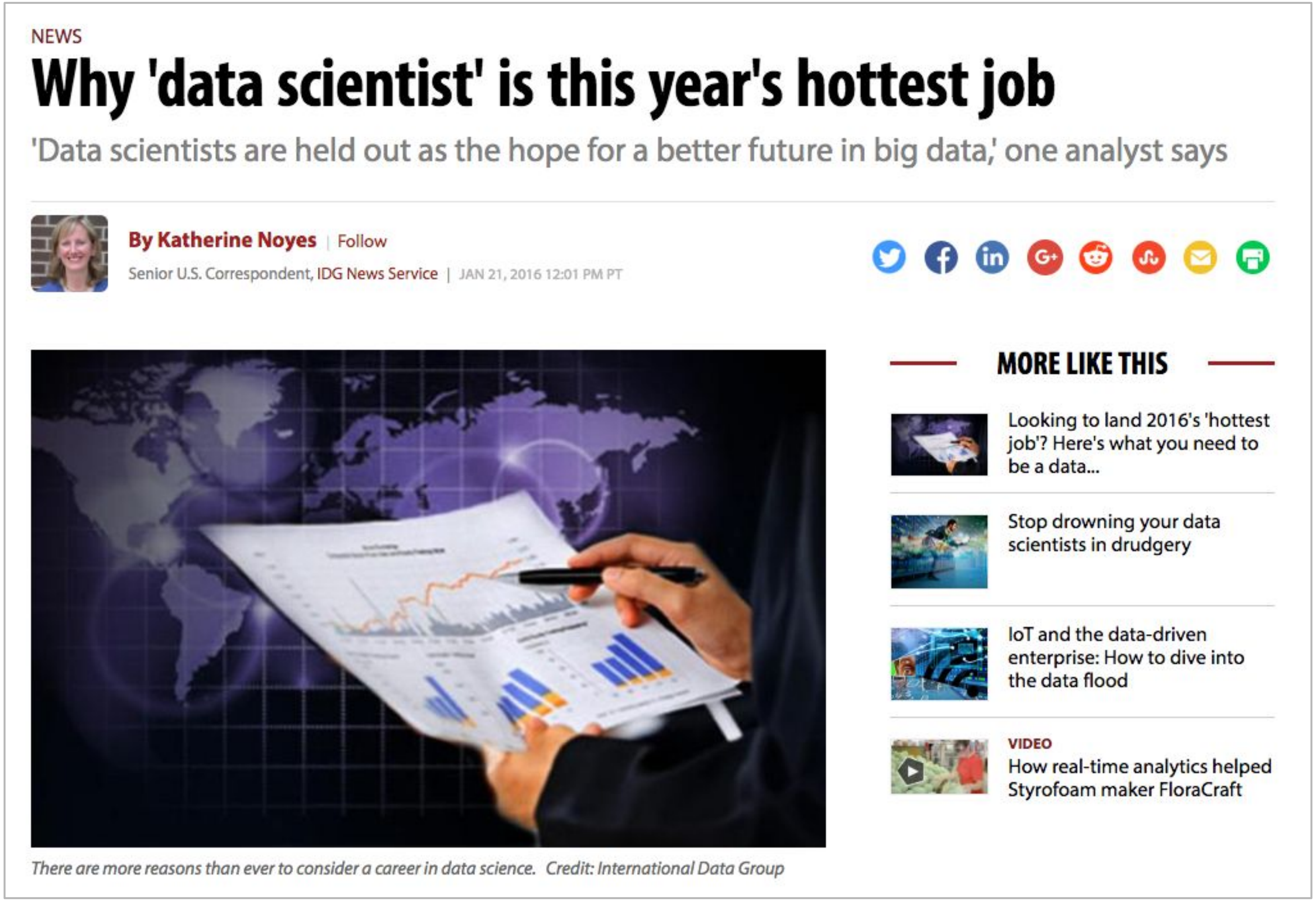
Introduction to Data Science

INFO 370



NEWS

Why 'data scientist' is this year's hottest job

'Data scientists are held out as the hope for a better future in big data,' one analyst says

By Katherine Noyes | Follow

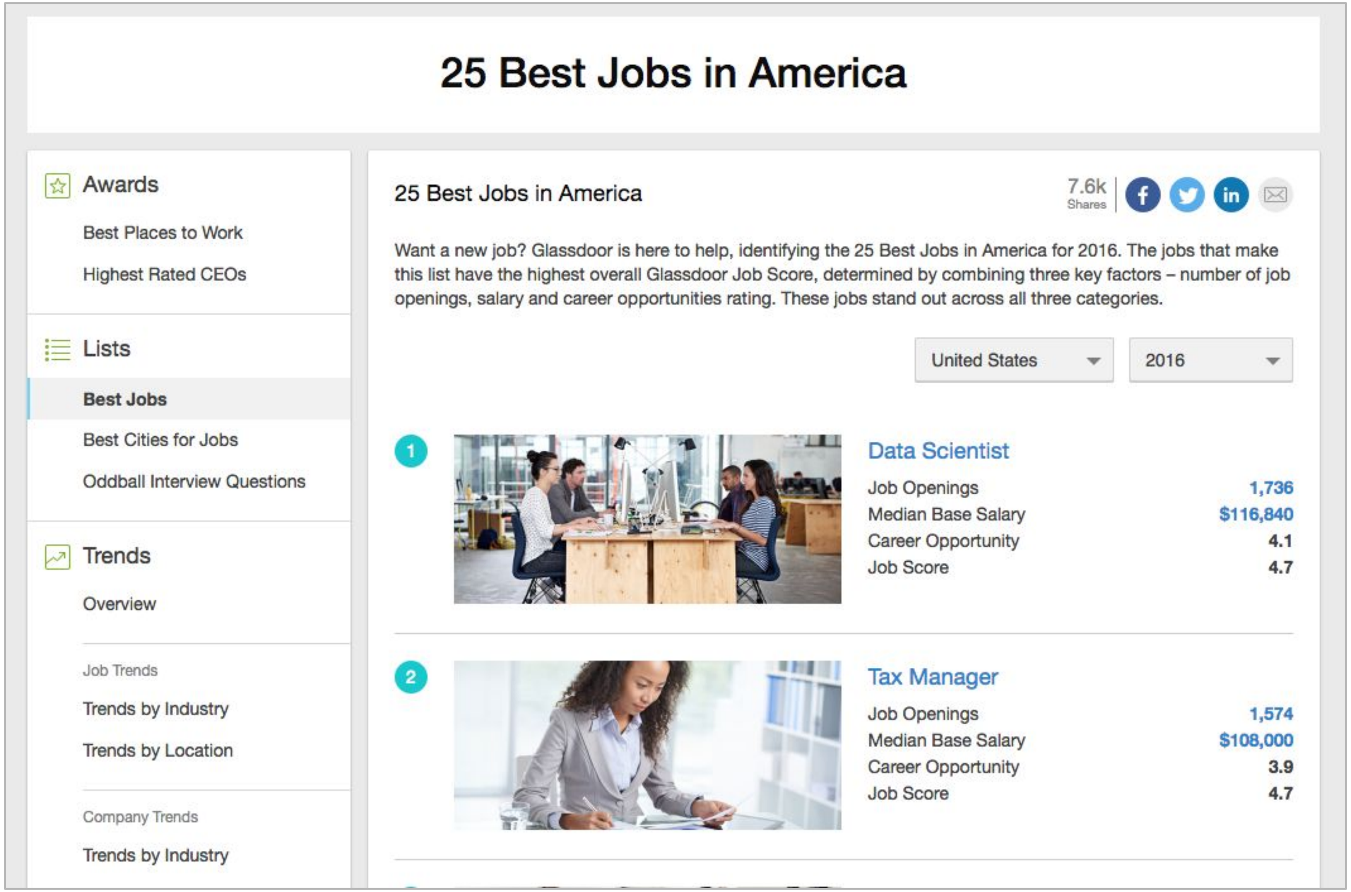
Senior U.S. Correspondent, IDG News Service | JAN 21, 2016 12:01 PM PT

