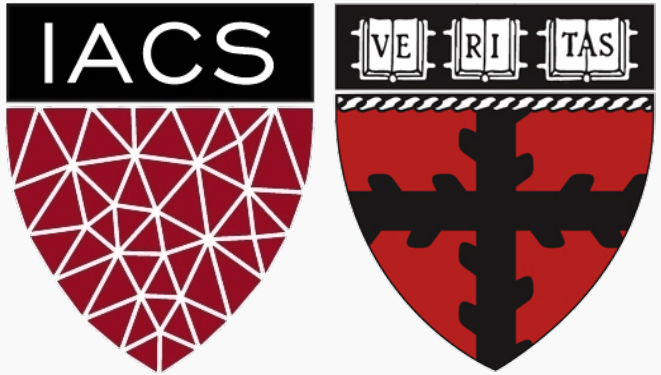
Lecture 1: Review of 109A Preview of 109B

CS109B Introduction to Data Science

Pavlos Protopapas and Mark Glickman



Outline

- Who
- What have we learned in 109a?
- What is covered in 109b
- Course Logistics

Outline

- **Who**
- What have we learned in 109a?
- What is covered in 109b
- Course Logistics

Who: Instructors

Mark Glickman: Senior Lecturer in Statistics

Who: Instructors (cont)

About Mark Glickman:

- BA in Statistics from Princeton; PhD in Statistics from Harvard
- Chess master, inventor of Glicko and Glicko-2 rating systems for head-to-head competition, ratings committee chair of US Chess
- Former Editor-in-Chief of the Journal of Quantitative Analysis in Sports (2015-2017)
- Director of the Harvard Sports Analytics Laboratory
- Senior Statistician at the Center for Healthcare Organization and Implementation Research, a Veterans Administration Center of Innovation
- Fellow of the American Statistical Association
- Board of Directors member of the American Statistical Association (ASA); Co-Chair of the Committee on Data Science of the ASA.

Who: Instructors (cont)

Pavlos Protopapas: Scientific Director of the Institute for Applied Computational Science (IACS)

About Pavlos Protopapas

- BSc in Physics, Imperial College, PhD in Theoretical Physics, UPENN
- Teaches CS109 and the IACS Capstone Course
- Active member of the astrostatistics community. Research at the intersection of astronomy, machine learning and statistics
- Member of Alerce, an intelligent broker for online annotating celestial objects from streaming data
- Loves classical music, hiking and anything adventurous

Who: Lab Instructors

- **Rahul Dave**

  Lecturer at IACS. PhD in cosmology and teaches AM207. He loves climbing, hiking and he is also known as the human Google.

  Lab: AWS and scaling up your calculations (Lab 4)

- **Eleni Kaxiras**

  Eleni has been the CS109a/b Head TF for 3 years. She is also a staff member at SEAS, advising courses in the use of computation for teaching and learning. She holds a Bachelor's in Physics and she produces her own olive oil.

  Labs: NN optimization (Lab 3) and CNNs (Lab5)

  Head TF

Who: Lab Instructors

- **Will Claybaugh**

  IACS Master's student, former social network analyst at Booz Allen Hamilton. Former fencer, built and flew on a cluster of 18 weather balloons,

  Labs: Setting up environments (Lab 1), Smoothing/GAM (Lab2), Clustering Lab 7), Bayes 2 (Lab 9)

- **Srivatsan Srinivasan**

  IACS Masters Student, Former summer data science intern at Facebook. Incoming Research Engineer at DeepMind. Enjoy occasional creative writing and theater.

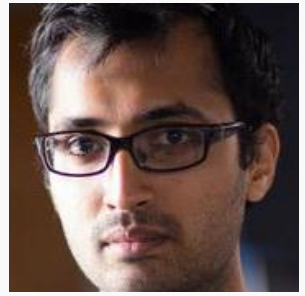
  Labs: RNNs (Lab 6) and GANS (Lab 11)

  Advanced Sections: Deep RL (a-sec 6), Variational Inference (a-sec 7)

Who: Lab Instructors

- **Vivek Hv**

  Vivek is a graduate student in the Design Engineering program. He has a background in product development, healthcare, and computer science. After his undergraduate studies in Aerospace Engineering, he joined Honeywell, where he worked on rapid prototyping and development of products for private jets. Beyond this, Vivek enjoys art, cats, soccer, waffles, programming, and trekking.

  Labs: Bayes (Lab 8), Autoencoders and variational autoencoders (Lab 10).

  Advanced Sections: GANS (A-sec 8)

Who: Advanced Section Leaders

- **Javier Zazo**

Postdoc at SEAS. Works in optimal transport and neural signal processing. Comes from Madrid. Loves going to the mountains and good weather, and being outdoors. Many hobbies, from playing Go, watching movies, and hanging out. Hates cooking. Survives on minimal effort cooked foodstuff. But still loves delicious food.

Advanced Sections: Optimization (a-sec 1), Dropout (a-sec 2), Advanced CNNs (a-sec 3), NN transfer learning (a-sec 5)

- **Marios Matthaiakis**

He is a postdoctoral fellow at IACS, computational physicist and trying to apply physical laws in Neural Network architectures.
I came from Crete, a beautiful island in Greece.

Advanced Section: LSTN, GRU in NLP + (a-sec 4)

Who: Teaching Fellows

- **Sol Girouard**

  Reaching Fellow for 109a/b, while a Top of Class and Award Wining Student graduating as part of Harvard Class of 2018. She is a Quant, Mathematical Economist and Data Scientist who channels her applied interdisciplinary background in the intersection of financial markets and technology. Sol is training for her 2nd degree black belt in full contact Tae KwonDo.

- **Brandon Walker**

  Principal data scientist for LexisNexis Risk Solutions Healthcare Analytics Group. He has TF'ed CS109a twice.

- **Yujiao Chen**

  Ph.D student at GSD. She loves TF'ing 109.

Who: Teaching Fellows

Rashmi Banthia

She has been TF for long time for CS109A/B. Interests - Indian food and latest - Orangetheory (doesn't mean I'm good at it)

Evan Mackay

Harvard College

Evan is from Florida and enjoys biking, podcasts, and sweet potatoes

Alex Lin

Harvard College

Alex enjoys working with Python(s)

Who: Teaching Fellows

Curtis Hsu

Curtis Hsu is a Senior at Harvard College living in Mather House studying statistics and computer science. He enjoys hip hop dancing in his free time!

Anirudh (Ani) Suresh

Ani (’20) is a Harvard undergraduate concentrating in Math & CS.

Outline

- Who
- **What have we learned in 109a?**
- What is covered in 109b
- Course Logistics

109A

- Scraping, skLearn, numpy, Pandas, matplotlib
- Visualization best practices
- Linear, multiple and polynomial regression
- Model Selection and regularization
- Logistic Regression, multiple and polynomial.
- kNN classification
- Decision Trees, RF, Boosting, Stacking
- SVM
- AB testing and experimental design

Outline

- Who
- What have we learned in 109a?
- **What is covered in 109b**
- Course Logistics

Topics

The semester is divided into 2 parts.

