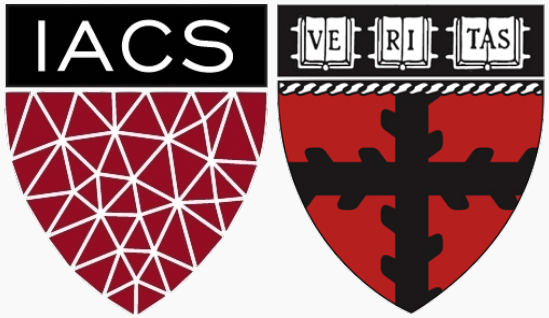
Lecture 0: Introduction to CS109B

CS 109B, STAT 121B, AC 209B, CSE 109B

Mark Glickman Pavlos Protopapas



Lecture Outline

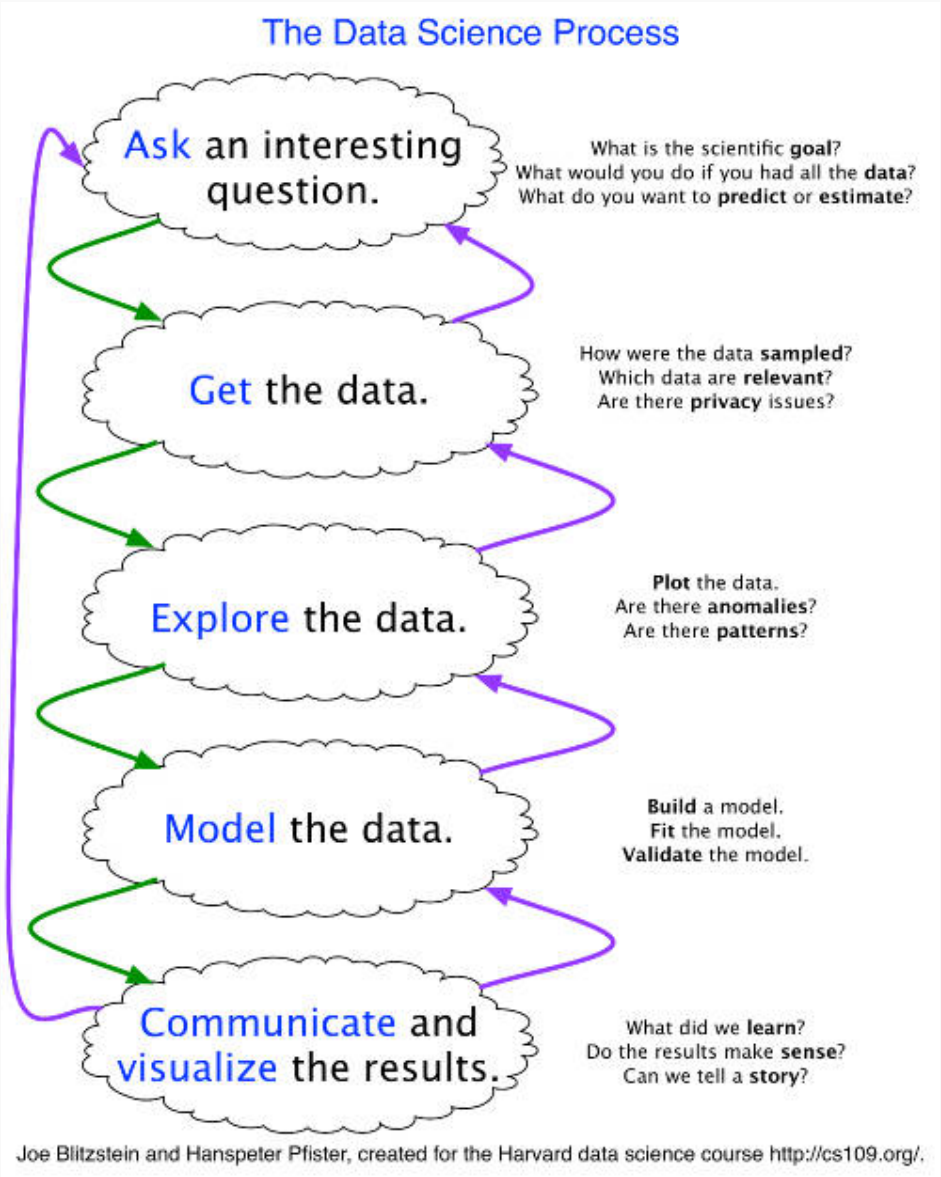
What have we learned in 109a?

This Class

What have we learned in 109a?

109a

109A Data Process



CS109A Topic

Topics:

- 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

This Class

Who

Mark Glickman, Senior Lecturer in Statistics

Who


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

Who

Pavlos Protopapas

Who

Pavlos Protopapas

- Scientific Director of the Institute for Applied Computational Science (IACS)
- CS109 and the Capstone course for the Data Science masters program. Research in astrostatistics and excited about the new telescopes coming online in the next few years.

Lab Instructors

Katy McKeough: kathrynmckeough@g.harvard.edu
G-3 PhD STATS. Lab1-Lab6 (R related)

Lab Instructors

Weiwei Pan: weiweipan@g.harvard.edu
IACS Postdoc. Lab8 and Lab9

Lab Instructors

David Wihl: davidwihl@g.harvard.edu
Lecture 10,20, Lab7, Lab10,11

Advanced Sections Instructors

Javier Zazo (IACS affiliate scientist):
javier.zazo.ruiz@gmail.com

Teaching Fellows

Eleni Kaxiras (Head TF): eleni@seas.harvard.edu

Teaching Fellows

Nicholas Ruta (Head TF DCE): nruta@g.harvard.edu

Teaching Fellows

- Kevin Wu: kewu93@gmail.com
- Eric Wu: eric_wu@g.harvard.edu
- Zona Kostic: zonakostic@g.harvard.edu
- Sol Girouard: solgirouard@g.harvard.edu
- Rashmi Banthia: rjain29@gmail.com
- Raghu Dhara: rdhara@college.harvard.edu

Topics

The semester is divided into 3 parts.

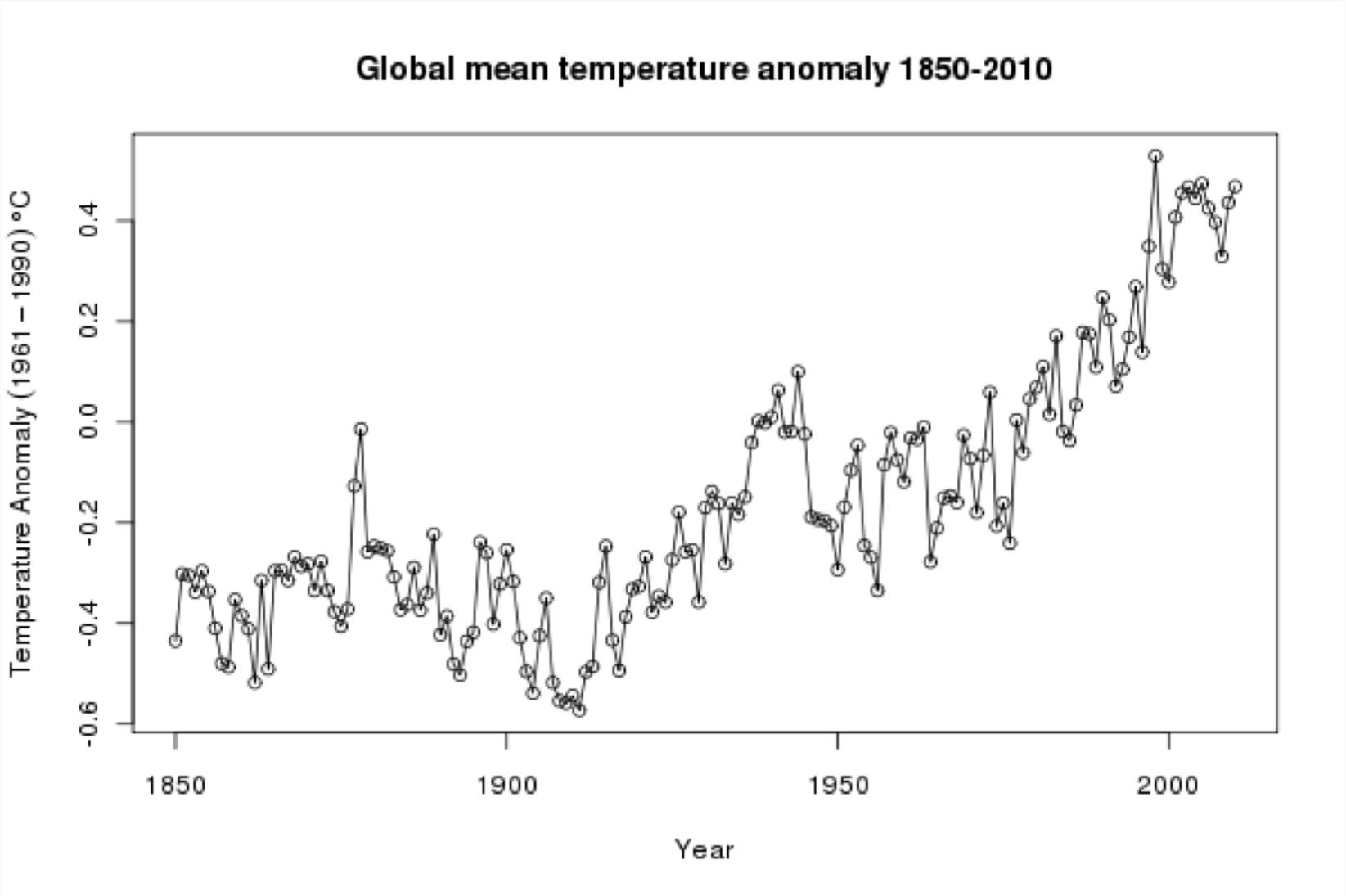
- **Part 1:** Smoothing, Unsupervised Learning and Bayesian inference in R and Stan
- **Part 2:** Deep Neural Networks, scalability issues in Keras and python
- **Project:** Incorporates everything from 109a and 109b

