Cultural Analytics

ENGL 64.05

Fall 2019

Prof. James E. Dobson

Nov 4, 2019

Chris Anderson

"The End of Theory: The Data Deluge Makes the Scientific Method Obsolete"
WIRED, June 23 2008

"This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves…There is now a better way. Petabytes allow us to say: 'Correlation is enough.' We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot."

Assessing Machine Learning

- We've started from the bottom up to learn about methods in cultural analytics.
- We've discussed issues related to locating the appropriate data for our studies and the inability to produce historical datasets that would provide broad representation.
- We've examined methods for the extraction of features from text and discussed the impact of our choices (ngrams, stopwords, accent correction, etc).
- We've learned about different methods of selecting from these features, for example minimum and maximum document frequency and limiting our vocabulary.
- We've learned how to use external dictionaries to add "knowledge" to our data (Stanford hard seeds, color lexicons, sentiment lexicons).
- We've experimented with techniques to transform these into higher-level statistics (topic modeling, multi-dimensional scaling, etc) and seen how delicate/brittle these models can be.
- Now we'll begin to learn about methods that can *learn* to recognize features from our data and then classify documents or small units of text.

But First,
Two Examples
and
Another Critique of
Obscure Algorithms

gender *Predict gender from first names using historical data*

Description

This function predicts the gender of a first name given a year or range of years in which the person was born. The prediction can use one of several data sets suitable for different time periods or geographical regions. See the package vignette for suggestions on using this function with multiple names and for a discussion of which data set is most suitable for your research question. When using certain methods, the `genderdata` data package is required; you will be prompted to install it if it is not already available.

Usage

```
gender(names, years = c(1932, 2012), method = c("ssa", "ipums", "napp",
  "kantrowitz", "genderize", "demo"), countries = c("United States", "Canada",
  "United Kingdom", "Denmark", "Iceland", "Norway", "Sweden"))
```

Arguments

`names`	First names as a character vector. Names are case insensitive.
`years`	The birth year of the name whose gender is to be predicted. This argument can be either a single year, a range of years in the form `c(1880, 1900)`. If no value is specified, then for the `"ssa"` method it will use the period 1932 to 2012; acceptable years for the SSA method range from 1880 to 2012, but for years before 1930 the IPUMS method is probably more accurate. For the `"ipums"` method the default range is the period 1789 to 1930, which is also the range of acceptable years. For the `"napp"` method the default range is the period 1758 to 1910, which is also the range of acceptable years. If a year or range of years is specified, then the names will be looked up for that period.
`method`	This value determines the data set that is used to predict the gender of the name. The `"ssa"` method looks up names based from the U.S. Social Security Administration baby name data. (This method is based on an implementation by Cameron Blevins.) The `"ipums"` method looks up names from the U.S. Census data in the Integrated Public Use Microdata Series. (This method was contributed by Ben Schmidt.) The `"napp"` method uses census microdata from Canada, Great Britain, Denmark, Iceland, Norway, and Sweden from 1801 to 1910 created by the North Atlantic Population Project. The `"kantrowitz"` method uses the Kantrowitz corpus of male and female names. The `"genderize"` method uses the Genderize.io <`http://genderize.io/`> API, which is based on "user profiles across major social networks." The `"demo"` method is uses the top 100 names in the SSA method; it is provided only for demonstration purposes when the `genderdata` package is not installed and it is not suitable for research purposes.

