Data mining & Machine Learning

CS 373
Purdue University

Dan Goldwasser
dgoldwas@purdue.edu

Why is learning possible?

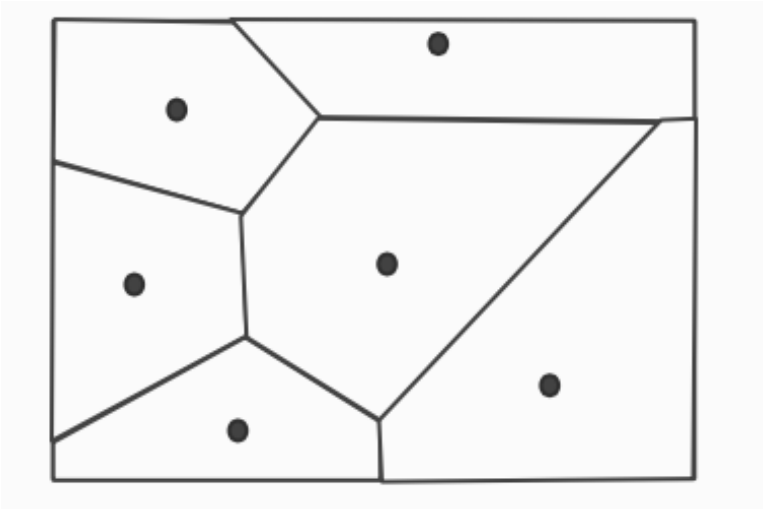
- **Learning is removal of remaining uncertainty**
  - If we know that the function is a "m-out-of-n", data can help find a function from that class
- **Finding a good hypothesis class is essential!**
  - You can start small, and enlarge it until you can find a hypothesis that fits the data

Question: *Can there be more than one function that is consistent with the data*?

How do you choose between them?

KNN analysis

- We discussed the importance of the model space
  - Expressive (we can represent the right model)
  - Constrained (we can search effectively, using available data)
- Let's try to characterize the model space, by looking at the **decision boundary**
- **How would it look if K=1?**

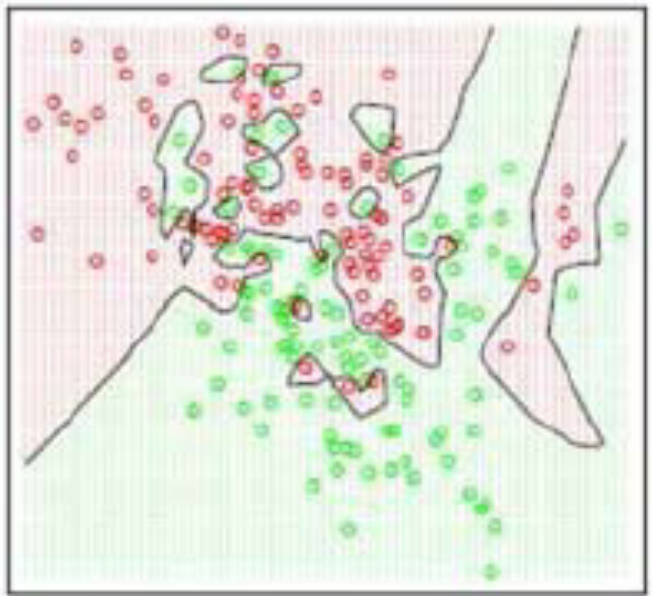


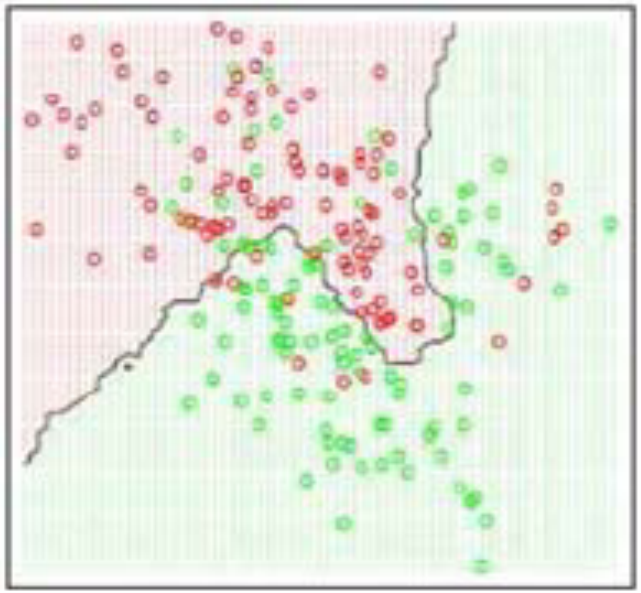
We define the model space to be our choice of K.

Does the complexity of the model space increase of decrease with K?

KNN analysis

- Which model has a higher K value?
- Which model is more complex?
- Which model is more sensitive to noise?





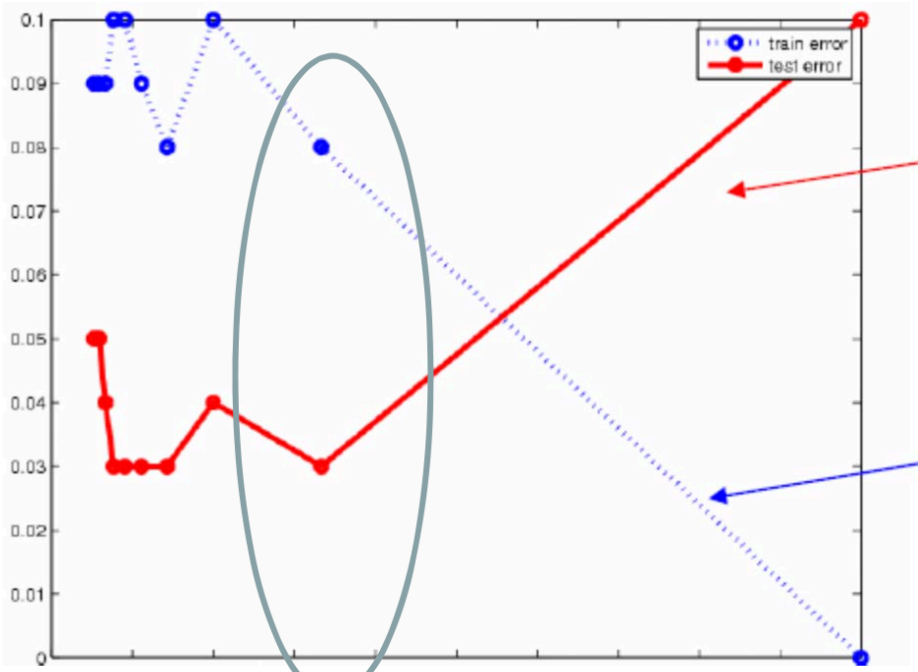
Determining the value of K

- Higher K result in less complex functions (less expressive)
- Lower K values are more complex (more expressive)
  - **How can we find the right balance between the two?**
- **Option 1**: *Find the K that minimizes the training error*.
  - Training error: after learning the classifier, what is the number of errors we get on the training data.
  - What will be this value for k=1, k=n, k=n/2?

Is this a good idea?

- **Option 2**: *Find K that minimizes the **validation error***.
  - Validation error: set aside some of the data (validation set). what is the number of errors we get on the validation data, after training the classifier.

Determining the value of K



Test error

Train error

k=20

k=1

Possibly a good choice for k

In general – using the training error to tune parameters will always result in a more complex hypothesis! **(why?)**

Questions

- Is KNN a supervised or unsupervised learning algorithm?

- We defined learning as search, ”where” is the search in KNN?