Part 1 topics: covered by Mark

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

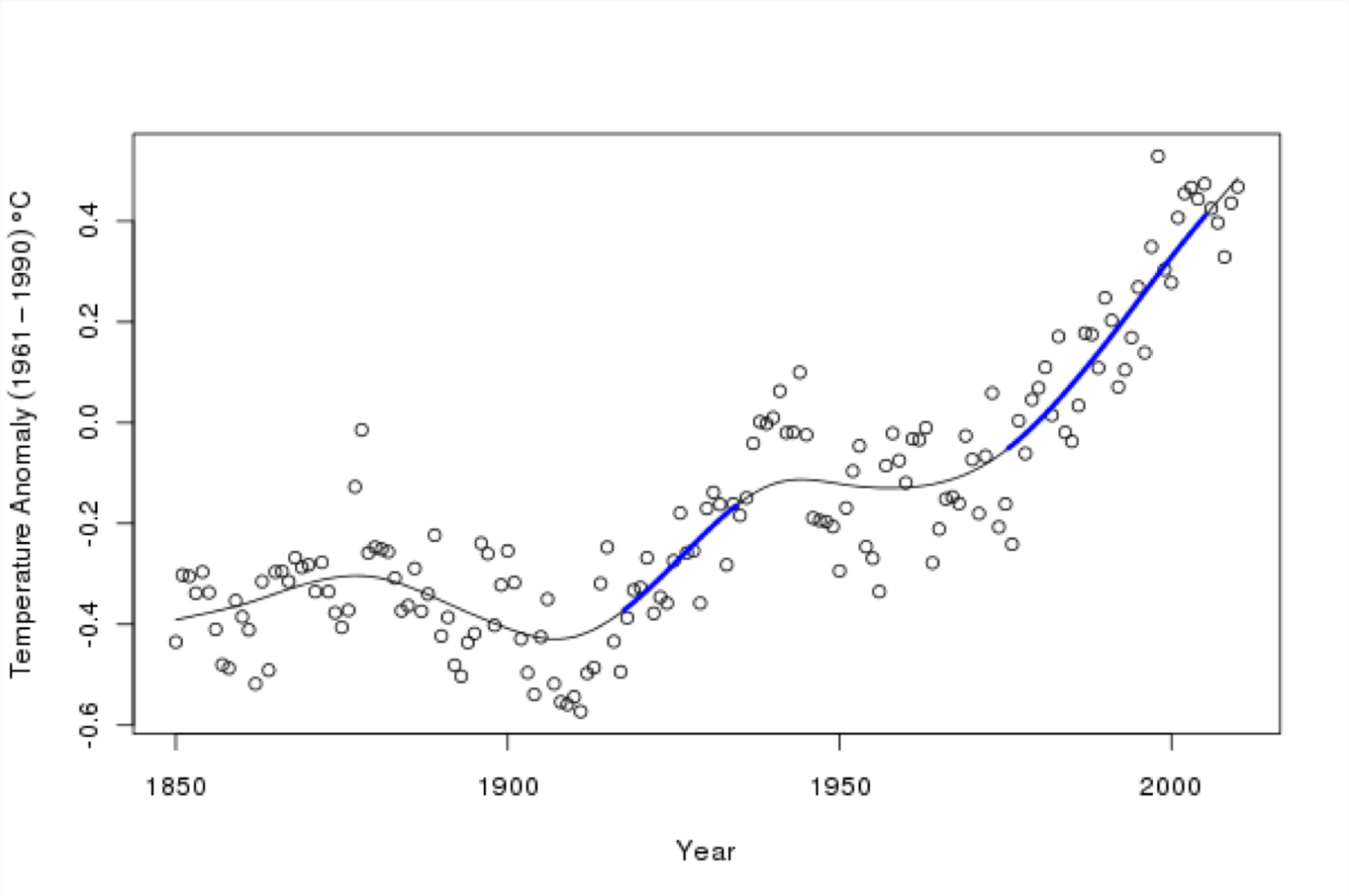
Part 1 topics: covered by Mark

Smoothers and GAMs: (raw data)



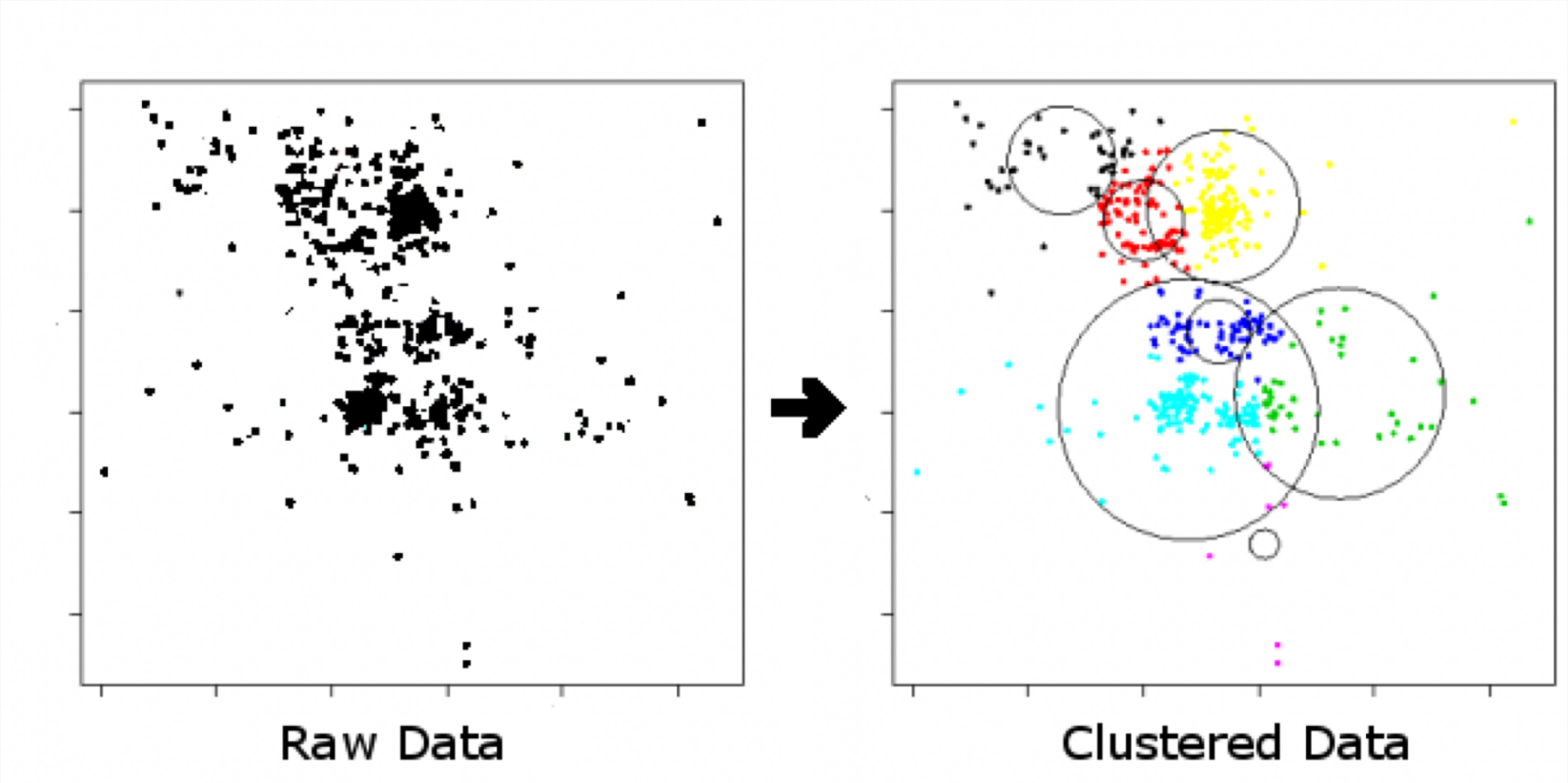
Part 1 topics: covered by Mark

Smoothers and GAMs: (smoothed fit)

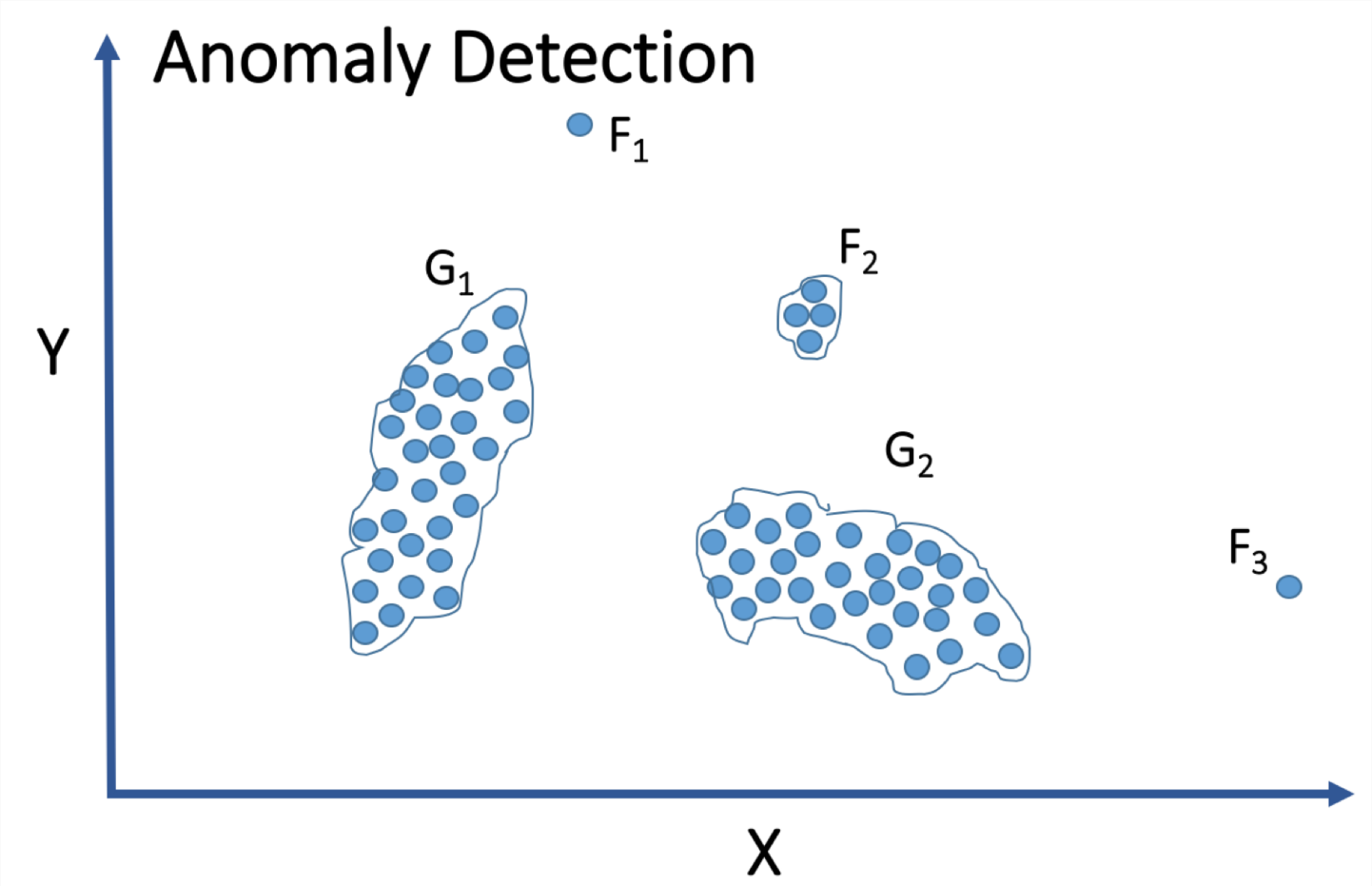


Part 1 topics: covered by Mark

Cluster analysis:

