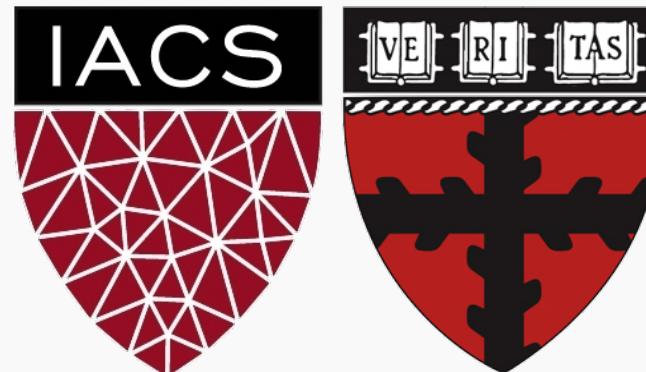


Lecture #1: Introduction to CS109A

aka STAT121A, AC209A, CSCIE-109A

CS109A Introduction to Data Science
Pavlos Protopapas, Kevin Rader and Chris Tanner



Lecture Outline

- Why data science? Why taking CS109A?
- What is data science?
- What is this class and what it is not?
- The data science process
- Example



Job Title, Keywords, or Company

Jobs

Location

Search

50 Best Jobs in America for 2020

Best Jobs

2020

United States

Share



Job Title	Median Base Salary	Job Satisfaction	Job Openings	
#1 Front End Engineer	\$105,240	3.9/5	13,122	View Jobs
#2 Java Developer	\$83,589	3.9/5	16,136	View Jobs
#3 Data Scientist	\$107,801	4.0/5	6,542	View Jobs
#4 Product Manager	\$117,713	3.8/5	12,173	View Jobs
#5 DevOps Engineer	\$107,310	3.9/5	6,603	View Jobs
#6 Data Engineer	\$102,472	3.9/5	6,941	View Jobs
#7 Software Engineer	\$105,563	3.6/5	50,438	View Jobs

Why?

Jobs!

50 Best Jobs in America

This report ranks jobs according to each job's Glassdoor Job Score, determined by combining three factors: number of job openings, salary, and overall job satisfaction rating.

Employers: Want to recruit better in 2017? [Get started](#) [about how.](#)

United States | 2017

12k Shares | [f](#) [t](#) [in](#) [e](#)

1 Data Scientist



4.8 / 5
Job Score

\$110,000
Median Base Salary

4.4 / 5
Job Satisfaction

4,184
Job Openings

[View Jobs](#)

2 DevOps Engineer



A large red arrow points from the top right towards the median base salary figure of \$110,000, which is highlighted with a red circle.



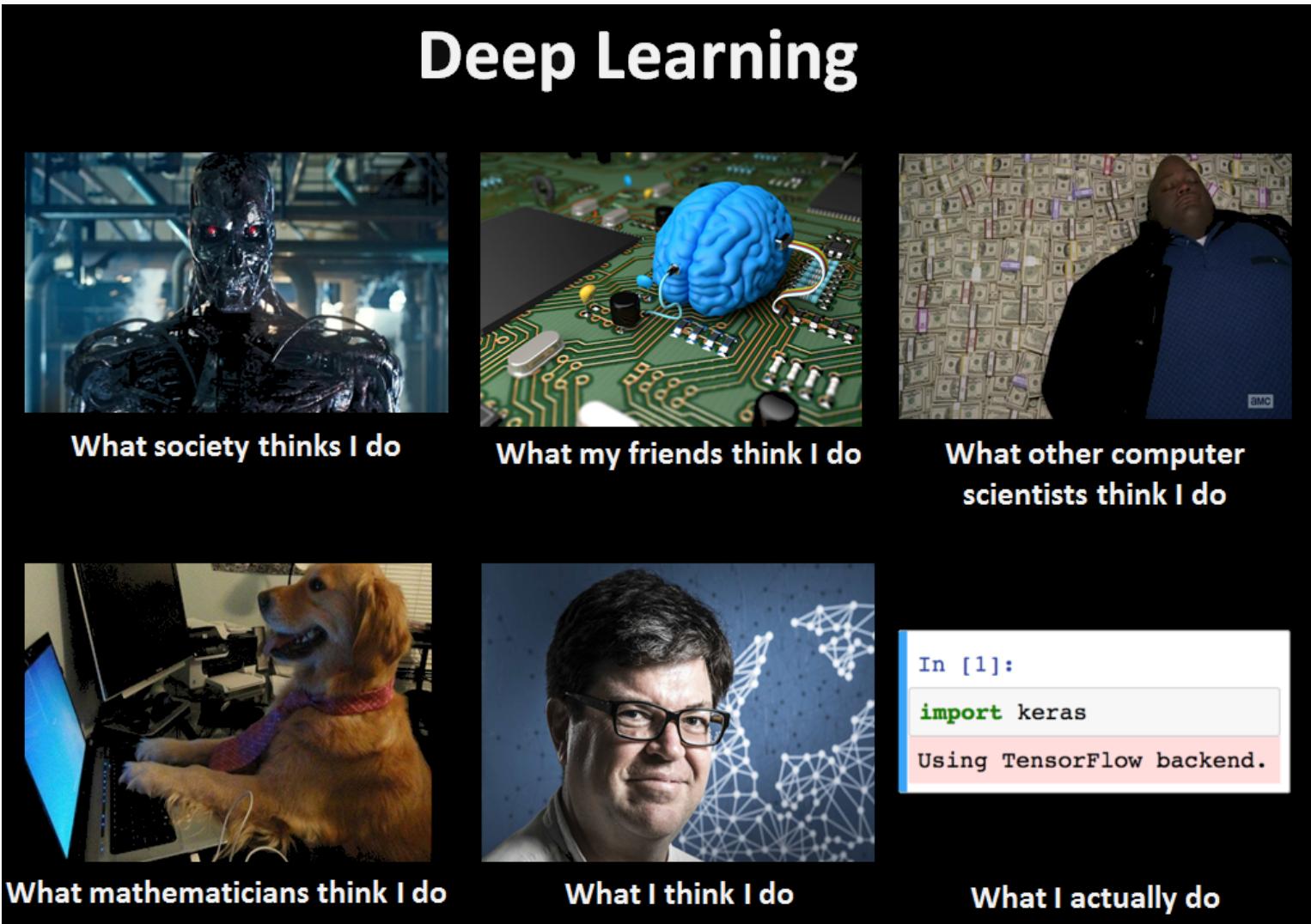
Why?

