Deconstructing Data Science

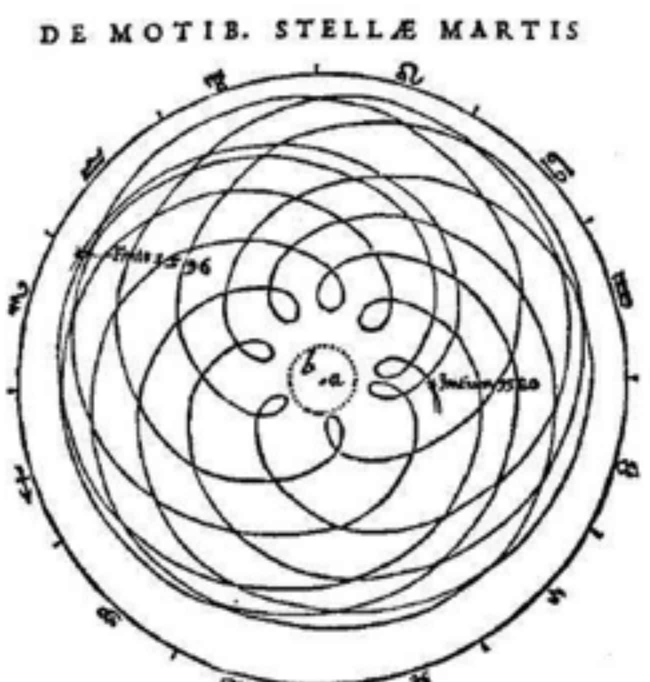
David Bamman, UC Berkeley

Info 290
Lecture 1: Introduction

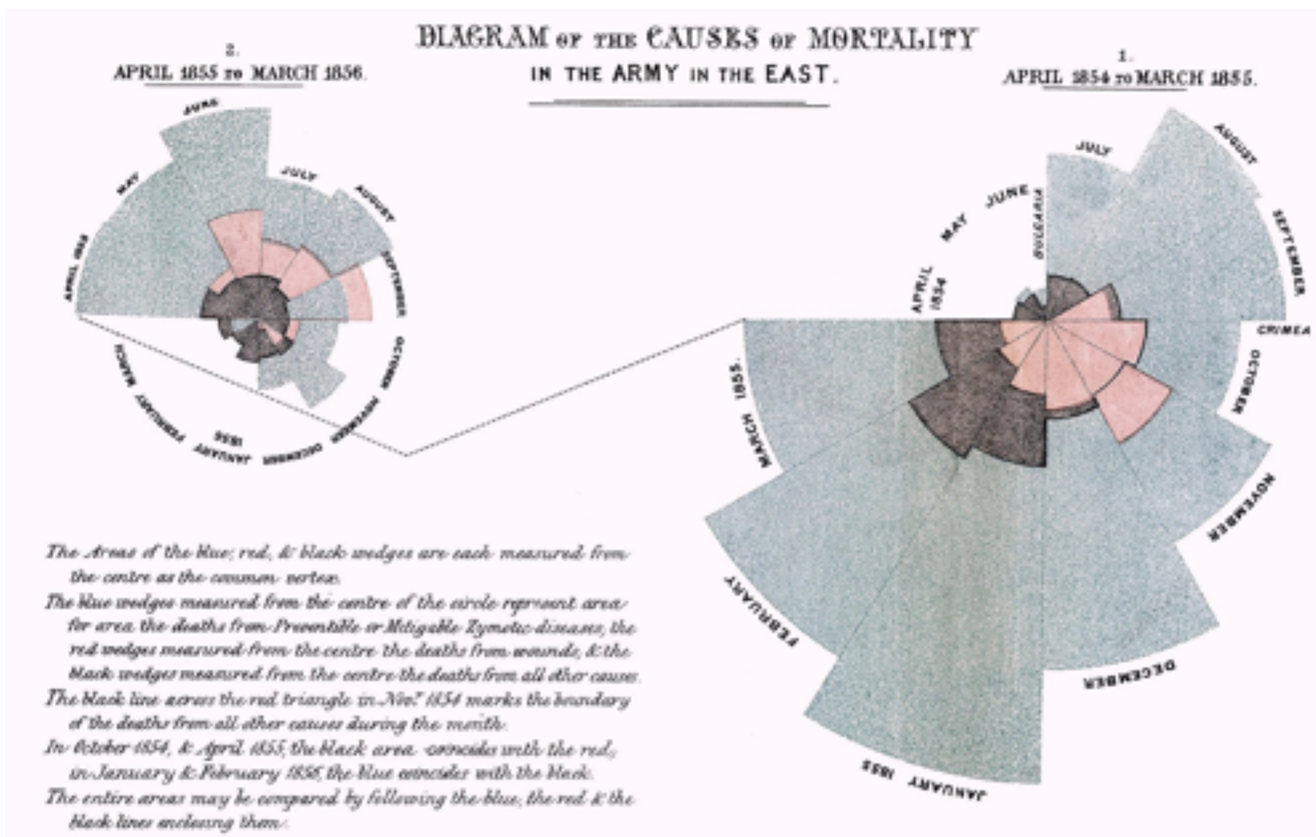
Jan 17, 2017

the “data scientist” trope

johannes kepler,
data scientist



florence nightingale,
data scientist



franz bopp,
data scientist

Software/Libraries

NumPy

SciPy

