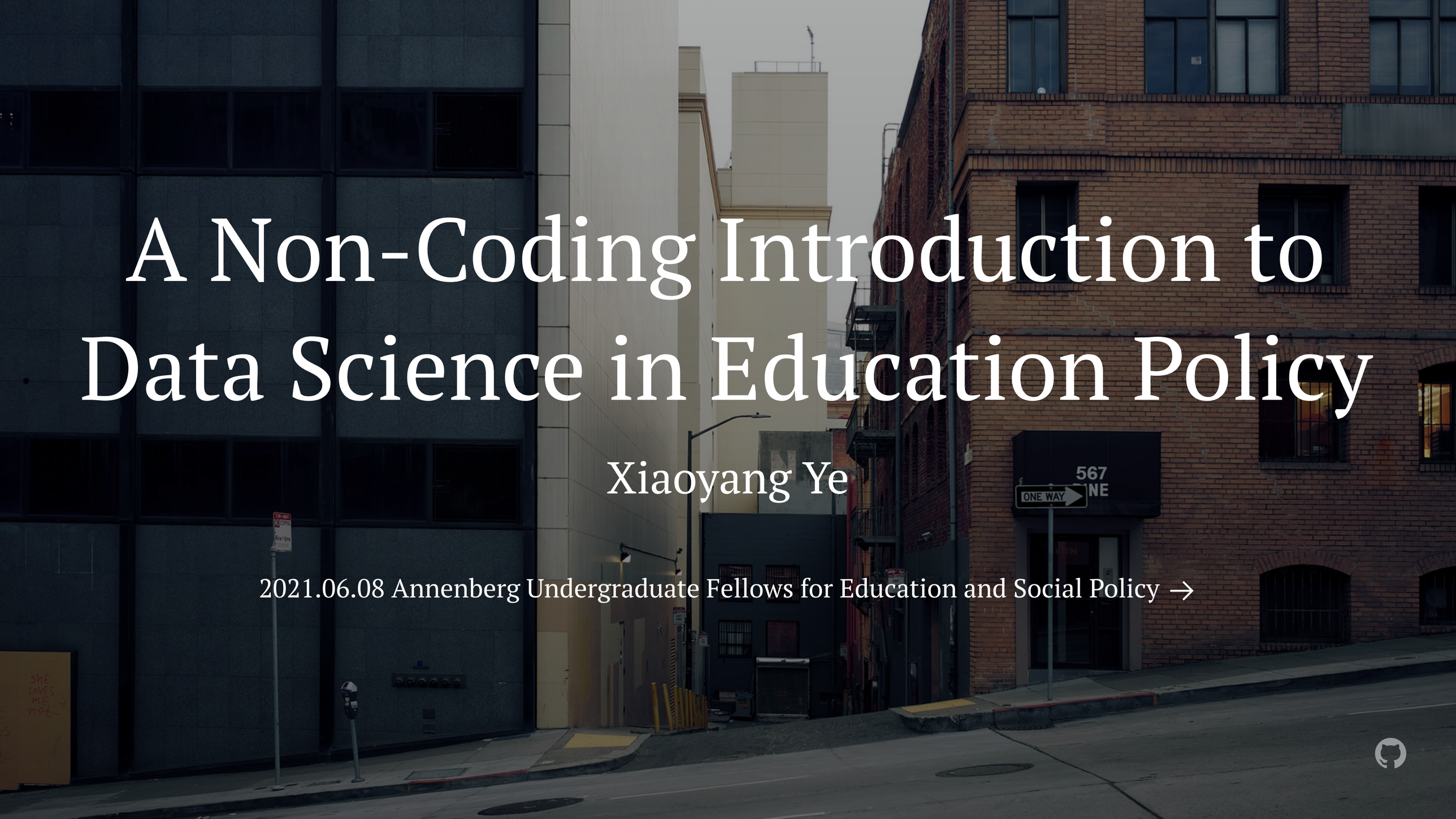


A Non-Coding Introduction to Data Science in Education Policy

Xiaoyang Ye

2021.06.08 Annenberg Undergraduate Fellows for Education and Social Policy →

The Rise of Data Scientists

BEFORE

nobody cared for a "math geek" in parties.

NOW

People love ~~math geeks~~ data scientists!

RK

Data science as sience?

