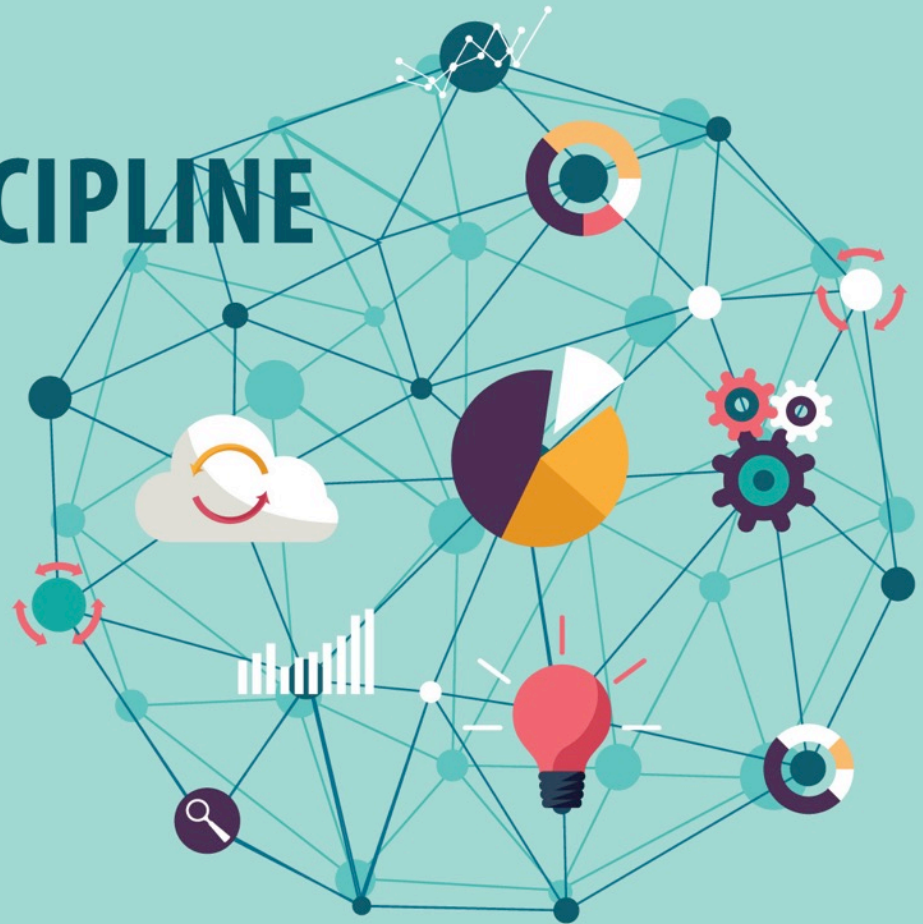


Envisioning the  
**DATA SCIENCE DISCIPLINE**  
The Undergraduate Perspective

Webinar Series  
Fall 2017



*The National  
Academies of*

SCIENCES  
ENGINEERING  
MEDICINE

[nas.edu/EnvisioningDS](https://nas.edu/EnvisioningDS)

# Envisioning the **DATA SCIENCE DISCIPLINE**

## The Undergraduate Perspective

**9/12/17** – Building Data Acumen  
*(recording posted)*

**9/19/17** – Incorporating Real-World  
Applications *(recording posted)*

**9/26/17** – Faculty Training and  
Curriculum Development  
*(recording posted)*

**10/3/17** – Communication Skills and  
Teamwork *(recording posted)*

**10/10/17** – Inter-Departmental  
Collaboration and Institutional  
Organization *(recording posted)*

**10/17/17** – Ethics

**10/24/17** – Assessment and Evaluation  
for Data Science Programs

**11/7/17** – Diversity, Inclusion, and  
Increasing Participation

**11/14/17** – Two-Year Colleges and  
Institutional Partnerships

**Provide input, download the interim  
report, and learn more about the  
study at [www.nas.edu/EnvisioningDS](http://www.nas.edu/EnvisioningDS)**

