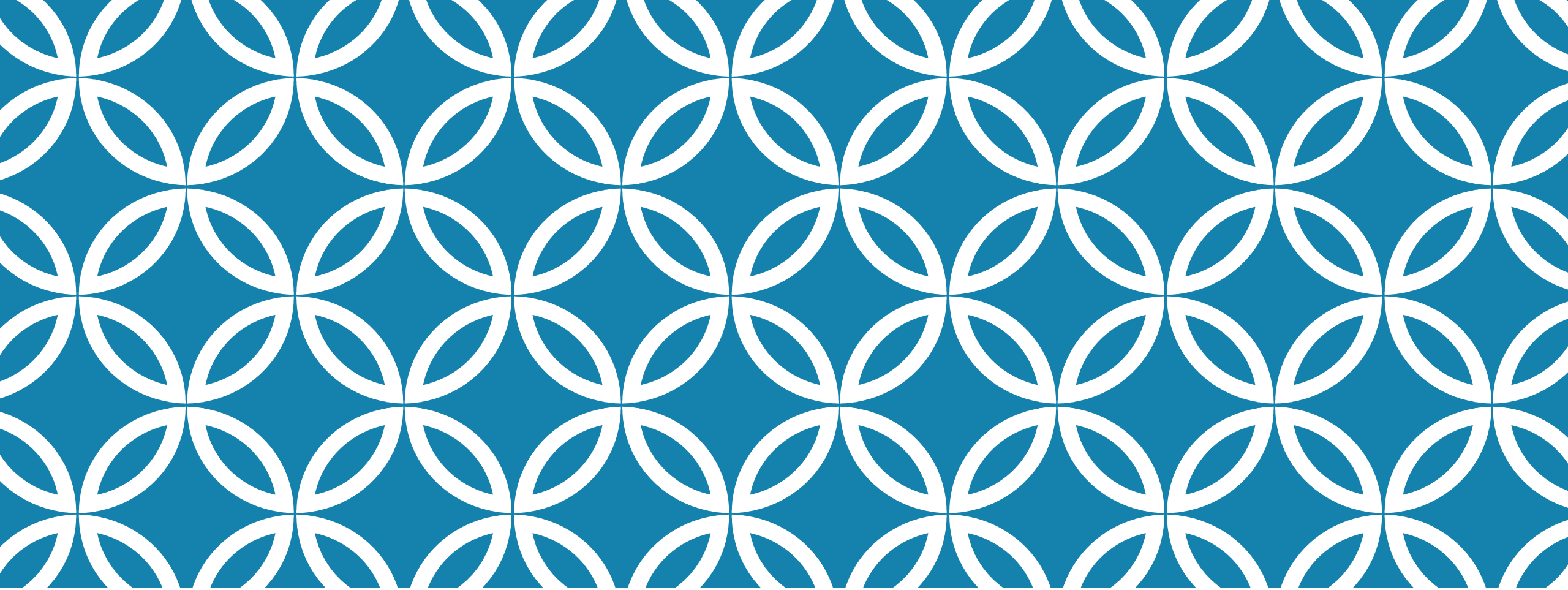
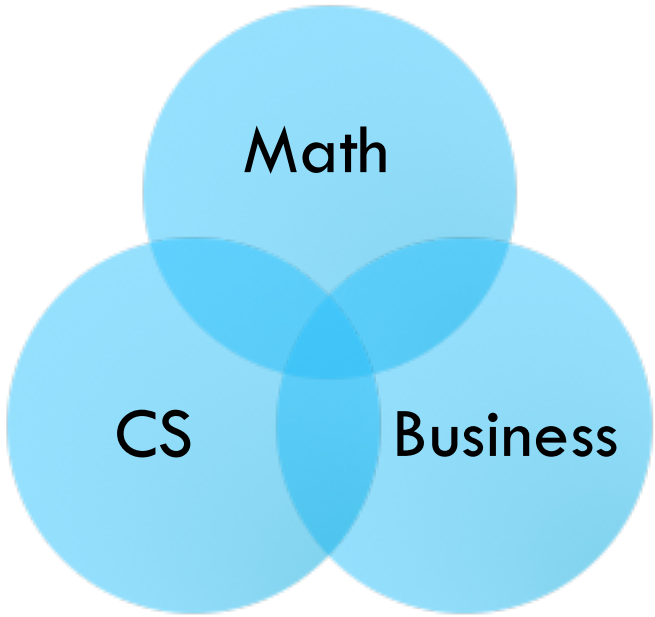


INTRODUCTION TO DATA MINING

Material based on Aggarwal book and Coursera's Data Scientist Toolbox

WHAT IS DATA MINING?

