CSCI 3022

intro to data science with probability & statistics

August 27, 2018

1. What (even) 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**

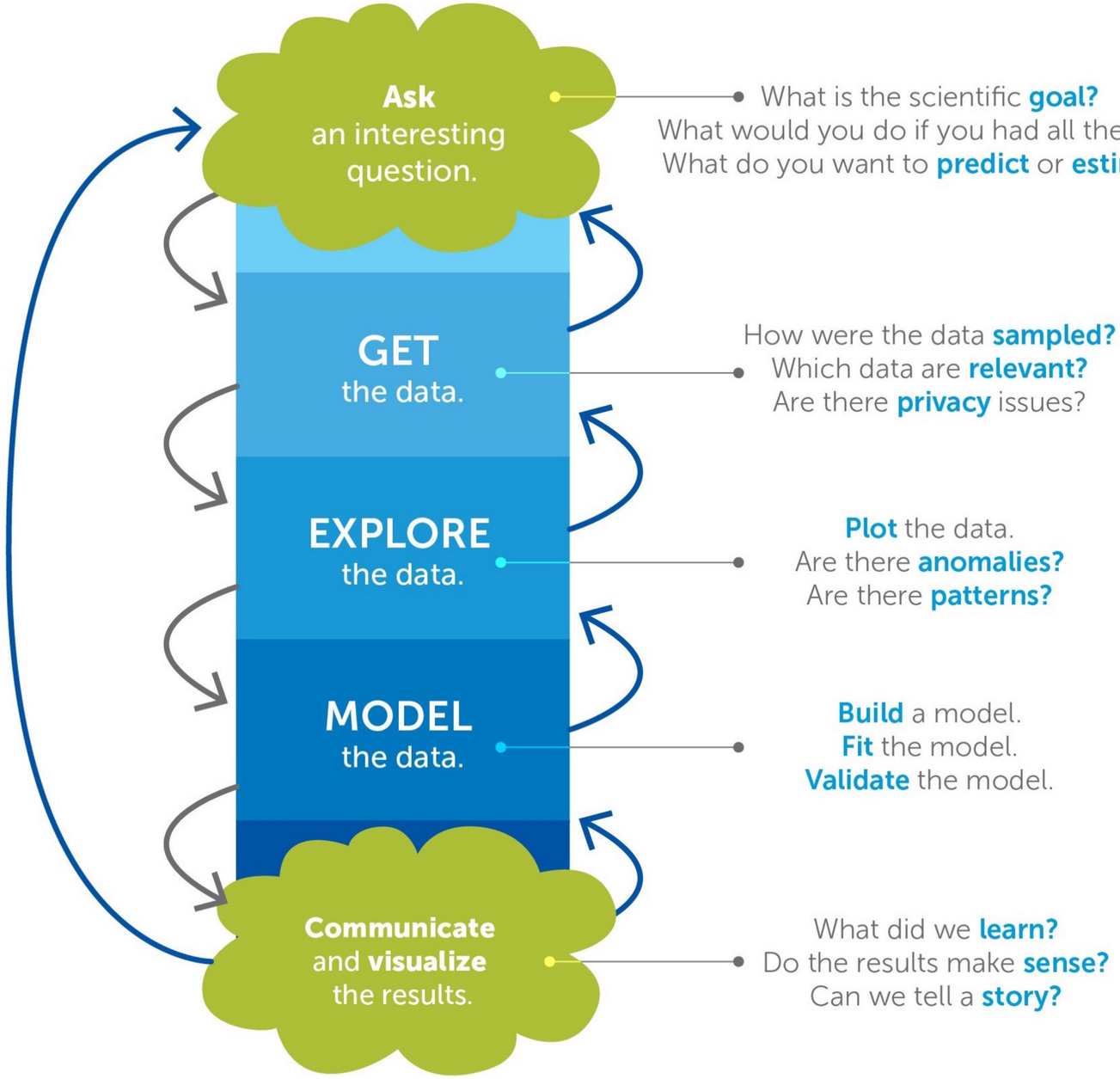
yes: **recovering insights/trends** that are hiding behind data

