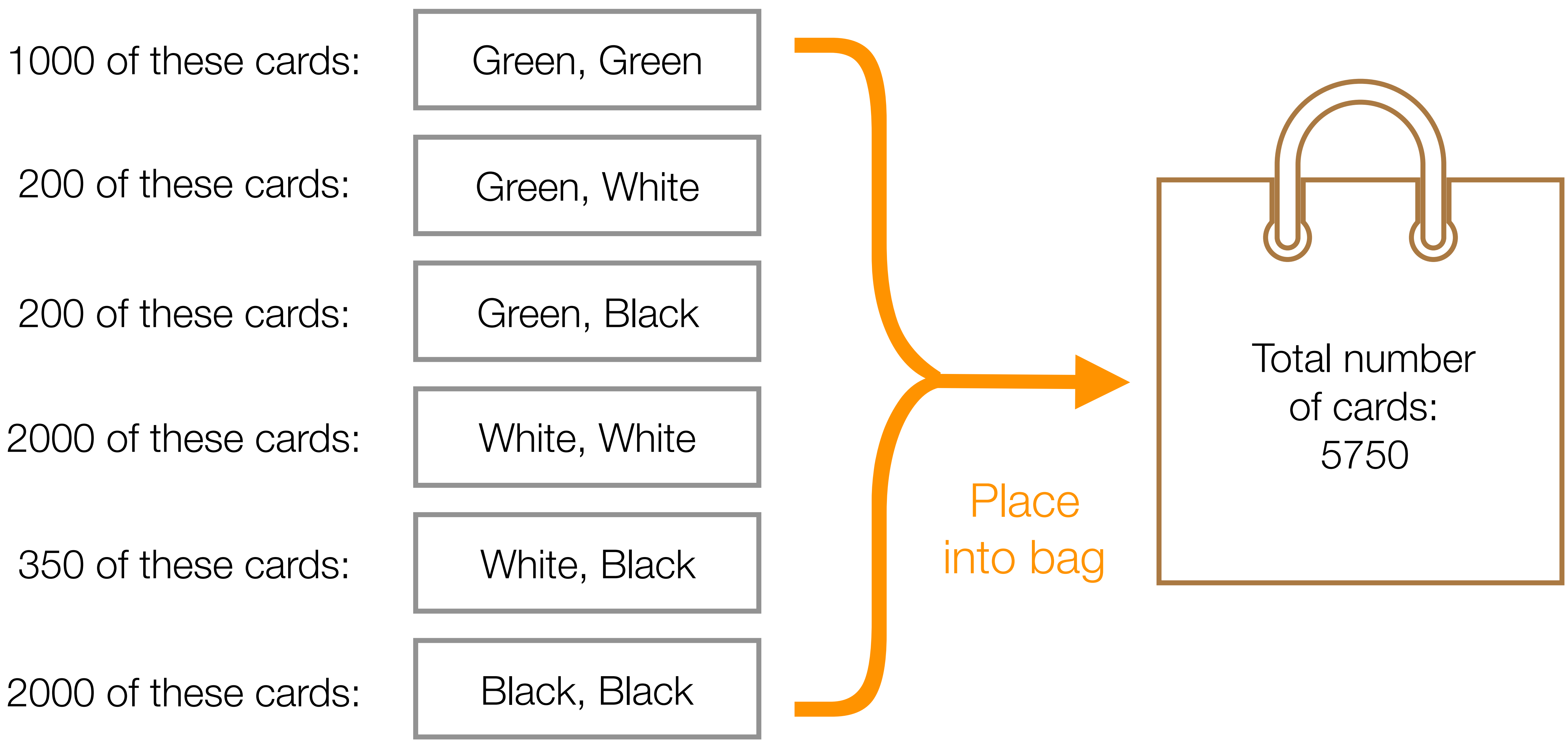
2000 of these cards: White, White

350 of these cards: White, Black

2000 of these cards: Black, Black



	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000



Probability of drawing "White, Black"?

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

1000 of these cards: Green, Green

200 of these cards: Green, White

200 of these cards: Green, Black

2000 of these cards: White, White

350 of these cards: White, Black

2000 of these cards: Black, Black

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

Probability of drawing "White, Black"?

350/5750

1000 of these cards: Green, Green

200 of these cards: Green, White

200 of these cards: Green, Black

2000 of these cards: White, White

350 of these cards: White, Black

2000 of these cards: Black, Black

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

Probability of drawing "White, Black"?

350/5750

Probability of drawing a card that has "White" on it?

1000 of these cards: Green, Green

200 of these cards: Green, White

200 of these cards: Green, Black

2000 of these cards: White, White

350 of these cards: White, Black

2000 of these cards: Black, Black

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

Probability of drawing "White, Black"?

350/5750

Probability of drawing a card that has "White" on it?

(200+2000+350)/5750

1000 of these cards: Green, Green

200 of these cards: Green, White

200 of these cards: Green, Black

2000 of these cards: White, White

350 of these cards: White, Black

2000 of these cards: Black, Black

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

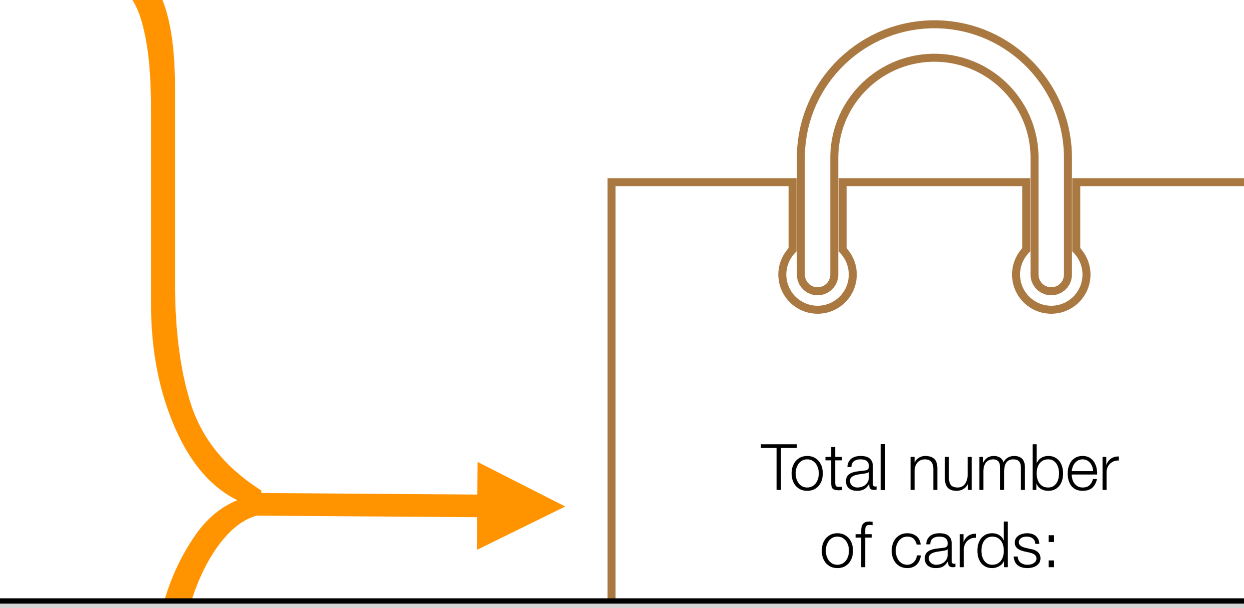
Probability of drawing "White, Black"?

350/5750

Probability of drawing a card that has "White" on it?

(200+2000+350)/5750

1000 of these cards:	Green, Green
200 of these cards:	Green, White
200 of these cards:	Green, Black
2000 of these cards:	White, White



$$P(\text{Green, White}) = \frac{200}{5750} \qquad P(\text{Green, Black}) = \frac{200}{5750} \qquad P(\text{White, Black}) = \frac{350}{5750}$$

$$P(\text{Green}) = \frac{1400}{5750} \qquad P(\text{White}) = \frac{2550}{5750} \qquad P(\text{Black}) = \frac{2550}{5750}$$

Measuring Surprise:
Pointwise Mutual Information (PMI)

Measuring Surprise:
Pointwise Mutual Information (PMI)

$P(A, B)$

Measuring Surprise:
Pointwise Mutual Information (PMI)

$$\frac{P(A, B)}{P(A)\ P(B)}$$

Measuring Surprise: Pointwise Mutual Information (PMI)

$$\text{PMI}(A, B) = \log \frac{P(A, B)}{P(A)\ P(B)}$$

Measuring Surprise: Pointwise Mutual Information (PMI)

$$\mathrm{PMI}(A, B) = \log \frac{P(A, B)}{P(A)\, P(B)}$$

Base of log doesn't really matter (we'll use base 2)

Measuring Surprise:
Pointwise Mutual Information (PMI)

$$\mathrm{PMI}(A, B) = \log_2 \frac{P(A, B)}{P(A)\ P(B)}$$

Base of log doesn't really matter (we'll use base 2)

Measuring Surprise: Pointwise Mutual Information (PMI)

$$\mathrm{PMI}(A, B) = \log_2 \frac{P(A, B)}{P(A)\ P(B)}$$

Base of log doesn't really matter (we'll use base 2)

PMI(Green, White) =

Measuring Surprise: Pointwise Mutual Information (PMI)

$$\text{PMI}(A, B) = \log_2 \frac{P(A, B)}{P(A)\ P(B)}$$

Base of log doesn't really matter (we'll use base 2)

$$\text{PMI}(\text{Green}, \text{White}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)}$$

Measuring Surprise: Pointwise Mutual Information (PMI)

$$\text{PMI}(A, B) = \log_2 \frac{P(A, B)}{P(A)\ P(B)}$$

Base of log doesn't really matter (we'll use base 2)

$$\text{PMI}(\text{Green, White}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = -1.63\ldots \text{ bits}$$

Measuring Surprise: Pointwise Mutual Information (PMI)

$$\text{PMI}(A, B) = \log_2 \frac{P(A, B)}{P(A)\ P(B)}$$

Base of log doesn't really matter (we'll use base 2)

$$\text{PMI}(\text{Green, White}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = -1.63\ldots \text{ bits}$$

PMI(Green, Black) =

Measuring Surprise: Pointwise Mutual Information (PMI)

$$\text{PMI}(A, B) = \log_2 \frac{P(A, B)}{P(A)\ P(B)}$$

Base of log doesn't really matter (we'll use base 2)

$$\text{PMI}(\text{Green, White}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = -1.63\ldots \text{ bits}$$

$$\text{PMI}(\text{Green, Black}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)}$$

Measuring Surprise: Pointwise Mutual Information (PMI)

$$\text{PMI}(A, B) = \log_2 \frac{P(A, B)}{P(A)\ P(B)}$$

Base of log doesn't really matter (we'll use base 2)

$$\text{PMI}(\text{Green, White}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = -1.63\ldots \text{ bits}$$

$$\text{PMI}(\text{Green, Black}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = -1.63\ldots \text{ bits}$$

Measuring Surprise: Pointwise Mutual Information (PMI)

$$\text{PMI}(A, B) = \log_2 \frac{P(A, B)}{P(A)\ P(B)}$$

Base of log doesn't really matter (we'll use base 2)

$$\text{PMI}(\text{Green, White}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = -1.63\ldots \text{ bits}$$

$$\text{PMI}(\text{Green, Black}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = -1.63\ldots \text{ bits}$$

$$\text{PMI}(\text{White, Black}) =$$

Measuring Surprise: Pointwise Mutual Information (PMI)

$$\text{PMI}(A, B) = \log_2 \frac{P(A, B)}{P(A)\,P(B)}$$

Base of log doesn't really matter (we'll use base 2)

$$\text{PMI}(\text{Green, White}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = -1.63\ldots \text{ bits}$$

$$\text{PMI}(\text{Green, Black}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = -1.63\ldots \text{ bits}$$

$$\text{PMI}(\text{White, Black}) = \log_2 \frac{350/5750}{(2550/5750)(2550/5750)}$$

Measuring Surprise: Pointwise Mutual Information (PMI)

$$\text{PMI}(A, B) = \log_2 \frac{P(A, B)}{P(A)\ P(B)}$$

Base of log doesn't really matter (we'll use base 2)

$$\text{PMI}(\text{Green}, \text{White}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = -1.63\ldots \text{ bits}$$

$$\text{PMI}(\text{Green}, \text{Black}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = -1.63\ldots \text{ bits}$$

$$\text{PMI}(\text{White}, \text{Black}) = \log_2 \frac{350/5750}{(2550/5750)(2550/5750)} = -1.69\ldots \text{ bits}$$

Measuring Surprise: Pointwise Mutual Information (PMI)

$$\text{PMI}(A, B) = \log_2 \frac{P(A, B)}{P(A)\ P(B)}$$

Base of log doesn't really matter (we'll use base 2)

$$\text{PMI}(\text{Green, White}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = -1.63\ldots \text{ bits}$$

$$\text{PMI}(\text{Green, Black}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = -1.63\ldots \text{ bits}$$

$-1.63 > -1.69$

$$\text{PMI}(\text{White, Black}) = \log_2 \frac{350/5750}{(2550/5750)(2550/5750)} = -1.69\ldots \text{ bits}$$

Measuring Surprise: Pointwise Mutual Information (PMI)

$$\text{PMI}(A, B) = \log_2 \frac{P(A, B)}{P(A)\,P(B)}$$

Base of log doesn't really matter (we'll use base 2)

Higher PMI ➔ more surprising

$$\text{PMI}(\text{Green}, \text{White}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = -1.63\ldots \text{ bits}$$

$$\text{PMI}(\text{Green}, \text{Black}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = -1.63\ldots \text{ bits}$$

$-1.63 > -1.69$

$$\text{PMI}(\text{White}, \text{Black}) = \log_2 \frac{350/5750}{(2550/5750)(2550/5750)} = -1.69\ldots \text{ bits}$$

Measuring Surprise: Pointwise Mutual Information (PMI)

$$\text{PMI}(A, B) = \log_2 \frac{P(A, B)}{P(A)\,P(B)}$$

Base of log doesn't really matter (we'll use base 2)

PMI can be positive or negative

Higher PMI ➔ more surprising

$$\text{PMI}(\text{Green, White}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = \text{-1.63... bits}$$

$$\text{PMI}(\text{Green, Black}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = \text{-1.63... bits}$$

-1.63 > -1.69

$$\text{PMI}(\text{White, Black}) = \log_2 \frac{350/5750}{(2550/5750)(2550/5750)} = \text{-1.69... bits}$$

Measuring Surprise: Pointwise Mutual Information (PMI)

$$\text{PMI}(A, B) = \log_2 \frac{P(A, B)}{P(A)\ P(B)}$$

PMI can be positive or negative

Higher PMI ➔ more surprising

Base of log doesn't really matter (we'll use base 2)

$$\text{PMI}(\text{Green}, \text{White}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = -1.63\ldots \text{ bits}$$

$$\text{PMI}(\text{Green}, \text{Black}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = -1.63\ldots \text{ bits}$$

-1.63 > -1.69

$$\text{PMI}(\text{White}, \text{Black}) = \log_2 \frac{350/5750}{(2550/5750)(2550/5750)} = -1.69\ldots \text{ bits}$$

$$P(\text{Green}, \text{White}) = \frac{200}{5750} \quad P(\text{Green}, \text{Black}) = \frac{200}{5750} \quad P(\text{White}, \text{Black}) = \frac{350}{5750}$$

$$P(\text{Green}) = \frac{1400}{5750} \quad P(\text{White}) = \frac{2550}{5750} \quad P(\text{Black}) = \frac{2550}{5750}$$

What is PMI Measuring?

What is PMI Measuring?

$$\frac{P(A, B)}{P(A)\ P(B)}$$

What is PMI Measuring?

Probability of A and B co-occurring

$$\frac{P(A, B)}{P(A)\ P(B)}$$

What is PMI Measuring?

Probability of A and B co-occurring

$$\frac{P(A, B)}{P(A)\,P(B)}$$

Probability of just A occurring

What is PMI Measuring?

Probability of A and B co-occurring

$$\frac{P(A, B)}{P(A)\,P(B)}$$

Probability of just A occurring

Probability of just B occurring

What is PMI Measuring?

Probability of A and B co-occurring

$$\frac{P(A, B)}{P(A)\,P(B)}$$

Probability of just A occurring

Probability of just B occurring

If A and B were "independent"

➔ probability of A and B co-occurring would be P(A)P(B)

What is PMI Measuring?

Probability of A and B co-occurring

$$\frac{P(A, B)}{P(A)\ P(B)}$$

What is PMI Measuring?

Probability of A and B co-occurring

$$\frac{P(A, B)}{P(A)\ P(B)}$$

Probability of A and B co-occurring *if they were independent*

What is PMI Measuring?

Probability of A and B co-occurring

$$\frac{P(A, B)}{P(A)\ P(B)}$$

Probability of A and B co-occurring *if they were independent*

PMI measures (the log of) a ratio that says how far A and B are from being independent

What is PMI Measuring?

