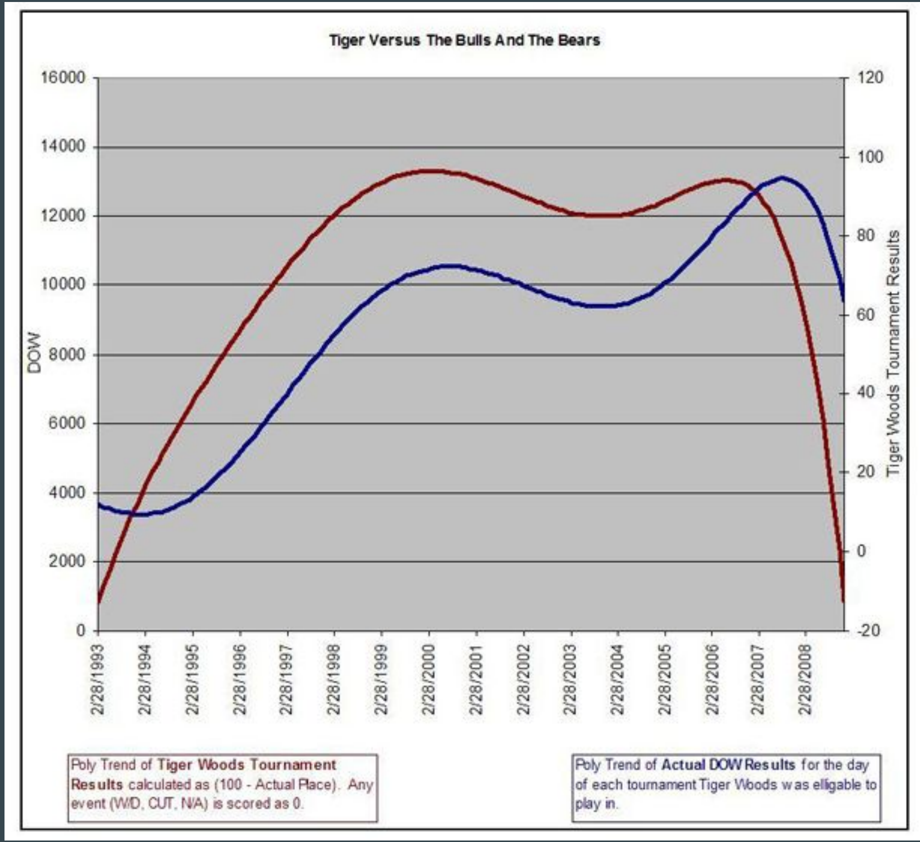
Statistical Inference: The Big Picture

PLSC 309

15 April 2019

First things first



Logistics

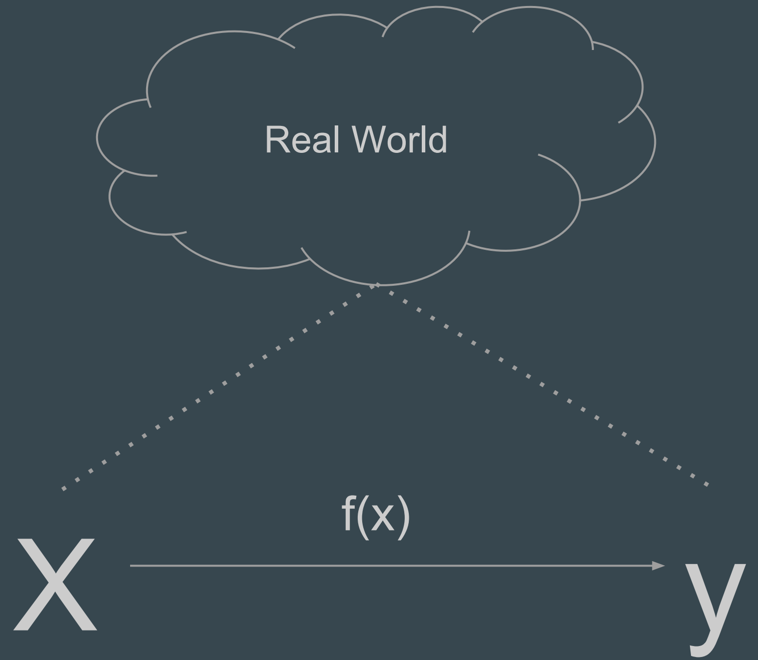
- There is no problem set this week
- There is a **two-day** lab
    - So both **Wednesday** and **Friday** will be lab days
- That means make sure to come to class on Wednesday, or email me ahead of time, or else you'll be working alone!