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

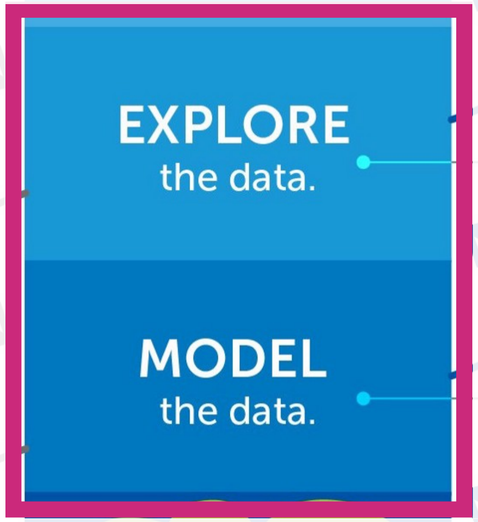
Observation

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	8.27			Course & Computing Introduction		
	8.29		16.1-3	EDA and Summary Statistics		
	1.26		2	Introduction to Probability		
2	9.03			**LABOR DAY - NO CLASS**		
	9.05		15.1-2,16.4	EDA and Data Visualization		hw1 posted
	9.07			**Data Wrangling**		
3	9.10		2,3	**How to Python**		
	9.12		6	Axioms and Theorems of Probability		
	9.14		3	**Stochastic Simulation**		hw1 due
4	9.17		4	Bayes' Rule and Intro to PDFs		hw2 posted
	9.19		4,5	Discrete RVs, PMFs, CMFs		
	9.21			Discrete RVs Strike Back		
5	9.24		5	**Return of the Discrete RVs**		
	9.26			Continuous RVs, Kolmogorov, PDFs, CDFs		
	9.28		7	**The Last Continuous RVs**		hw2 due
6	10.01		7	Expectation		hw3 posted
	10.03			Variance		
	10.05		5.5	**More Expectation & Variance**		
7	10.08			The Normal Distribution		
	10.10		14	**MIDTERM EXAM REVIEW**		
	10.10			The Central Limit Theorems		
	10.12			**MIDTERM EXAM (PM)**		hw3 due
8	10.15		23,24	**The Central Limit Theorem and You**		hw4 posted
	10.17		23,24	Inference and CI Intro		
	10.19			Two-Sample CIs		
9	10.22		25,26	**CIs in the Wild**		
	10.24		25,26	Hypothesis Testing Intro		
	10.26			p-Values		hw4 due
10	10.29		27	**Practical HT & p**		hw5 posted
	10.31			Small-sample HT		
	11.02			TBD		
11	11.05		18,23.3	Bootstrap Intro		
	11.07			**Bootstrap and Small n HT**		
	11.09		27	OLS/SLR Regression		hw5 due
12	11.12			Inference in SLR		hw6 posted
	11.14			*Hands on inference in SLR*		
	11.16			MLR		
13	11.19			**FALL BREAK - NO CLASS**		
	11.21			**FALL BREAK - NO CLASS**		
	11.23			**FALL BREAK - NO CLASS**		
14	11.26		ISL Ch3	Inference in MLR		practicum posted
	11.28		ISL Ch3	More MLR and ANOVA I		
	11.30		ISL Ch3	ANOVA II		hw6 due
15	12.03			**ANOVA + Inference in MLR**		
	12.05			Logistic Regr. & Classification		
	12.07			**Logistic Regr. & Classification**		
16	12.10			Solution Techniques and SGD		
	12.12			**FINAL EXAM REVIEW**		practicum due
X	12.XX			**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/fall2018/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