Probability of A and B co-occurring

$$\frac{P(A, B)}{P(A)\ P(B)}$$

if equal to 1
➔ A, B are indep.

Probability of A and B co-occurring *if they were independent*

PMI measures (the log of) a ratio that says how far A and B are from being independent

What is PMI Measuring?

Probability of A and B co-occurring

$$\frac{P(A, B)}{P(A)\ P(B)}$$

Probability of A and B co-occurring *if they were independent*

if equal to 1
➔ A, B are indep.

PMI measures (the log of) a ratio that says how far A and B are from being independent

There are *lots* of connections of information theory to prediction

What is PMI Measuring?

Probability of A and B co-occurring

$$\frac{P(A, B)}{P(A)\ P(B)}$$

if equal to 1
➔ A, B are indep.

Probability of A and B co-occurring *if they were independent*

PMI measures (the log of) a ratio that says how far A and B are from being independent

There are *lots* of connections of information theory to prediction

Rough intuition:
Something surprising ⟷ less predictable ⟷ more bits to store

Looking at All Pairs of Outcomes

Looking at All Pairs of Outcomes

- PMI measures how P(A, B) differs from P(A)P(B) using a **log ratio**

Looking at All Pairs of Outcomes

- PMI measures how P(A, B) differs from P(A)P(B) using a **log ratio**
- **Log ratio** isn't the only way to compare!

Looking at All Pairs of Outcomes

- PMI measures how P(A, B) differs from P(A)P(B) using a **log ratio**
- **Log ratio** isn't the only way to compare!
- Another way to compare:

Looking at All Pairs of Outcomes

- PMI measures how P(A, B) differs from P(A)P(B) using a **log ratio**
- **Log ratio** isn't the only way to compare!
- Another way to compare:

$$\frac{[\ P(A, B) - P(A)\ P(B)\ ]^2}{P(A)\ P(B)}$$

Looking at All Pairs of Outcomes

- PMI measures how P(A, B) differs from P(A)P(B) using a **log ratio**
- **Log ratio** isn't the only way to compare!
- Another way to compare:

$$\frac{[\ P(A, B) - P(A)\ P(B)\ ]^2}{P(A)\ P(B)}$$

$$\text{Phi-square} = \sum_{A, B} \frac{[\ P(A, B) - P(A)\ P(B)\ ]^2}{P(A)\ P(B)}$$

Looking at All Pairs of Outcomes

- PMI measures how P(A, B) differs from P(A)P(B) using a **log ratio**
- **Log ratio** isn't the only way to compare!
- Another way to compare:

$$\frac{[\ P(A, B) - P(A)\ P(B)\ ]^2}{P(A)\ P(B)}$$

$$\text{Phi-square} = \sum_{A,\ B} \frac{[\ P(A, B) - P(A)\ P(B)\ ]^2}{P(A)\ P(B)}$$

Chi-square = N × Phi-square

N = sum of all co-occurrence counts (in upper right of triangle earlier)

Looking at All Pairs of Outcomes

- PMI measures how P(A, B) differs from P(A)P(B) using a **log ratio**
- **Log ratio** isn't the only way to compare!
- Another way to compare:

$$\frac{[\ P(A, B) - P(A)\ P(B)\ ]^2}{P(A)\ P(B)}$$

$$\text{Phi-square} = \sum_{A,\ B} \frac{[\ P(A, B) - P(A)\ P(B)\ ]^2}{P(A)\ P(B)}$$

$$\text{Chi-square} = N \times \text{Phi-square}$$

Measures how close *all* pairs of outcomes are close to being indep.

N = sum of all co-occurrence counts (in upper right of triangle earlier)

Looking at All Pairs of Outcomes

- PMI measures how P(A, B) differs from P(A)P(B) using a **log ratio**
- **Log ratio** isn't the only way to compare!
- Another way to compare:

$$\frac{[\ P(A, B) - P(A)\ P(B)\ ]^2}{P(A)\ P(B)}$$

Phi-square is between 0 and 1

$$\text{Phi-square} = \sum_{A,\ B} \frac{[\ P(A, B) - P(A)\ P(B)\ ]^2}{P(A)\ P(B)}$$

$$\text{Chi-square} = N \times \text{Phi-square}$$

Measures how close *all* pairs of outcomes are close to being indep.

N = sum of all co-occurrence counts (in upper right of triangle earlier)

Looking at All Pairs of Outcomes

- PMI measures how P(A, B) differs from P(A)P(B) using a **log ratio**
- **Log ratio** isn't the only way to compare!
- Another way to compare:

$$\frac{[\, P(A, B) - P(A)\, P(B) \,]^2}{P(A)\, P(B)}$$

$$\text{Phi-square} = \sum_{A,\, B} \frac{[\, P(A, B) - P(A)\, P(B) \,]^2}{P(A)\, P(B)}$$

Chi-square = N × Phi-square

Phi-square is between 0 and 1

0 → pairs are all indep.

Measures how close *all* pairs of outcomes are close to being indep.

N = sum of all co-occurrence counts (in upper right of triangle earlier)

$$\text{Phi-square} = \sum_{A,\,B} \frac{[\,P(A, B) - P(A)\,P(B)\,]^2}{P(A)\,P(B)}$$

$$\text{Chi-square} = N \times \text{Phi-square}$$

N = sum of all co-occurrence counts (in upper right of triangle earlier)

Example: Phi-Square Calculation

$$\text{Phi-square} = \sum_{A,\,B} \frac{[\ P(A, B) - P(A)\ P(B)\ ]^2}{P(A)\ P(B)}$$

Chi-square = N × Phi-square

N = sum of all co-occurrence counts (in upper right of triangle earlier)

Example: Phi-Square Calculation

$$P(\text{Green, White}) = \frac{200}{5750} \qquad P(\text{Green, Black}) = \frac{200}{5750} \qquad P(\text{White, Black}) = \frac{350}{5750}$$

$$P(\text{Green}) = \frac{1400}{5750} \qquad P(\text{White}) = \frac{2550}{5750} \qquad P(\text{Black}) = \frac{2550}{5750}$$

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

$$\text{Phi-square} = \sum_{A,\,B} \frac{[\,P(A, B) - P(A)\,P(B)\,]^2}{P(A)\,P(B)}$$

$$\text{Chi-square} = N \times \text{Phi-square}$$

N = sum of all co-occurrence counts (in upper right of triangle earlier)

Example: Phi-Square Calculation

$$P(\text{Green, White}) = \frac{200}{5750} \qquad P(\text{Green, Black}) = \frac{200}{5750} \qquad P(\text{White, Black}) = \frac{350}{5750}$$

$$P(\text{Green}) = \frac{1400}{5750} \qquad P(\text{White}) = \frac{2550}{5750} \qquad P(\text{Black}) = \frac{2550}{5750}$$

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

$$\text{Phi-square} = \sum_{A,\,B} \frac{[\,P(A, B) - P(A)\,P(B)\,]^2}{P(A)\,P(B)}$$

$N = 5750$

$$\text{Chi-square} = N \times \text{Phi-square}$$

N = sum of all co-occurrence counts (in upper right of triangle earlier)

Example: Phi-Square Calculation

$$P(\text{Green, White}) = \frac{200}{5750} \quad P(\text{Green, Black}) = \frac{200}{5750} \quad P(\text{White, Black}) = \frac{350}{5750}$$

$$P(\text{Green}) = \frac{1400}{5750} \quad P(\text{White}) = \frac{2550}{5750} \quad P(\text{Black}) = \frac{2550}{5750}$$

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

$N = 5750$

$$\text{Phi-square} = \sum_{A,\,B} \frac{[\,P(A, B) - P(A)\,P(B)\,]^2}{P(A)\,P(B)}$$

Example: Phi-Square Calculation

$$P(\text{Green, White}) = \frac{200}{5750} \qquad P(\text{Green, Black}) = \frac{200}{5750} \qquad P(\text{White, Black}) = \frac{350}{5750}$$

$$P(\text{Green}) = \frac{1400}{5750} \qquad P(\text{White}) = \frac{2550}{5750} \qquad P(\text{Black}) = \frac{2550}{5750}$$

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

N = 5750

Sum comprises of 6 terms

Green, Green:

Green, White:

Green, Black:

White, White:

White, Black:

Black, Black:

$$\text{Phi-square} = \sum_{A,\,B} \frac{[\,P(A, B) - P(A)\,P(B)\,]^2}{P(A)\,P(B)}$$

Example: Phi-Square Calculation

$$P(\text{Green, White}) = \frac{200}{5750} \qquad P(\text{Green, Black}) = \frac{200}{5750} \qquad P(\text{White, Black}) = \frac{350}{5750}$$

$$P(\text{Green}) = \frac{1400}{5750} \qquad P(\text{White}) = \frac{2550}{5750} \qquad P(\text{Black}) = \frac{2550}{5750}$$

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

N = 5750

Sum comprises of 6 terms

Green, Green: $\frac{\left[\frac{1000}{5750} - \left(\frac{1400}{5750}\right)\left(\frac{1400}{5750}\right)\right]^2}{\left(\frac{1400}{5750}\right)\left(\frac{1400}{5750}\right)}$

Green, White:

Green, Black:

White, White:

White, Black:

Black, Black:

$$\text{Phi-square} = \sum_{A,\,B} \frac{[\,P(A, B) - P(A)\,P(B)\,]^2}{P(A)\,P(B)}$$

Example: Phi-Square Calculation

$P(\text{Green, White}) = \frac{200}{5750}$ $P(\text{Green, Black}) = \frac{200}{5750}$ $P(\text{White, Black}) = \frac{350}{5750}$

$P(\text{Green}) = \frac{1400}{5750}$ $P(\text{White}) = \frac{2550}{5750}$ $P(\text{Black}) = \frac{2550}{5750}$

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

N = 5750

Sum comprises of 6 terms

Green, Green: $\frac{\left[\frac{1000}{5750} - \left(\frac{1400}{5750}\right)\left(\frac{1400}{5750}\right)\right]^2}{\left(\frac{1400}{5750}\right)\left(\frac{1400}{5750}\right)} = 0.2216\ldots$

Green, White:

Green, Black:

White, White:

White, Black:

Black, Black:

$$\text{Phi-square} = \sum_{A,\,B} \frac{[\,P(A, B) - P(A)\,P(B)\,]^2}{P(A)\,P(B)}$$

Example: Phi-Square Calculation

$$P(\text{Green, White}) = \frac{200}{5750} \quad P(\text{Green, Black}) = \frac{200}{5750} \quad P(\text{White, Black}) = \frac{350}{5750}$$

$$P(\text{Green}) = \frac{1400}{5750} \quad P(\text{White}) = \frac{2550}{5750} \quad P(\text{Black}) = \frac{2550}{5750}$$

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

N = 5750

Sum comprises of 6 terms

Green, Green: $\frac{[\frac{1000}{5750} - (\frac{1400}{5750})(\frac{1400}{5750})]^2}{(\frac{1400}{5750})(\frac{1400}{5750})} = 0.2216...$

Green, White: $\frac{[\frac{200}{5750} - (\frac{1400}{5750})(\frac{2550}{5750})]^2}{(\frac{1400}{5750})(\frac{2550}{5750})}$

Green, Black:

White, White:

White, Black:

Black, Black:

$$\text{Phi-square} = \sum_{A, B} \frac{[\, P(A, B) - P(A)\, P(B)\,]^2}{P(A)\, P(B)}$$

Example: Phi-Square Calculation

$$P(\text{Green, White}) = \frac{200}{5750} \quad P(\text{Green, Black}) = \frac{200}{5750} \quad P(\text{White, Black}) = \frac{350}{5750}$$

$$P(\text{Green}) = \frac{1400}{5750} \quad P(\text{White}) = \frac{2550}{5750} \quad P(\text{Black}) = \frac{2550}{5750}$$

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

$N = 5750$

Sum comprises of 6 terms

Green, Green: $\dfrac{\left[\frac{1000}{5750} - \left(\frac{1400}{5750}\right)\left(\frac{1400}{5750}\right)\right]^2}{\left(\frac{1400}{5750}\right)\left(\frac{1400}{5750}\right)} = 0.2216\ldots$

Green, White: $\dfrac{\left[\frac{200}{5750} - \left(\frac{1400}{5750}\right)\left(\frac{2550}{5750}\right)\right]^2}{\left(\frac{1400}{5750}\right)\left(\frac{2550}{5750}\right)} = 0.0496\ldots$

Green, Black:

White, White:

White, Black:

Black, Black:

$$\text{Phi-square} = \sum_{A,\,B} \frac{[\,P(A, B) - P(A)\,P(B)\,]^2}{P(A)\,P(B)}$$

Example: Phi-Square Calculation

$$P(\text{Green, White}) = \frac{200}{5750} \qquad P(\text{Green, Black}) = \frac{200}{5750} \qquad P(\text{White, Black}) = \frac{350}{5750}$$

$$P(\text{Green}) = \frac{1400}{5750} \qquad P(\text{White}) = \frac{2550}{5750} \qquad P(\text{Black}) = \frac{2550}{5750}$$

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

N = 5750

Sum comprises of 6 terms

Green, Green: $\frac{[\frac{1000}{5750} - (\frac{1400}{5750})(\frac{1400}{5750})]^2}{(\frac{1400}{5750})(\frac{1400}{5750})} = 0.2216\ldots$

Green, White: $\frac{[\frac{200}{5750} - (\frac{1400}{5750})(\frac{2550}{5750})]^2}{(\frac{1400}{5750})(\frac{2550}{5750})} = 0.0496\ldots$

Green, Black:

White, White:

White, Black:

Black, Black:

⋮

$$\text{Phi-square} = \sum_{A,\,B} \frac{[\,P(A, B) - P(A)\,P(B)\,]^2}{P(A)\,P(B)}$$

Example: Phi-Square Calculation

$$P(\text{Green, White}) = \frac{200}{5750} \qquad P(\text{Green, Black}) = \frac{200}{5750} \qquad P(\text{White, Black}) = \frac{350}{5750}$$

$$P(\text{Green}) = \frac{1400}{5750} \qquad P(\text{White}) = \frac{2550}{5750} \qquad P(\text{Black}) = \frac{2550}{5750}$$

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

$N = 5750$

Sum comprises of 6 terms

Green, Green: $\frac{\left[\frac{1000}{5750} - \left(\frac{1400}{5750}\right)\left(\frac{1400}{5750}\right)\right]^2}{\left(\frac{1400}{5750}\right)\left(\frac{1400}{5750}\right)} = 0.2216\ldots$

Green, White: $\frac{\left[\frac{200}{5750} - \left(\frac{1400}{5750}\right)\left(\frac{2550}{5750}\right)\right]^2}{\left(\frac{1400}{5750}\right)\left(\frac{2550}{5750}\right)} = 0.0496\ldots$

Green, Black: $= 0.0496\ldots$

White, White: $= 0.1161\ldots$

$\vdots$

White, Black: $= 0.0937\ldots$

Black, Black: $= 0.1161\ldots$

$$\text{Phi-square} = \sum_{A,\,B} \frac{[\,P(A, B) - P(A)\,P(B)\,]^2}{P(A)\,P(B)}$$

Example: Phi-Square Calculation

$$P(\text{Green, White}) = \frac{200}{5750} \qquad P(\text{Green, Black}) = \frac{200}{5750} \qquad P(\text{White, Black}) = \frac{350}{5750}$$

$$P(\text{Green}) = \frac{1400}{5750} \qquad P(\text{White}) = \frac{2550}{5750} \qquad P(\text{Black}) = \frac{2550}{5750}$$

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

N = 5750

Sum comprises of 6 terms

$$\text{Green, Green: } \frac{\left[\frac{1000}{5750} - \left(\frac{1400}{5750}\right)\left(\frac{1400}{5750}\right)\right]^2}{\left(\frac{1400}{5750}\right)\left(\frac{1400}{5750}\right)} = 0.2216\ldots$$

$$\text{Green, White: } \frac{\left[\frac{200}{5750} - \left(\frac{1400}{5750}\right)\left(\frac{2550}{5750}\right)\right]^2}{\left(\frac{1400}{5750}\right)\left(\frac{2550}{5750}\right)} = 0.0496\ldots$$

Green, Black: = 0.0496…

White, White: = 0.1161…

⋮

White, Black: = 0.0937…

Black, Black: = 0.1161…

$$\text{Phi-square} = \sum_{A,\,B} \frac{[\,P(A, B) - P(A)\,P(B)\,]^2}{P(A)\,P(B)}$$

Add these up to get:
Phi-square = 0.6470…

Example: Phi-Square Calculation

$P(\text{Green, White}) = \frac{200}{5750}$ $P(\text{Green, Black}) = \frac{200}{5750}$ $P(\text{White, Black}) = \frac{350}{5750}$

$P(\text{Green}) = \frac{1400}{5750}$ $P(\text{White}) = \frac{2550}{5750}$ $P(\text{Black}) = \frac{2550}{5750}$

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

$N = 5750$

Sum comprises of 6 terms

Green, Green: $\dfrac{[\frac{1000}{5750} - (\frac{1400}{5750})(\frac{1400}{5750})]^2}{(\frac{1400}{5750})(\frac{1400}{5750})} = 0.2216\ldots$

Green, White: $\dfrac{[\frac{200}{5750} - (\frac{1400}{5750})(\frac{2550}{5750})]^2}{(\frac{1400}{5750})(\frac{2550}{5750})} = 0.0496\ldots$

Green, Black: $= 0.0496\ldots$

White, White: $= 0.1161\ldots$

$\vdots$

White, Black: $= 0.0937\ldots$

Black, Black: $= 0.1161\ldots$

$$\text{Phi-square} = \sum_{A,\,B} \frac{[\,P(A, B) - P(A)\,P(B)\,]^2}{P(A)\,P(B)}$$

Add these up to get:
Phi-square = 0.6470…

Interpretation: neighboring pixels not close to being indep.