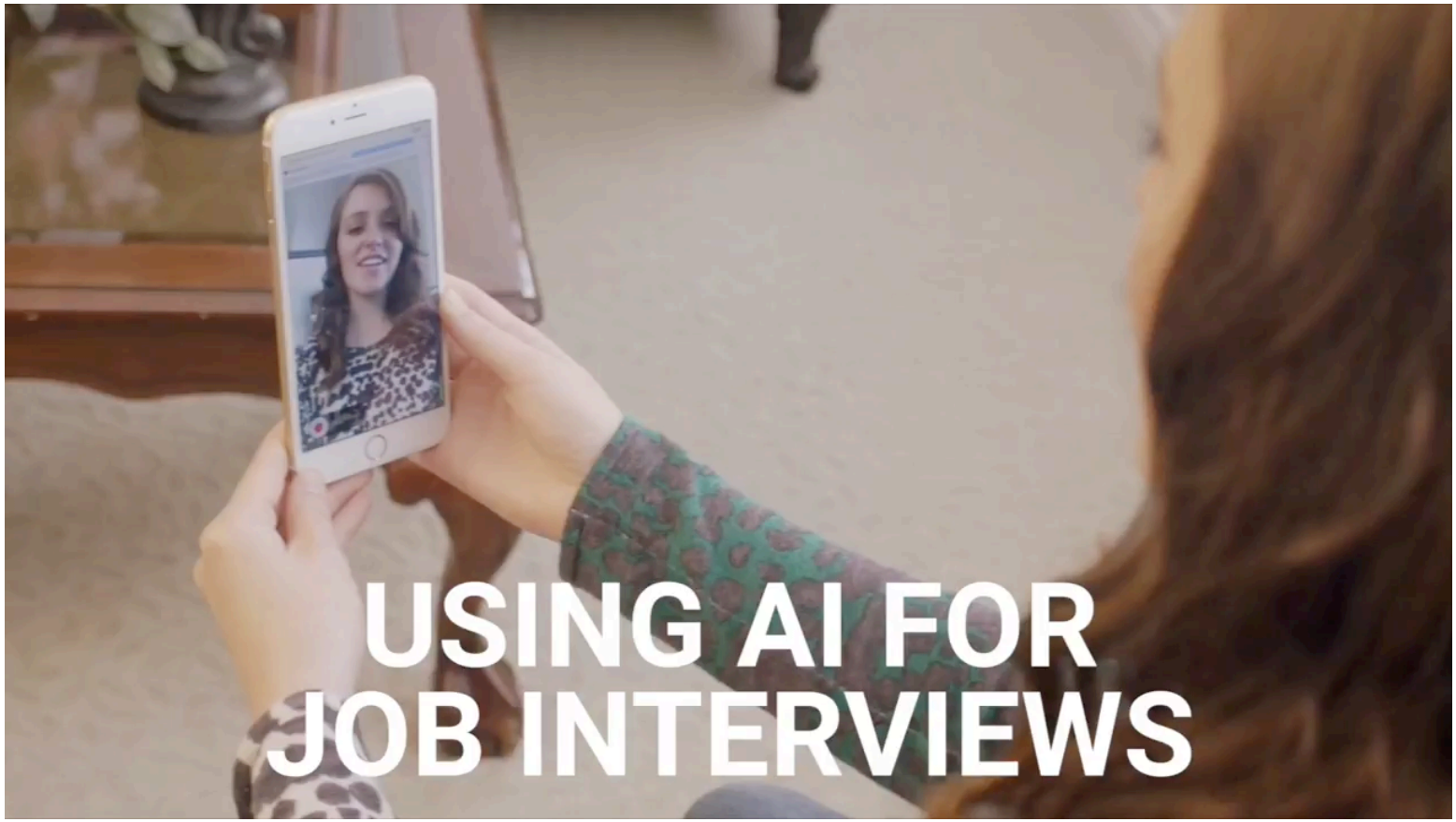
Cultural Analytics

For Enders of Gender

`gendrendr` contains a simple set of functions designed to highlight the inaccuracy and violence of assigning genders to others.

The premise is as follows:

1. the assignment of gender to another in the absence of personal confirmation is an act of violence that perpetuates hierarchical systems of oppression and can be personally traumatizing;
2. the assumption of the correctness of gender assigned at birth reinforces archaic medical views and state-sanctioned violence;
3. gender is a construct that varies over space, time, culture, and ethnicity, and assuming that data from one context apply to another reinforces gendered imperialist violence and perpetuates cultural stereotypes;
4. of specific relevance is the fact that gender is not a binary, and use of data that assume a gender binary reinforces that norm, which does violence to individuals who are non-binary and erases culurues that embrace a diversity of genders;
5. gender cannot be accurately inferred from names, presentations, pronouns or other such factors, and assuming it can and that the consequences of any failure are trivial speaks to the devaluation of transgender, non-binary, and gender-non-conforming life...and this is not absolved by using large data sets and fancy statistics; and
6. if it is important for some reason to know what someone's gender is, the only way to accurately, respectfully, and definitively obtain that information is from that person; and
7. gender and sex are both imperfect constructs and drawing distinctions between them creates unnecessary hierarchies--sex is as problematic (if not moreso) a concept than gender and holding it in any higher regard elevates out-dated medicalized views and perpetuates systematic oppressions.



Weapons of Math Destruction
(WMDs)

Cathy O'Neil's Three Big Questions:

1. Even if the participant is aware of being modeled, or what the model is used for, is the model opaque, or even invisible? (28).
2. Does the model work against the subject's interest? In short, is it unfair? Does it damage or destroy lives? (29).
3. Does the model have the capacity to grow exponentially? As a statistician would put it, can it scale? (29-30).

Sabermetrics vs. LSI-R

- What makes the example of statistics in baseball a reasonable use of big data?
- What is the problem with using data from the LSI-R to predict recidivism?

Transparency Ideal

- Opening the "black box" and examining data are ideals but not the *only* answer.
- Among other problems, seeing doesn't mean understanding.
- Not all models are subject to explanation.
- Code and algorithms are embedded in larger social systems (an "assemblage") and we need to understand these as well.

Mike Ananny and Kate Crawford, "Seeing without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability," *New Media & Society* 20, no. 3 (2018): 973–989.