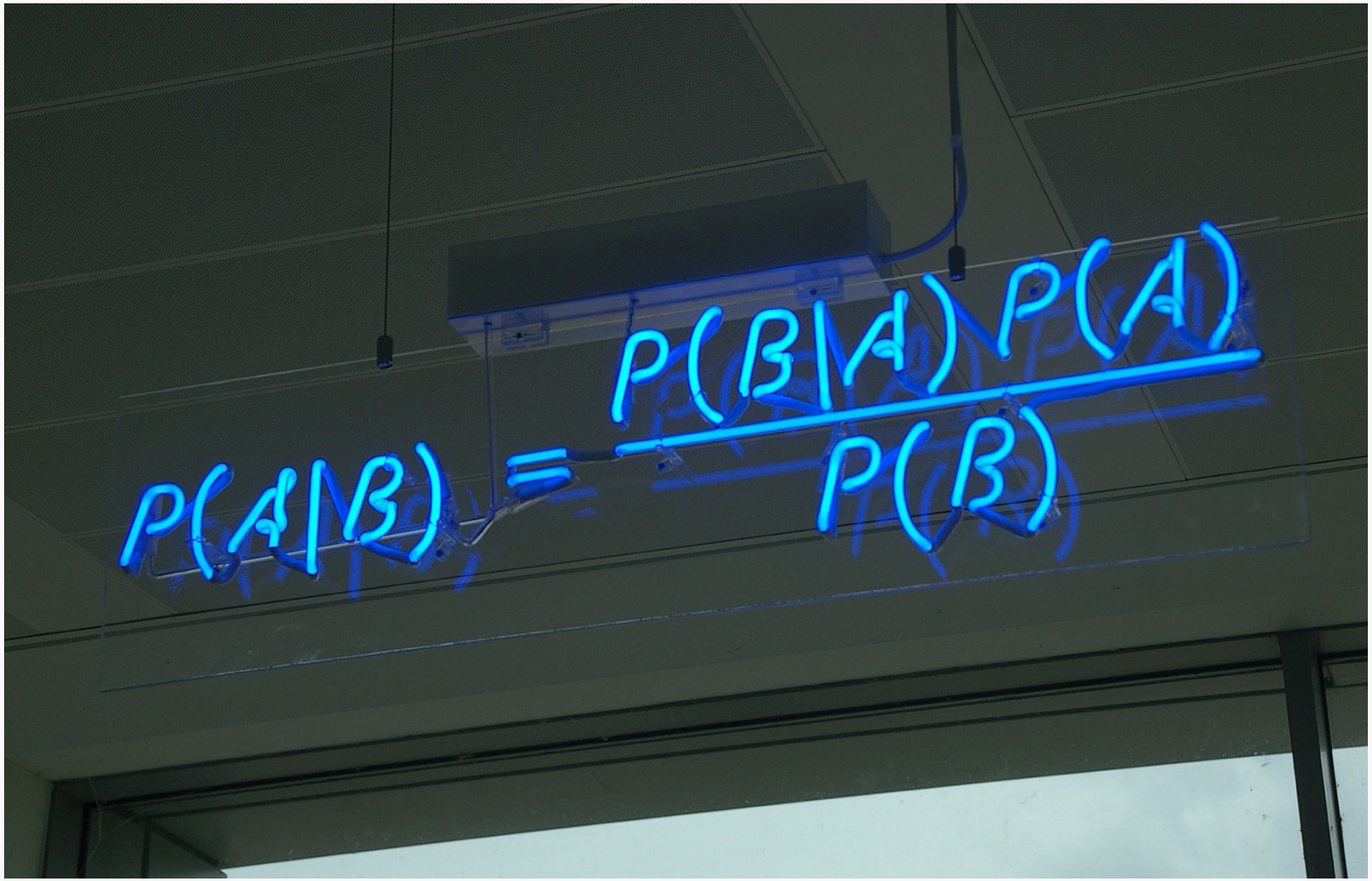


Example use of cluster analysis:



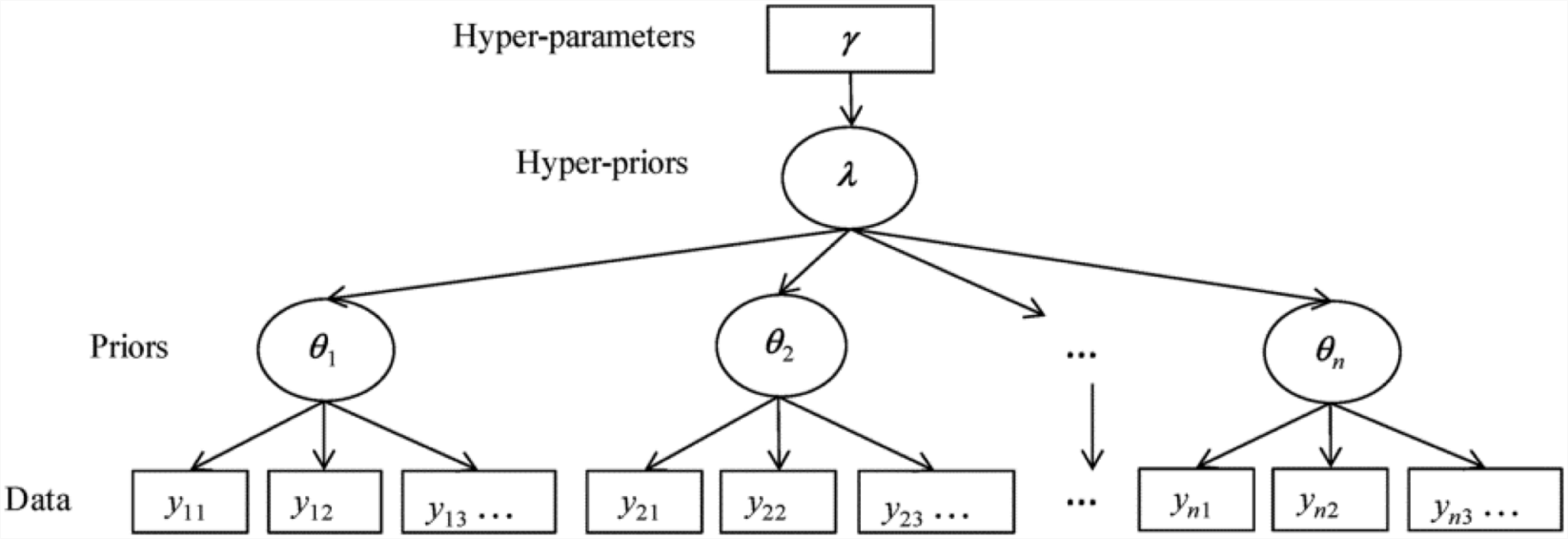
Part 1 topics: covered by Mark

Bayesian statistics:



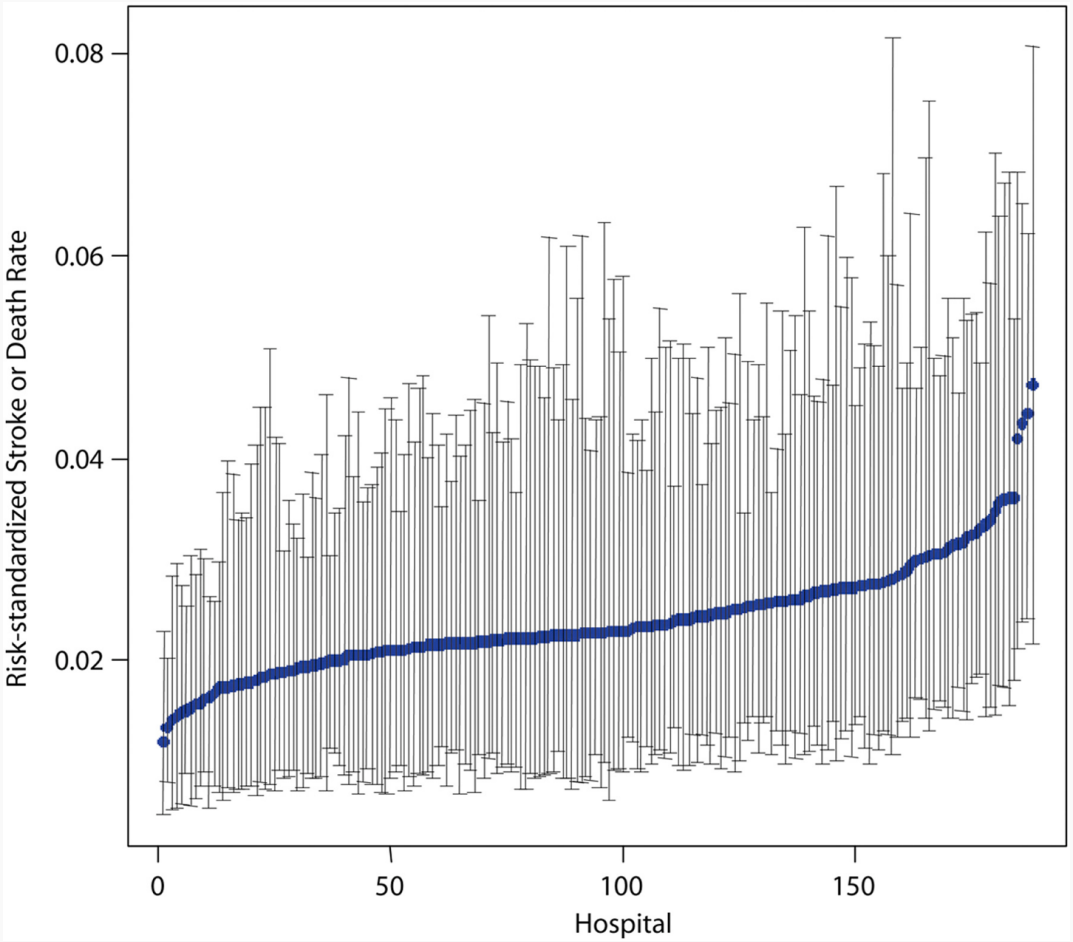
Part 1 topics: covered by Mark

Bayesian statistics: Hierarchical modeling :



