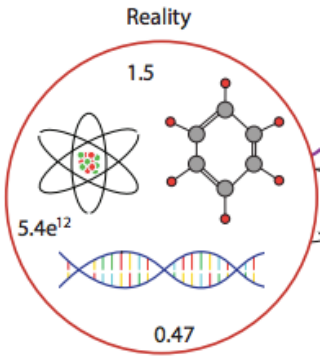
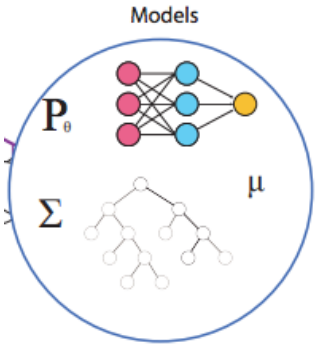


Future reality



Statistics 154, Spring 2019

Modern Statistical Prediction and Machine Learning

Lecture 1: Introduction and pre-requisites

Instructor: Bin Yu
(binyu@berkeley.edu); office hours: Tu: 9:30-10:30 am; Wed: 9:00-10:00 am
office: 409 Evans

GSIs: Yuansi Chen (Mon: 10-12; 12-2); Raaz Dwivedi (Mon: 2-4; 4-6)
yuansi.chen@berkeley.edu; raaz.rsk@berkeley.edu
(office hours to be announced)

Logistics

- Lecture: TU/TH 8:00 am - 9:29 am, VLS 2040
- Lab: Two hours / week
- Homework: biweekly (including math and simulation parts)
- Projects: 2 projects total, in groups of 2
- One midterm + One final (Group 16, May 16, Thurs, 7-10 pm)
- Pre-requisites: math 53, 54; stat 135 (134). R proficiency

- **Grading**

5% attendance + 25% final + 20% midterm + 20% project + 30% homework

- **Policy**

No late homework or late project, no make-up midterm or make-up final

- **Final exam**

Group 16, Thurs, **May 16, 2019 at 7–10 pm**

In-class midterm, Piazza, Gradescope

- One midterm in class on March 21 (Thurs) before spring break

- Piazza (for discussions related to the course)

Search stat154 Spring 2019

https://piazza.com/berkeley/spring2019/stat154/home

- Gradescope (where to submit homework)

Create an account and use the entry code 9J7BJB for the class

Coordination between lectures, labs, hw and projects

- Lecture materials are reinforced in labs, hw, projects and exams
- Lectures show results of computation and visualization, but don't teach R or R commands
- R materials are covered in labs, and reinforced in hw and projects
- Two projects have some open-ended questions to emulate real work situations
- Guided report write-ups are important part of solving real problems

Attendance of lectures and labs

- Attendance is necessary to learn in this class, do hw and projects, prepare and take exams – slides are not self-contained notes and blackboard will be used

- Randomization algorithms will be used to select a date and a random number of students to do roster calls in lectures/labs

154 learning environment

- Supportive, respectful, collaborative, and intellectually honest
- Everyone here is smart, but with different strengths and weaknesses – try to better yourself, while being inspired by and inspiring others in the class
- No question is stupid – getting out of one's comfort zone is how we grow
- Shy? Sitting in the front helps
- Introduce yourself to people around you and say hi next time

Importance of continuous feedback

- Raise your hand in class if things get confusing
- Talk to me before or after class and in office hours
- Talk to Raaz and Yuansi before and after labs, and in office hours
- Ask questions on Piazza
- Both negative and positive comments are helpful

What to do if you are under severe stress?

- Talk to friends and us (Bin, Yuansi and Raaz) after lectures or labs
- Send emails to us
- Send emails and talk to us

Academic integrity

- **Discussions are encouraged with credits attributed appropriately in submitted work**
- **No copying homework or project – this is not learning**
- **GSIs are trained to detect misconducts -- our GSIs have a proven record**

http://sa.berkeley.edu/sites/default/files/Code%20of%20Conduct_January%202016.pdf

Berkeley student code of conduct

- V. GROUNDS FOR DISCIPLINE

The Chancellor may impose discipline for the commission or attempted commission (including aiding or abetting in the commission or attempted commission) of the following types of violations by students (as specified by University Policy 100.00, http://www.ucop.edu/ucophome/coordrev/ucpolicies/), as well as such other violations as may be specified in campus regulations:

- 102.01 Academic Misconduct

All forms of academic misconduct including but not limited to cheating, fabrication, plagiarism, or facilitating academic dishonesty. See Appendix II of this Code for further explanation of academic misconduct.

- 102.02 Other Dishonesty

Other forms of dishonesty including but not limited to fabricating information, bribery, furnishing false information, or reporting a false emergency to the University.

- 102.03 Forgery

Forgery, alteration, or misuse of any University document, record, key, electronic device, or identification

- 102.04 Theft

Theft of, conversion of, destruction of, or damage to any property of the University, or any property of others while on University premises, or possession of any property when the student had knowledge or reasonably should have had knowledge that it was stolen.

- 102.05 Electronic Resources

Theft or abuse of University computers and other University electronic resources such as computer and electronic communications facilities, systems, and services. Abuses include (but are not limited to) unauthorized entry, use, transfer, or tampering with the communications of others; interference with the work of others and with the operation of computer and electronic communications facilities, systems, and services; or copyright infringement (for example, the illegal file-sharing of copyrighted materials).

Use of University computer and electronic communications facilities, systems, or services that violates other University policies or campus regulations.

What to do if you are under severe stress?

- Talk to friends and us (Bin, Yuansi and Raaz) after lectures or labs
- Send emails to us
- Send emails and talk to us

Course description

- This course aims at training students to solve real world prediction problems. We achieve our goal through data experience and training of critical thinking, use of domain knowledge, data visualization, machine learning algorithms and mathematics. We take a holistic view of prediction in the data science life cycle that consists of problem formulation, data collection, exploratory data analysis (EDA), unsupervised learning, supervised learning, data results, validation, and conclusion.

- We separate the real world from the world of mathematics and algorithm, and cover systematic methods for seeking evidence to connect the two worlds (sometimes well but often not). In particular, to help connect the two worlds, we emphasize the concepts PQR-S: P for population, Q for question, R for representativeness and S for scrutiny, and we follow the PCS framework: P for predictability, C for computability and S for Stability.

How to achieve learning under our philosophy – learn ``how to learn"

- **Attendance in class is necessary to do well in the class (5% of grade)**
- **Use of R-studio**
- **Two real data projects in groups of 2**
- **Biweekly homework: math+simulations**
- **Reading selected chapters in textbooks**
- An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics) [2017 Edition] (**required**) (**on-line version available**)

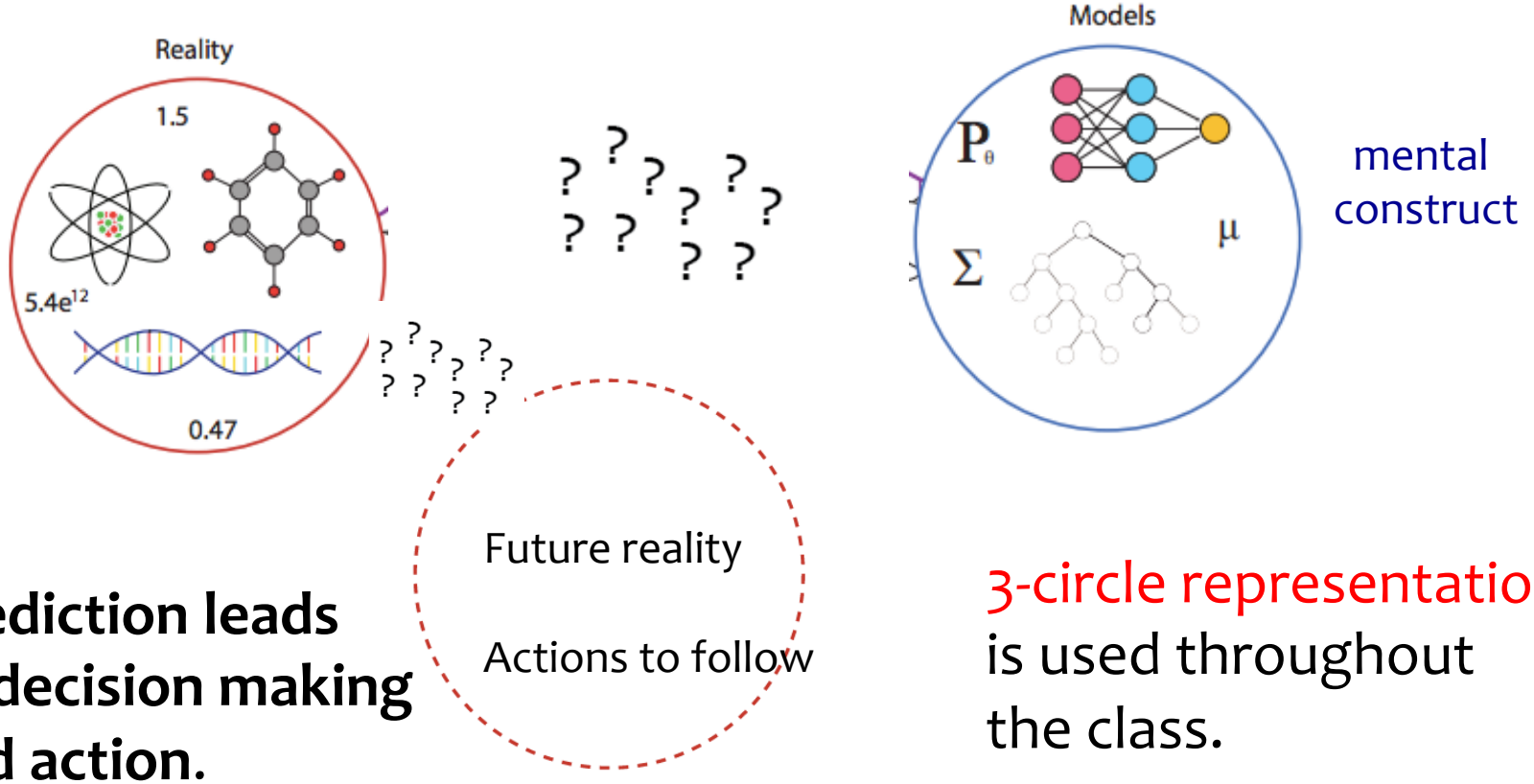
  Authors: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