Final week of semester

- April 22-26 there will be **no class** and **no assignments**
- Starting this week, I will accept up to **3** revisions per week
- **Final due date for revisions is 12:00 AM April 29**
- Office hours will be MWF next week
    - 9-12

Modern Statistical Analysis

What is statistical inference?



Find g(x) from X_D and y_D, where...

- D is a given dataset drawn from the real world
- g(x) is the best guess for f(x)

We've learned a lot of g(x) this semester

- OLS represents g(x) as a straight line
- GLMs allow us to extend that to all different kinds of functions
    - Poisson
    - Binomial
    - Multinomial
    - Exponential
- But let's forget about that and think of g(x) as *any type* of approximation of a real world function

The Guessing Game: Hypothesis space

$$h_1, \ldots, h_m \in \mathcal{H}$$

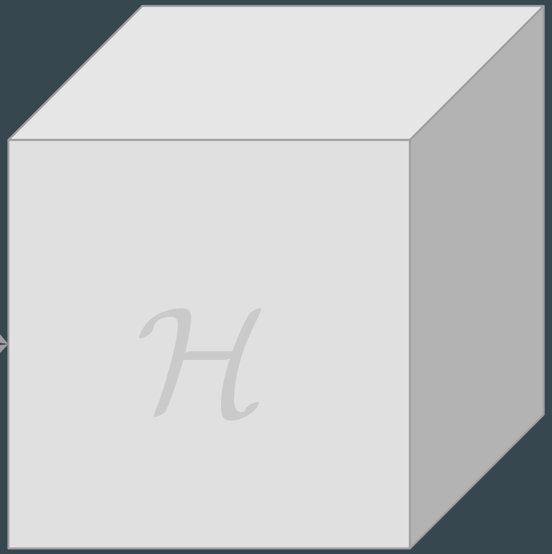
$$h_i = (\mathbf{X}, \mathbf{w}, \hat{y})$$

