Data Science as an Organizing System

David Bamman

Info 202: Information Organization and Retrieval

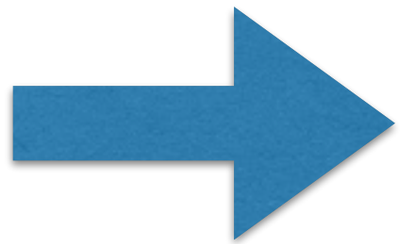
Lecture 2, August 29, 2016

what is data science?

“data science”

raw data

algorithm



knowledge

- a lot of data science is focused here

Algorithms

- Classification: decision trees, random forests, probabilistic models (naive bayes, logistic regression), SVM, neural networks

- Clustering: latent variable models (topic models), PCA, factor analysis, K-means, hierarchical clustering

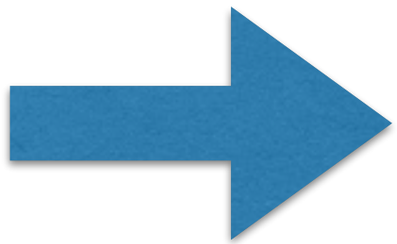
- Linear regression

- Networks (structural properties, diffusion)

- Temporal data: time series forecasting and survival analysis

“data science”

- causal inference



knowledge

raw data

- what’s the right data to analyze?
- which aspects of it?

algorithm

- what assumptions underlie the methods?

- what’s the right question to ask?

what is data science?

- Data science involves empirical sensemaking (learning from observations/experience)

- Algorithms/methods are one half of this; but equally important are the fundamental choices that go into the design of experiments.

- How do we design an experiment that can use data to answer some question of interest?

data science as information organization

- The selection of data
- The description of data
- Leveraging relationships between data points
- To enable interactions: classification, prediction, recommendation, inference, hypothesis testing

Data Science

software

algorithms

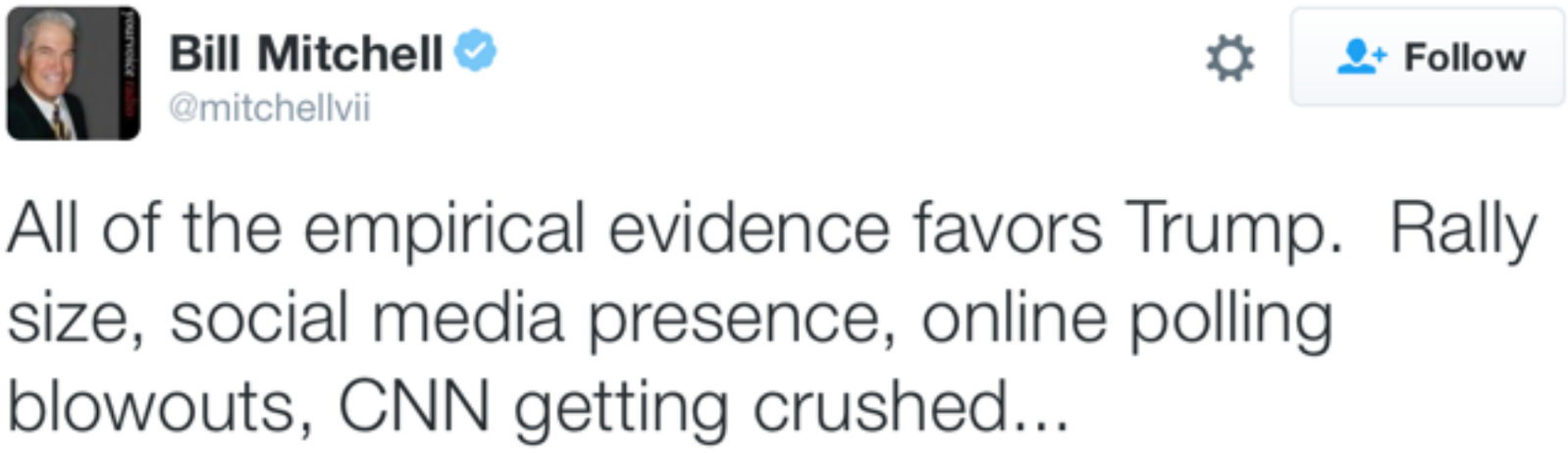
classification, regression, clustering, network analysis, prediction, hypothesis testing,

critical thinking

data selection, representation, experimental design, validation

two case studies

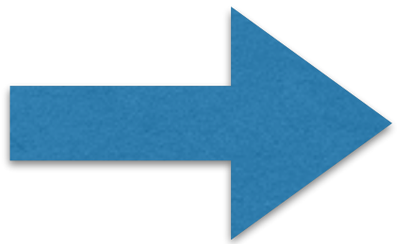
case study: predicting elections



case study: predicting elections

- Goal: predict the future (the outcome of an election)
- Many resources we can marshall to make this prediction.
  - Descriptive: call people up and ask them (which people?)
  - Some polls, in retrospect, are better predictors than others; consider many polls in one model and weight accordingly (538)
  - Other features are also better predictors than others (e.g., incumbency, historical state voting). Twitter followers? Rally size?

“data science”



knowledge

~~raw~~ data

algorithm

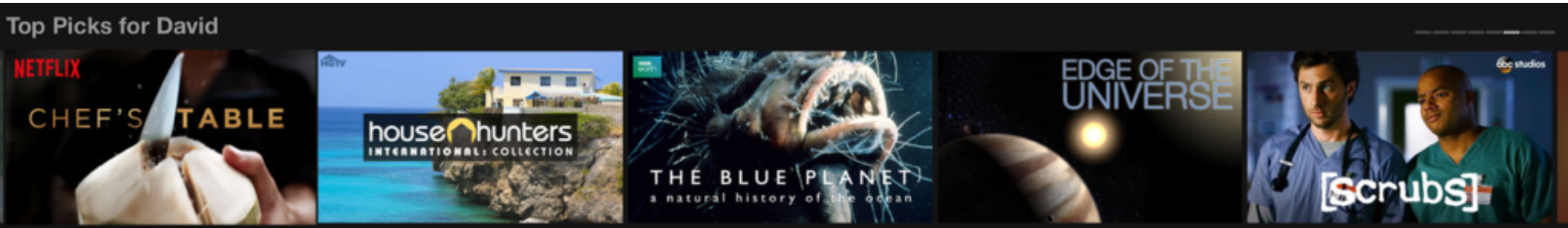
- mediated
- selection criteria
- multiple (noisy) sources

case study: predicting elections

Information organization here involves selecting data and describing it to enable an interaction: prediction

- what is being organized?
- why is it being organized?
- how much is it being organized?
- when is it being organized?
- how (or by whom) is it being organized?
- where is it being organized?

case study: recommendation systems



Goal: recommend other items that users will rate favorably/buy

case study: recommendation systems

- Many resources we can marshall to make this prediction.
  - Descriptions of the items themselves
  - Data points given to us by company catalog
  - But considerable flexibility in resource description

case study: recommendation systems

- Many resources we can marshall to make this prediction.
  - Users who rate movies
  - Recommend movies through the relationships they hold to the people who watch them.

case study: recommendation systems

Information organization here involves selecting and describing data, leveraging relationships among data points to enable an interaction: recommendation

- what is being organized?
- why is it being organized?
- how much is it being organized?
- when is it being organized?
- how (or by whom) is it being organized?
- where is it being organized?