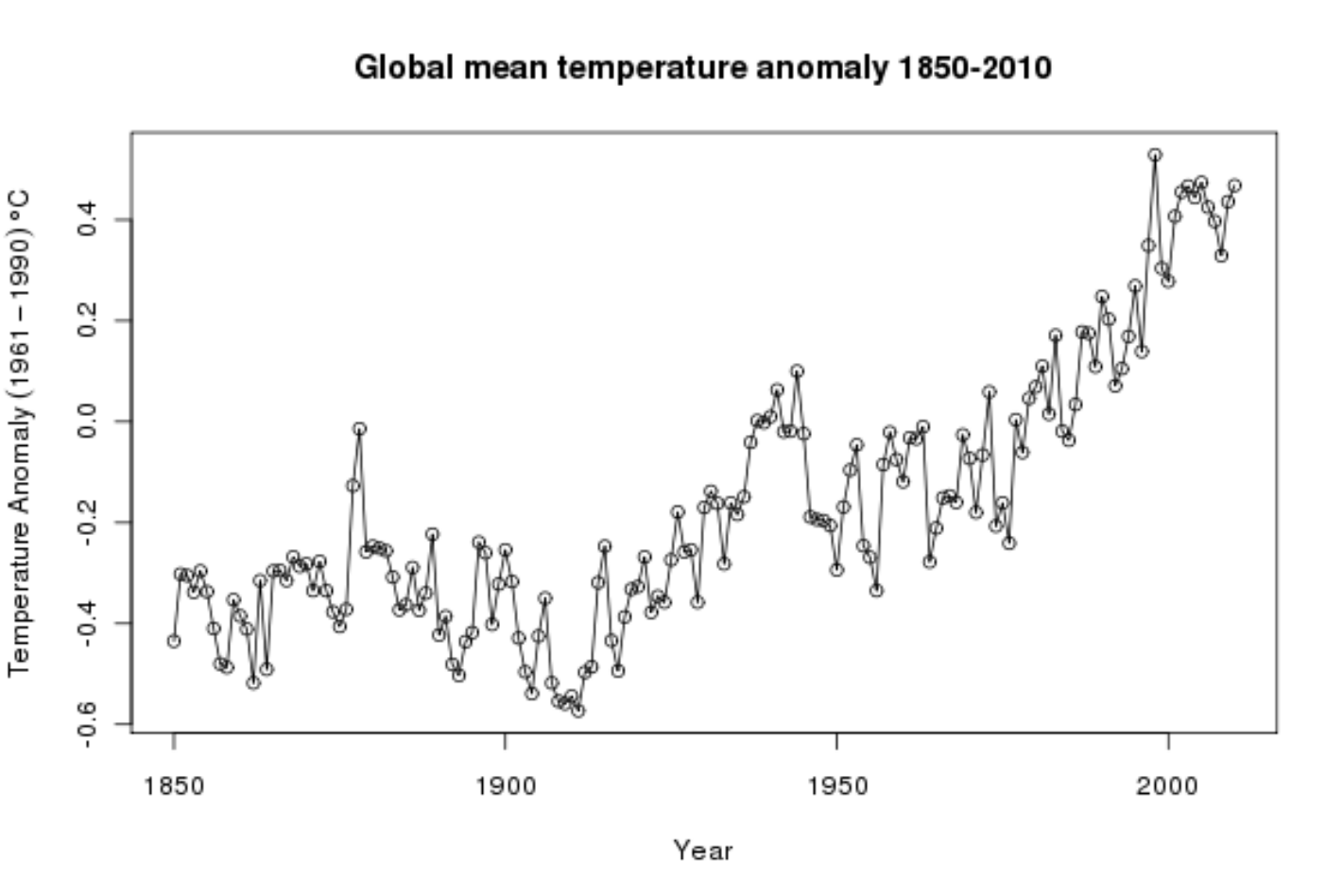
- **Part 1:** Smoothing, Unsupervised Learning and Bayesian inference all in python, Neural Networks in python and Keras.
- **Modules:** Incorporates everything from 109a and 109b into modules.

Course topics covered by Glickman

- Regression splines, smoothers, additive and generalized additive models
- Unsupervised learning and cluster analysis
- Introduction to Bayesian methods
  - ➢ Hierarchical modeling
  - ➢ Latent Dirichlet Allocation (topic modeling)

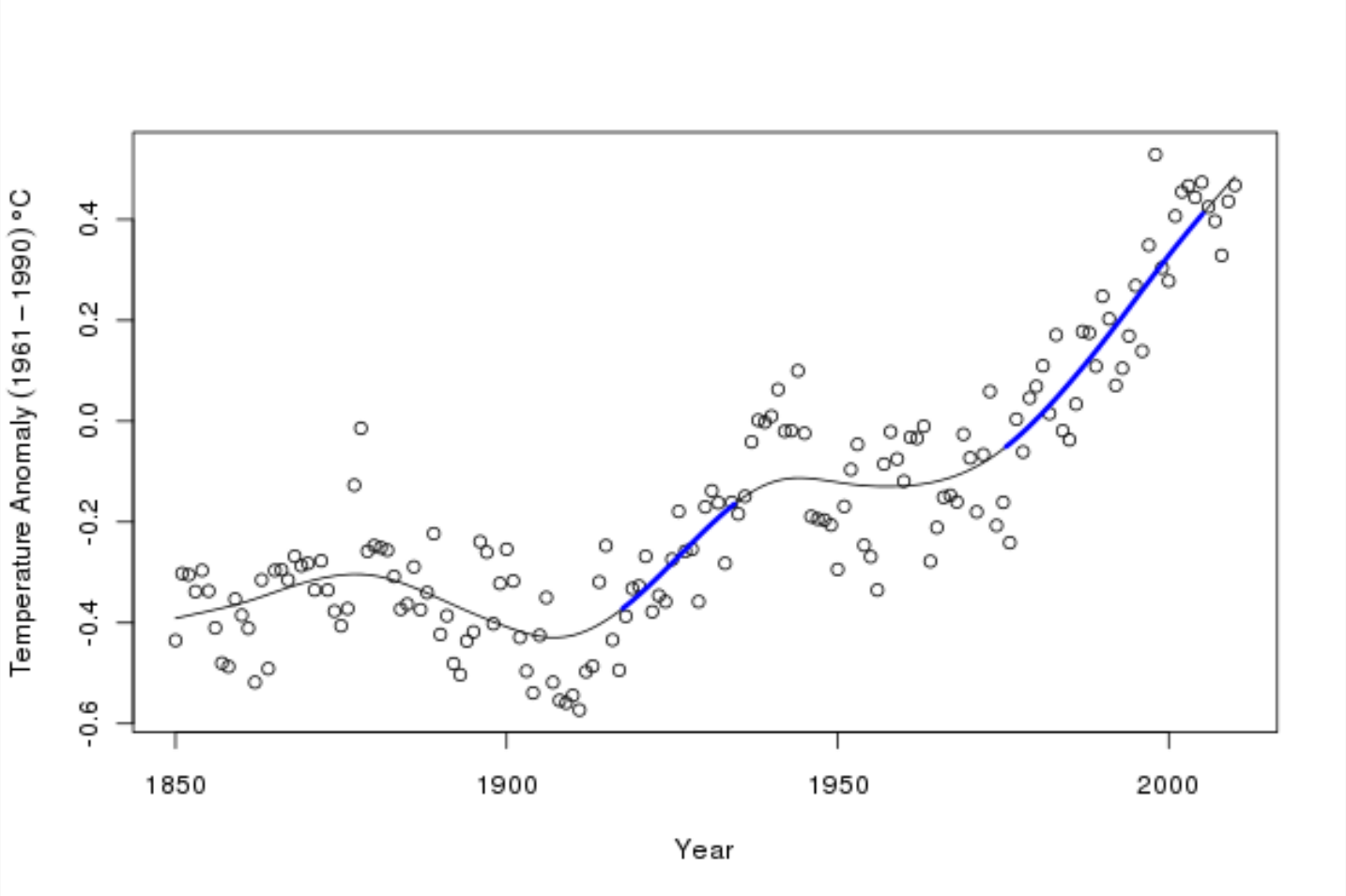
Course topics covered by Glickman (cont)

Smoothers and GAMs: (raw data)



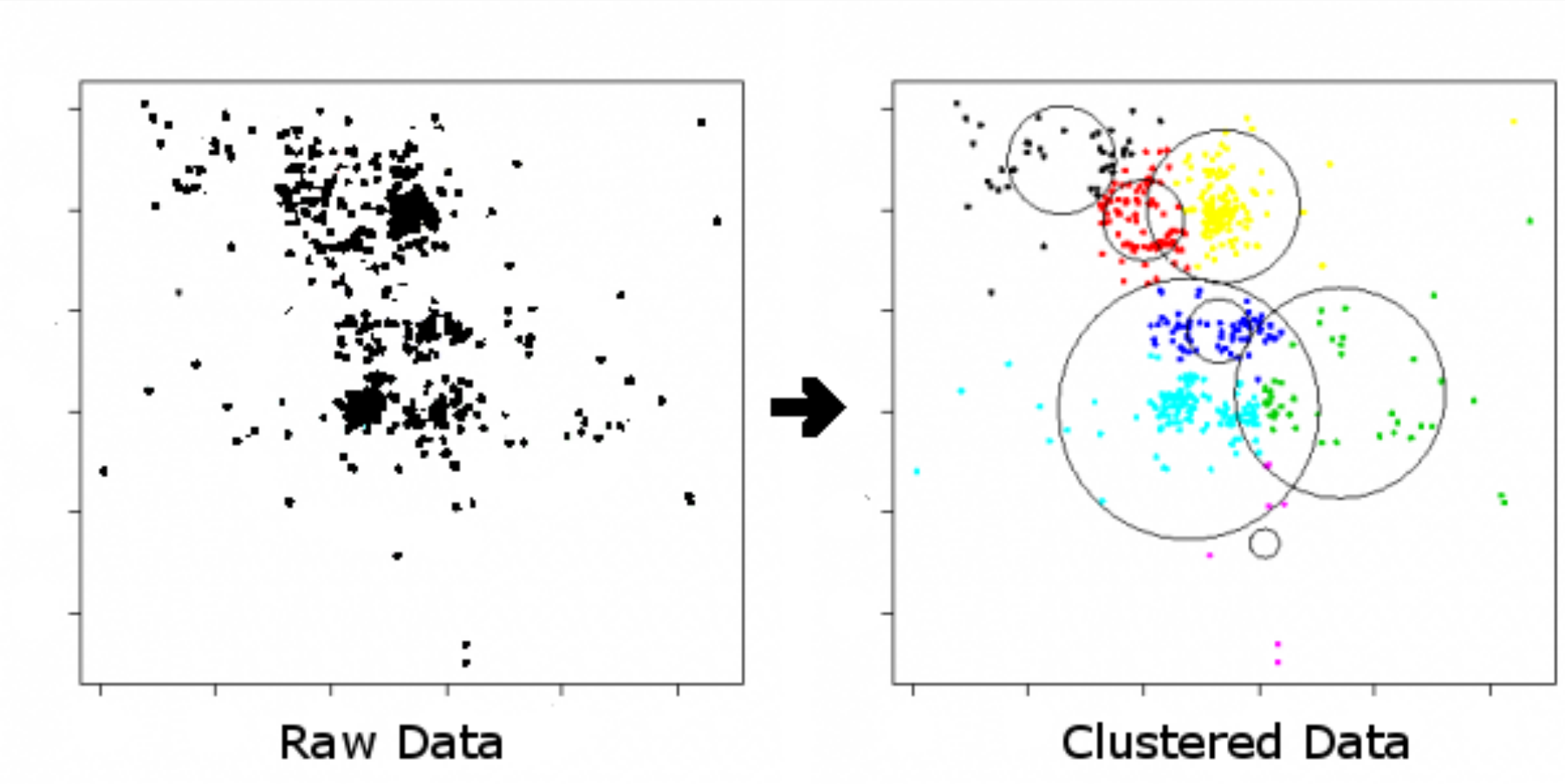
Course topics covered by Glickman (cont)