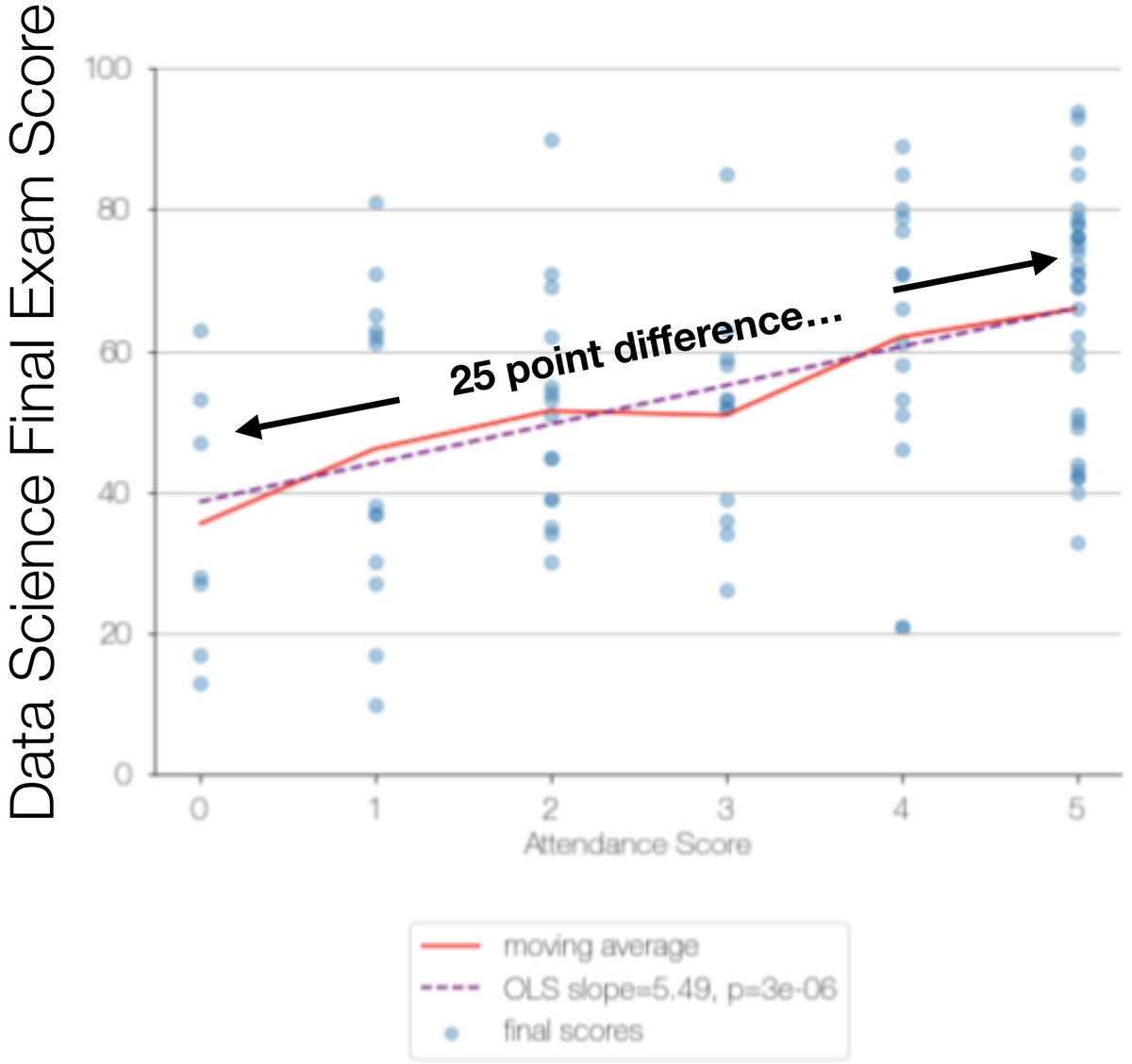
Midterm Exam (20%)