Today's Lecture

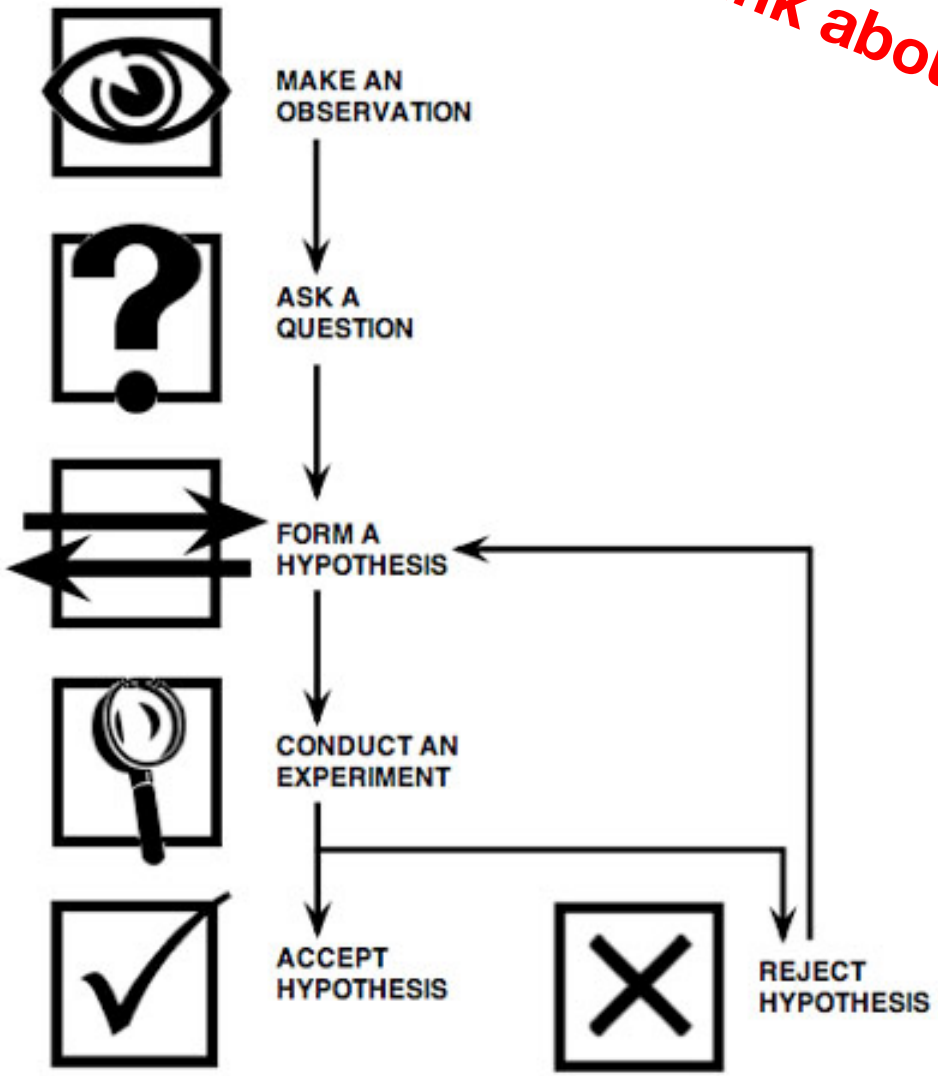
*Now that we understand **machine hypotheses** better, let's go back to **hypotheses humans make***

- *We all see data, **all the time.***
- *We all reach conclusions based on it, **all the time.***

Can we trust this mechanism?

Primer on hypotheses

Let's think about an example!



What is a hypothesis?

- **Hypotheses** are tentative statements of the expected relationships between two or more variables
  - **Inductive** hypotheses are formed through inductively reasoning from many specific observations to tentative explanations (*bottom-up*)
  - **Deductive** hypotheses are formed through deductively reasoning implications of theory (*top-down*)

- **Reasons for using hypotheses**
  - Provides focus and directs research investigation
  - Allows the investigator to confirm or not confirm relationships
  - Provides a useful framework for organizing and summarizing results and conclusions

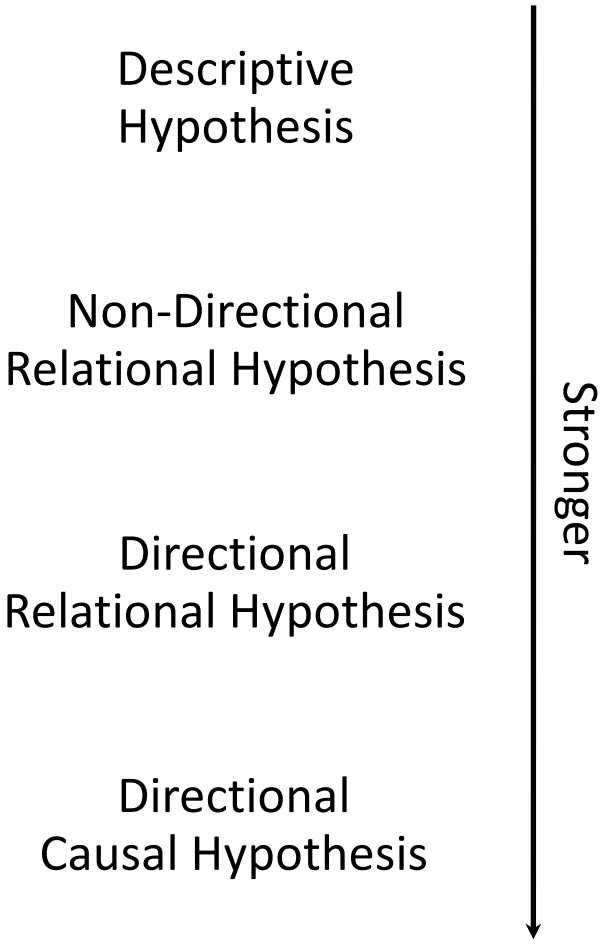
Types of hypotheses

Broad categories

- **Descriptive**: propositions that describe a characteristic of an object
- **Relational**: propositions that describe the relationship between 2+ variables
- **Causal**: propositions that describe the effect of one variable on another

Specific characteristics

- **Non-directional**: a differential outcome is anticipated but the specific nature of it is not known (e.g., the tuning parameter will affect algorithm performance)
- **Directional**: a specific outcome is anticipated (e.g., the use of pruning will increase accuracy of models compared to no pruning)



From claims to testable hypotheses

- *Ever since 1980, when Ronald Reagan inspired more men than women, the difference in the way men and women vote has been a significant part of American politics.*
- **Step 1**: Express data as random variables

$$X := gender$$
$$Y := \%\ voted\ Democrat$$

- **Step 2**: Restate claim as a hypothesis about the relationship between the random variables, e.g.,
  - (X=male) is associated with smaller Y
- **Step 3**: Determine type of hypothesis (and consider whether you can make it stronger), e.g.,
  - Directional-relational

From claims to testable hypotheses

- *Ever since 1980, when Ronald Reagan inspired more men than women, the difference in the way men and women vote has been a significant part of American politics.*
- **Types of hypotheses**:
  - ***Descriptive***: Voting practices vary throughout the population (i.e., Y varies).
  - ***Non-directional relational***: Voting behavior varies based on gender (i.e., X and Y are associated)
  - ***Directional-relational***: Women vote more for democrats (i.e., X=female is associated with larger Y)
  - ***Causal-relational***: Women vote more for democrats because they are more likely to be poor (i.e., X=female is associated with larger Y, but if you control for salaries this effect may disappear)