- An interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems (wikipedia)



DATA MINING APPLICATIONS

Business

- Data: Purchase information, web site browsing habits, social network data
- Goals:
  - "Is this really the customer's credit card?"
  - "How do I target my actual customers?"
  - "What will revenue be next year?"

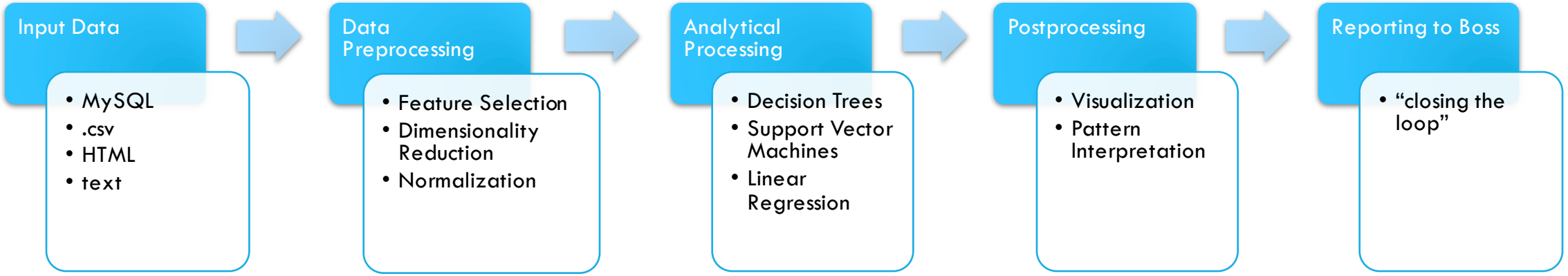
Science

- Data: weather observations (collecting land surface, ocean, atmosphere readings), genomic data, crop/agricultural information
- Goals:
  - "How is land surface precipitation and temperature affected by ocean surface temperature?"
  - "How well can we predict the beginning and end of the growing season for a region?"

WHAT IS *NOT* DATA MINING?

- looking up records in a MySQL database (*database*)
- finding relevant web pages based on a Google search query *(information retrieval)*