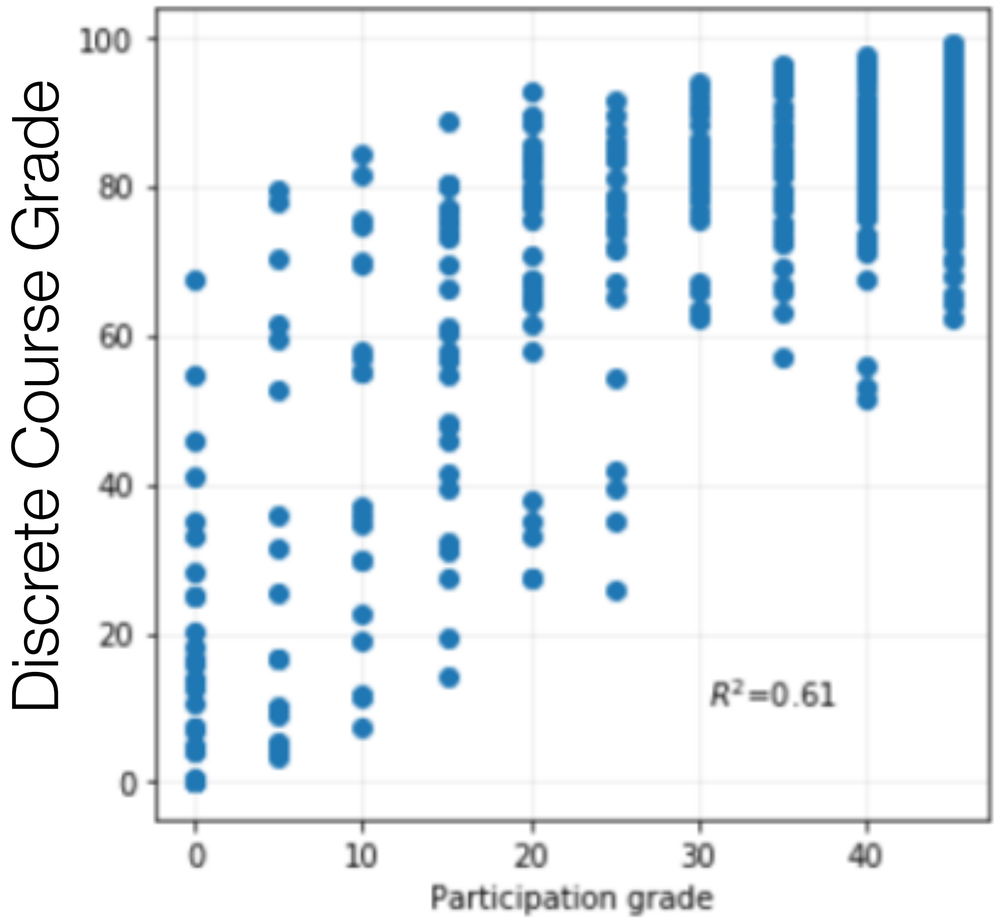
Practicum (15%)

Final Exam (25%)

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

how to pass (any class): show up to class and dig in





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).
- Details on syllabus. [link]

Genius Bar Work Authorization

Repair No: R285992581

Customer Information

United States

Product Information

Warranty Status: Protection Plan (A2)
Model: MACBOOK PRO (RETINA, 15-INCH, MID 2014)
Date of Purchase: 29-Nov-15
Serial No:

Problem Description/Diagnosis

Issue: Customer reports device will sometimes boot to user but most of the time boot and shut down showing the power button logo.
Steps to Reproduce: Verified issue in store. Attempted SMC reset and NVRAM RESET and issue persists. Took machine in RR and performed
down. MagSafe power adapter
hardware issue. Not able to bo
Cosmetic Condition: No notab
Proposed Resolution: Send to
Estimated Turn Around Time:
Mac OS Version: 10.13.x
Hard Drive Size: stock
Memory Size: stock
iLife Version: 13
Contact Apple Support Case:

Employee 291057521

Repair Estimate

Item Number	Description	Price	Amount Due	Customer KBB
S1586LL/A	Labor Charge, PBG4/MBP15"	$ 100.00	$ 0.00	
S5741LL/A	Flat Rate 2 Repair Charge MBP15/MBP17	$ 475.00	$ 0.00	
	Total (Tax not included)	$ 575.00	$ 0.00	

By signing below, I agree that:

- the Repair Terms and Conditions on the reverse side of this page will apply to the service of the product identified above;

- Apple is not responsible for any loss, corruption, or breach of the data on my product during service; and

- as loss of data may occur as a result of the service, it is my responsibility to make a backup copy of my data before bringing my product to Apple for service

Recommendation:

Don't be like this person.

Save your late days for when you need them.

Repair Status

ID: R285992581

Product received by repair center | Repair completed | Product shipped to Apple Store

MacBook Pro

Your product has shipped.

April 3, 2018:
We've shipped your product to the Apple Store where you requested service.