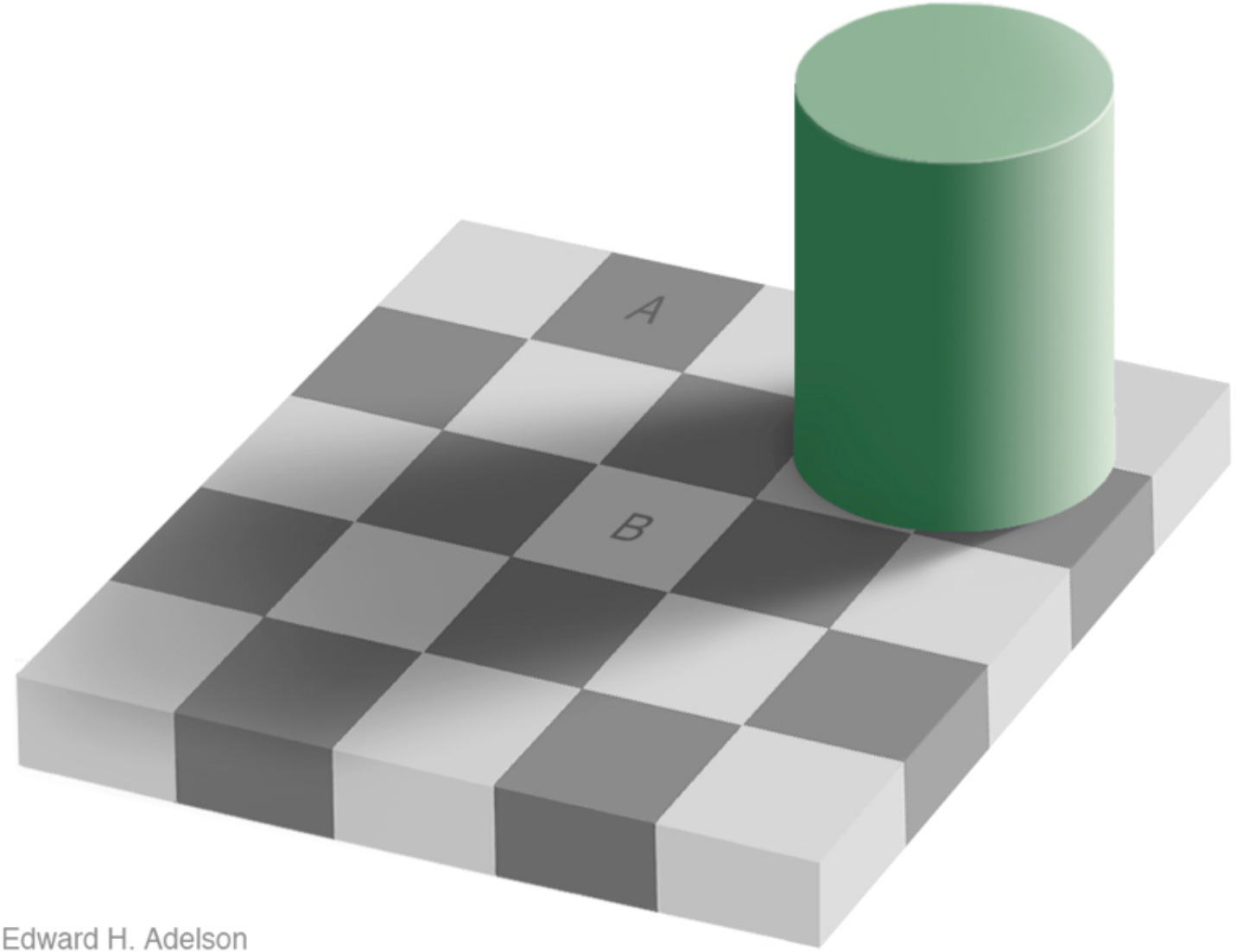
Example

- *The Princeton researchers say the experiments suggest that high-fructose corn syrup prompts more weight gain than sucrose, at least in rats, even when the animals eat the same number of calories over all.*
  - Define the random variables:
  - *Descriptive*:
  - *Non-directional relational*:
  - *Directional-relational*:
  - *Causal-relational*:

https://www.sciencedaily.com/releases/2010/03/100322121115.htm

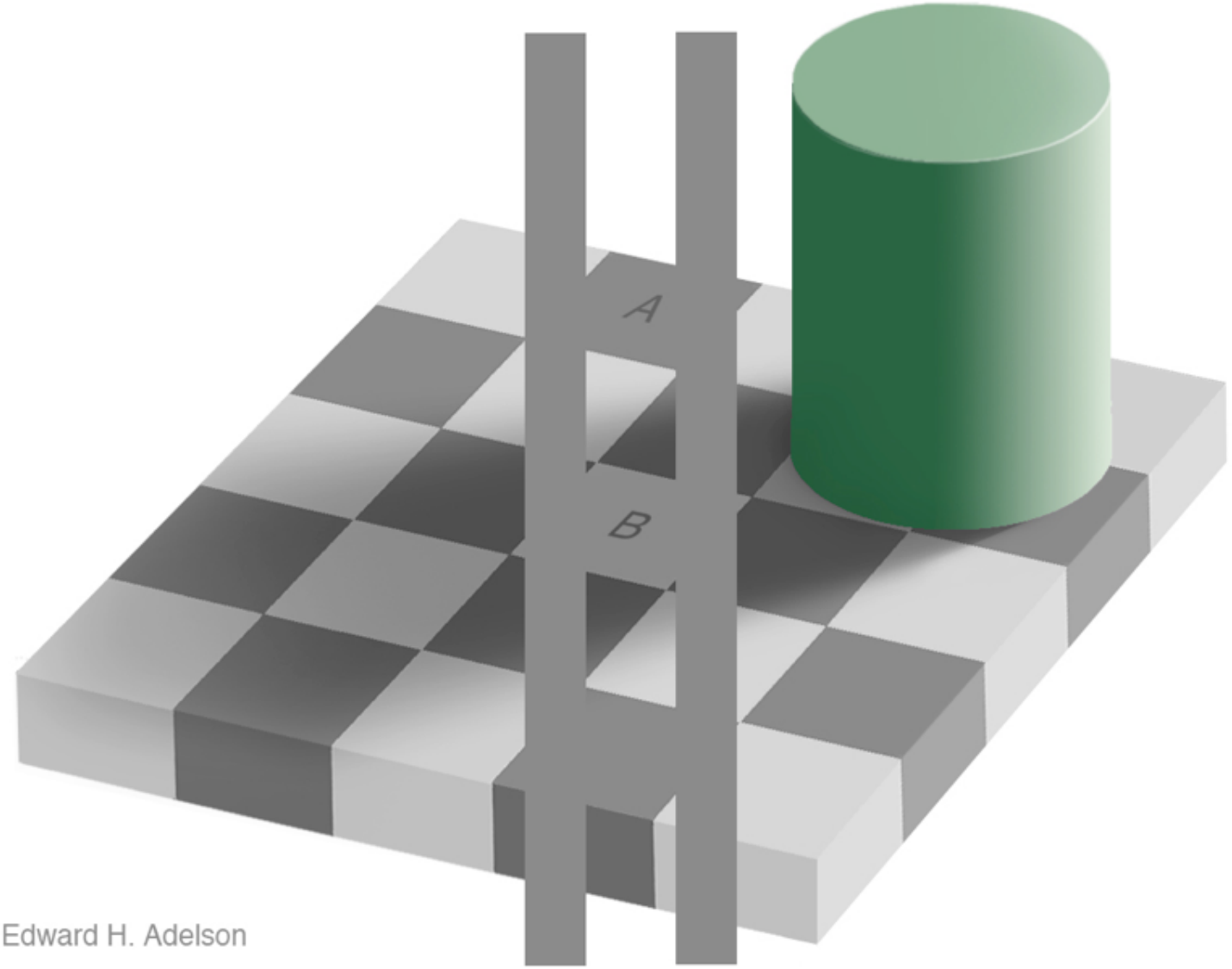
Decision making

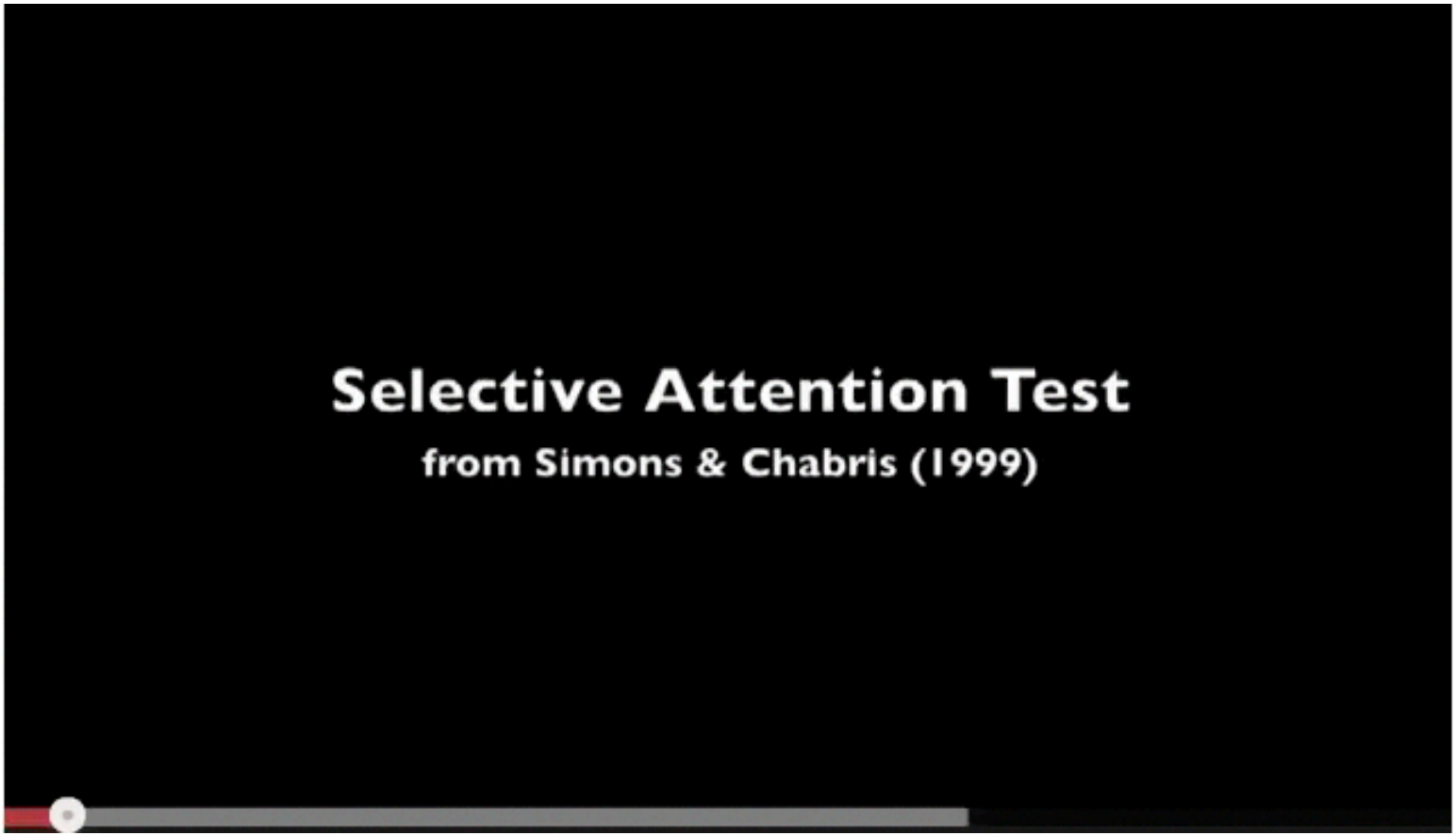
Are A and B the same color?



Edward H. Adelson

The trick uses the “biases” in the human visual system




Selective Attention Test
from Simons & Chabris (1999)

Heuristics and biases

- Tversky & Kahneman, psychologists, propose that people often do not follow rules of probability when making decisions
- Instead, decision making may be based on heuristics
  - Lowers cognitive load but may lead to systematic errors and biases

Can you find an example of such heuristics?

- Examples:
  - Representativeness heuristic
  - Availability heuristic
  - Confirmation bias
  - Conjunction fallacy (we will not cover this)
  - Numerosity heuristic (we will not cover this)

Gambler's fallacy

- **Definition:** belief that if deviations from expected behavior are observed in repeated independent trials, then future deviations in the opposite direction are then more likely
- This is an example of the **representativeness heuristic**—where the probability of an event is judged by its similarity to the population from which sample is drawn
- The sequence "H T H T T H" is seen as more representative of a prototypical coin sequence. **Why**?
  - When people are asked to make up random sequences, they tend to make the proportion of H and T closer to 50% than would be expected by random chance
- **T&K interpretation**: *people believe that short sequences should be representative of longer ones*

Base Rate Study (Kahneman & Tversky '73)

- Participants told that for a set of 100 people are either:
  - 30% engineers/70% lawyers, or
  - 70% engineers/30% lawyers
- *Given*: A description of a person Jack, which is representative of a prototypical engineer (e.g., likes carpentry and mathematical puzzles, careful, conservative)
- *Question*: Is Jack more likely to be a lawyer or engineer?
- *Results*: Participants in the 30% condition judged Jack just as likely to be an engineer as participants in the 70% condition.

Base rate study (cont)

- People use the representative heuristic to make inferences...
  - Inferences is based solely on similarity of target to category members
  - Base rates (70%-30%) are ignored
- ...rather than using formal statistical rules to make inferences
  - Inferences should be based on similarity of target to category members AND base rates (70%-30%)

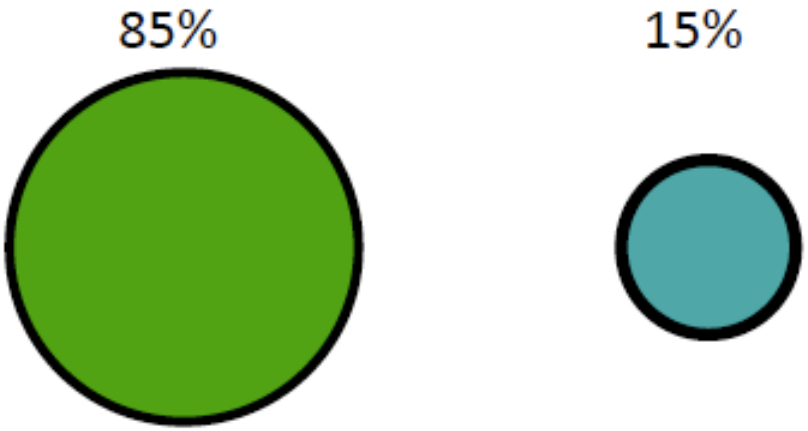
- **Representative heuristic**: categorizations made on the basis of similarity between instance and category members

Neglecting base rates

- Taxi-cab problem *(Tversky & Kahneman '72)*
  - 85% of the cabs are Green
  - 15% of the cabs are Blue
  - An accident eyewitness reports a Blue cab
  - But she is wrong 20% of the time.
- What is the probability that the cab is Blue?
  - Participants tend to overestimate probability, most answer 80%
  - They ignore baseline prior probability of blue cabs.

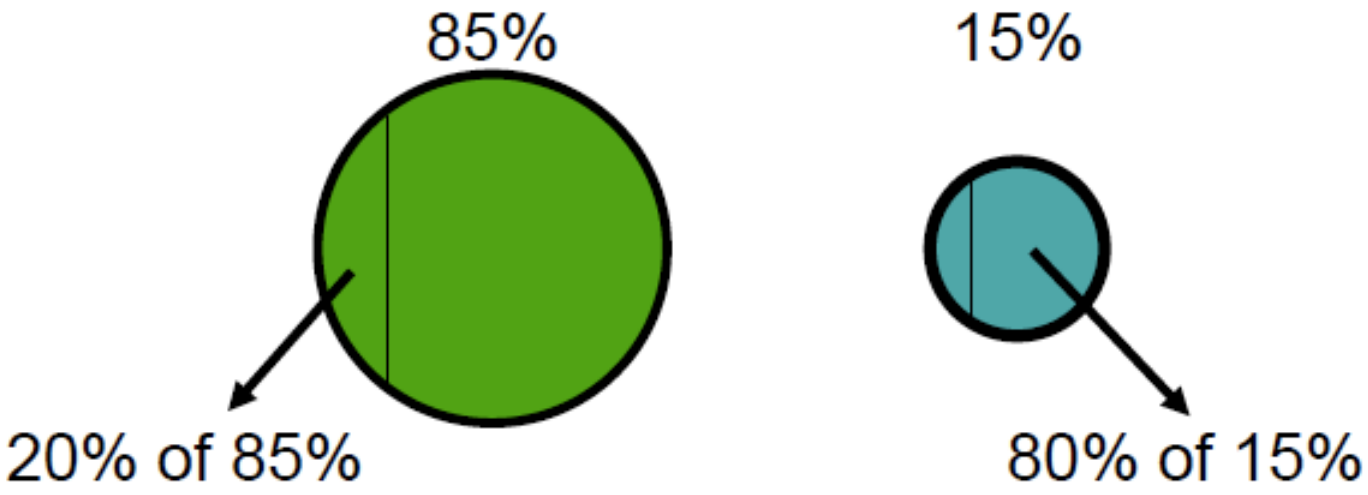
A priori (beforehand)

$P(green) = 0.85$
$P(blue) = 0.15$



$P(seeBlue|blue) = 0.80$
$P(seeBlue|green) = 0.20$

After accident (only cars reported as being blue)



More on neglecting base rates

- How to compute probability

$$
\begin{aligned}
P(blue|seeBlue) &= \frac{P(blue \wedge seeBlue)}{P(seeBlue)} \\
&= \frac{P(seeBlue|blue)P(blue)}{P(seeBlue)} \\
&= \frac{P(seeBlue|blue)P(blue)}{P(seeBlue|blue)P(blue) + P(seeBlue|green)P(green)} \\
&= \frac{0.80 \cdot 0.15}{(0.80 \cdot 0.15) + (0.20 \cdot 0.85)} \\
&= 0.41
\end{aligned}
$$

Most people answered 80%

TheUpshot

EDITED BY DAVID LEONHARDT

A Quick Puzzle to Test Your Problem Solving

By **DAVID LEONHARDT** and **YOU** JULY 2, 2015

A short game sheds light on government policy, corporate America and why no one likes to be wrong.