Why do I love data science?

Why are you here?



Memes !



Why?

Why are you here?



What is data science?



A little bit of history

History

Long time ago (thousands of years) science was only empirical and people counted stars



History (cont)

Long time ago (thousands of years) science was only empirical and people counted stars or crops



History (cont)

Long time ago (thousands of years) science was only empirical and people counted stars or crops and used the data to create machines to describe the phenomena



History (cont)

Few hundred years: theoretical approaches, try to derive equations to describe general phenomena.

$$1. \quad \nabla \cdot \mathbf{D} = \rho_v$$

$$2. \quad \nabla \cdot \mathbf{B} = 0$$

$$3. \quad \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

$$4. \quad \nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} + \mathbf{J}$$

$$T^2 = \frac{4\pi^2}{GM} a^3$$

can be expressed
as simply

$$T^2 = a^3$$

If expressed in the following units:

T Earth years

a Astronomical units AU
($a = 1$ AU for Earth)

M Solar masses M_\odot

$$\text{Then } \frac{4\pi^2}{G} = 1$$

$$H(t)|\psi(t)\rangle = i\hbar \frac{\partial}{\partial t} |\psi(t)\rangle$$

History (cont)

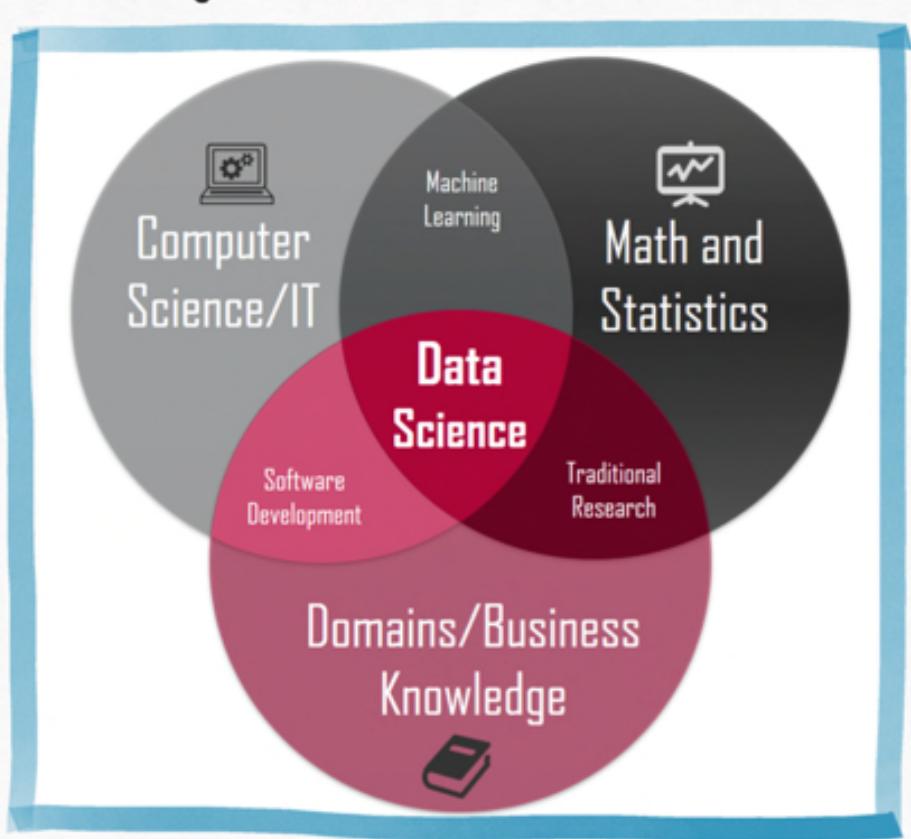
About a hundred years ago: computational approaches