- Statistical Models: Theory and Practice [2009 Edition]

  (**recommended**) *Author: David Freedman*

Why are we here?

To solve prediction problems in real world by connecting the two solid circles below in a justifiable way to say things about the third circle



mental construct

Prediction leads to decision making and action.

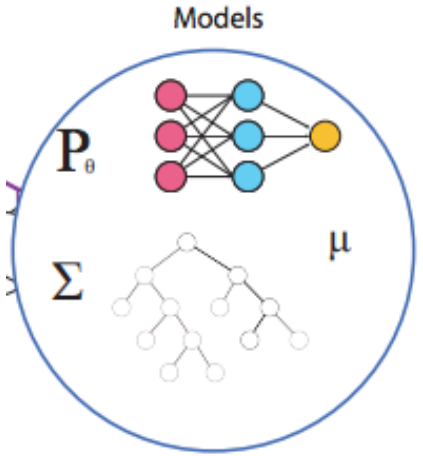
3-circle representation is used throughout the class.

Example 1: 2016 Presidential election prediction

Stats/ML algorithms
- mental constructs

Gallup survey data
on Oct. 30, 2016
of 1000 people

? ? ? ? ?
? ? ? ?
? ?



? ? ? ? ?
? ? ? ?
? ?

Election results
on Nov. 8, 2016.

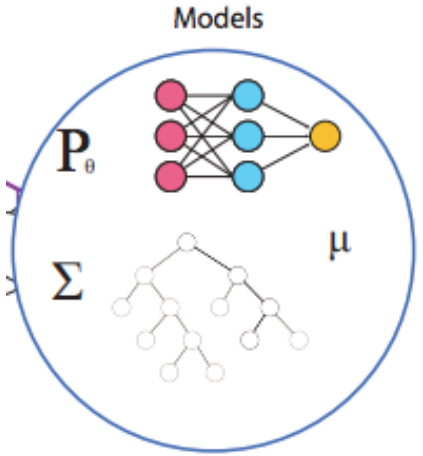
Politics change…

Example 2:
Ames house price prediction for a new comer

Stats/ML algorithms
- mental constructs

House prices in 2018 in Ames, Ohio

? ? ? ? ?
? ? ? ?
? ?



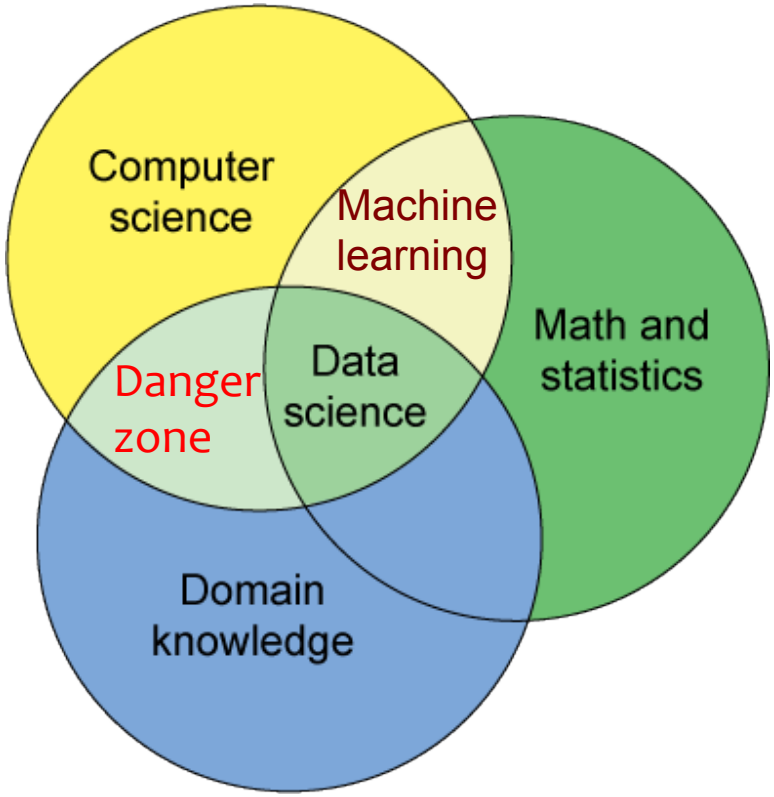
? ? ? ? ?
? ? ? ?
? ?

House prices in 2019 for a new comer.

Whether to purchase

Importance of Stats/ML in data science

Conway's Venn Diagram for DS



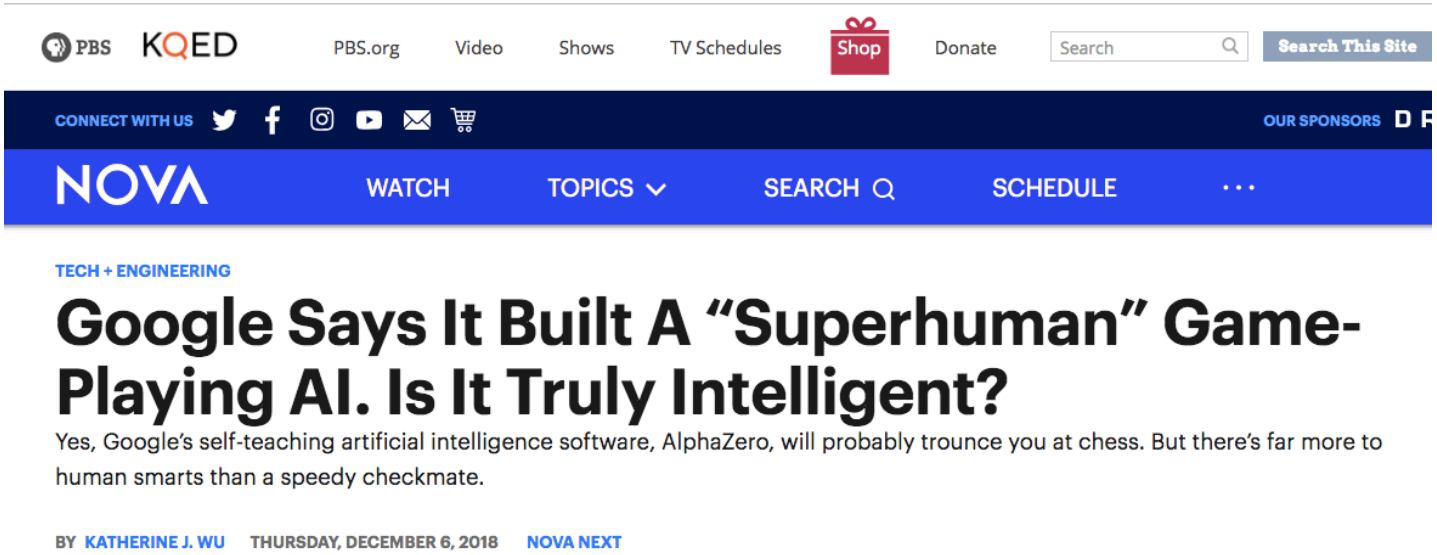
Statistician, Inventor **H. Hollerith**

1890's Hollerith Tabulating Machine

Founding father of modern statistics and statistical genetics, **R. A. Fisher**

Data science is the re-merging of **computational** and **statistical** thinking in the context of domain problems

Stats/ML is also key to the current practice of AI



https://www.pbs.org/wgbh/nova/article/google-alphazero-artificial-intelligence

"The absolute **energy consumption** of AlphaZero must be taken into consideration, adds Bin Yu, ... AlphaZero is powerful, but might not be good bang for the buck—especially when adding in the person-hours that went into its creation and execution."