# Envisioning the **DATA SCIENCE DISCIPLINE**

## The Undergraduate Perspective

### Ethics



**Sorin Adam Matei, Purdue University**  
*Professor and Director of Purdue Data  
Storytelling Network*



**Brittany Fiore-Gartland,**  
**University of Washington**  
*Director of Data Science Ethnography  
eScience Institute*

Provide input and learn more about the study at [www.nas.edu/EnvisioningDS](http://www.nas.edu/EnvisioningDS)

# Envisioning the **DATA SCIENCE DISCIPLINE**

The Undergraduate Perspective

Ethics



**Sorin Adam Matei, Purdue University**  
*Professor and Director of Purdue Data  
Storytelling Network and Kredible.net Project*

An ethical  
reasoning  
framework for  
data science  
education

# Goals of presentation

- Define ethics and big data (the realm of data science)
- Define principles for ethical reasoning
- Discuss the principles in a data science context
- Present the ethical reasoning process in context, using an example
- Discuss the implications / answer questions

# A principled approach to ethics

- Ethics is a methodology. Depends on morality, which is rooted in human values.
  - Deals with “right conduct.”
    - Necessitates values: What values should direct ethical conduct?
    - Used in context: What types of trade-offs are possible and when?

# Ethical conduct in the context of data science

- What are the main ethical challenges involved in data science?
- What moral principles are affected by these challenges?
- How do contexts of data science use or potential abuse aggravate or mitigate the urgency of ethical decision making?



# Data Science and its challenges

- Challenges derived from the automatic collection of big data
  - Reuse
    - Genomic data collected for one purpose (disease prevention) used for drug discovery
  - Recombination
    - Joining multiple datasets to obtain fuller profiles of users (Google, Facebook)
  - Repurposing
    - Data collected for billing purposes is used for anti-terrorism surveillance (the NSA meta-data scandal)
  - Reanalysis
    - Raw data can be anonymous, analyzed data can become personally identifiable (behavior forensics: login patterns, Wikipedia editing patterns, etc.)



# Challenge 1: Privacy

- The 4 Rs (recombining, repurposing, reuse, reanalysis) implicit in data science may affect individually profoundly and directly
- Identity, anonymity may be compromised -> **Privacy** main concern
- But is this all?

# Privacy is not sufficient

Concern for privacy is a serious issue but does not suggest a path to action

Strategic question: what does loss of privacy do to us?

# Loss of privacy and core human values

- Privacy loss is in fact part of a family of much general concerns.
  - Limits or denies human autonomy
  - May lead to treating people unfairly
  - Breaks the bonds of trust in society
  - Makes some people worse off or exposes them to risk

# Core values in data science ethical reasoning

- Autonomy
- Justice
- Trust
- Non-maleficence / beneficence
- How do we make practical use of them?

# Ethical reasoning process

Does this procedure, be it of any form (collection, analysis, repurposing, recombination, etc.) lead to:

- Loss of autonomy?
- Loss of trust between individuals or between individuals and the researcher?
- Create unjust allocation of rewards (recognition, access to resources, social advancement) or deleterious effects (stigma, loss of face, material loss, etc)?
- Harm and / creates a net benefit for the users?

# So what?

- Asking questions is easy....
- Answering them is hard.
- How do we go about answering them?
- In a contextual manner....

# Contextual use of values

- Human values are (should be) absolutes
- Value judgements in themselves are non-negotiable (justice, trust, non-maleficence are preferable under all circumstances.)
- When considering the effect of values on our actions we should establish a minimum threshold under which we should and we cannot go
- Yet, for certain situations and above a certain level of moral certainty, ethical actions should be pragmatically directed by trade-off analysis



# Project example

- Large scale harvesting of video-streams from publically available web-cams in national parks by non-governmental actors (university researchers)
- Declared goals:
  - Reduction of human-wildlife interaction – less human and animal limb and life loss incidents
  - Open up the data to the research community to predict incidents and target potential victims via social media to reduce human-animal encounters
  - Data contains tens of thousands of hours of footage, including faces, acts, and personally identifiable information about park visitors