Back to Earlier Example

Back to Earlier Example

	Elon Musk	Tesla	Apple	Tim Cook
Elon Musk		300	1	4
Tesla	300		5	1
Apple	1	5		195
Tim Cook	4	1	195	

Back to Earlier Example

	Elon Musk	Tesla	Apple	Tim Cook
Elon Musk		300	1	4
Tesla	300		5	1
Apple	1	5		195
Tim Cook	4	1	195	

Often we know what kind of named entities are found

Back to Earlier Example

	Elon Musk	Tesla	Apple	Tim Cook
Elon Musk		300	1	4
Tesla	300		5	1
Apple	1	5		195
Tim Cook	4	1	195	

Often we know what kind of named entities are found

Example: Elon Musk and Tim Cook are people,
Tesla and Apple are companies

Back to Earlier Example

	Elon Musk	Tesla	Apple	Tim Cook
Elon Musk		300	1	4
Tesla	300		5	1
Apple	1	5		195
Tim Cook	4	1	195	

Often we know what kind of named entities are found

Example: Elon Musk and Tim Cook are people,
Tesla and Apple are companies

➔ can ask what people are related to what companies

Back to Earlier Example

	[redacted]	Tesla	Apple	[redacted]
Elon Musk	[redacted]	300	1	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Tim Cook	[redacted]	1	195	[redacted]

Often we know what kind of named entities are found

Example: Elon Musk and Tim Cook are people,
Tesla and Apple are companies

➔ can ask what people are related to what companies

Back to Earlier Example

	Tesla	Apple
Elon Musk	300	1
Tim Cook	1	195

Back to Earlier Example

	Tesla	Apple
Elon Musk	300	1
Tim Cook	1	195

PMI, phi-square, chi-square calculations are done the same way

Back to Earlier Example

	Tesla	Apple
Elon Musk	300	1
Tim Cook	1	195

PMI, phi-square, chi-square calculations are done the same way

Main things to calculate first:

Back to Earlier Example

	Tesla	Apple
Elon Musk	300	1
Tim Cook	1	195

PMI, phi-square, chi-square calculations are done the same way

Main things to calculate first:

P(Elon Musk, Tesla)

P(Elon Musk, Apple)

P(Tim Cook, Tesla)

P(Tim Cook, Apple)

P(Elon Musk)

P(Tim Cook)

P(Tesla)

P(Apple)

Back to Earlier Example

	Tesla	Apple
Elon Musk	300	1
Tim Cook	1	195

PMI, phi-square, chi-square calculations are done the same way

Main things to calculate first:

P(Elon Musk, Tesla)	P(Elon Musk)
P(Elon Musk, Apple)	P(Tim Cook)
P(Tim Cook, Tesla)	P(Tesla)
P(Tim Cook, Apple)	P(Apple)

The math here is actually a bit easier to think about
because the rows and columns aren't indexing the same items

Back to Earlier Example

	Tesla	Apple
Elon Musk	300	1
Tim Cook	1	195

Back to Earlier Example

	Tesla	Apple
Elon Musk	300	1
Tim Cook	1	195

Total: 497

Back to Earlier Example

	Tesla	Apple
Elon Musk	300	1
Tim Cook	1	195

Total: 497

↓ Divide by total

	Tesla	Apple
Elon Musk	300/497	1/497
Tim Cook	1/497	195/497

Back to Earlier Example

	Tesla	Apple
Elon Musk	300	1
Tim Cook	1	195

Total: 497

↓ Divide by total

These are the joint probabilities!

	Tesla	Apple
Elon Musk	300/497	1/497
Tim Cook	1/497	195/497

Back to Earlier Example

	Tesla	Apple
Elon Musk	300	1
Tim Cook	1	195

Total: 497

↓ Divide by total

These are the joint probabilities!

	Tesla	Apple
Elon Musk	300/497	1/497
Tim Cook	1/497	195/497

P(Elon Musk, Tesla) → 300/497

P(Elon Musk, Apple) → 1/497

P(Tim Cook, Tesla) → 1/497

P(Tim Cook, Apple) → 195/497

Back to Earlier Example

	Tesla	Apple
Elon Musk	300	1
Tim Cook	1	195

Total: 497

↓ Divide by total

These are the joint probabilities!

Compute "marginals"

	Tesla	Apple
Elon Musk	300/497	1/497
Tim Cook	1/497	195/497

P(Elon Musk, Tesla)

P(Elon Musk, Apple)

P(Tim Cook, Tesla)

P(Tim Cook, Apple)

Back to Earlier Example

	Tesla	Apple
Elon Musk	300	1
Tim Cook	1	195

Total: 497

↓ Divide by total

These are the joint probabilities!

Compute "marginals"

	Tesla	Apple	
Elon Musk	300/497	1/497	300/497+1/497
Tim Cook	1/497	195/497	1/497+195/497
	300/497+1/497	1/497+195/497	

P(Elon Musk, Tesla) → 300/497

P(Elon Musk, Apple) → 1/497

P(Tim Cook, Tesla) → 1/497

P(Tim Cook, Apple) → 195/497

Back to Earlier Example

	Tesla	Apple
Elon Musk	300	1
Tim Cook	1	195

Total: 497

↓ Divide by total

These are the joint probabilities!

Compute "marginals"

	Tesla	Apple	
Elon Musk	300/497	1/497	300/497+1/497
Tim Cook	1/497	195/497	1/497+195/497
	300/497+1/497	1/497+195/497	

P(Elon Musk, Tesla)

P(Elon Musk, Apple)

P(Tim Cook, Tesla)

P(Tim Cook, Apple)

P(Elon Musk)

P(Tim Cook)

P(Tesla)

P(Apple)

Back to Earlier Example

	Tesla	Apple
Elon Musk	300	1
Tim Cook	1	195

Total: 497

Divide by total

These are the joint probabilities!

	Tesla	Apple	Compute "marginals"
Elon Musk	300/497	1/497	300/497+1/497
Tim Cook	1/497	195/497	1/497+195/497
	300/497+1/497	1/497+195/497	

P(Elon Musk, Tesla)

P(Elon Musk, Apple)

P(Tim Cook, Tesla)

P(Tim Cook, Apple)

P(Elon Musk)

P(Tim Cook)

P(Tesla)

P(Apple)

Not just for 2 by 2 tables

Back to Earlier Example

	Tesla	Apple
Elon Musk	300	1
Tim Cook	1	195

Total: 497

↓ Divide by total

These are the joint probabilities!

	Tesla	Apple	
Elon Musk	300/497	1/497	300/497+1/497
Tim Cook	1/497	195/497	1/497+195/497
	300/497+1/497	1/497+195/497	

Compute "marginals"

P(Elon Musk, Tesla)

P(Elon Musk, Apple)

P(Tim Cook, Tesla)

P(Tim Cook, Apple)

P(Elon Musk)

P(Tim Cook)

P(Tesla)

P(Apple)

Not just for 2 by 2 tables (e.g., we could have many people, many companies)

Recap: Co-Occurrences

Recap: Co-Occurrences

- Joint probability P(A, B) can be poor indicator of whether A and B co-occurring is “interesting”

Recap: Co-Occurrences

- Joint probability P(A, B) can be poor indicator of whether A and B co-occurring is “interesting”
- Find interesting relationships between pairs of items by looking at PMI

Recap: Co-Occurrences

- Joint probability P(A, B) can be poor indicator of whether A and B co-occurring is “interesting”
- Find interesting relationships between pairs of items by looking at PMI
    - Intuition: “Interesting” co-occurring events should occur more frequently than if they were to co-occur independently

Recap: Co-Occurrences

- Joint probability P(A, B) can be poor indicator of whether A and B co-occurring is "interesting"
- Find interesting relationships between pairs of items by looking at PMI
    - Intuition: "Interesting" co-occurring events should occur more frequently than if they were to co-occur independently
- In practice: some times it is helpful to generalize PMI and look instead at

$$\text{PMI}_\rho(A, B) = \log_2 \frac{P(A, B)^\rho}{P(A)\ P(B)}$$

Recap: Co-Occurrences

- Joint probability P(A, B) can be poor indicator of whether A and B co-occurring is "interesting"
- Find interesting relationships between pairs of items by looking at PMI
  - Intuition: "Interesting" co-occurring events should occur more frequently than if they were to co-occur independently
- In practice: some times it is helpful to generalize PMI and look instead at

$$\text{PMI}_\rho(A, B) = \log_2 \frac{P(A, B)^\rho}{P(A)\, P(B)}$$

Tune parameter
$\rho > 0$

Recap: Co-Occurrences

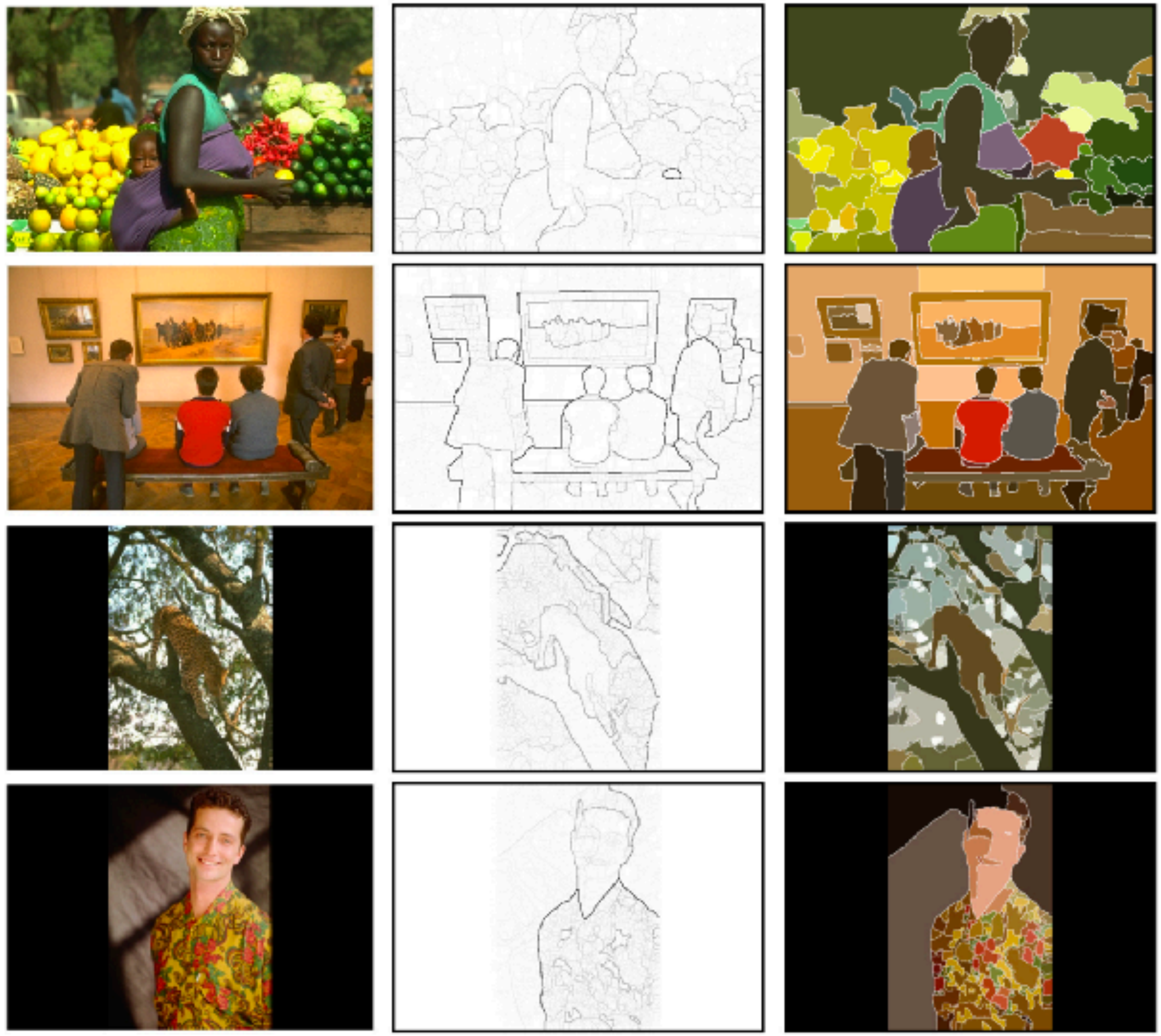
- Joint probability P(A, B) can be poor indicator of whether A and B co-occurring is “interesting”
- Find interesting relationships between pairs of items by looking at PMI
  - Intuition: “Interesting” co-occurring events should occur more frequently than if they were to co-occur independently
- In practice: some times it is helpful to generalize PMI and look instead at

$$\text{PMI}_\rho(A, B) = \log_2 \frac{P(A, B)^\rho}{P(A)\, P(B)}$$

Tune parameter
$\rho > 0$

(we'll talk about parameter tuning later in the course)

Example Application of PMI: Image Segmentation



Phillip Isola, Daniel Zoran, Dilip Krishnan, and Edward H. Adelson. Crisp boundary detection using pointwise mutual information. ECCV 2014.

Example Application of PMI: Word Embeddings

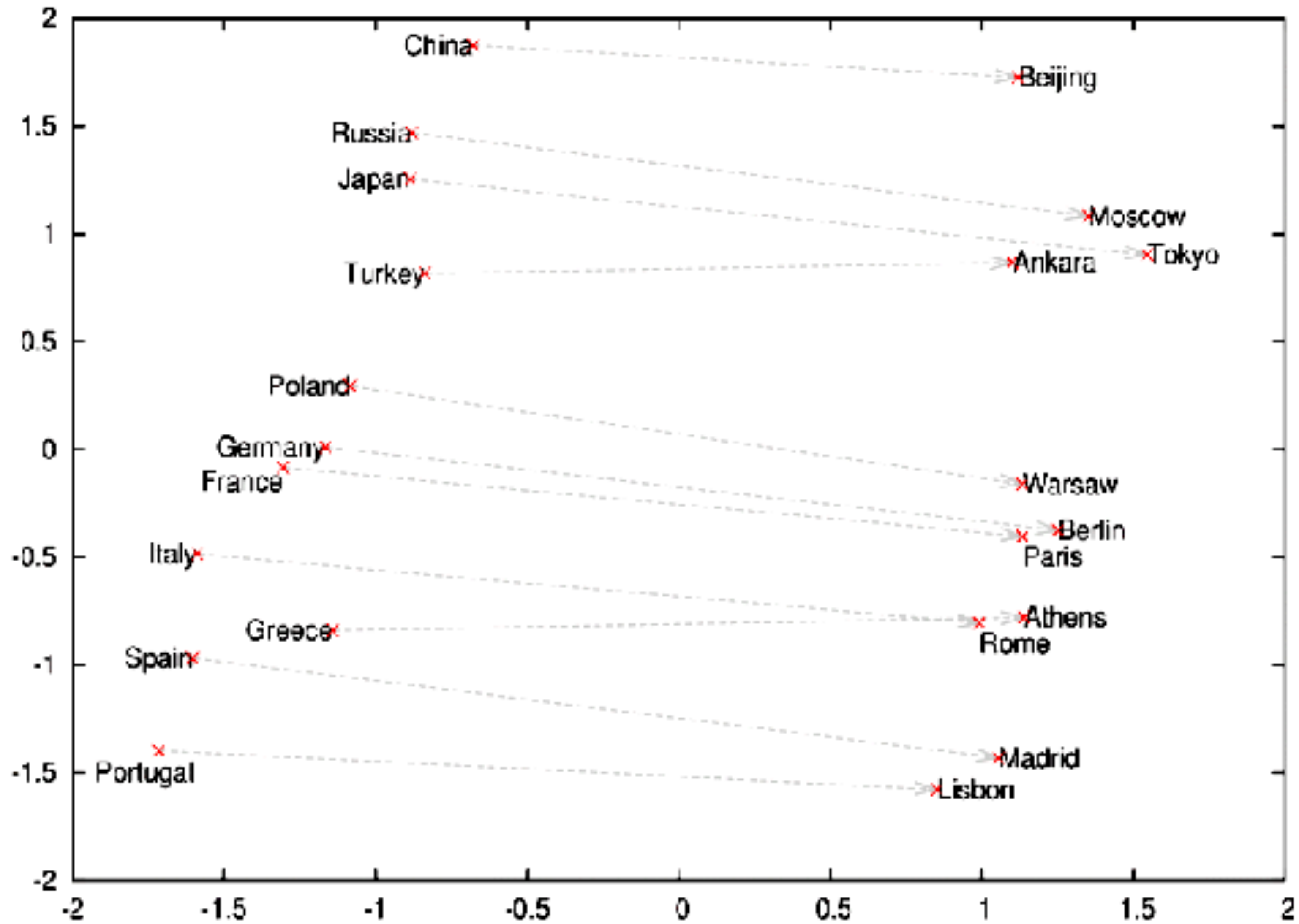


Image source: https://deeplearning4j.org/img/countries_capitals.png

Omer Levy and Yoav Goldberg. Neural word embeddings as implicit matrix factorization. NIPS 2014.