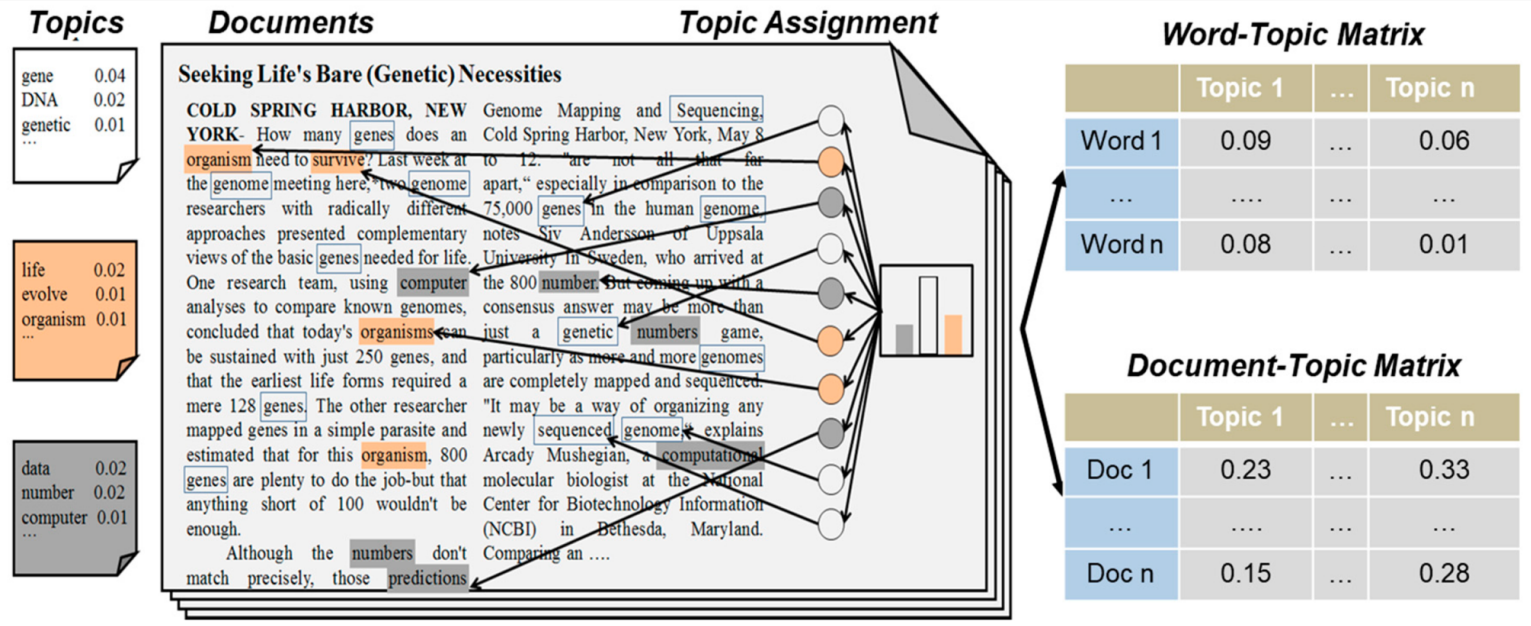
Part 1 topics: covered by Mark

Bayesian statistics: Hierarchical modeling Hospital Variation in Carotid Stenting Outcomes:



Part 1 topics: covered by Mark

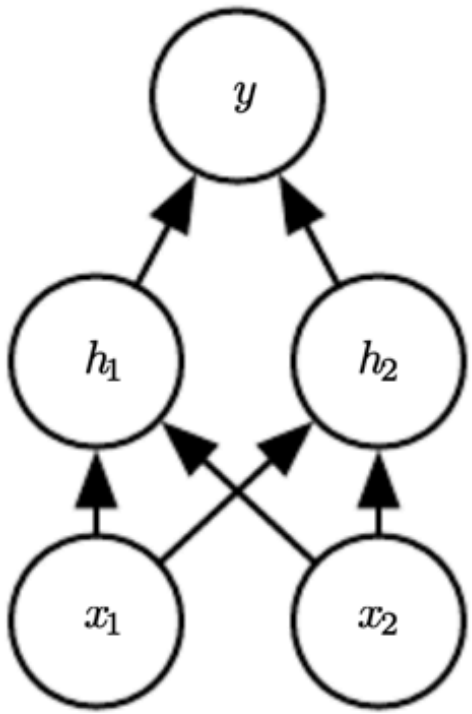
Bayesian statistics: Latent Dirichlet Allocation:



Part 2 topics, covered by Pavlos

- Deep Neural Network
- Neural Net Basics & Math
- Deep Feed Forward
- Regularization
- Optimization
- CNNs
- RNNs
- Autoencoders
- Generative Models and GANs