"... of planning for, acquisition, management, analysis of, and inference from data" (NSF 2014)

- A verty short history of data science
  - 1950, Alan Turing, "Computing Machinery and Intelligence"

    a seminal paper on the topic of artificial intelligence

  - 1962, John Tukey "The Future of Data Analysis"

    "... data analysis is intrinsically an empirical science"

  - 1997, C. F. Jeff Wu, the H. C. Carver Chair in Statistics at the University of Michigan

    calls for statistics to be renamed data science and statisticians to be renamed data scientists

Data science for decisions

- 1999, Jacob Zahavi, “Mining Data for Nuggets of Knowledge”

  1. “Conventional statistical methods work well with small data sets. Today’s databases, however, can involve millions of rows and scores of columns of data… Scalability is a huge issue in data mining.
  2. Another technical challenge is developing models that can do a better job analyzing data, detecting non-linear relationships and interaction between element…
  3. Special data mining tools may have to be developed to address web-site decisions.”

- 2006, Thomas Davenport, Competing on Analytics (publised at *Harvard Business Review*)

  “Employees hired for their expertise with numbers or trained to recognize their importance are armed with the best evidence and the best quantitative tools. As a result, they make the best decisions.”

Data scientists do all the data work

- 2009, Hal Varian, Google's Chief Ecnomist

> "I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s? The ability to take data - **to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it** that's going to be a hugely important skill in the next decades…"

- 2009, Nathan Yau, "Rise of Data Scientist"

> "…if you went on to read the rest of Varian's interview, you'd know that by statisticians, he actually meant it as a general title for someone who is able to extract information from large datasets and then present something of use to non-data experts… We're seeing **data scientists** - people who can do it all - emerge from the rest of the pack."

Data science is the new sexy

- 2009, Mike Driscoll, “The Three Sexy Skills of Data Geeks”

> “…with the Age of Data upon us, those who can model, munge, and visually communicate data—call us statisticians or data geeks—are a hot commodity.”

- 2009, Troy Sadkowsky, LinkedIn

> created one of the first data scientists groups

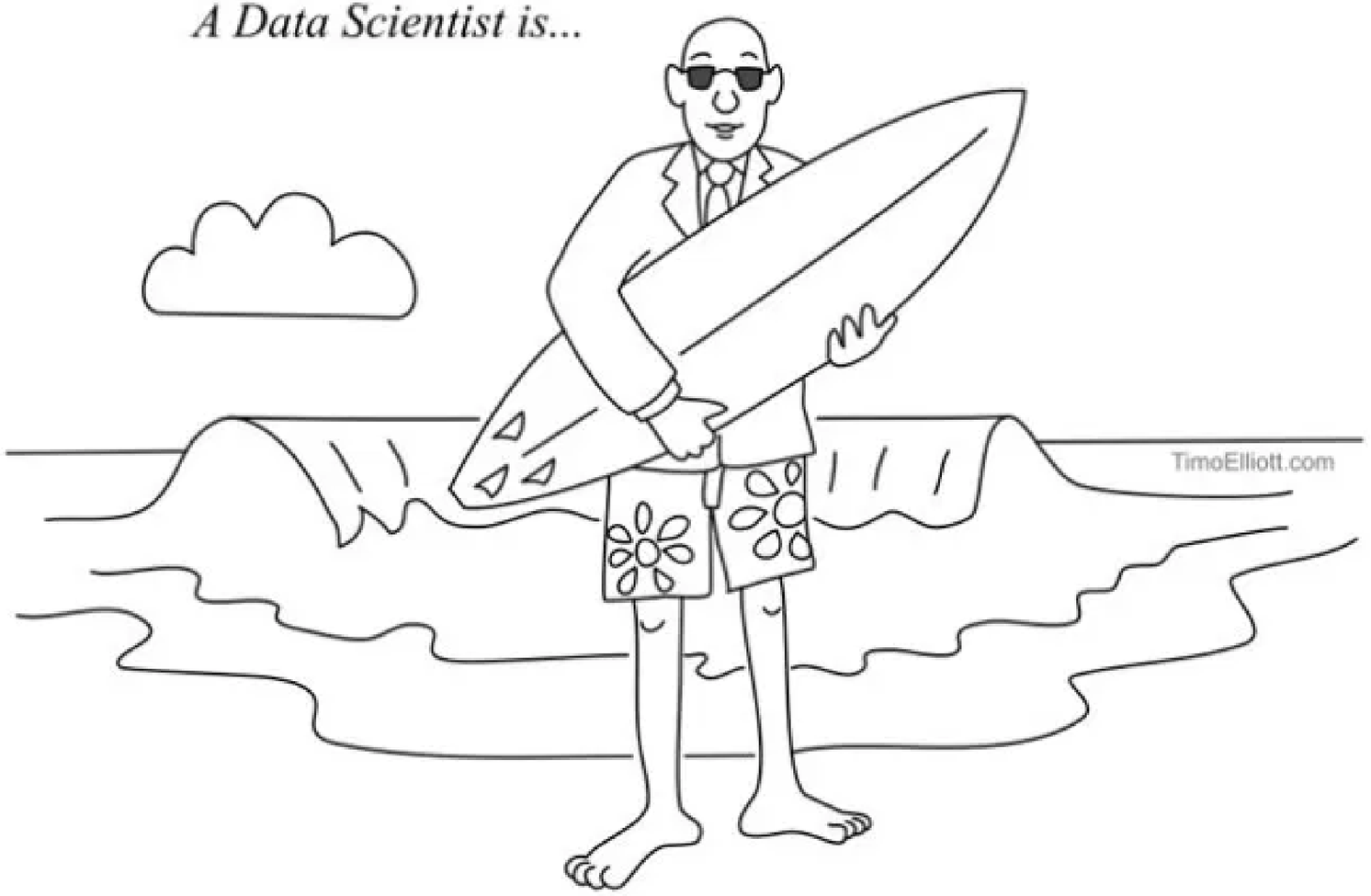
- 2010, Kenneth Cukier, “Data, Data Everywhere” (published at *The Economist*)