History (cont)

And then data science

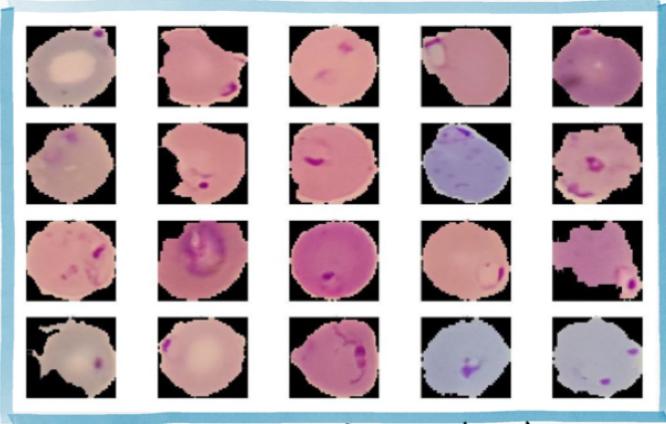
In both data science and machine learning we extract pattern and insights from data.



- Inter-disciplinary
- Data and task focused
- Resource aware
- Adaptable to changes in the environment and needs

The Potential of Data Science

Disease Diagnosis



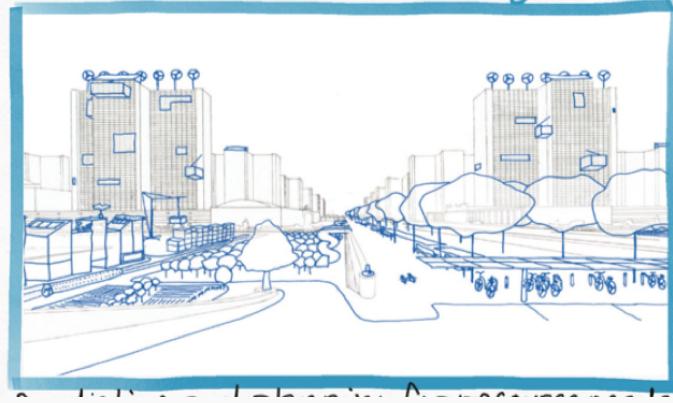
Detecting malaria from blood smears

Drug Discovery



Quickly discovering new drugs for COVID

Urban Planning



Predicting and planning for resource needs
Agriculture



Precision agriculture

The Potential of Data Science

Gender Bias



Some DS models for evaluate job applications show bias in favor of male candidate

Racial Bias



Risk models used in US courts have shown to be biased against non-white defendants

What?

The Data Science Process

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results



What?

The Data Science Process

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

What is the scientific goal?

What would you do if you had **all** of the data?

What do you want to predict or estimate?

What?

The Data Science Process

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

How were the data sampled?

Which data are relevant?

Are there privacy issues?

What?