Feed Forward



$$h_1 = \sigma(w_1^T x + c_1)$$

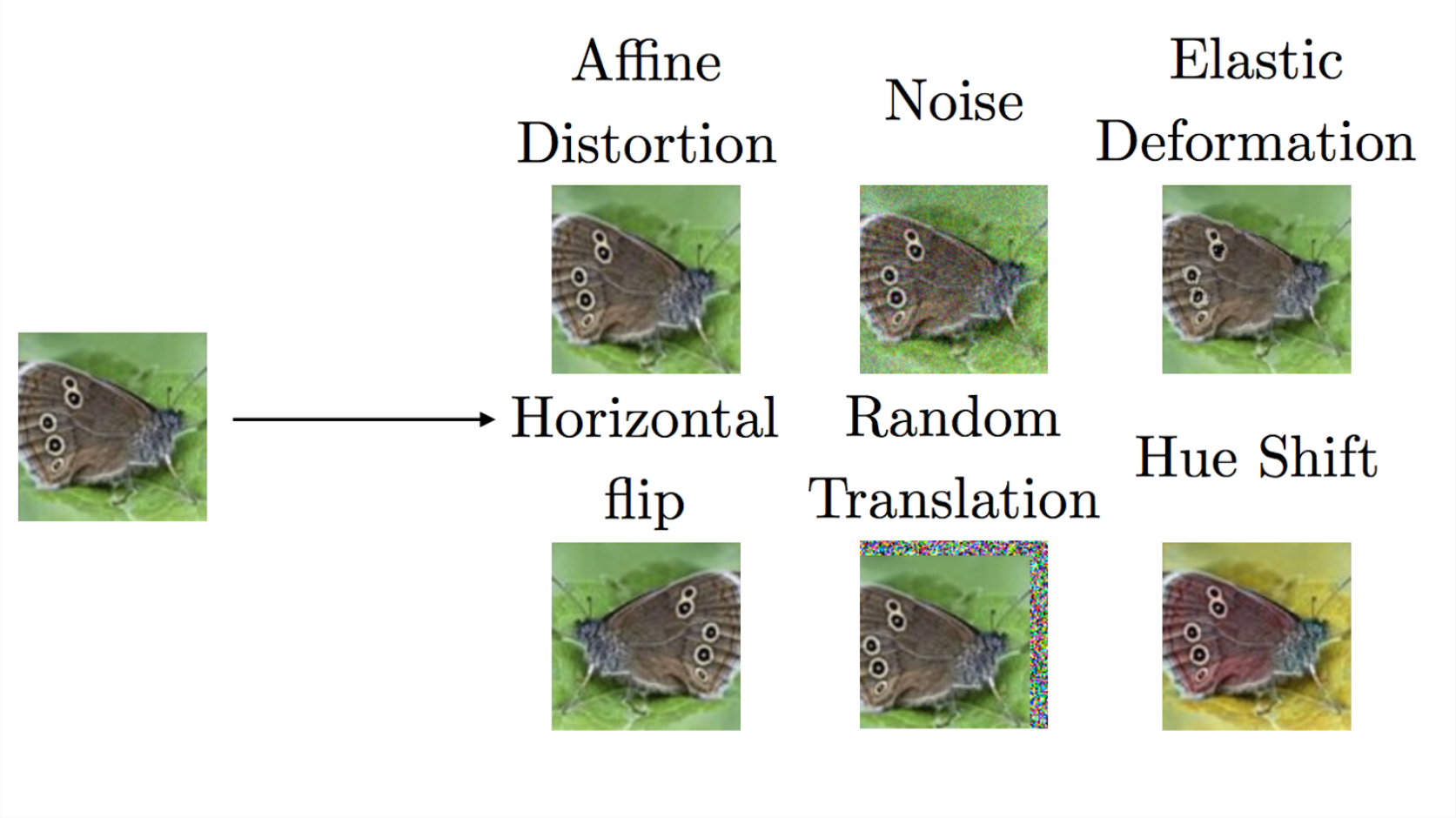
$$h_2 = \sigma(w_2^T x + c_2)$$

$$y = \sigma(w^T h + b)$$

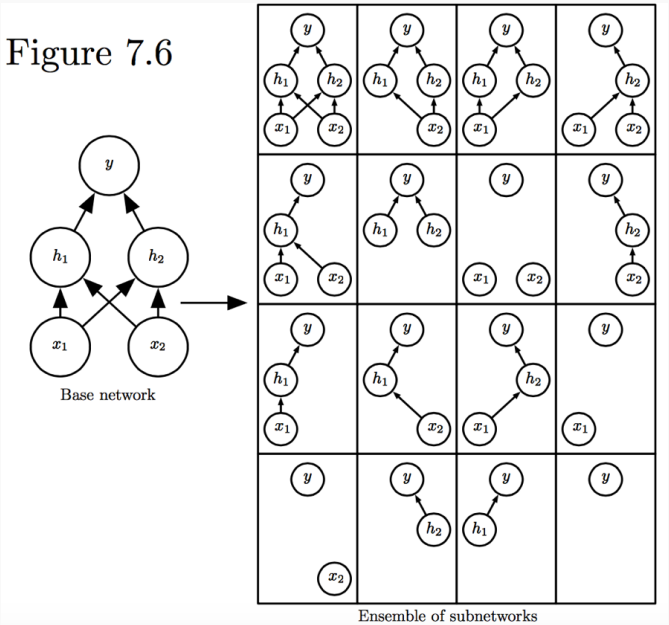
where,

$$\sigma(z) = \max\{0, z\}$$

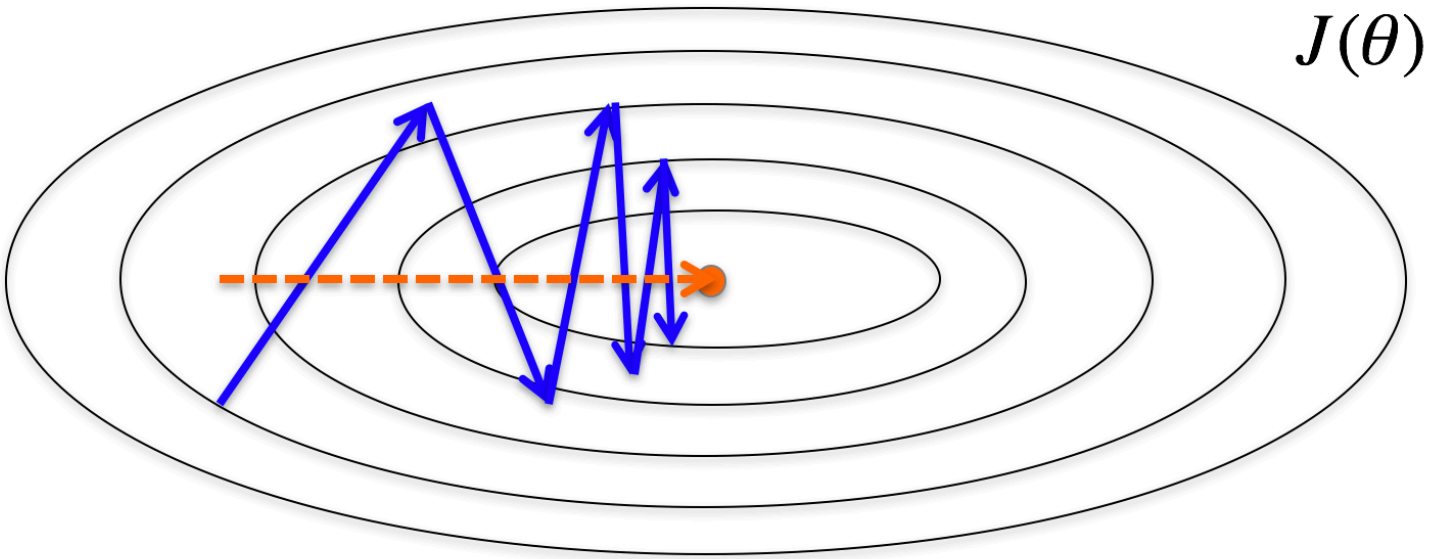
Data Augmentation



Dropout

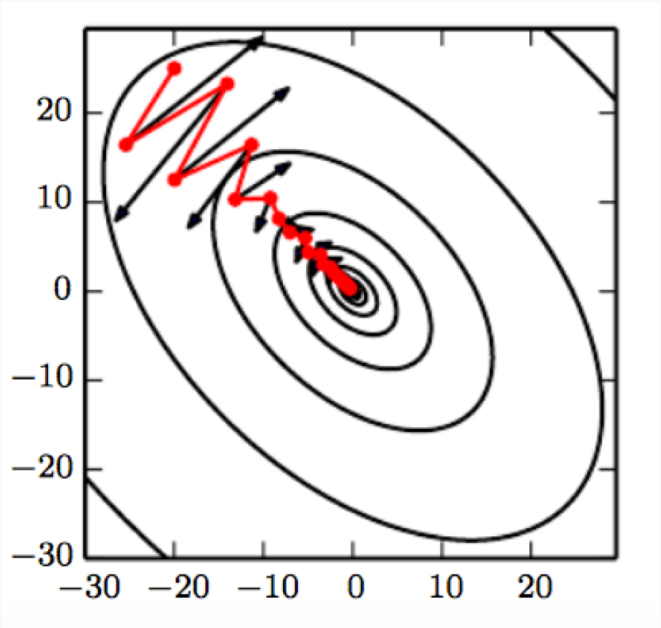


Momentum

- SGD is slow when there is high curvature
- 

- Average gradient presents faster path to opt:
  - vertical components cancel out

Momentum

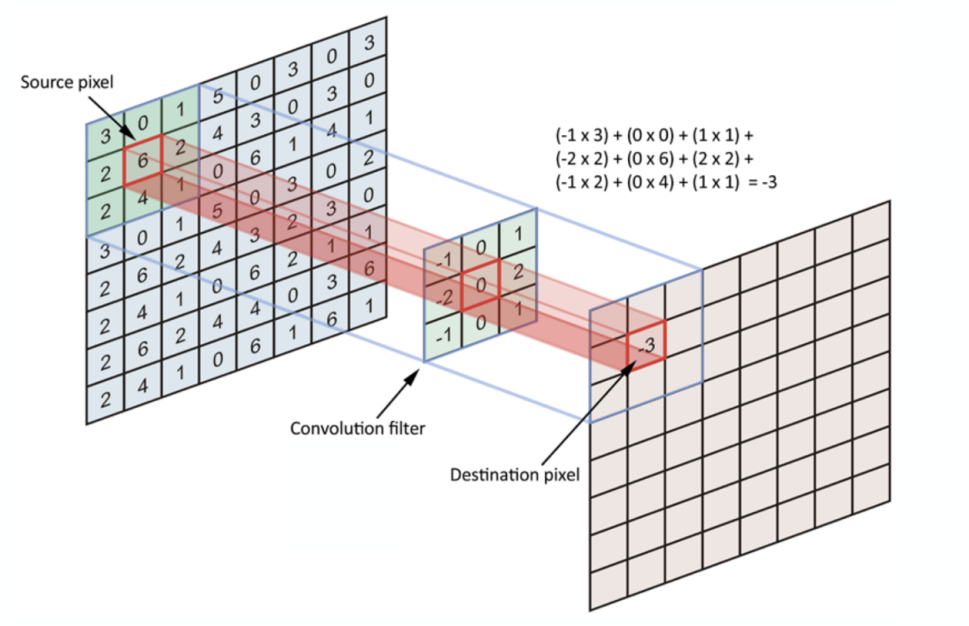


Images are Local and Hierarchical

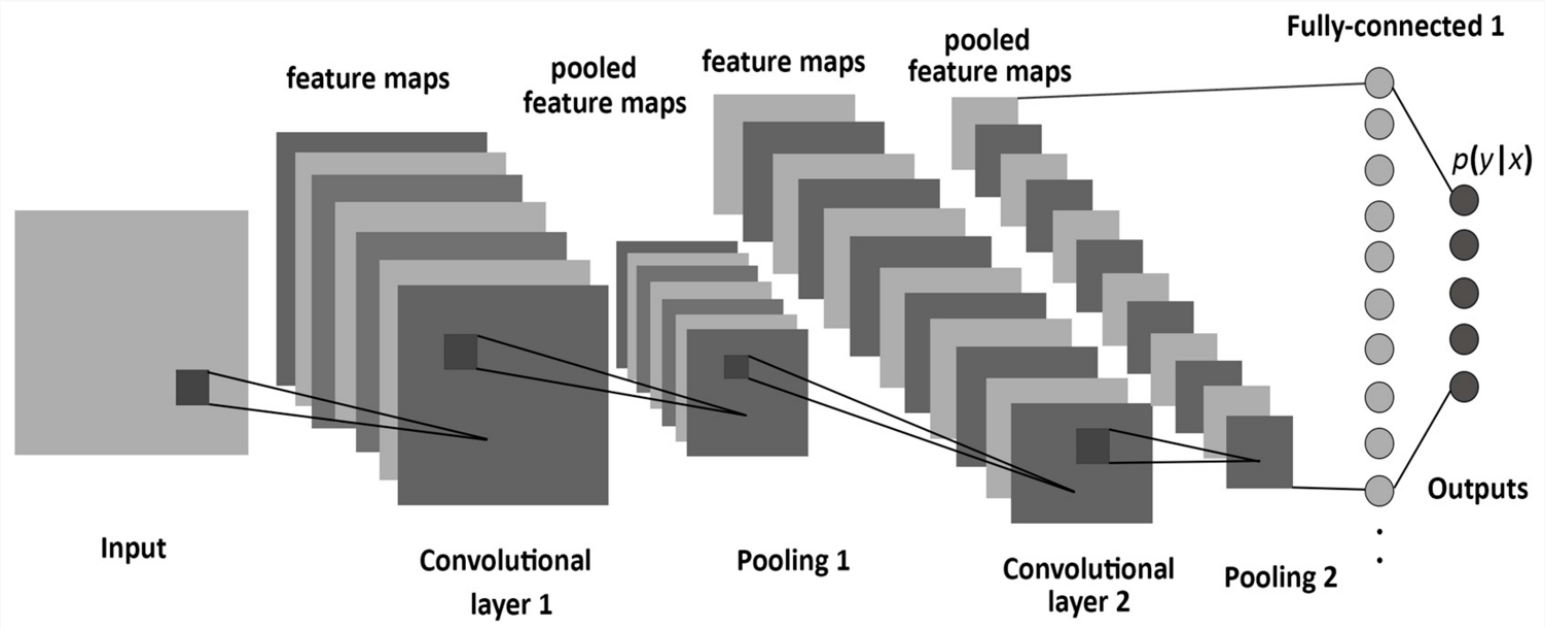
Nearby pixels are more strongly related than distant ones.

