Activities in Data Science II

David Bamman
Info 202: Information Organization and Retrieval

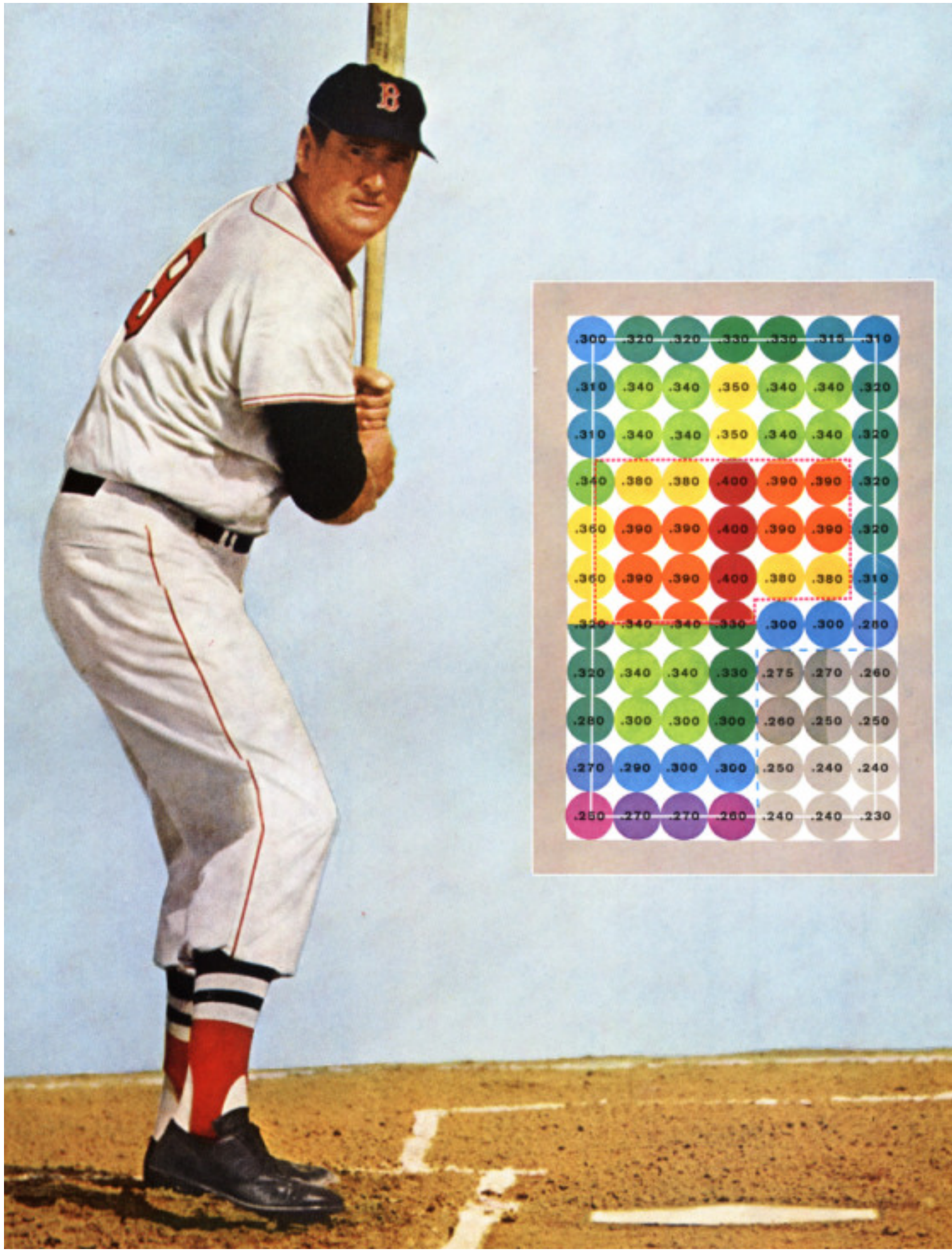
Lecture 4, September 7, 2016

Activities

- Selecting resources
- Organizing resources
- Designing resource-based interactions
- Maintaining resources
- Data science as an organizing system vs. organizing via data science

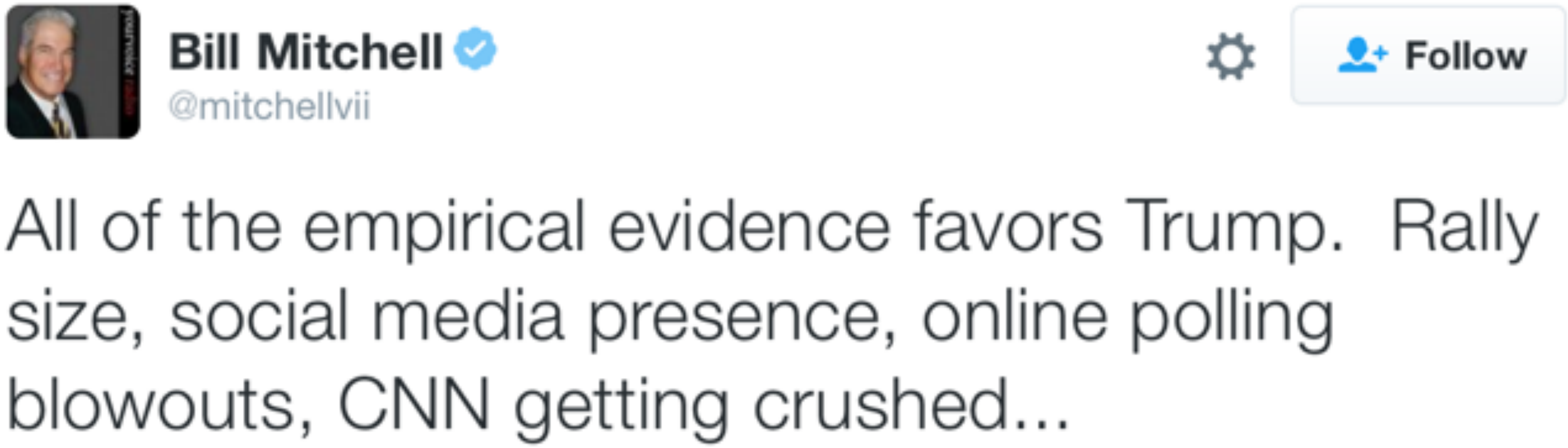
Selection via data

- Moneyball effect. Personnel selection is becoming data-driven
- Hiring decisions based on features that make “successful employees”



http://www.hardballtimes.com/wp-content/images/tht/williamsgraphic.png

Selection in data science



Bill Mitchell
@mitchellvii

All of the empirical evidence favors Trump. Rally size, social media presence, online polling blowouts, CNN getting crushed...