The Data Science Process

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

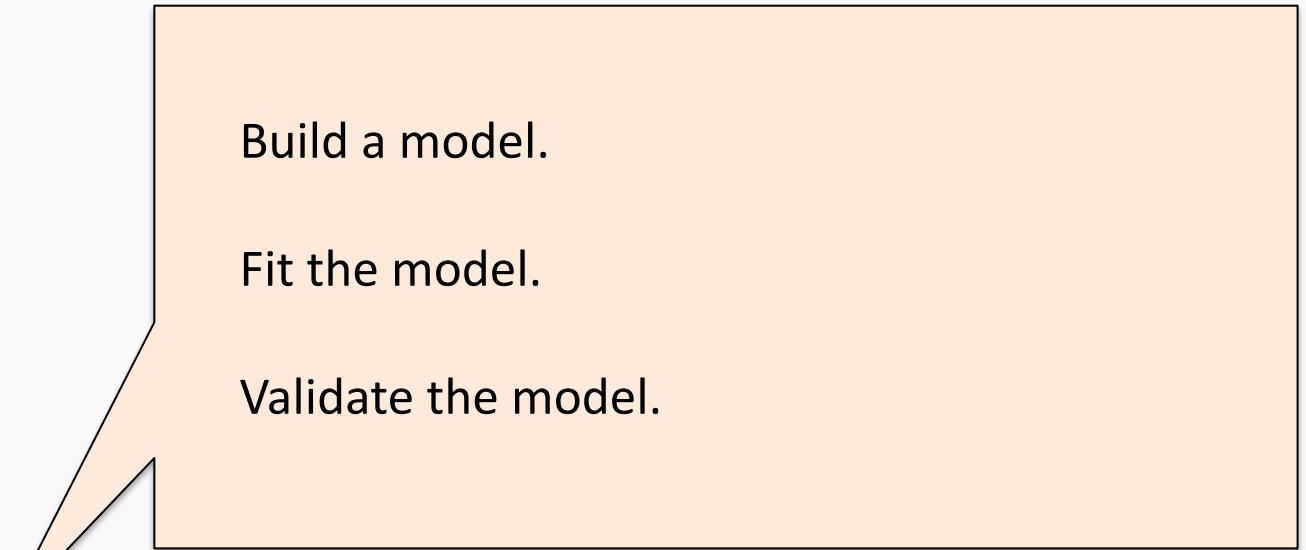
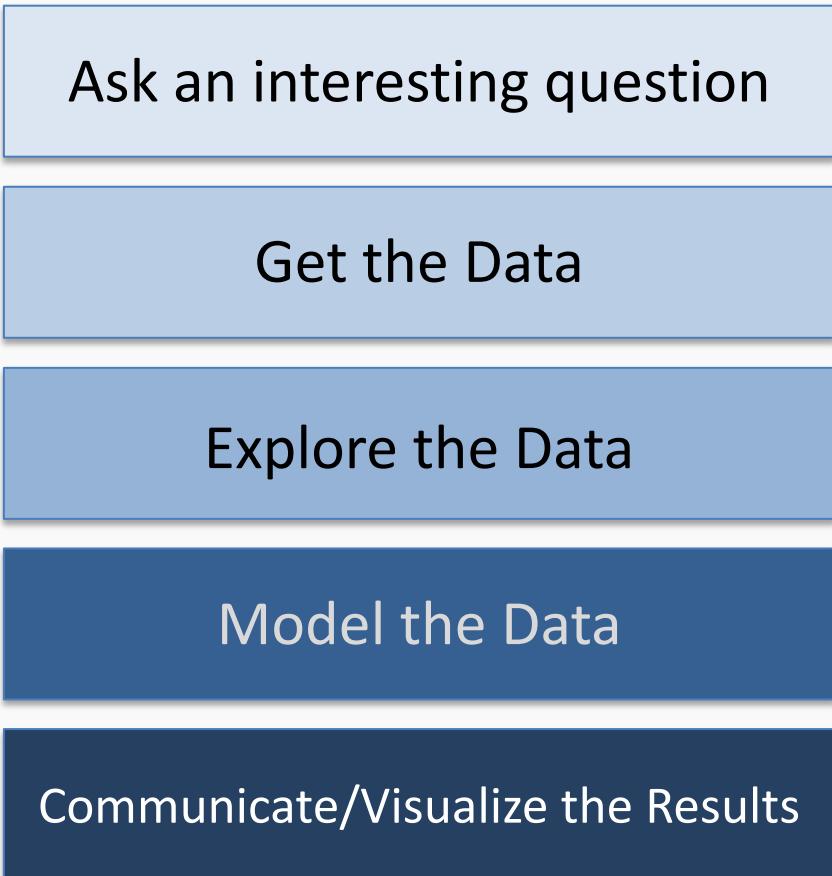
Plot the data.

Are there anomalies or egregious issues?

Are there patterns?

What?

The Data Science Process



What?

The Data Science Process

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

What did we learn?

Do the results make sense?

Can we effectively tell a story?

What?

The material of the course will integrate the five key facets of an investigation using data:

1. data collection; data wrangling, cleaning, and sampling to get a suitable data set.
2. data management; accessing data quickly and reliably.
3. exploratory data analysis; generating hypotheses and building intuition.
4. prediction or statistical learning.
5. communication; summarizing results through visualization, stories, and interpretable summaries.

Goal of the course

Theory

1. Key Machine Learning concept
2. Important metrics for evaluation
3. Handling different kinds of data
4. Extracting insights from analysis of the models

Practice

1. Implement ML and deep learning models using python libraries
2. Using free online tools and resources for data science

Impact

1. Solving real-life problems using DS
2. Evaluating the social impact of DS

Weeks 1-2: Data

Data Formats + Web Scraping
Pandas

Weeks 3-5: Regression

kNN Regression
Linear Regression
Multi and Poly Regression
Model Selection and Cross Validations
Inference
Bootstrap
Ridge and Lasso Regularization

Weeks 6-7: Classification

Weeks 8: Data

Data Imputation
PCA

Weeks 9-10: Trees

Decision Trees
Bagging
Random Forest
Boosting Methods

Weeks 11-12: Neural Networks

Multi-Layer Perceptron
Architecture of NN
Fitting NN, backprop and SGD
Regularization of NN