$$h_i = g'(X)$$

E.g., voter turnout

LPM for turnout with two explanatory variables

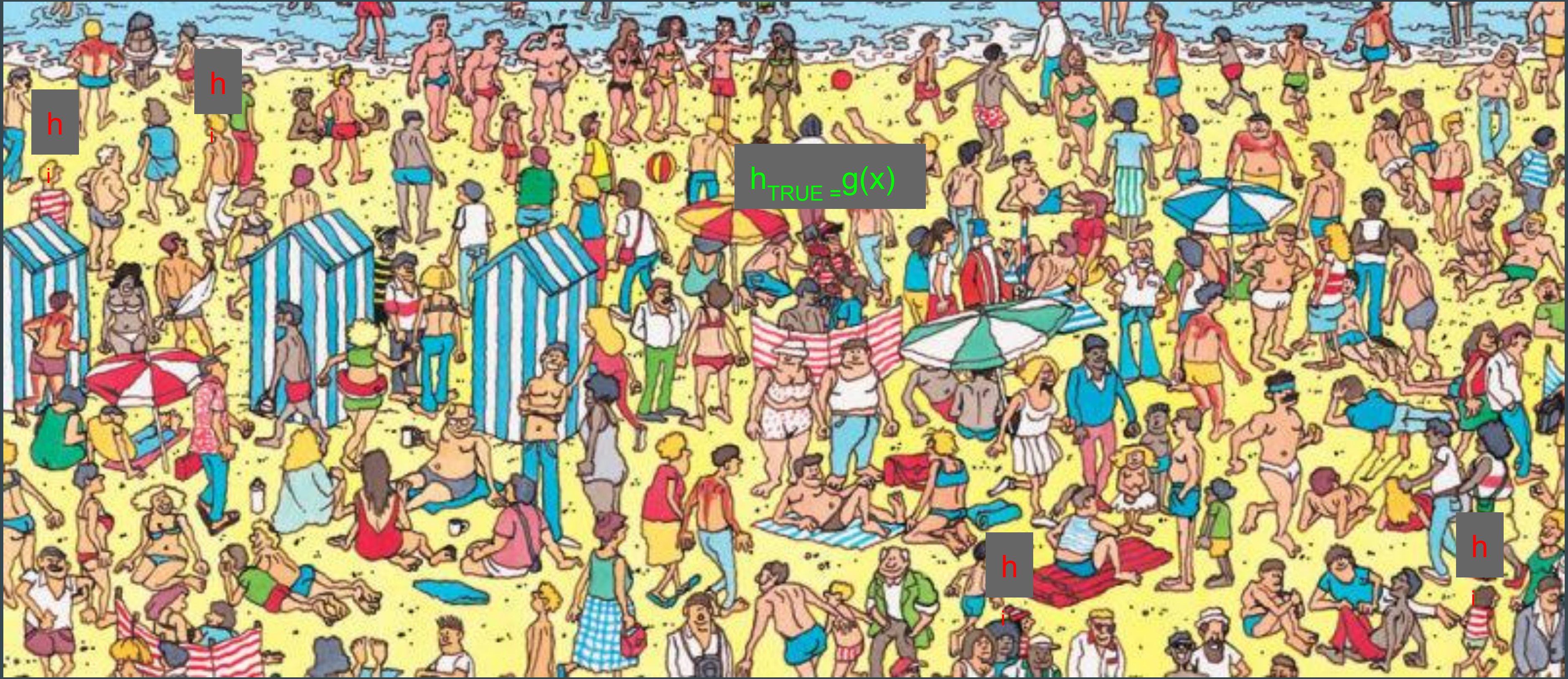
Logistic Regression for turnout



What affects the size of your hypothesis space?

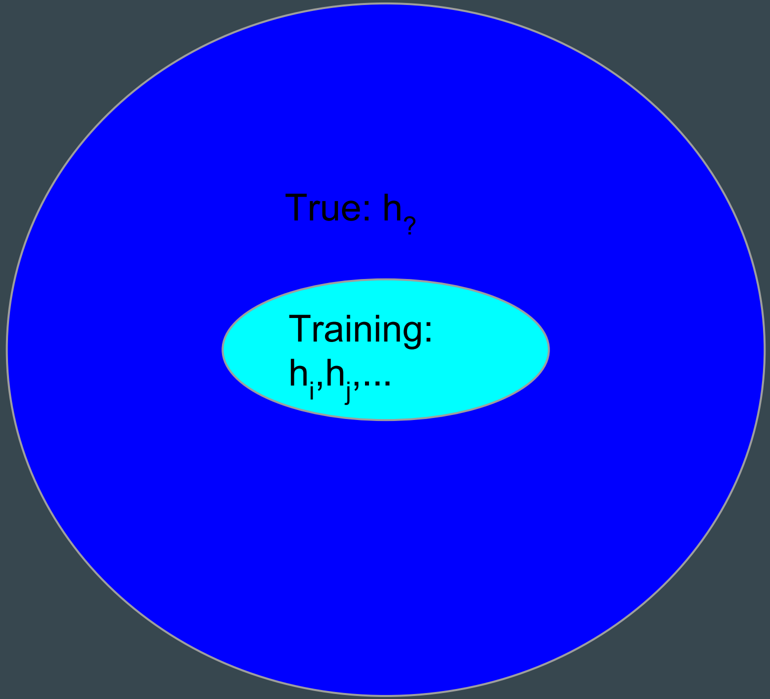
1. What functional form you use
   a. Linear, logistic, etc.
2. Are there any interactions?
   a. Including exponentiated terms
3. What data do you use?
   a. Which explanatory variables
   b. Measurement and operationalization
4. How do you calculate your errors?