> “… a new kind of professional has emerged, the data scientist, who combines the skills of software programmer, statistician and storyteller/artist to extract the nuggets of gold hidden under mountains of data.”

All the buzzwords

- Data science
- Big data
- Artificial Intelligence (AI)
  - Machine learning
    - Deep learning - the next big thing

A Data Scientist is...



A Business Analyst that lives in California.

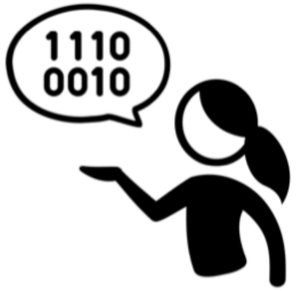
The State of Data, May 2021

by Gil Press, https://whatsthebigdata.com/2021/06/01/the-state-of-data-may-2021/

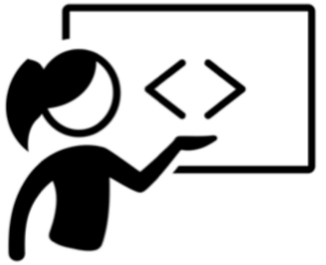
- 30 million
  - the number of predictions per second a single AI model can make in order to match Pinterest users' interests with relevant ads
- 33.5%
  - of radiologists using AI in their clinical practices
- 74%
  - reduction in application review time achieved by the GRaduate ADmissions Evaluator, an AI evaluation system built and used by the graduate program in computer science at the University of Texas at Austin
- $150,000
  - the average base salary for machine-learning engineers

What is a Data Scientist?

Scientist
Research & explore
Deal with ambiguity
Manage risk



Programmer
Wrangle Data
Design for Scale
Automate



Communicator
Influence sans-authority
Convey value simply
Empathize

What data scientists actually do

- 3%: Building training sets
- 4%: Refining algorithms
- 5%: Others
- 9%: Mining data for patterns
- 19%: Collecting data sets
- 60%: Cleaning and organizing data

Today: Data science for education policy research

- Three tasks in data science/machine learning

1. Description
    - Example: Measure teacher practices at scale (Liu & Cohen, 2021)
2. Prediction
    - Example: Guide college choice using prediceted admissions probability (Ye, 2021)
3. Causal inference
    - Example: Identify treatment effect of a growth mindset intervention (Athey & Wagner, 2019)

1. Description

Example: Measure teacher practices at scale

Liu, J., & Cohen, J. (2021). Measuring Teaching Practices at Scale: A Novel Application of Text-as-Data Methods. *Educational Evaluation and Policy Analysis*.