Weeks 13-14

Ethics
Model Interpretation



After CS109A

CS109B

A. Neural Networks:

- CNNs
- RNNs
- Generative models

B. Unsupervised Clustering

C. Piecewise Linear Regression

D. Bayesian Modeling

AC295

A. Production Data Science, from notebooks to the cloud

B. Big models, transfer learning and architecture learning

C. Visualization tools for interpreting models



Not an exclusive list

- CS171 (Visualization)
- CS182 (AI)
- CS181 (ML)
- CS 187 (NLP)
- Stat 110 (Probability)
- Stat 111 (Inference)
- Stat 139 (Linear Models)
- Stat 149 (generalized linear models)
- Stat 131 (Time Series)
- Stat 171 (Stochastic Processes)
- Stat 195 (Statistical Machine Learning).



Who? Instructors



Pavlos Protopapas

Scientific director
Institute of Applied
Computational Science



Kevin Rader

Senior preceptor in
Statistics



Chris Tanner

Lecturer at Institute of
Applied Computational
Science

Who?



Eleni Kaxiras

Supportive Instructor

Assistant Director for
Data Science and
Computation at SEAS



Marios Mattheakis

Section Leader

Post-doctoral Fellow
IACS



Chris Gumb

Head TF

Graduate student of Data
Science at Harvard
Extension School

Who? Teaching Fellows



Course Components



Lectures, Advanced Sections, Sections and Office Hours

During lecture will cover the material which you will need to complete the homework, and to survive the rest of your life in CS109A.

We will use a mix of notes and exercises via edstem.

1. Lecture notes and associated notebooks will be posted before lecture on GitHub and on edstem.
2. Lectures will be video taped (and live streamed) and posted approximately within 24 hours on web page.

We will have two ‘shows’ , morning and matinee (A, B)

A: Morning Mon/Wed/Fri 9:00-10:15am @Zoom

B: Matinee Mon/Wed/Fri 3:00-4:15pm @Zoom



Lecture format

ASYNCHRONOUS

- Quiz
- Finish exercises from previous lecture
- Reading or Video Watching for next lecture

SYNCHRONOUS

The “hands-on exercises” part will be longer during the Friday lecture as opposed to Mon/Wed.

Questions from asynchronous material, review of quiz and homework

Live Lecture

Q&A

Hands-on exercises in breakout rooms

Discussion about the exercises

Repeat

⋮

Summary and conclusions



Advanced Sections, Sections and Office Hours

Lectures are supplemented by 1.25 hours sections led by teaching fellows.

There are [two types](#) of sections:

- Standard Sections will be a mix of review of material and practice problems similar to the homework.

[Friday 1:30-2:45 pm, and Mon 8:30-9:45 pm @zoom](#)

- Advanced Sections (**A-Sections**) will cover advanced topics like the mathematical underpinnings of the methods seen in lectures and labs.

[Weds 12:01 pm - 1:15 pm @zoom.](#) A-sections are required for AC209 students.

Note: Sections are not held every week. Consult the course calendar for exact dates.



Advanced Sections topics

Topics

1. Linear Algebra and Hypothesis Testing: The Short Versions
2. Methods of regularization and their justifications
3. Generalized Linear Models
4. Mathematics of PCA
5. Ensemble methods
6. Stochastic Gradient Descent and solvers

NOTE 1: The materials in the Advanced Sections are required for all AC 209A students. There will be one extra question in most homework for AC 209 students which will be based on the A-Section materials.

NOTE 2: No additional quizzes for A-section.

NOTE 3: A-sections and Friday's regular section will be live streamed to everyone.



Office Hours

FALL 2019 CS109A WEEKLY SCHEDULE							
	OH ON CAMPUS IACS Lobby	OH ONLINE zoom meeting room	B-SECTION	A-SECTION	LAB Pierce 301	LECTURE NW B-103	
	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SATURDAY	SUNDAY
9 - 9:30 AM							
9:30 - 10 AM							
10 - 10:30 AM							
10:30 - 11 AM							
11 - 11:30 AM							
11:30 AM - 12 PM							
12:00 - 12:30 PM							
12:30 - 1 PM							
1 - 1:30 PM							
1:30 - 2 PM		LECTURE DCE live streamed					
2 - 2:30 PM							
2:30 - 3 PM							
3 - 3:30 PM	3-5 OH Kevin & Pavlos						
3:30 - 4 PM							
4 - 4:30 PM							
4:30 - 5 PM			8-section 4:30-5:45 Science Center Room 110		4:30-5 OH Brandon	4:30-5 OH Chris T MO B120	
5 - 5:30 PM						4 - 5:15 A-Section Video recorded 1 Story St, Room 305	
5:30 - 6 PM							
6 - 6:30 PM	6-7:30 OH			5:30-7 OH	5:30 - 7:00 OH	4:30-5:45 DCE live streamed TFa: TBD	
6:30 - 7 PM							
7 - 7:30 PM							
7:30 - 8 PM							
8 - 8:30 PM							
8:30 - 9 PM							
9 - 9:30 PM							
9:30 - 10 PM							
							7:30-8 TBD



Assignments



Four Graded Components

Homework: 63%

Homework zero: 1%
Individual Homework (2): 10%
Paired Homework (6): 42%

HW4 and HW7 are the indiv. HW

Exercises: 6%

During lecture.