Smoothers and GAMs: (smoothed fit)



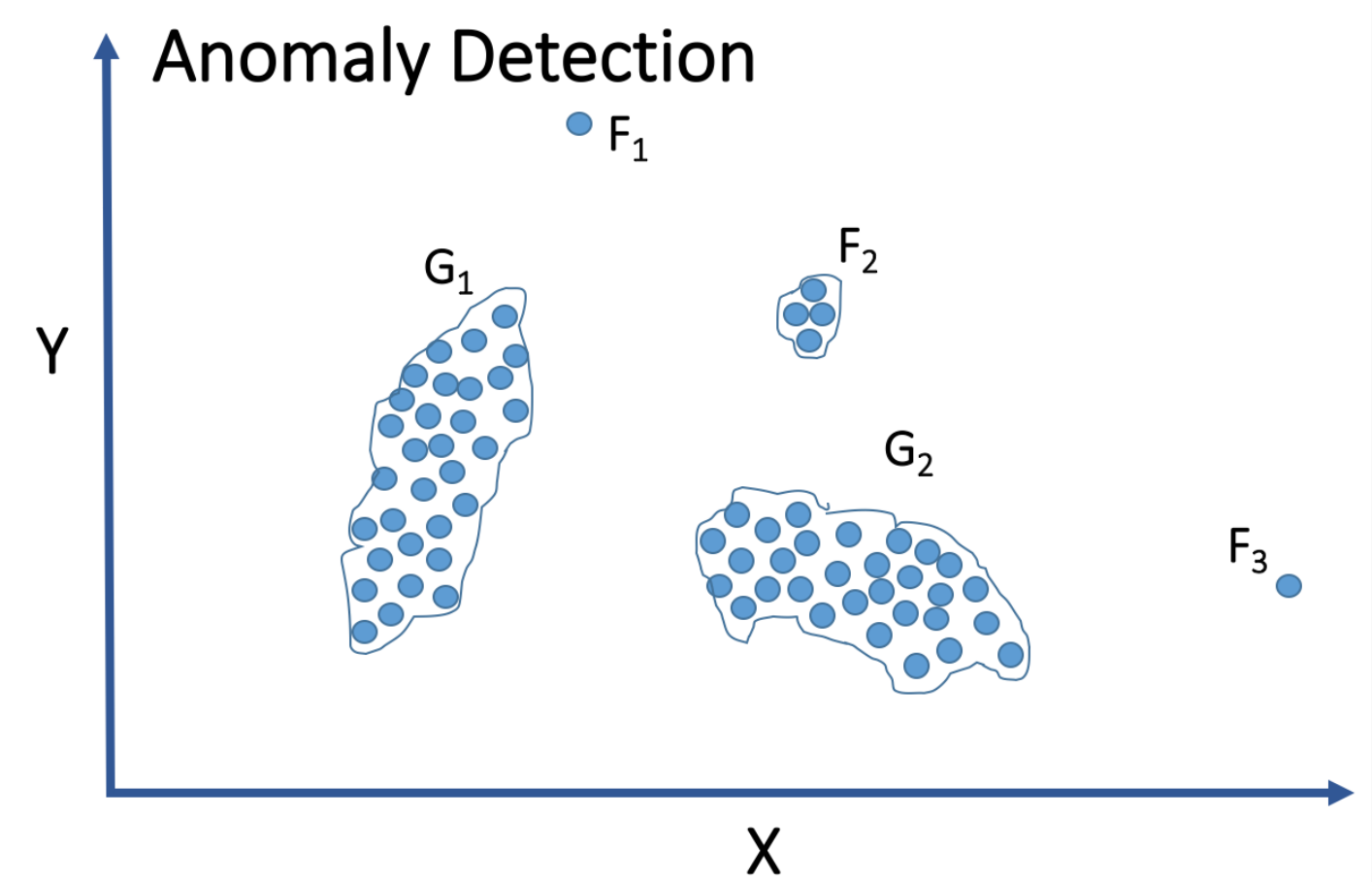
Course topics covered by Glickman (cont)

Cluster analysis:



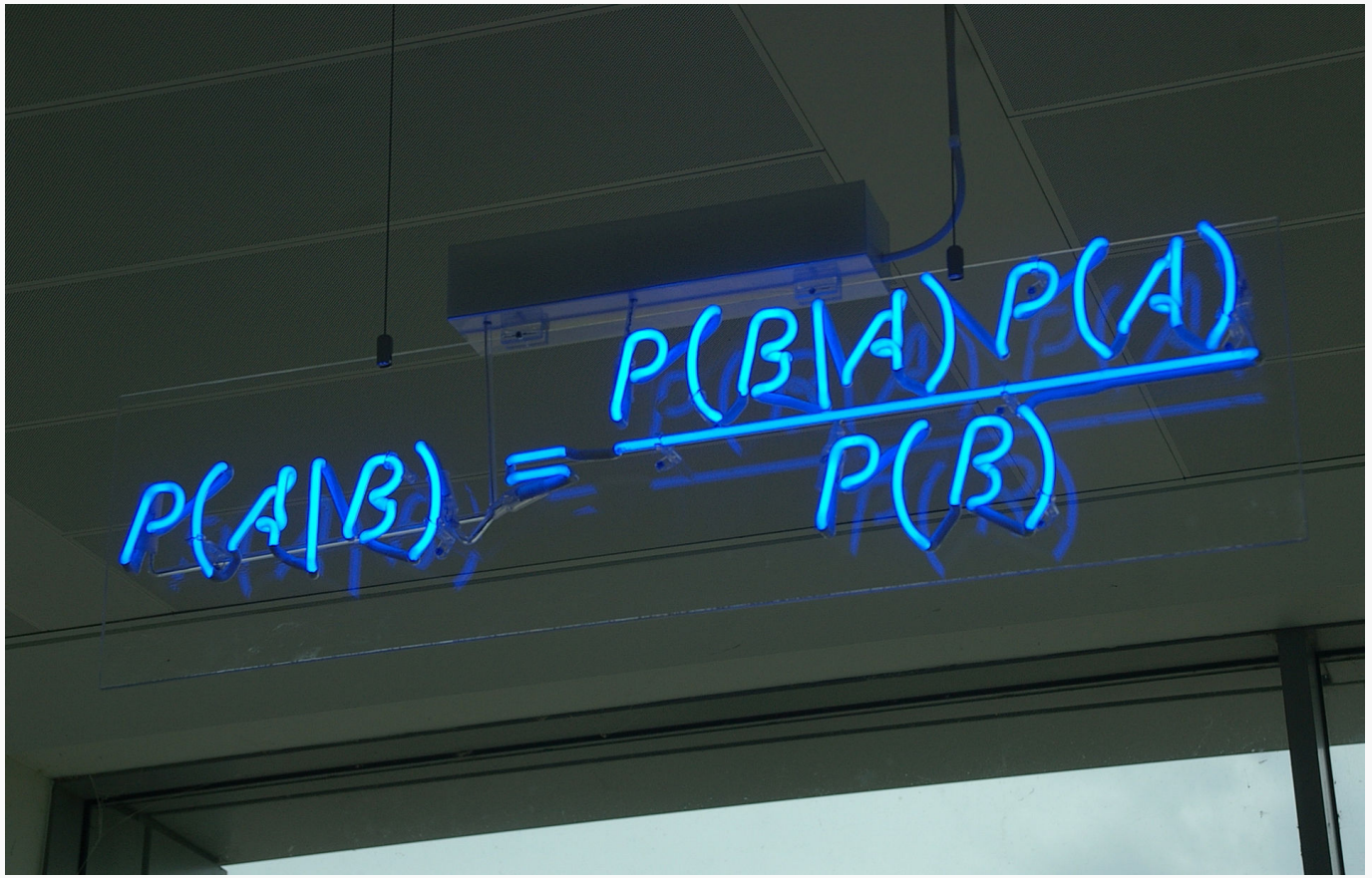
Course topics covered by Glickman (cont)