- Context:
  - Valid and reliable measurements of teaching quality facilitate school-level decision-making and policies pertaining to teachers
- Empirical problem:
  - Traditional measure: classroom observations by human raters -> cost/unscalability, rater bias
- Method:
  - text-as-data (natural language processing) + unsupervised learning (factor analysis)

Text-as-data

Appendix A

Student A: [00:00:02] Tell you the definition of cause? [00:00:03]

Teacher: [00:00:03] Yeah, what's the definition of cause, what does it mean? [00:00:05]

Student A: [00:00:06] It's like something that happens and that the something that happens and the cause is what happens. [00:00:13]

Teacher: [00:00:16] Okay, you said the effect, the cause is why something happens. Is that what you're saying? And the effect is why something happens. Or the effect is what happens, the cause is why it happens. Okay, keep that in mind as we watch this short video. You need to listen, okay, because it's going to talk to you. [00:00:34]

[Silence]

Teacher: [00:00:41] Okay, so keep that in mind what she's saying. Cause is why something happens, the effect is what actually happens. [00:00:43]

Language style matching

an index that measures whether two people in a natural conversation match each other's speaking behavior or style using functional words

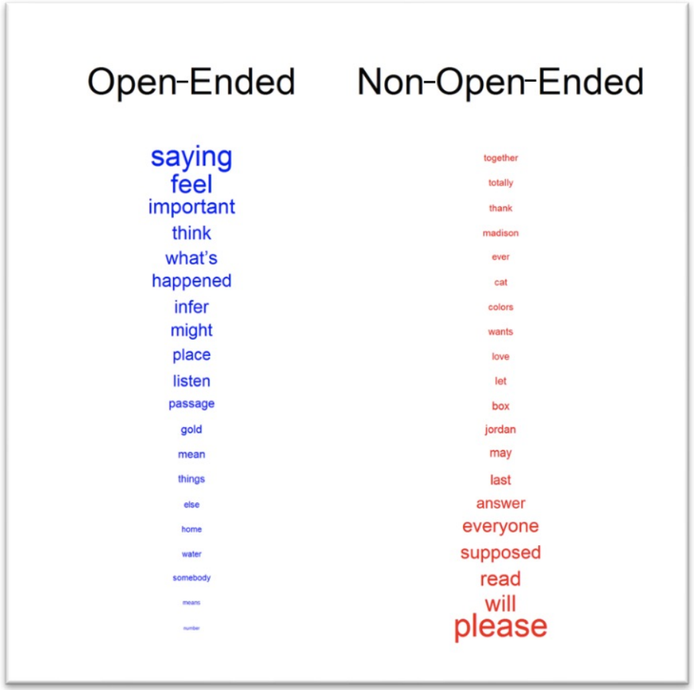
Word Categories Used for Calculating Language Style Matching

Category	Examples
Personal pronouns	I, his, their
Impersonal pronouns	it, that, anything
Articles	a, an, the
Conjunctions	and, but, because
Prepositions	in, under, about
Auxiliary verbs	shall, be, was
High-frequency adverbs	very, rather, just
Negations	no, not, never

Predicting questioning styles (supervised learning)

plays a key role in eliciting rich discussion and engaging students, and both the quantity and quality of questions play an integral role.

Figure C1: Features of Open-Ended and Non-Open-Ended Questions



Note. Results are from Lasso regression. Word size is proportionate to coefficient size.

Results

Liu, J., & Cohen, J. (2021). Measuring Teaching Practices at Scale: A Novel Application of Text-as-Data Methods. *Educational Evaluation and Policy Analysis*.

- Data:
  - 976 word-to-word video transcriptions (29,436 minutes) of 4th- and 5thgrade English language arts classes
- Finding:
  - the detection of three instructional factors: classroom management, interactive instruction, and teacher-centered instruction
- Implication:
  - The text-as-data approach has the potential to enhance existing classroom observation systems through collecting far more data on teaching **with a lower cost, higher speed, and the detection of multifaceted classroom practices**

What We Teach About Race and Gender: Representation in Images and Text of Children's Books

https://bfi.uchicago.edu/wp-content/uploads/2021/04/BFI_WP_2021-44.pdf

WORKING PAPER · NO. 2021-44

What We Teach About Race and Gender: Representation in Images and Text of Children's Books

Anjali Adukia, Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha, and Teodora Szasz

APRIL 2021

2. Prediction

Widely used in many predictive analytics

- **Examples**
  - Predicting school dropout
  - Predicting college admissions
  - Predciting online learning behaviors
- **Challenge**
  - Individualization vs. out-of-sample accuracy
  - All the models work on minimizing the prediction errors
    - this is why deep learning outperforms
- vs. applied econometrics

How does machine learn?