Selection in data science

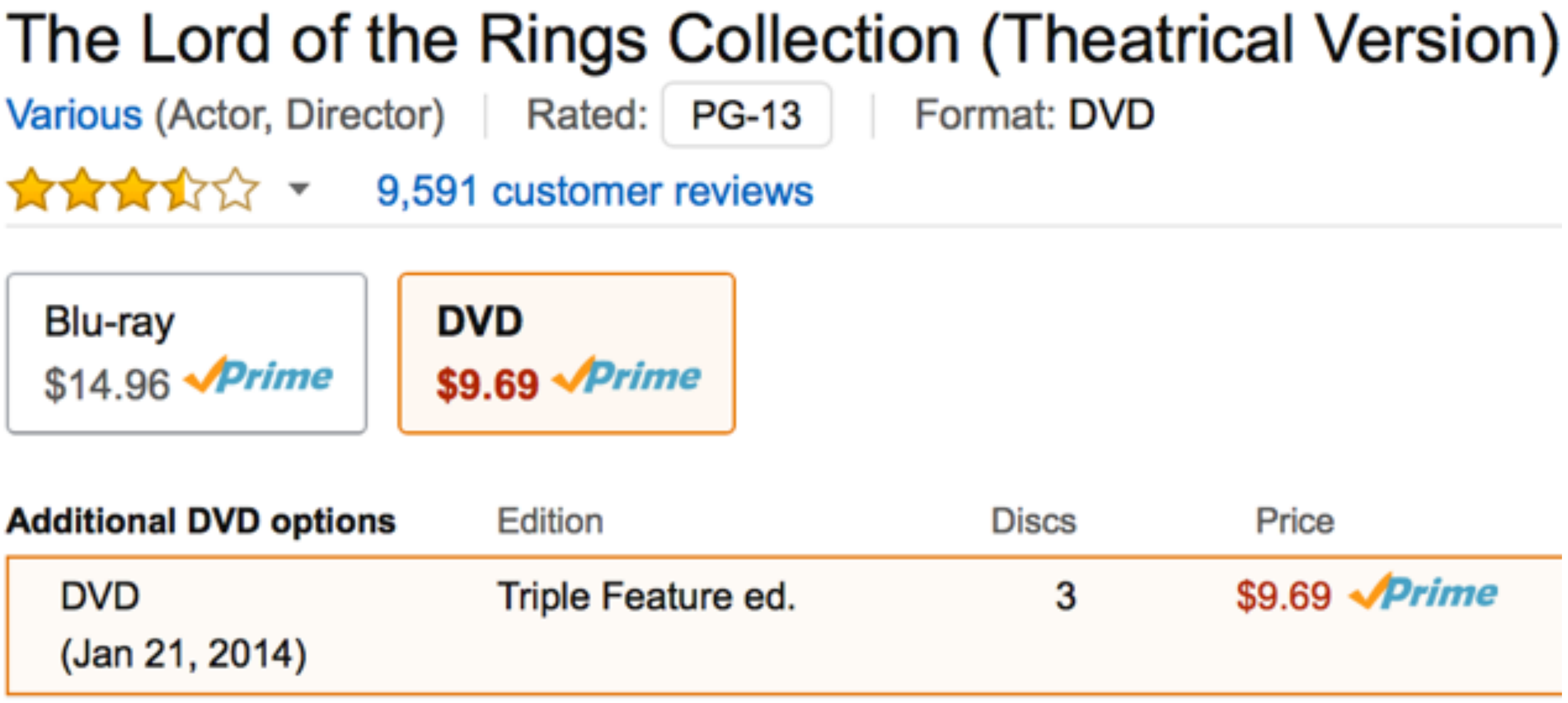
- Most analyses use a sample of a target population
- Sources of sample bias in data selection:
    - Misspecification the target population
    - Sample size too small for generalization
    - Selection bias: some members of the target population are more/less likely to be included.
    - Response bias: some members of the target sample are more/less likely to exclude themselves

Selection in data science

- Data quality as selection criterion
    - How far upstream can you track the provenance?
    - Anomalous/duplicate data
    - What external checks do you have on its quality?

Organizing via data

- Descriptive statistics
  - count
  - central tendency (mean, median, mode)
  - variance



The Lord of the Rings Collection (Theatrical Version)