Y. and Kumbier (2018). "Artificial intelligence and Statistics".

Machine learning (ML): part of statistics and CS

Prediction: part of statistics, which also invented Cross-validation (CV) in the 70's.

First generation ML: **prediction** + **optimization**, with a heavy use of CV

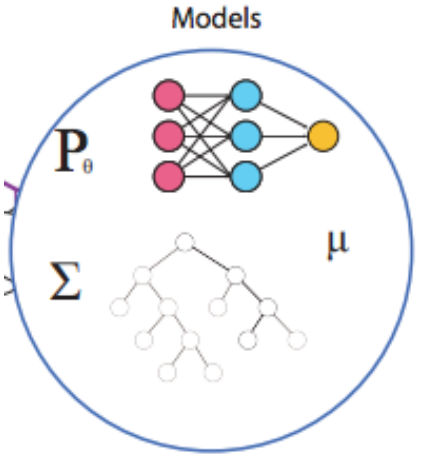
Prediction leads to decision making and action

Reasons for ML success

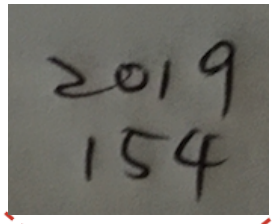
- Prediction and cross validation are both natural and simple conceptually (related to prediction and replication principles in science)
- Prediction based decision making and actions are ubiquitous in real life
- Data availability
- Computing resource availability
- Open-source software

Example 3:
Digit recognition from images of hand-written digits

Mental constructs

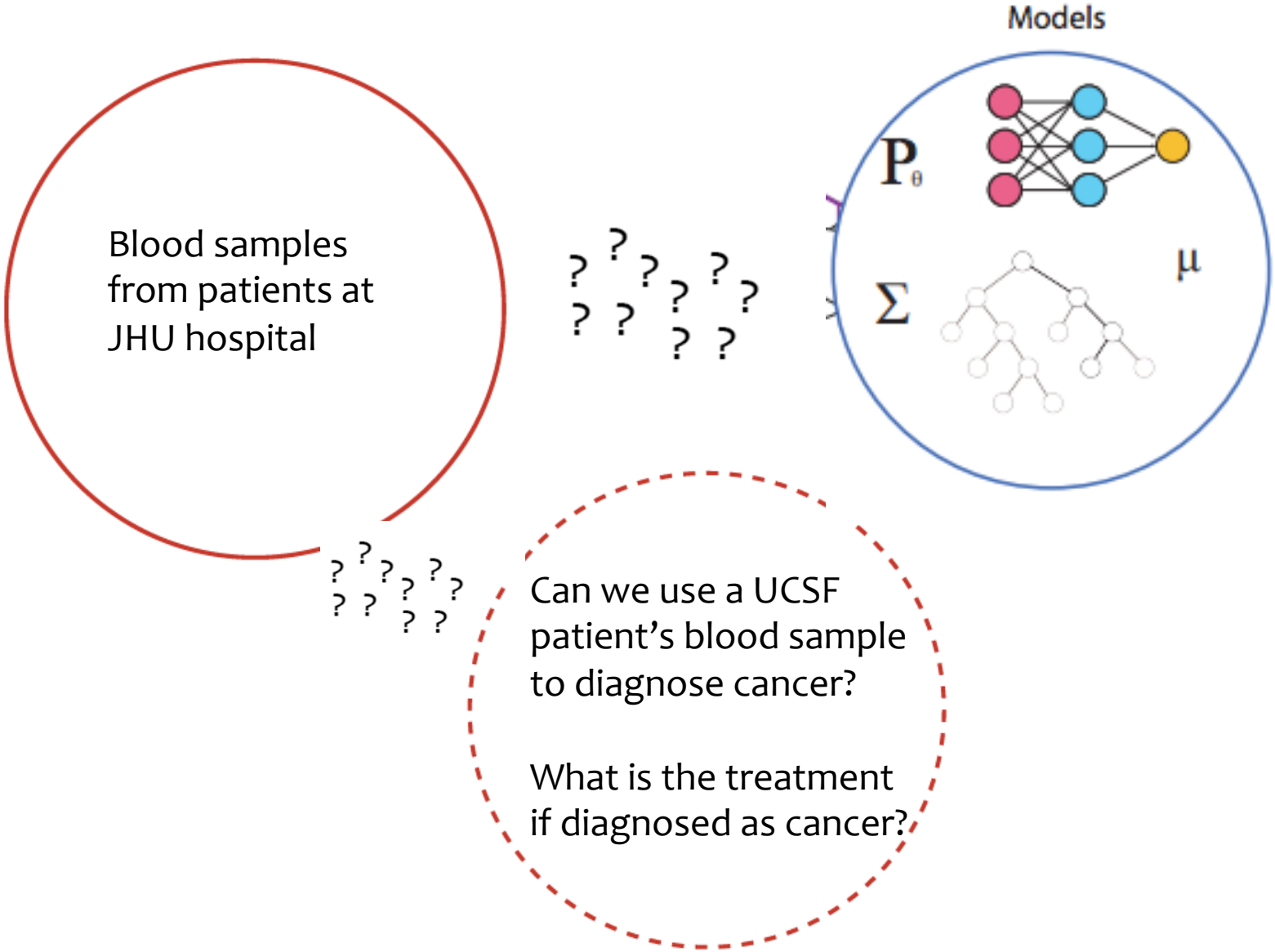


Recognition of Bin's digits



Will she trust digit recognition?

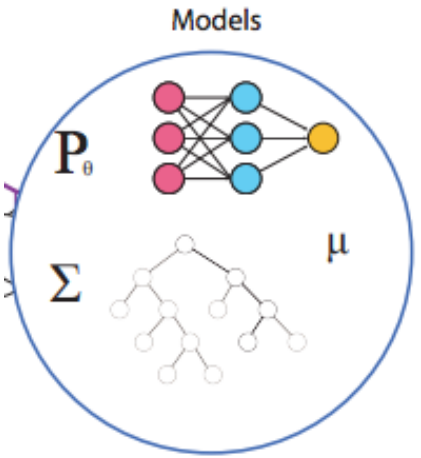
Example 4: early cancer detection



Example 5:
pet image recognition using human-labeled data

? ? ? ? ? ? ? ?