Example use of cluster analysis:



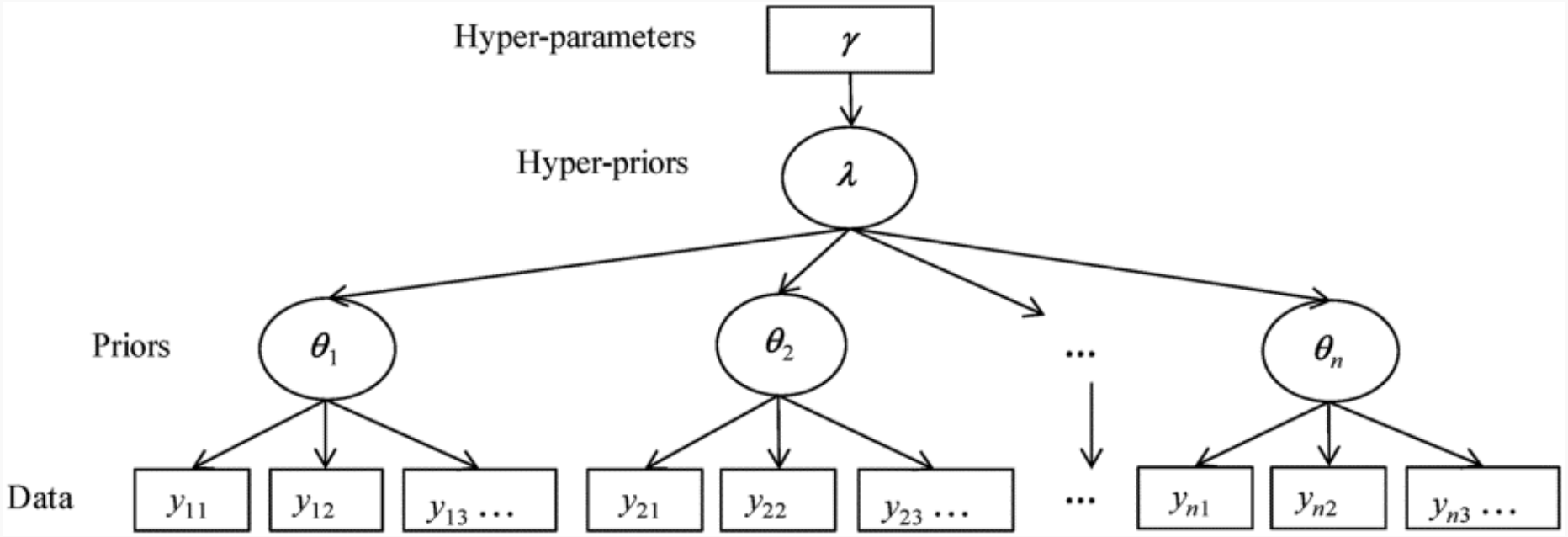
Course topics covered by Glickman (cont)

Bayesian statistics:



Course topics covered by Glickman (cont)

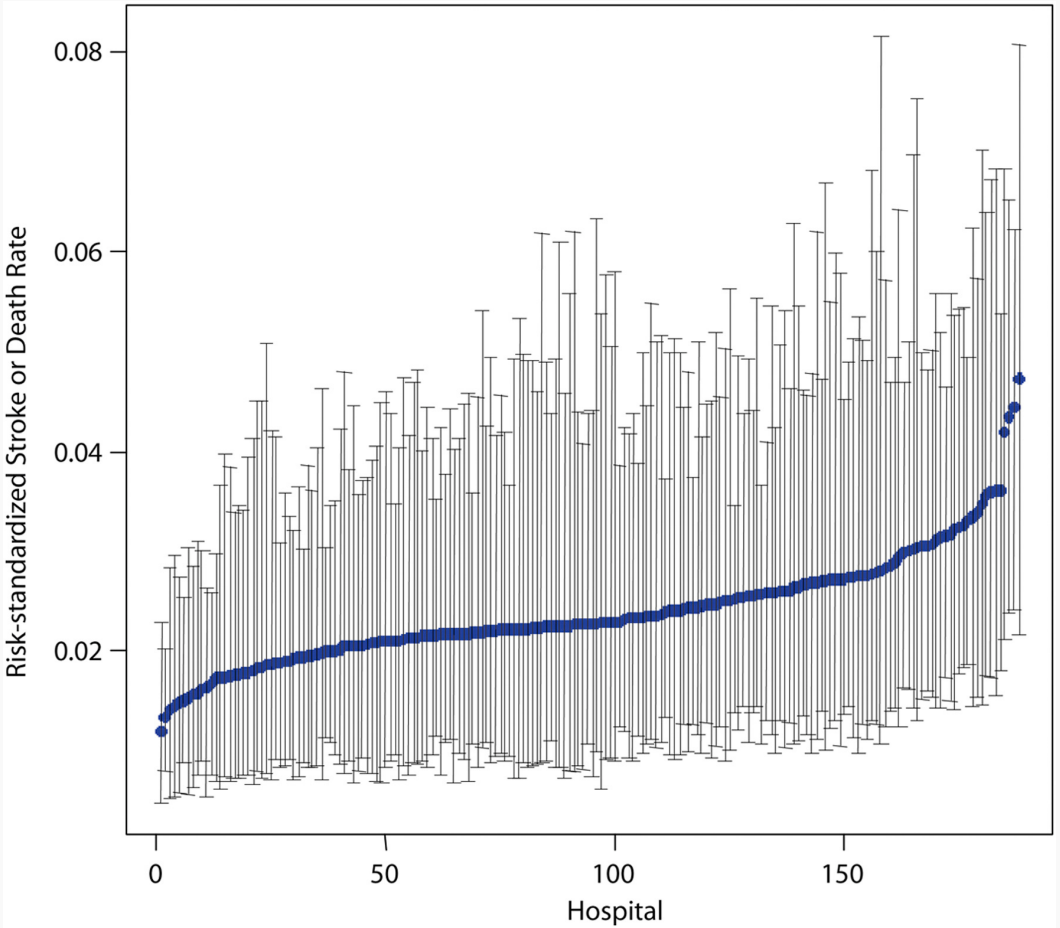
Bayesian statistics: Hierarchical modeling



Course topics covered by Glickman (cont)

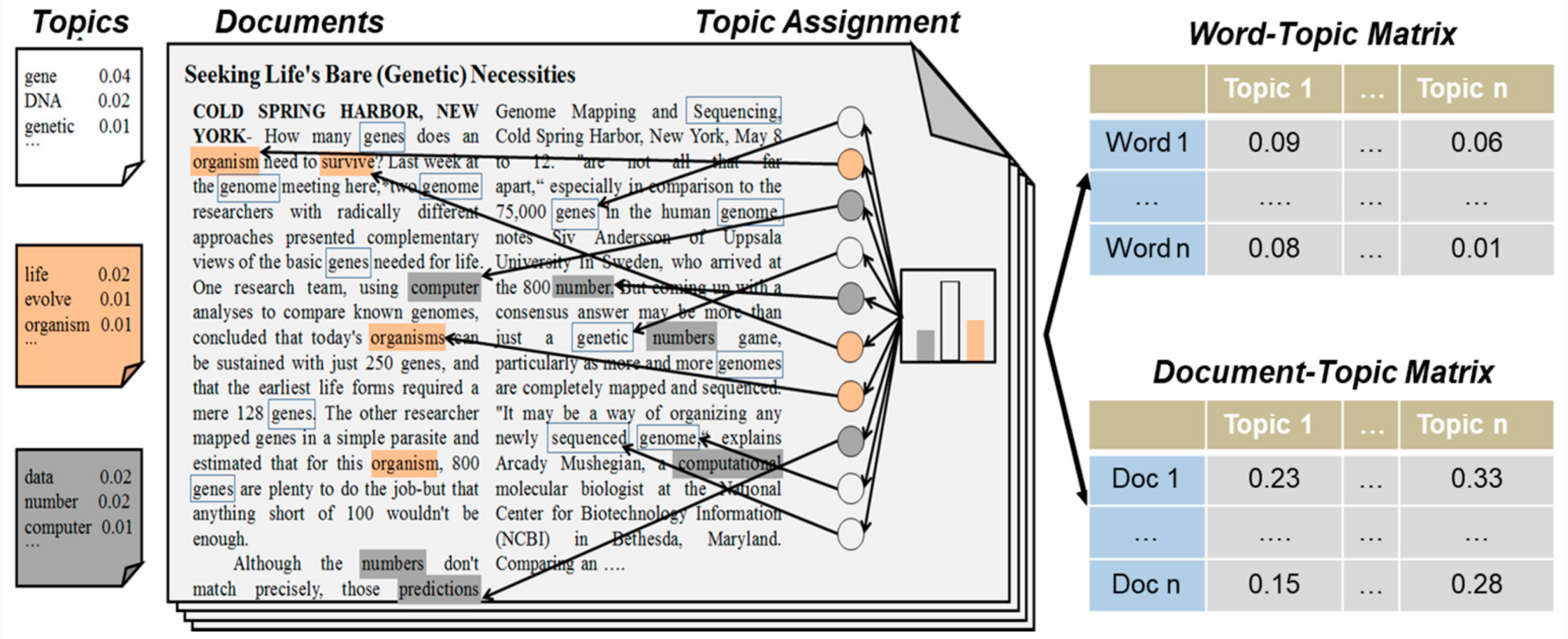
Bayesian statistics: Hierarchical modeling