Various (Actor, Director) | Rated: PG-13 | Format: DVD

★★★½☆ 9,591 customer reviews

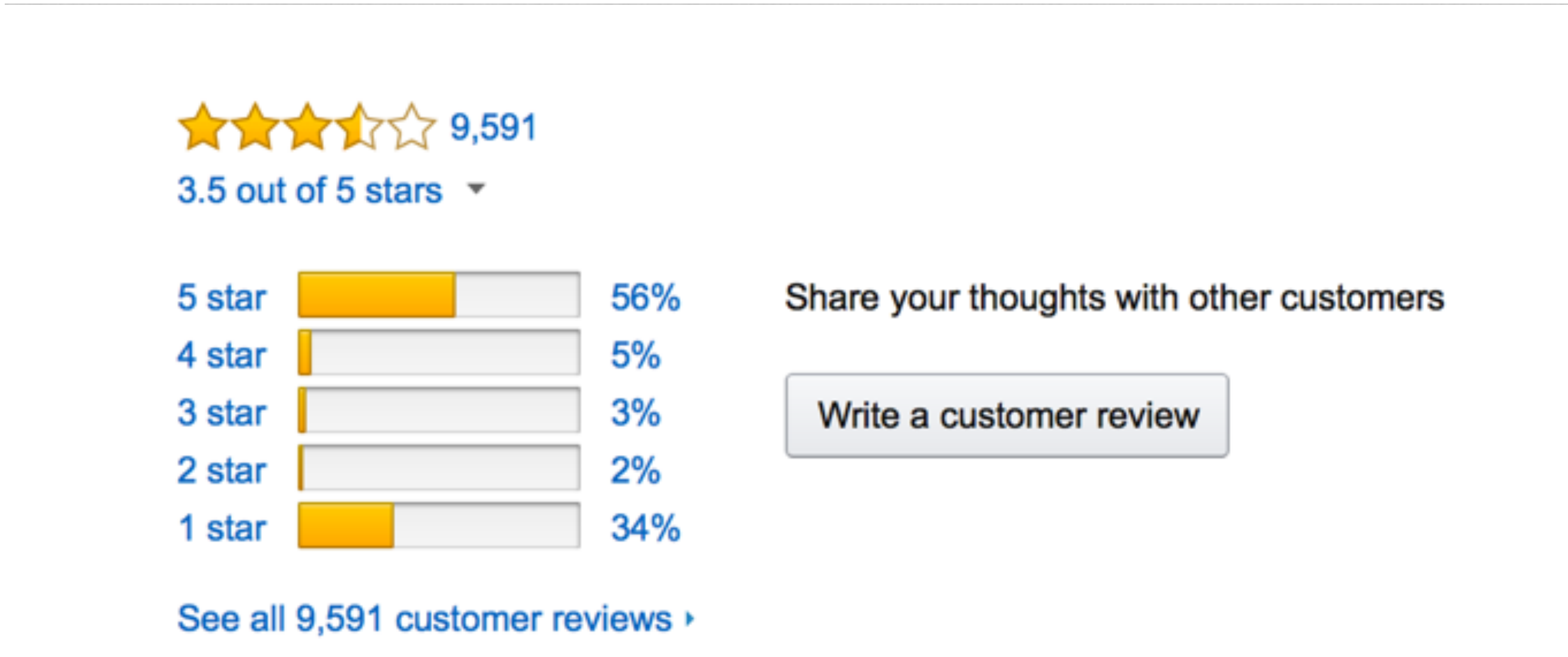
Blu-ray $14.96 ✓Prime

DVD **$9.69** ✓Prime

Additional DVD options	Edition	Discs	Price
DVD (Jan 21, 2014)	Triple Feature ed.	3	$9.69 ✓Prime

Organizing via data

- Descriptive statistics
  - count
  - central tendency (mean, median, mode)
  - variance



9,591

3.5 out of 5 stars

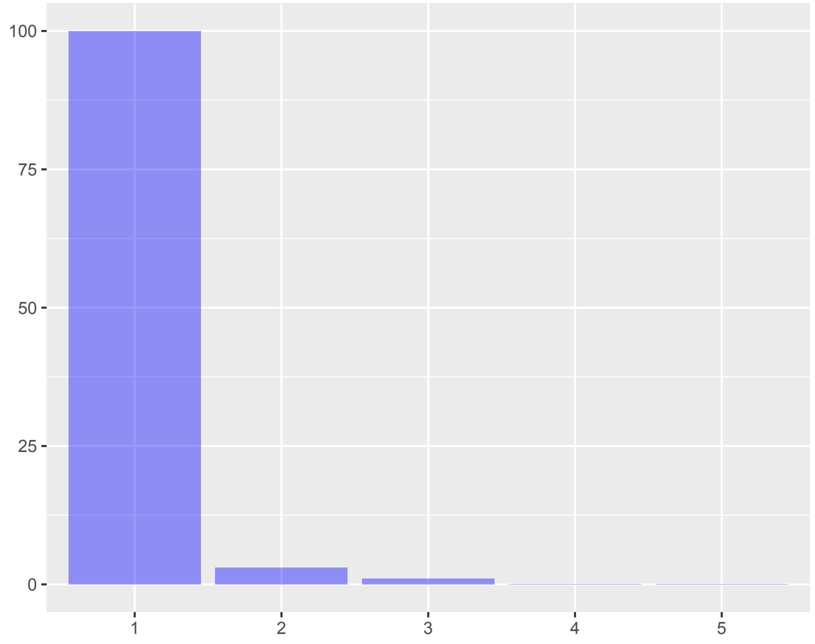
5 star	56%
4 star	5%
3 star	3%
2 star	2%
1 star	34%

See all 9,591 customer reviews

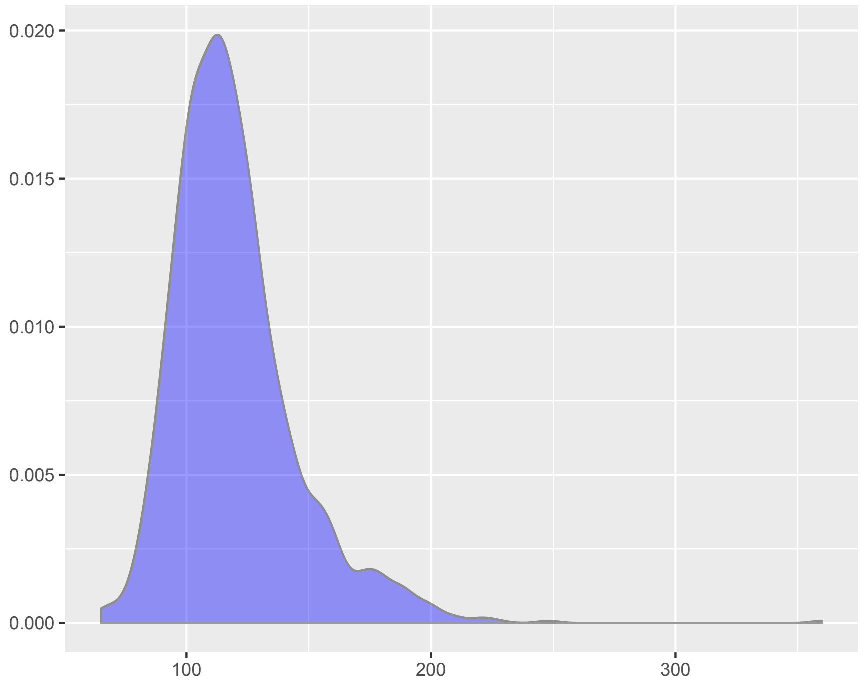
Share your thoughts with other customers