# Project outcomes

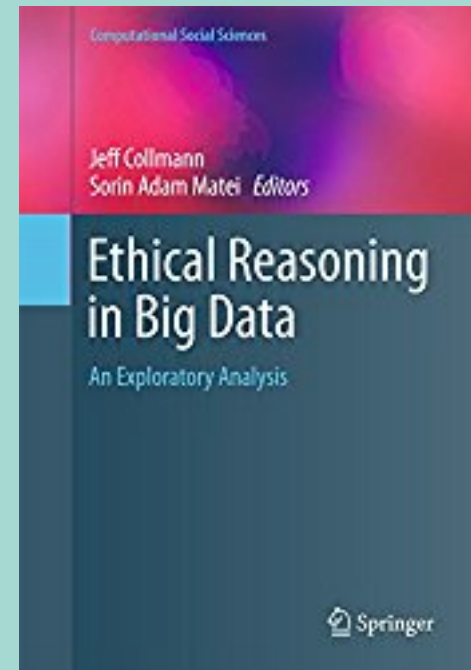
- Findings
  - Reanalysis – reveals who is more likely to interact with animals (gender, race, age)
  - Recombining with social media data – identifies recent visitors at a location of given gender, race, or age
  - Repurposing – identified visitors are targeted with customized messages via Twitter to avoid future encounters

# Ethical reasoning process

- Yes, if:
  - The goal is non-maleficent or even beneficent
    - In this case to limit loss of human and animal life or limb
  - Autonomy is not limited
    - In this case, individuals are not denied access to the park, they are simply informed
  - Justice
    - A case can be made that some individuals may be targeted by race, gender, age affiliation regardless if they did not interact with animals in an unsafe way
      - Trade-off analysis – is the societal benefit greater than the mild annoyance of the individuals of receiving informative tweets? were some fundamental rights denied to these individuals or some privileges bestowed to some and not to them?
  - Trust
    - Does this campaign inspire or erode trust?
      - Trade off analysis – while some might find the campaign intrusive, if rightly prepared by posting notes next to cameras that inform about their use, it may lead to an increase of trust in those that are trying to limit human or animal life and limb loss

# What did we just do?

- We engaged in ethical reasoning
- Value based, with a minimum threshold of value concern and context-based trade-off analysis
- The process is described in Ethical Reasoning in Big Data, co-edited with Jeff Collmann
- The ethical process described was shaped in a major way by Michael Steinmann from Stevens Institute of Technology



<http://amzn.to/2kOwTDz>

# Questions?

- More information
  - <http://kredible.net>
  - <http://matei.org/ithink>

# Envisioning the **DATA SCIENCE DISCIPLINE**

The Undergraduate Perspective

Ethics

## Ethical Thinking for Data Science Education



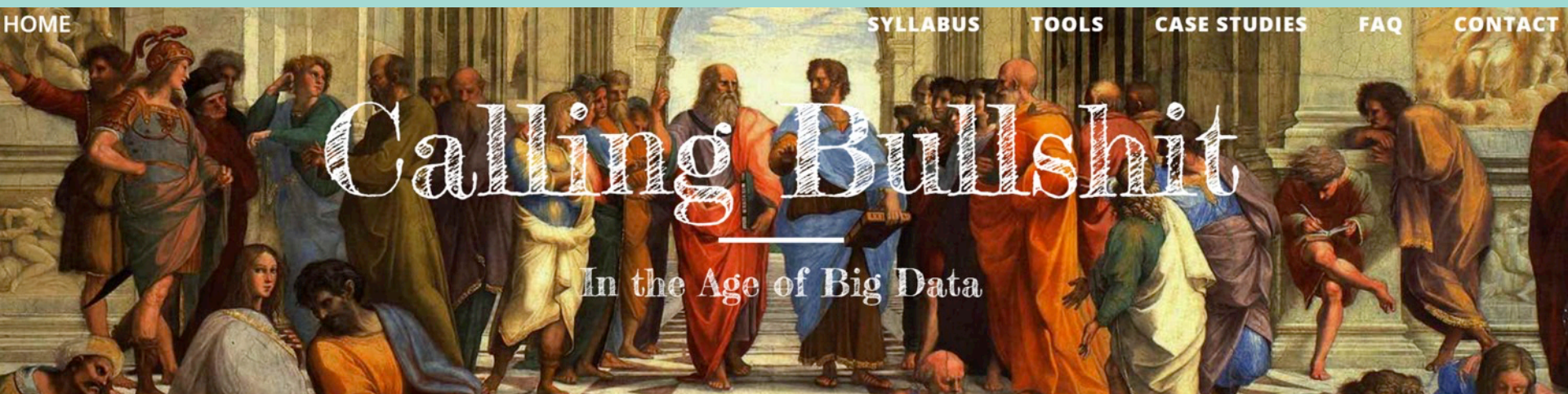
**Brittany Fiore-Gartland,**  
**University of Washington**  
*Director of Data Science Ethnography*  
*eScience Institute*

