CSCI 3022

intro to data science with probability & statistics

Lecture 1
January 17, 2018

1. What is data science?
2. What will we learn in this course?
3. My friend Anna's instagram

Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Dan Larremore

What is data science?

is there *non*-data science?

yes: using data to **understand the world**

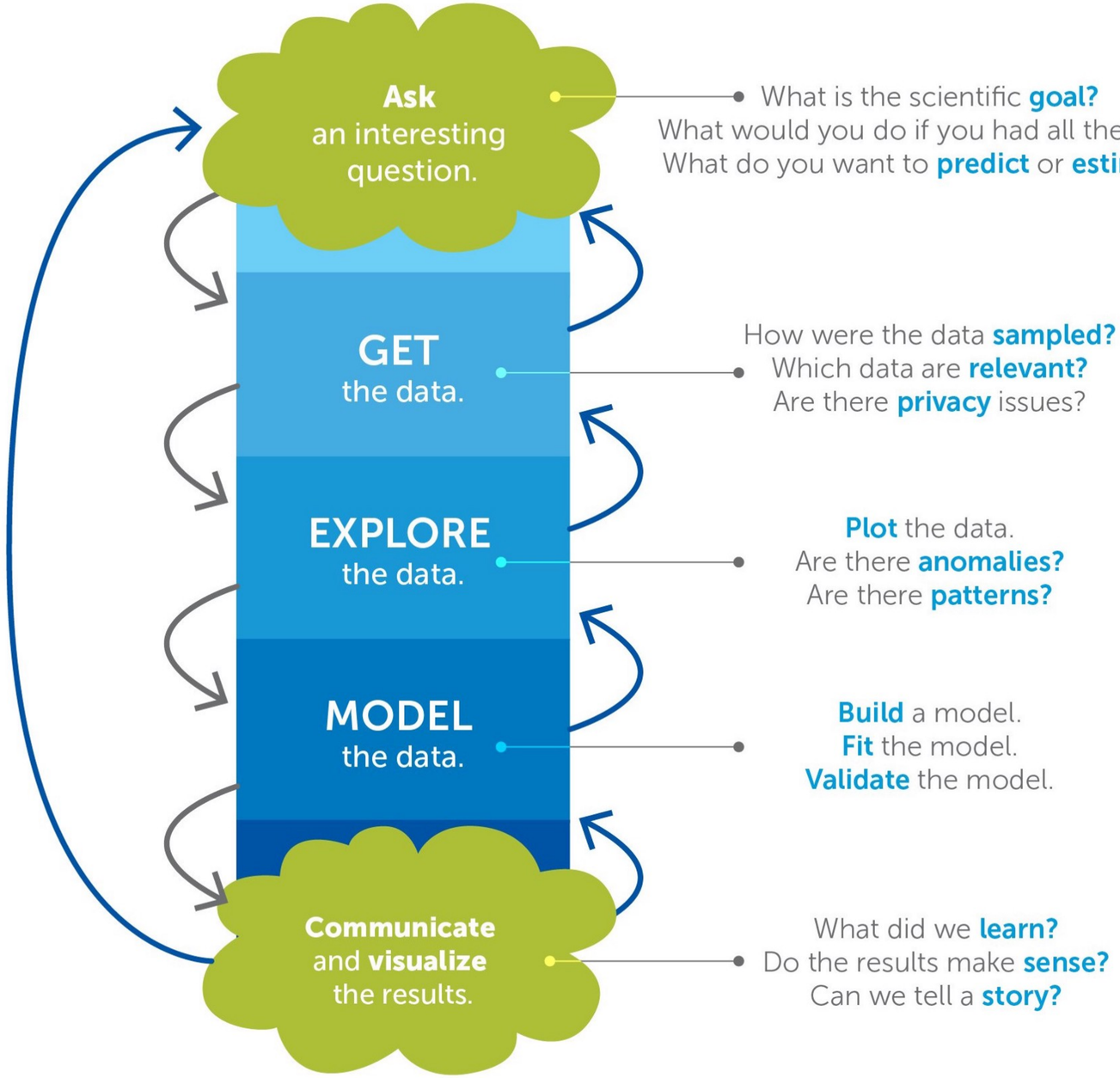
yes: **recovering insights/trends** that are hiding behind data

yes: applying statistically rigorous techniques to data to **find answers to questions**

no: more about data than science

no: storytelling with data

Data science sounds a lot like…science!