Continuous Measurements

Continuous Measurements

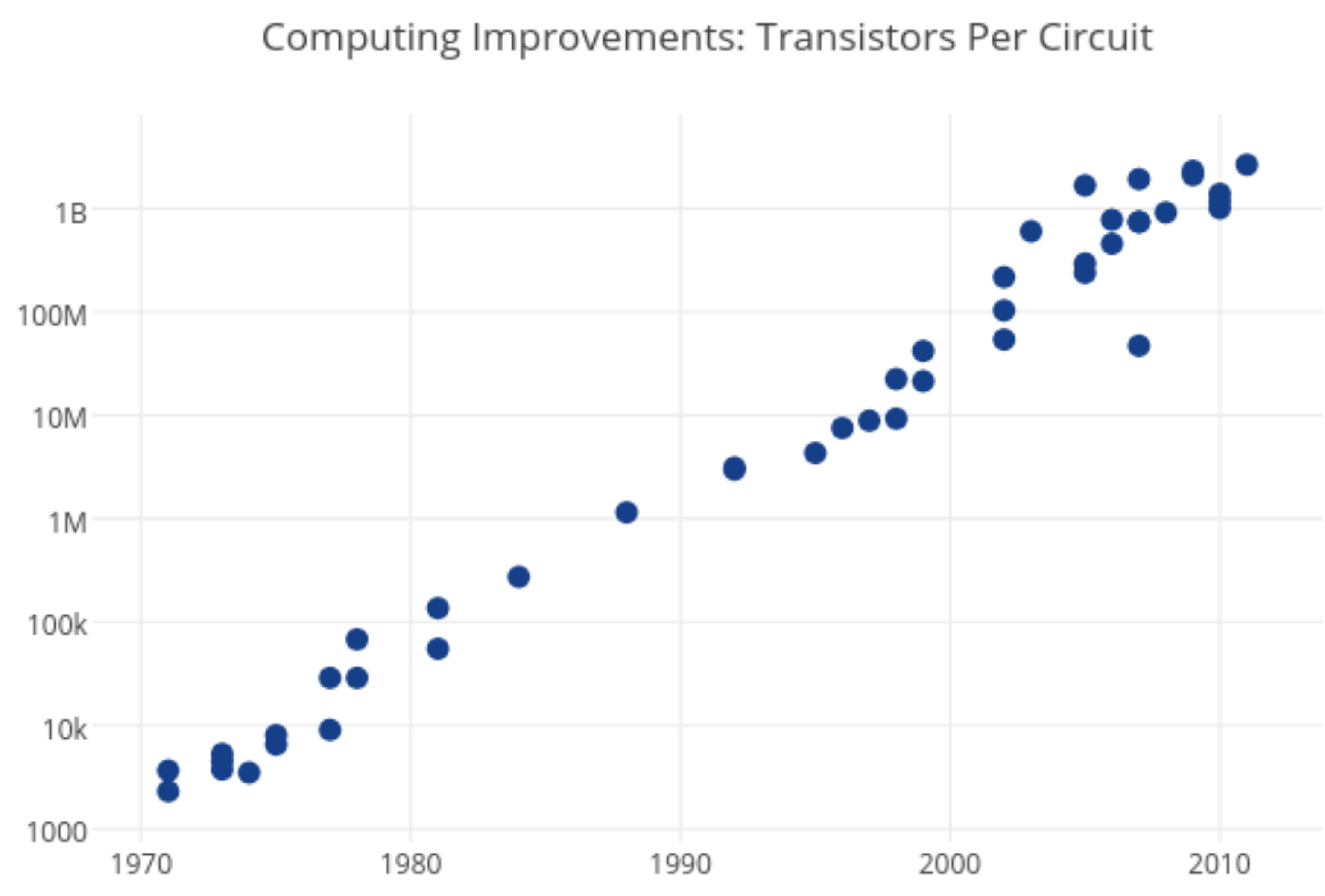
- So far, looked at relationships between *discrete* outcomes

Continuous Measurements

- So far, looked at relationships between *discrete* outcomes
- For pair of *continuous* outcomes, use a **scatter plot**

Continuous Measurements

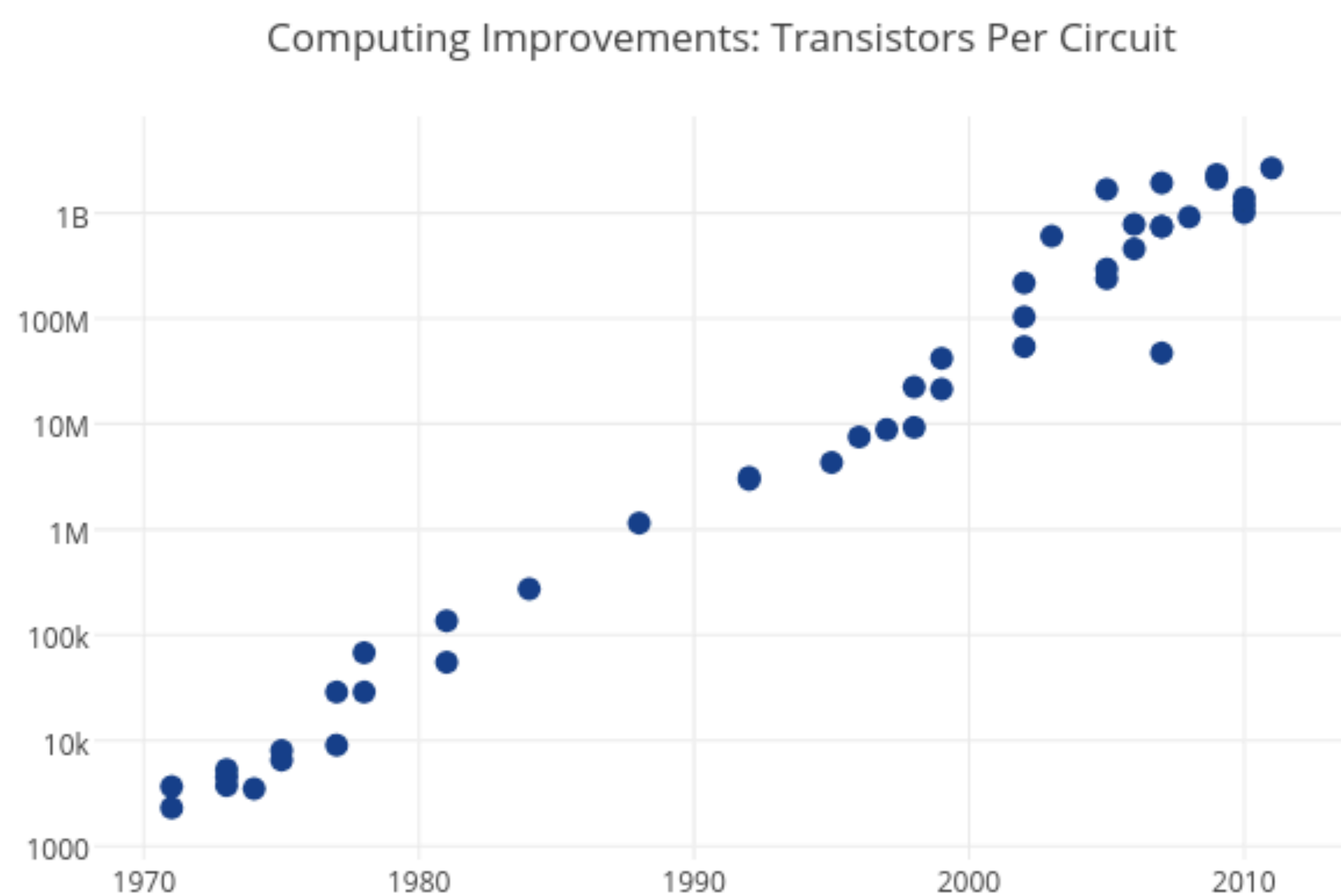
- So far, looked at relationships between *discrete* outcomes
- For pair of *continuous* outcomes, use a **scatter plot**



https://plot.ly/~MattSundquist/5405.png

Continuous Measurements

- So far, looked at relationships between *discrete* outcomes
- For pair of *continuous* outcomes, use a **scatter plot**