Carrier: FedEx

Save tracking results

Print Help

Actual delivery:

Wed 4/04/2018 10:27 am

Tue 4/03/2018

Houston, TX US

Delivered

Signed for by: A.BRADLEY

LONE TREE, CO US

Request Notifications | Hold at Location | **Obtain Proof of Delivery**

More actions

Travel History

Help Hide

▲ Date/Time	Activity	Location
− 4/04/2018 - Wednesday		
10:27 am	Delivered	LONE TREE, CO
8:12 am	On FedEx vehicle for delivery	ENGLEWOOD, CO

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/fall2018/csci3022/home

about me

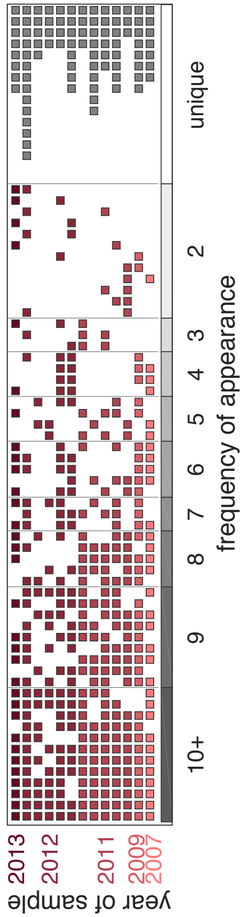
Office Hrs: FLMG 417 | W 4-530 | F 4-5

Assistant Professor, BioFrontiers Institute & Department of Computer Science

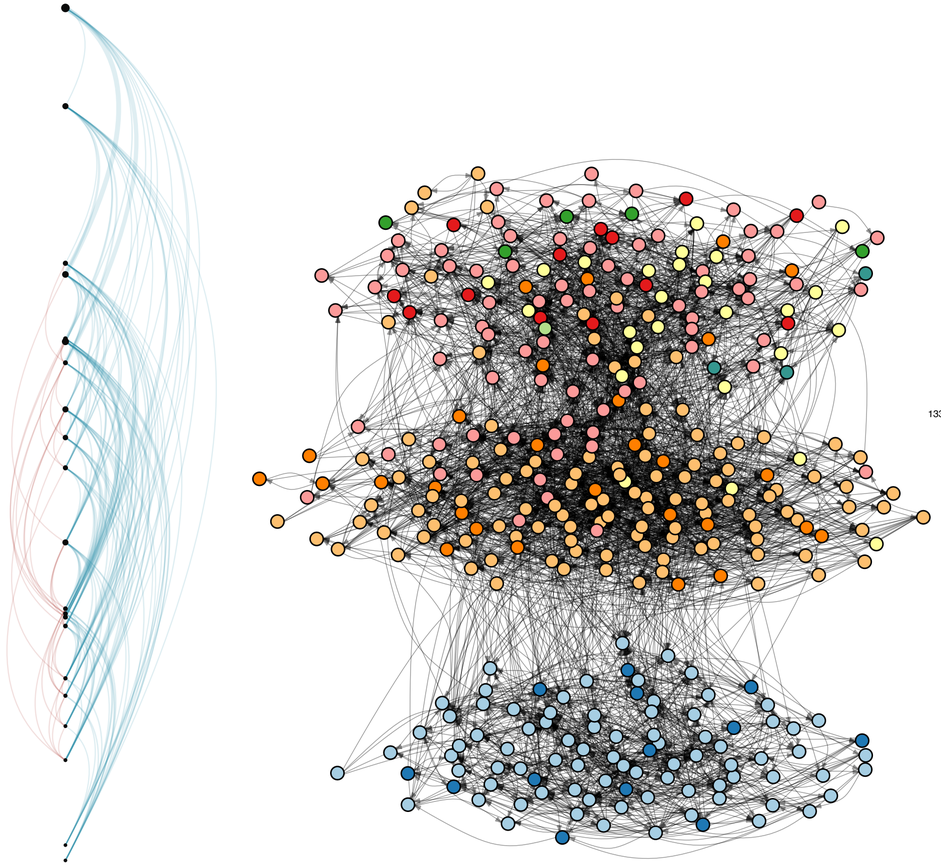
Previously: fellowships at Harvard, Santa Fe Institute; PhD Applied Math