scikit learn

Dato

R

pandas $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

WEKA The University of Waikato

TensorFlow

MALLET machine learning for language toolkit

theano

Data Science

software

algorithms

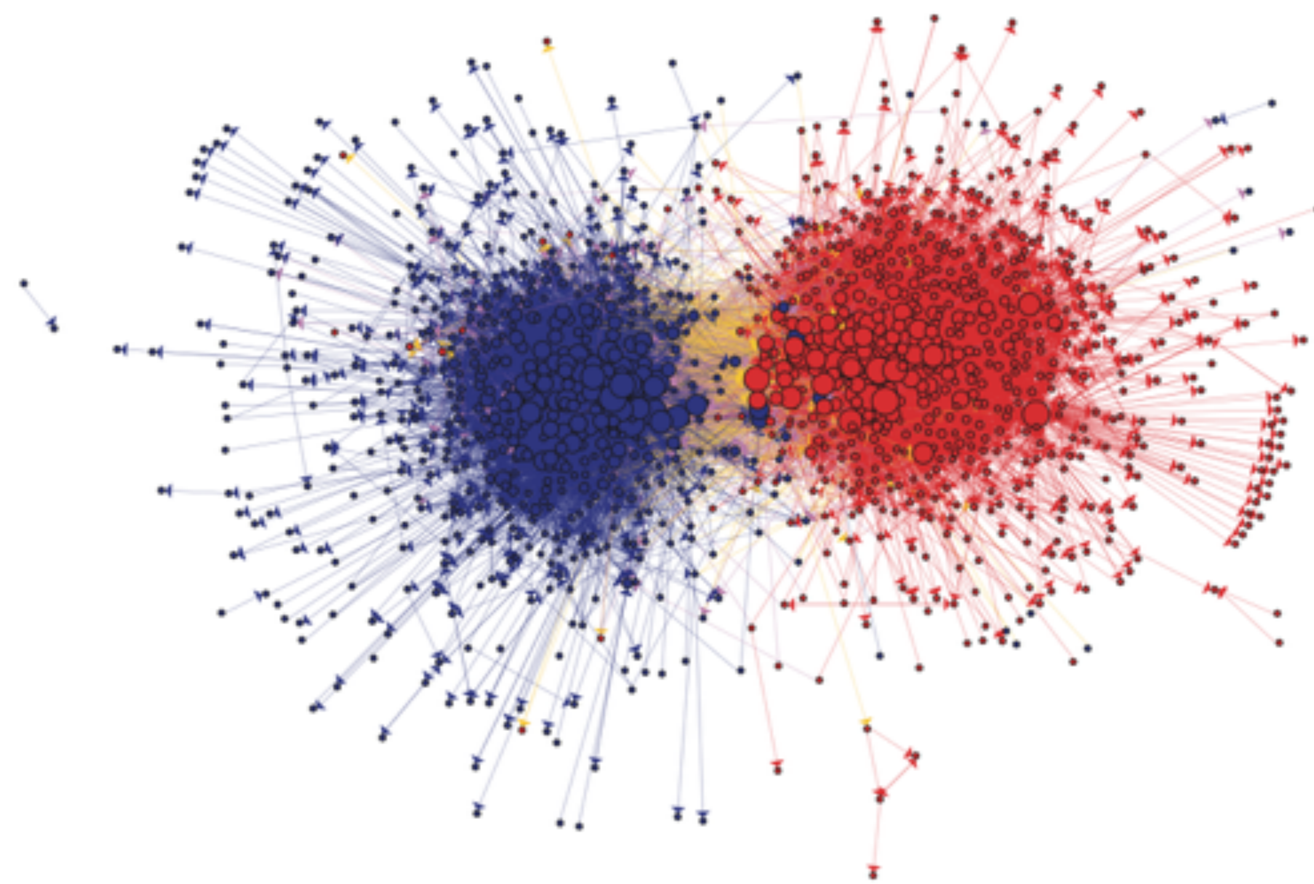
classification, regression, clustering, network analysis, prediction, hypothesis testing,