Mental constructs

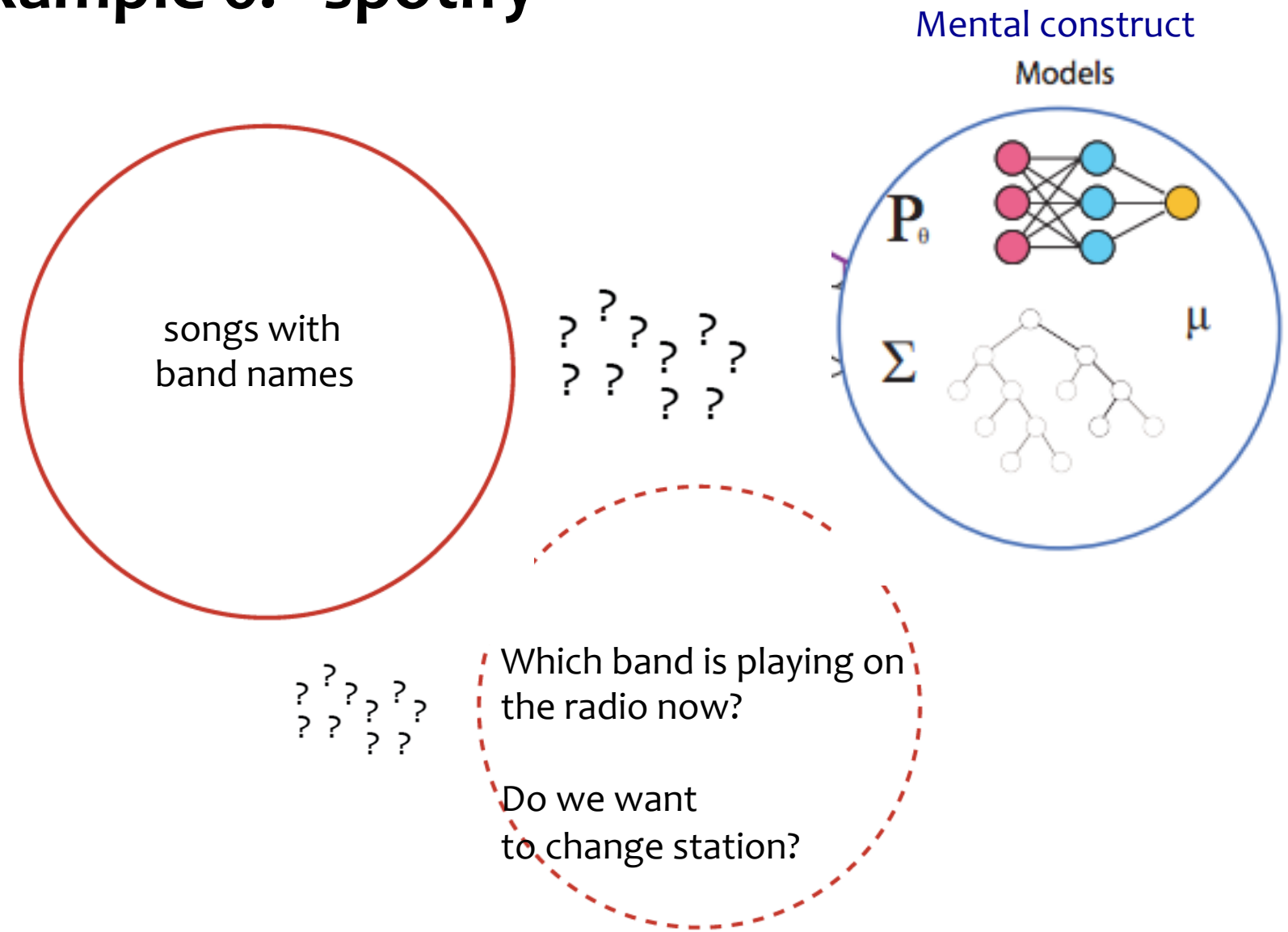


? ? ? ? ? ? ? ? ? ?

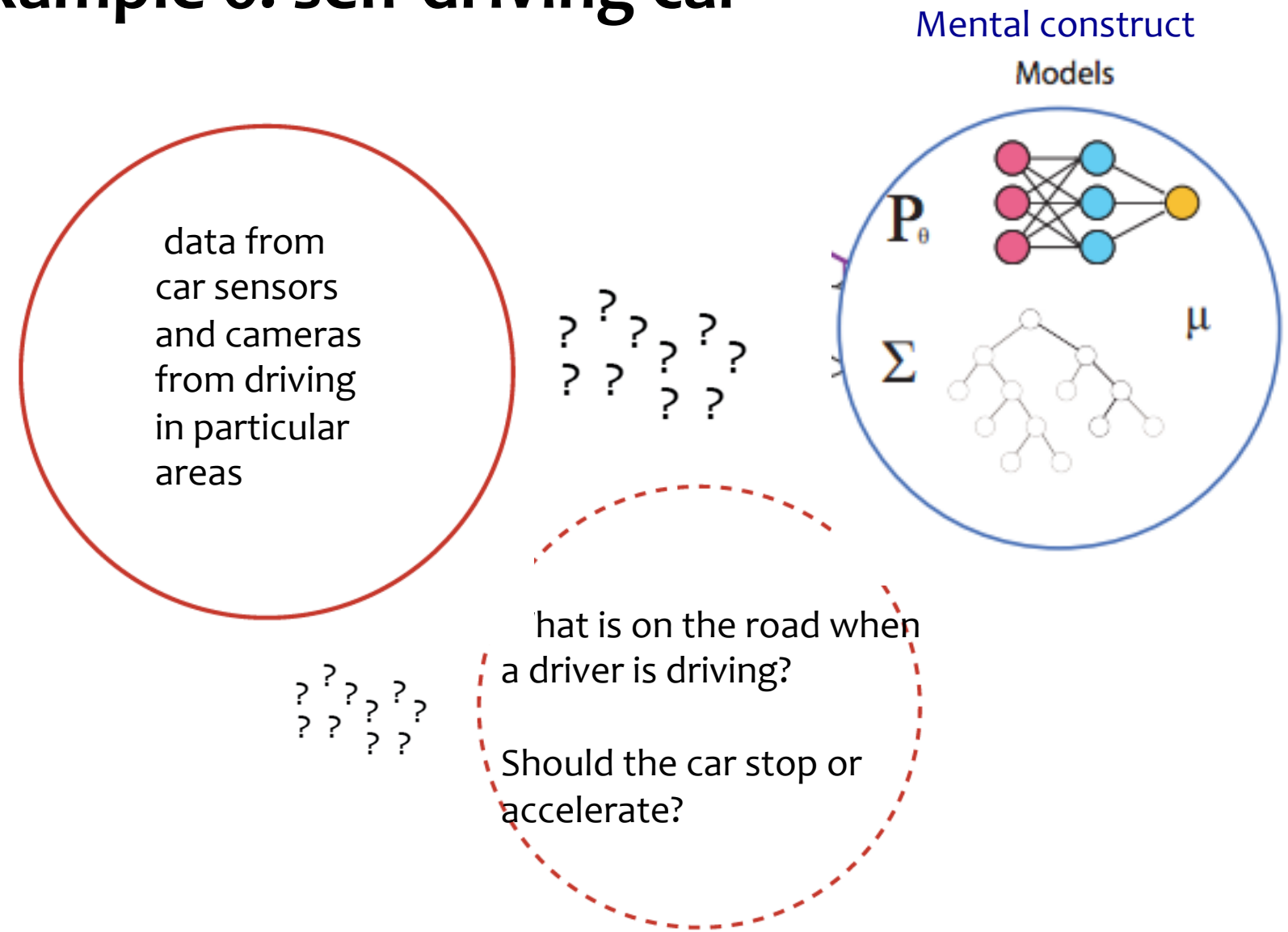
What pets does Bin have?

does she want more?