Objects are built up out of smaller parts.

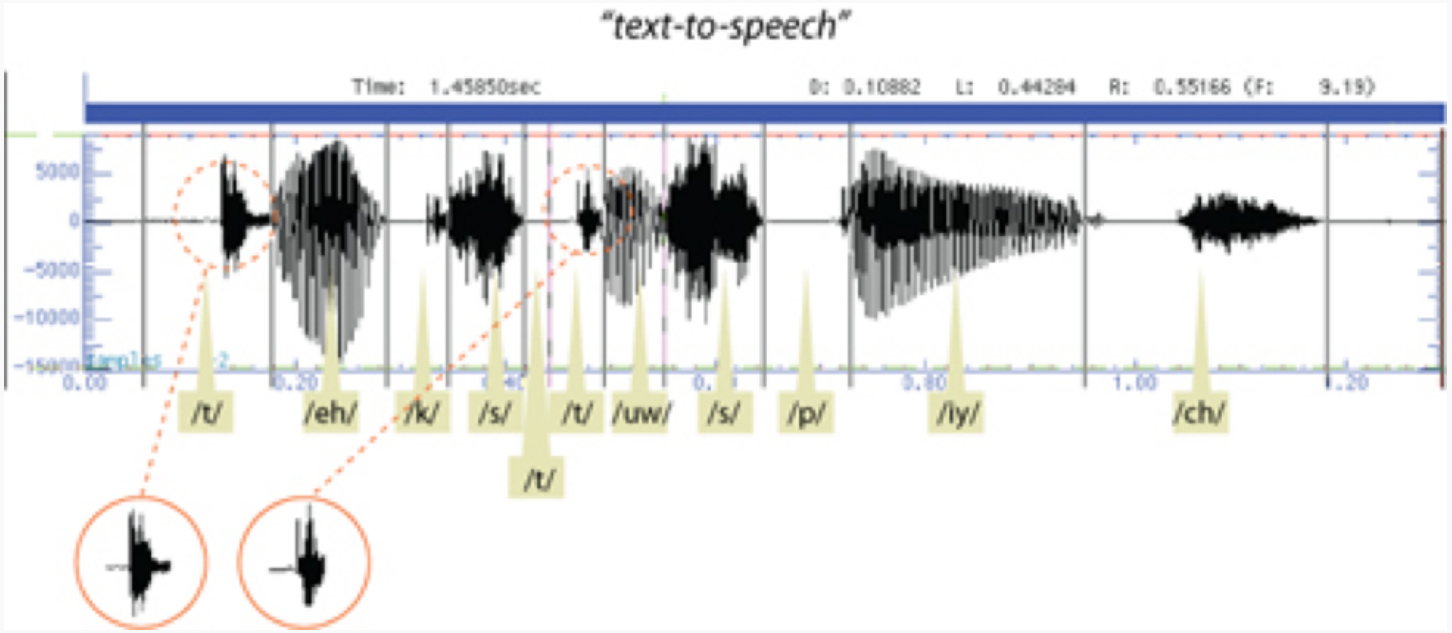
Convolution Operation



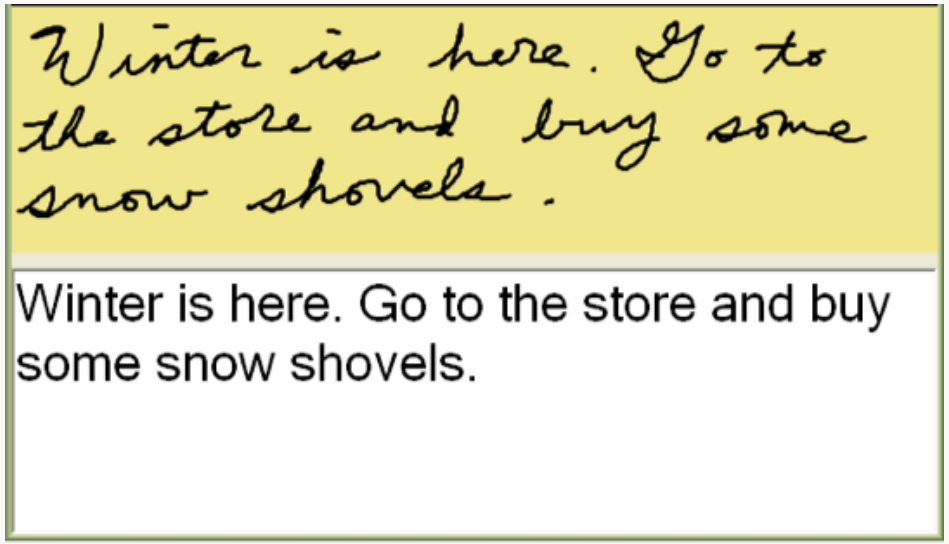
Convolution Operation



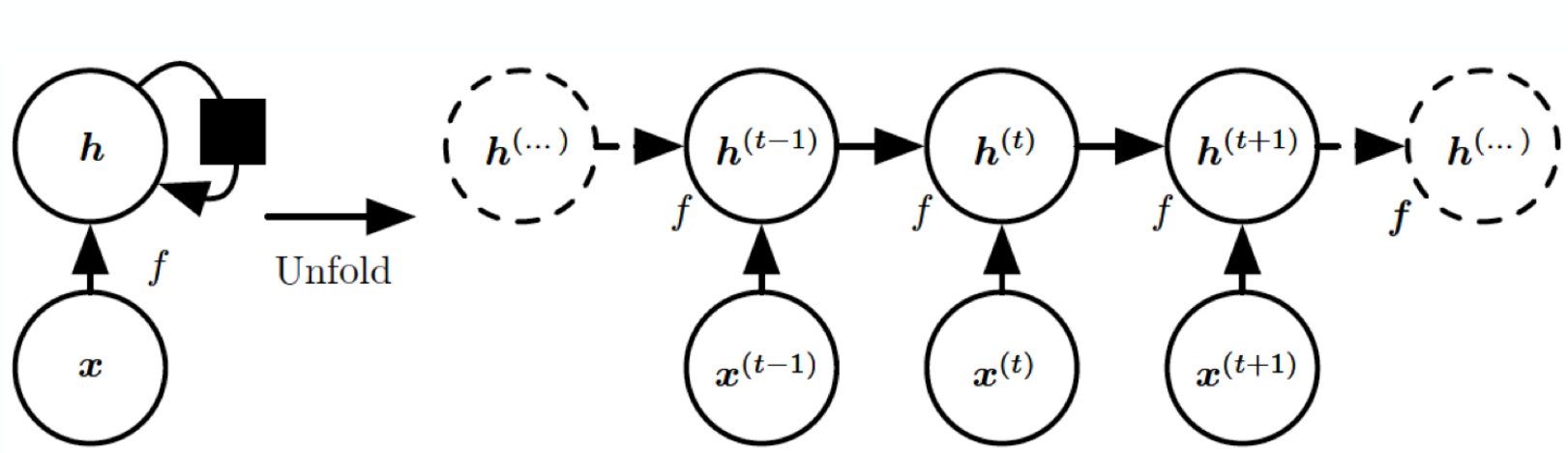
Sequence Modeling



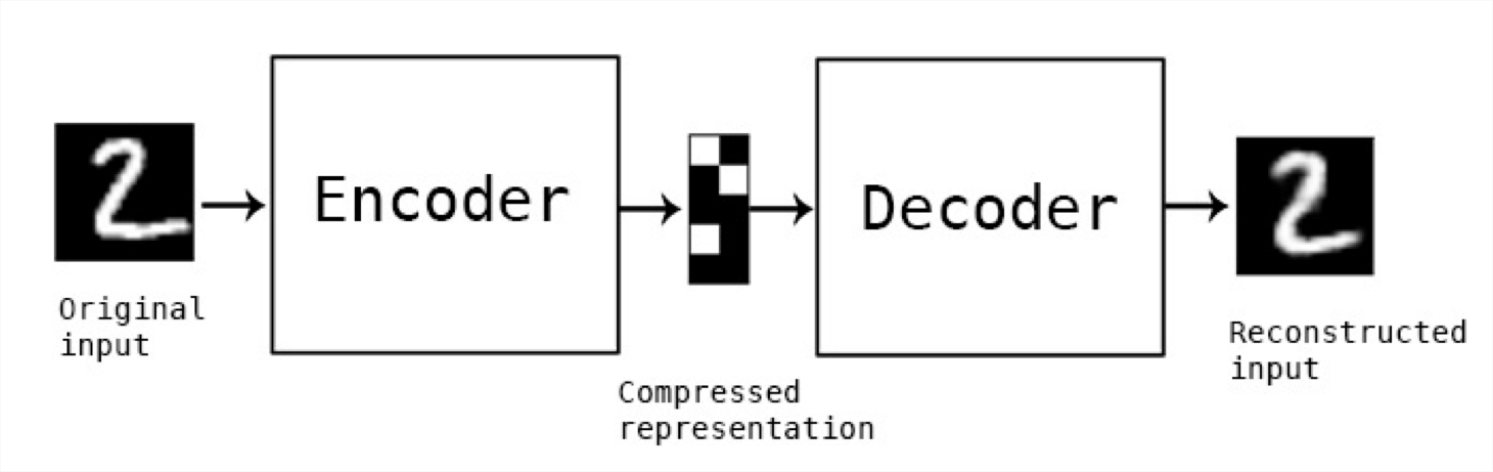
Sequence Modeling



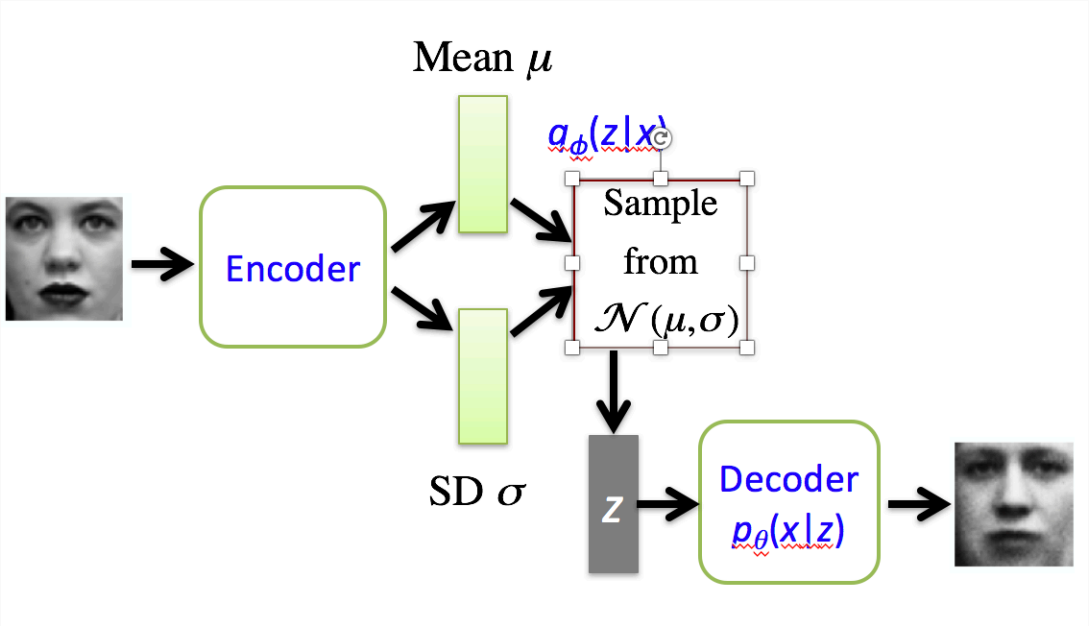
Unfolding the network



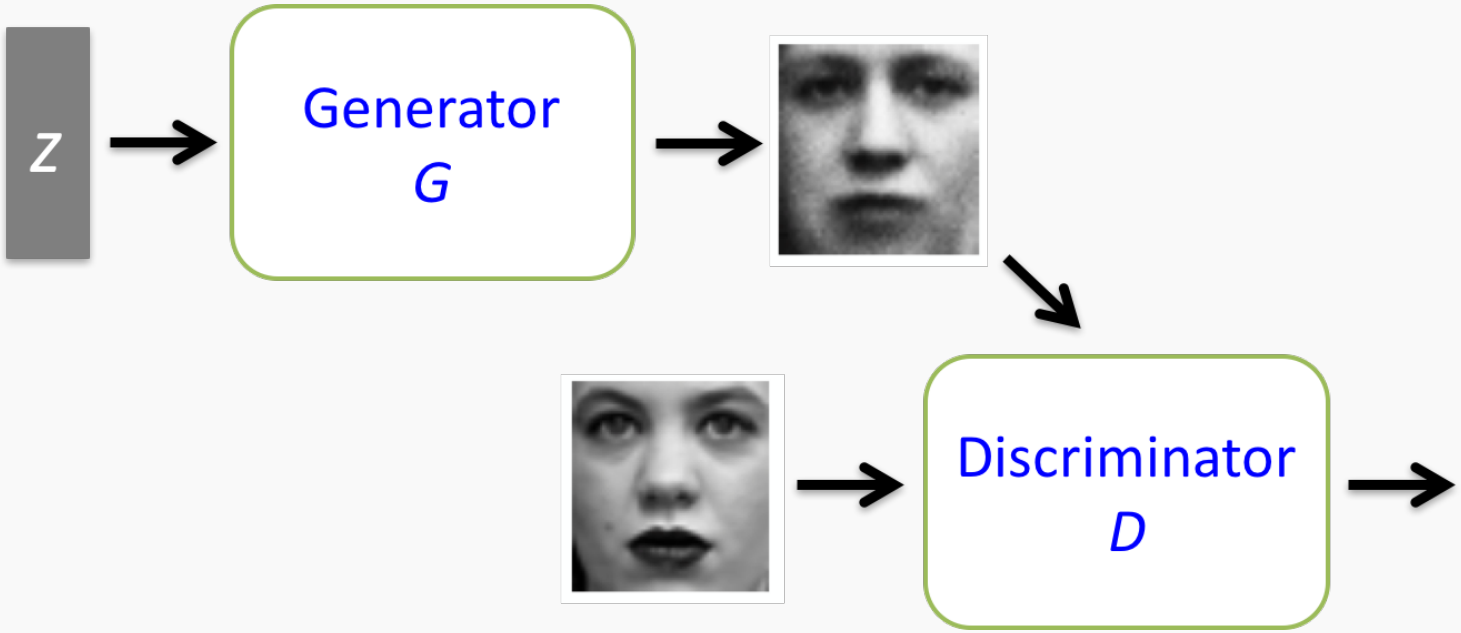
Autoencoders



Variational Auto Encoder Architecture



GAN Overview



Other Topics

- Scalability issues: AWS, Spark, etc
- Databases: SQL etc
- Two guest lectures from industry

Lectures, Labs and Office hours

Lectures: Mondays and Wednesdays 1:00-2:30pm @ Northwest Building B103.
During lecture will cover the material which you will need to complete the homework, midterms and to survive the rest of your life. 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
2. Lectures will be video taped and posted approximately in 24 hours on Canvas

Lectures, Labs and Office hours

Labs: Wednesday 4-5:30pm and Thursdays 4:00-5:30pm, location TBD.
Labs are meant to help you understand the lecture materials better via examples.

1. These two labs will be the same and therefore you need to only attend one of the two
2. Thursday lab will be video taped (and live streamed for DCE students) and posted approximately in 24 hours on Canvas

Lectures, Labs and Office hours

Instructors Office Hours:

- **Mark:** By appointment
- **Pavlos:** Monday 3:00-4:00pm, MD G109

Lectures, Labs and Office hours

TF Office Hours (tentative schedule): Check canvas in the next day or so

AC 209 Students

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

1. Attend A-Sections (5 starting Feb 28)
2. Complete extra questions in homework 1-6
3. Complete extra questions in midterm
4. Expand the scope of the final project

Advance Sections Topics:

1. SVM, Logistic Regression, Preceptor and Feed Forward
2. Dropout, batch normalization and gradient checking
3. Advanced deep neural networks, such as LeNet, AlexNet, VGG-15, Inception and ResNet
4. Neural style transfer learning
5. Deep GANs

Homework(s)

There will be 6 homework (not including Homework 0)

1. Homework 1 on smoothing
2. Homework 2 on clustering
3. Homework 3 on Bayesian
4. Homework 4 on Scalability
5. Homework 5 on Basic DNNs
6. Homework 6 on CNN and RNNs

Homework(s)

Homework(s)

You are encouraged but not required to submit in pairs. We will be using the Groups function on canvas to do this. Instructions on how to submit in pairs are on canvas.
If you work with someone else but not submitting in pair you should indicate that - instructions are also on canvas

All assignments will be posted on Wed. at 11:59pm and will be due Wed. at 11.59pm.

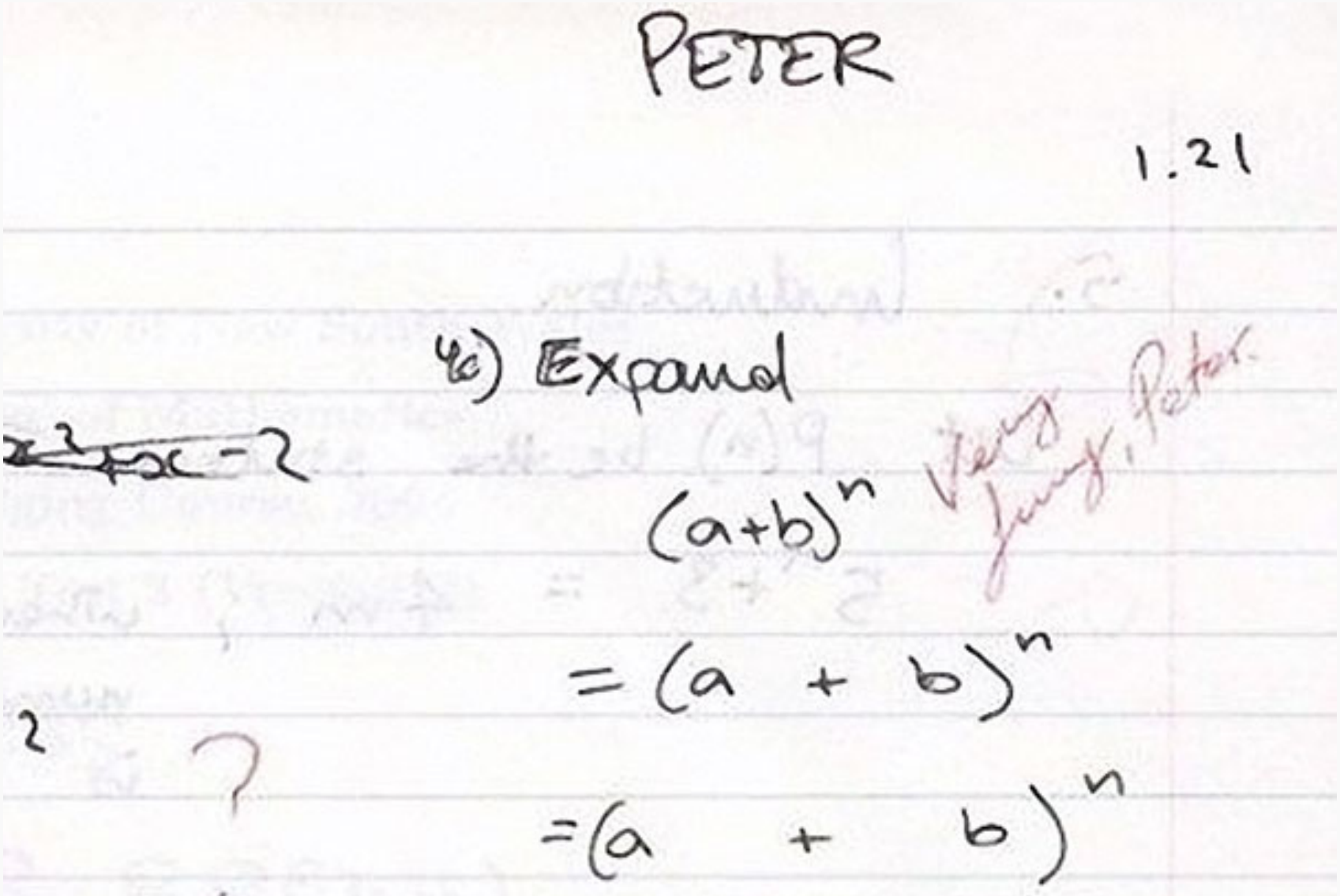
Homework(s)

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.