Organizing via data

- Descriptive statistics in exploratory data analysis can be used to understand the distribution of values for a property to determine whether it can be used to organize the data



number of discs



runtime

Exploratory data analysis

- Exploring the data can help inform the application you put it to

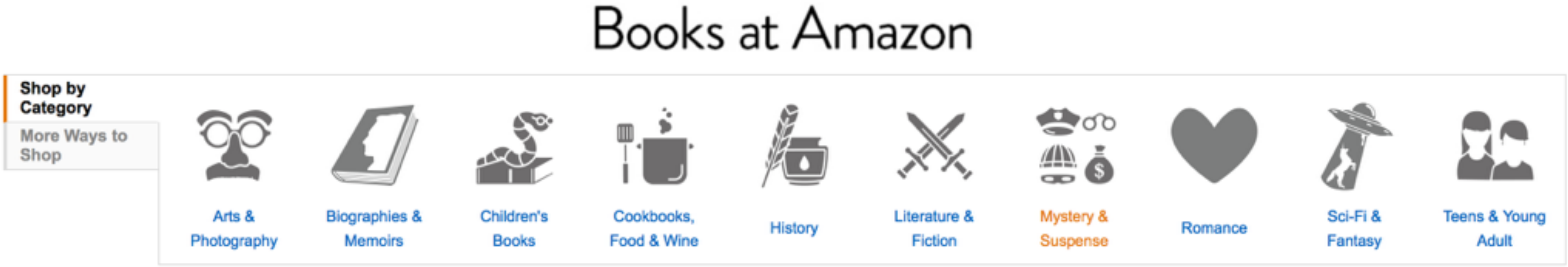
- e.g., understanding its central tendency (mean, variance), the structure between variables (correlations, PCA), visualizations.

- Complement to classical hypothesis testing (though need to take care to keep the two separate on the same data).

Tukey (1977), Exploratory Data Analysis

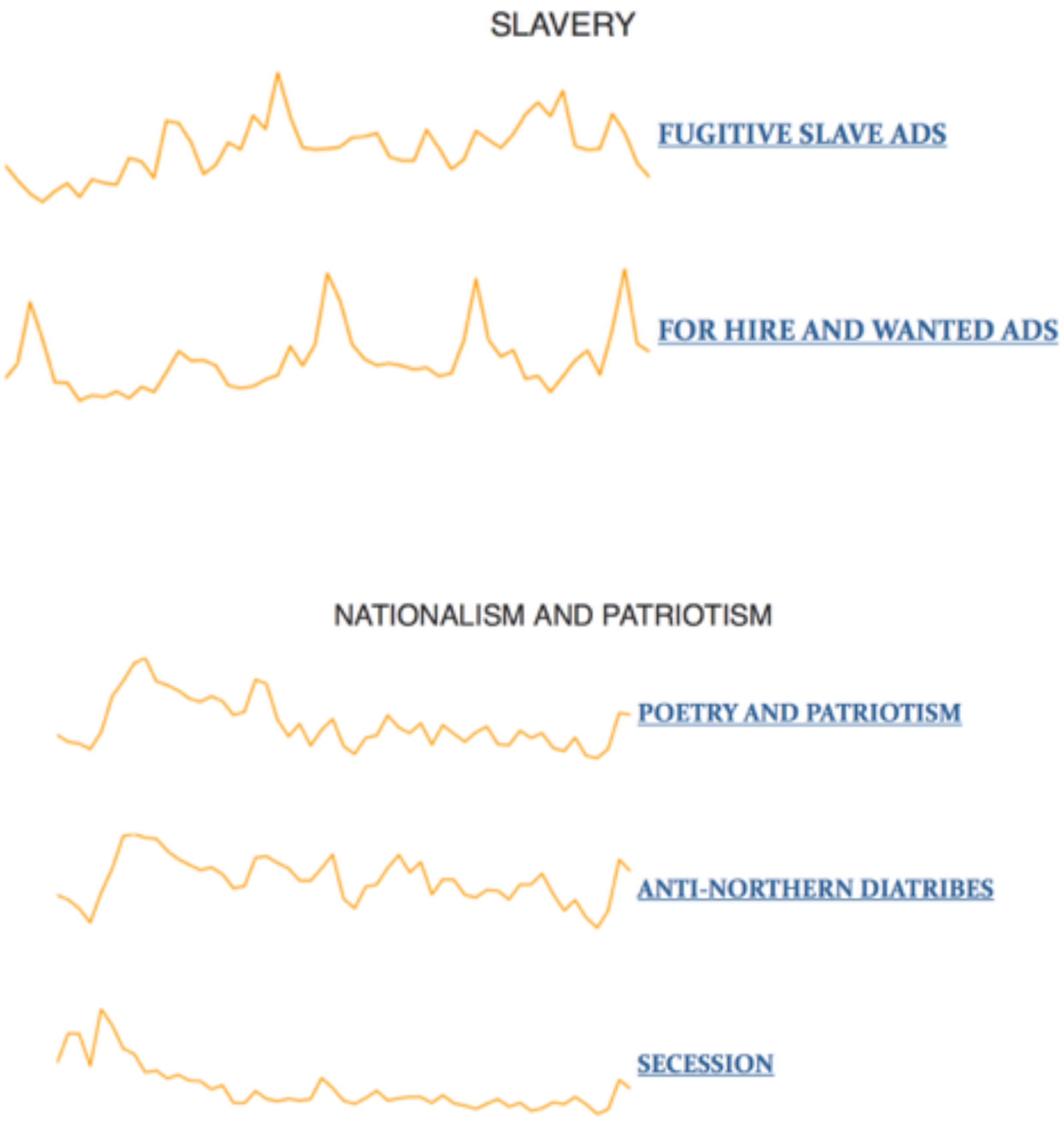
Organizing via data

- Organizing via explicit attributes
- Organization imposed through human-labeled subject/topic codes



Organizing via data

- Organizing via latent attributes
- Organization facilitated by the machine identification of “topics”