Example 6: spotify

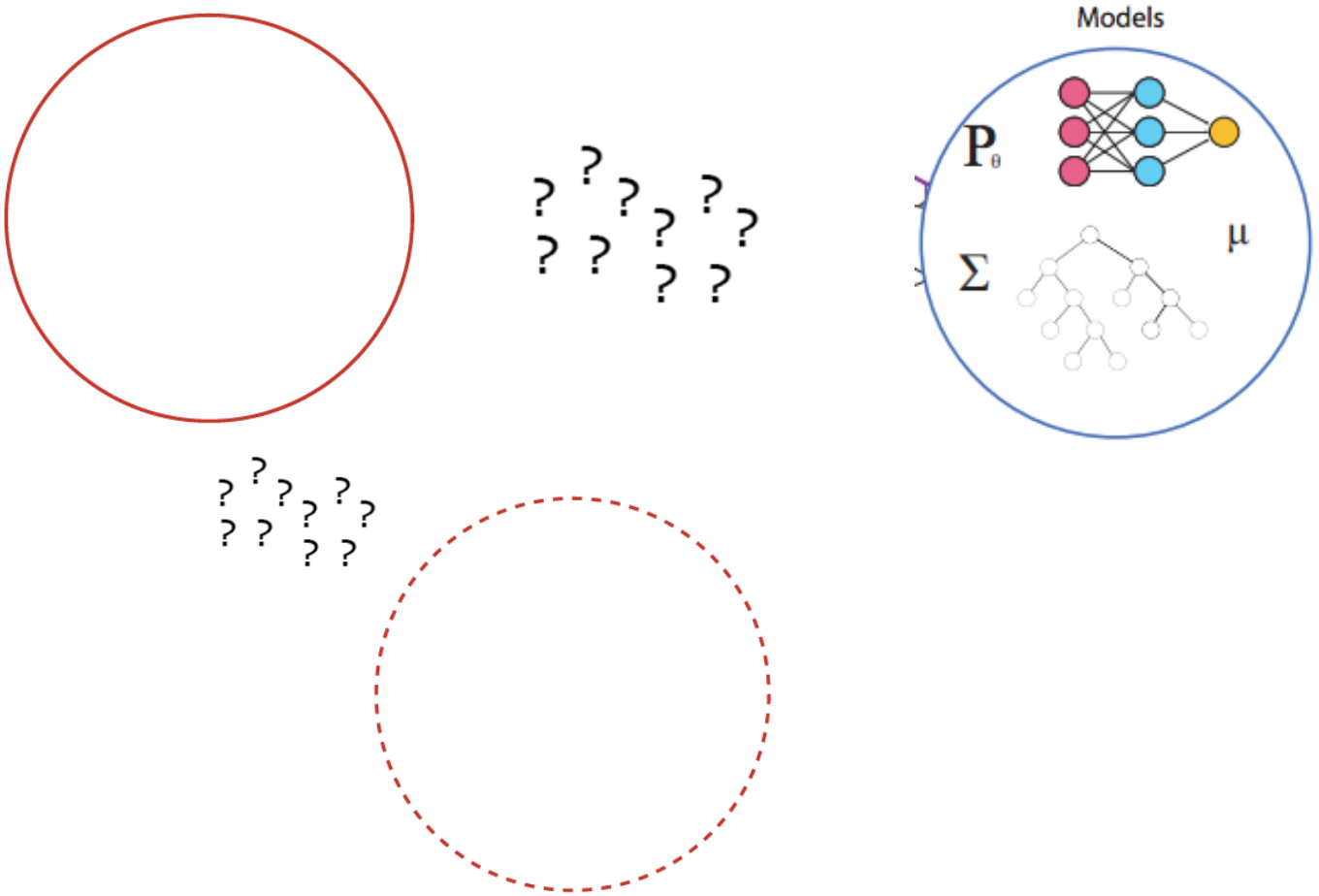


Example 6: self-driving car



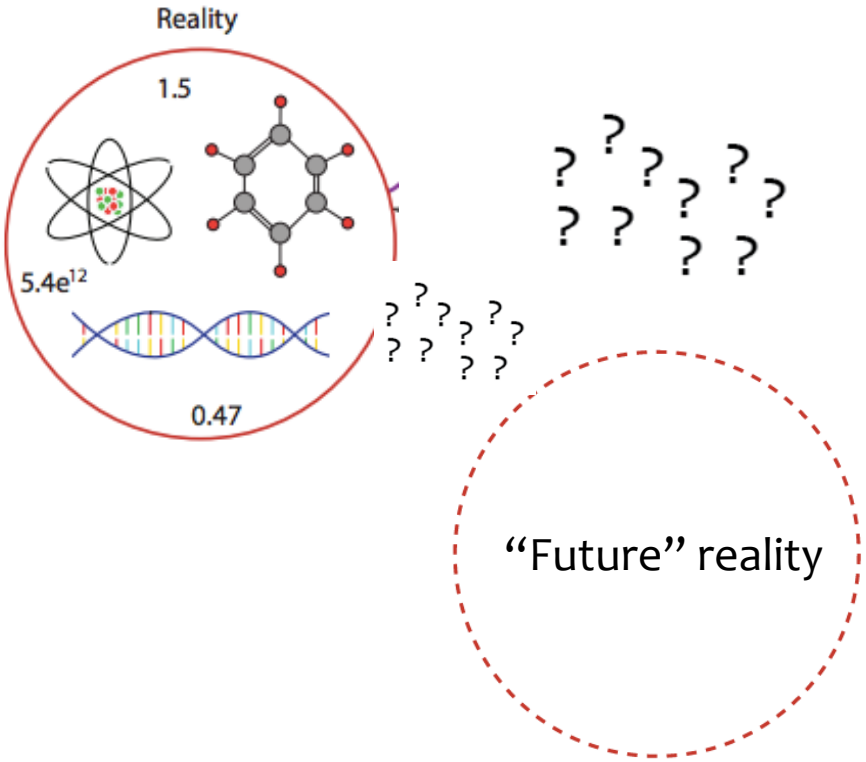
Example 8: class?

Machine learning/Stats algorithms
-- mental constructs

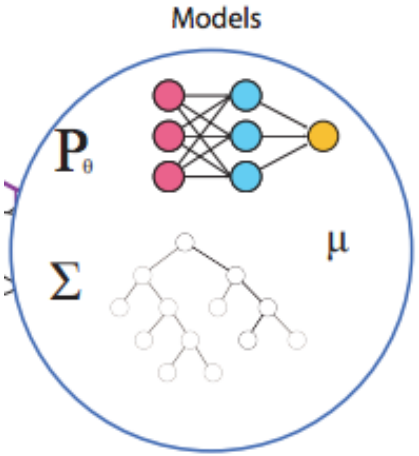


Why are we here?

To solve prediction problems in real world
by connecting the two solid circles below
in a justifiable way to say things about the third circle



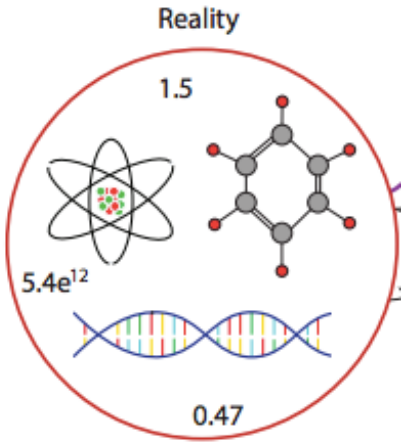
Machine learning/Stats algorithms
-- mention constructs



3-circle representation
is used throughout
the class.

What will you learn?

Future reality



- Problem formulation
- Data collection, data cleaning
- EDA (exploratory data analysis, visualization)
- Unsupervised learning (e.g. PCA, clustering)
- Supervised learning (LS, regularized LS, Kernel regression, logistic regression, Support Vector Machines (SVMs), Nearest Neighbor (NN), Decision trees, Random Forests, Deep Learning)
- Data results, validation, conclusions

mental constructs

Important concepts to frame the solution for connecting justifiably the reality and mental construct (models/algorithms) circles

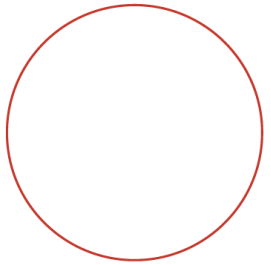
- **PQR-S**: P – population; Q - question; R – representativeness; S– scrutiny

- **PCS**: P – predictability; C – computability, S – stability

Context decides the meaning of P and S – used in both in different ways

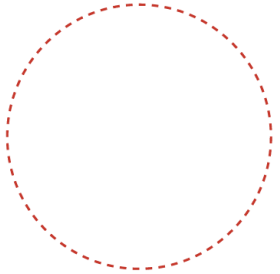
Setting up the prediction problem

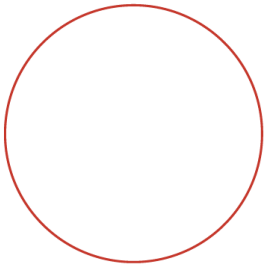
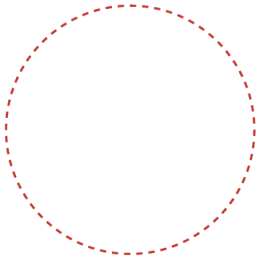
For each data unit (say in a database), we have



- Predictor (feature or covariate) vector
- Response variable
- Prediction error (or prediction performance metric) (often additive across data units, but not necessarily a good idea)

Goal: predict response for new data units (e.g. new patients) with a prediction function



Q: Is  similar to  ?

A: Domain knowledge and/or experience is needed – and ways to evaluate.

Some math notations

Given n data units (or observations) indexed by i :

$$\{(x_i, y_i)\}_{i=1}^{n}$$

x_i predictor vector (or feature, or covariate, or attribute)

y_i response (variable) (continuous or discrete)

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \in R^p$$

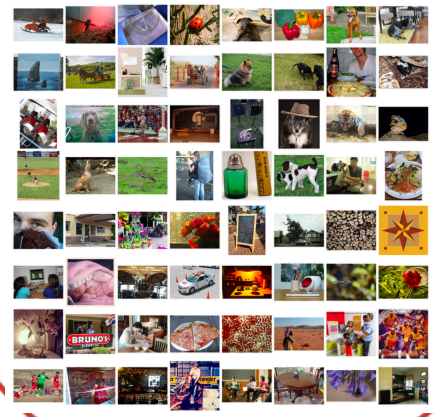
$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}_{n \times p} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in R^n$$

$$\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_p)$$

Example 5:
pet image recognition using human-labeled data

n=1.2 million
p=224*224
=50K

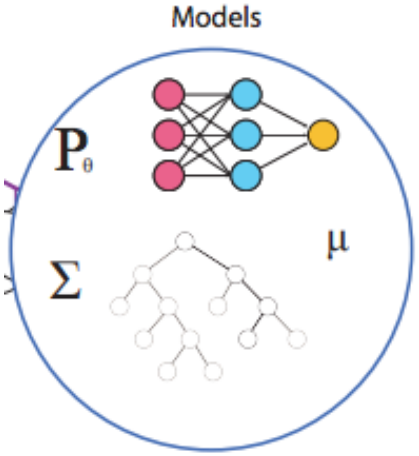
ImageNet



y: discrete or categorical
1000 categories

? ? ? ? ? ? ? ?