Hypothesis

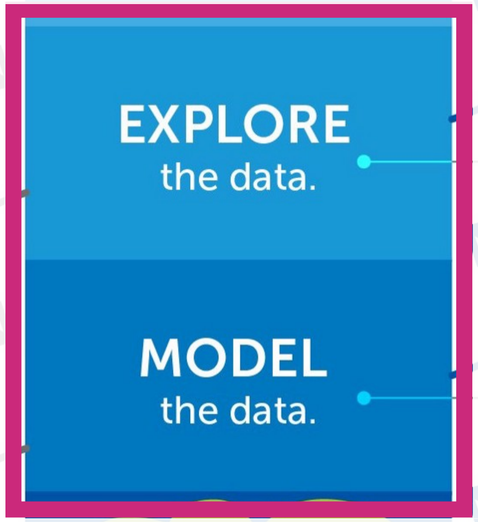
Observation

Analysis - *what?*

Analysis - *why* & *how*?

Conclusions

Derived from the work of Joe Blitzstein and Hanspeter Pfister,
originally created for the Harvard data science course http://cs109.org/.



in this class, we're going to build toward the core topics in exploration & modeling

1. data mining [discover]
2. statistical analysis [understand]
3. machine learning [predict]

in this class, we're going to build toward
the core topics in exploration & modeling

1. data mining [discover]
2. statistical analysis [understand]
3. machine learning [predict]

foundations:

probability	EDA, null models & null hypotheses, decision trees
statistical inference	averages, regression models, max. likelihood estimates
optimization & calculus	model fitting, math shortcuts
linear algebra	any time we've got a matrix… (or can make one!)
computer science	data structures, rapid estimation, simulation

Week	Date	nb	txt	Topic	Slides	Hmwk
1	01.17			Course & Computing Introduction		
	01.19		16.1-3	EDA and Summary Statistics		
2	01.22		15.1-2,16.4	EDA and Data Visualization		hw1 posted
	01.24			**Data Wrangling**		
	01.26		2	Introduction to Probability		
3	01.29		2,3	Axioms and Theorems of Probability		
	01.31		6	**Stochastic Simulation**		
	02.2		3	Bayes' Rule and Intro to PDFs		hw1 due
4	02.5		4	Discrete RVs, PMFs, CMFs		hw2 posted
	02.7		4,5	Discrete RVs Strike Back		
	02.9			**Return of the Discrete RVs**		
5	02.12		5	Continuous RVs Awaken, PDFs, CDFs		
	2.14			**The Last Continuous RVs**		
	02.16		7	Expectation		hw2 due
6	02.19		7	Variance		hw3 posted
	02.21			**More Expectation & Variance**		
	02.23		5.5	The Normal Distribution		
7	02.26			**MIDTERM EXAM REVIEW**		
	02.28		14	The Central Limit Theorems		
	02.28			**MIDTERM EXAM (PM)**		
	03.2			**The Central Limit Theorem and You**		hw3 due
8	3.5		23,24	Inference and CI Intro		hw4 posted
	3.7		23,24	Two-Sample CIs		
	03.9			**CIs in the Wild**		
9	03.12		25,26	Hypothesis Testing Intro		
	03.14		25,26	p-Values		
	03.16			**Practical HT & p**		hw4 due
10	3.19		27	Small-sample HT		hw5 posted
	3.21		18,23.3	Bootstrap Theory		
	3.23			**Bootstrap Practice**		
12	04.2		22	OLS/SLR Regression		
	04.4			**OLS/SLR Regression**		
	04.6		27	Inference in SLR		hw5 due
13	4.9		ISL Ch3	MLR		hw6 posted
	04.11		ISL Ch3	Inference in MLR		
	04.13			**Practical MLR**		
14	04.16			ANOVA		practicum posted
	04.18			**ANOVA**		
	04.20			Logistic Regr. & Classification		hw6 due
15	04.23			**Logistic Regr. & Classification**		
	04.25			Solution Techniques for OLS & LogReg		
	04.27			**Solution Techniques for OLS & LogReg**		
16	04.30					
	05.2			**FINAL EXAM REVIEW**		practicum due
X	05.X			**FINAL EXAM**		

- exploratory data analysis
- probability theory & simulation
- hypothesis testing & inferential statistics
- modeling, classification, prediction
- cleaning, munging, wrangling data

the plan

Goal: Fluency in the theoretical and computational aspects of data analysis.