Mining the Richmond Times Dispatch (1860-1865)
http://dsl.richmond.edu/dispatch/Topics

Designing Interactions

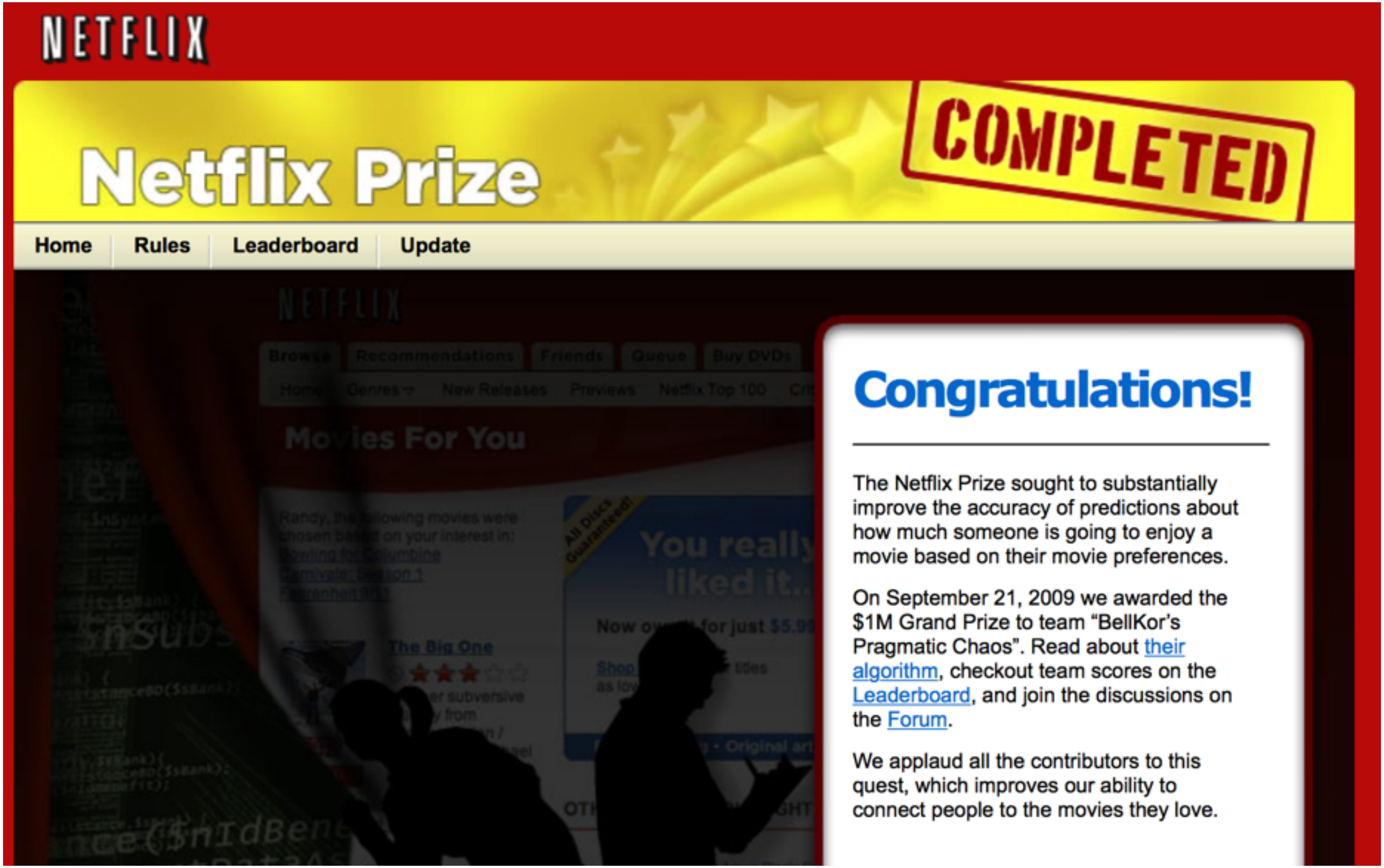
- Prediction
- Inference
- Recommendation
- Hypothesis testing

Access policies

- Who is allowed to access the data we're analyzing?
    - Public domain (US gov't data)
    - Proprietary data (corporate, IP, etc.)
    - Data restricted by privacy/policy/ethical considerations
- Distinction between restricted access to data and restricted uses of it.
    - Decisionmaking based on restricted attributes (gender, ethnicity, etc.)

How Companies Learn Your Secrets

By CHARLES DUHIGG FEB. 16, 2012



Congratulations!

The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.

On September 21, 2009 we awarded the $1M Grand Prize to team "BellKor's Pragmatic Chaos". Read about their algorithm, checkout team scores on the Leaderboard, and join the discussions on the Forum.

We applaud all the contributors to this quest, which improves our ability to connect people to the movies they love.

Maintaining Resources

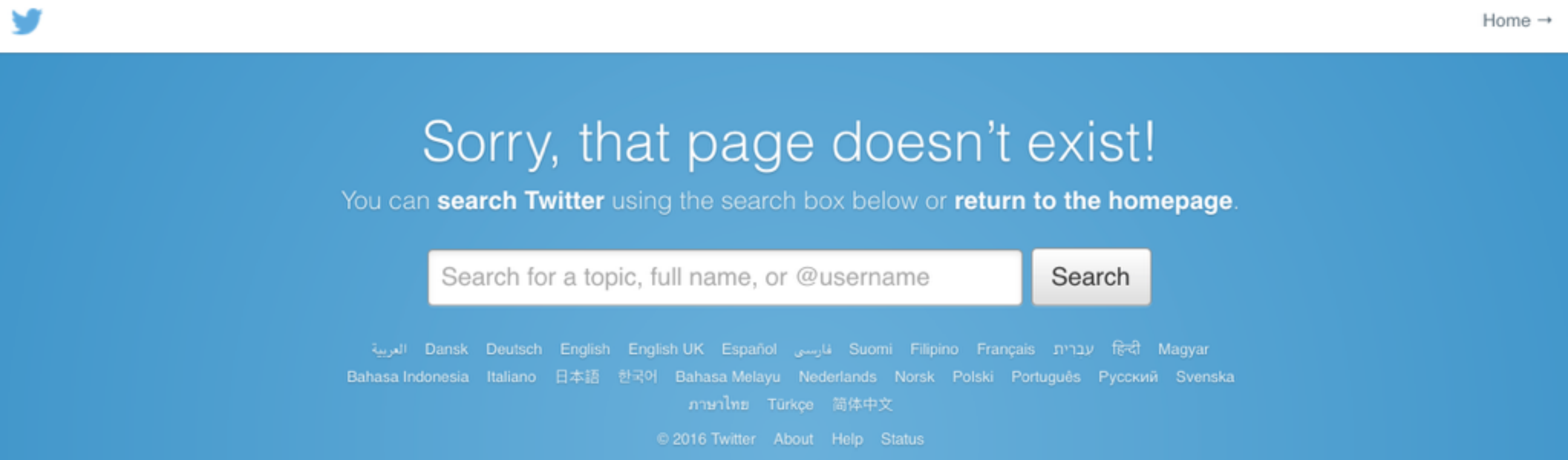
- Storage
- Preservation
- Curation
- Governance