Here's how it works:

We've chosen a rule that some sequences of three numbers obey — and some do not. Your job is to guess what the rule is.

We'll start by telling you that the sequence 2, 4, 8 obeys the rule:

2	4	8	Obeys the rule

Now it's your turn. Enter a number sequence in the boxes below, and we'll tell you whether it satisfies the rule or not. You can test as many sequences as you want.

Enter your first sequence here:

			Check

I don't want to play; just tell me the answer.

Arthritis study (Redelmeier & Tversky '96)

- Common belief:
  - **Arthritis pain is associated with changes in weather**
- Experiment:
  - Followed 18 arthritis patients for 15 months
  - 2 x per month assessed: (1) pain and joint tenderness, and (2) weather
- Results:
  - No correlation between pain/tenderness and weather
  - Patients saw correlation that did not exist... why?

Arthritis study (cont)

- Patients noticed when bad weather and pain co-occurred, but failed to notice when they didn't.
  - Better memory for times that bad weather and pain co-occurred.
  - Worse memory for times when bad weather and pain did not co-occur
- **Confirmation bias**: People often seek information that **confirms** rather than disconfirms their original hypothesis

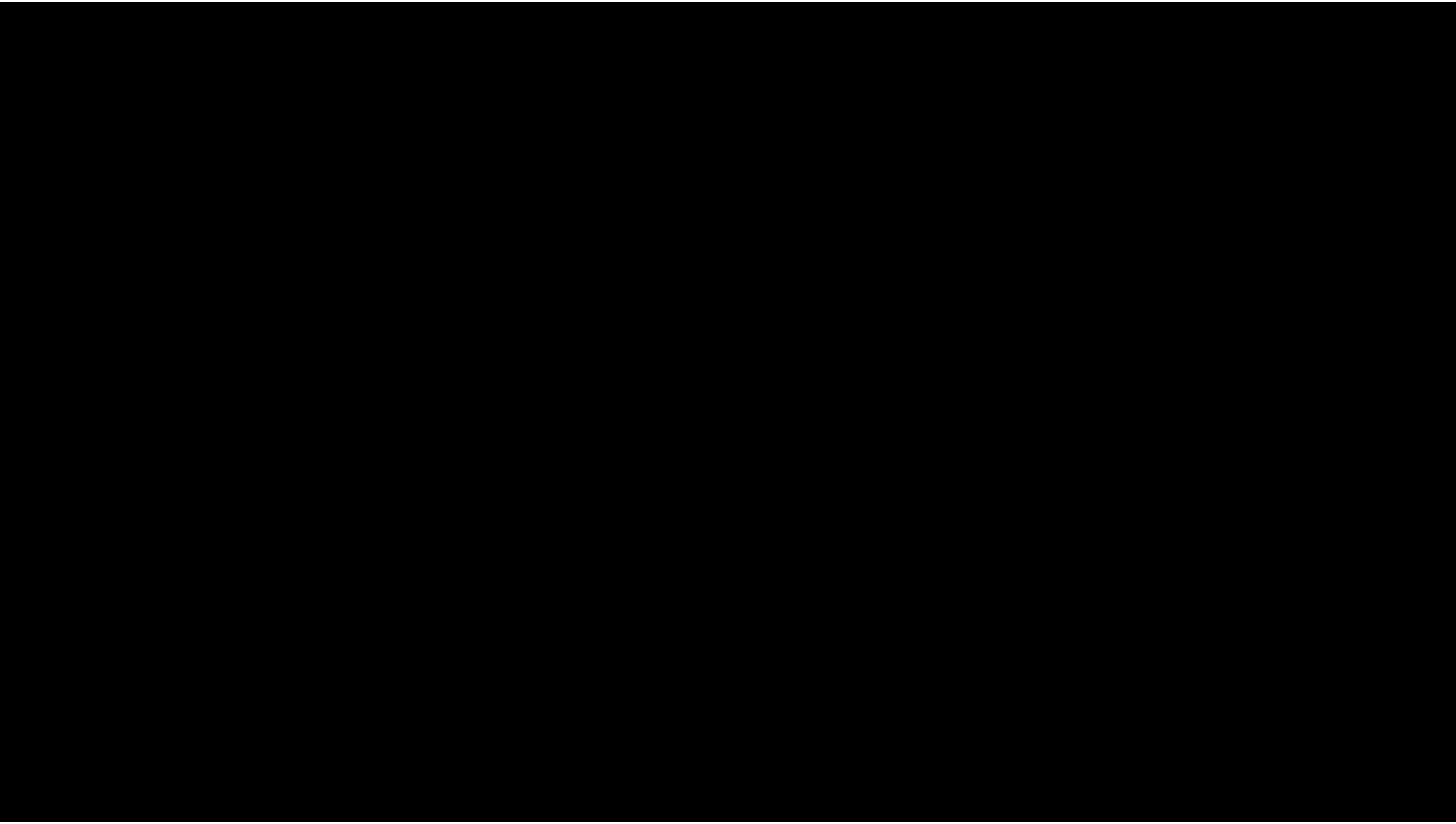
Estimating probabilities (Tversky & Kahneman '73/'74)

- ***Question***: *Is the letter **R** more likely to be the 1st or 3rd letter in English words?*
- ***Results***: *Most said **R** more probable as 1st letter*
- ***Reality***: ***R** appears much more often as the 3rd letter, but it's easier to think of words where **R** is the 1st letter*

Estimating probabilities (cont)

- ***Question***: Which causes more deaths in developed countries?
  (a) traffic accidents or (b) stomach cancer
- ***Typical guess***: *traffic accident = 4X stomach cancer*
- ***Actual***: *45,000 traffic, 95,000 stomach cancer deaths in US*
- **Ratio of newspaper reports on each subject:**
  137 (traffic fatality) to 1 (stomach cancer death)
- **Availability heuristic**: Tendency for people to make judgments of frequency on basis of how easily examples come to mind

Pie math

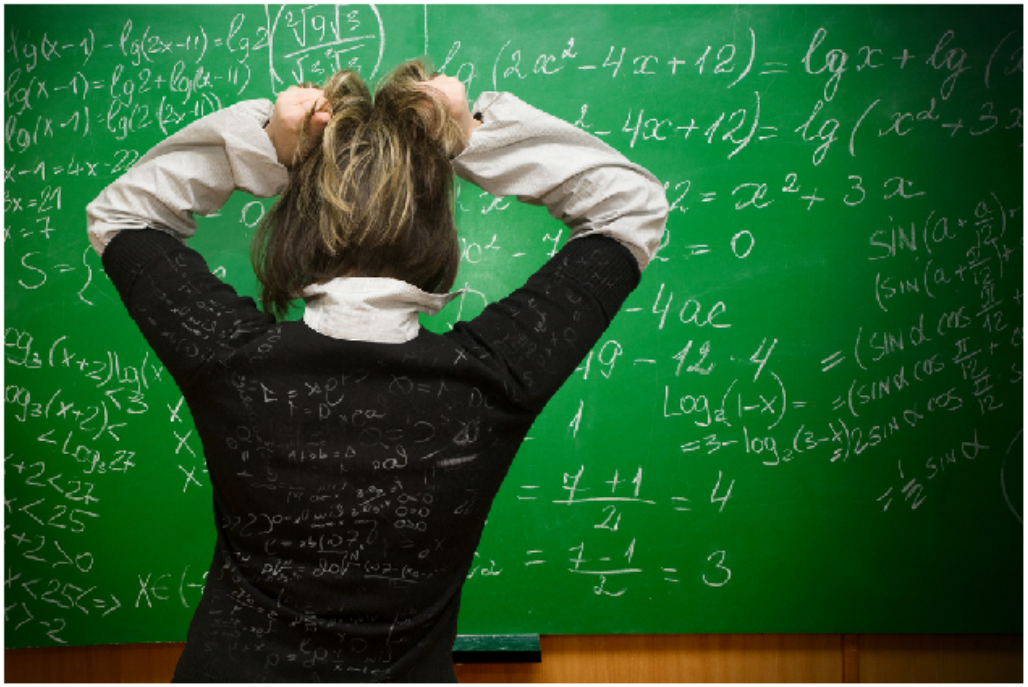


POLITICS
→ *Climate Change, Climate Desk, Guns, Science*

Science Confirms: Politics Wrecks Your Ability to Do Math

Farewell, Enlightenment: New research suggests that people even solve math problems differently if their political ideology is at stake.