Where’s waldo?



Bad guesses and more complex $\mathcal{H}$

What we want to know is how likely we are to guess the wrong thing?



$P[h_{train} \neq h_?] =$

$P(h_i \neq h_?) \vee P(h_i \neq h_?) \vee \ldots$

Things will obviously start to get out of hand as the complexity of H increases!

Intuition behind complex $\mathcal{H}$ and difficulty of modelling

Chess

VS.

Topic Model

- Complicated models designed for messy target function and very large, sparse matrix
- Gold standard is human coded

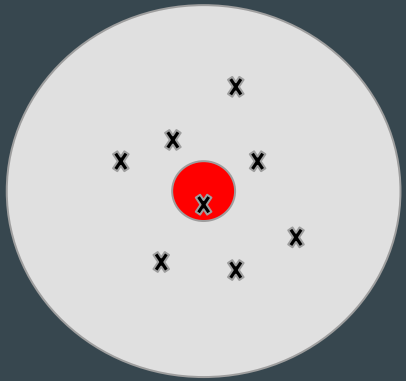
Bias and variance

Bias

- g(x) makes wrong predictions
- g(x) should be *flexible*
- Being really good at making really bad predictions

Variance

- g(x) captures too much noise
- g(x) should be *simple*
- Being really bad about making really good predictions



Cost = loss + regularization

- How we manage the bias-variance tradeoff: tells us when too much fit is a bad thing

sum of squared errors: $\sum(\mathbf{w}^T\mathbf{x} - \hat{y})^2$

w/ regularization: $\sum(\mathbf{w}^T\mathbf{x} - \hat{y})^2 + \lambda||\mathbf{w}||^2$

- We know this intuitively. We need to find a way to shrink **w**, or else we risk overfitting.

Regularization = adding noise, reducing weights

- Regularization is when we deliberately induce noise into the data
- In other words, we “tone down” our conclusions
- This is to avoid overfitting and improve predictive validity

Common cost functions

Loss and Regularizer	Classification
1.Ordinary Least Squares $\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^\top x_i - y_i)^2$	• Squared Loss • No Regularization
2.Ridge Regression $\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^\top x_i - y_i)^2 + \lambda\|w\|_2^2$	• Squared Loss • l_2-Regularization
3.Lasso $\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^\top \mathbf{x}_i - \vec{y_i})^2 + \lambda\|\mathbf{w}\|_1$	• + sparsity inducing (good for feature selection) • + Convex • - Not strictly convex (no unique solution) • - Not differentiable (at 0)
4.Logistic Regression $\min_{\mathbf{w},b} \frac{1}{n}\sum_{i=1}^{n}\log\left(1 + e^{-y_i(\mathbf{w}^\top \mathbf{x}_i + b)}\right)$	• Often l_1 or l_2 Regularized

- Most cost functions are used because of their “pliability”
- But there are many other that we can optimize without an analytical solution!

Source: Kilian Weinberger

So far...