There are more reasons than ever to consider a career in data science. Credit: International Data Group

MORE LIKE THIS

- Looking to land 2016's 'hottest job'? Here's what you need to be a data...
- Stop drowning your data scientists in drudgery
- IoT and the data-driven enterprise: How to dive into the data flood
- VIDEO How real-time analytics helped Styrofoam maker FloraCraft

Hot jobs



25 Best Jobs in America

25 Best Jobs in America

Want a new job? Glassdoor is here to help, identifying the 25 Best Jobs in America for 2016. The jobs that make this list have the highest overall Glassdoor Job Score, determined by combining three key factors – number of job openings, salary and career opportunities rating. These jobs stand out across all three categories.

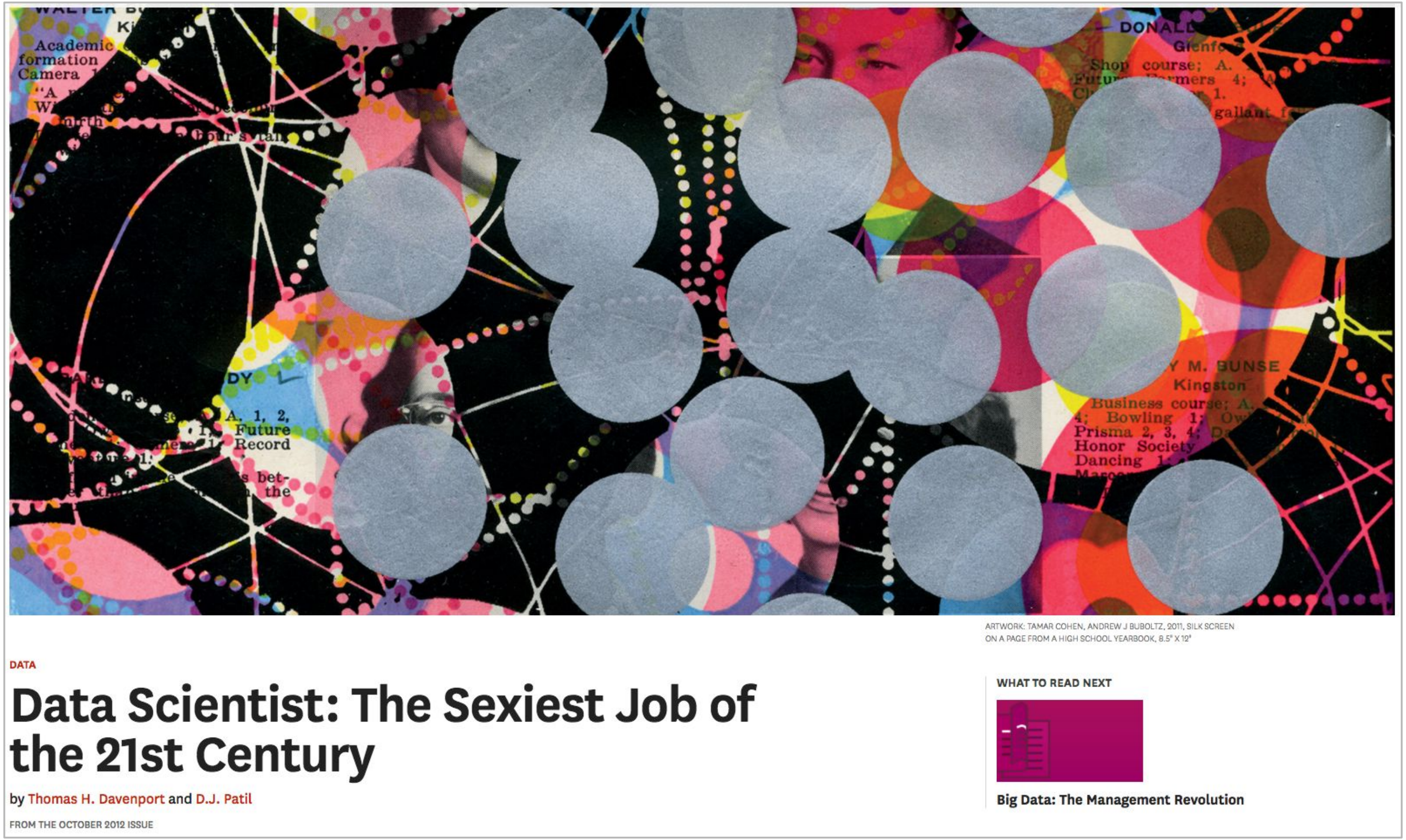
1 Data Scientist

Job Openings	1,736
Median Base Salary	$116,840
Career Opportunity	4.1
Job Score	4.7

2 Tax Manager

Job Openings	1,574
Median Base Salary	$108,000
Career Opportunity	3.9
Job Score	4.7