—By **Chris Mooney** | Wed Sep. 4, 2013 12:59 PM EDT



A new study finds that even how you solve a difficult math problem can depend on your politics. *AlenKadr/Shutterstock*

Everybody knows that our political views can sometimes get in the way of thinking clearly. But perhaps we don't realize how bad the problem actually is. According to a new psychology paper, our political passions can even undermine our very basic reasoning skills. More specifically, the study finds that people who are otherwise very good at math may totally flunk a problem that they would otherwise probably be able to solve, simply because giving the right answer goes against their political beliefs.

Kahan et al. (2013) "Motivated numeracy and enlightened self-government." *Social Science Research Network*.

1000+ participants were asked about their political views and also asked a series of questions to gauge their mathematical reasoning ability.

Participants were then asked to solve a fairly difficult problem that involved interpreting the results of a (fake) scientific study.

One group was given a problem involving the effectiveness of a new **skin cream**. The other group was given a mathematically similar problem, but the data involved the effectiveness of a **gun control** measure.

What was the result?
Highly numerate people were more susceptible to letting politics skew their reasoning than were those with less mathematical ability.

Does this look familiar?

- Economist Dan Arieli showed that we are willing to pay more for a product when something free is involved
  - ”we’ll throw in X for free, if you buy Y”
- Advertisers make sophisticated use of these biases!

Relevant TED talks

- Informal, but interesting!

- **Daniel Kahneman: The riddle of experience vs. memory**
- https://www.ted.com/talks/daniel_kahneman_the_riddle_of_experience_vs_memory

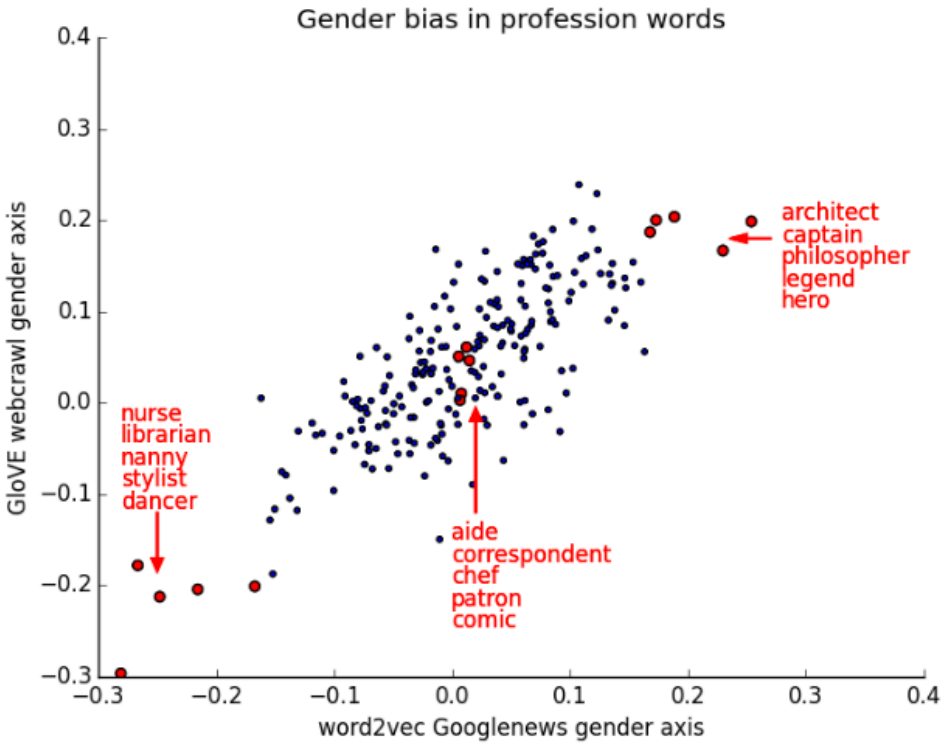
- **Dan Ariely: Are we in control of our own decisions?**
  - https://www.ted.com/talks/dan_ariely_asks_are_we_in_control_of_our_own_decisions

Interpretation of these findings

- People do not use proper statistical/probabilistic reasoning... instead people use heuristics which can **bias** decisions
- Heuristics can often be very effective (and efficient) for social inferences and decision-making
  - E.g., the book "Simple Heuristics That Make Us Smart" summarizes research by Gigerenzer and Todd
- ... but be aware that **heuristics can bias** results from **exploratory data analysis and other modeling efforts**

Is data the answer?

- Well, **yes and no..**



Quantifying and Reducing Stereotypes in Word Embeddings.
Bolukbasiet-al, 2016

Is data the answer?

Data driven analogies between concepts, based on word embedding

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}.$$

Gender stereotype *she-he* analogies.

sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber

Gender appropriate *she-he* analogies.

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

And now –

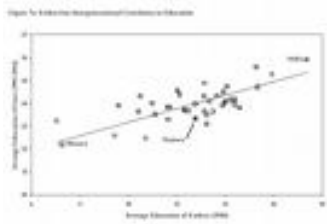
Diggin' into Data!