Mental constructs



? ? ? ? ? ? ? ? ? ?

What pets does Bin have?

30 sec. meditation …

Prediction function and mean sq. error (MSE)

Ex: $\hat{y}$ = cat

$$prediction\ \ function$$
$$\hat{y} = \hat{f}(x), x \in R^p$$

$$MSE(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - y_i)^2$$

Test MSE
or prediction
error at (x_0, y_0)

$$(\hat{f}(x_0) - y_0)^2$$

why hat on f?

Why additive cross data units or observations?

When is it not a good Idea?

Why sq. error?

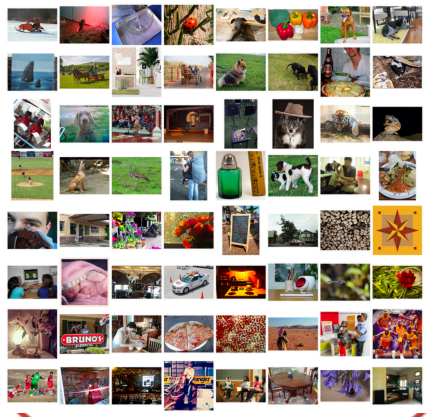
When is it not a good idea?

Prediction error is also called loss function

Example 5:
pet image recognition using human-labeled data

n=1.2 million
p=224*224
=50K

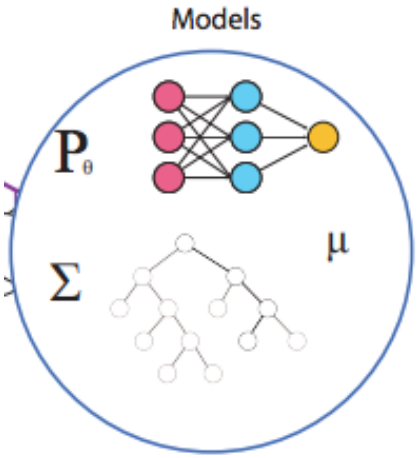
ImageNet



y: discrete or categorical
1000 categories

? ? ? ? ? ? ? ?

Mental constructs



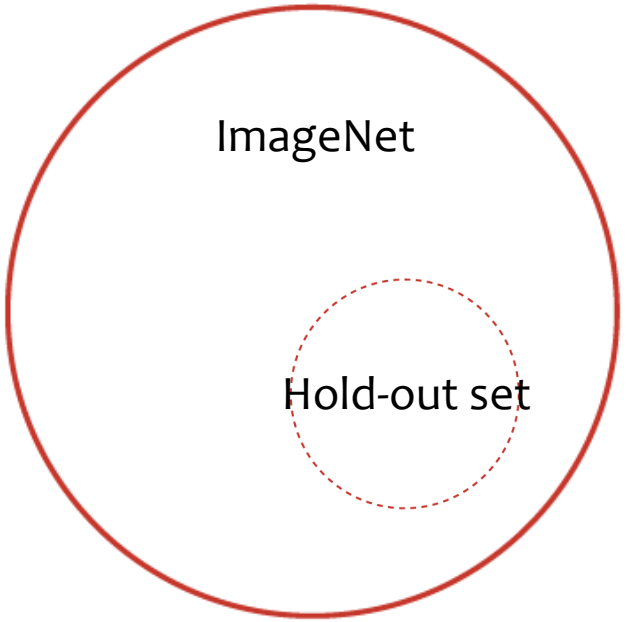
? ? ? ? ? ? ? ? ?

What pets does Bin have?

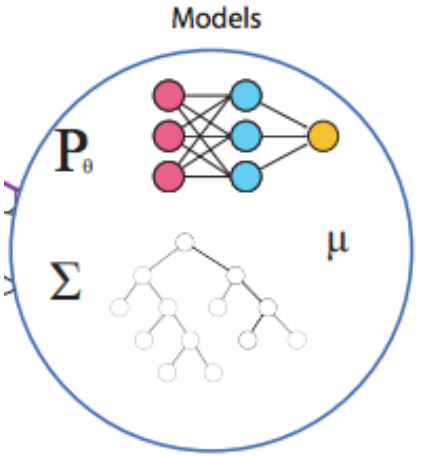
0-1 prediction error
= 0 correct category
1 otherwise

A few jargons

Training data set



Stats/ML algorithms



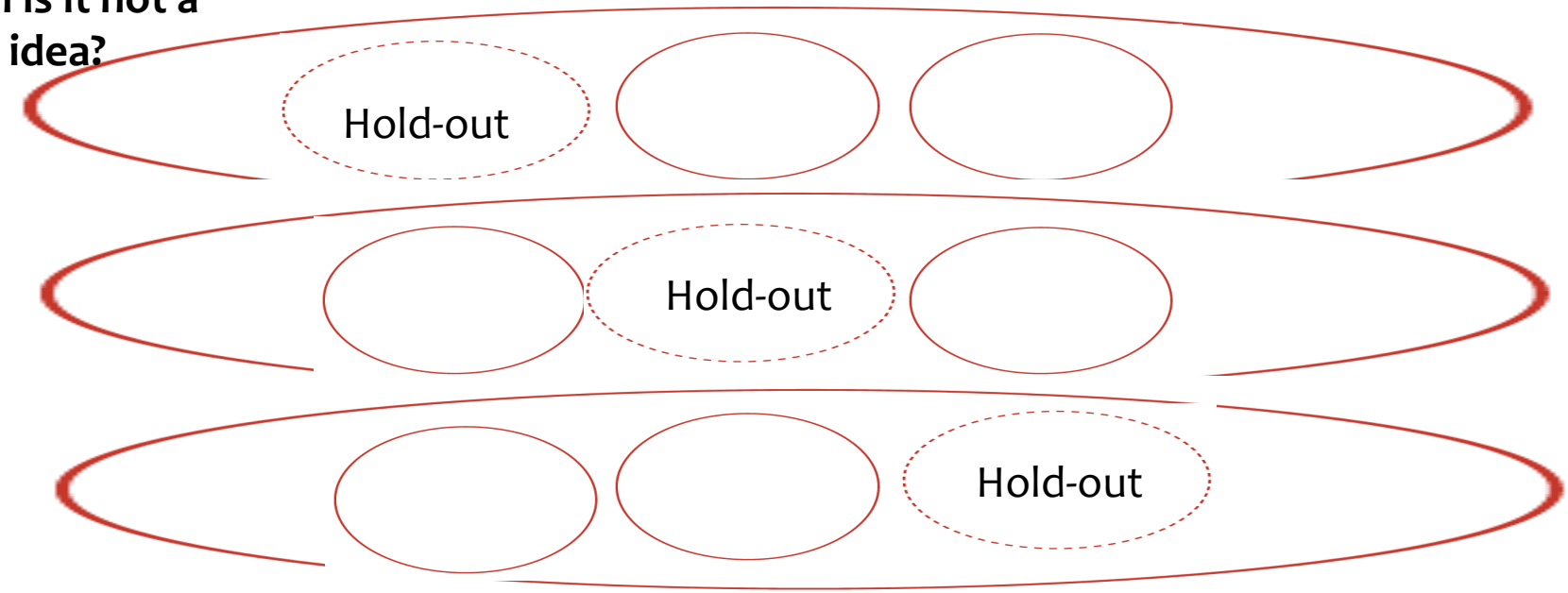
What pets does Bin have?

Cross-validation (CV)

to estimate prediction error within one data set

Given a prediction problem with an "exchangeable" data set, CV creates k "pseudo-replicated" prediction problems or it creates K hold-out sets. K=3 below.

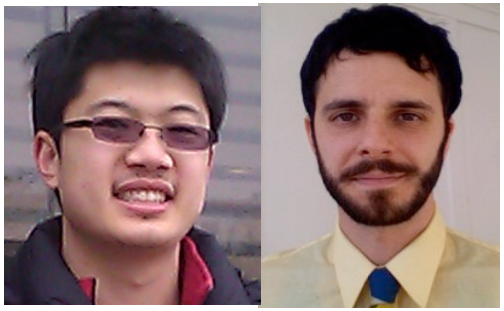
When is it not a good idea?