Hospital Variation in Carotid Stenting Outcomes



Course topics covered by Glickman (cont)

Bayesian statistics: Latent Dirichlet Allocation



Course topics covered by Pavlos

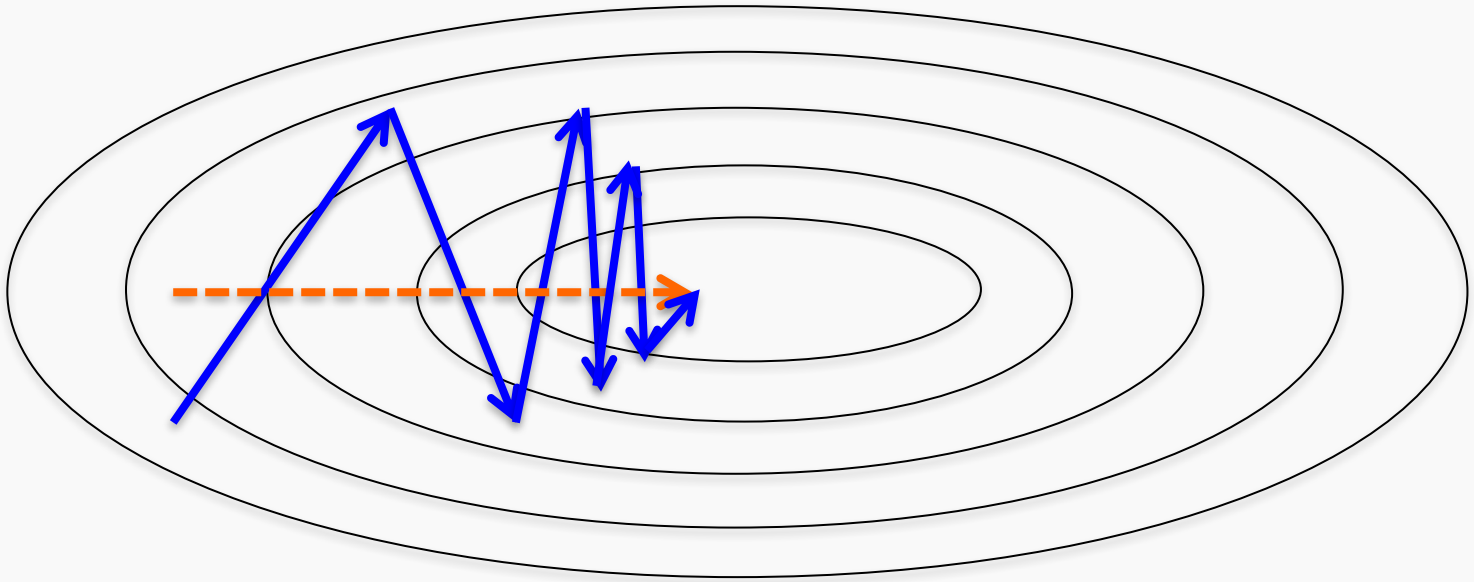
Deep Neural Network

- Review from 109a: Neural Net Basics & Math, Deep Feed Forward, Regularization
- Optimization
- CNNs
- RNNs
- Autoencoders
- Variational Autoencoders
- GANs
- Deep reinforcement learning

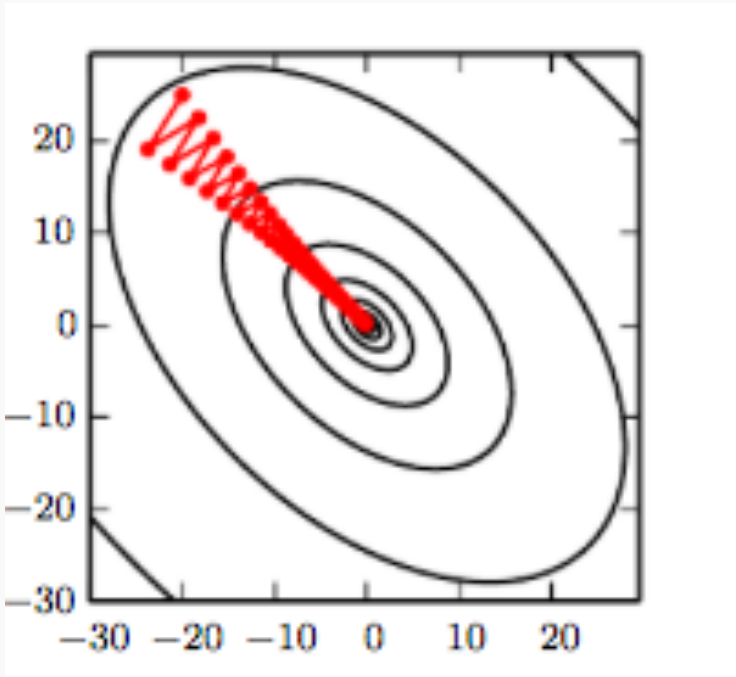
Course topics covered by Pavlos (cont)

SGD is slow when there is high accuracy

SGD



SGD with momentum



LiveSlides web content

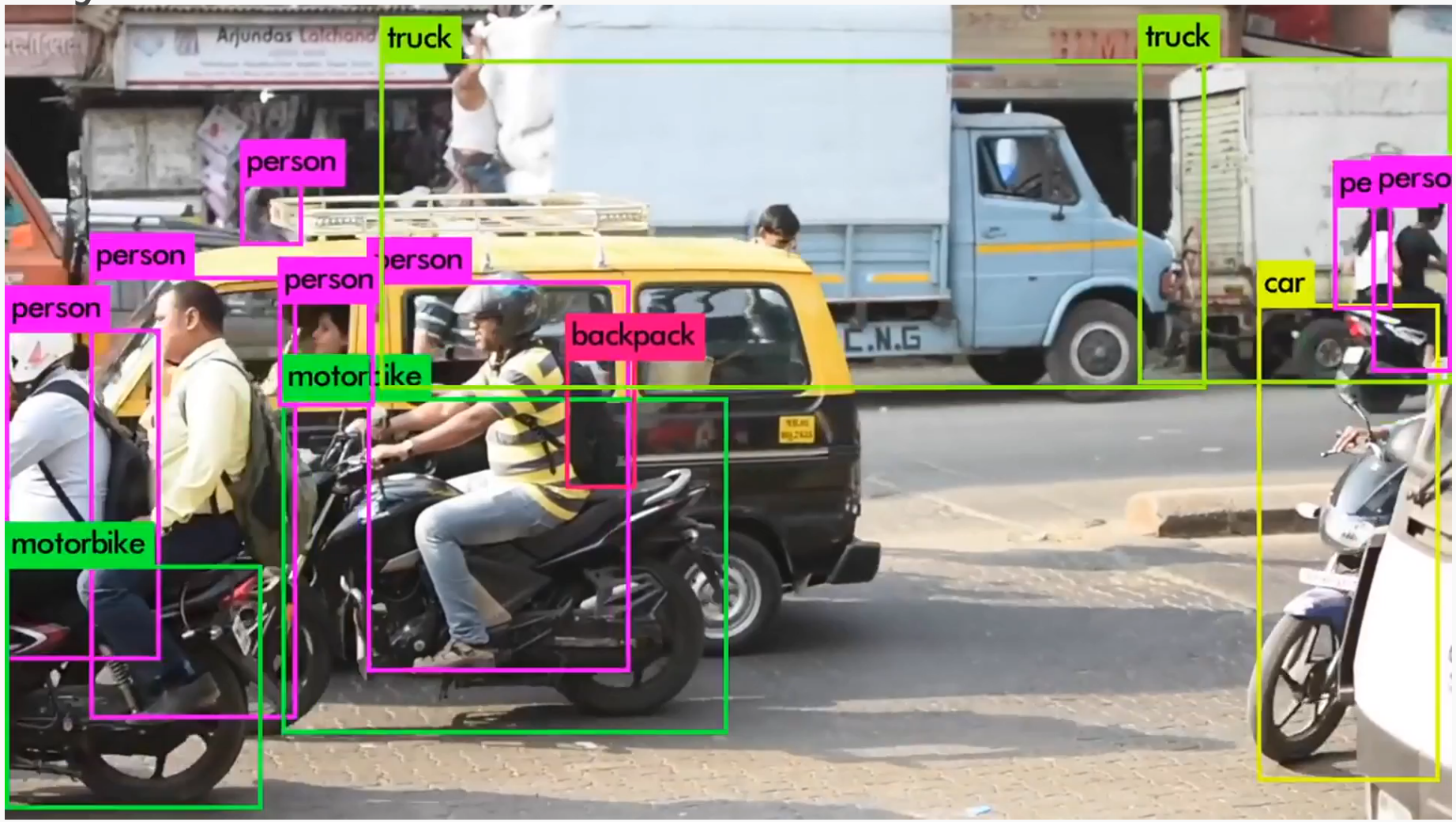
To view

Download the add-in.
liveslides.com/download

Start the presentation.