DATA MINING PROCESS



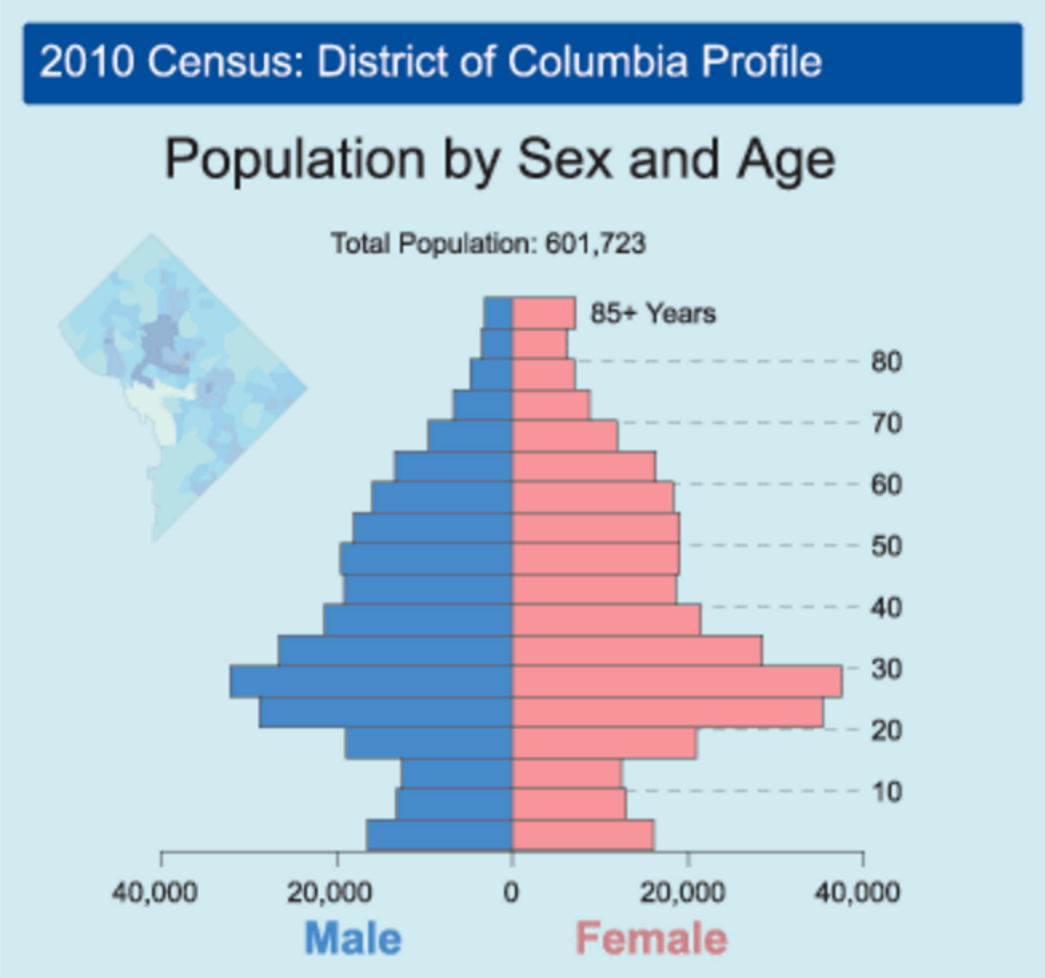
TYPES OF QUESTIONS

From easiest to hardest type of analysis:

- Descriptive
- Exploratory
- Inferential
- Predictive
- Causal
- Mechanistic

DESCRIPTIVE ANALYSIS

- The first type of analysis performed
- Description and interpretation are two different/separate steps
- Conclusions cannot typically be drawn from descriptive analysis without additional statistical modeling



EXPLORATORY ANALYSIS

- Goal is to find relationships you didn't know about
- New connections – useful for defining future studies
- Not the final conclusion
- Important: correlation does not imply causation

INFERENTIAL ANALYSIS

- Use a small sample to say something more general about a larger population
- Inference is commonly the goal of statistical modeling
- Important to estimate your uncertainty

Epidemiology. 2013 Jan;24(1):23-31. doi: 10.1097/EDE.0b013e3182770237.

Effect of air pollution control on life expectancy in the United States: an analysis of 545 U.S. counties for the period from 2000 to 2007.

Correia AW[1], Pope CA 3rd, Dockery DW, Wang Y, Ezzati M, Dominici F.