# Data justice and responsibility



- Democratization of data science tools and techniques
- Computational researchers working in unfamiliar fields or contexts
- Secondary uses of data; Data collected for one purpose and used for another
- Shift culture of ethical compliance to culture of ethical thinking





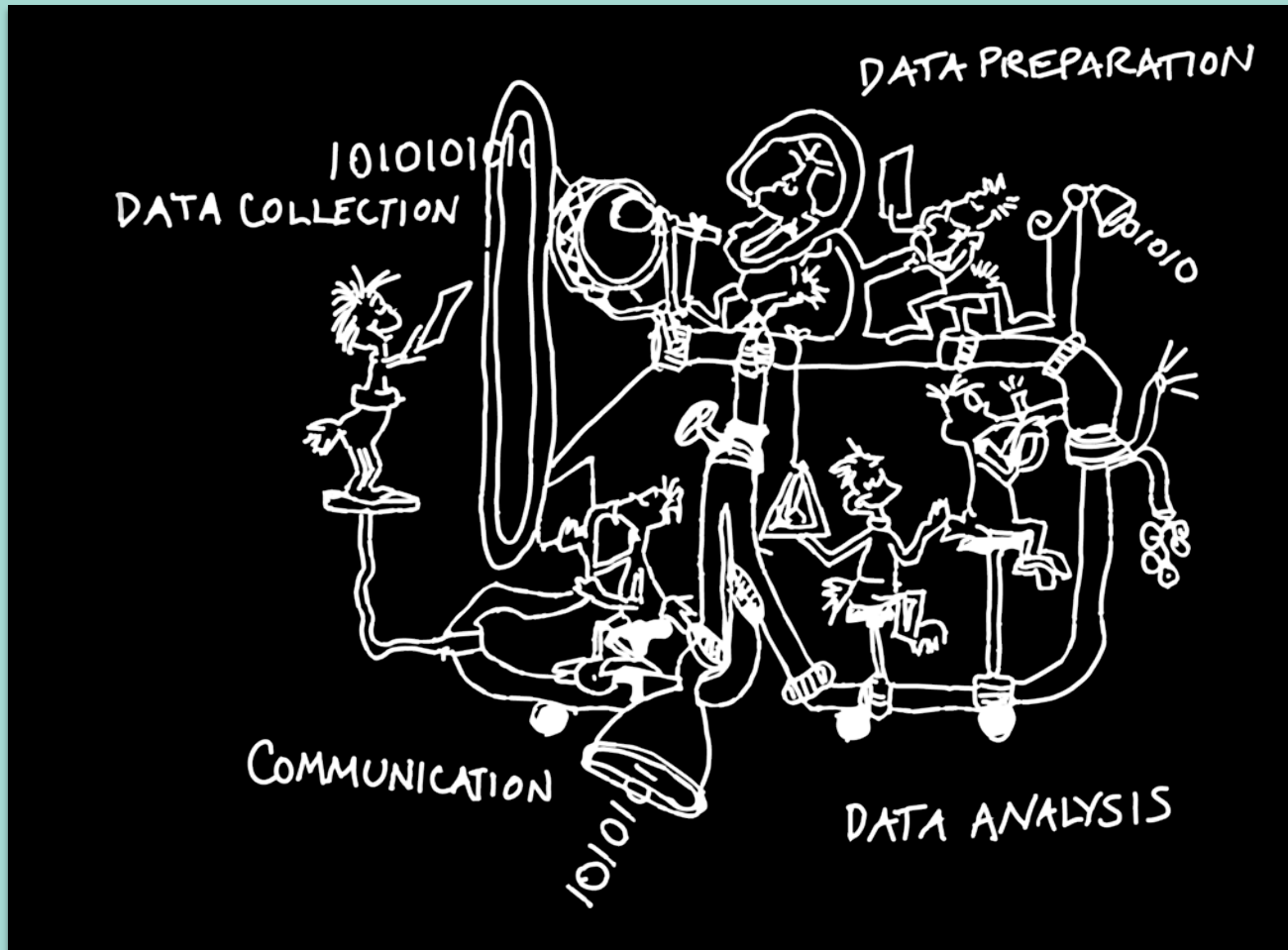
**The world is awash in bullshit.** Politicians are unconstrained by facts. Science is conducted by press release. Higher education rewards bullshit over analytic thought. Startup culture elevates bullshit to high art. Advertisers wink conspiratorially and invite us to join them in seeing through all the bullshit —

