INFO 370

Introduction to Data Science

Learning Objectives

What have we done?

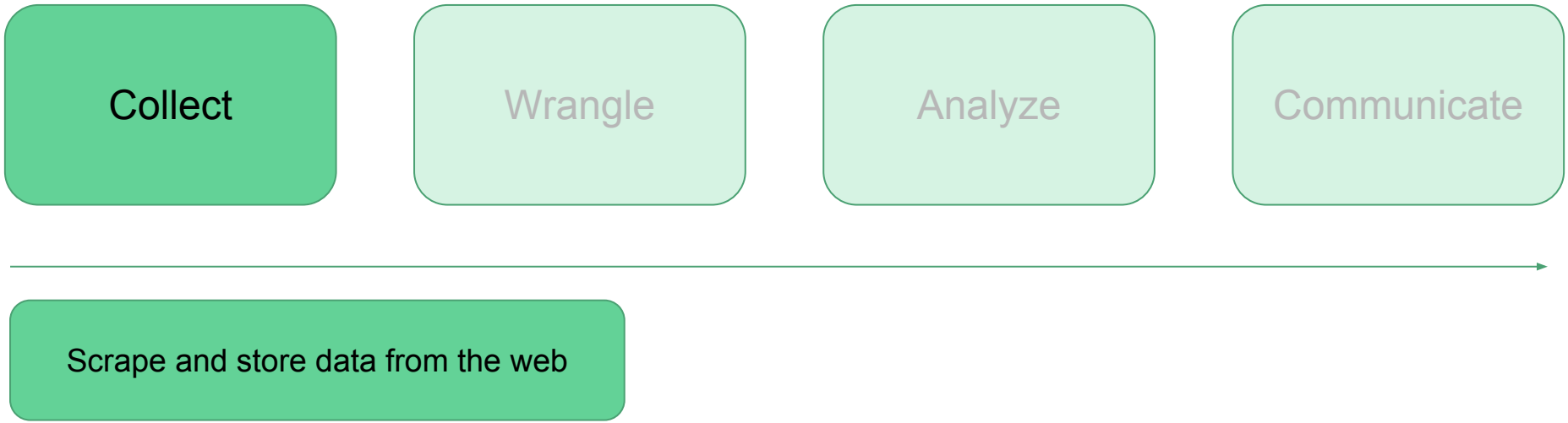
What can we do better?

What can you do after this class?

What have we done?