Preservation

- Replicability of analysis
  - The original data/methodology should be preserved so that an analysis can be recreated.
  - Proper techniques so that same methodology used on similar but new data should yield similar results (generalizability).

Preservation

- Recreating analyses on Twitter data



Home →

Sorry, that page doesn't exist!

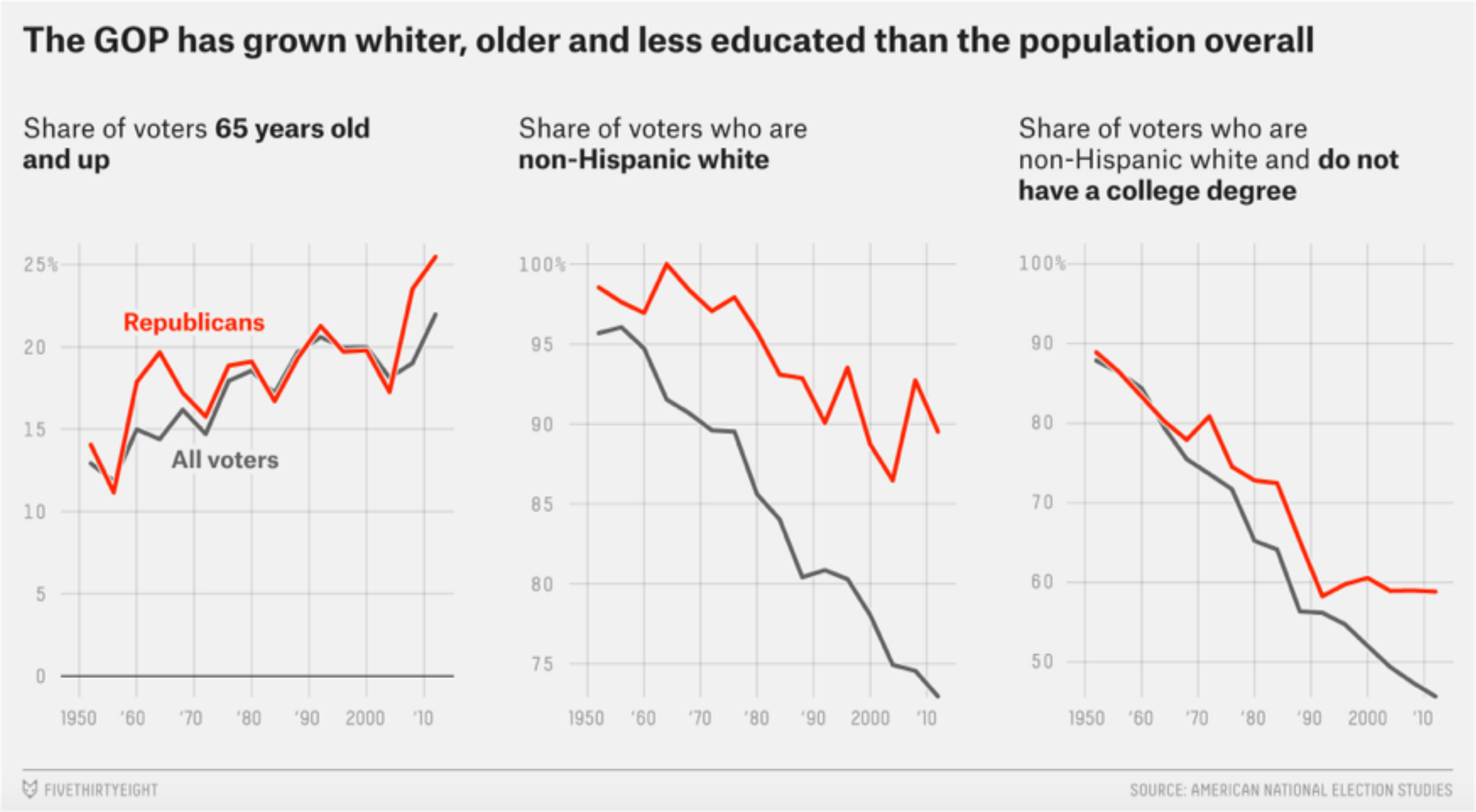
You can **search Twitter** using the search box below or **return to the homepage**.