Author information

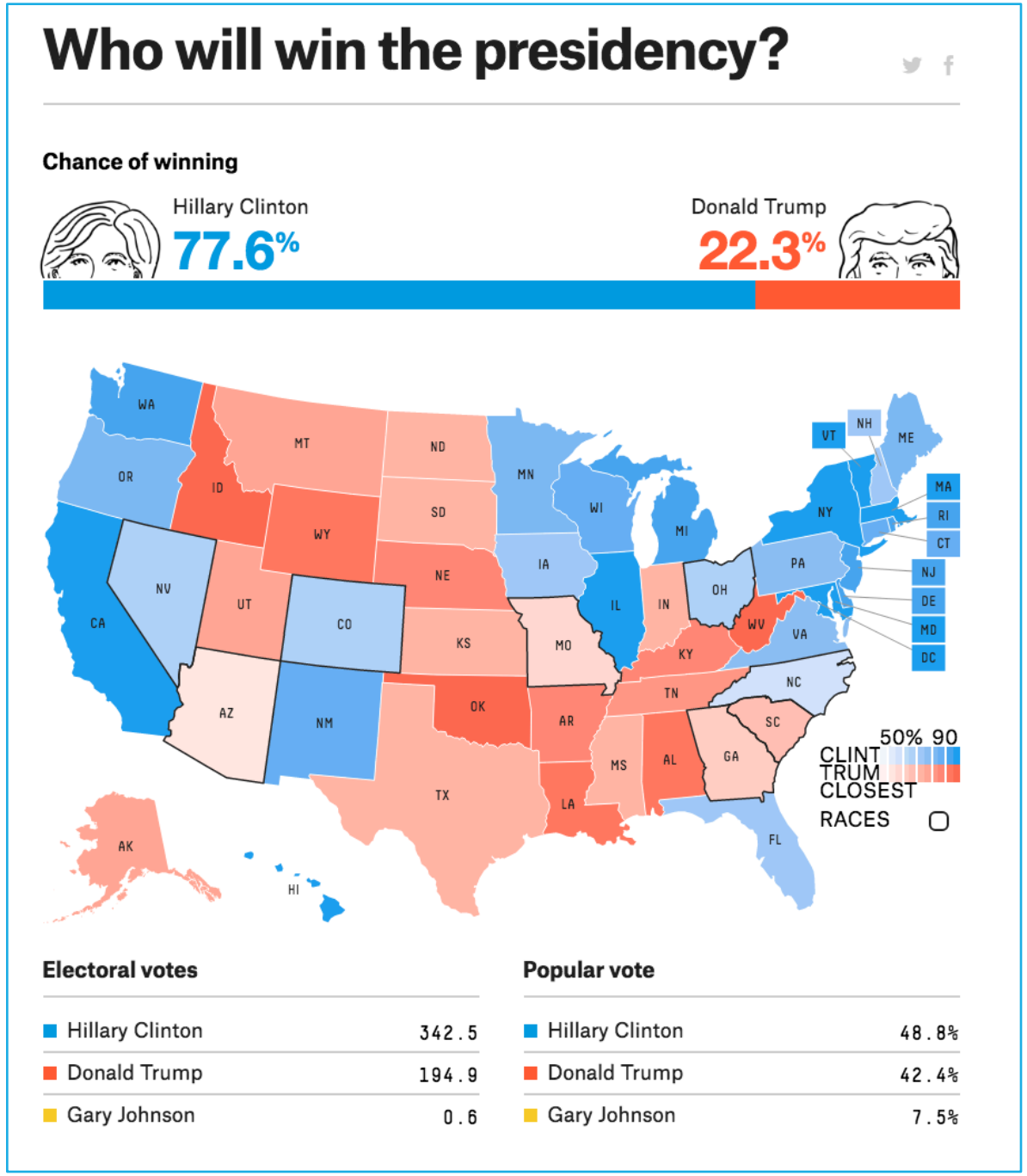
Abstract

BACKGROUND: In recent years (2000-2007), ambient levels of fine particulate matter (PM2.5) have continued to decline as a result of interventions, but the decline has been at a slower rate than previous years (1980-2000). Whether these more recent and slower declines of PM2.5 levels continue to improve life expectancy and whether they benefit all populations equally is unknown.

METHODS: We assembled a data set for 545 U.S. counties consisting of yearly county-specific average PM2.5, yearly county-specific life expectancy, and several potentially confounding variables measuring socioeconomic status, smoking prevalence, and demographic characteristics for the years 2000 and 2007. We used regression models to estimate the association between reductions in PM2.5 and changes in life expectancy for the period from 2000 to 2007.