research: danlarremore.com

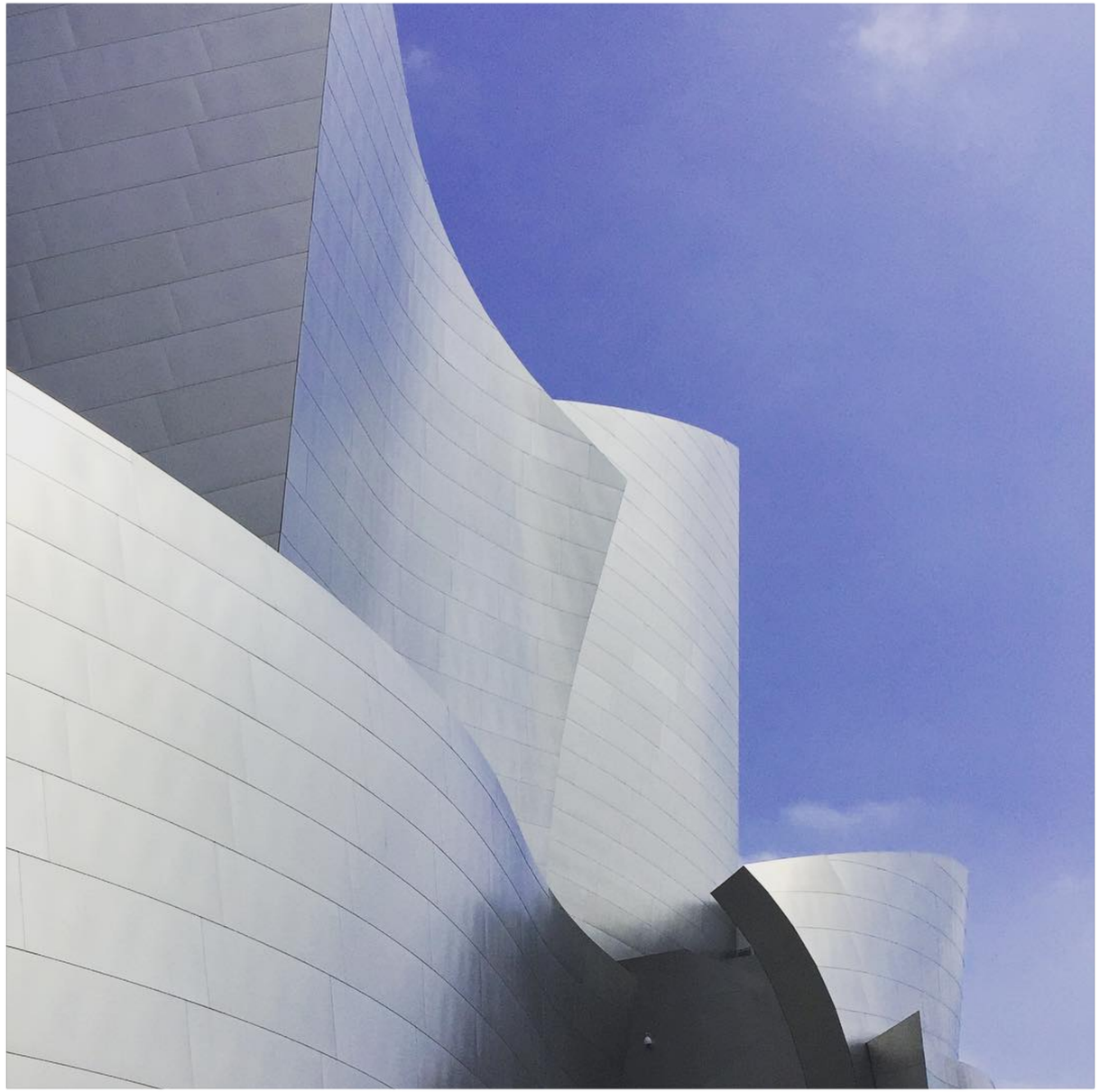
malaria parasite evolution and epidemiology