All questions are weighted equally.

Due at the beginning of the next morning lecture.

Quizzes: 6%

End of each lecture.

25% of the quizzes will be dropped from your grade.
All questions are weighted equally.

Due at the beginning of the next morning lecture.

Projects: 25%

Three milestones plus final presentation and a report in the form of a blog.
More details soon.



Homework(s)

There will be 8 homework (not including Homework 0):

- Homework 0 (due Sept 9)
- Homework 1: Web scraping, Beautiful Soup
- Homework 2: Regression kNN and LinReg
- Homework 3: Multi-regression, polynomial reg and model selection
- **Homework 4*: Log Reg**
- Homework 5: PCA and ethics
- Homework 6: Random Forest, Boosting and Neural Networks
- **Homework 7*: Neural Networks**
- Homework 8: Ethics and model interpretation



Homework(s)

You are encouraged but not required to submit in pairs, except homework 4 and homework 7, which you must work individually.

We will be using the Groups function in Canvas to do this, details to be announced later.

All homework are **due 11:59 pm Wednesdays**, and homework will be released on Wednesdays.



Final Project

There will be a final group project (2-4 students) due during exams period.

- We will provide 7 pre-defined projects which you could use for your final project.
- In some very special cases you can use your own (public) data set and your own project definition (to be approved by the instructors)
- Project topics will be announced September 10th.

Help

The process to get help is:

1. Post the question in Ed, and hopefully, your peers will answer. We monitor the posts, and we will respond within 8 hours from the posting time.
2. Attend the Office Hours; this is the best way to get help.
3. For private matters, send an email to the Helpline: cs109a2020@gmail.com. All the instructors and TFs monitor the Helpline.
4. For personal matters, send an email to Pavlos, Kevin, and Chris.

Sundays will be slow days, so please be patient!

Tools for the course

Web page

CS109A: Introduction to Data Science

Fall 2020
Pavlos Protopapas, Kevin A. Rader, and Chris Tanner

Additional Instructor: Elisa Keefer

Welcome to CS109a/STAT121a/AC109a, also offered by the DCE as CS109, Introduction to Data Science. This course is the first half of a one-year course in data science. We will focus on the analysis of data to perform predictions using statistical and machine learning methods. Topics include data scraping, data management, data visualization, regression and classification methods, and deep neural networks. You will get ample practice through weekly homework assignments. The class material integrates the five key facets of an investigation using data:

1. data collection - data scraping, cleaning, and sampling to get a suitable data set
2. data management - accessing data quickly and reliably
3. exploratory data analysis - generating hypotheses and building intuition
4. prediction or statistical learning
5. communication - summarizing results through visualization, stories, and interpretive summaries

Only one of CS 109a, AC 209a, or Stat 121a can be taken for credit. Students who have previously taken CS 109, AC 209, or Stat 121 cannot take CS 109a, AC 209a, or Stat 121a for credit.

Helpdesk: cs109a2020@gmail.com

Lectures: Wed, Fri 9:30-10:15 am & 1:30-2:14 pm (identical material in a single slot)
Section: Fri 1:30-2:45 pm (identical material) [starts 9/11]
Adapted Sections: Wed 1:30pm (separate) [starts 9/11]
Office Hours: TBA

Course materials can be viewed in the public [GitHub repository](#).

edstem

CS109a - Discussion

Forum

Answer

Pavlos Protopapas posted a question: Wednesday Edition is the first lecture. Friday

Wednesday Edition is the first lecture. Friday

Pavlos Protopapas posted a response: Multivariate Calculus

What will be the scope of this course?

Wednesday Edition is the first lecture. Friday

Pavlos Protopapas posted a response: Multivariate Calculus

What will be the scope of this course?

Wednesday Edition is the first lecture. Friday

Pavlos Protopapas posted a response: Multivariate Calculus

What will be the scope of this course?

Wednesday Edition is the first lecture. Friday

Canvas

COMPSCI 109A - Syllabus

Home | Announcements | Grades | People | Pages | Files | Academic Integrity Policy | Support Resources | Manage Course | Modules | Discussions | Collaborations | Outcomes | Rubrics | Settings

Our [Public Course Page](#) is the primary source for course info and materials. There you will find the [syllabus](#), [FAQ](#), [announcements](#), etc., and other resources. Course Helpdesk: [cs109a2020@gmail.com](#)

Prerequisites: In order to get the most out of CS109A, knowledge of multivariate calculus, probability theory, statistics, and some basic linear algebra (e.g., matrix operations, eigenvalues, etc.) is suggested but not required. Below are some resources for self-assessment and review:

- Multivariate Calculus: [multiple exams w/ solutions](#)
- Linear Algebra: [multiple exams w/ solutions](#)
- Probability: [exams w/ solutions](#)
- Statistics: [multiple pairs of exams w/ solutions](#)

Here is a useful notebook for reviewing many of the above topics: [Mathematics for Machine Learning](#)

Note: you can be successful in the course (assignments, quizzes, etc.) with the listed pre-requisites, but some of the material presented in lecture may be more easily understood with more background.