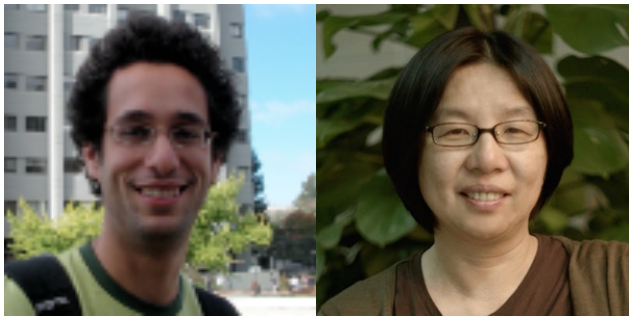
CV prediction error is the average over k-fold
(not always a good estimate of the pred. error)

Success of ML from Gallant/Yu labs: "mind-reading"

S. Nishimoto **J. Gallant**



A. Vu T. Naselaris



Yuval Benjamini Bin Yu

Current Biology 21, 1641–1646, October 11, 2011 ©2011 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2011.08.031

Report

Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies

Data collection
(the Gallant Lab)

fMRI

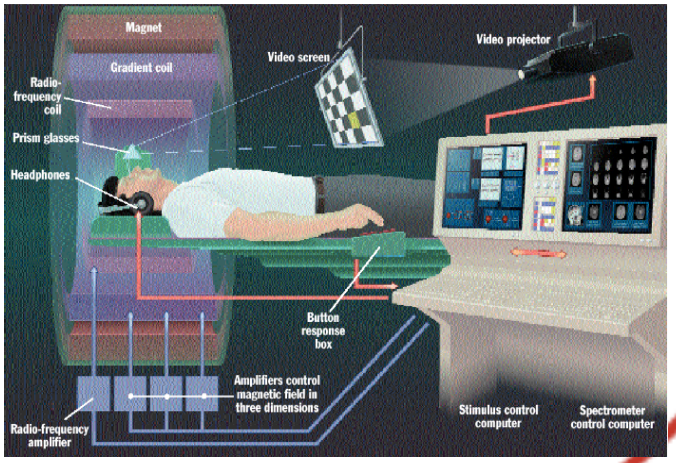
- **Non invasive and indirect recording technique**
- Low temporal resolution, a few seconds
- High spatial resolution
  (**voxel = 1x1x1 mm cube**)
  - 10,000 voxels in early visual areas
  - each voxel covers > 100,000 neurons

Data:

input movie clips

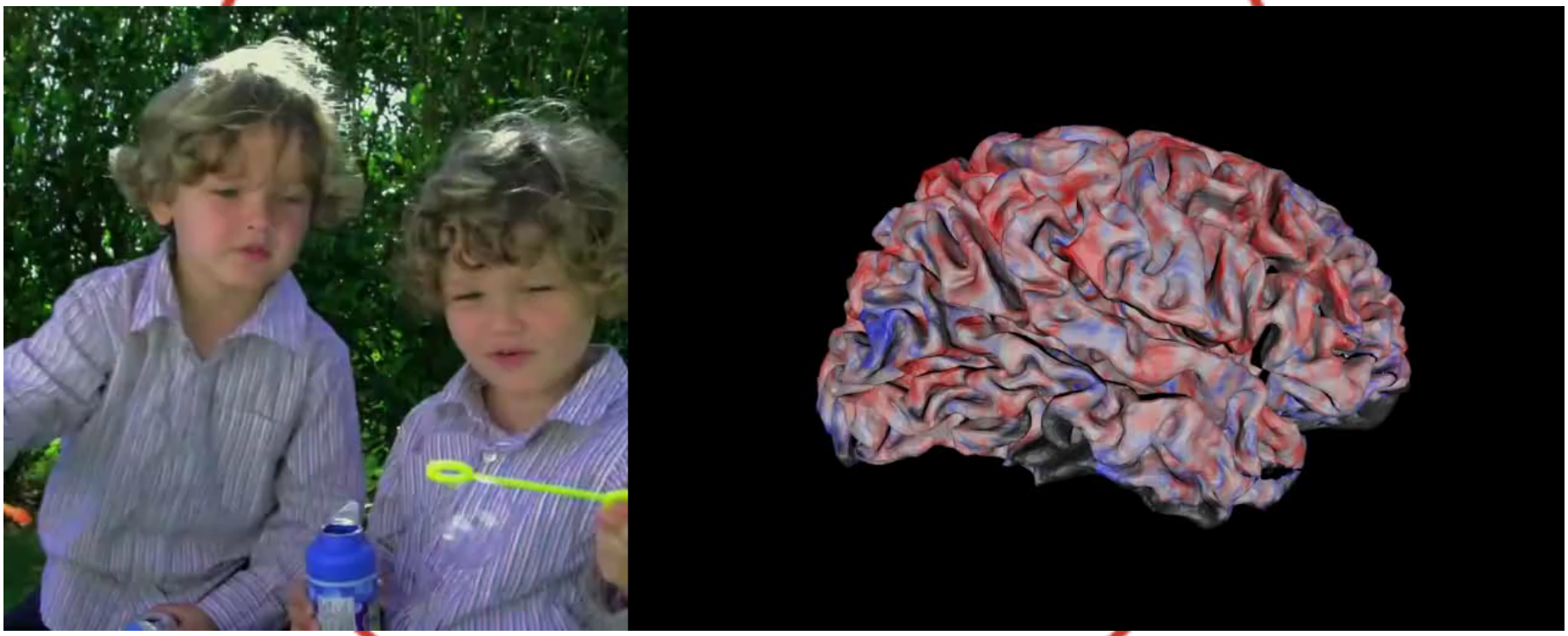
and corresponding fMRI

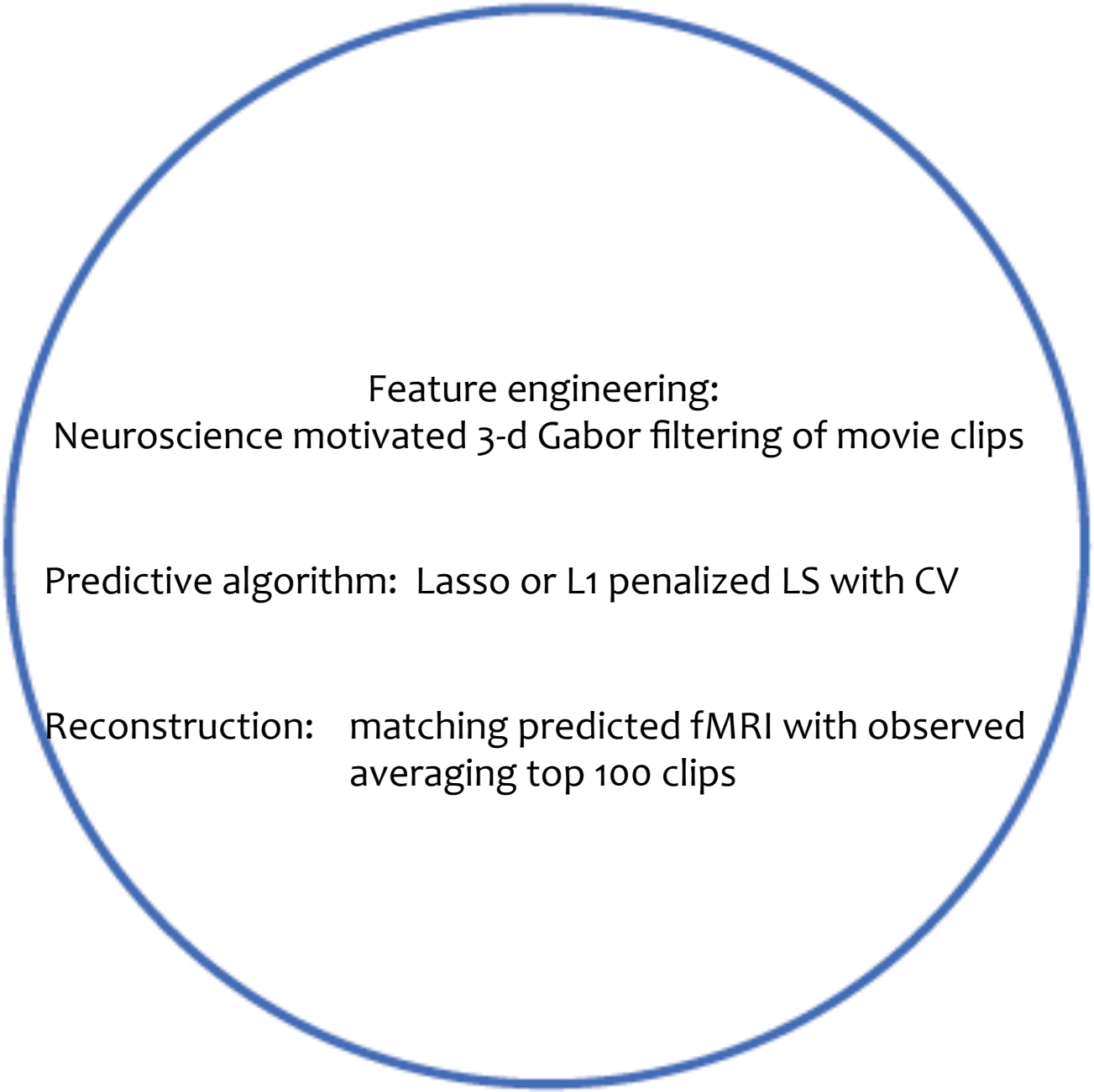
brain signals of subjects

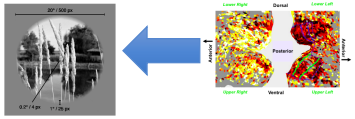


Our movie-fMRI data

7200s training (1 replicate) and 5400s test (10 replicates).