critical thinking

data selection, representation, experimental design, validation

Computational Social Science

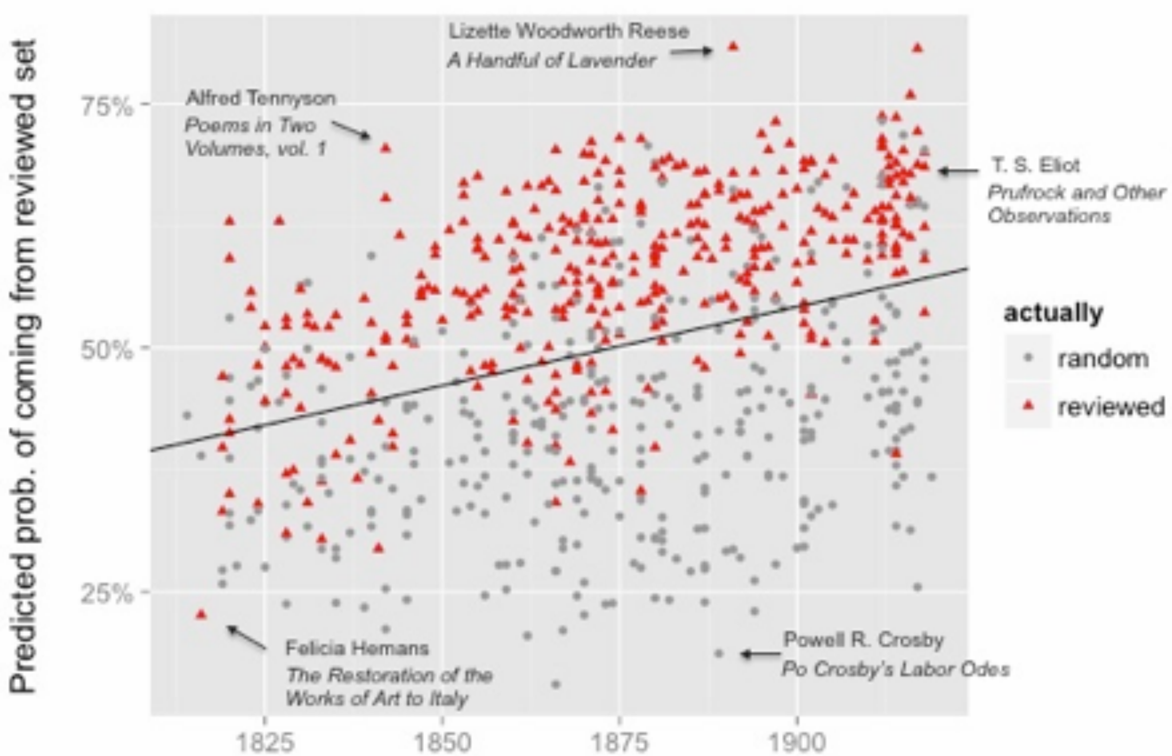
- Inferring ideal points of politicians based on voting behavior, speeches
- Detecting the triggers of censorship in blogs/ social media
- Inferring power differentials in language use



Link structure in political blogs
Adamic and Glance 2005

Digital Humanities

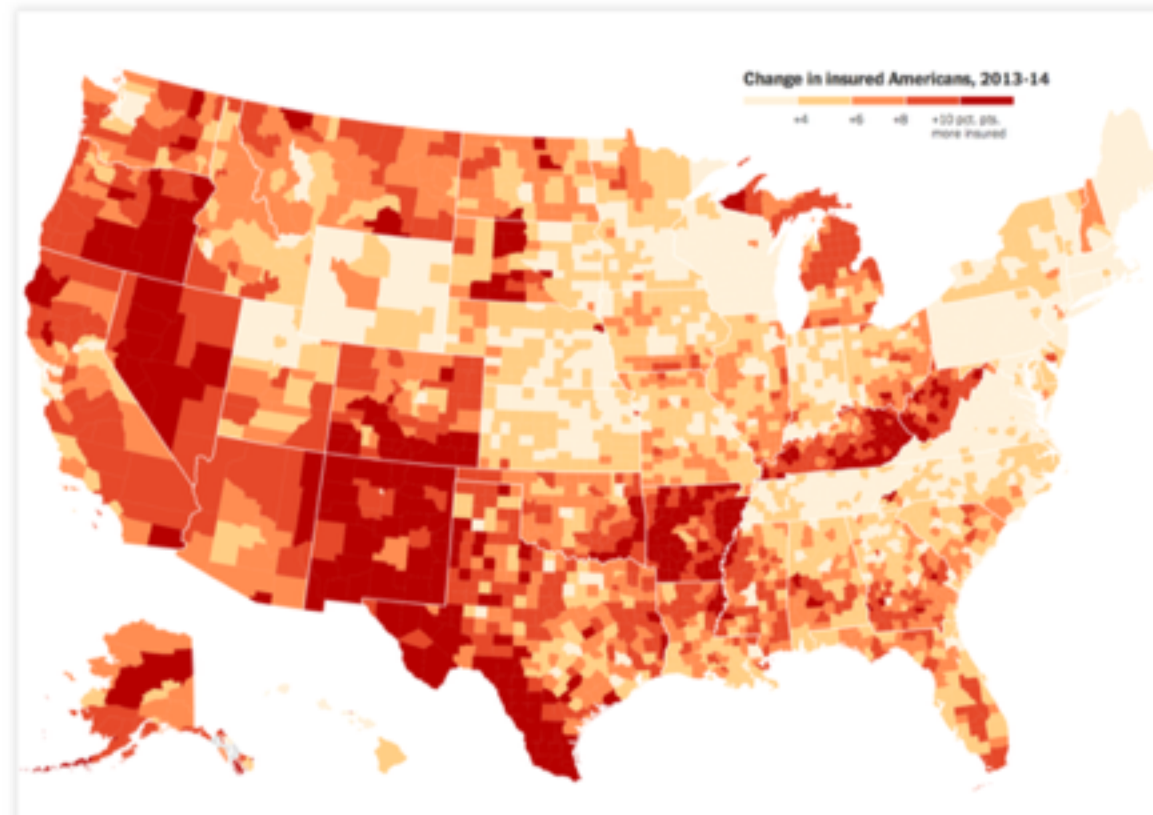
- Authorship attribution (literary texts, paintings, music)
- Genre classification (literary genre, music genre)
- Inferring plot, character types