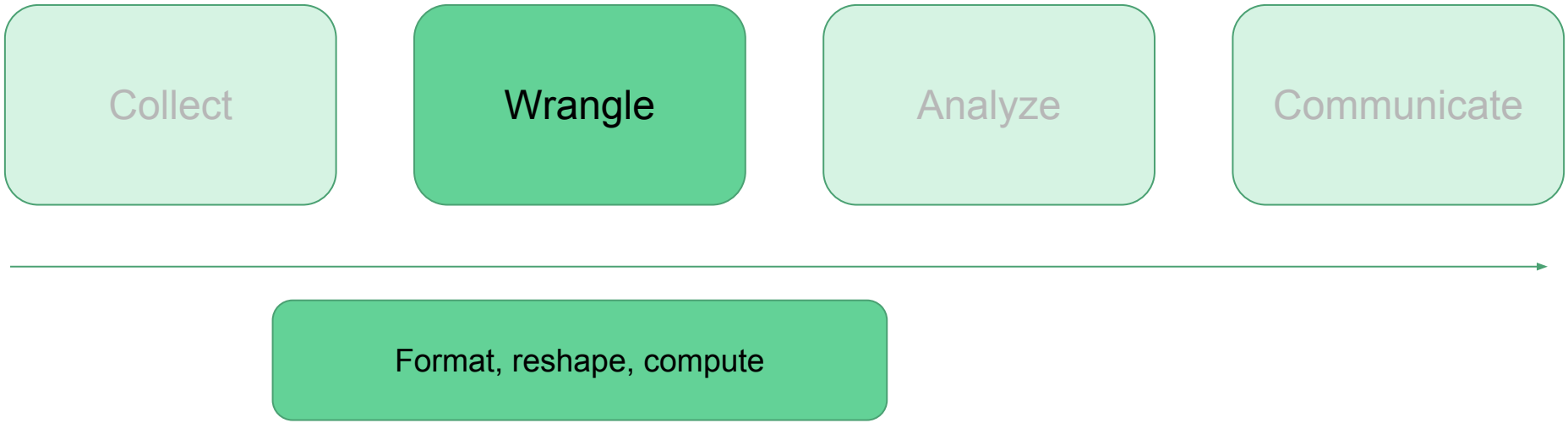
Studied data science

Collect
Wrangle
Analyze
Communicate


Collect
Wrangle
Analyze
Communicate
Scrape and store data from the web

Most Common 3 Words	URL of song lyrics	album	artist	lyrics
[('the', 16), ('my', 15), ('me', 9)]	http://www.lyrics.com/lyric/24034647	+ [2011]	Ed Sheeran	This city never sleeps I hear the people v
[('the', 11), ('drunk', 10), ('I', 10)]	http://www.lyrics.com/lyric/24034653	+ [2011]	Ed Sheeran	I wanna be drunk when I wake up On the
[('my,', 81), ('me', 38), ('give', 27)]	http://www.lyrics.com/lyric/24034643	+ [2011]	Ed Sheeran	Give me love like her 'Cause lately I've b
[('you', 26), ('never', 25), ('let', 18)]	http://www.lyrics.com/lyric/24034651	+ [2011]	Ed Sheeran	My mind is a warrior, My heart is a foreig
[('in', 18), ('be', 14), ('wanna', 12)]	http://www.lyrics.com/lyric/24034644	+ [2011]	Ed Sheeran	Settle down with me Cover me up Cuddl
[('out', 21), ('of', 21), ('you', 19)]	http://www.lyrics.com/lyric/24034646	+ [2011]	Ed Sheeran	I'm gonna pick up the pieces And build a
[('to', 23), ('in', 14), ('the', 11)]	http://www.lyrics.com/lyric/24034654	+ [2011]	Ed Sheeran	White lips, pale face Breathing in snowfla
[('the', 8), ('me', 5), ('of', 5)]	http://www.lyrics.com/lyric/24034648	+ [2011]	Ed Sheeran	This is the start of something beautiful Th
[('I', 34), ('the', 16), ('that', 14)]	http://www.lyrics.com/lyric/24034652	+ [2011]	Ed Sheeran	I found your hairband on my bedroom flo
[('you', 22), ('a', 13), ('the', 12)]	http://www.lyrics.com/lyric/24034650	+ [2011]	Ed Sheeran	I should ink my skin With your name And
[('need', 47), ('I', 39), ('the', 24)]	http://www.lyrics.com/lyric/24034645	+ [2011]	Ed Sheeran	Now I'm in town, break it down, thinking
[('little', 14), ('I', 11), ('to', 10)]	http://www.lyrics.com/lyric/30285896	+ [2011]	Ed Sheeran	If we take this bird in With its broken leg