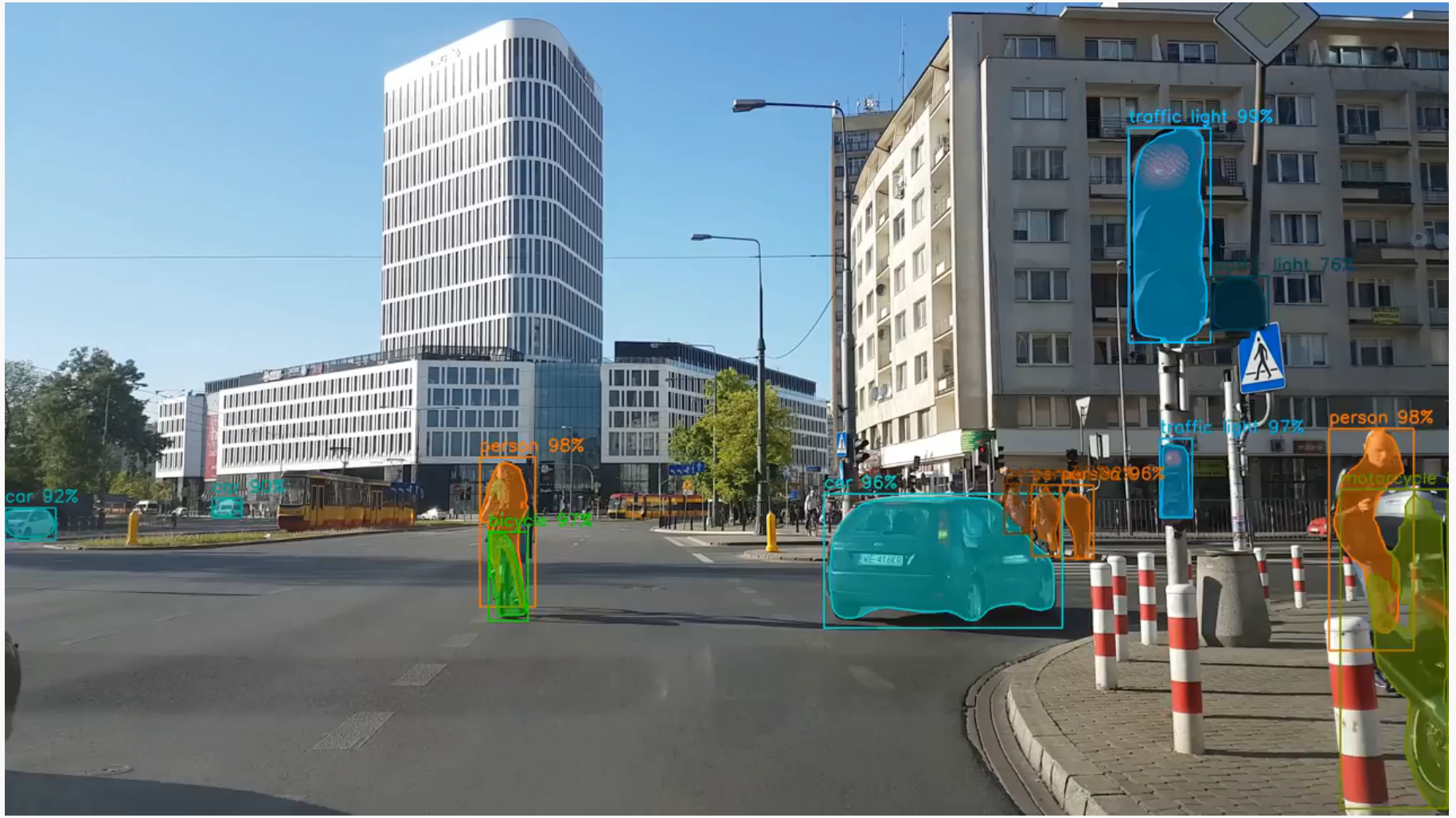
Course topics covered by Pavlos (cont)

You Only Look Once (YOLO) - 2016



Course topics covered by Pavlos (cont)

Mask- RCNN - 2017



Course topics covered by Pavlos (cont)

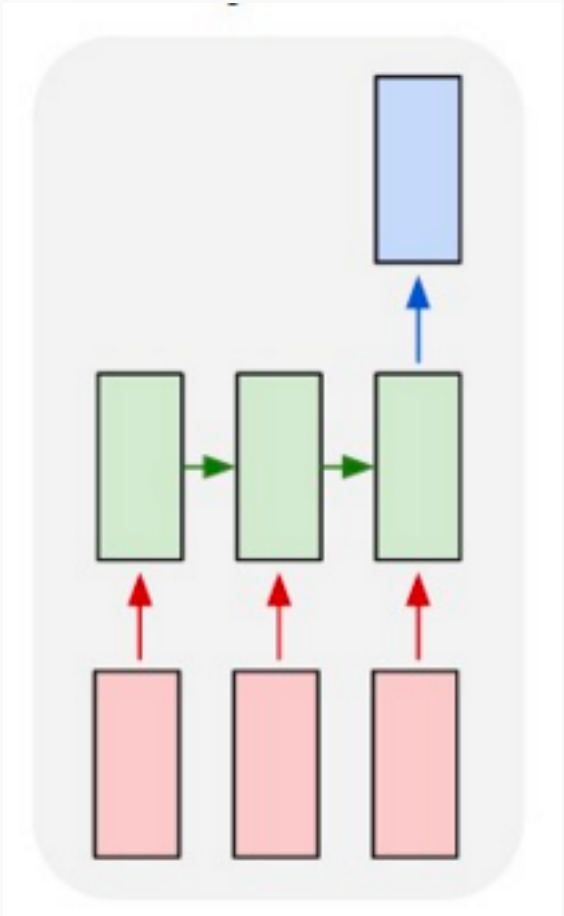
RNN classification, e.g. sentiment analysis

Sentence:

While the music was great, the screenplay was not so engaging and hence even if I started to enjoy it, the movie failed to work for me eventually.

Actual Sentiment: Negative

Predicted Sentiment: ?



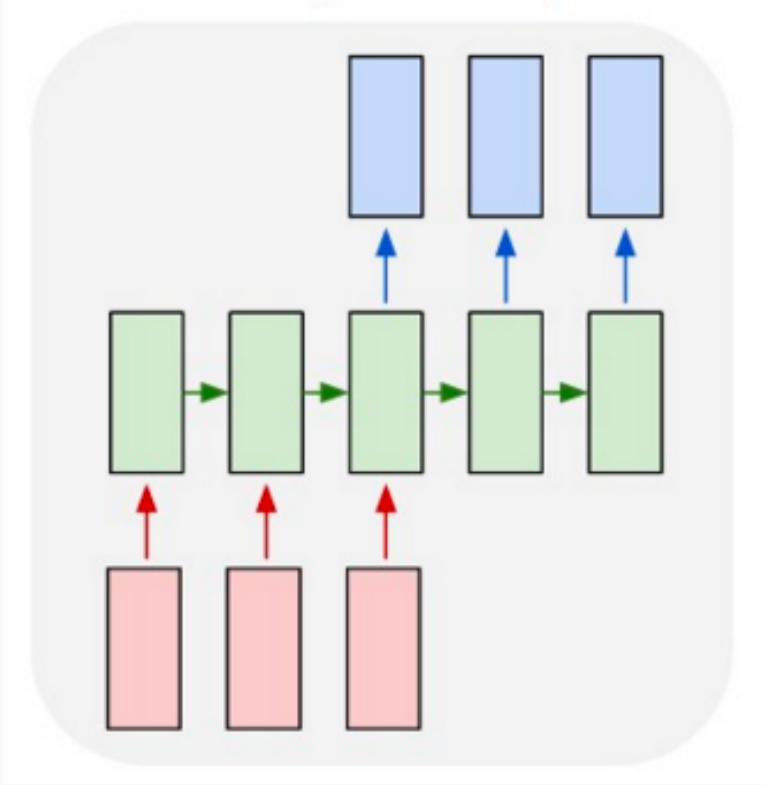
Course topics covered by Pavlos (cont)