- We have defined the goal of statistical modelling as minimizing *generalization erro*
- Intuitively, generalization is trade-off between flexibility and simplicity
- Shown this in two different cases
  - *Computationally*: as hypothesis space gets more complex, it is more likely to contain the correct hypothesis, but you are less likely to be able to find it
  - *Statistically*: models for a particular dataset are designed to balance informative and uninformative information (simplicity vs flexibility)
- There are two things we have to worry about
  - *Underfitting*: f(x) cannot be identified within the hypothesis (not flexible, or not enough data)
  - *Overfitting*: g(x) thinks that error in y is actually part of f(x)
- Use regulization and resampling

Modelling = Representation + Evaluation + Optimization

1. Representation
    - Express real world phenomena as informative features and an outcome of interest
    - Choose a class of models to relate features and outcomes
2. Evaluation
    - *Cost function* helps determine which g(x) to pick
    - Regularization helps reduce generalization error
3. Optimization
    - Out of all the different representations, find the one that minimizes the total cost of all of our screw ups

Source: Pedros Domingos

Two problems for learning from social data

- Your data sucks!
  a. *Too much* of the wrong information
  b. *Not enough* of the right information
- The world is messy
  a. Complex models require flexible model classes
  b. Need to adjust for comparing different strategies

Your data sucks!

Not enough of the right information

- You are missing critical information that would explain f(x)
- In other words, f(x) could be outside of your hypothesis space

Too much of the wrong information

- Curse of dimensionality
- Need to fit simpler g(x)

To make things worse, these often happen together. Because you're lacking very important information, you end up including lots of features that you don't need.

Social Data in the 21st century

- There is a ton of data available
  - Social-media platforms
  - Micro-behavioral data
- But none of the right data is available
  - Impossible to get a random sample