Scraping lyrical data


Collect
Wrangle
Analyze
Communicate
Format, reshape, compute

Interaction and Isolation

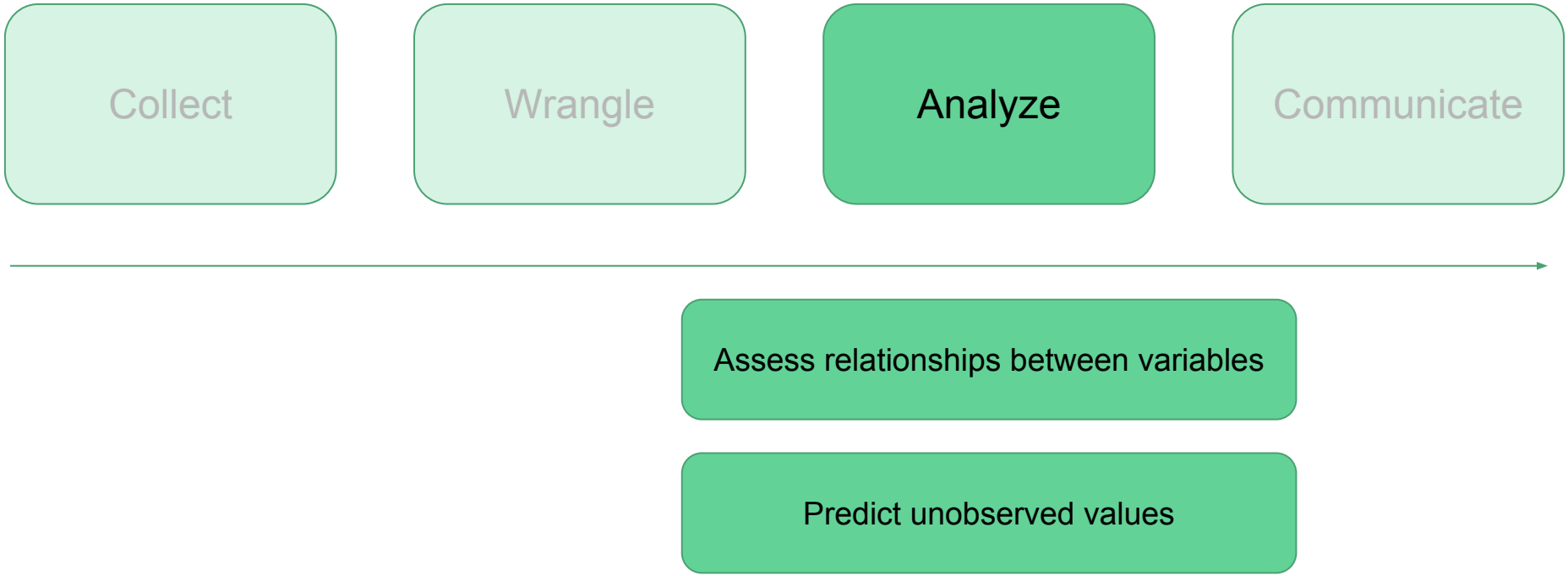
Interaction and Isolation metrics work hand in hand, since both measures the "probabilities that a minority person shares a unit area with a majority person or with another minority person." (U.S. Census Bureau) Interaction meansures the likelihood of a minority person interacting with a majority person. Vice-versa, isolation measures the likelihood of a minority person being exposed to another minority person. Below are its formulas:

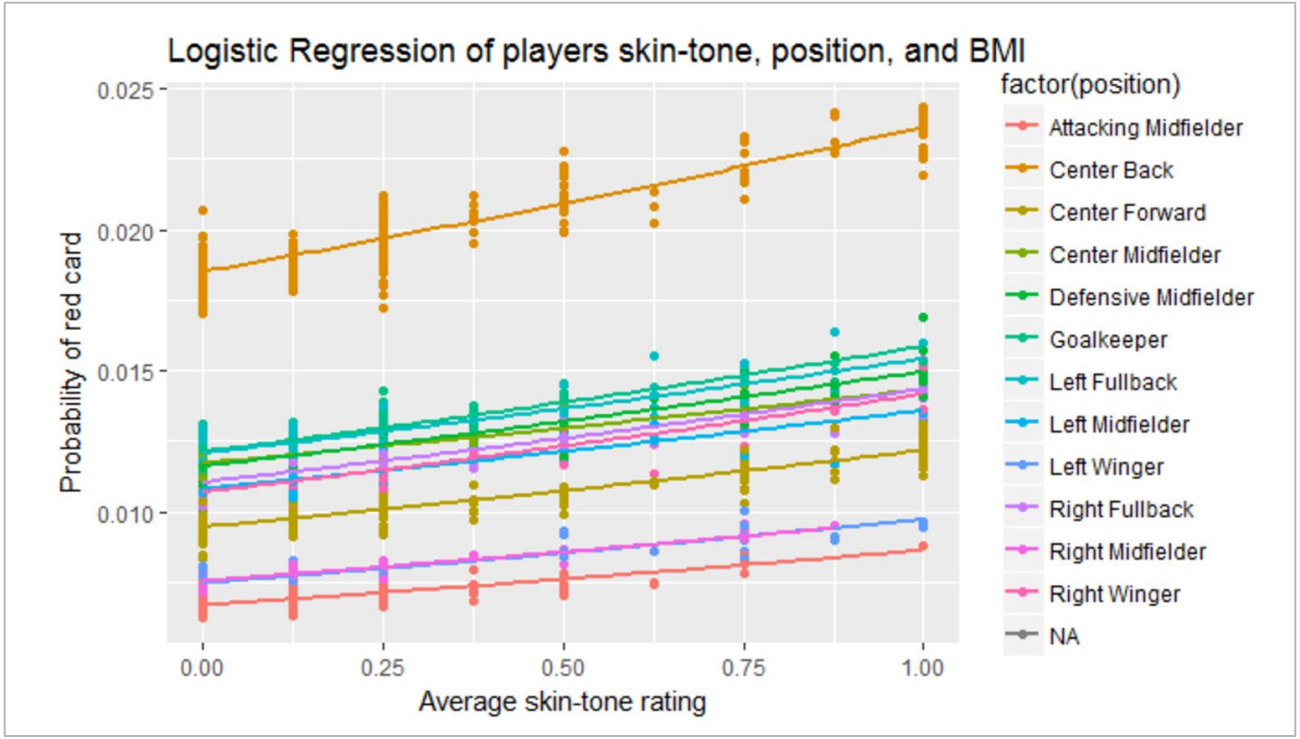
Interaction:

$$\sum_{i=1}^{n}\left[\left(\frac{x_i}{X}\right)\left(\frac{y_i}{t_i}\right)\right]$$

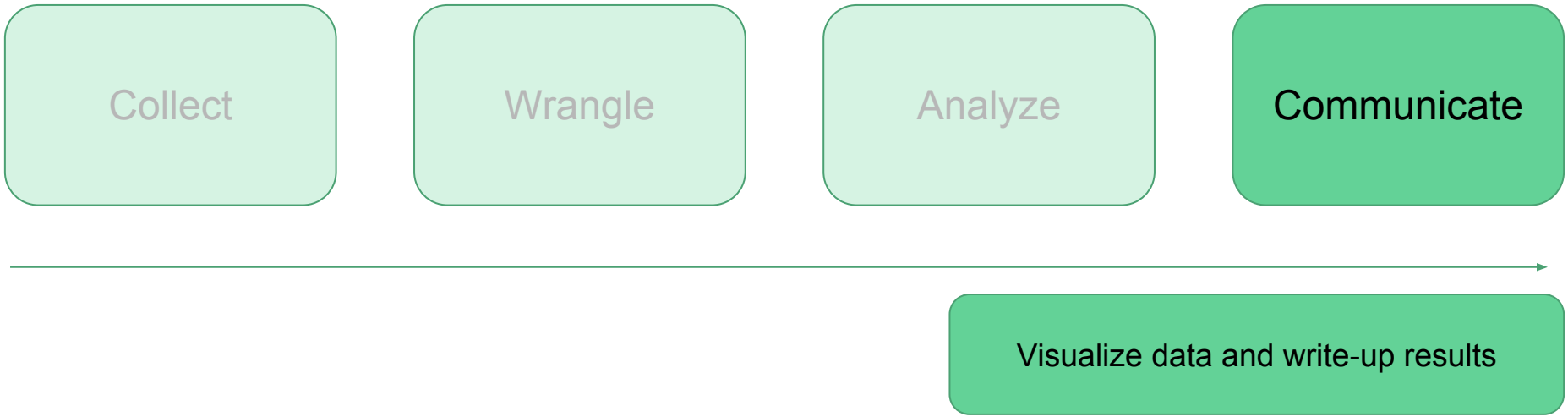
Isolation:

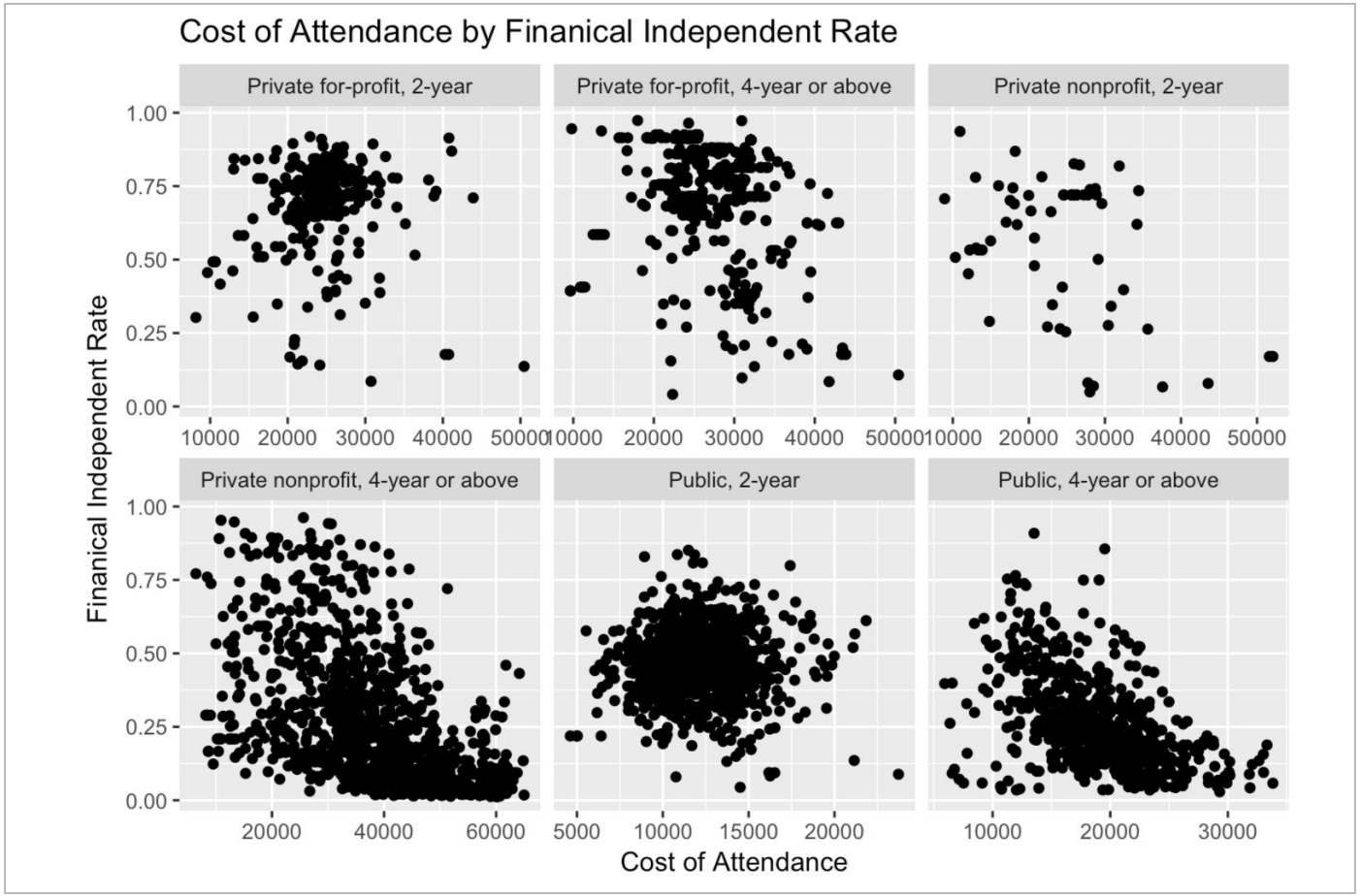
$$\sum_{i=1}^{n}\left[\left(\frac{x_i}{X}\right)\left(\frac{x_i}{t_i}\right)\right]$$