RNN sequence to sequence modeling

English:
I love this course

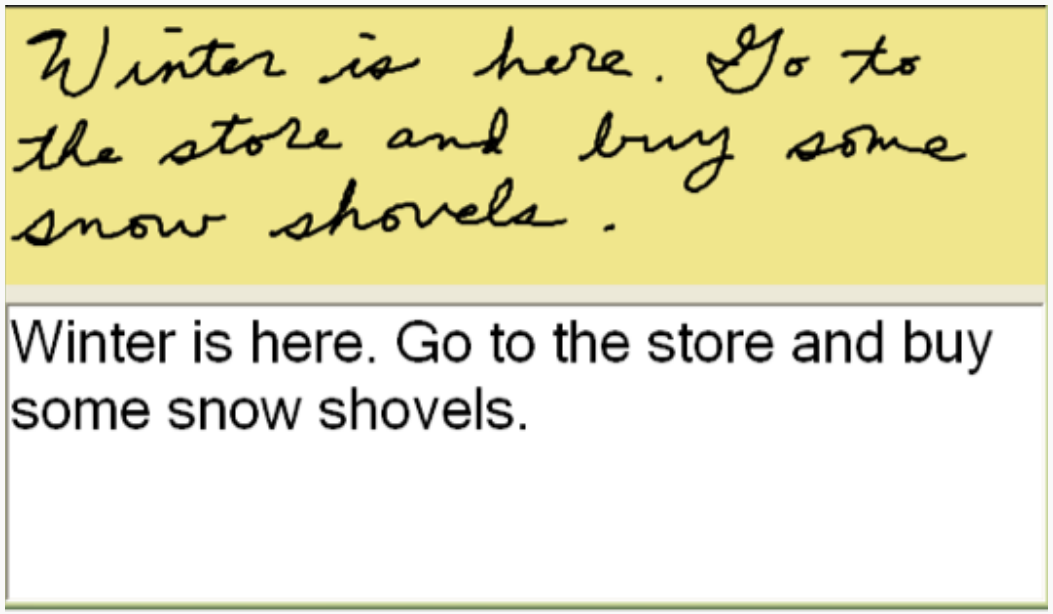
Spanish:
Me encanta esta clase

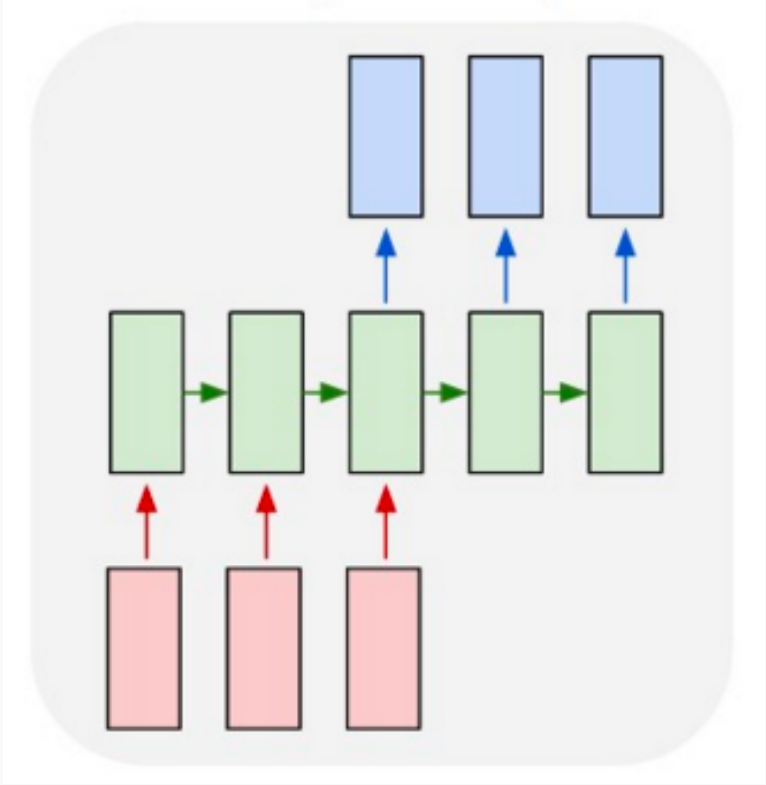
Greek:
Λατρεύω αυτή την τάξη



Course topics covered by Pavlos (cont)

RNN sequence to sequence modeling

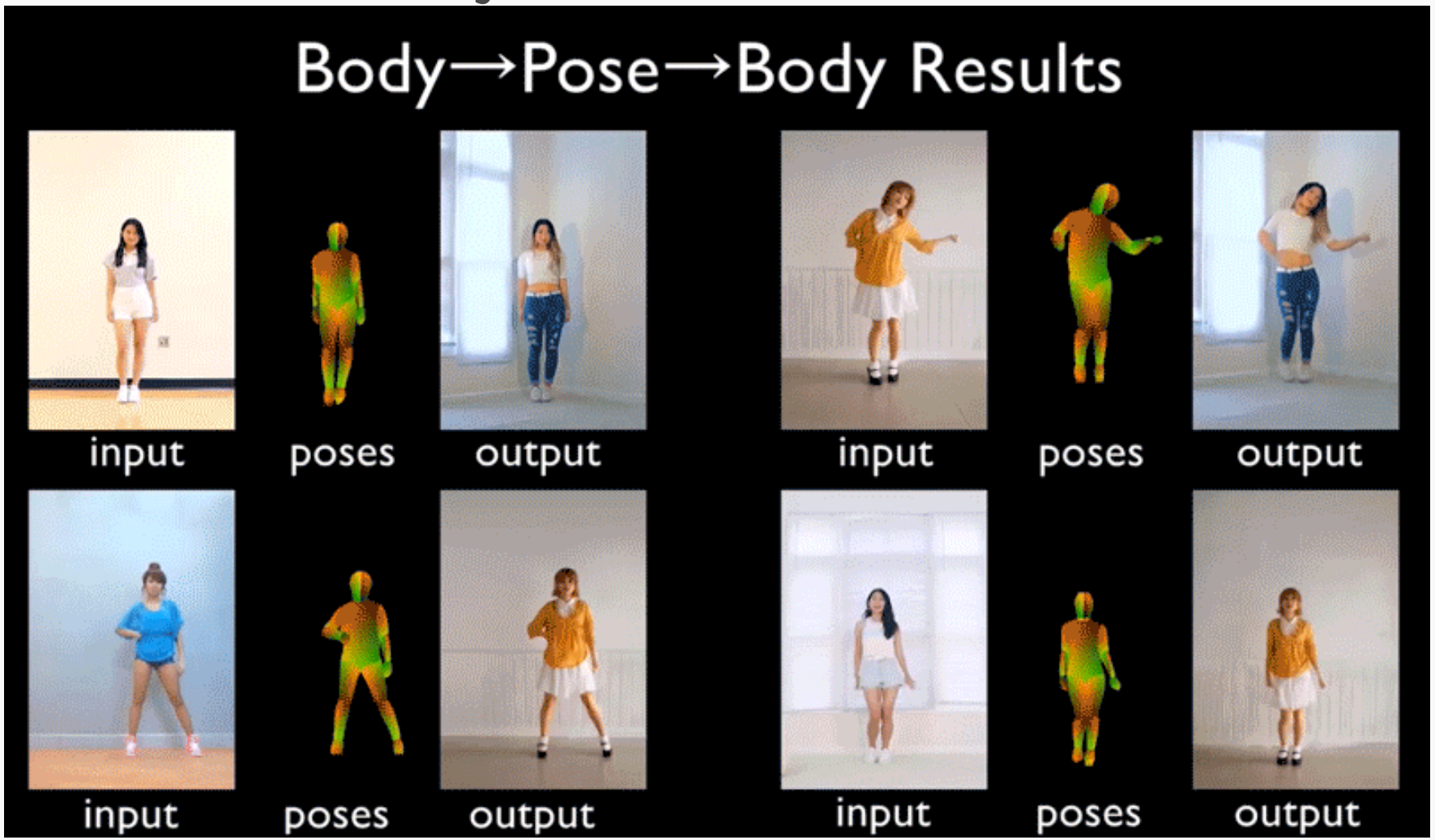




Course topics covered by Pavlos (cont)

Course topics covered by Pavlos (cont)

NVIDIA Video to Video Synthesis - 2018



Other Topics

- Scalability issues: AWS
- Databases: SQL

Projects/Modules

Campus students:

During the final four (4) weeks of the course, students will be divided in break-out thematic sections where they will study topics such as medicine, law, astronomy, e-commerce, and government. Each section will include lectures by faculty, experts on the field, followed by project work also led by that faculty. You will get to present your projects in the SEAS Design Fair at the end of the semester.

Projects/Modules

DCE students:

The goal of the project is to have a complete end-to-end data science process encompassing both semesters of subject material while working as a 3-4 person team. We will supply a small set of project choices within the thematic categories. Teams may propose a different project with sufficient notice and will be subject to approval by the course staff.