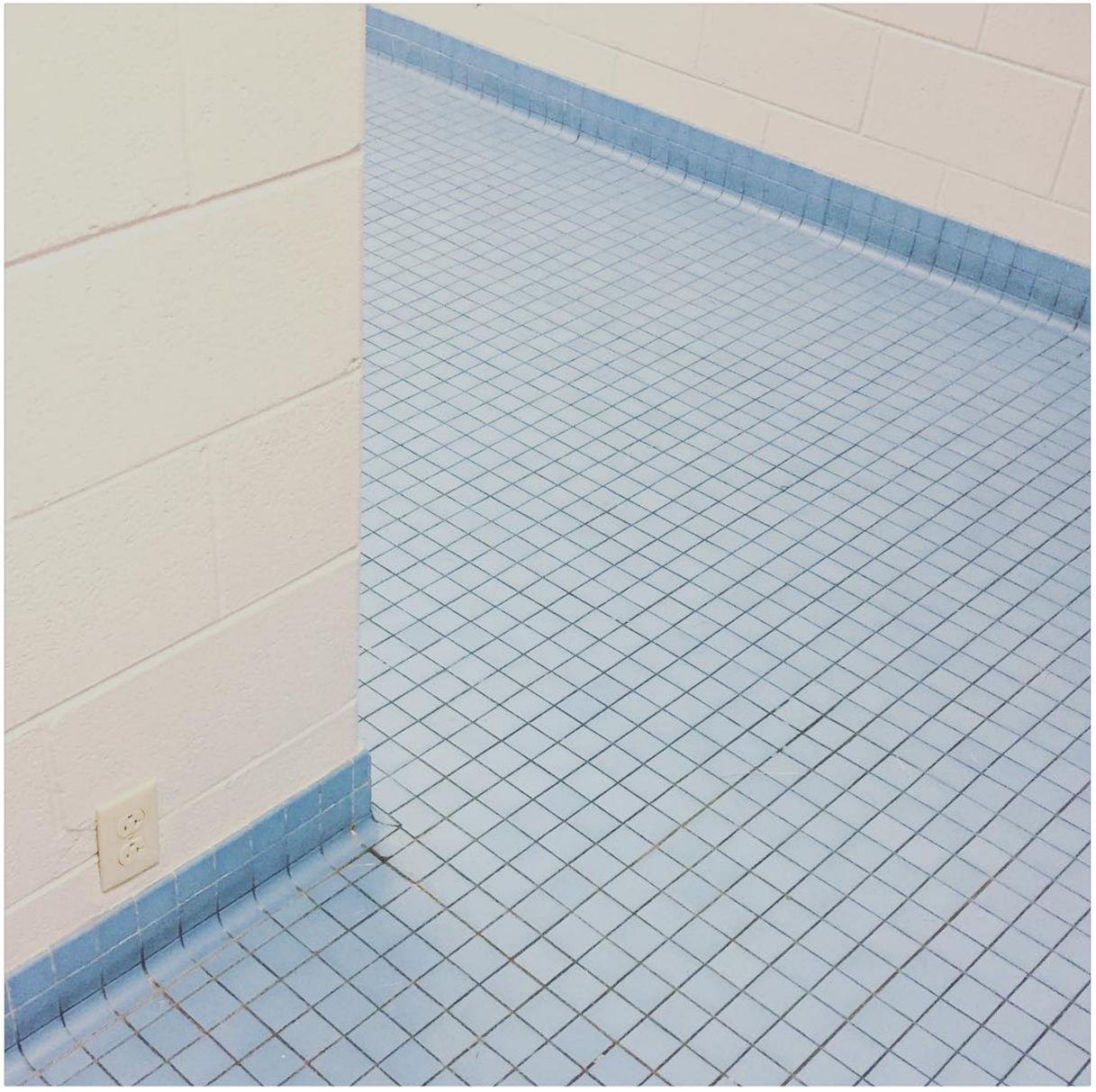


mathematical methods for statistical inference/analysis



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. Numpy & Pandas tutorial (github/notebooks)
2. Lecture01 notebook (github/notebooks)

Before next class

1. Moodle yourself. *csci3022-Dan*
2. Accept invitation to Piazza (check email)
3. Install anaconda 3.6 (Piazza/Resources/resources)
4. Review & complete Numpy & Pandas tutorial (github/notebooks)
5. [optional] explore nb1 (github/notebooks)