Collect
Wrangle
Analyze
Communicate
Assess relationships between variables
Predict unobserved values



Analyzing implicit racial bias


Collect
Wrangle
Analyze
Communicate
Visualize data and write-up results



Answer questions with visualization

Learn how to acquire the ***programming skills*** and ***statistical knowledge*** to ***ask questions*** of data.

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

Learning Objectives

What can be done better?

Please fill out course evaluation.

Course Improvements: INFO 370, Winter 2018

Summary

Students gained the technical and conceptual skills necessary to engage in the data science process (collect, wrangle, analyze, communicate data). This was achieved through weekly (jupyter) notebooks, readings with written responses, and assignments that brought together these technical and conceptual skills. A few ways in which these can be improved:

Notebooks

Notebooks ensured the acquisition of the necessary (python) skills. Some improvements:

- **Foundational Skills:** More time was needed to practice python fundamentals.
- **Instruction Specificity**: Instructions need to be more specific (at times), though expectations of "completion" can be more clearly stated.
- **Time commitment**: For some students, these were a burden -- by increasing foundational skills, however, the time commitment should be appropriate.

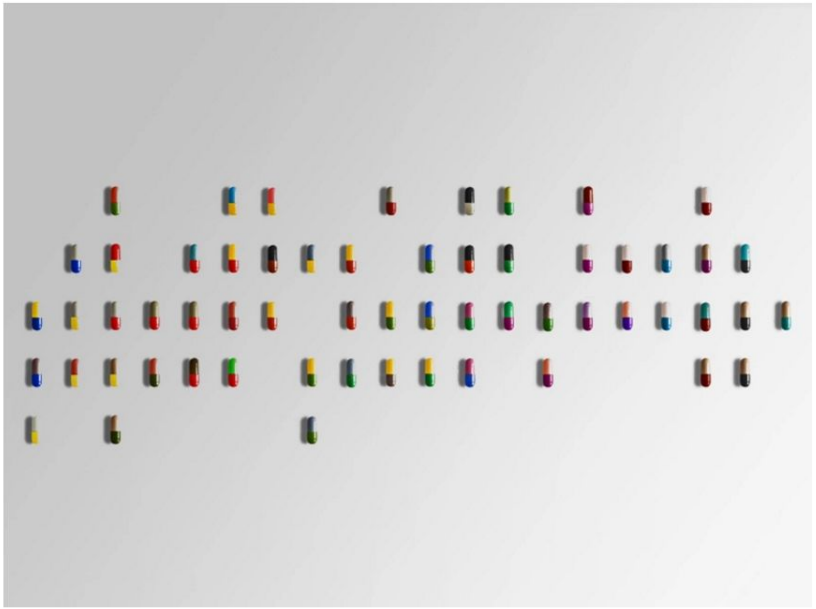
What's next?

Use data science to answer ***pressing social questions***, and be wary of the ***unintended consequences*** of your actions.

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

By Anna Maria Barry-Jester

Filed under Public Health



REBECCA WHITEHURST

Get involved in a variety of fields (healthcare)

Be really careful (source)

Be really careful (source, code)

Thanks, Amy and Aaditya!

Upcoming...

Final resource due ***before class on Thursday***

Final paper due ***Sunday night at midnight***

Peer evaluations are due ***next Tuesday by midnight***

Please fill out your course evaluations

This week

- Final resource fair

Thanks.