Predicting reviewed texts
Underwood and Sellers (2015)

Computational Journalism

- Exploratory data analysis for lead generation
- Information extraction from unstructured text
- Data-driven stories



Change in insured Americans under the ACA, NY Times (Oct 29, 2014)

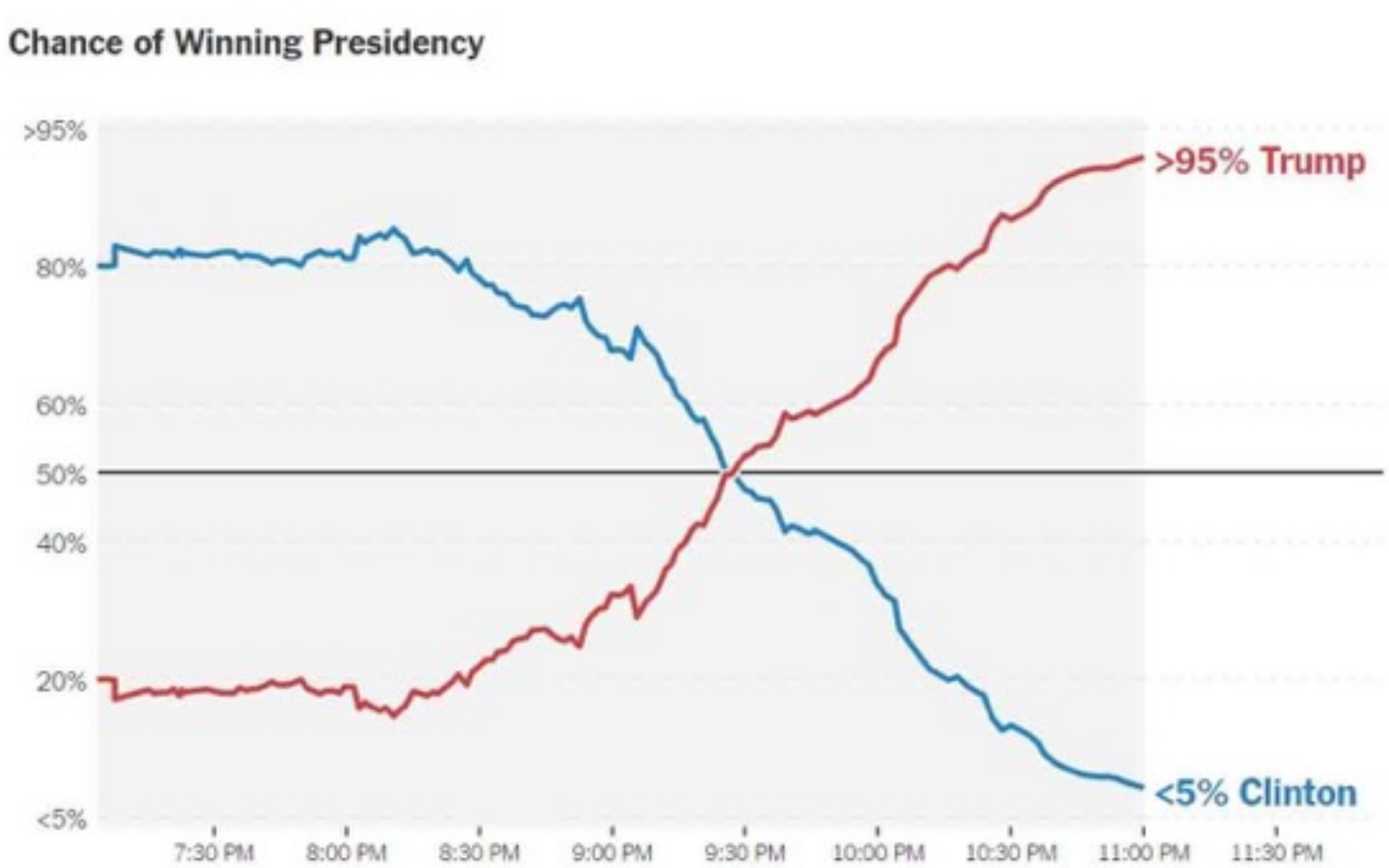
What to expect

- Each class: learn about a technical method (e.g., random forests), and an discuss application area that makes use of it.
- As the course goes on, we’ll compare methods with those we’ve already learned to critically assess the assumptions that they make and understand what methods are appropriate for different contexts.
- We will learn by example: Lots of reading.