At the end of this course you'll be able to

1. Clean, munge, and **wrangle data** in Python and perform Exploratory Data Analysis.
2. **Draw insight** from data by computing and interpreting classic summary statistics.
3. Know the ins-and-outs of probability and how to use it to **solve real-world problems**.
4. Perform statistical tests to **determine if your conclusions are real** or due to chance.
5. Construct and analyze simple models to **make predictions** and inferences about data.
6. **Tell compelling stories** about data using modern visualization and presentation tools.

course logistics 1 - web resources

Favorite the course pages now (Piazza & GitHub)

Piazza: https://piazza.com/colorado/spring2018/csci3022

No emails plz. Send me a private message on Piazza.

GitHub: https://github.com/dblarremore/csci3022

In-class work posted here. Homework posted here.
Clone the repo and then do a pull every day before coming to class.
Git tutorials:
http://rogerdudler.github.io/git-guide/
https://github.com/rochelleterman/PS239T/blob/master/15_Git/quick-n-dirty-git.md

course logistics 2 - grades

Homework (35%)

Every 2 weeks.
Lowest score dropped.
3 *total* late days. Rounded up: anything from 1s - 24 hours late = 1 late day

Class participation (5%)

Tutorial problems & short Moodle Quizzes

Midterm Exam (20%)

Practicum (15%)

Final Exam (25%)

Note: 55% average on the two exams is required to pass.

course logistics 3 - collaboration policy

- Data science is a collaborative field. Discuss problems with classmates & instructors
- But you *must* do your own work. **Write solutions and code on your own**.
- Give **hints**, not solutions, on Piazza.
- Make repositories that contain your homework private (GitHub, Azure).
- Details on syllabus. [link]

course logistics 4 - python & jupyter

- We'll use *python3*—with lots of *numpy* and *pandas*.
- We'll work exclusively in Jupyter Notebooks.
- Easiest way to get both is **Anaconda Python 3.6**
- I strongly recommend that you install a local copy (i.e. on your computer)
- We'll often work on problems in groups in class.
- Bring a laptop or buddy up!

let’s syllabus: https://piazza.com/colorado/spring2018/csci3022/home

about me

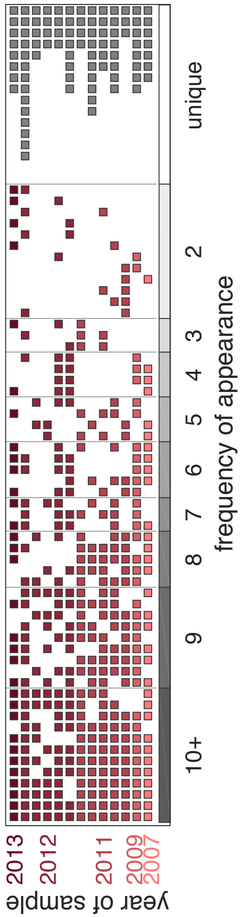
Office Hrs: FLMG 417 | W 11-1 | F 8-9:50

Assistant Professor, BioFrontiers Institute & Department of Computer Science

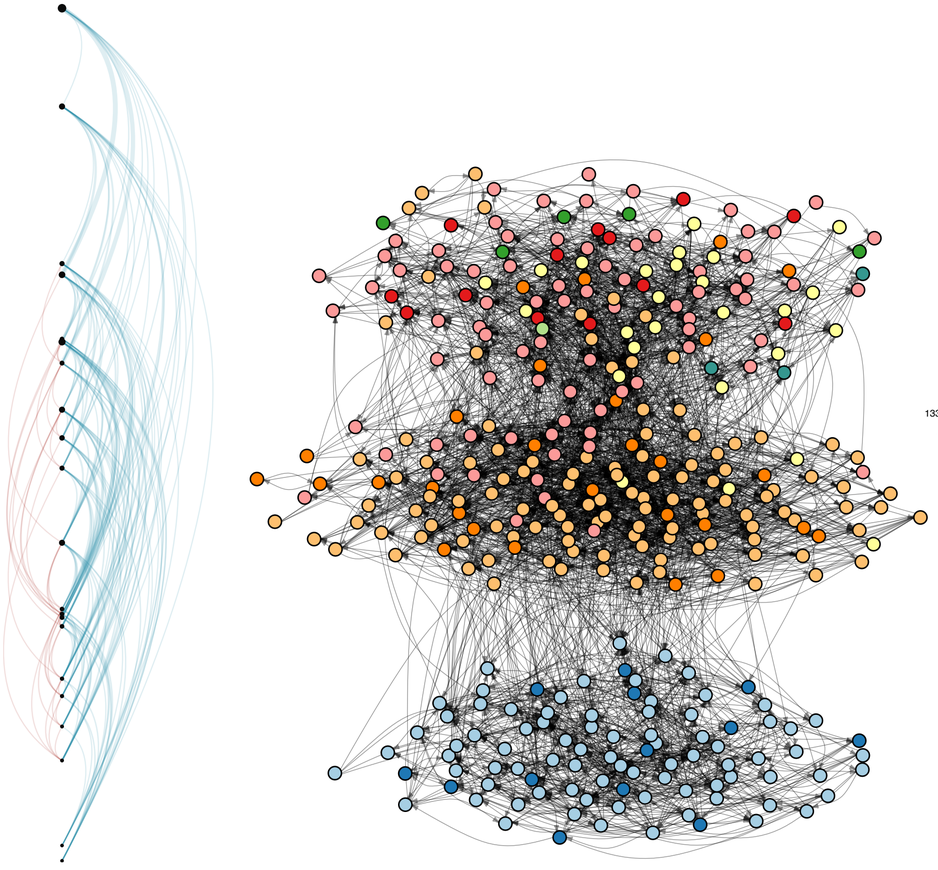
Previously: fellowships at Harvard, Santa Fe Institute; PhD CU Applied Math