Best jobs



ARTWORK: TAMAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

WHAT TO READ NEXT

Big Data: The Management Revolution

Sexiest jobs

More importantly…

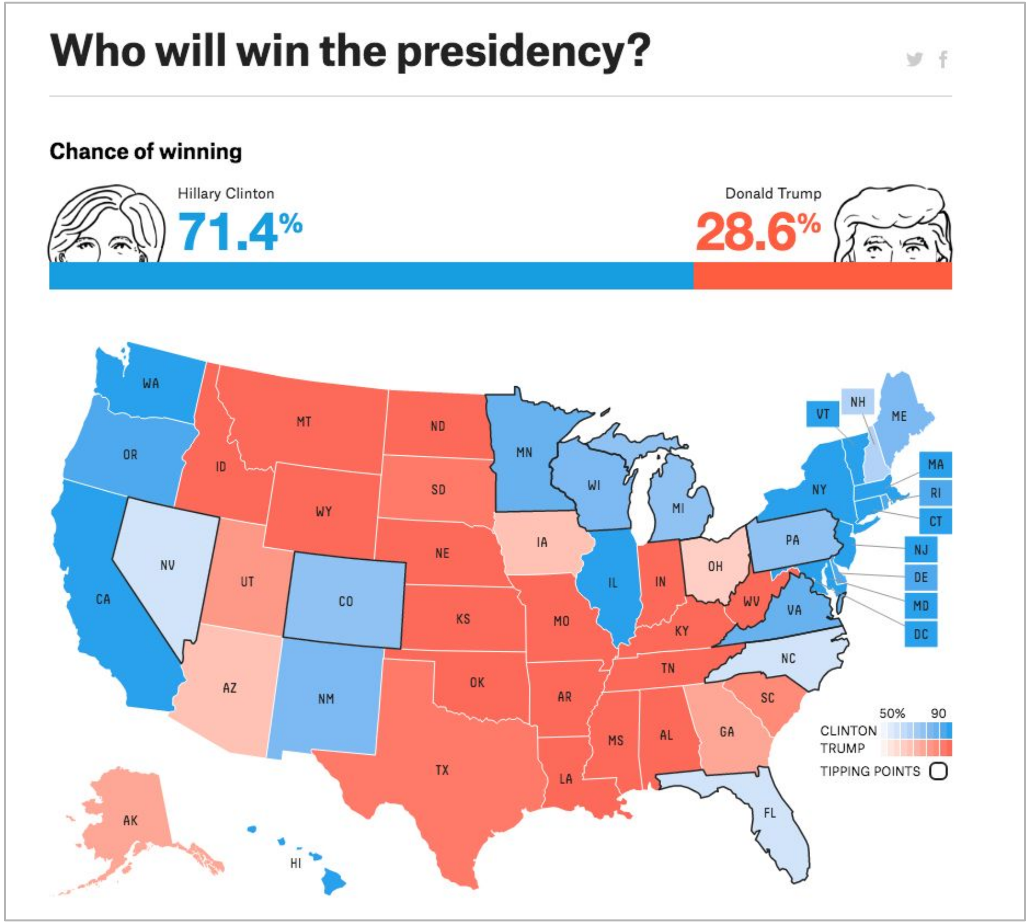
An Algorithm Could Know You Have A Genetic Disease Before You Do

By Anna Maria Barry-Jester

Filed under Public Health

REBECCA WHITEHURST

Healthcare



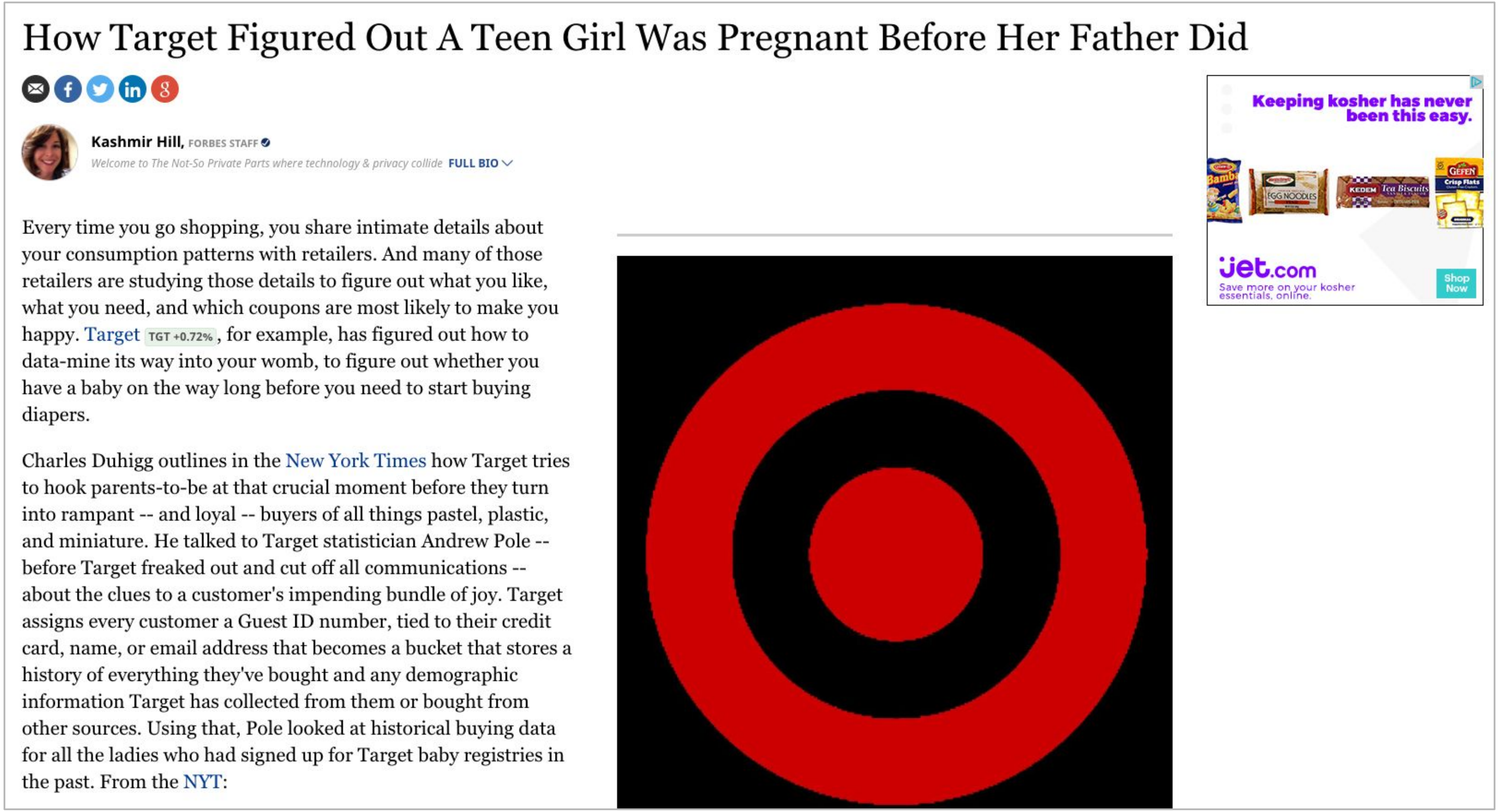
Who will win the presidency?