Themes

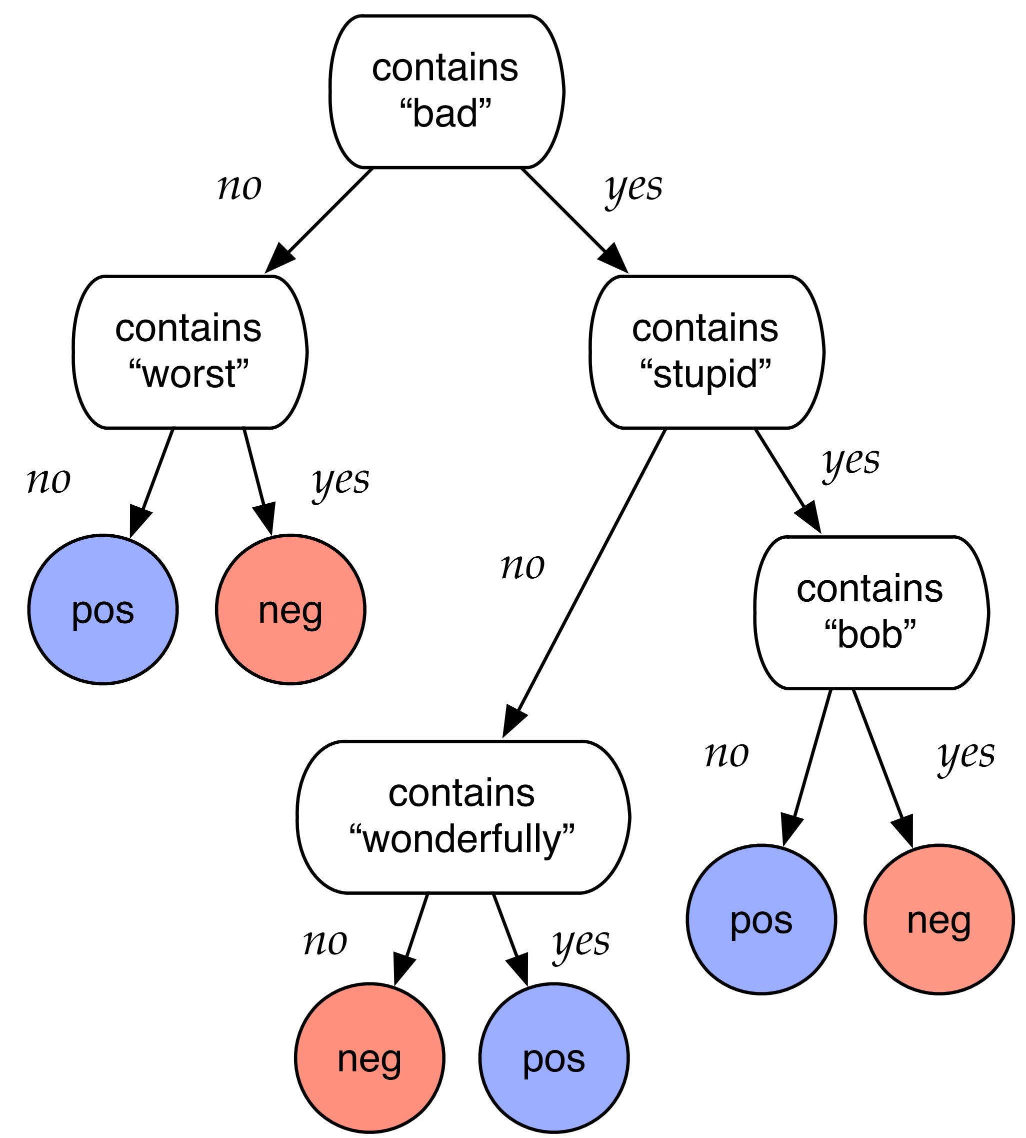
1. Validity

How do we assess that a model is valid?