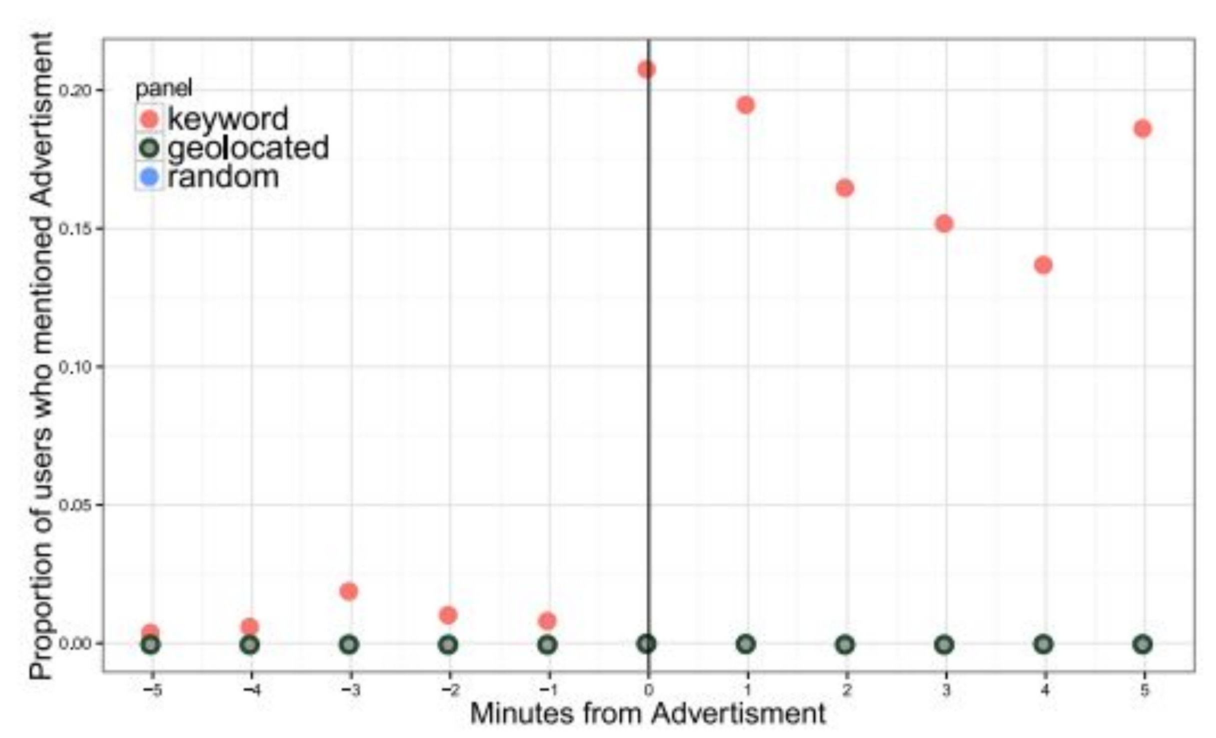
Social Data in the 21st century

Who uses social networking sites

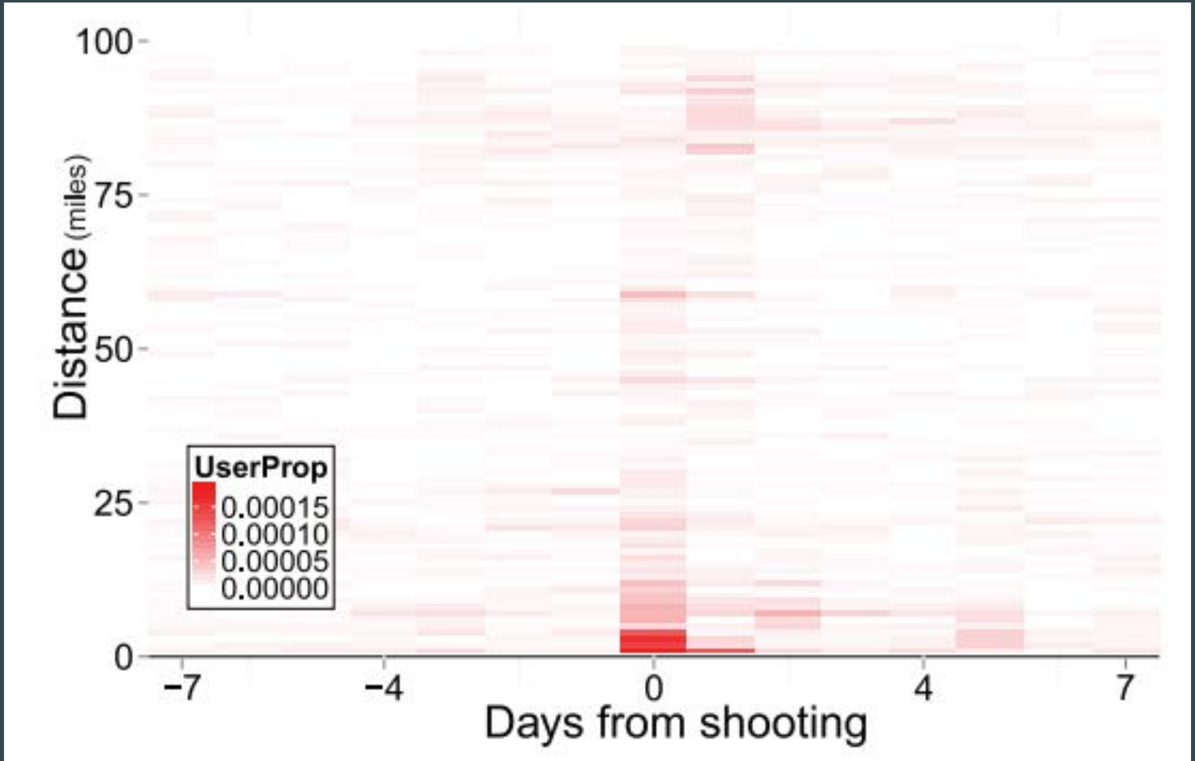
% of internet users within each group who use social networking sites

