data-ppf.github.io apr 2, 2019

lecture 10 of 14: data science, 1962-2017

chris wiggins + matt jones, Columbia

themes for today (1962-2018)

themes for today (1962-2018)

1. industry and academia

themes for today (1962-2018)

1. industry and academia

- academic power, “jobs” power

themes for today (1962-2018)

1. industry and academia
   - academic power, "jobs" power
2. truth, people, and practice

themes for today (1962-2018)

1. industry and academia
   - academic power, "jobs" power
2. truth, people, and practice
   - constant theme of this course: data as rhetorical claim

themes for today (1962-2018)

1. industry and academia
   - academic power, "jobs" power
2. truth, people, and practice
   - constant theme of this course: data as rhetorical claim
3. where in "STEM" is "data"?

themes for today (1962-2018)

1. industry and academia
   - academic power, "jobs" power
2. truth, people, and practice
   - constant theme of this course: data as rhetorical claim
3. where in "STEM" is "data"?
   - inconvenient truths this week: truth is negotiated

historical & social context

- industrial data powers

historical & social context

- industrial data powers
- academic mathematical statistics

historical & social context

- industrial data powers
- academic mathematical statistics
  - recall neyman/fisher 1955-1956

historical & social context

- industrial data powers
- academic mathematical statistics
  - recall neyman/fisher 1955-1956
- rise of data outside silos

historical & social context

- industrial data powers
- academic mathematical statistics
  - recall neyman/fisher 1955-1956
- rise of data outside silos
  - EPA, ETS, etc. . . ..

historical & social context

- industrial data powers
- academic mathematical statistics
  - recall neyman/fisher 1955-1956
- rise of data outside silos
  - EPA, ETS, etc. . . ..
- "data science" 2001, 2010, ...

contemporary context/modern day relevance

- data science everywhere as Donoho says

student observations

student observations

This week's readings felt like an examination of the identity crisis of the field of data science.

- role of truth and subjectivity

what are the new capabilities this week?

- ML as “technology”

power

- academic

power

- academic
- jobs

readings: Tukey, Breiman, Donoho, Neff

Tukey: bio

- mathematician

Tukey: bio

- mathematician
- turned statistician

Tukey: bio

- mathematician
- turned statistician
- split career

Tukey: bio

- mathematician
- turned statistician
- split career
  - consulting

Tukey: bio

- mathematician
- turned statistician
- split career
  - consulting
  - intelligence work

Tukey's FoDA

- opens with imposter line

Tukey's FoDA

- opens with imposter line
- attack on mathematization

Breiman: bio

- mathematician

Breiman: bio

- mathematician
- turned consultant

Breiman: bio

- mathematician
- turned consultant
- turned computational statistician!

Breiman's two cultures

- C.P. Snow reference

Breiman's two cultures

- C.P. Snow reference
- generative v. predictive

Breiman's two cultures

- C.P. Snow reference
- generative v. predictive
- predictive v. interpretable

Breiman's two cultures

- C.P. Snow reference
- generative v. predictive
- predictive v. interpretable
- what is "best" ML model?

Breiman's two cultures

- C.P. Snow reference
- generative v. predictive
- predictive v. interpretable
- what is "best" ML model?
- curses, dimensionality, complexity

Donoho: bio

- worked with Tukey as undergrad

Donoho: bio

- worked with Tukey as undergrad
- Berkeley briefly

Donoho: bio

- worked with Tukey as undergrad
- Berkeley briefly
- Stanford

Donoho: bio

- worked with Tukey as undergrad
- Berkeley briefly
- Stanford
  - also: consultant

Donoho: bio

- worked with Tukey as undergrad
- Berkeley briefly
- Stanford
  - also: consultant
  - CA connection

Donoho: 50 years

- Traces history, not only JWT+LB

Donoho: 50 years

- Traces history, not only JWT+LB
- GLS'93

Donoho: 50 years

- Traces history, not only JWT+LB
- GLS'93
- Cleveland'01

Donoho: 50 years

- Traces history, not only JWT+LB
- GLS'93
- Cleveland'01
- interest: baptizing DS as Stats; cakeism

Cleveland

The focus of the plan is the practicing data analys

One outcome of the plan is that computer science joins mathematics as an area of competency for the field of data science. This enlarges the intellectual foundations. It implies partnerships with computer scientists just as there are now partnerships with mathematicians.

The primary agents for change should be university departments themselves. But it is reasonable for departments to look both to university administrators and to funding agencies for resources to assist in bringing about the change.

Chambers GLS 93

Greater statistics: learning from data Three broad categories characterize work in greater statistics:

- preparing data, including planning, collection, organization, and validation

Chambers GLS 93

Greater statistics: learning from data Three broad categories characterize work in greater statistics:

- preparing data, including planning, collection, organization, and validation
- analysing data, by models or other summaries

Chambers GLS 93

Greater statistics: learning from data Three broad categories characterize work in greater statistics:

- preparing data, including planning, collection, organization, and validation
- analysing data, by models or other summaries
- presenting data in written, graphical or other form

Chambers GLS 93 on exhaust

Many mundane commercial and so social activities generate large quantitites ties of potentially valuable data. Examples from business include retail sales, billing, and inventory management. The data were not generated for the purpose of learning; however, the potential for learning is great, if we can cope with some major challenges.

THe data usually pass through a computer system nowadays, but aside from the enourmous quantity the data are typically thrown away farily quickly. The computaional challeng of collecting and organizing such data is huge. A more clearly statistical challenge is that the data may represent only a portion of the conceptually relevant data; if so, the sample is often biased in crucial ways.

Neff: bio

- CC'93!

Neff: bio

- CC'93!
- PhD ethnography

Neff: messages

- critiques benefit from understanding process

A key insight of critical data studies is that interpretation is cooked into the very structures of data and that the work of claiming something as data or a dataset becomes a "rhetorical" move.

Data, as a word, although ends up sounding more authoritative than perhaps those who produce it ever intended

Neff: messages

- critiques benefit from understanding process
  - e.g., not reflecting reflexive data scientists

A key insight of critical data studies is that interpretation is cooked into the very structures of data and that the work of claiming something as data or a dataset becomes a "rhetorical" move.

Data, as a word, although ends up sounding more authoritative than perhaps those who produce it ever intended

Neff: messages

- critiques benefit from understanding process
  - e.g., not reflecting reflexive data scientists
- data scientists benefit from critical data studies

  A key insight of critical data studies is that interpretation is cooked into the very structures of data and that the work of claiming something as data or a dataset becomes a "rhetorical" move.

  Data, as a word, although ends up sounding more authoritative than perhaps those who produce it ever intended

Neff: observations

- subjective design choices

Neff: observations

- subjective design choices
- problem choice $>$ problem solution (cf. JWT)

Neff: observations

- subjective design choices
- problem choice $>$ problem solution (cf. JWT)
- communcation is everything

Neff: observations

- subjective design choices
- problem choice $>$ problem solution (cf. JWT)
- communcation is everything
- prominence of ethics (tho not defined. . .)

contemporary article(s) on the themes

power and principles

how did this capability rearrange power? who can now do what, from what, to whom?

role of rights, harms, justice?

foreshadowing data for Thursday

reminder of themes/big main takeaways

themes

up next