Search for a topic, full name, or @username
Search
العربية Dansk Deutsch English English UK Español فارسی Suomi Filipino Français עברית हिन्दी Magyar
Bahasa Indonesia Italiano 日本語 한국어 Bahasa Melayu Nederlands Norsk Polski Português Русский Svenska
ภาษาไทย Türkçe 简体中文
© 2016 Twitter About Help Status

Curation: Drift

- Resources in data analysis are often a sample of the population
- What you care about is not the specific data points you have, but the population they represent
- Semantic drift

Curation: Drift



The GOP has grown whiter, older and less educated than the population overall

Share of voters **65 years old and up**

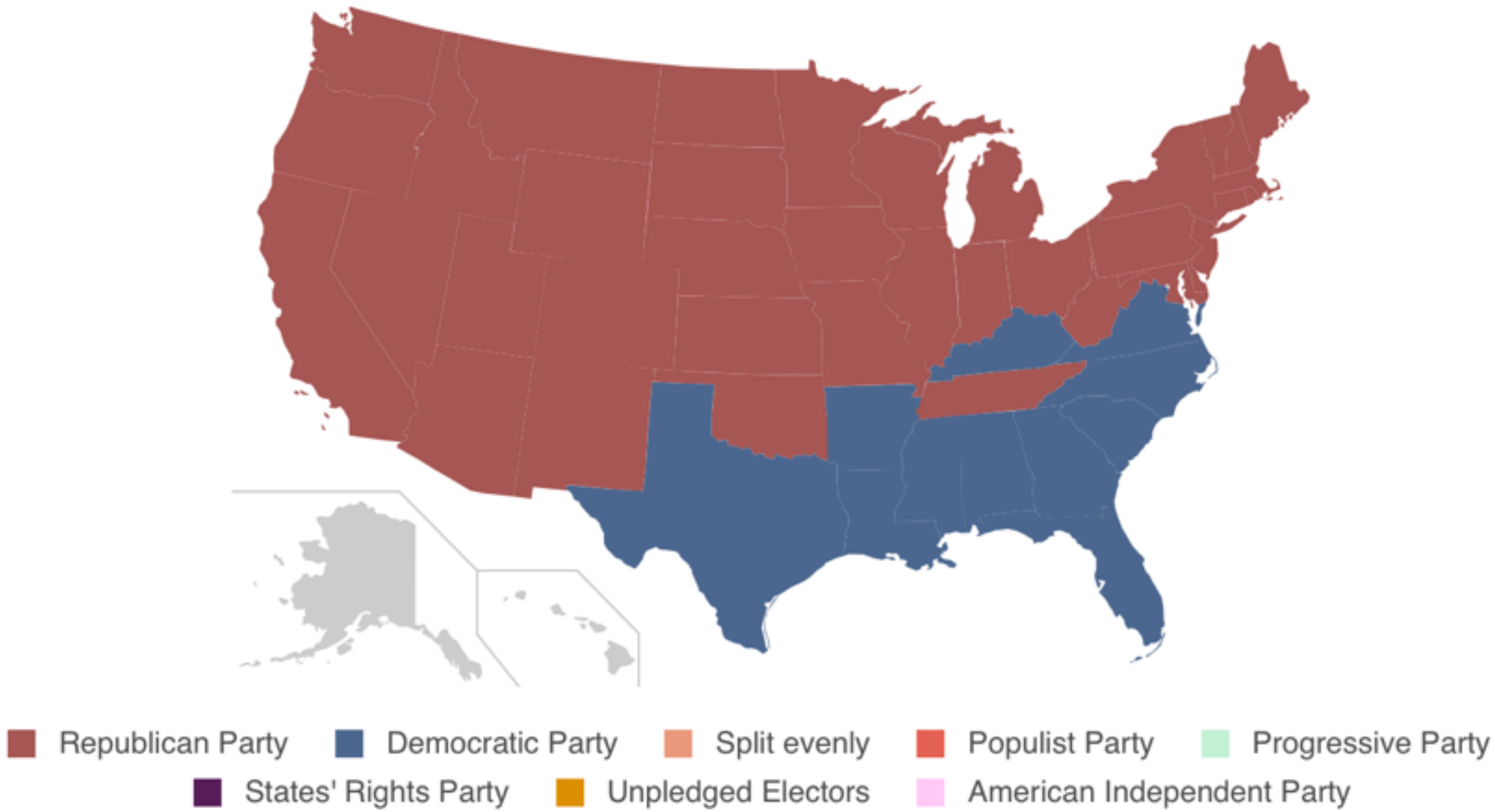
Share of voters who are **non-Hispanic white**

Share of voters who are non-Hispanic white and **do not have a college degree**

SOURCE: AMERICAN NATIONAL ELECTION STUDIES

http://fivethirtyeight.com/features/the-end-of-a-republican-party/

Curation: Drift



1920 presidential election

http://us-presidents.insidegov.com/stories/3613/republicans-democrats-switch-platform