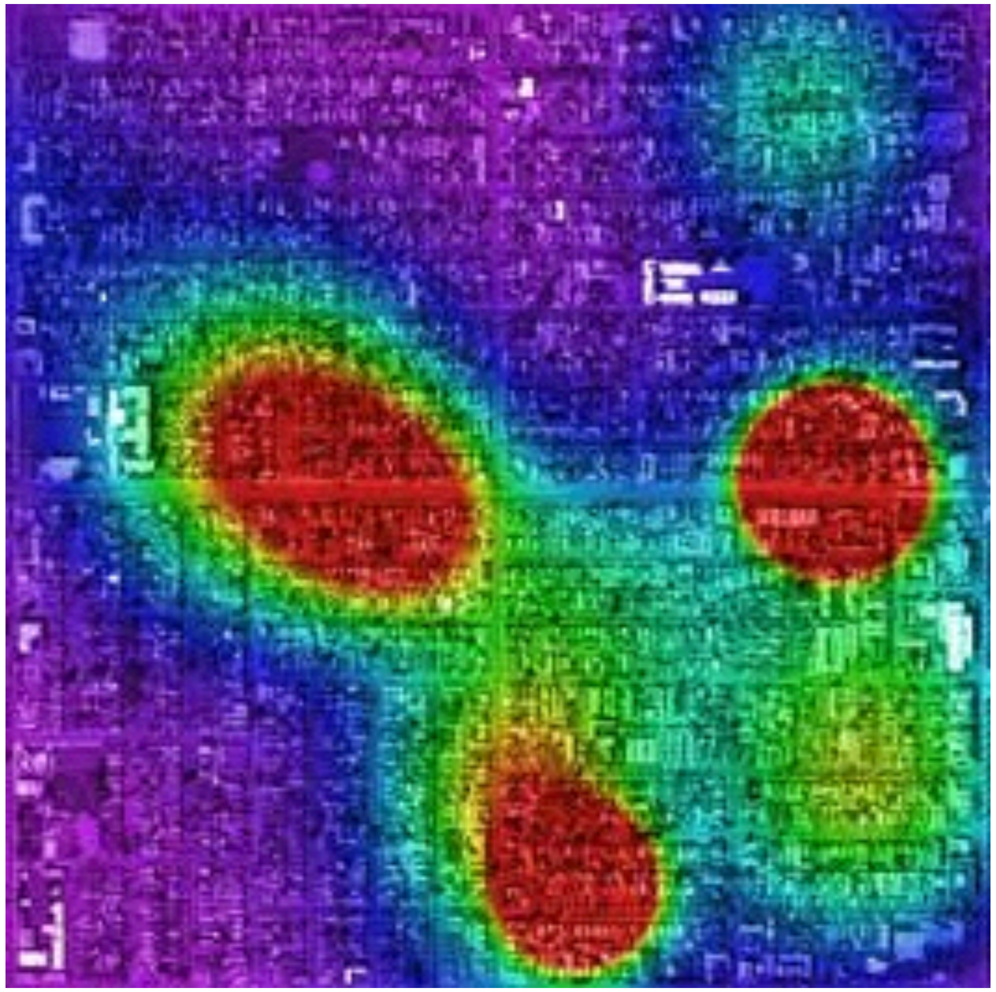
2. Transparency

How do we understand what a model is learning?



3. Fairness

To what degree does a problem translate biases in the input data into biases in its the output?



Predictive policing; heat map indicating increased risk of certain crimes
http://magazine.ucla.edu/depts/quicktakes/a-weapon-to-fight-crime-try-math/

Topics

- Overview of methods (classification, regression, clustering)
- Classification: decision trees, random forests, probabilistic models (naive bayes, logistic regression), neural networks
- Clustering: latent variable models (topic models), PCA, factor analysis, K-means, hierarchical clustering
- Linear regression
- Networks (structural properties, diffusion)
- Causal inference

Applications

- Authorship attribution
- Latent attribute prediction
- Predicting movie revenue
- Recommender systems
- Music genre classification
- Word embeddings
- Visual style classification
- Text reuse
- Genre clustering
- Predicting high school dropout rates

… in medias res

- Task: predict political preference of Twitter users.

- Assume access to training data <x, y> where:

  x = set of Twitter users
  y = {Democrat, Republican}

Representation

- How can you best represent a data point to enable learning?

TWEETS 508 | FOLLOWING 400 | FOLLOWERS 799 | LIKES 133 | LISTS 2

David Bamman

@dbamman

Assistant Professor, School of Information, UC Berkeley. Natural language processing, machine learning, computational social science, digital humanities.

Berkeley, CA

people.ischool.berkeley.edu/~dbamman/

Joined October 2009

10 Photos and videos

Tweets | Tweets & replies | Photos & videos

David Bamman Retweeted

Ted Underwood @Ted_Underwood · 6h

How have the differences between descriptions of men and women in fiction changed over the last 200 yrs? (ICYMI) tedunderwood.com/2016/01/09/the...

8 13 View summary

David Bamman @dbamman · Jan 6

"Figure Eights" (Max Roach/Buddy Rich, 1959) is just dazzling. Probably no video of them anywhere? open.spotify.com/track/23EssvWY...

View summary

David Bamman Retweeted

Anders Søgaard @soegaarducph · Jan 6

@stanfordnlp @brendan642 @jacobeisenstein Here goes: twitter-research.ccs.neu.edu/language/

x = feature vector

Feature	Value
follow clinton	0
follow trump	0
“benghazi”	0
negative sentiment + “benghazi”	0
“illegal immigrants”	0
“republican” in profile	0
“democrat” in profile	0
self-reported location = Berkeley	1

$$\sum_{i=1}^{F} x_i \beta_i = x_1\beta_1 + x_2\beta_2 + \ldots + x_F\beta_F$$

$$= x^\top \beta \qquad \text{(dot product, inner product)}$$

$$\hat{y}_i = \begin{cases} 1 & \text{if } \sum_i^F x_i\beta_i \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

x = feature vector

Feature	Value
follow clinton	0
follow trump	0
“benghazi”	0
negative sentiment + “benghazi”	0
“illegal immigrants”	0
“republican” in profile	0
“democrat” in profile	0
self-reported location = Berkeley	1

β = coefficients

Feature	β
follow clinton	-3.1
follow trump	6.8
“benghazi”	1.4
negative sentiment + “benghazi”	3.2
“illegal immigrants”	8.7
“republican” in profile	7.9
“democrat” in profile	-3.0
self-reported location = Berkeley	-1.7

	"benghazi"	follows trump	follows clinton	$\sum x_i \beta_i$	prediction
β	1.4	6.8	-3.1		
x^1	1	1	0	8.2	1
x^2	0	0	1	-3.1	-1
x^3	1	0	1	-1.7	-1

$$(1 \times 1.4) + (1 \times 6.8) + (0 \times -3.1) = 8.2$$

Learning

How do get good values for β?