Course developed by Jevin West and Carl Bergstrom, University of Washington

# Data Pipeline



# Human-Data Pipeline



Inspired by Dr. Seuss

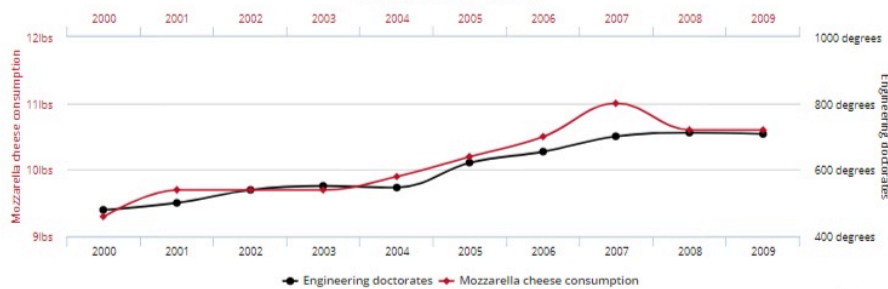
Provide input and learn more about the study at [www.nas.edu/EnvisioningDS](http://www.nas.edu/EnvisioningDS)



# Spurious correlations (Tyler Vigen)

Per capita consumption of mozzarella cheese  
correlates with  
Civil engineering doctorates awarded

Correlation: 95.86% ( $r=0.958648$ )

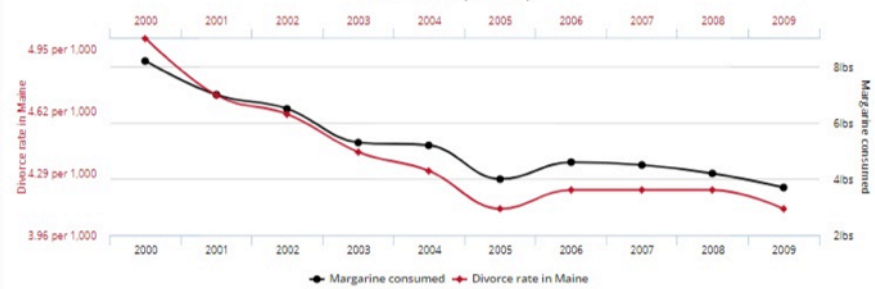


Data sources: U.S. Department of Agriculture and National Science Foundation

tylervigen.com

Divorce rate in Maine  
correlates with  
Per capita consumption of margarine

Correlation: 99.26% ( $r=0.992558$ )

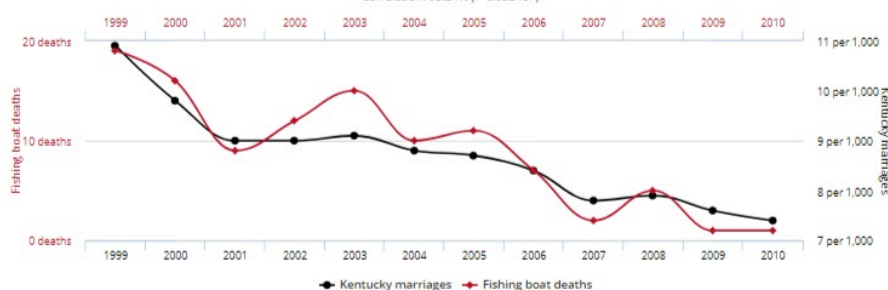


Data sources: National Vital Statistics Reports and U.S. Department of Agriculture

tylervigen.com

People who drowned after falling out of a fishing boat  
correlates with  
Marriage rate in Kentucky