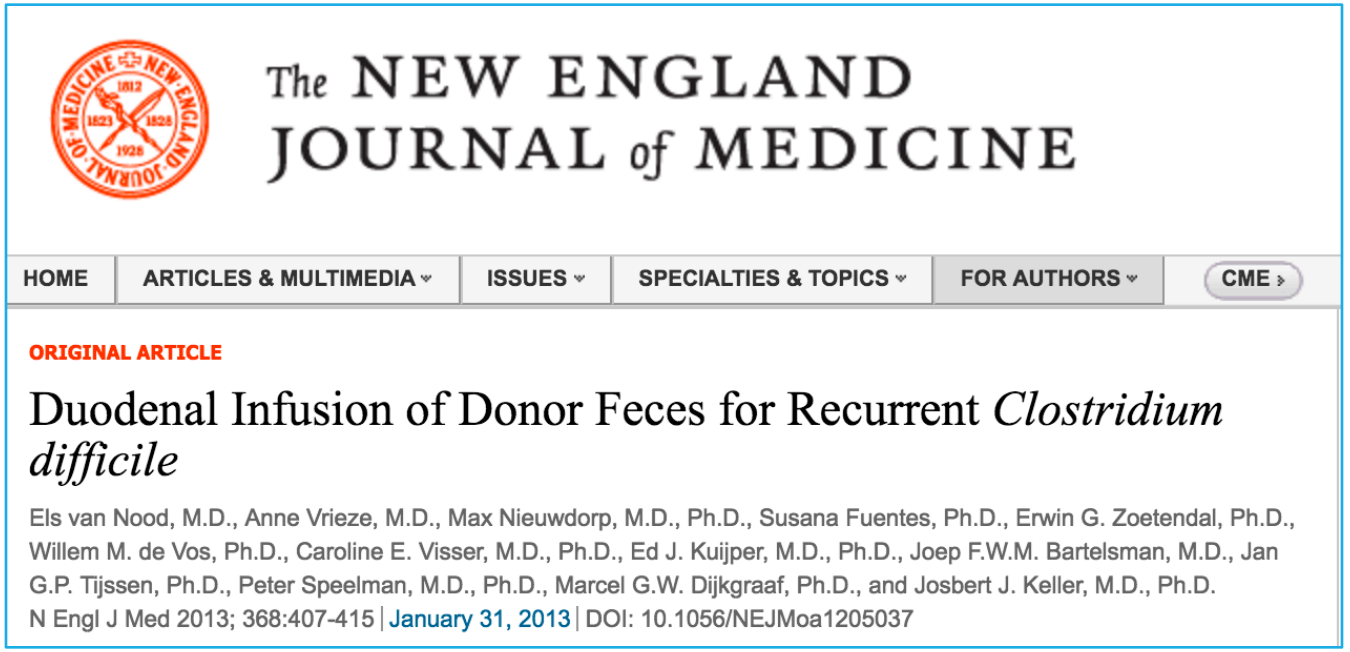
PREDICTIVE ANALYSIS

- Use data on some object to predict values for another object
- Remember: If x predicts y, it doesn't mean x causes y!
- More data and a simple model tends to work best
- "Prediction is very difficult, especially if it's about the future" – Niels Bohr
- http://projects.fivethirtyeight.com/2016-election-forecast/



CAUSAL ANALYSIS

- Goal: To find out what happens to one variable when you change another variable
- Usually randomized studies are required (other approaches possible, but very sensitive to assumptions)
- Causal models viewed as the "gold standard" for data analysis
- Usually identified as "average effects" – may not apply on individual basis



MECHANISTIC ANALYSIS

- Goal: To understand the exact changes in one variable that lead to the exact changes in others
- Really hard to infer, except in very simple situations
- Usually modeled by a set of deterministic equations (physical/engineering sciences)
- Only random component is measurement error

WHAT IS DATA?

"Data are values of qualitative or quantitative variables, belonging to a set of items." (Wikipedia)

- Set of items: population or set of items that you are interested in
- Variables: measurements or characteristics of an item
- Qualitative: measures of 'types' and may be represented by a name, symbol, or a number code (sex, country of origin, treatment option)
- Qualitative: information about quantities; that is, information that can be measured and written down with numbers. Usually measured on a continuous scale (weight, height, age)

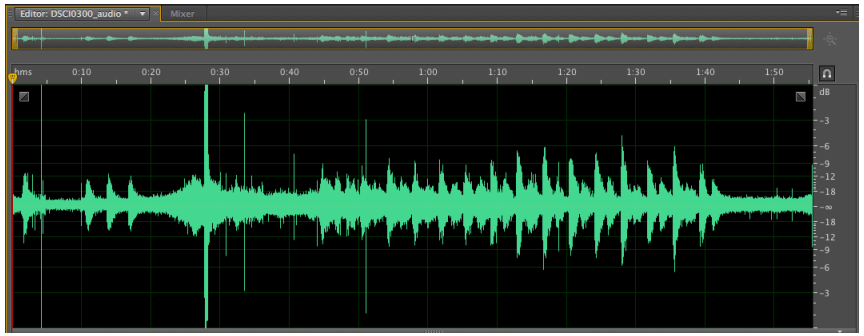
WHAT DOES DATA LOOK LIKE?

```
<patient>
    <name>
        <prefix> Mr.</prefix>
        <suffix>Jr</suffix>
        <family>Thomson</family>
        <given>Marcus</given>
        <given>Joshua</given>
    <name>
    <administrativeGenderode code="Male"/>
    <birthTime value="20140801"/>
</patient>
```

Element

Atribute

Value



RARELY DO WE START HERE:

1	Operation Date	Operation Direction	Operation Type	Release Time	Close Time	Maturity Start	Maturity End	# Issues	Total Accepted (M)	Total Submitted (M)
2	8/25/2005 0:00	P	COUPON	10:30 AM	11:00 AM	9/30/2005 0:00	8/15/2006 0:00	22	1298	11683.5
3	9/7/2005 0:00	P	COUPON	10:30 AM	11:00 AM	11/15/2008 0:00	5/15/2009 0:00	9	1201	6597
4	9/8/2005 0:00	P	TIPS	11:00 AM	11:30 AM	1/15/2007 0:00	4/15/2032 0:00	17	450	1908
5	9/12/2005 0:00	P	COUPON	10:30 AM	11:00 AM	11/15/2007 0:00	10/15/2008 0:00	13	1100	6029
6	9/21/2005 0:00	P	BILL	10:30 AM	11:00 AM	1/12/2006 0:00	3/16/2006 0:00	10	1247	11445
7	9/22/2005 0:00	P	COUPON	10:45 AM	11:15 AM	6/15/2009 0:00	1/15/2010 0:00	11	1104	6491
8	10/4/2005 0:00	P	COUPON	10:30 AM	11:00 AM	10/31/2006 0:00	7/31/2007 0:00	18	1193	8334.8
9	10/25/2005 0:00	P	COUPON	10:30 AM	11:00 AM	1/31/2006 0:00	10/31/2006 0:00	18	1000	9746
10	10/26/2005 0:00	P	COUPON	10:45 AM	11:15 AM	11/15/2013 0:00	2/15/2026 0:00	35	902	10949
11	11/4/2005 0:00	P	COUPON	10:30 AM	11:00 AM	3/15/2010 0:00	8/15/2013 0:00	17	800	9652
12	11/14/2005 0:00	P	COUPON	10:30 AM	11:00 AM	1/31/2006 0:00	10/31/2006 0:00	18	1096	14944.9
13	11/18/2005 0:00	P	COUPON	10:30 AM	11:00 AM	6/15/2009 0:00	4/15/2010 0:00	13	1095	9182
14	1/4/2006 0:00	P	COUPON	10:30 AM	11:00 AM	5/15/2010 0:00	2/15/2013 0:00	15	844	9413
15	1/5/2006 0:00	P	TIPS	10:30 AM	11:00 AM	1/15/2007 0:00	4/15/2032 0:00	17	448	2550
16	1/12/2006 0:00	P	BILL	10:30 AM	11:00 AM	2/23/2006 0:00	5/4/2006 0:00	11	1249	10480
17	1/23/2006 0:00	P	COUPON	10:30 AM	11:00 AM	2/15/2014 0:00	2/15/2026 0:00	35	947	8191.5
18	1/25/2006 0:00	P	COUPON	10:30 AM	11:00 AM	11/15/2008 0:00	7/15/2009 0:00	13	1090	7008
19	1/26/2006 0:00	P	COUPON	10:30 AM	11:00 AM	9/30/2007 0:00	10/15/2008 0:00	15	1190	9743
20	2/8/2006 0:00	P	COUPON	10:15 AM	10:45 AM	11/30/2006 0:00	8/31/2007 0:00	18	1250	13775
21	2/10/2006 0:00	P	BILL	10:30 AM	11:00 AM	5/11/2006 0:00	8/10/2006 0:00	14	1244	12935
22	2/21/2006 0:00	P	COUPON	10:30 AM	11:00 AM	2/15/2007 0:00	11/30/2007 0:00	19	1248	8695

DATA IS THE SECOND MOST IMPORTANT THING

- Don't be driven by the data, be driven by the QUESTION
- Data might enable or limit particular questions but having data can't save you if you don't have a question!

BUILDING BLOCKS

- Association Pattern Mining
- Data Clustering
- Outlier Detection
- Data Classification

ASSOCIATION PATTERN MINING

- Looking for patterns of attribute values that frequently occur together
- Example: items frequently purchased together or by the same customer
- How often does the pattern occur?
- How confident are we that one attribute value occurring indicates the other will occur?

DATA CLUSTERING

- Given some dataset, group elements so that items in a group are “similar” to one another
- Example: Customer segmentation (Grouping customers that are similar to one another based on some measure)
- Unsupervised – we don’t know the clusters or even relevant attributes for the clusters

OUTLIER DETECTION

- "An outlier is an observation that deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism." (Hawkins)
- Example: Intrusion-detection systems, credit card fraud
- We may use data clustering as an intermediate step in detecting outliers

DATA CLASSIFICATION

- Like data clustering, the entities are grouped, but here the process is supervised – that is, we know the classifications ahead of time
- Example: Classifying customer buying behavior, intrusion detection, insurance ratings