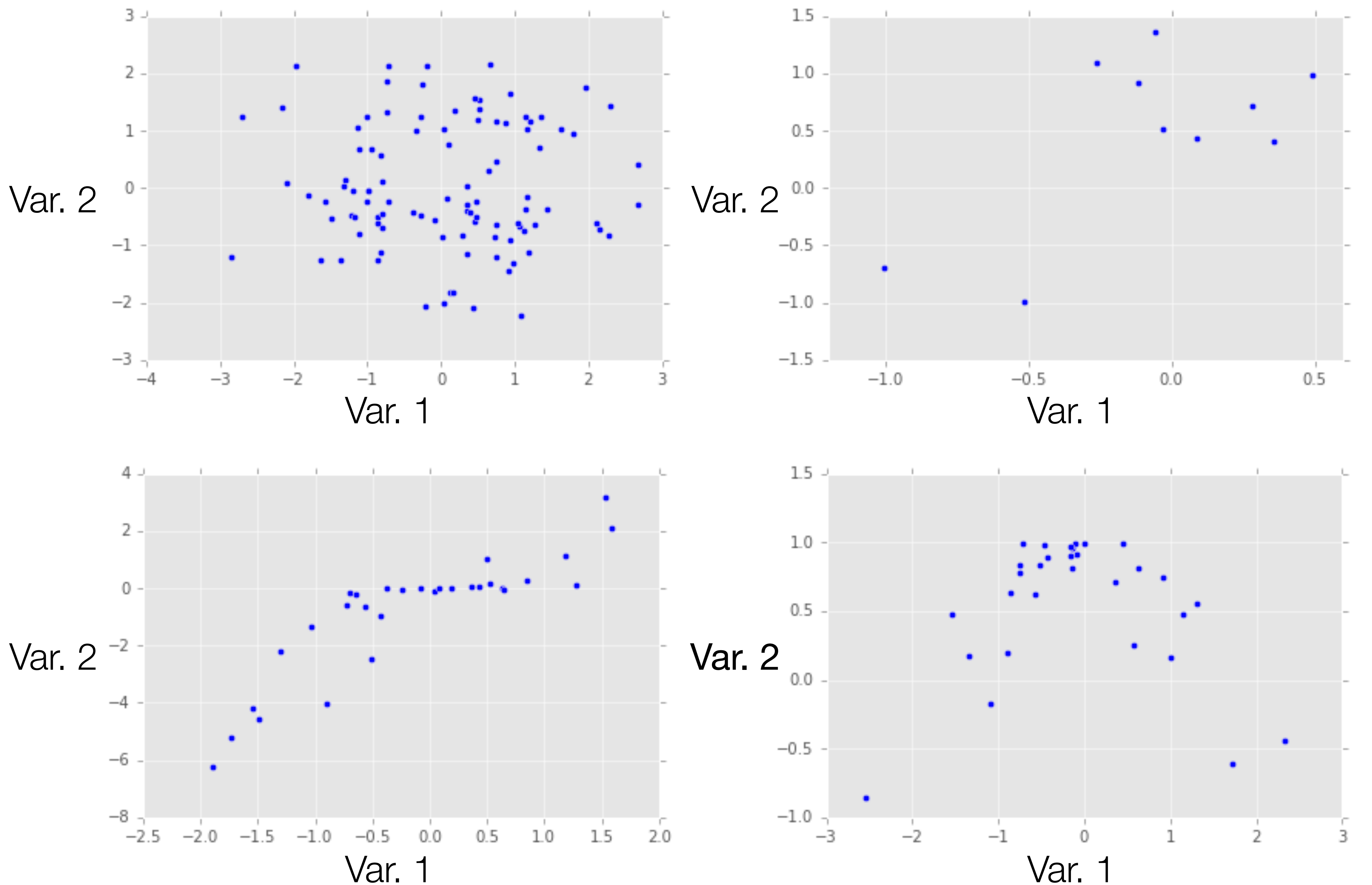


Of course, not all trends look like a line

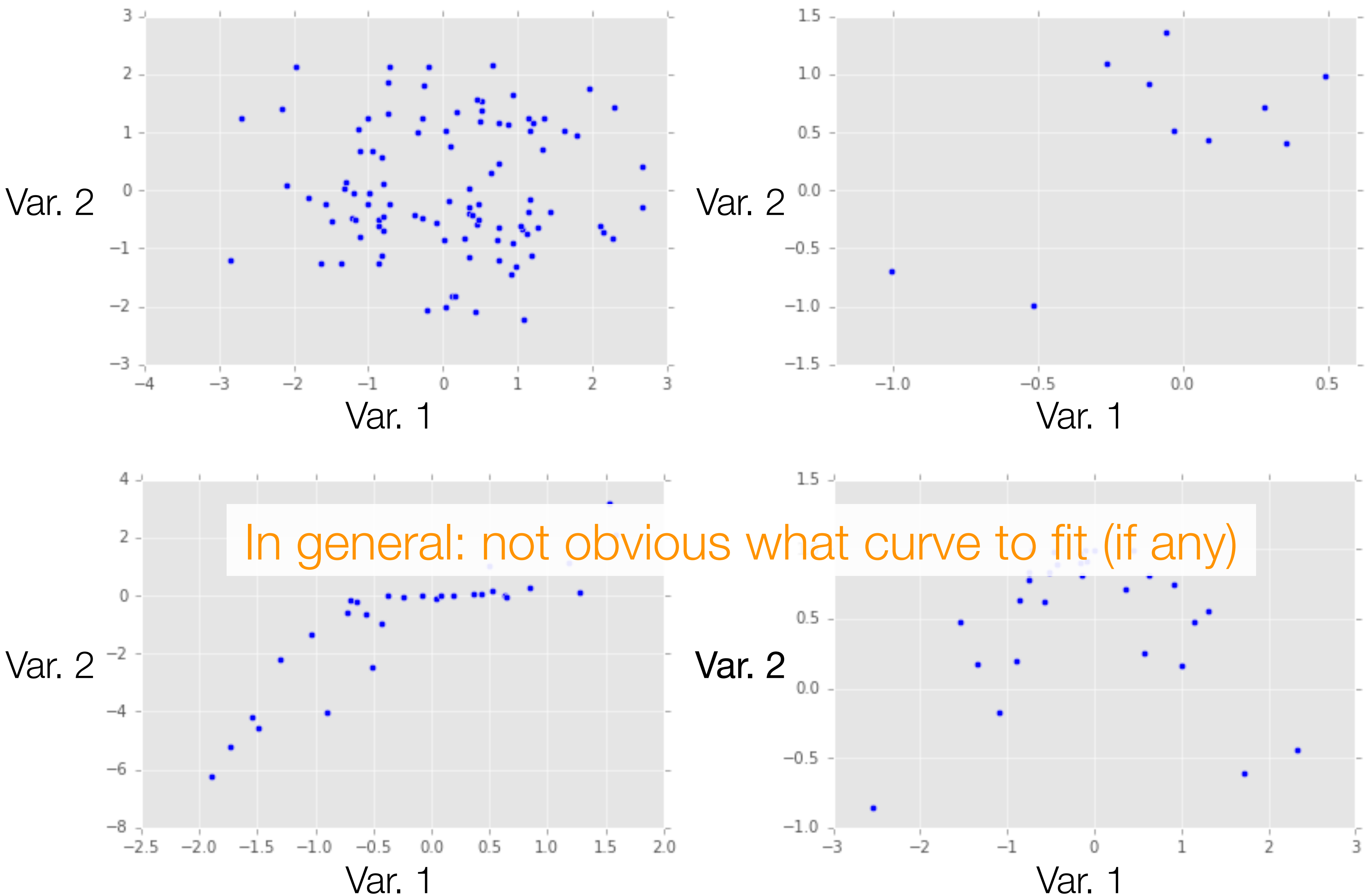
https://plot.ly/~MattSundquist/5405.png

The Importance of Staring at Data

The Importance of Staring at Data

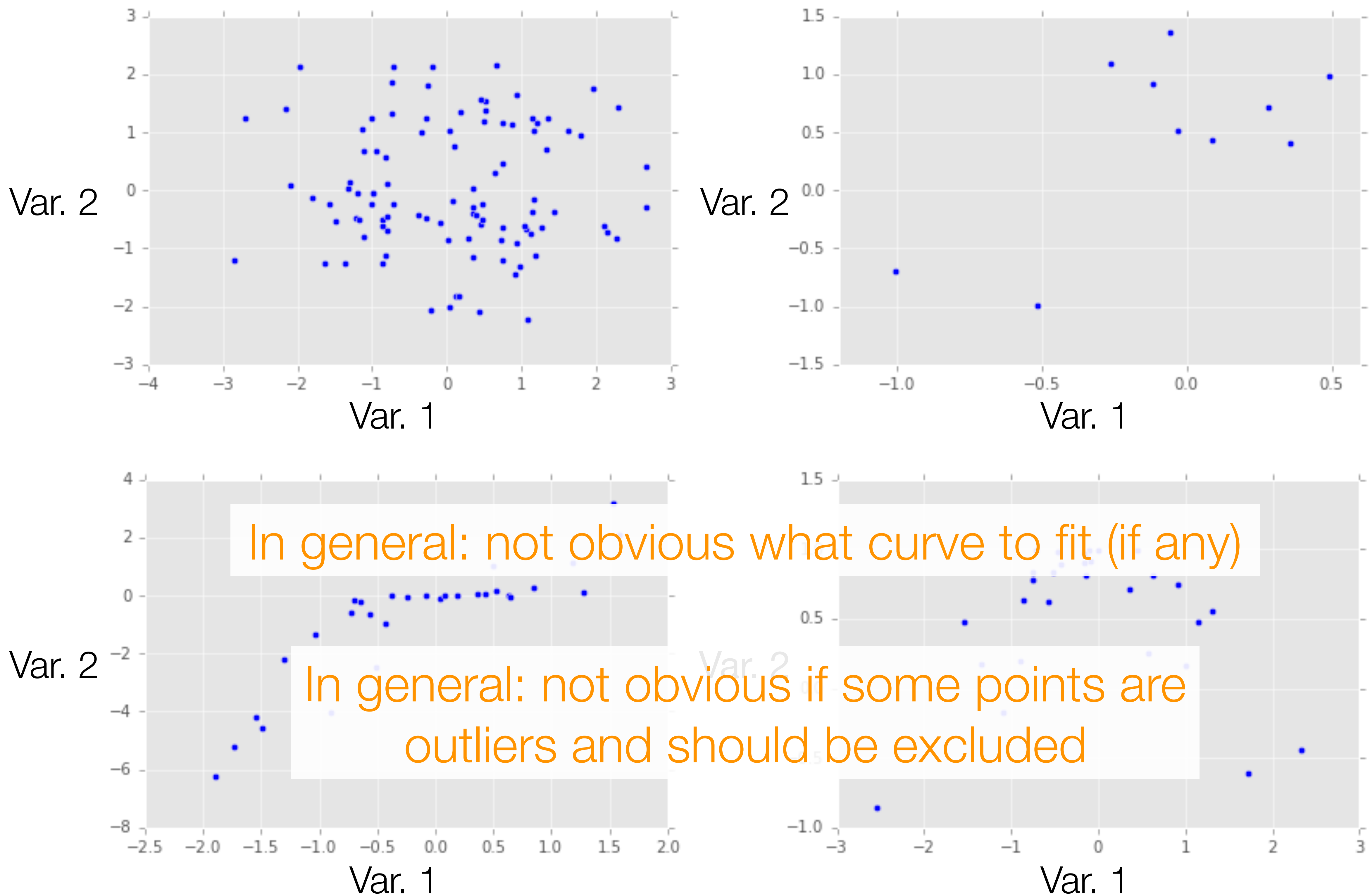


The Importance of Staring at Data



In general: not obvious what curve to fit (if any)

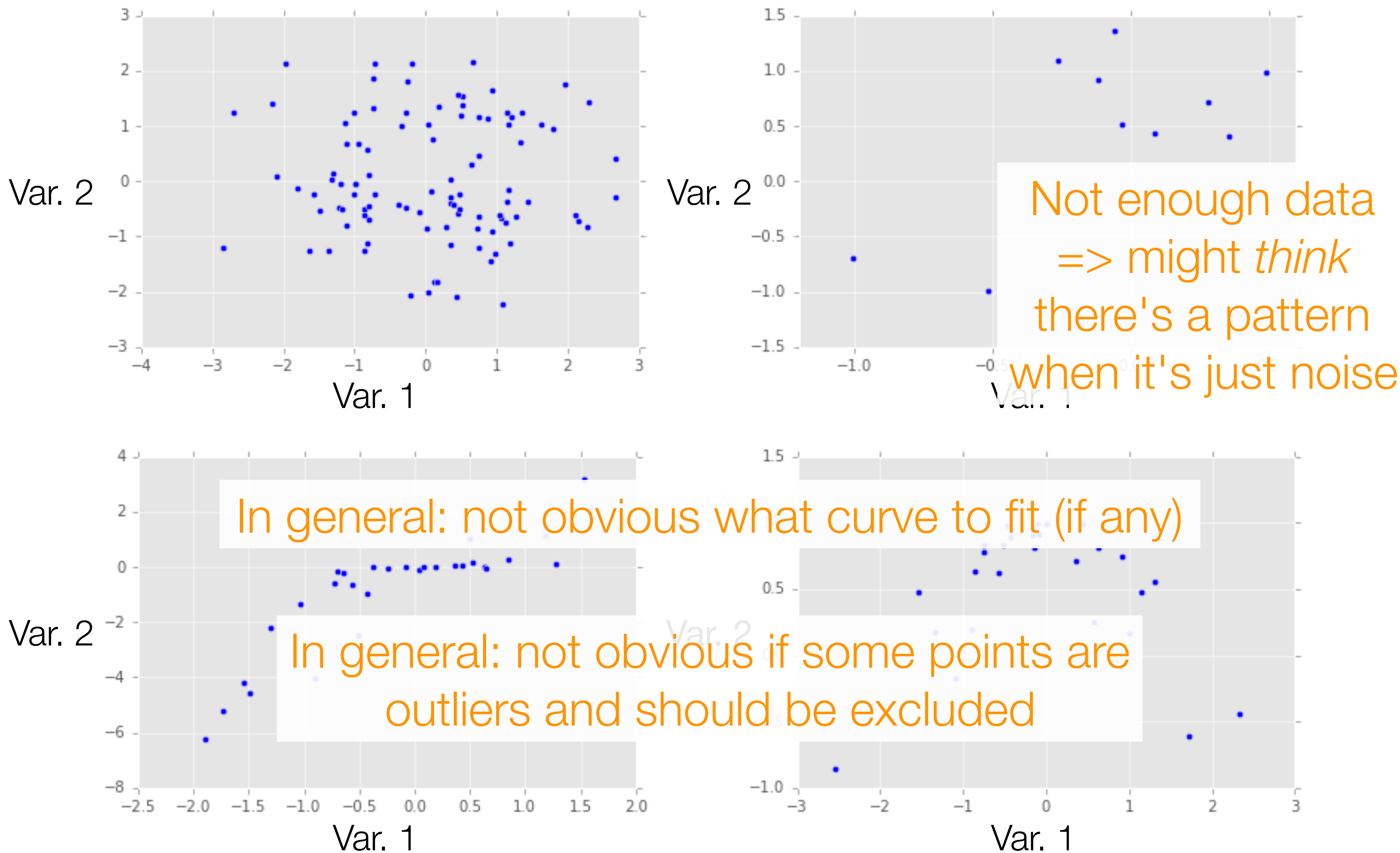
The Importance of Staring at Data



In general: not obvious what curve to fit (if any)

In general: not obvious if some points are outliers and should be excluded

The Importance of Staring at Data

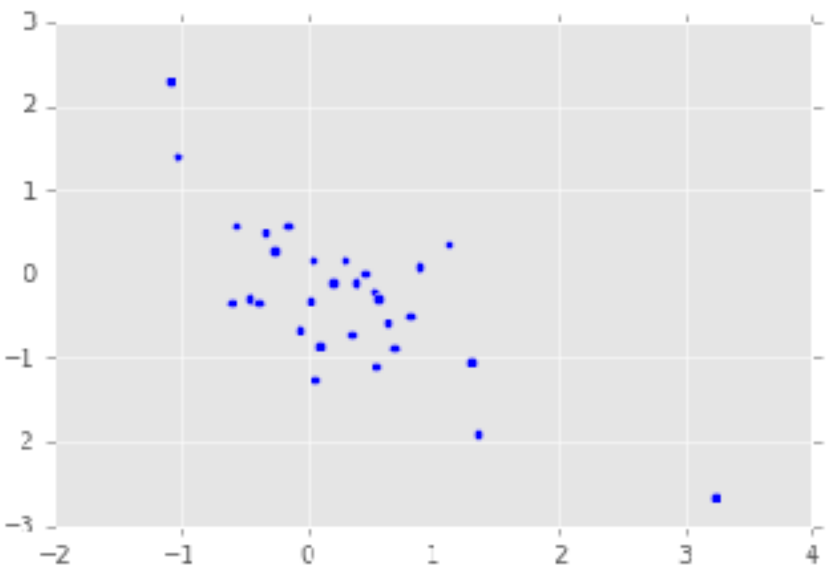


Not enough data => might *think* there's a pattern when it's just noise

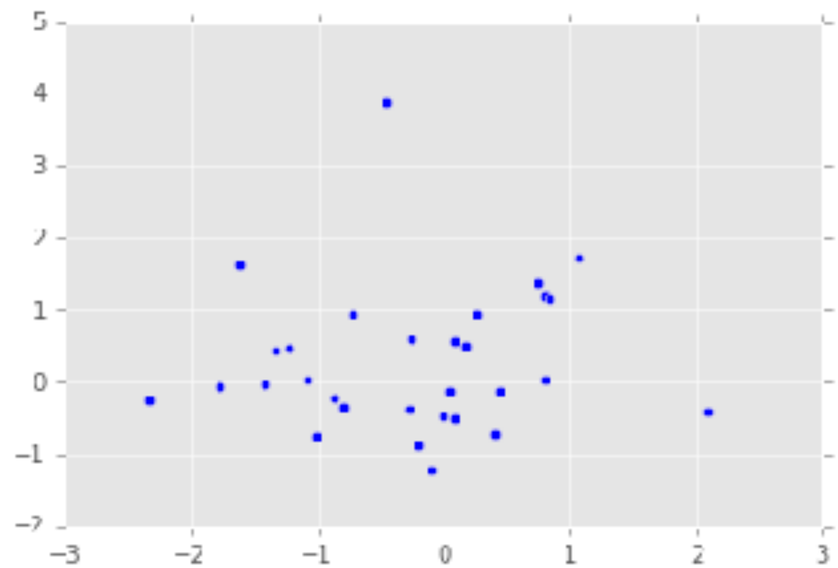
In general: not obvious what curve to fit (if any)