research: danlarremore.com

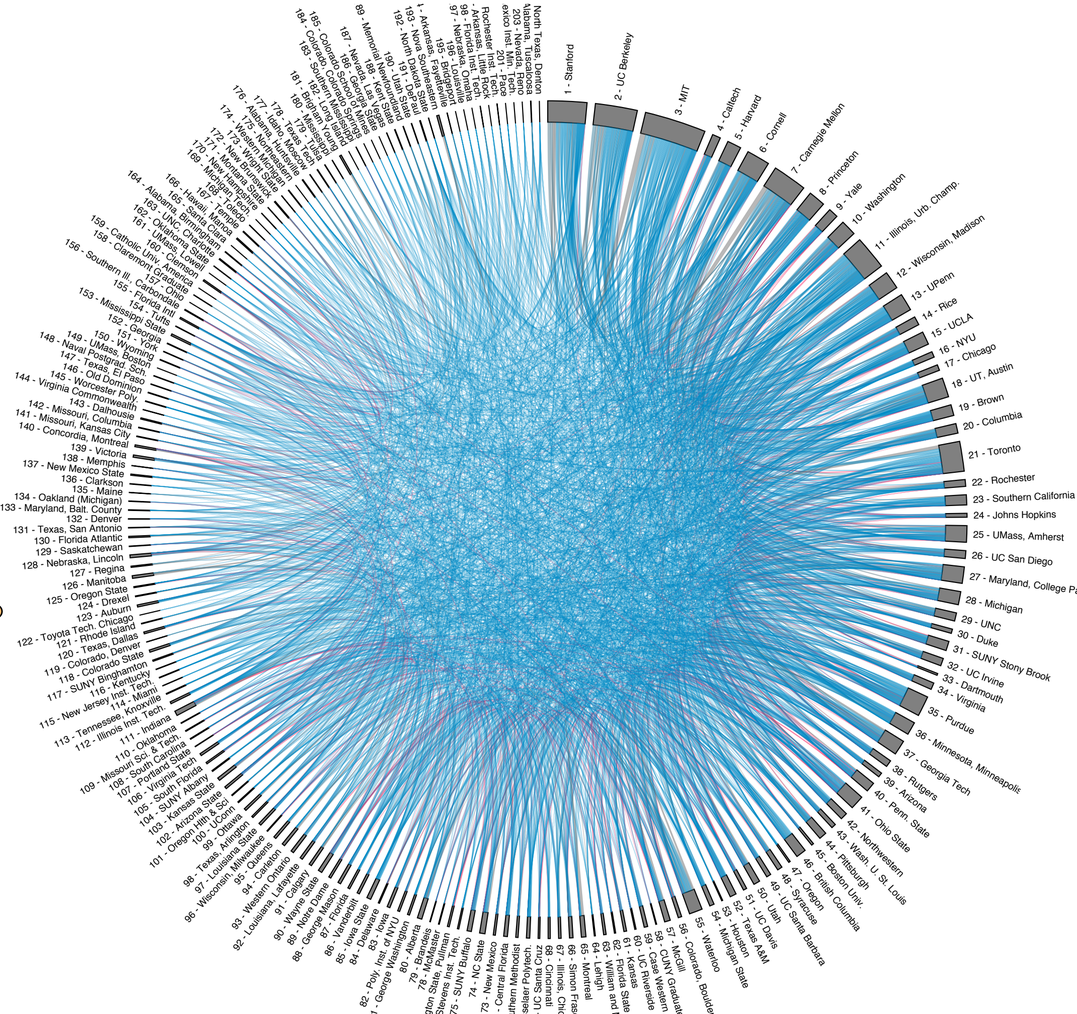
malaria parasite evolution and epidemiology