Harvard Extension School Policies

The Extension School is committed to providing an accessible academic community. The Accessibility Office offers a variety of accommodations and services to students with documented disabilities. Please visit [https://diversity.extension.harvard.edu/resource-centers/resources/accessibility-services/accessibility](#) for more information.

You are responsible for understanding Harvard Extension School policies on academic integrity: [https://diversity.extension.harvard.edu/resource-centers/student-conduct/academic-integrity/](#) and how to use sources responsibly. Not knowing the rules, misunderstanding the rules, running out of time, submitting the wrong draft, or being overwhelmed with multiple demands are not acceptable excuses. There are no excuses for failure to uphold academic integrity. To support your learning about academic citation rules, please visit the Harvard Extension School Tips to Avoid Plagiarism: [https://diversity.extension.harvard.edu/resource-centers/academic-integrity/tips-against-plagiarism/](#). When you'll find links to the Harvard Guide for citing sources and two free online 15-minute tutorials to test your knowledge of academic citation policy. The tutorials are anonymous open-learning tools.

- Syllabus
- Calendar
- Link to materials

- Forum
- Quizzes
- Reading assignments
- Hands on exercises
- Links to lectures

- Homework
- Grades



Misc

Video release form:

<https://canvas.harvard.edu/courses/74056/quizzes/182045>

Backup plan if zoom dies:

We will do lecture the next lecture (A or B). If it's still not up by then, we will record and upload to canvas.

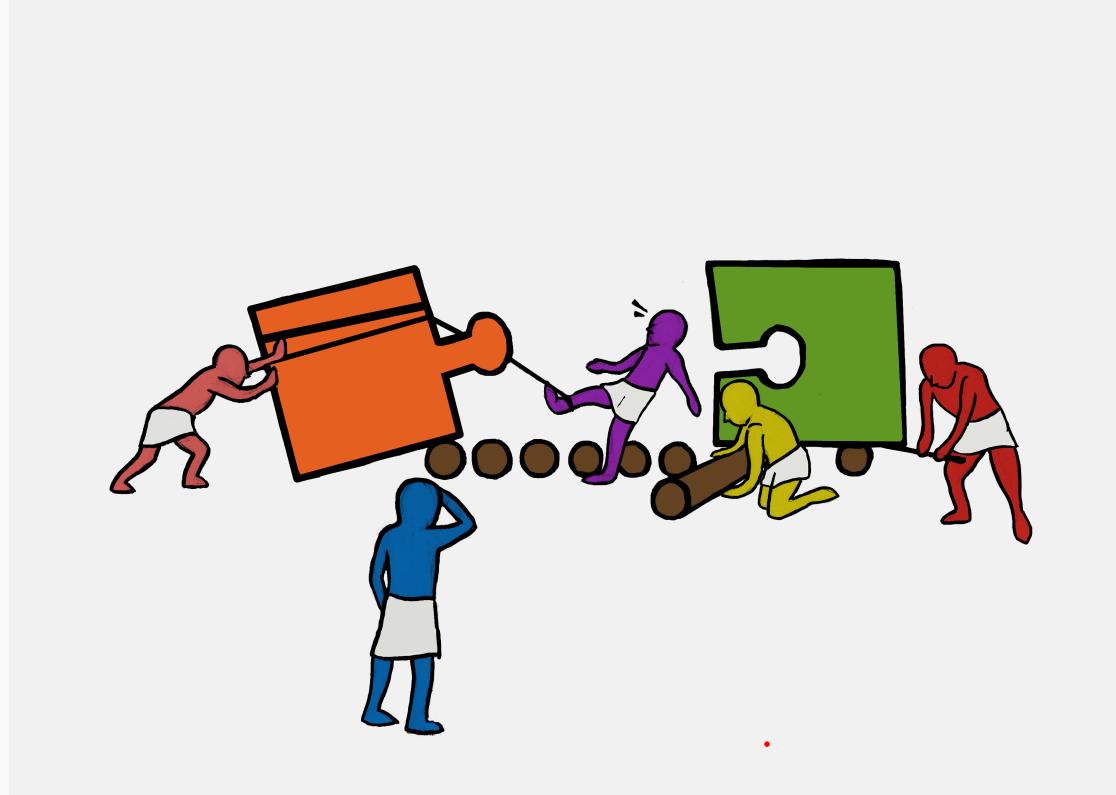
Forming groups:

<https://docs.google.com/forms/d/e/1FAIpQLSfIsQyxdCwUCbJmyWyotp30anrKsuGHfIHt-DQEKMnK8iE4TA/viewform>

Study breaks:

Thurs 9/10 @8:30pm & Fri 9/11 @10:15am





Breakout rooms and in-class exercises

Some students are having issues with Safari.