In general: not obvious if some points are outliers and should be excluded

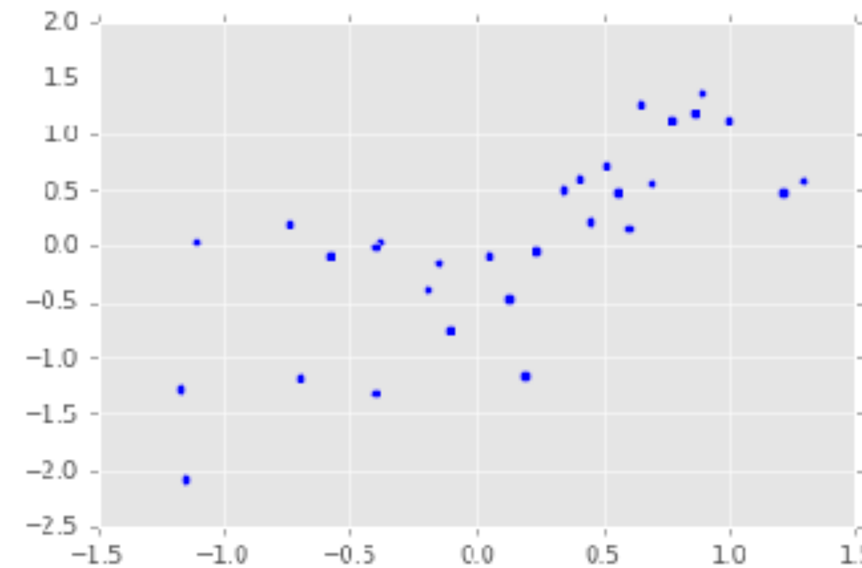
Correlation



Negatively correlated

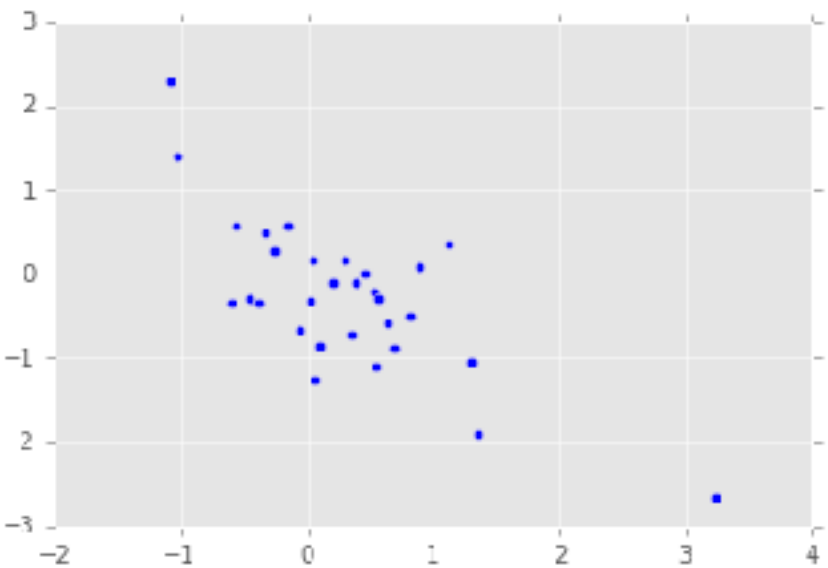


Not really correlated

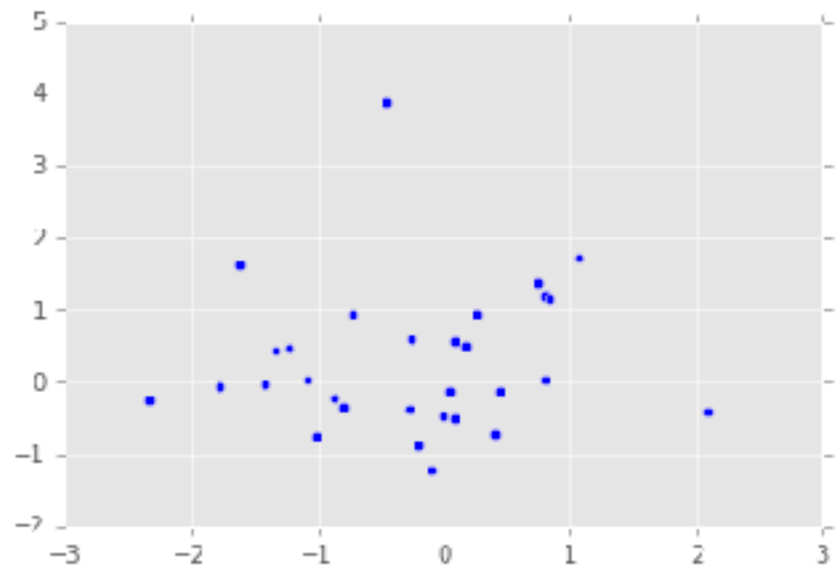


Positively correlated

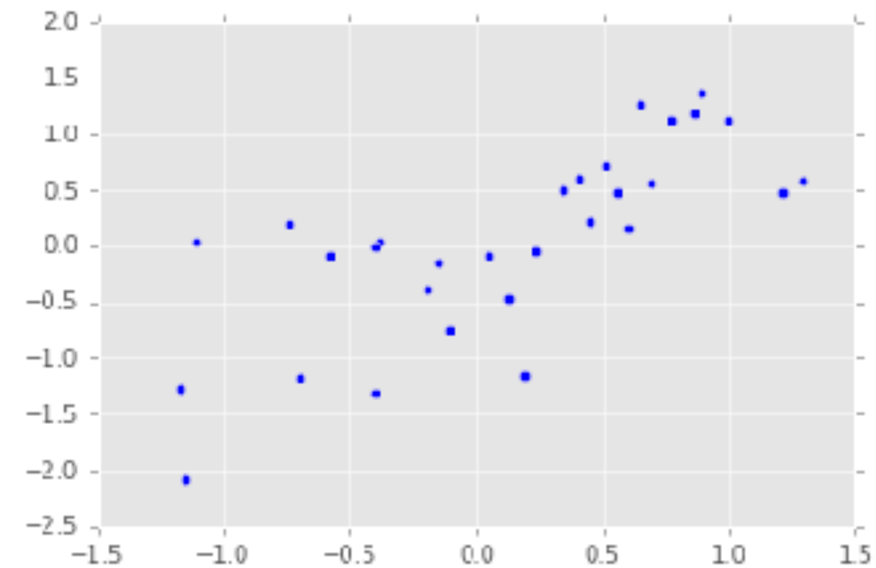
Correlation



Negatively correlated

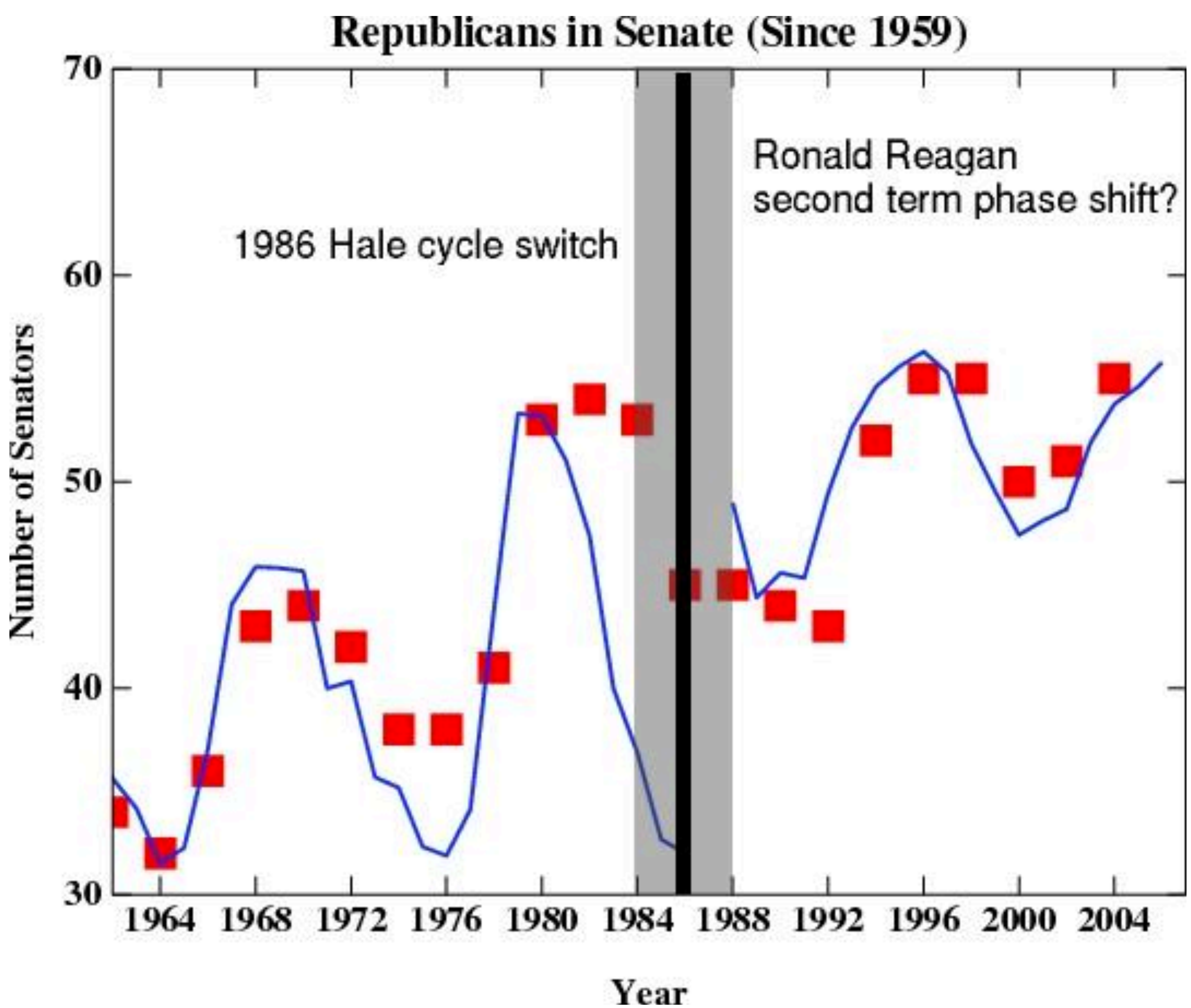


Not really correlated



Positively correlated

Beware: Just because two variables appear correlated
doesn't mean that one can predict the other

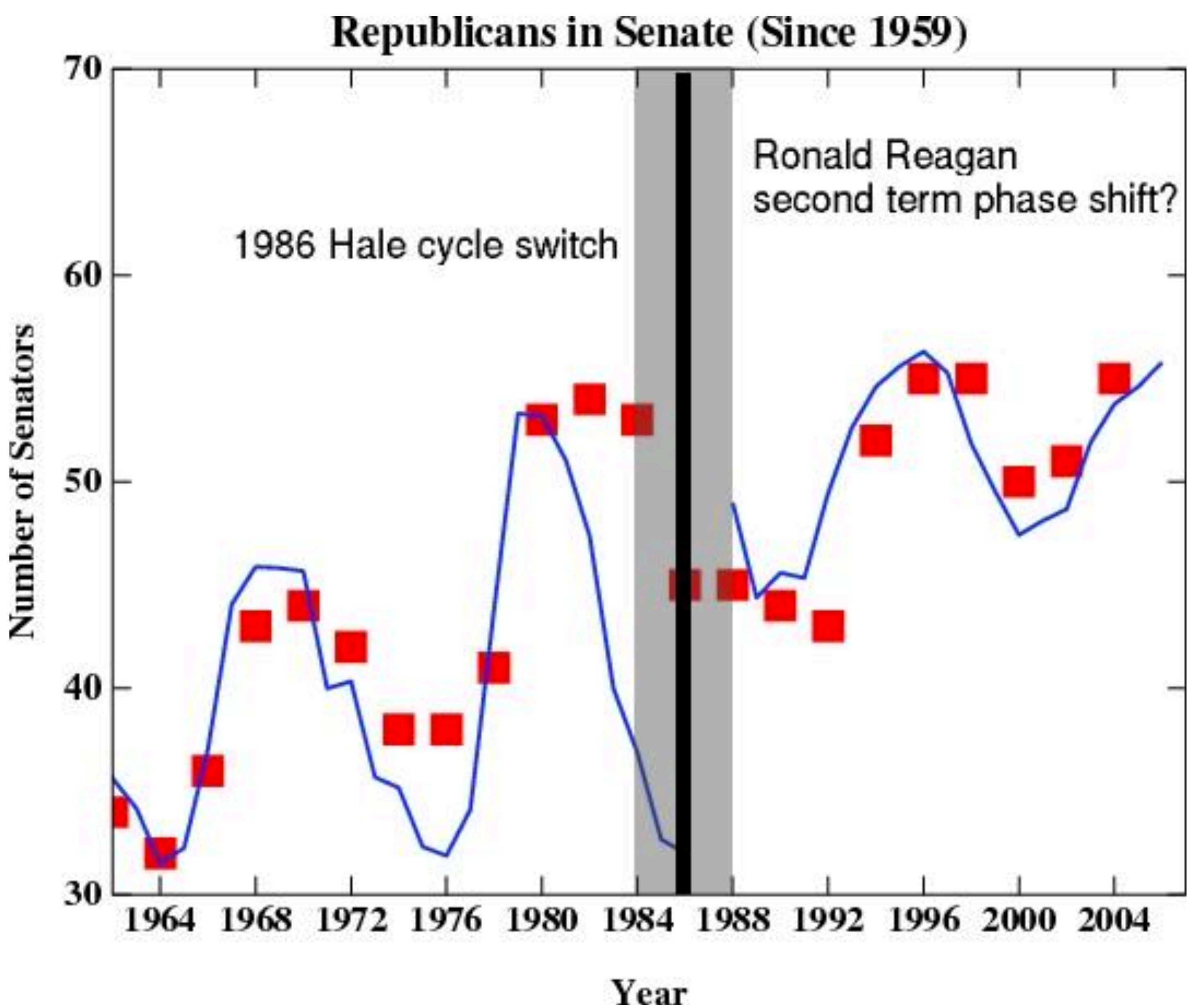


Blue: Scaled sunspot number (inverted after Reagan's 2nd term)

Red: Number of Republican senators

http://www.realclimate.org/index.php/archives/2007/05/fun-with-correlations/

Correlation ≠ Causation

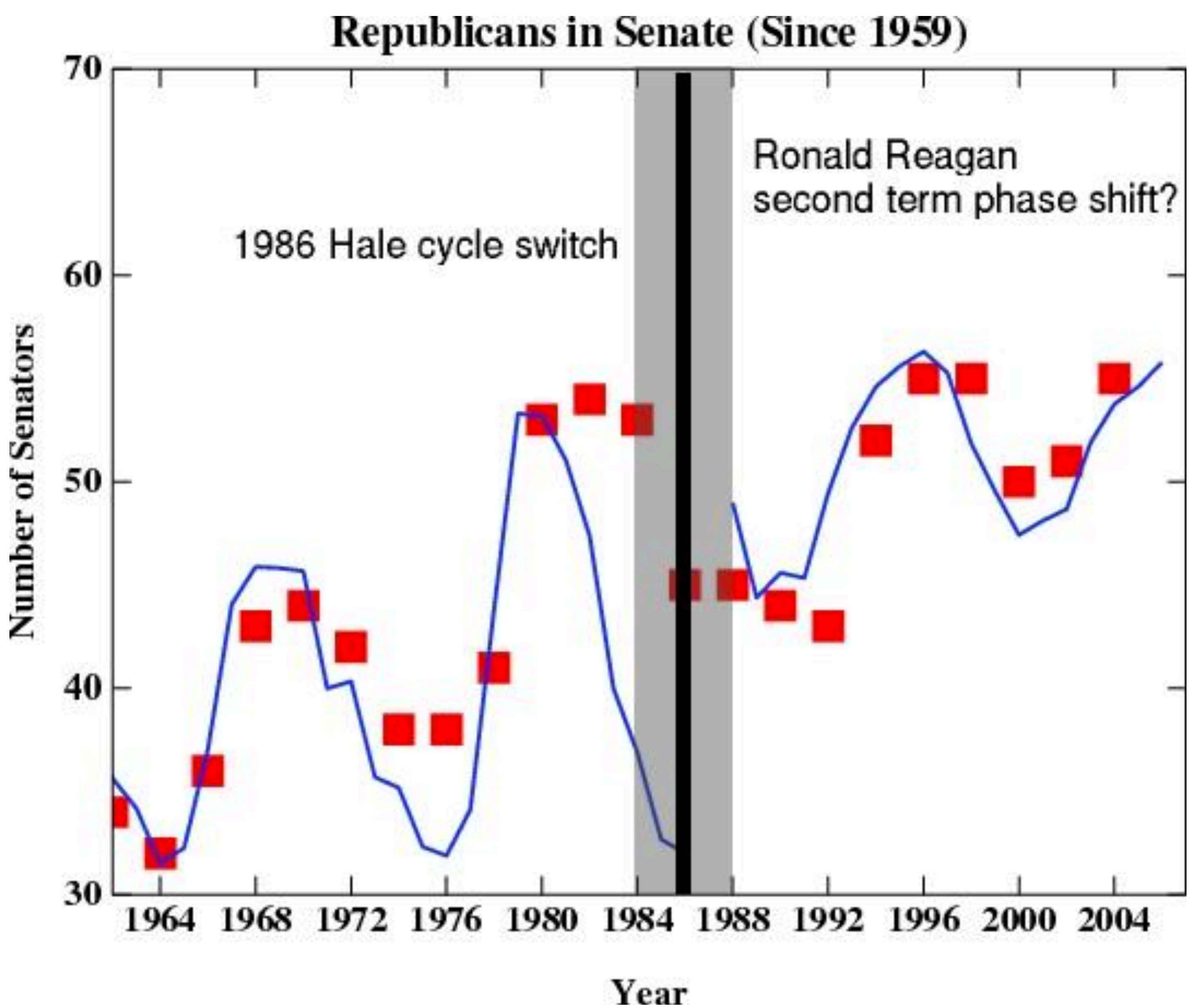


Blue: Scaled sunspot number (inverted after Reagan's 2nd term)

Red: Number of Republican senators

http://www.realclimate.org/index.php/archives/2007/05/fun-with-correlations/

Correlation ≠ Causation



Blue: Scaled sunspot number (inverted after Reagan's 2nd term)

Red: Number of Republican senators

Moreover, just because we find correlation in data doesn't mean it has predictive value!

http://www.realclimate.org/index.php/archives/2007/05/fun-with-correlations/

Important: At this point in the course, we are finding *possible* relationships between two entities

Important: At this point in the course, we are finding *possible* relationships between two entities

We are *not* yet making statements about prediction (we'll see prediction later in the course)

Important: At this point in the course, we are finding *possible* relationships between two entities

We are *not* yet making statements about prediction (we'll see prediction later in the course)

We are *not* making statements about causality (beyond the scope of this course)

Causality

Causality

Causality

Studies in 1960's: Coffee drinkers have higher rates of lung cancer

Causality

Studies in 1960's: Coffee drinkers have higher rates of lung cancer

Can we claim that coffee is a cause of lung cancer?

Causality

Studies in 1960's: Coffee drinkers have higher rates of lung cancer

Can we claim that coffee is a cause of lung cancer?

Back then: coffee drinkers also tended to smoke more than non-coffee drinkers (smoking is a **confounding variable**)

Causality

Studies in 1960's: Coffee drinkers have higher rates of lung cancer

Can we claim that coffee is a cause of lung cancer?

Back then: coffee drinkers also tended to smoke more than non-coffee drinkers (smoking is a **confounding variable**)

To establish causality, groups getting different treatments need to appear similar so that the only difference is the treatment