Measurement

Real world

Data

Relationship
in real world

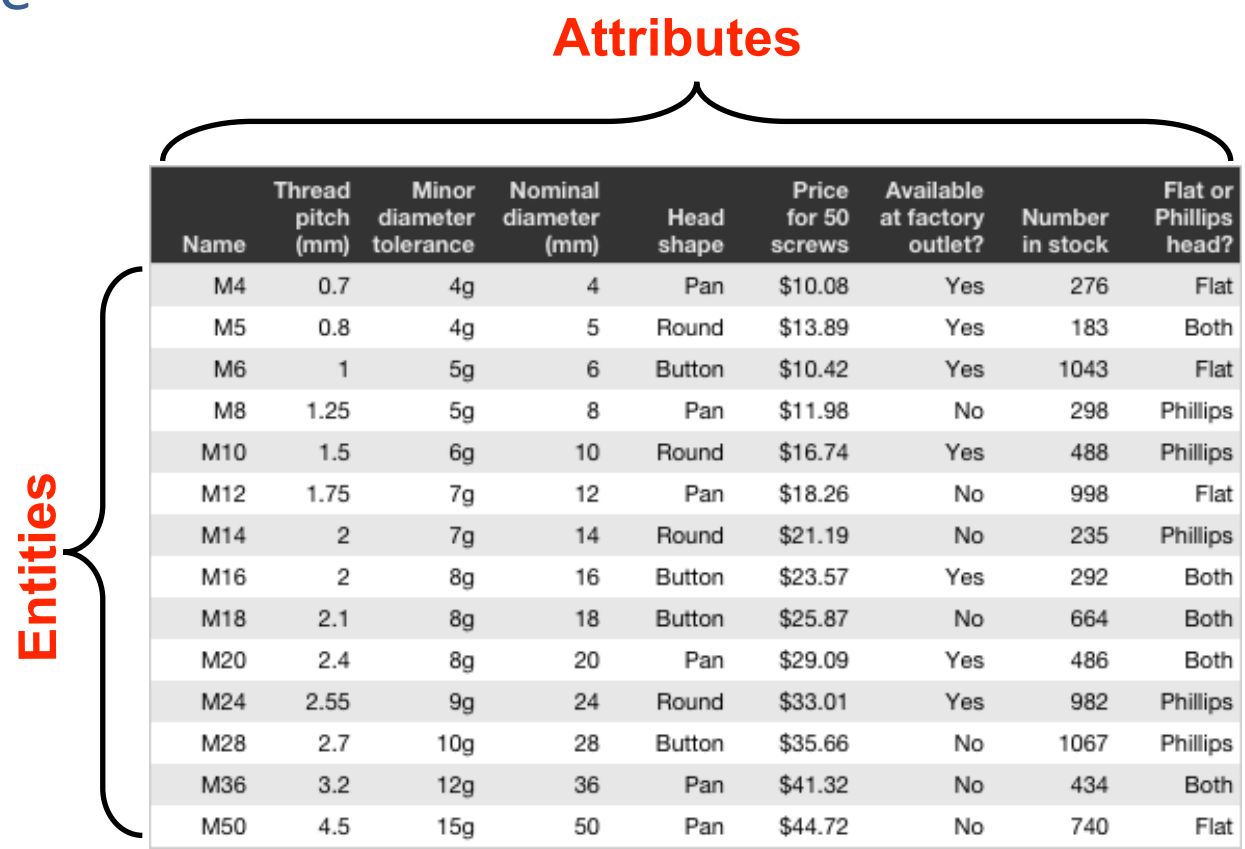


Relationship
in data

Goal: map domain entities to symbolic representations

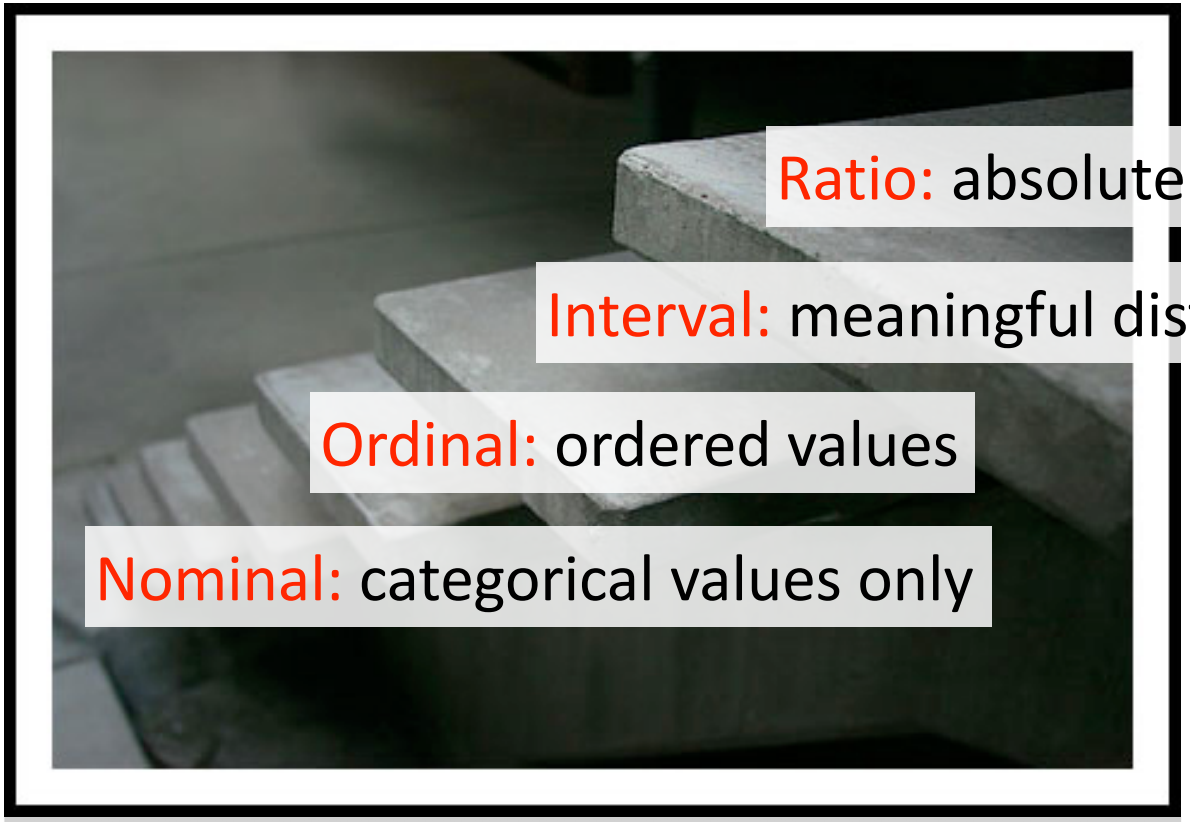
What is data?

- Collection of entities and their attributes
- **Attribute**: property or characteristic of an entity (e.g., eye color, temperature)
- **Entity**: collection of attributes
  Aka: record, point, case, sample, object, or instance



Name	Thread pitch (mm)	Minor diameter tolerance	Nominal diameter (mm)	Head shape	Price for 50 screws	Available at factory outlet?	Number in stock	Flat or Phillips head?
M4	0.7	4g	4	Pan	$10.08	Yes	276	Flat
M5	0.8	4g	5	Round	$13.89	Yes	183	Both
M6	1	5g	6	Button	$10.42	Yes	1043	Flat
M8	1.25	5g	8	Pan	$11.98	No	298	Phillips
M10	1.5	6g	10	Round	$16.74	Yes	488	Phillips
M12	1.75	7g	12	Pan	$18.26	No	998	Flat
M14	2	7g	14	Round	$21.19	No	235	Phillips
M16	2	8g	16	Button	$23.57	Yes	292	Both
M18	2.1	8g	18	Button	$25.87	No	664	Both
M20	2.4	8g	20	Pan	$29.09	Yes	486	Both
M24	2.55	9g	24	Round	$33.01	Yes	982	Phillips
M28	2.7	10g	28	Button	$35.66	No	1067	Phillips
M36	3.2	12g	36	Pan	$41.32	No	434	Both
M50	4.5	15g	50	Pan	$44.72	No	740	Flat

Hierarchy of measurements

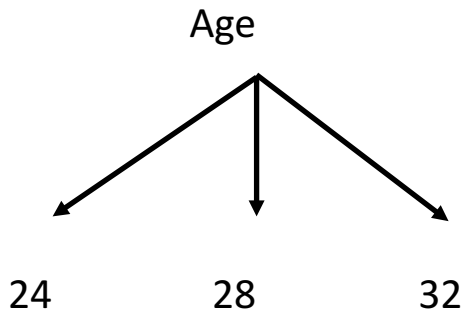


Discrete and continuous attributes

- Discrete
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, set of words in a collection of documents
  - Often represented as integer variables
- Continuous
  - Has real numbers as attribute values
  - Examples: temperature, height
  - Continuous attributes are typically represented as floating-point variables

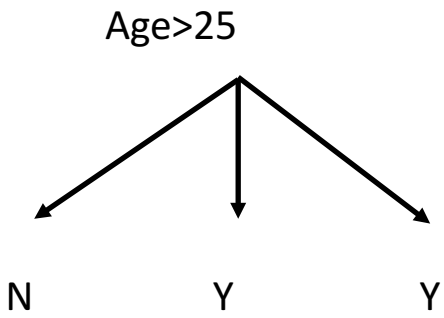
Naming conventions

Attribute/ variable



Values

Feature



Values

Tabular data

- Collection of records, each of which consists of a fixed set of attributes

Name	Thread pitch (mm)	Minor diameter tolerance	Nominal diameter (mm)	Head shape	Price for 50 screws	Available at factory outlet?	Number in stock	Flat or Phillips head?
M4	0.7	4g	4	Pan	$10.08	Yes	276	Flat
M5	0.8	4g	5	Round	$13.89	Yes	183	Both
M6	1	5g	6	Button	$10.42	Yes	1043	Flat
M8	1.25	5g	8	Pan	$11.98	No	298	Phillips
M10	1.5	6g	10	Round	$16.74	Yes	488	Phillips
M12	1.75	7g	12	Pan	$18.26	No	998	Flat
M14	2	7g	14	Round	$21.19	No	235	Phillips
M16	2	8g	16	Button	$23.57	Yes	292	Both
M18	2.1	8g	18	Button	$25.87	No	664	Both
M20	2.4	8g	20	Pan	$29.09	Yes	486	Both
M24	2.55	9g	24	Round	$33.01	Yes	982	Phillips
M28	2.7	10g	28	Button	$35.66	No	1067	Phillips
M36	3.2	12g	36	Pan	$41.32	No	434	Both
M50	4.5	15g	50	Pan	$44.72	No	740	Flat