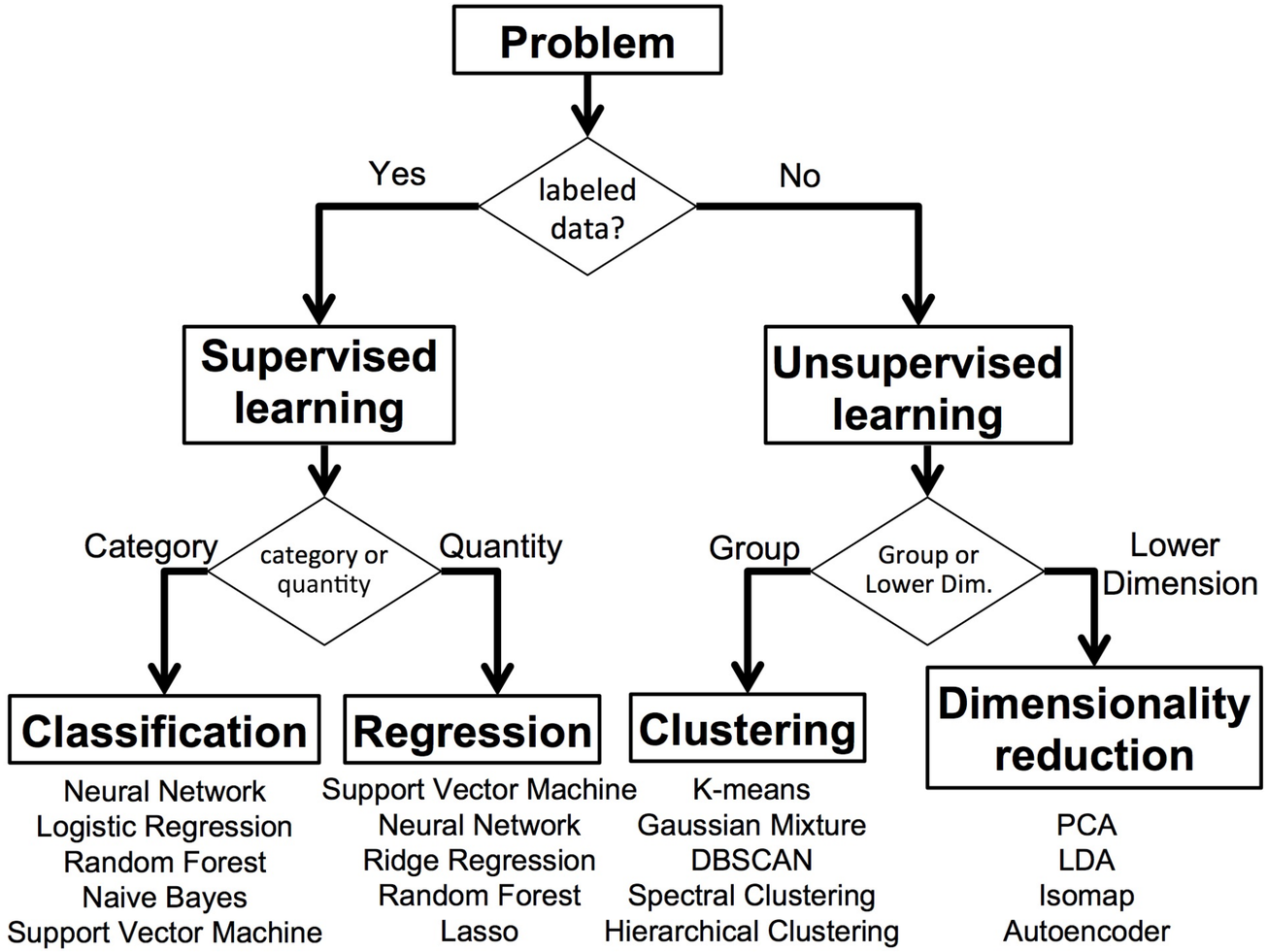
$$Y = f(X) + \epsilon$$

- **Why estimate *f*?**
  - prediction
    - given a set of inputs X, what is Y?
  - inference
    - what is the relationship between each x and Y?
- **Example: *f* = linear regression**

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon$$

Types of machine learning

note: one more group is reinforcement learning (e.g., self-driving cars)



Example: Predicting college admissions

Ye, X. (2021). Personalized advising for college match: Experimental evidence on the use of human expertise and machine learning to improve college choice.