Medicine: Microbiome Module

Georg Gerber, Assistant Professor of Pathology, Brigham and Women's Hospital

Through this module, you will have an opportunity to learn about the microbiome, one of the most exciting research fields in science and medicine today. You will also gain experience with analyzing complex time-series data

Modules

Physical Sciences. Potentially hazardous object

Matthew Holman, Harvard-Smithsonian Center for Astrophysics, Director of the Minor Planet Center

Through this module, you will have an opportunity to save the world by discovering, identifying and characterizing Potentially hazardous object

Modules

Law: Most likely NLP

David A. Colarusso, Clinical Fellow, and Director of the Legal Innovation & Technology Lab

Details to come in the next few weeks.

Modules

Business School:

Srikanth Jagabathula, and **Ashwin Jagabat**, HBS

Details to come in the next few weeks.

Projects/Modules

Government:

David Eaves, HKS

Details to come in the next few weeks.

Outline

- Who
- What have we learned in 109a?
- What is covered in 109b
- **Course Logistics**

Outline

- Who
- What have we learned in 109a?
- What is covered in 109b
- **Course Logistics**
  - Lectures, Labs and Office Hours
  - AC209
  - Homeworks
  - Content Organization
  - Grades

Outline

- Who
- What have we learned in 109a?
- What is covered in 109b
- Course Logistics
  - **Lectures, Labs and Office Hours**
  - AC209
  - Homeworks
  - Content Organization
  - Grades

Lectures, Labs and Office Hours

Lectures: Mondays and Wednesdays 1:30-2:45pm @ Maxwell-Dworkin G-115.

During lecture will cover the material which you will need to complete the homework. Attending lectures is required - quizzes at the end of each lecture (drop 40% of them) .

1. Lecture notes and associated notebooks will be posted before lecture on Canvas and GitHub (Pavlos Lectures only)

2. Lectures will be video tapes (and live streamed for DCE students) and posted approximately in 24 hours on Canvas.

Note: For DCE students the quizzes are optional

Lectures, Labs and Office Hours (cont)

Labs: Thursdays 4:30-6:00pm @ Pierce 301

Labs are meant to help you understand the lecture materials better via examples.

Labs will also be video taped (and live streamed for DCE students) and posted approximately in 24 hours on Canvas

Lectures, Labs and Office Hours (cont)

Instructors Office Hours:

- **Mark:** By appointment
- **Pavlos:** Monday 3:00-4:00pm, MD G108

TF Office Hours: Check canvas in the next day or so

Outline

- Who
- What have we learned in 109a?
- What is covered in 109b
- Course Logistics
  - Lectures, Labs and Office Hours
  - **AC209**
  - Homeworks
  - Content Organization
  - Grades

AC 209 Students

Students enrolled for the AC 209B course have the following extra requirements:

1. Attend A-Sections
2. Complete extra questions in homework (HW3-HW8)
3. Expand the scope of the final project (modules)

AC 209 Students (cont)

Advance Sections Topics:

1. Optimization/EMD
2. Dropout
3. ConvNets: LeNet, AlexNet, VGG-15, ResNet and Inception (HW)
4. LSTN, GRU in NLP (HW)
5. Neural style transfer learning (HW)
6. Deep RL (HW)
7. Variational Inference (HW)
8. GANS (HW)

Outline

- Who
- What have we learned in 109a?
- What is covered in 109b
- Course Logistics
  - Lectures, Labs and Office Hours
  - AC209
  - **Homeworks**
  - Content Organization
  - Grades

Homeworks

See schedule on canvas for released dates and due dates (http://bitly.com/2FWLM0v)

- Homework 0 **(Due Feb 6 11:59pm) Must submit**
- Homework 1 on smoothing
- Homework 2 on FF NN and various optimizations
- Homework 3 on CNNs
- Homework 4 on RNNs
- Homework 5 on Clustering and Autoencoders (individual)
- Homework 6 on LDA and Bayes
- Homework 7 on VAE+GANS (individual)

Homeworks (cont)

You are encouraged but not required to submit in pairs.
Instructions on how to submit in pairs are on canvas.

If you work with someone else but not submitting in pair you should indicate who you have worked with.

All assignments will be posted on Tues. at 11:59pm and will be due Wed. at 11.59pm - one or few weeks later (see schedule for details).

You only need to submit .ipynb (no need to submit pdf)

Homeworks (cont)

Grading: The homework provides an opportunity to learn advanced data science skills and to bolster your understanding of the material. See the homework as an opportunity to learn, and not to earn points. The homework will be graded to reflect this objective.

WE WILL BE PROVIDING LENGTHY FEEDBACK AND DETAILS HOW GRADING IS DONE (different from 109a)

Homeworks (cont)

Late Policy: No homework assignments will be accepted for credit after the deadline. If you have a verifiable medical condition or other special circumstances that interfere with your coursework please let us know as soon as possible by sending an email to the *Helpline*.

Homeworks (cont)

Late Policy: No homework assignments will be accepted for credit after the deadline. If you have a verifiable medical condition or other special circumstances that interfere with your coursework please let us know as soon as possible by sending an email to the *Helpline*.

Homeworks (cont)

Regrades: Our graders and instructors make every effort in grading accurately and in giving you a lot of feedback.

If you discover that your answer to a homework problem was correct but it was marked as incorrect, send an email to the *Helpline* with a description of the error. Please do not submit regrade requests based **on what you perceive is overly harsh grading**.

The points we take off are based on a grading rubric that is being applied uniformly to all assignments.

If you decide to send a regrade request, send an email to the *Helpline* with subject line "Regrade HW1: Grader=johnsmith" within **48 hours of the grade release**.

Content Organization

Github: https://github.com/Harvard-IACS/2019-CS109B

- Lab files and solutions
- Advanced Section materials
- Lectures (Pavlos' lectures only)

Assignments and Mark's lectures will only be posted on Canvas.

Grades

- Paired-option Homeworks: **45%** (5 homework for which you have the option to work in pairs)
- Individual Homeworks: **25%** (2 homework which you must complete individually, HW5 and HW7)
- Quizzes: **10%** (you may drop 40% of the quizzes)
- Project: **20%**

Academic Integrity

Ethical behavior is an important trait of a Data Scientist, from ethically handling data to attribution of code and work of others. Thus, in CS109 we give a strong emphasis to Academic Honesty.

As a student your best guidelines are to be reasonable and fair. We encourage teamwork for problem sets, but you should **not** split the homework and you should work on all the problems together.

Please be responsible and when in doubt ask either Pavlos, Mark, Eleni or Sol.

Auditing

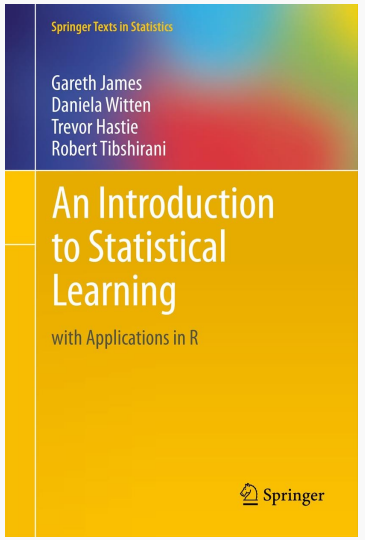
Auditors are welcomed!

If you would like to audit the class, please send an email to the *Helpline* indicating who you are and why you want to audit the class. You need a HUID to be included to Canvas.

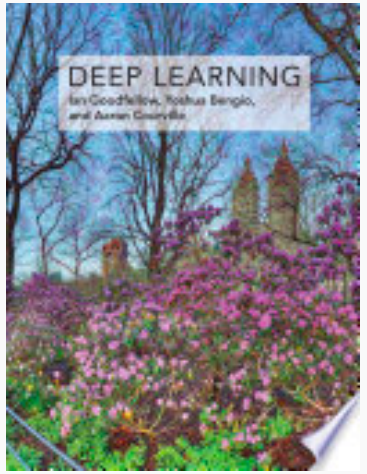
The only stipulation is that you don't plan to ask the teaching staff for help, attend office hours, etc., as these resources are intended for registered students.

Recommended Textbooks

ISLR: An Introduction to Statistical Learning, by James, Witten, Hastie, Tibshirani (Springer: New York, 2013)



DL: Deep Learning (http://www.deeplearningbook.org/) by Goodfellow, Bengio and Courville.



Accommodations for students with disabilities

Students needing academic adjustments or accommodations because of a documented disability must present their Faculty Letter from the Accessible Education Office (AEO) and speak with the professor by the end of the second week of the term.

Failure to do so may result in the Course Head's inability to respond in a timely manner. All discussions will remain confidential.

Diversity and Inclusion

- If you have a name and/or set of pronouns that differ from those that appear in your official Harvard records, please let us know!
- If you feel like your performance in the class is being impacted by your experiences outside of class, please don't hesitate to come and talk with us. If you prefer to speak with someone outside of the course, you may find helpful resources at the Harvard Office of Diversity and Inclusion.
- We (like many people) are still in the process of learning about diverse perspectives and identities. If something was said in class (by anyone) that made you feel uncomfortable, please talk to us about it.

See our diversity and inclusion statement on canvas

Looking forward to a fun and productive semester!