Chance of winning

Hillary Clinton
71.4%

Donald Trump
28.6%

Politics



How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill, FORBES STAFF

Welcome to The Not-So Private Parts where technology & privacy collide FULL BIO

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target TGT +0.72%, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Charles Duhigg outlines in the New York Times how Target tries to hook parents-to-be at that crucial moment before they turn into rampant -- and loyal -- buyers of all things pastel, plastic, and miniature. He talked to Target statistician Andrew Pole -- before Target freaked out and cut off all communications -- about the clues to a customer's impending bundle of joy. Target assigns every customer a Guest ID number, tied to their credit card, name, or email address that becomes a bucket that stores a history of everything they've bought and any demographic information Target has collected from them or bought from other sources. Using that, Pole looked at historical buying data for all the ladies who had signed up for Target baby registries in the past. From the NYT:

Ethical concerns

Today's Learning Objectives

Spend a few minutes **introducing** ourselves

Discuss the **learning objectives** of this course

Familiarize yourself with the **course structure**

Mutually agree on **course expectations**

Discuss **predicting recidivism** as a data science case study

Introductions

Faculty member at UW iSchool

Courses I Teach

INFO 200: Intellectual Foundations of Informatics

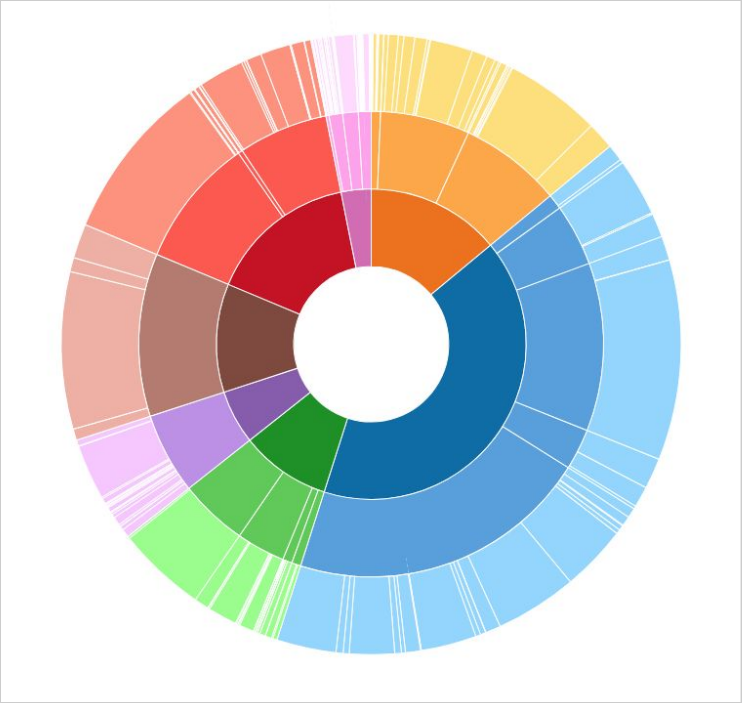
INFO 201: Technical Foundations of Informatics