Feature	β
follow clinton	-3.1
follow trump	6.8
“benghazi”	1.4
negative sentiment + “benghazi”	3.2
“illegal immigrants”	8.7
“republican” in profile	7.9
“democrat” in profile	-3.0
self-reported location = Berkeley	-1.7

Online learning

- Go through the training data <x, y> one data point at a time.
- Make a prediction $\hat{y}$ with current estimate of β; if it's right ($y = \hat{y}$), do nothing.
- If the prediction is wrong ($y \neq \hat{y}$), change β to make it slightly less wrong.

$$\hat{y}_i = \begin{cases} 1 & \text{if } \sum_i^F x_i \beta_i \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

"benghazi"	follows trump	follows clinton	y
1	1	0	1
0	0	1	-1
1	0	1	-1

training data

$$\hat{y}_i = \begin{cases} 1 & \text{if } \sum_i^F x_i \beta_i \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

"benghazi"	follows trump	follows clinton	y	ŷ
1	1	0	1	1
0	0	1	-1	-1
1	1	1	1	-1

true y = -1
predicted ŷ = 1

$$\sum_{i}^{F} x_i \beta_i$$

We want this value (function of β) to be small

$$\frac{\partial}{\partial \beta_i} \sum_{i}^{F} x_i \beta_i = x_i$$

The derivative tells us the direction to go to make it bigger or smaller

$$\beta_{t+1} = \beta_t - x$$

Update rule

true   y = 1
predicted ŷ = -1

$$\sum_{i}^{F} x_i \beta_i$$

We want this value (function of β) to be big

$$\frac{\partial}{\partial \beta_i} \sum_{i}^{F} x_i \beta_i = x_i$$

The derivative tells us the direction to go to make it bigger or smaller

$$\beta_{t+1} = \beta_t + x$$

Update rule

if $\hat{y} = 1$ and $y = -1$

$$\beta_{t+1} = \beta_t - x$$

	β_t	x	β_{t+1}
	3.6	0	3.6
	3.4	1	2.4
	1.2	1	0.2
	0.7	0	0.7
$\sum x_i \beta_i$	4.6		2.6
$\hat{y}$	1		1

if $\hat{y} = -1$ and $y = 1$

$$\beta_{t+1} = \beta_t + x$$

	β_t	x	β_{t+1}
	3.6	0	3.6
	-3.4	1	-2.4
	1.2	1	2.2
	0.7	0	0.7
$\sum x_i\beta_i$	-2.2		-0.2
$\hat{y}$	-1		-1

if $\hat{y} = 1$ and $y = -1$

$$\beta_{t+1} = \beta_t - x$$

if $\hat{y} = -1$ and $y = 1$

$$\beta_{t+1} = \beta_t + x$$

$$\beta_{t+1} = \beta_t + yx$$

Why $\beta_{t+1} = \beta_t + yx$?

[Approximation of stochastic gradient in binary logistic regression (lecture 9)]