Establishing Causality

Establishing Causality

If you control data collection

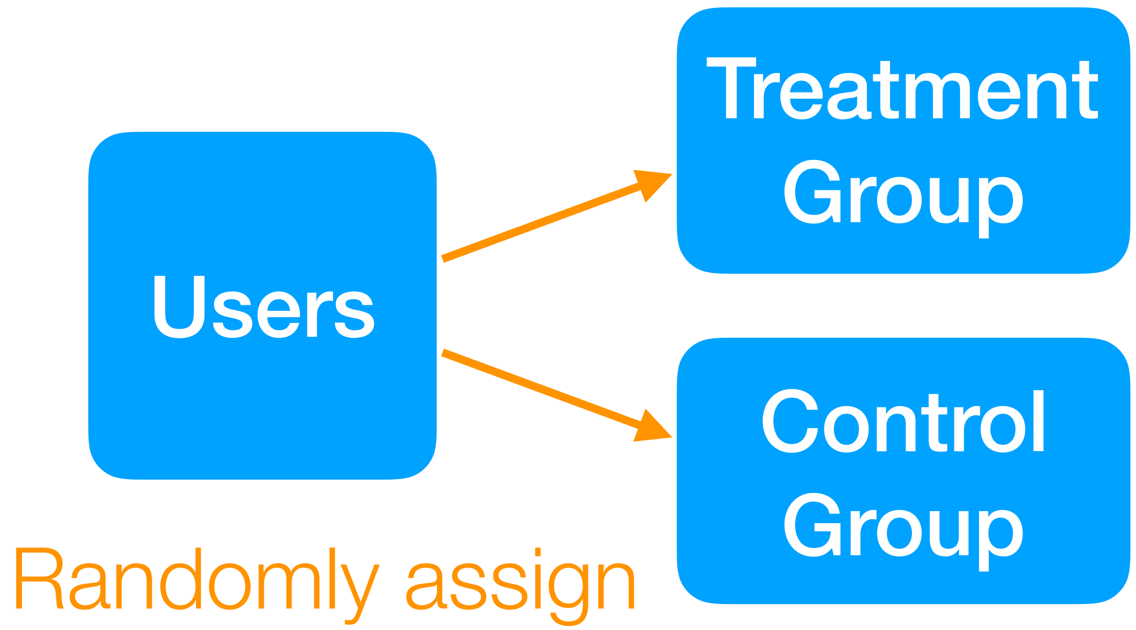
Establishing Causality

If you control data collection

Users

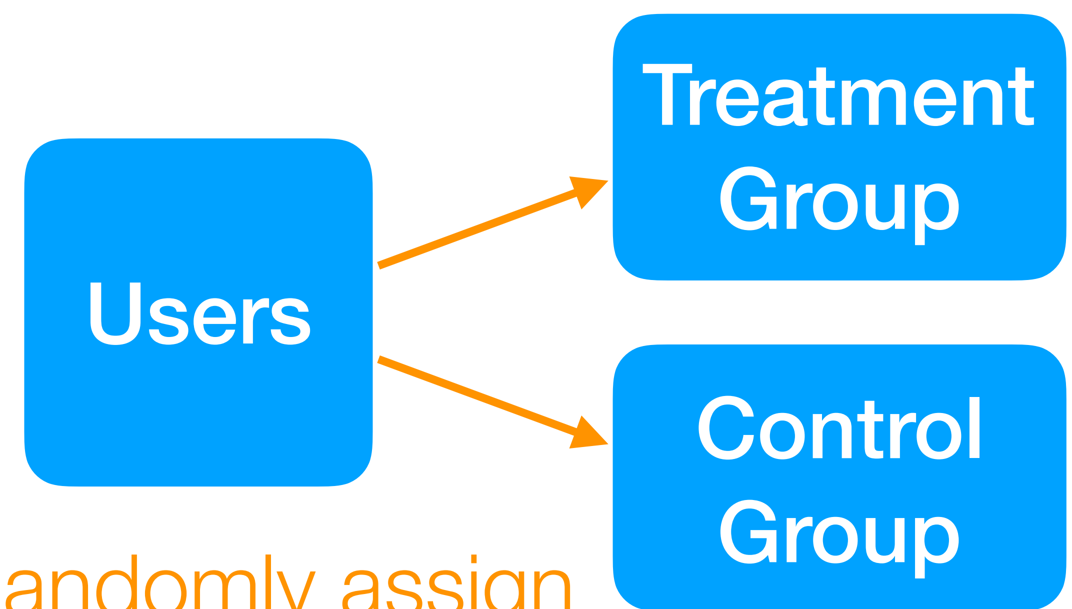
Establishing Causality

If you control data collection



Establishing Causality

If you control data collection

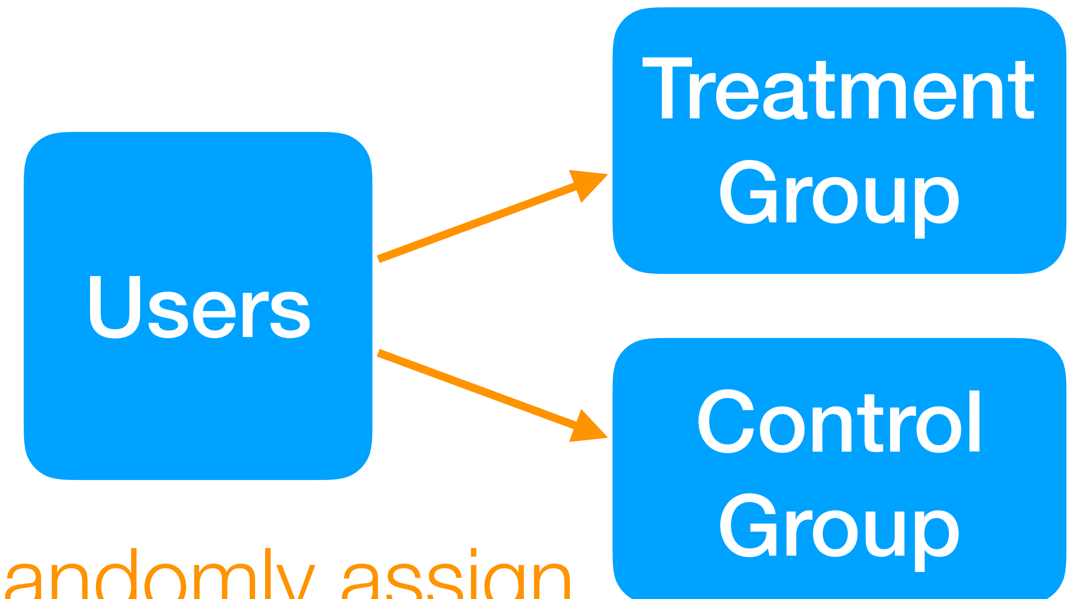


Randomly assign

Compare outcomes of two groups

Establishing Causality

If you control data collection



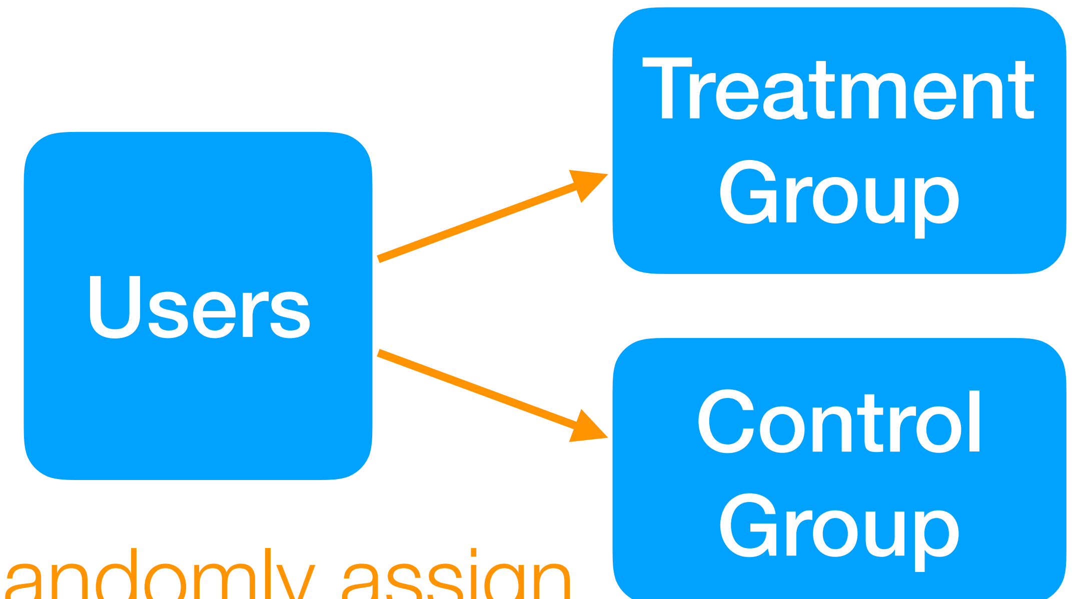
Randomly assign

Compare outcomes of two groups

Randomized controlled trial (RCT)
also called **A/B testing**

Establishing Causality

If you control data collection



Randomly assign

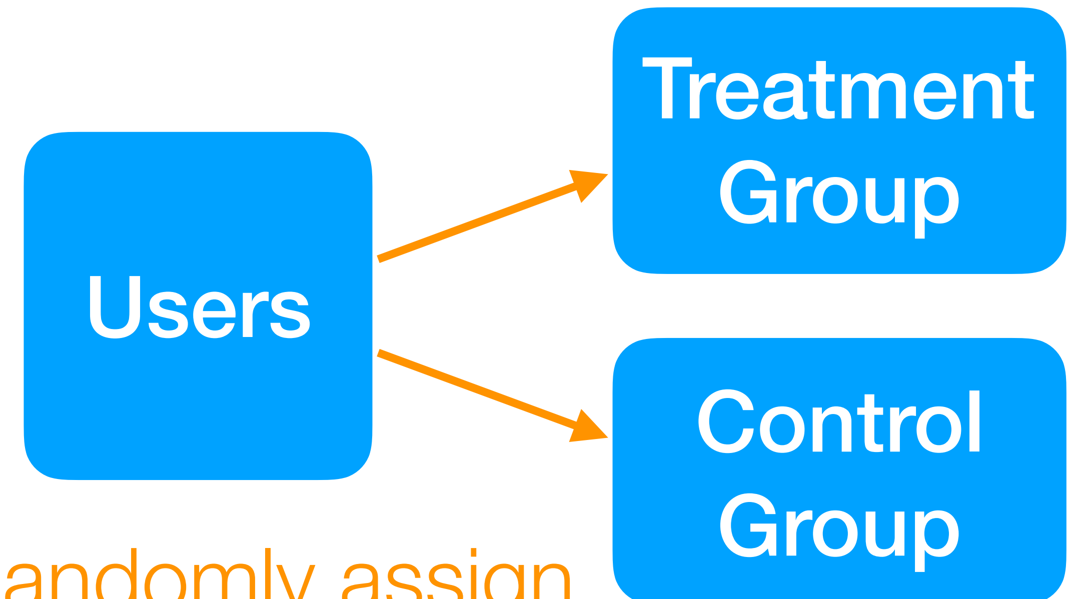
Compare outcomes of two groups

Randomized controlled trial (RCT)
also called **A/B testing**

Example: figure out webpage layout to maximize revenue (Amazon)

Establishing Causality

If you control data collection



Randomly assign

Compare outcomes of two groups

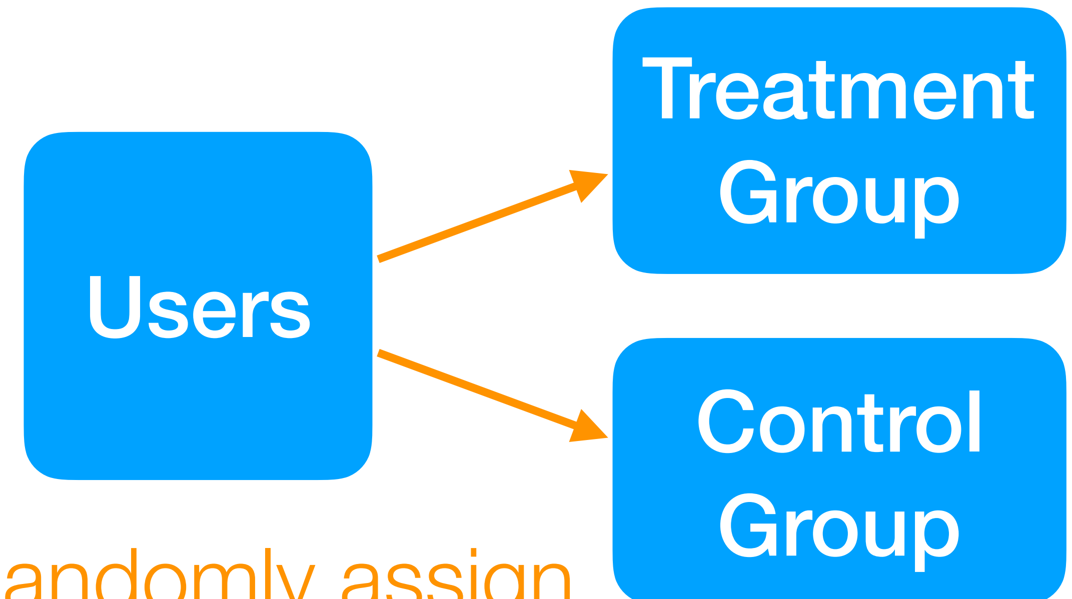
Randomized controlled trial (RCT)
also called **A/B testing**

Example: figure out webpage layout to maximize revenue (Amazon)

Example: figure out how to present educational material to improve learning (Khan Academy)

Establishing Causality

If you control data collection



Randomly assign

Compare outcomes of two groups

Randomized controlled trial (RCT)
also called **A/B testing**

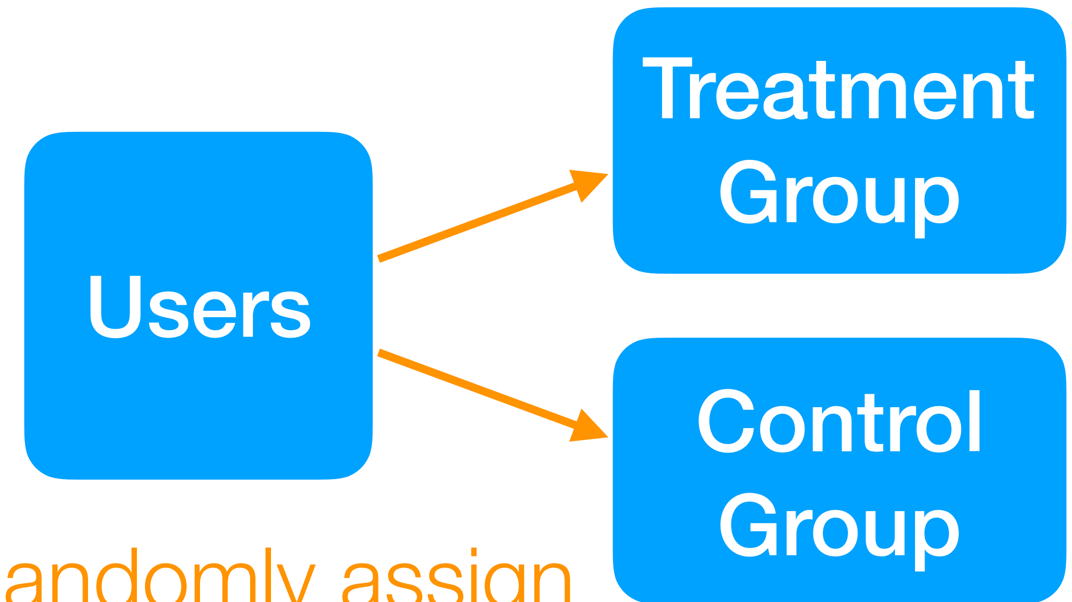
Example: figure out webpage layout to maximize revenue (Amazon)

Example: figure out how to present educational material to improve learning (Khan Academy)

If you do not control data collection

Establishing Causality

If you control data collection



Randomly assign

Compare outcomes of two groups

Randomized controlled trial (RCT)
also called **A/B testing**

Example: figure out webpage layout to maximize revenue (Amazon)

Example: figure out how to present educational material to improve learning (Khan Academy)

If you do not control data collection

In general: *not* obvious establishing what caused what

Course Outline (Tentative)

Part 1: Identify structure present in "unstructured" data

Exploratory data analysis

- Frequency and co-occurrences
- Clustering
- Topic modeling (special kind of clustering)

Unsupervised learning

Part 2: Make predictions using structure found in part 1

Predictive data analysis

Supervised learning

- Basic classification and regression models
- Adaptive nearest neighbor methods
- Deep learning models for classification

Course Outline (Tentative)

Part 1: Identify structure present in “unstructured” data

Exploratory data analysis

- Frequency and co-occurrences *Basic probability theory & stats*
- Clustering
- Topic modeling (special kind of clustering)

Unsupervised learning

Part 2: Make predictions using structure found in part 1

Predictive data analysis *Supervised learning*

- Basic classification and regression models
- Adaptive nearest neighbor methods
- Deep learning models for classification

We just did a crash course on basic probability theory and statistics

We just did a crash course on basic probability theory and statistics

What is the difference between probability theory and statistics?

Bag of words model:

Bag of words model:

- Suppose we know how many cards are in the bag with each token/symbol

Bag of words model:

- Suppose we know how many cards are in the bag with each token/symbol
- Can reason about probability of drawing different outcomes

Bag of words model:

- Suppose we know how many cards are in the bag with each token/symbol
- Can reason about probability of drawing different outcomes

"Fitting the model":

Bag of words model:

- Suppose we know how many cards are in the bag with each token/symbol
- Can reason about probability of drawing different outcomes

"Fitting the model":

- We set the number of cards of each token/symbol based on *observed frequencies in the data*

Probabilistic model

~~Bag of words~~ model:

- Suppose we know how many cards are in the bag with each token/symbol
- Can reason about probability of drawing different outcomes

"Fitting the model":

- We set the number of cards of each token/symbol based on *observed frequencies in the data*

Probabilistic model

~~Bag of words model~~:

- Suppose we know ~~how many cards are in the bag with each token/symbol~~ the model parameters
- Can reason about probability of drawing different outcomes

"Fitting the model":

- We set the number of cards of each token/symbol based on *observed frequencies in the data*

Probabilistic model

~~Bag of words model~~:

- Suppose we know ~~how many cards are in the bag with each token/symbol~~ the model parameters
- Can reason about probability of drawing different outcomes

"Fitting the model":

- ~~We set the number of cards of each token/symbol based on *observed frequencies in the data*~~

In general: often not as simple as using frequencies in the data

Probabilistic model

~~Bag of words model~~:

- Suppose we know ~~how many cards are in the bag with each token/symbol~~ the model parameters
- Can reason about probability of drawing different outcomes

"Fitting the model":

- ~~We set the number of cards of each token/symbol based on *observed frequencies in the data*~~

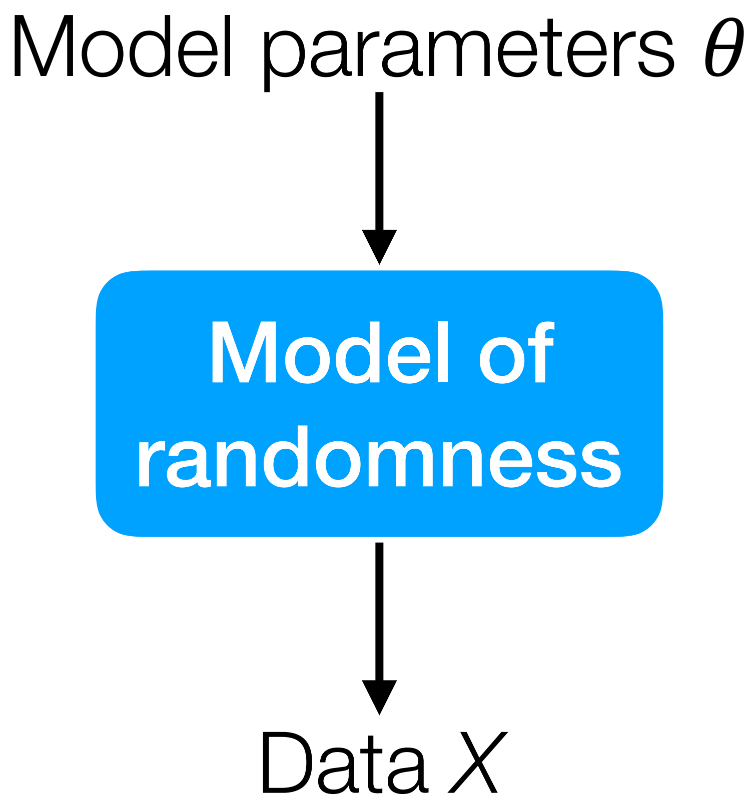
In general: often not as simple as using frequencies in the data

Also: how do we know unigram bag of words is the "right" model?