INFO 328: Population Health Metrics

INFO 343: Client-side Web Development

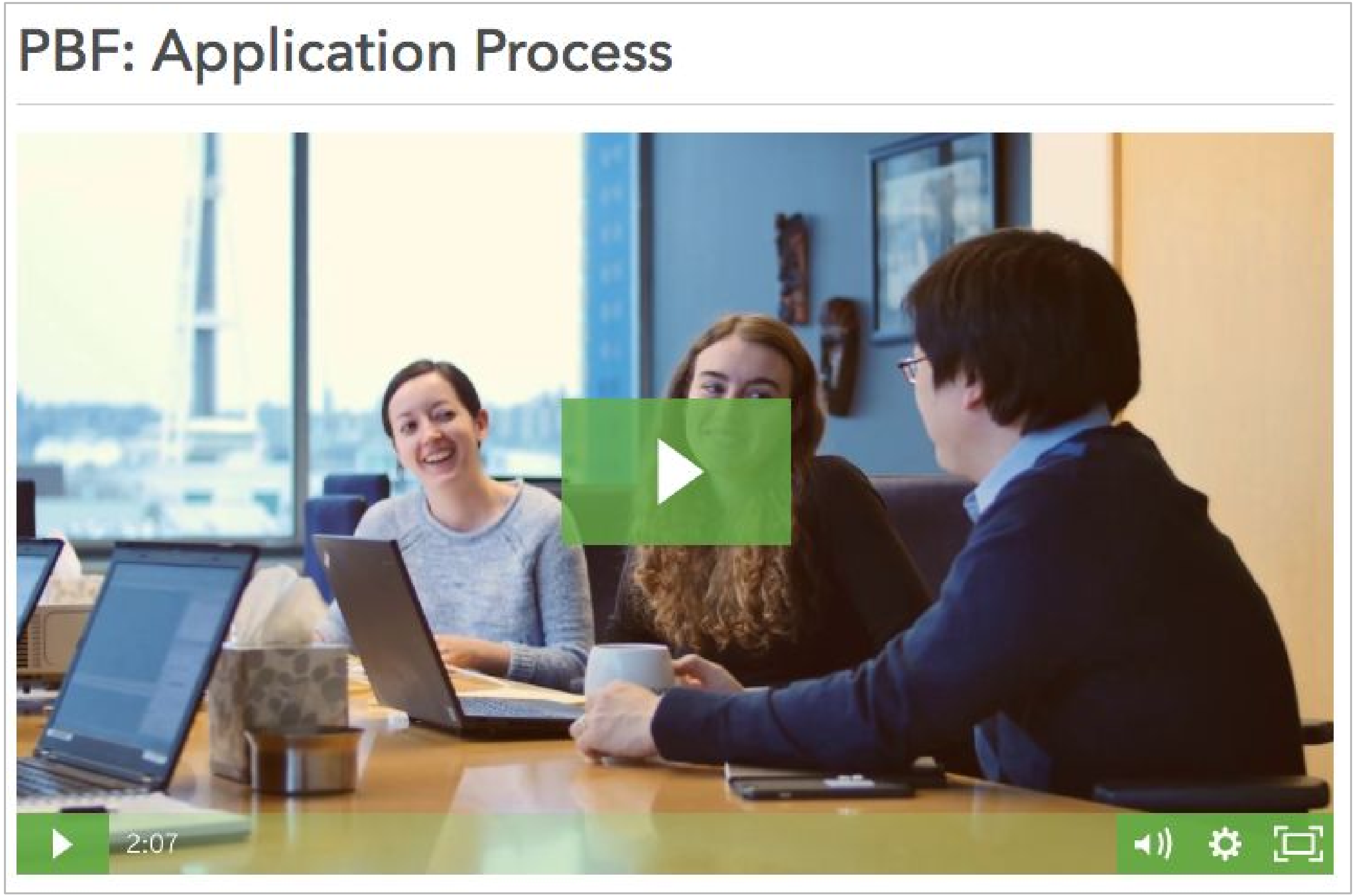
INFO 370: Introduction to Data Science

INFO 474: Interactive Data Visualization



Institute for Health Metrics and Evaluation

My Background (link)



PBF: Application Process

Post-bachelor Fellowship Program (Applications due 1/10, link)

Meet Amy and Aaditya, your helpful staff!

Brief Introductions

Name / pronoun(s) / year / major

Interest in the course

A non-academic interest of yours

Learning Objectives

What is data science?

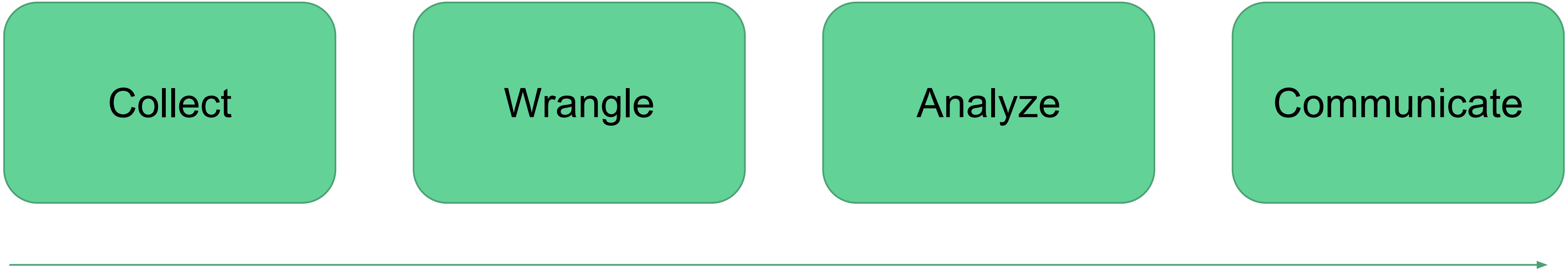
Foundation of being a data scientist

Use the **scientific method** to leverage **data** to answer questions

- Develop testable hypothesis for a given topic area
- Map from your concept (*segregation*) to a measurable outcome (*gini-index*)
- Perform appropriate tests on the dataset to assess hypothesis

Requires **programming skills** to comprehensively collect and interact with data