		% who use social networking sites
All internet users 18+ (n=5,112)		**73%**
a	Men (n=2,368)	69
b	Women (n=2,744)	78[a]
Race/ethnicity		
a	White, Non-Hispanic (n=3,617)	72
b	Black, Non-Hispanic (n=532)	73
c	Hispanic (n=571)	79[ab]
Age		
a	18-29 (n=929)	90[bcd]
b	30-49 (n=1,507)	78[cd]
c	50-64 (n=1,585)	65[d]
d	65+ (n=1,000)	46
Education attainment		
a	No high school diploma (n=243)	74
b	High school grad (n=1,238)	69
c	Some College (n=1,461)	75[b]
d	College + (n=2,144)	75[b]
Household income		
a	Less than $30,000/yr (n=1,212)	77
b	$30,000-$49,999 (n=886)	73
c	$50,000-$74,999 (n=746)	73
d	$75,000+ (n=1,600)	75
Urbanity		
a	Urban (n=1,605)	76[bc]
b	Suburban (n=2,585)	72
c	Rural (n=922)	70

Social Data in the 21st century



Social Data in the 21st century



The Caveman Effect

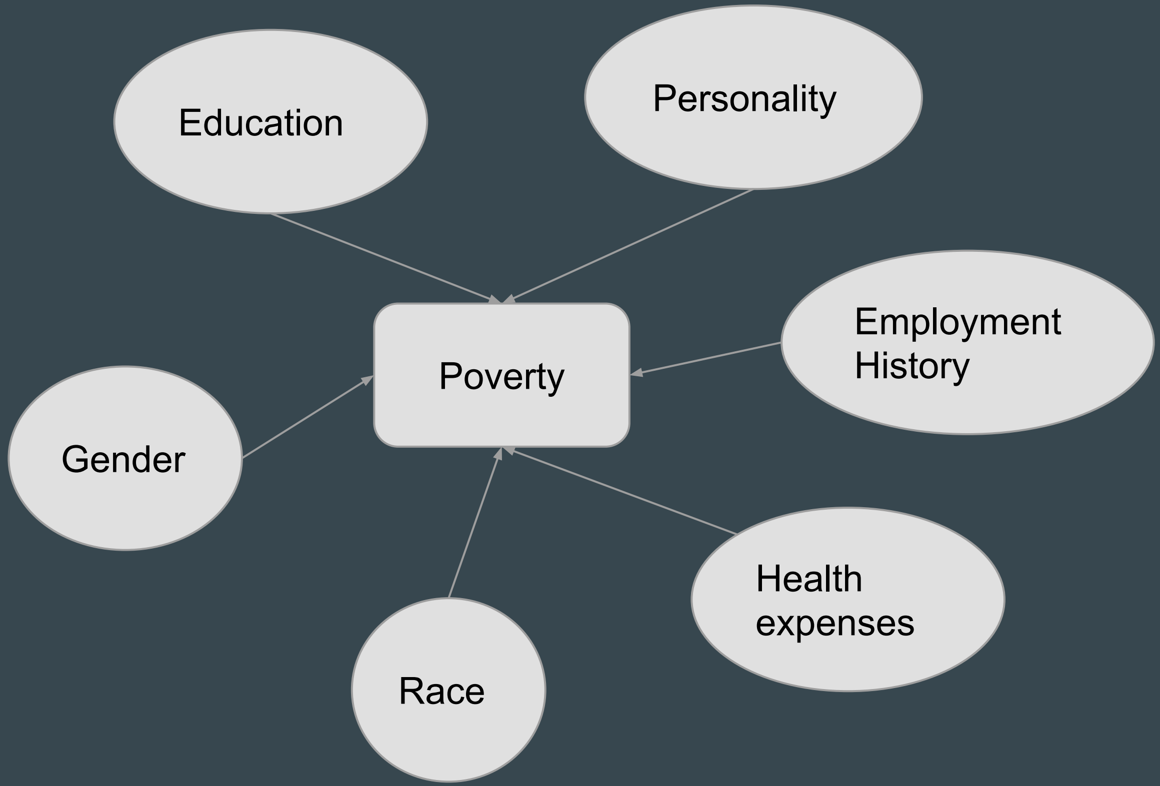
The caveman effect in political science

- What is the effect of ideology on whether or not a candidate wins an election?
- Do economic reforms increase GDP growth?
- What is the effect of media bias on political beliefs?