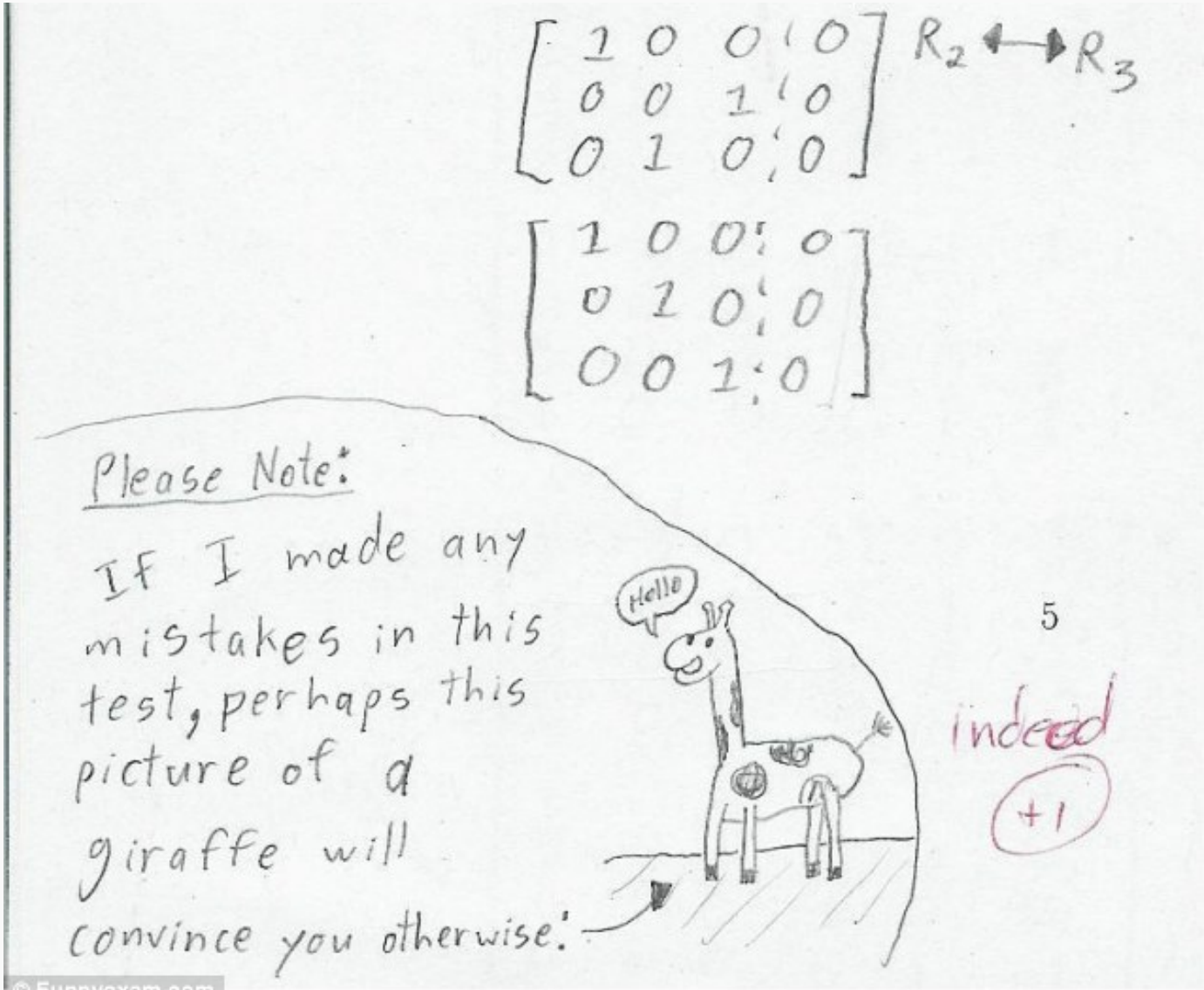
Late Policy: No homework assignments or project milestones will be accepted for credit after the deadline. If you have a verifiable medical condition or other special circumstances that interfere with your coursework please let us know as soon as possible by sending an email to the Helpline.

Homework(s)

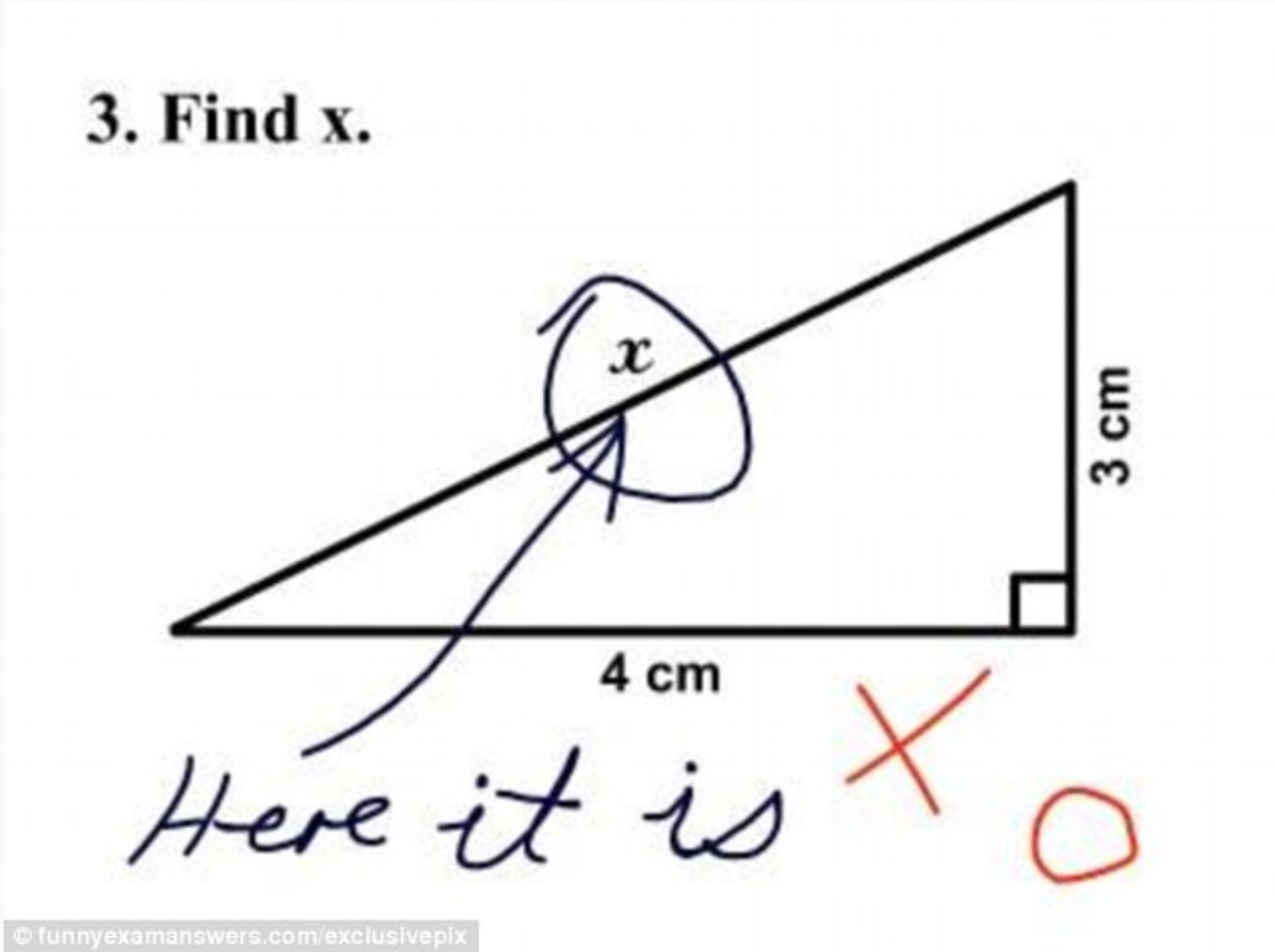
Grading



Grading


$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$ $R_2 \leftrightarrow R_3$
$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$
Please Note:
IF I made any mistakes in this test, perhaps this picture of a giraffe will convince you otherwise.
Hello
5
indeed
+1

Grading


3. Find x.
x
3 cm
4 cm
Here it is
X
O
© funnyexamanswers.com/exclusivepix

Midterm

There will be one midterm (take-home) to be done individually, which counts for 20% of the final grade.

- Released on Mar. 5 @2:30pm (after class), due on Mar. 8 @2:30pm
- 24 hours to complete it
- Extra questions for the AC209 students

Final Project

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

- We will provide approx 5-7 projects with associated datasets which you could use for your final project
- You can propose your project definition
- There will be different expectations for the AC209 students

Final Project Key Dates

- Projects released: Feb 7
- Projects proposals by students due: Feb 21
- Milestone 1, Group formation and sign up: Feb 28
- Milestone 2, EDA and SOW: April 4
- Milestone 3, Literature review and baseline model, April 16
- Final Project Due: May 2

Where to find content

Github: https://github.com/cs109/2018-cs109b
There you will find:

- Lab files and solutions
- Advanced Section materials
- Lectures

Where to find content

Canvas: There you will find:

- Lecture notes under *Lectures*
- Homework and solutions under *Assignments*
- Quizzes under *Quizzes*
- All videos, lectures, labs and a-sections under *Videos*

Everything under *Modules*

Help

Help

The process to get help is:

1. Post the question in Piazza. We monitor the posts but we will respond no later than 24 hours from the posting time
2. Go to Office Hours, **this is the best way to get help**
3. For private matters send an email to the Helpline: cs109b2018@gmail.com, cs-e109b2018@gmail.com for DCE students. The Helpline is monitored by all the instructors and TFs
4. For personal matters send an email to Mark and/or Pavlos

Grade

- Homework 45%
- Quizzes 10%
- Midterm 20%
- Project 25%

Looking forward to a fun and productive semester!

Not like this