Leverages **statistics** and **machine learning** to perform tests

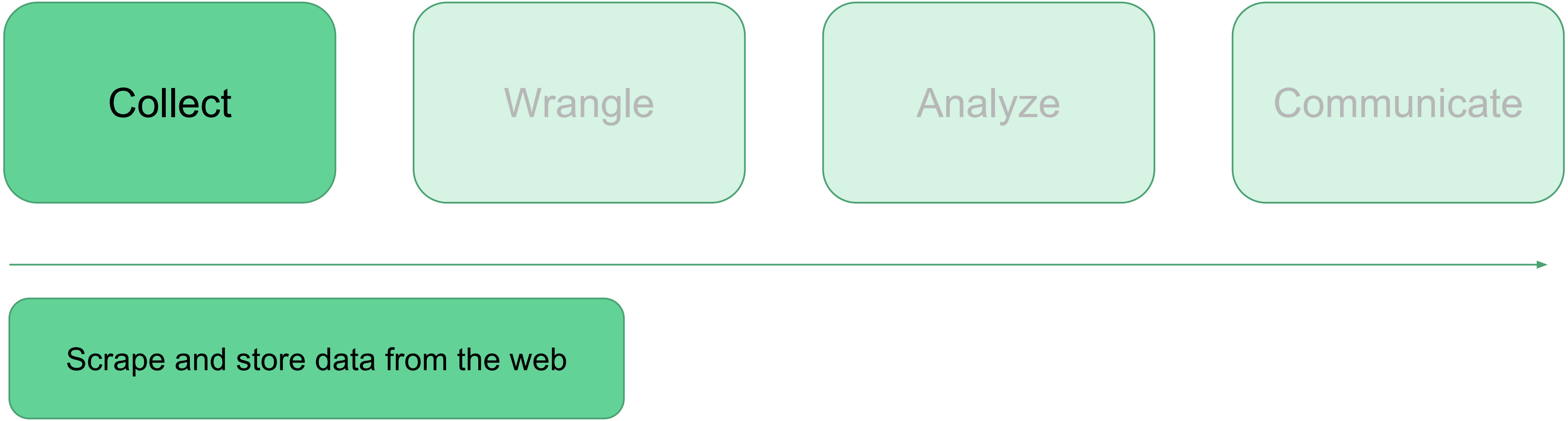


Collect

Wrangle

Analyze

Communicate



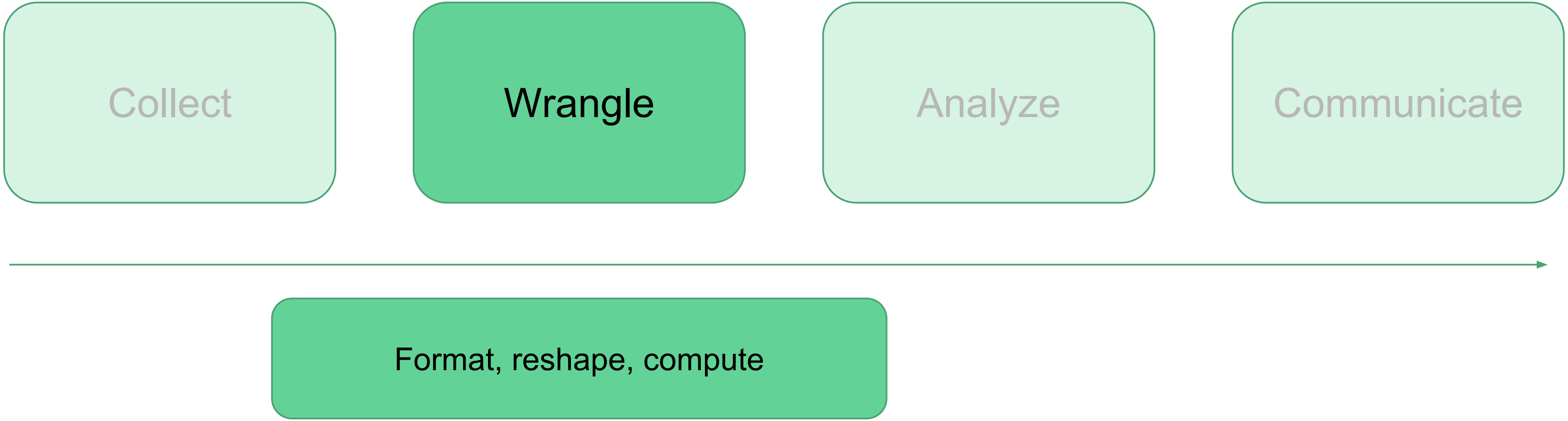
Collect

Wrangle

Analyze

Communicate

Scrape and store data from the web



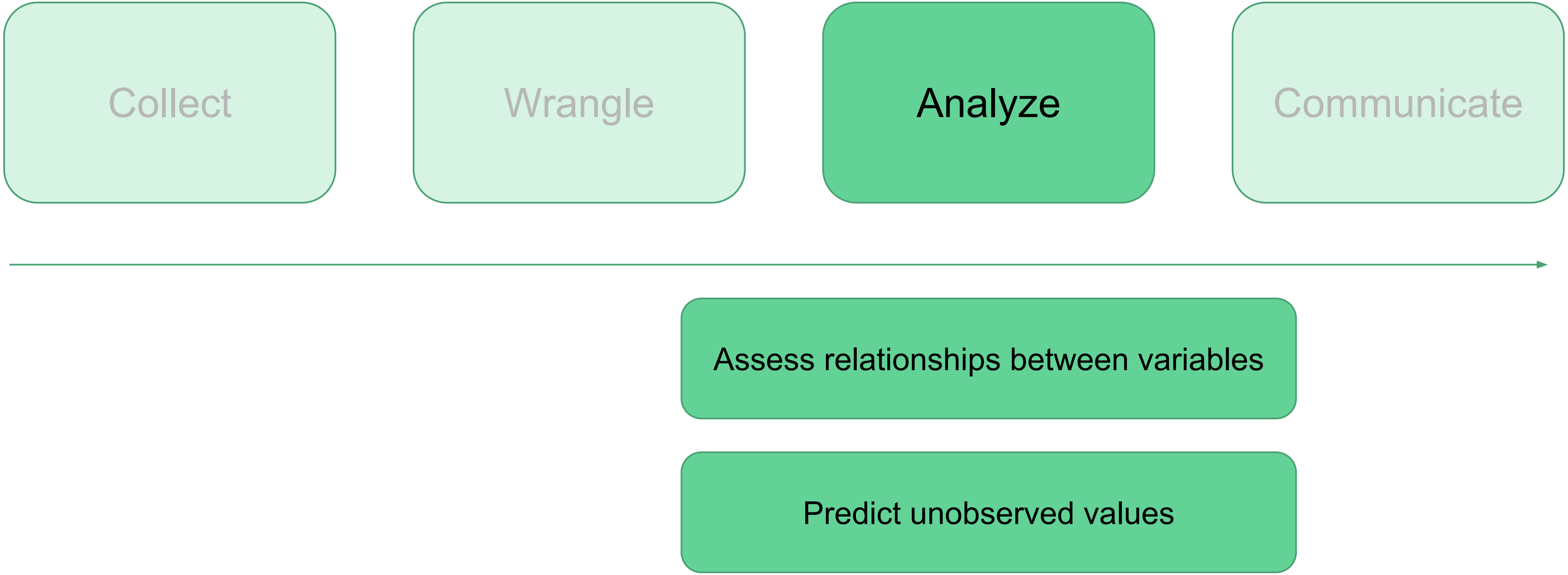
Collect

Wrangle

Analyze

Communicate

Format, reshape, compute



Collect

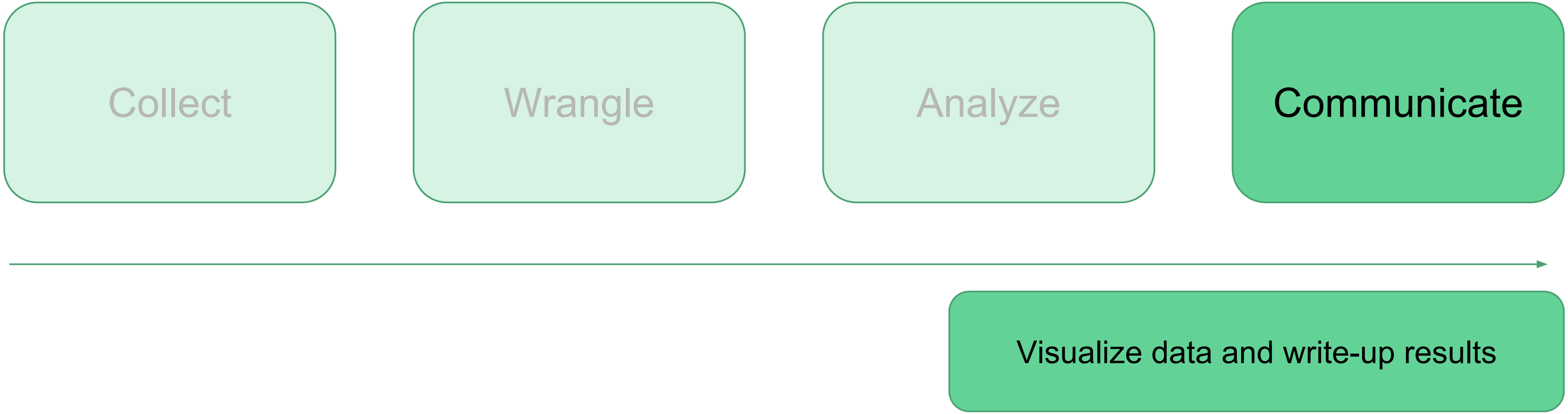
Wrangle

Analyze

Communicate

Assess relationships between variables

Predict unobserved values



Collect

Wrangle

Analyze

Communicate

Visualize data and write-up results

Learning Objectives

Conceptual	**Statistical Learning**	**Programming**
Purpose of data science	Probability Distributions	R/Python Basics, Code management
Mapping questions to methods	Central limit theorem	Data wrangling
Developing appropriate metrics	Hypothesis testing	Web-scraping
Basic visualization principles	Linear/Logistic regression	Visualization
Ethics in data science	Introductory machine learning methods	Statistical method implementation
		Machine learning implementation

This course will give you the foundational skills necessary to ***identify, implement, and interpret*** modern techniques.