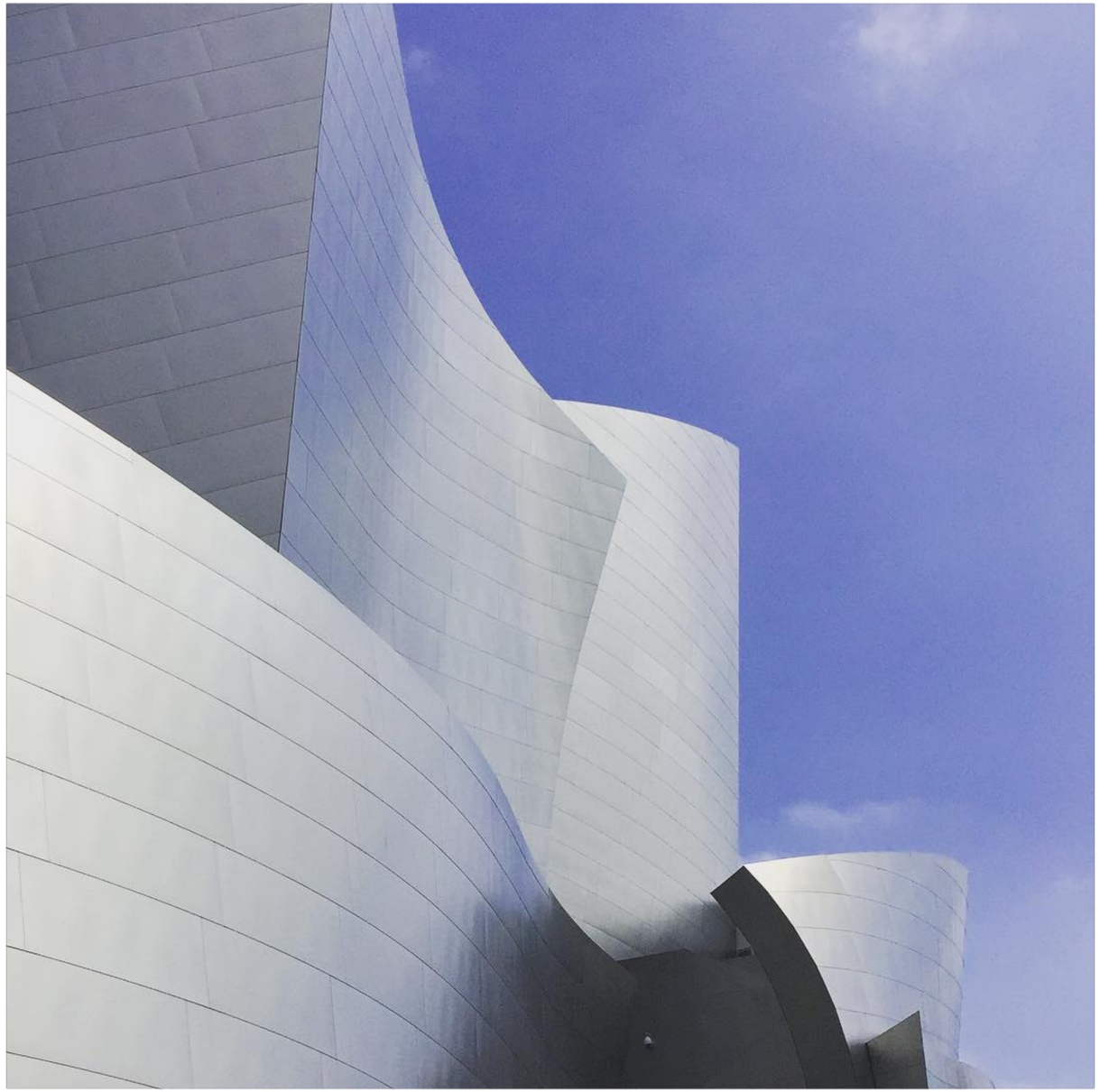


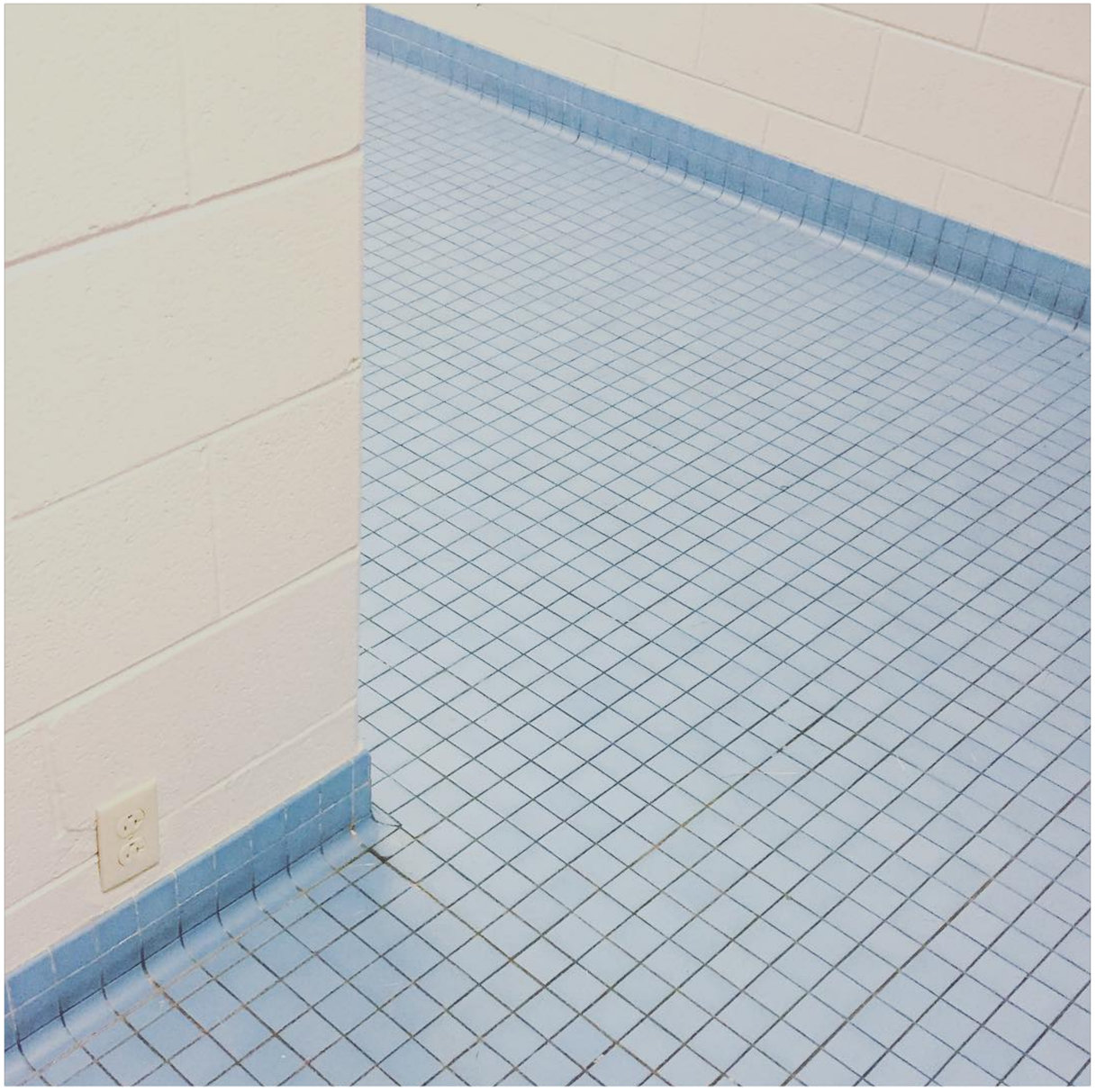
mathematical methods for statistical inference/analysis



inequality in networked labor markets







What is data science?

In other sciences, we have ideas and we conduct experiments.

In data science, with the data from natural experiments all around us,
we often just need to find a way to *see* the things are are right in front of us.

Time to get cracking.

Now

1. Azure Numpy & Pandas tutorial (notebooks.azure.com/ketelsen/libraries/csci3022)
2. Lecture01 notebook (github/notebooks)

Before next class

1. Accept invitation to Piazza (check email)
2. Install anaconda 3.6 (Piazza/Resources/resources)
3. Review & complete Numpy & Pandas tutorial (github/notebooks/)
4. [optional] explore nb1 (github/notebooks)