Probability Theory vs Statistics

Probability Theory vs Statistics

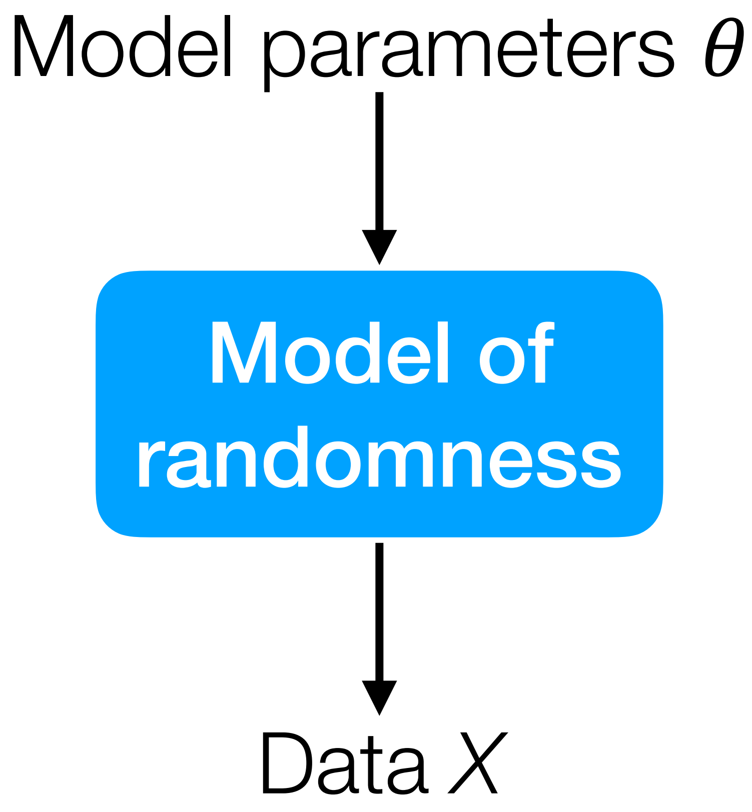
Probabilistic model



Probability Theory vs Statistics

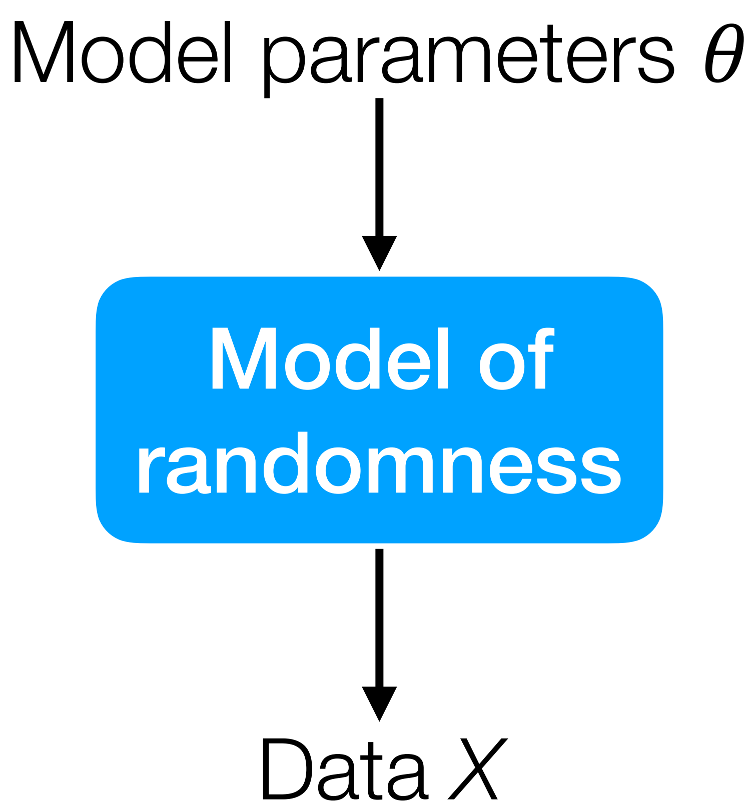
Probability theory:

Probabilistic model



Probability Theory vs Statistics

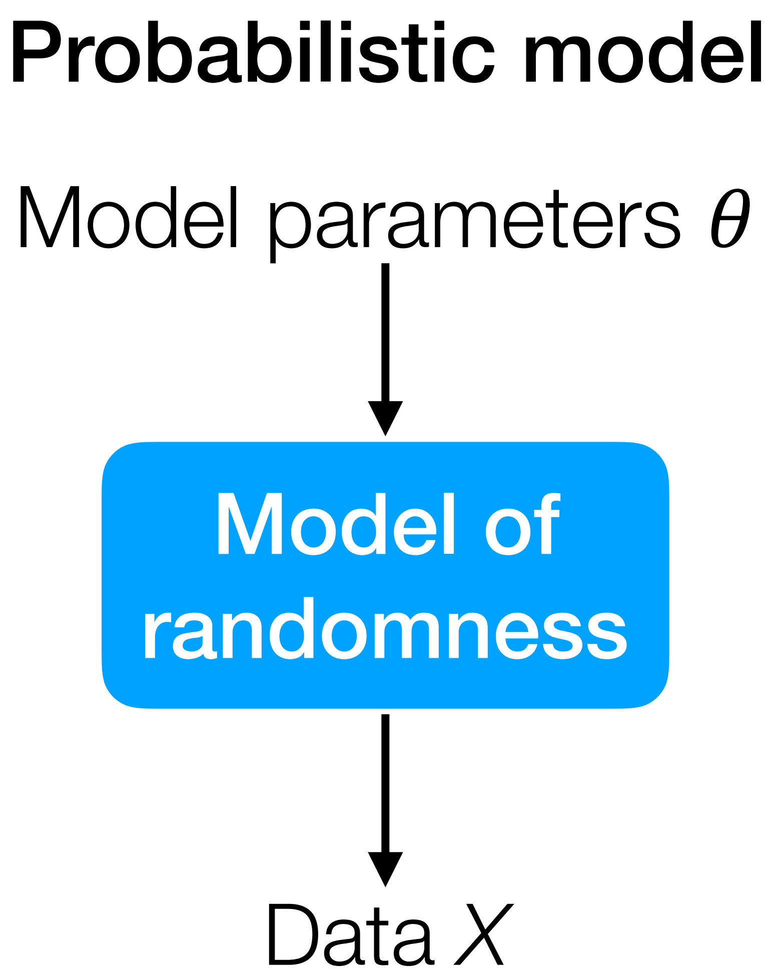
Probabilistic model



Probability theory:

- Assume we know model of randomness and parameters θ

Probability Theory vs Statistics

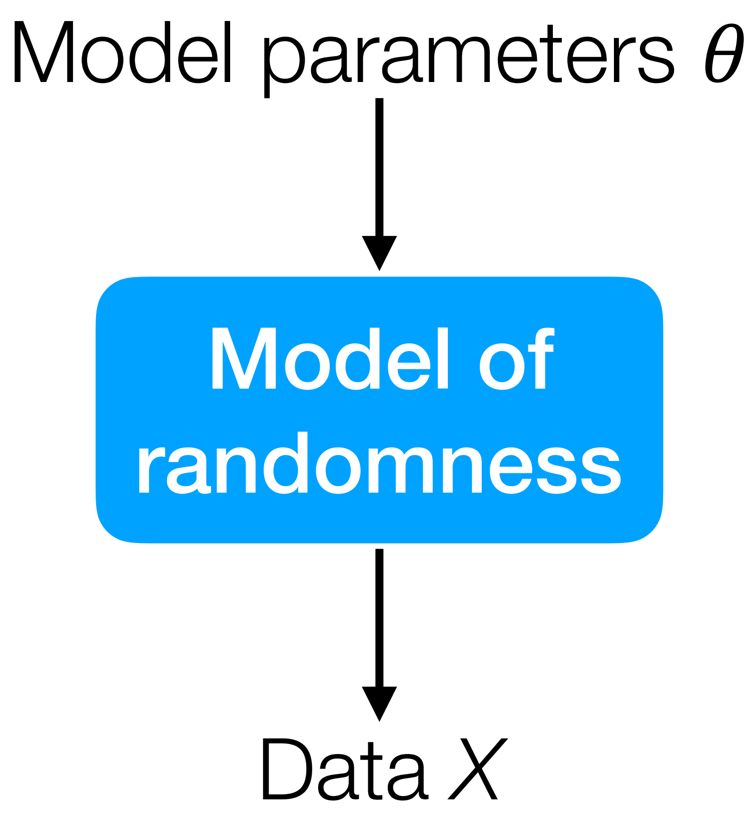


Probability theory:

- Assume we know model of randomness and parameters θ
- Reason about what happens in the model, what data X look like

Probability Theory vs Statistics

Probabilistic model



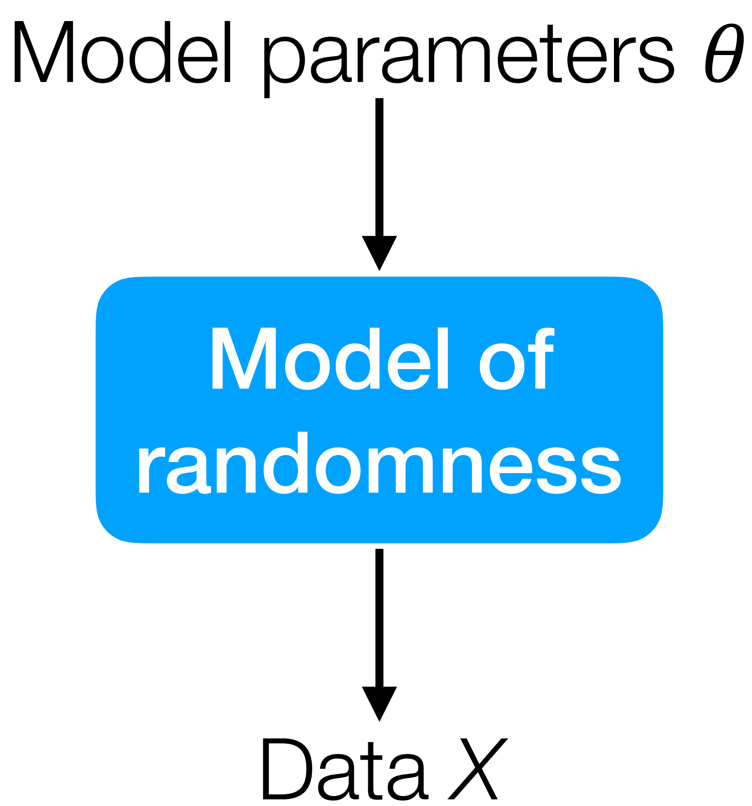
Probability theory:

- Assume we know model of randomness and parameters θ
- Reason about what happens in the model, what data X look like

Statistics:

Probability Theory vs Statistics

Probabilistic model



Probability theory:

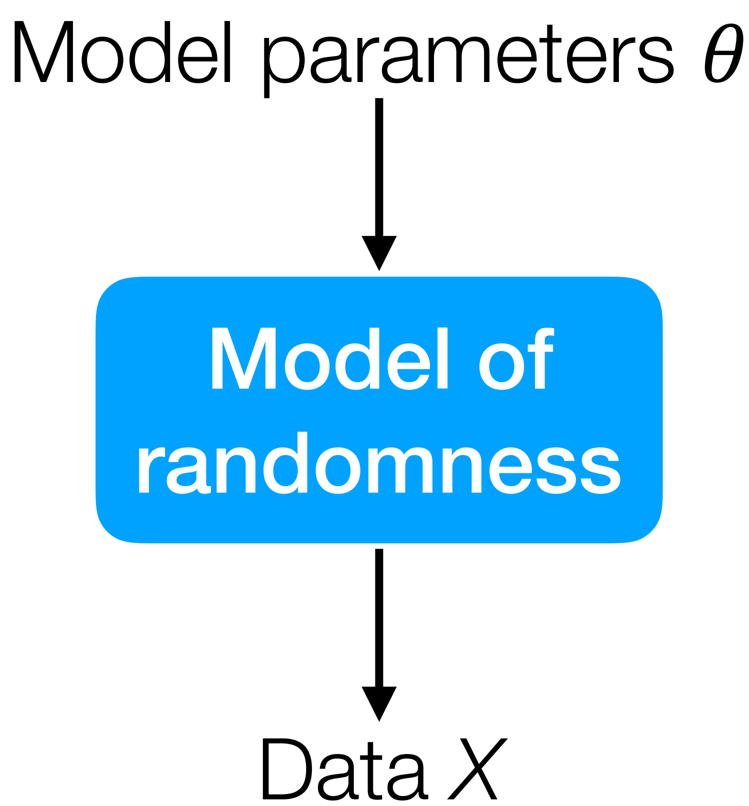
- Assume we know model of randomness and parameters θ
- Reason about what happens in the model, what data X look like

Statistics:

- Assume we collect data X

Probability Theory vs Statistics

Probabilistic model



Probability theory:

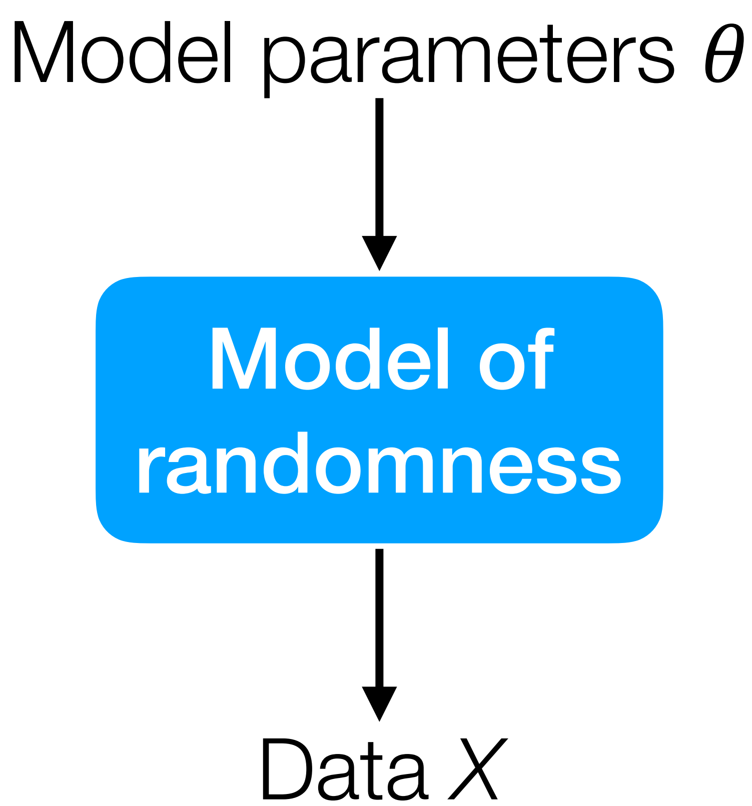
- Assume we know model of randomness and parameters θ
- Reason about what happens in the model, what data X look like

Statistics:

- Assume we collect data X
- Reason about what model of randomness makes sense, and what its parameters θ are

Probability Theory vs Statistics

Probabilistic model



Probability theory:

- Assume we know model of randomness and parameters θ
- Reason about what happens in the model, what data X look like

Statistics:

- Assume we collect data X
- Reason about what model of randomness makes sense, and what its parameters θ are

We will be seeing these ideas a lot in this course!