Course Structure

Course Resources

Canvas: used for *submitting* **assignments** and accessing **slides** and **policies**

GitHub**:** where you will *save* **assignments** and access **class/lab activities**

Slack: how you will **collaborate**, ask **questions**, and see **announcements**

Python Basics: A short free/online book on python basics, written as an introductory programming resource

UW Resources

Listed on this Canvas page

- **Disability Resources**: DRS Office
- **Physical and Mental health**: Hall Health and UW Counseling Center
- **Academic Support:** Tutoring centers
- **Legal Support**: Student legal services

Assignments

Due at the *end* of each week

- **Notebooks**: In-class/lab activities graded on completion basis (20%)

Due at the *beginning* of each week

- **Readings:** 4 written responses, **not accepted late** (25%)
- **Assignments**: 4 hands-on assignments, require analysis + interpretation (30%)

Due at the end of the *quarter*

- **Group project:** Group research project, paper + online resource (25%)

Assignment Policies

Assignments/Notebooks: Penalized 10% each 24 hour period, down to 50%, until final lecture

Readings: ***will not*** be accepted after the deadline

Late-days*:* *three days* to use at your discretion (for **assignments** only)

Academic integrity

Collaboration is ***encouraged*** (especially for notebooks):

- Discuss and debate high-level ideas
- Work through challenging syntax issues

Plagiarism ***is not*** acceptable, including (but not limited to):

- Sharing code / written responses
- Representing someone else's work as your own

Often result in failing the assignment

Consequences may extend beyond this class

Time Breakdown

A 5 credit course is a **15 hour per week** commitment (link)

Class + Lab (5 hours)

- Working on notebooks (50%)
- Discussions (20%)
- Code-alongs // exercises (10%)
- Lecture (20%)

At Home (10 hours)

- Assignments (50%)
- Notebooks (10%)
- Readings (40%)

Course Expectations

What expectations do you have of one another and the teaching team?

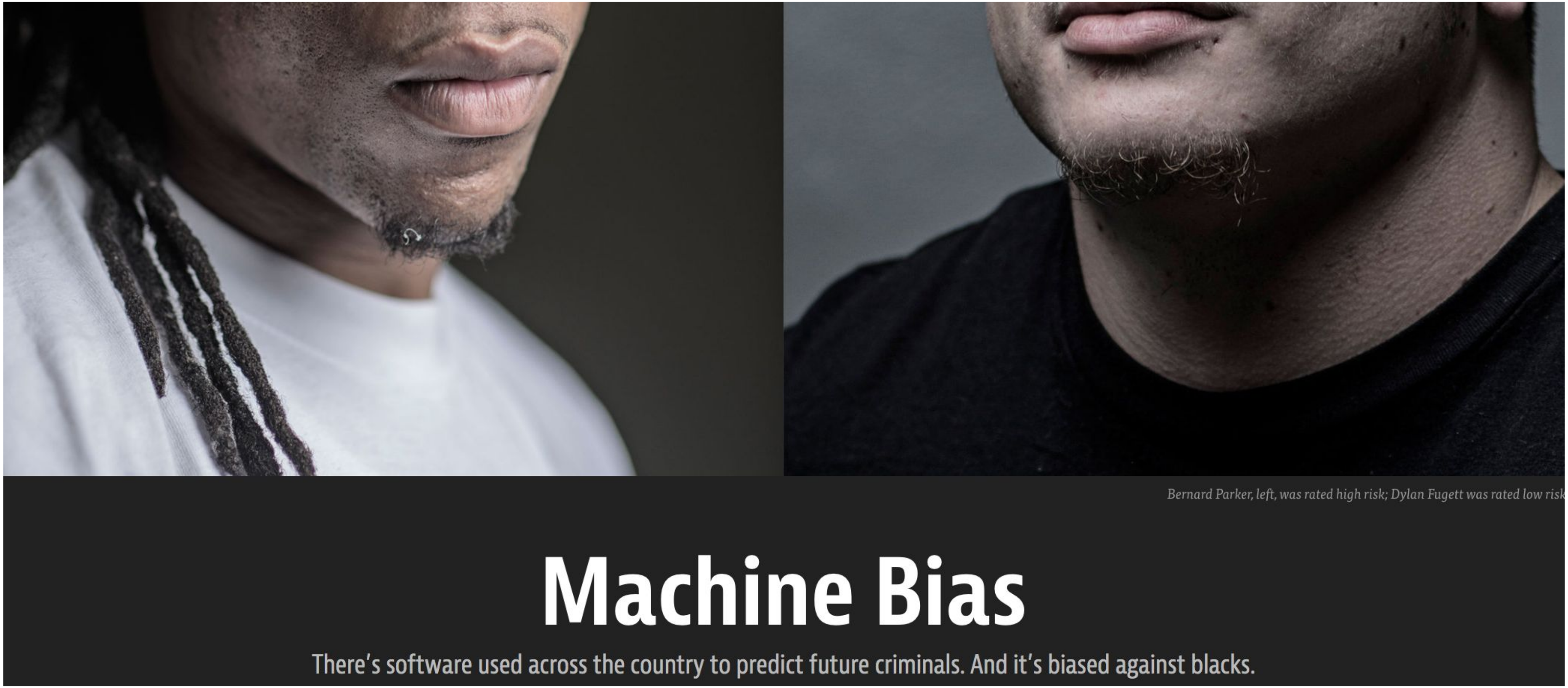
Expectations (to be detailed collectively)

Respect each others' time, intelligence, and experiences

Work collaboratively

Provide honest, timely, and direct feedback

Predicting Recidivism



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

Predicting recidivism (link)