Perceptron

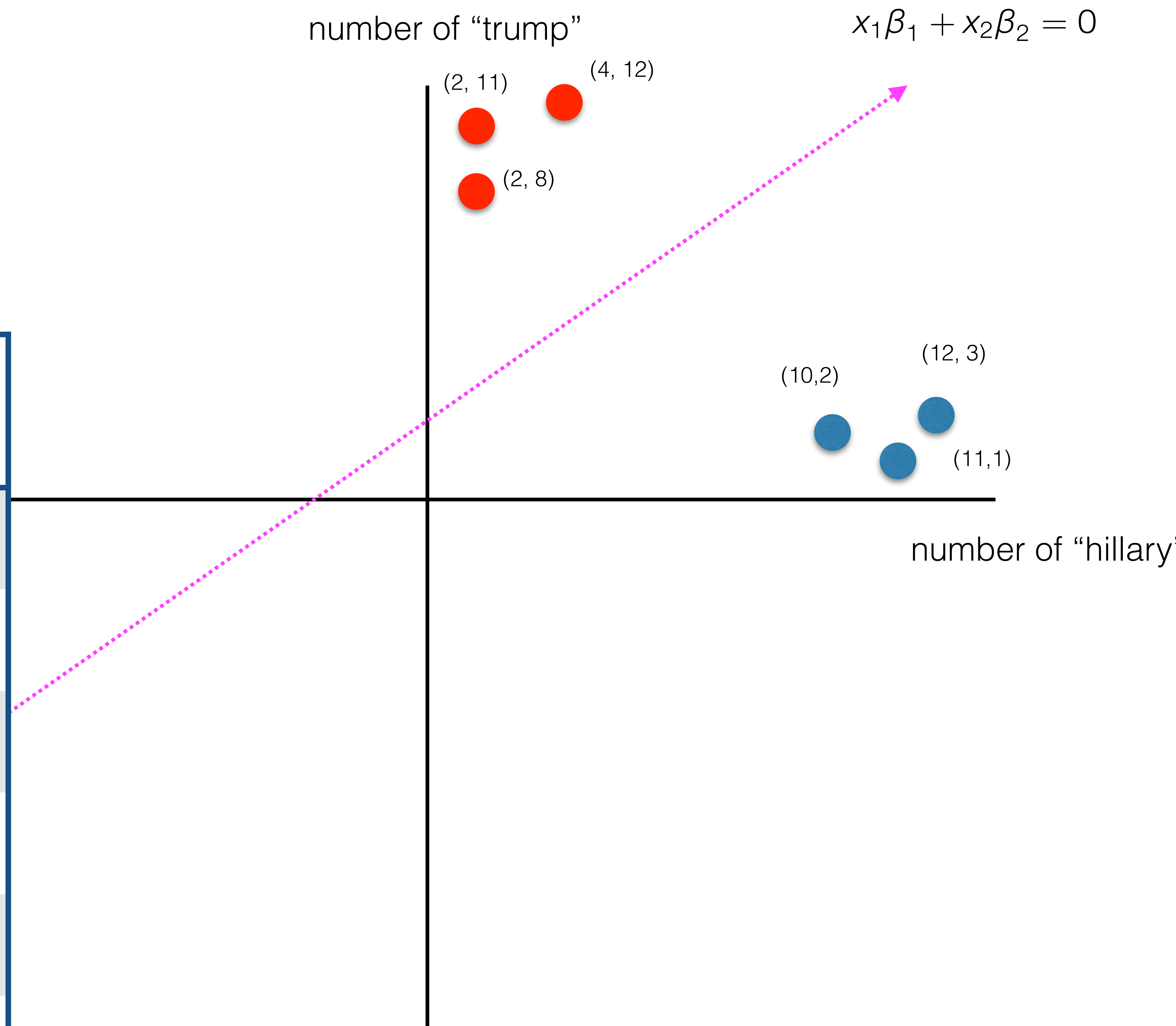
Data: training data $x \in R^F$, $y \in \{-1, +1\}$, i = 1 ... N;
initialize $\beta_0 = 0^F$;
k=0;
while *not converged* **do**
    k = k + 1;
    i = k mod N;
    if $\hat{y}_i \neq y_i$ **then**
        $\beta_{t+1} = \beta_t + y_i x_i$
    else
        do nothing;
    end
end

Rosenblatt 1957

Code

decision boundary in 2 dimensions

# of "hillary"	# of "trump"	y
2	8	1
2	11	1
4	12	1
10	2	-1
11	1	-1
12	3	-1



Trends

- Counts later points more than earlier points (voted perceptron, averaged perceptron)
- Only linear decision boundaries
- Prone to overfitting
- Extraordinarily simple and accurate classifier

Problem assumptions

- Is this the right task (classification vs. clustering vs. regression, time series forecasting etc.)
- Is the data appropriate for the problem?

Administrivia

- David Bamman
  dbamman@berkeley.edu

  Office hours: Thursdays 10am-noon, 314 SH
  — or by appointment

- Rob Kuvinka, TA

  Office hours: Tuesday, 5-7pm, 110 South Hall

Grading

- Class participation (10%)
- Homeworks (4 x 12.5%)
- Project (40%)

All deliverables (homeworks, project components) have deadlines; late work not accepted after 2 “free days” used up

Free days

- You have a total of 2 “free days” to use over the entire semester
- Each free day gives you an extra 24 hours to turn in a homework assignment
- A free day is used up once you cross the deadline for a homework being due (e.g., 12:01am for a 12:00am deadline)
- Use them wisely!

Homeworks, broadly

A

- Implement a quantitative method and evaluate it on a dataset

B

- Write an analysis/critique of an algorithm and published work that has used it

Homework Example

Binary perceptron classifies into two classes. For inferring political preference, this corresponds to a simple {Democrat, Republican} distinction. Assume rather that the training data you have is hierarchical. Design a perceptron-style algorithm that can exploit this hierarchical structure during learning.

y1	Republican > Tea Party Republican
y2	Republican > Social Conservatives
y3	Republican > Neoconservative
y4	Republican > Social Conservative
y5	Democrat > Centrist Democrat
y6	Democrat > Progressive

A

Code and evaluate on test data

B

What are the comparative advantages and disadvantages of binary vs. multiclass vs. hierarchical categories? Under what circumstances should either be used? (2 pages, single-spaced)

Participation

- Most classes will include discussion of an application as documented in a research paper.
- While everyone is expected to read these papers, one student each class will act as a discussion leader, coming prepared with questions and discussion topics for the class a whole to discuss.

Project

- Use methods learned in class to draw inferences about the world and critically assess the quality of the results.

- Collaborative (2-3 students). Choose wisely! Everyone in group will receive the same grade; you will be evaluated both on the empirical methodology and the domain questions you're asking

Project

- Milestones:
  - Proposal and literature review (5%). 2 pages, 5 sources.
  - Midterm report (10%). 4 pages, 10 sources.
  - Final report (20%). 10 pages.
  - Presentation (5%). 15-20 min. conference-style talk in front of peers.
- Evaluated according to standards for conference publication—clarity, originality, soundness, substance, evaluation, meaningful comparison, impact.