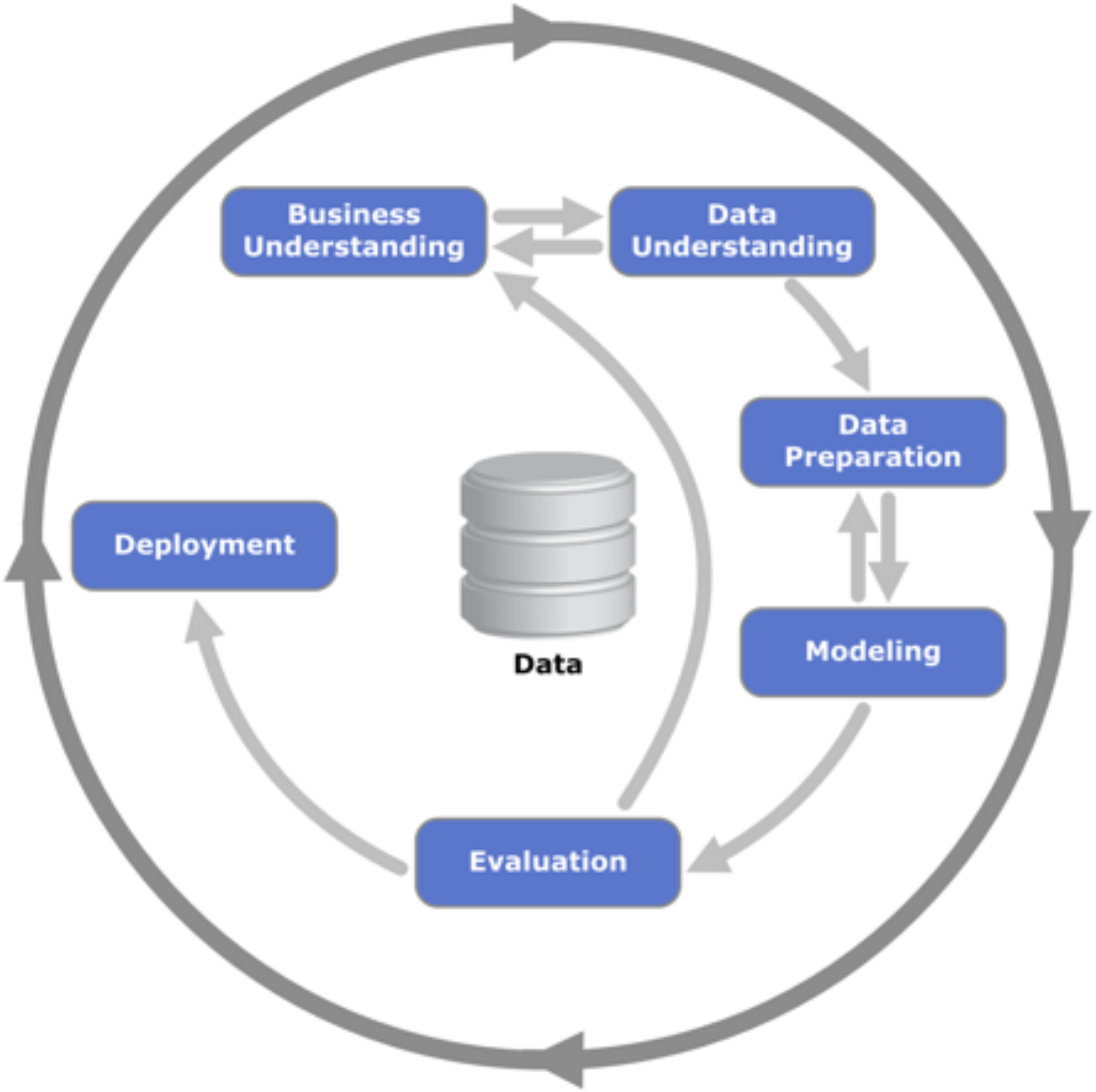
Organizing system lifecycle

- Defining and scoping the domain
- Identifying requirements
- Design and implementation
- Operations and maintenance

Organizing system lifecycle

- Defining and scoping the domain
- Identifying requirements
- Design and implementation
- Operations and maintenance

Data science lifecycle

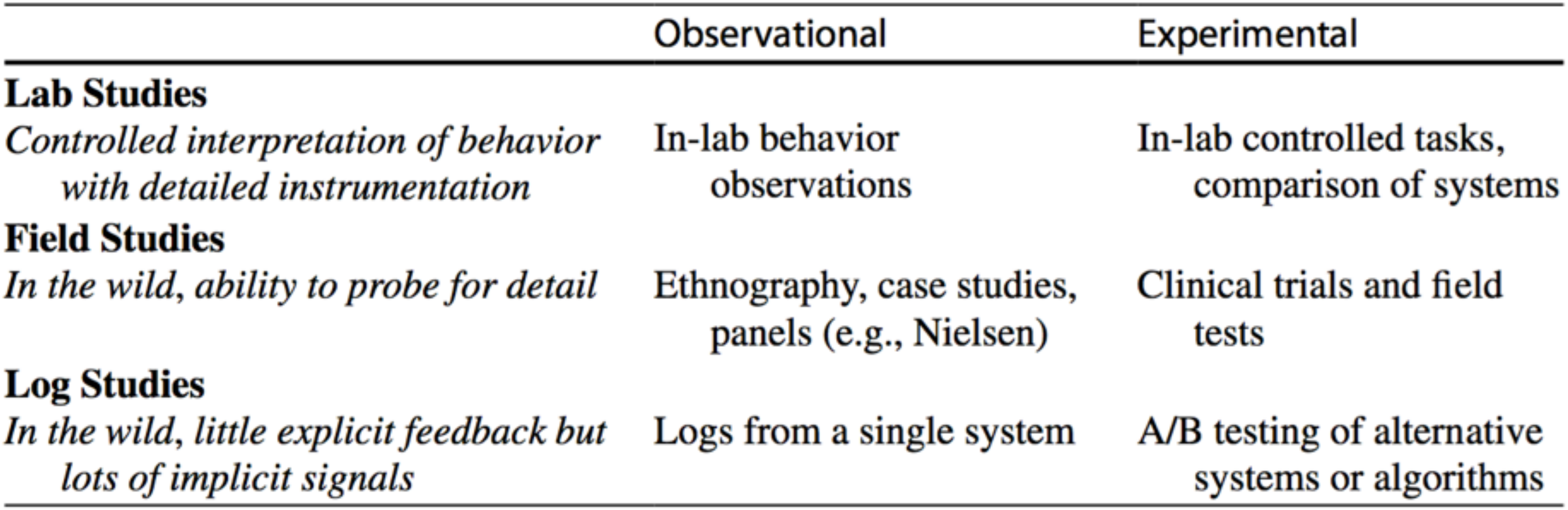


Cross Industry Standard Process for Data Mining (CRISP-DM)

Data science lifecycle

- The goal of data analysis drives the organizing activities you make use of.
    - what data to collect
    - how to describe it
    - how to
- Data collected/employed for other goals may have hidden biases lurking with it.

Case study: Log file analysis



	Observational	Experimental
Lab Studies *Controlled interpretation of behavior with detailed instrumentation*	In-lab behavior observations	In-lab controlled tasks, comparison of systems
Field Studies *In the wild, ability to probe for detail*	Ethnography, case studies, panels (e.g., Nielsen)	Clinical trials and field tests
Log Studies *In the wild, little explicit feedback but lots of implicit signals*	Logs from a single system	A/B testing of alternative systems or algorithms

Dumais et al. (2014), “Understanding User Behavior Through Log Data and Analysis”

Case study: Log file analysis

- Data collection
    - What data to collect?
    - Recording time
    - Identifying users (cookies, IP addresses, logins)
- Data cleaning
    - Missing data
    - Data transformations
    - Outliers

Dumais et al. (2014), "Understanding User Behavior Through Log Data and Analysis"

Audit trail (traceability)

- As in other organizing systems, preserving the chain of decisions make can improve reproducibility and trust in an analysis.

- Trust extends to the interpretability of algorithms