Correlation: 95.24% ( $r=0.952407$ )

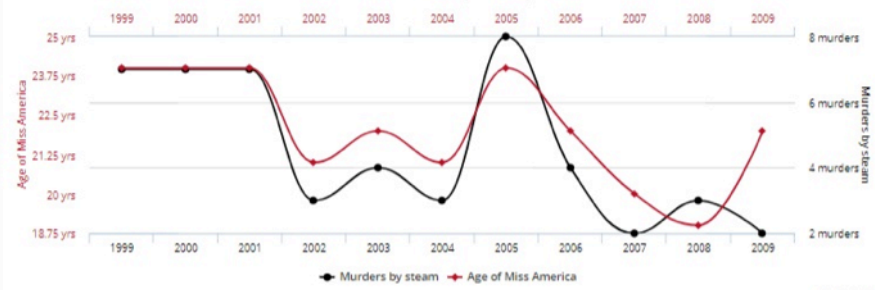


Data sources: Centers for Disease Control & Prevention and National Vital Statistics Reports

tylervigen.com

Age of Miss America  
correlates with  
Murders by steam, hot vapours and hot objects

Correlation: 87.01% ( $r=0.870127$ )



Data sources: Wikipedia and Centers for Disease Control & Prevention

tylervigen.com

# Big Data Truthiness

Implicit assumption and expectation that big data has the quality of seeming or being felt to be true, even if not necessarily true.

(Adapted from Stephen Colbert's "Truthiness")

# Data Science challenges current ethical frameworks

The perils!

Case Study:  
Uber “Rides of Glory”  
analysis

# Uber's "Rides of Glory" Analysis



- Rides from 10 pm-4 am on Fri/Sat, followed by ride from same location 4-6 hrs later
- Implies people “found love that you might immediately regret upon waking up the morning after”
- Calls these “Rides of Glory”



# Uber's "Rides of Glory" Analysis



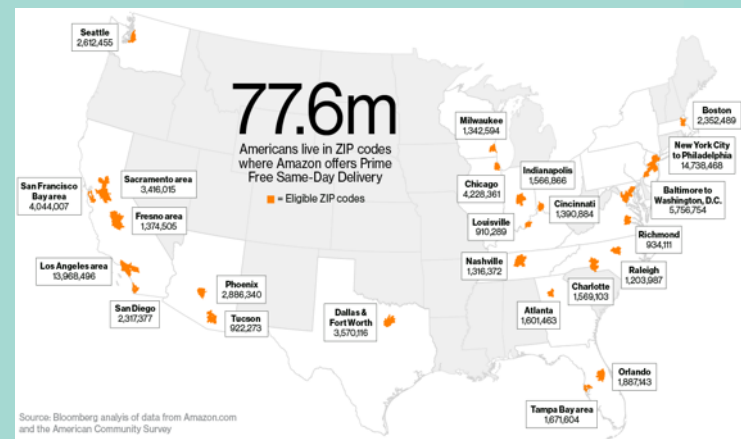
- Consent
- Privacy
- Interpretation



# Data Science as Ethical Intervention

The promise!

Case Study: Amazon same-day delivery service: Can algorithms be racist?