Initial Discussion Questions

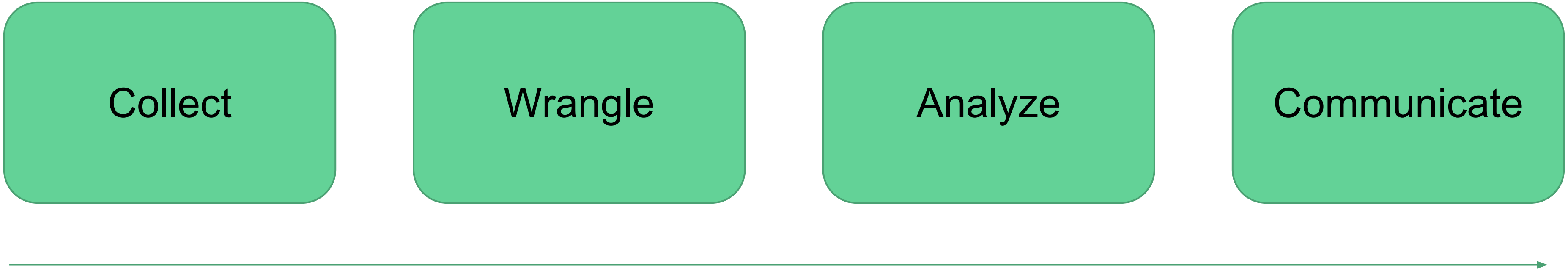
Why would you want to use data driven methods in criminal sentencing?

Who is responsible for assessing software used in the justice system?

If you don't use race in your analysis, how can you create unfair results by race?

What skills do you need to perform this analysis?

(and which ones are you missing?)

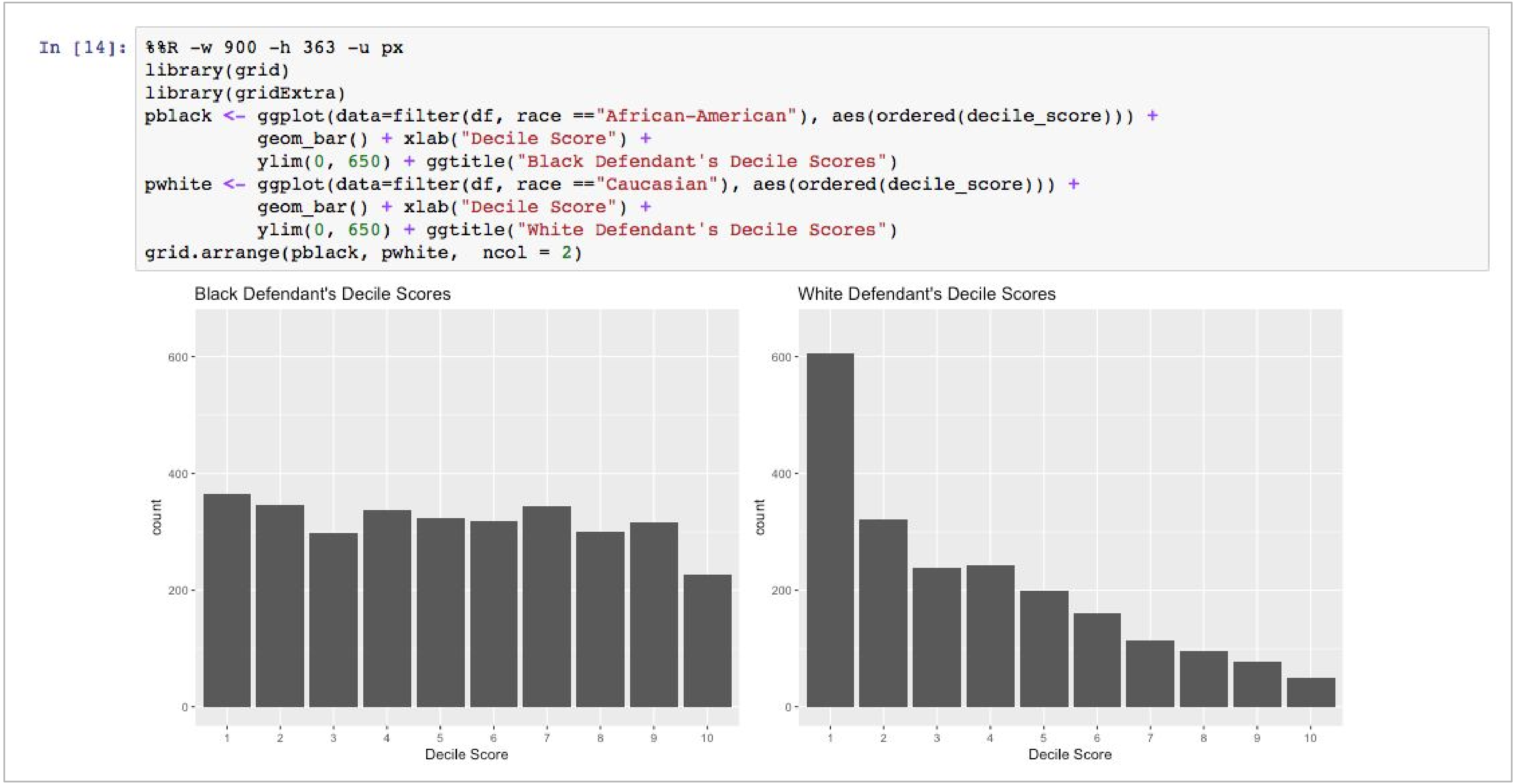


Collect

Wrangle

Analyze

Communicate



```
In [14]: %%R -w 900 -h 363 -u px
         library(grid)
         library(gridExtra)
         pblack <- ggplot(data=filter(df, race =="African-American"), aes(ordered(decile_score))) +
                   geom_bar() + xlab("Decile Score") +
                   ylim(0, 650) + ggtitle("Black Defendant's Decile Scores")
         pwhite <- ggplot(data=filter(df, race =="Caucasian"), aes(ordered(decile_score))) +
                   geom_bar() + xlab("Decile Score") +
                   ylim(0, 650) + ggtitle("White Defendant's Decile Scores")
         grid.arrange(pblack, pwhite,  ncol = 2)
```

How they did it (link)

Reading 1

Read and write responses to two data science articles

Short, but dense

Will require a bit of outside research

Responses written in Markdown

Due **before class** on Tuesday (not accepted after deadline!)

Upcoming...

Reading assignment 1 due ***next Tuesday (1/9) before class***

Install necessary software ***before class Tuesday*** (see book)

Next week

- Discuss reading assignment
- Python foundations for Data Science
- Read the python basics book (you might want to get started this weekend)