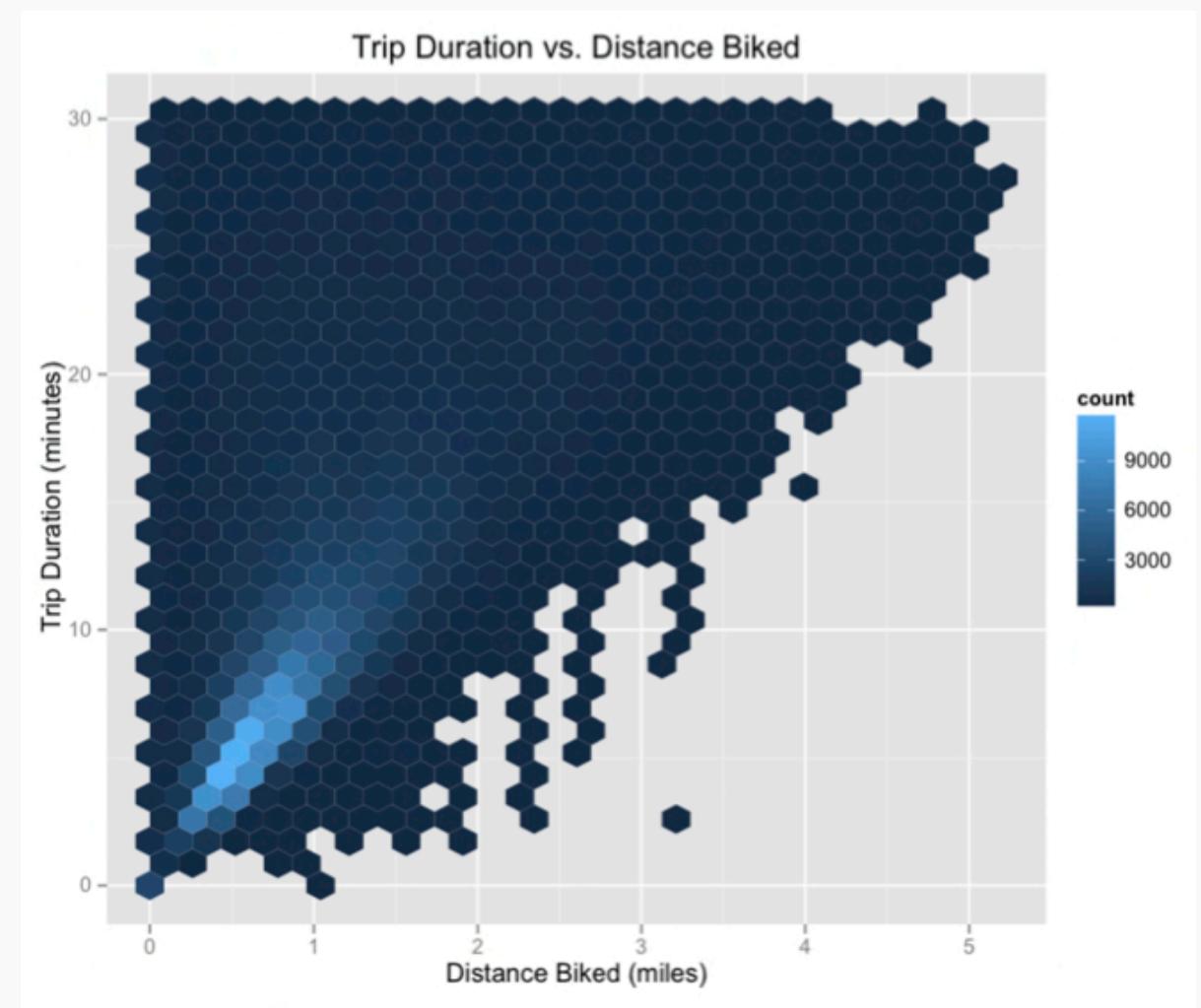


Inspirations for Data Viz/Exploration

So how well did we do in formulating creative hypotheses and manipulating the data for answers?

Check out the winners of the Hubway Challenge:

<http://hubwaydatachallenge.org>



Statistics. Math. Computer Science. Physics. Long ago, the four disciplines lived together in harmony. Then, everything changed when the Computer Science attacked. Only a master of all four elements, could stop them, but when the world needed it most, it was not invented. A few years ago the world discovered the new master, a scientist called data scientist, a master of all four elements