Document data

- Each document is represented as a **term** vector, where each attribute records the number of times the term occurs in the document

Terms	Documents													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14
abnormalities	0	0	0	0	0	0	0	1	0	1	0	0	0	0
age	1	0	0	0	0	0	0	0	0	0	0	1	0	0
behavior	0	0	0	0	1	1	0	0	0	0	0	0	0	0
blood	0	0	0	0	0	0	0	1	0	0	1	0	0	0
close	0	0	0	0	0	0	1	0	0	0	1	0	0	0
culture	1	1	0	0	0	0	0	1	1	0	0	0	0	0
depressed	1	0	1	1	1	0	0	0	0	0	0	0	0	0
discharge	1	1	0	0	0	1	0	0	0	0	0	0	0	0
disease	0	0	0	0	0	0	0	0	1	0	1	0	0	0
fast	0	0	0	0	0	0	0	0	0	1	0	1	1	1
generation	0	0	0	0	0	0	0	0	1	0	0	0	1	0
oestrogen	0	0	1	1	0	0	0	0	0	0	0	0	0	0
patients	1	1	0	1	0	0	0	1	0	0	0	0	0	0
pressure	0	0	0	0	0	0	0	0	0	0	1	0	0	1
rats	0	0	0	0	0	0	0	0	0	0	0	0	1	1
respect	0	0	0	0	0	0	0	1	0	0	0	1	0	0
rise	0	0	0	1	0	0	0	0	0	0	0	0	0	1
study	1	0	1	0	0	0	0	0	1	0	0	0	0	0

Transaction data

- Each record corresponds to a transaction involving a set of items
- E.g., in a grocery store purchase, the set of products purchased by a customer constitute a transaction, while the individual products that were purchased are the items

Table 6.22. Example of market basket transactions.

Customer ID	Transaction ID	Items Bought
1	0001	{a,d,e}
1	0024	{a,b,c,e}
2	0012	{a,b,d,e}
2	0031	{a,c,d,e}
3	0015	{b,c,e}
3	0022	{b,d,e}
4	0029	{c,d}
4	0040	{a,b,c}
5	0033	{a,d,e}
5	0038	{a,b,e}

Ordered data

- Genomic sequence data

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Graph data

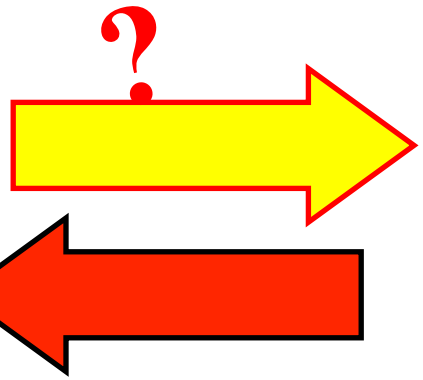
- Nodes correspond to entities, edges correspond to relationships
- E.g.: Web graph with HTML links, molecules with atoms and bonds

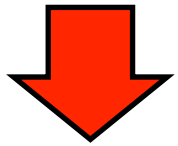
Measurement

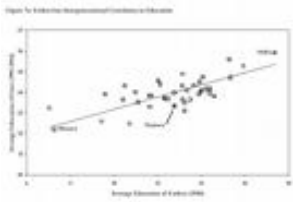
Real world

Data

Relationship
in real world







Relationship
in data

Go [illegible] tations

Does the data representation provide the appropriate abstraction for answering questions about the real world?

Document Data

Document =
words frequencies

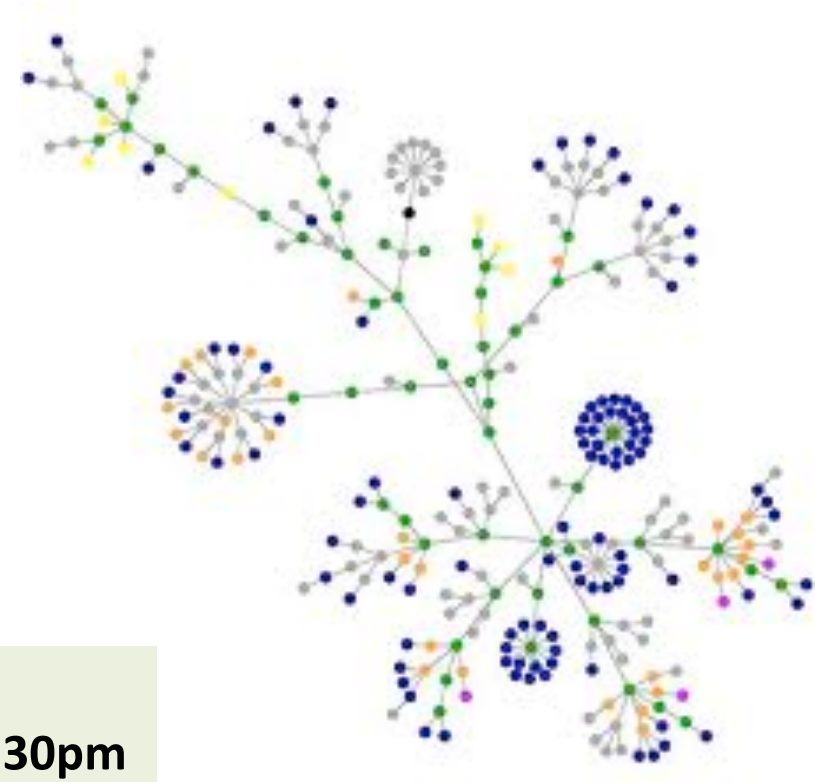
Terms	Documents													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14
abnormalities	0	0	0	0	0	0	0	1	0	1	0	0	0	0
age	1	0	0	0	0	0	0	0	0	0	0	1	0	0
behavior	0	0	0	0	1	1	0	0	0	0	0	0	0	0
blood	0	0	0	0	0	0	0	1	0	0	1	0	0	0
close	0	0	0	0	0	0	1	0	0	0	1	0	0	0
culture	1	1	0	0	0	0	0	1	1	0	0	0	0	0
depressed	1	0	1	1	1	0	0	0	0	0	0	0	0	0
discharge	1	1	0	0	0	1	0	0	0	0	0	0	0	0
disease	0	0	0	0	0	0	0	0	1	0	1	0	0	0
fast	0	0	0	0	0	0	0	0	0	1	0	1	1	1
generation	0	0	0	0	0	0	0	0	1	0	0	0	1	0
oestrogen	0	0	1	1	0	0	0	0	0	0	0	0	0	0
patients	1	1	0	1	0	0	0	1	0	0	0	0	0	0
pressure	0	0	0	0	0	0	0	0	0	0	1	0	0	1
rats	0	0	0	0	0	0	0	0	0	0	0	0	1	1
respect	0	0	0	0	0	0	0	1	0	0	0	1	0	0
rise	0	0	0	1	0	0	0	0	0	0	0	0	0	1
study	1	0	1	0	0	0	0	0	1	0	0	0	0	0

Task 1: based on this representation, identify the “hot topics” in the media, in the span of a month.

Task 2: based on this representation, identify the general sentiment about the new IPhone in the 12 hours after its release

Take Home Quiz

- Nodes are users in a social network, edges represent interactions between users.
- Edges are weighted as follows:
  - No interaction: no edge
  - Otherwise – edge weight: # interactions.



Will be posted on Piazza.
Please respond (privately) until **Monday 12:30pm**

Take Home Quiz

- You should:
  - *Find one example of a question (task) that can be answered using this representation, and one that cannot.*
  - *How would you modify the network representation to answer the second question?*
    - *What should you consider when changing the representation? What are the tradeoffs involved?*

Will be posted on Piazza.
Please respond (privately) until **Monday 12:30pm**