- Context:
  - Students make college choice decisions (partly) based on their predictions of admissions probabilities
- Empirical problem:
  - Many students make costly mistakes in predictions, resulting in undiserable college-going outcomes
- Proposed solutions: teach students to predict vs. help them predict
- Findings:
  - 1. Machine learning largely increased advising efficiency through the combination of increased access to human expertise, data, and optimal decision algorithms

3. Causal inference

Key questions

1. Who is the control group
    - self-selection bias
2. Who benefits most from a treatment
    - heterogeneous treatment effects
3. What is the correct choice of variables and models
4. How to assign the most effective intervention?
    - multi-arm bandit treatment status assignment

Example

Observational Studies 5 (2019) 21-35 Submitted 7/19; Published 7/19

Assessing Treatment Effect Variation in Observational Studies: Results from a Data Challenge

Carlos Carvalho **carlos.carvalho@mccombs.utexas.edu**
Department of Information, Risk and Operations Management
The University of Texas at Austin
Austin, TX 78712, USA

Avi Feller **afeller@berkeley.edu**
Goldman School of Public Policy
The University of California, Berkeley
Berkeley, CA 94720, USA

Jared Murray **jared.murray@mccombs.utexas.edu**
Department of Information, Risk and Operations Management
The University of Texas at Austin
Austin, TX 78712, USA

Spencer Woody **spencer.woody@utexas.edu**
Department of Statistics and Data Science
The University of Texas at Austin
Austin, TX 78712, USA

David Yeager **dyeager@utexas.edu**
Department of Psychology
The University of Texas at Austin
Austin, TX 78712, USA

Example: Treatment effect variation in observational studies

- Context:
  - A randomized experiment of a low-cost growth mindset intervention
  - growth mindset (vs. fixed mindset): the belief that people can develop intelligence
- Empirical problem:
  - Model choice in observational studies
- **Data chanllenge:**
  - 8 teams worked independently

True effect = 0.24

Author	ATE estimate (95 % C.I.)
Athey & Wager	0.25 (0.21, 0.29)
Carnegie et al.	0.25 (0.23, 0.27)
Johannsson	0.27 (0.22, 0.31)
Keele & Pimentel	0.27 (0.25, 0.30)
Keller et al.	0.26 (0.22, 0.30)
Künzel et al.	0.25 (0.22, 0.27)
Parikh et al.	0.26 (0.25, 0.26)
Zhao & Panigrahi	0.26 (0.24, 0.28)

Table 2: Submitted estimates for the average treatment effect and corresponding 95% uncertainty intervals.

Athey & Wager

Observational Studies 5 (2019) 36-51 Submitted 7/19; Published 8/19

Estimating Treatment Effects with Causal Forests: An Application

Susan Athey **athey@stanford.edu**
Stanford Graduate School of Business, Stanford, CA-94305

Stefan Wager **swager@stanford.edu**
Stanford Graduate School of Business, Stanford, CA-94305

Abstract

We apply causal forests to a dataset derived from the National Study of Learning Mindsets, and discusses resulting practical and conceptual challenges. This note will appear in an upcoming issue of *Observational Studies*, Empirical Investigation of Methods for Heterogeneity, that compiles several analyses of the same dataset.

Causal inference in observational data

- Potential outcome: $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$
- Conditional on $X_i = x$:

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)] \quad | \quad X_i = x$$

- **Unconfoundedness**: treatment assignment (Z_i) is as good as random conditional on covariates

$$Y_i(0), Y_i(1) \perp\!\!\!\perp Z_i | X_i$$

- **Empirical task**:
  - Find the most comparable/similar groups based on X_i

The only codes we see today!

X, Y, Z are input variables

```
#define outcome model
  Y.forest = regression_forest (X, Y)
  Y.hat = predict(Y.forest)$predictions

#define selection model
  Z.forest = regression_forest (X, Z)
  Z.hat = predict(Z.forest)$predictions

#train the Causal Forest algorithm
  cf = causal_forest(X, Y, Z,
                     Y.hat = Y.hat, Z.hat = Z.hat)

#get the predicted treatment effect (tau)
  tau.hat = predict(cf)$predictions
```

Summary

Data science has been seen as “a black box”

but there is no hidden magic…

- The key is to answer important research questions with data
  - Description (also, get the new/big data)
  - Predcition
  - Causal inference

- Data science as new tools
  - to answer new types of questions
  - to answer questions more efficiently

Data science is evolving very fast!

- It uses the same “evidence-based” logic as applied econometrics that we (education policy researchers) are more familiar with
- **More work is needed** to understand how to best apply data science for education policy research
- We also need to address issues of bias, fairness, and ethnics as we do in all types of empirical research

Thank you very much!