Homo Economicus

- Most of political science and economics assumes rational behavior
- We find much evidence in advanced, industrial democracies to support this behavior
- This actually doesn’t apply in most other cultures in the world

The world is messy



Do we live in this world?

The world is messy

... or this one?

Complicated Models Require Big Data

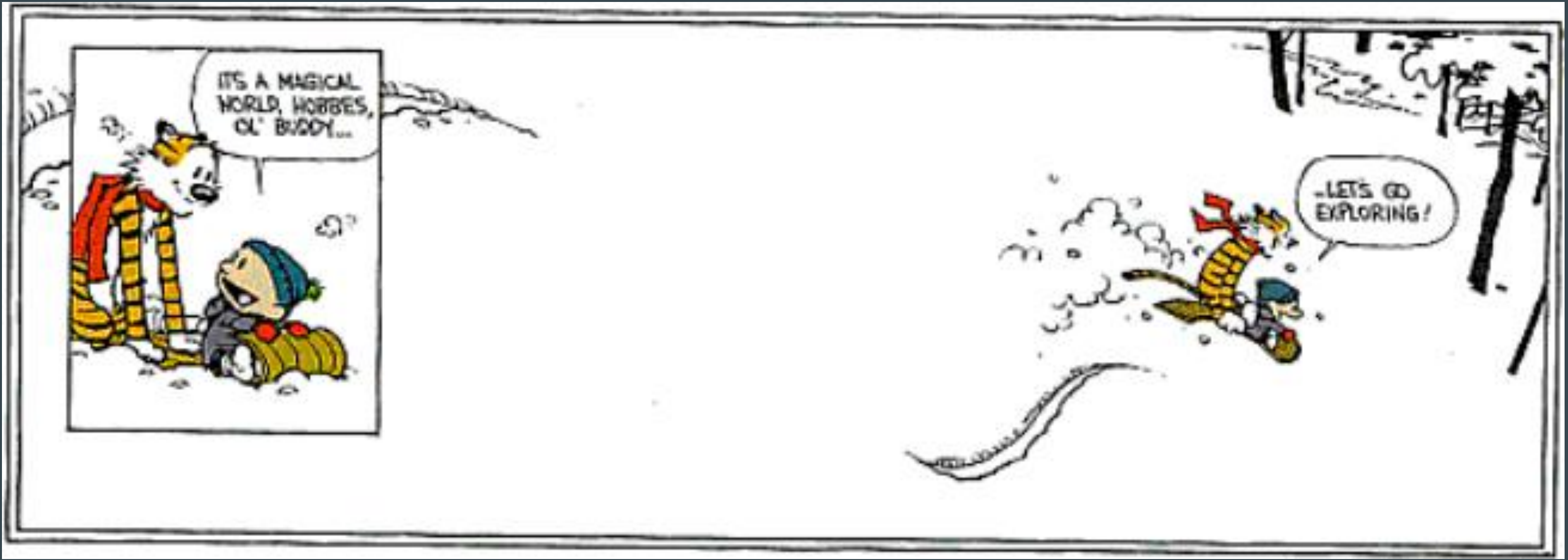
- If you are interested in age and state of residence as explanatory variables, you only need data that varies along those two dimensions
  - Young - Old People
  - Every state
- If you are interested in an interaction between the two, you need data that covers every possible intersection
  - Young and old people **in** every state

A Note on Assumptions

- We have been quite clear about the amount of assumptions required for models to work
- Assumptions are not a binary choice, there are degrees of adherence to assumptions
- The proof is in the pudding
- If you get better predictions with bad assumptions, then ignore the violations of those assumptions!

Conclusion

- Statistical modelling is fundamentally about balancing complexity and parsimony
- A model is only as good as the data that goes into it
- Social phenomena are messy and complex
- Always be critical about your data and where it comes from



Thanks for listening!