Provide input and learn more about the study at [www.nas.edu/EnvisioningDS](http://www.nas.edu/EnvisioningDS)

### Atlanta



### Boston



### Chicago



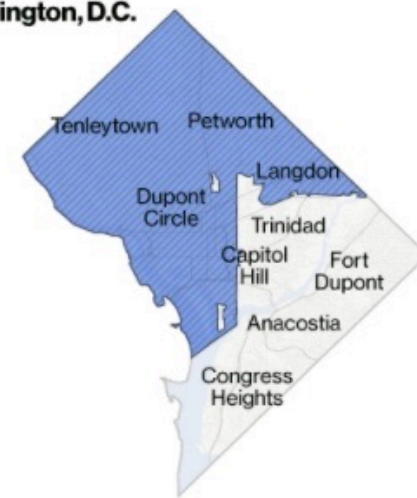
### Dallas



### New York City



### Washington, D.C.

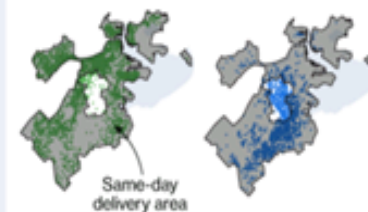


## Boston

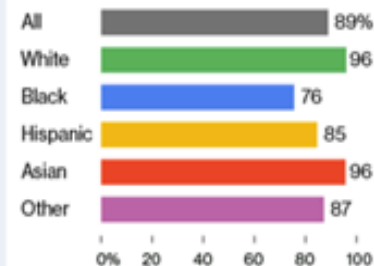


Three ZIP codes in the center of Boston, including the Roxbury neighborhood, are excluded from same-day coverage.

**White residents** **Black residents**



**Percentage of residents living in ZIP codes with same-day delivery**



*Update 1 — Following the publication of this story, Boston Mayor Martin J. Walsh and Massachusetts Senator Ed Markey called upon Amazon to provide Prime Free Same-Day Delivery to Boston's excluded Roxbury neighborhood. On April 26, Amazon agreed to expand the same-day delivery area to include all Boston neighborhoods. On May 4, all Boston ZIP codes were included. [Read More](#)*

*Update 2 — Following criticism from Bronx Borough President Ruben Diaz Jr. and New York State Assemblyman Jeff Dinowitz, Amazon announced on May 1 it will bring same-day delivery to every ZIP code in New York City in the coming weeks. Expanded coverage will include the Bronx, the only borough entirely excluded from the service. [Read More](#)*

*Update 3 — After demands from U.S. Rep. Bobby Rush, whose district covers excluded neighborhoods in Chicago's South Side, Amazon announced on May 2 it will extend the same-day service area to cover every ZIP code in Chicago in the coming weeks. [Read More](#)*

# Thank you!

**Contact: [fioreb@uw.edu](mailto:fioreb@uw.edu)**

## **Resources:**

- Calling Bullshit [syllabus and readings](#), [tools](#), [edited videos](#) of lectures and [case studies](#) are all available at [www.callingbullshit.org](http://www.callingbullshit.org)
- Spurious correlations @ [www.tylervigen.com](http://www.tylervigen.com)
- Link to data science ethics case studies and further reading, <https://tinyurl.com/y9meel3s>



# Envisioning the **DATA SCIENCE DISCIPLINE**

The Undergraduate Perspective

Ethics – Q&A



**Sorin Adam Matei, Purdue University**  
*Professor and Director of Purdue Data  
Storytelling Network*



**Brittany Fiore-Gartland,**  
**University of Washington**  
*Director of Data Science Ethnography  
eScience Institute*

Provide input and learn more about the study at [www.nas.edu/EnvisioningDS](http://www.nas.edu/EnvisioningDS)



# Envisioning the **DATA SCIENCE DISCIPLINE**

## The Undergraduate Perspective

**9/12/17** – Building Data Acumen  
*(recording posted)*

**9/19/17** – Incorporating Real-World  
Applications *(recording posted)*

**9/26/17** – Faculty Training and  
Curriculum Development  
*(recording posted)*

**10/3/17** – Communication Skills and  
Teamwork *(recording posted)*

**10/10/17** – Inter-Departmental  
Collaboration and Institutional  
Organization *(recording posted)*

**10/17/17** – Ethics

**10/24/17** – Assessment and Evaluation  
for Data Science Programs

**11/7/17** – Diversity, Inclusion, and  
Increasing Participation

**11/14/17** – Two-Year Colleges and  
Institutional Partnerships

**Provide input, download the interim  
report, and learn more about the  
study at [www.nas.edu/EnvisioningDS](http://www.nas.edu/EnvisioningDS)**