Feature engineering:
Neuroscience motivated 3-d Gabor filtering of movie clips
Predictive algorithm: Lasso or L1 penalized LS with CV
Reconstruction: matching predicted fMRI with observed
averaging top 100 clips



Movie reconstruction through ML

Nishimoto, Vu, Naselaris, Benjamini, Yu and Gallant (2011)

Presented clip

Clip reconstructed from brain activity

Math pre-requisites
I. Matrix

- Matrix operations
- Matrix inversion
- Matrix single value decomposition (SVD)

Math pre-requisites
II. Multivariate gradient

- Multivariate derivatives
- Multivariate Taylor series

Math pre-requisites
III. Probability theory

- Special distributions: Gaussian (univariate and multivariate), Laplacian, t-distribution, Bernoulli, Binomial, Poisson, etc

- Expectation operations, variance, covariance, correlation, etc

- Central Limit Theorem (CLT)

Computing pre-requisites

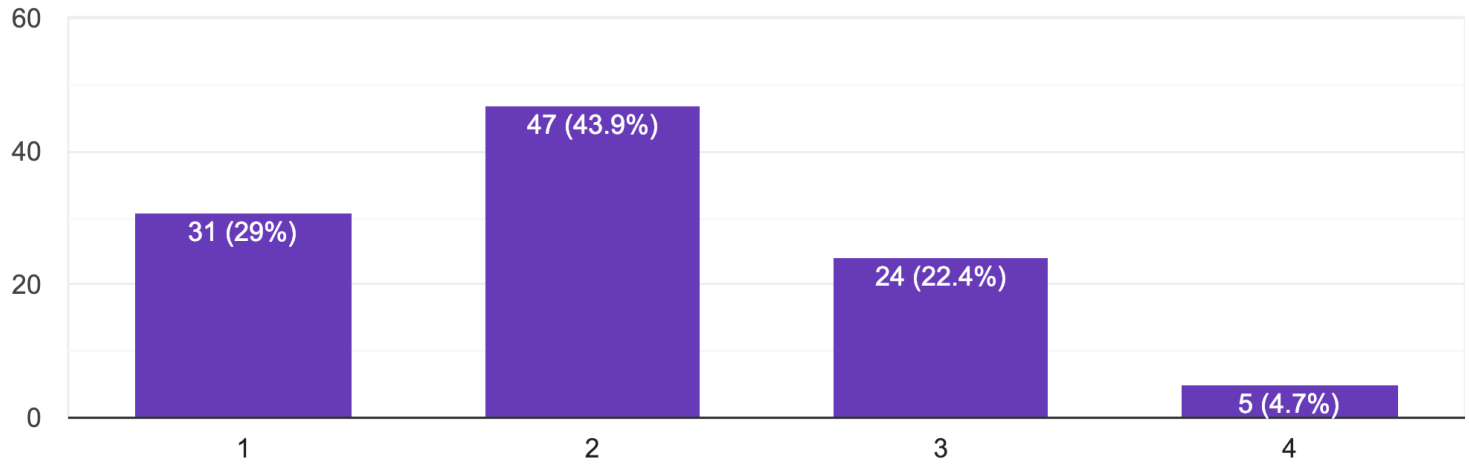
- R-studio

Survey results

Thanks for filling it out with a short notice!

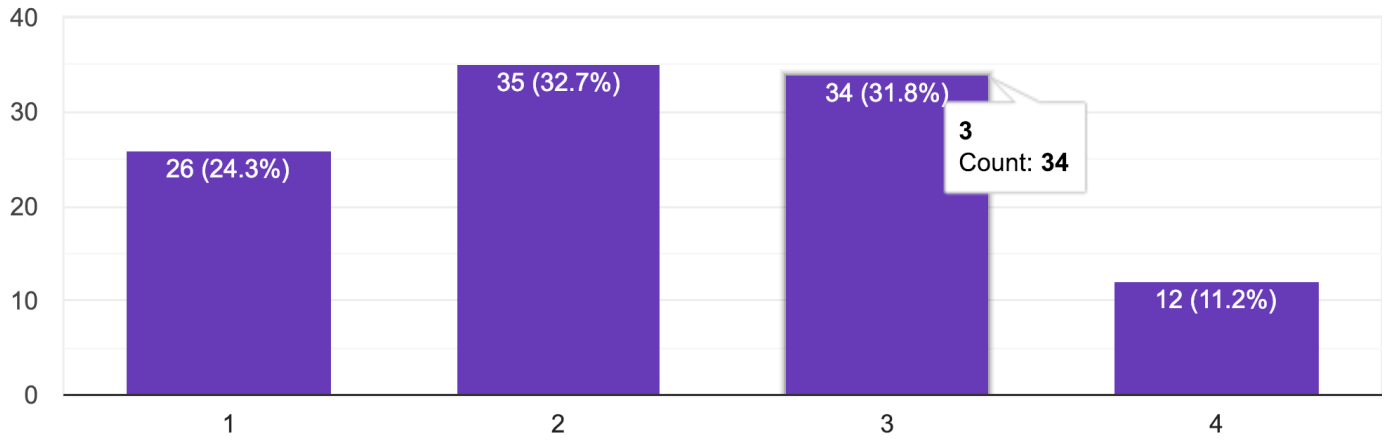
Do you have background in multivariate calculus? E.g., how familiar are you with concepts covered in MATH 53 like gradient, derivatives?

107 responses



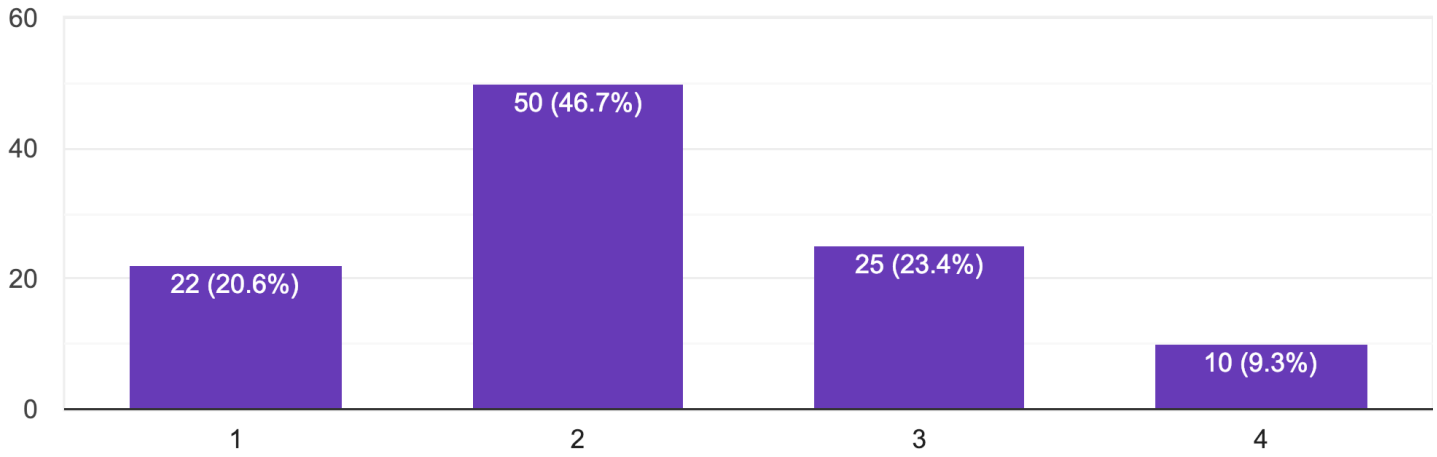
Do you have background in linear/matrix algebra? E.g., how familiar are you with concepts covered in MATH 54 like SVD, Eigen-decomposition

107 responses



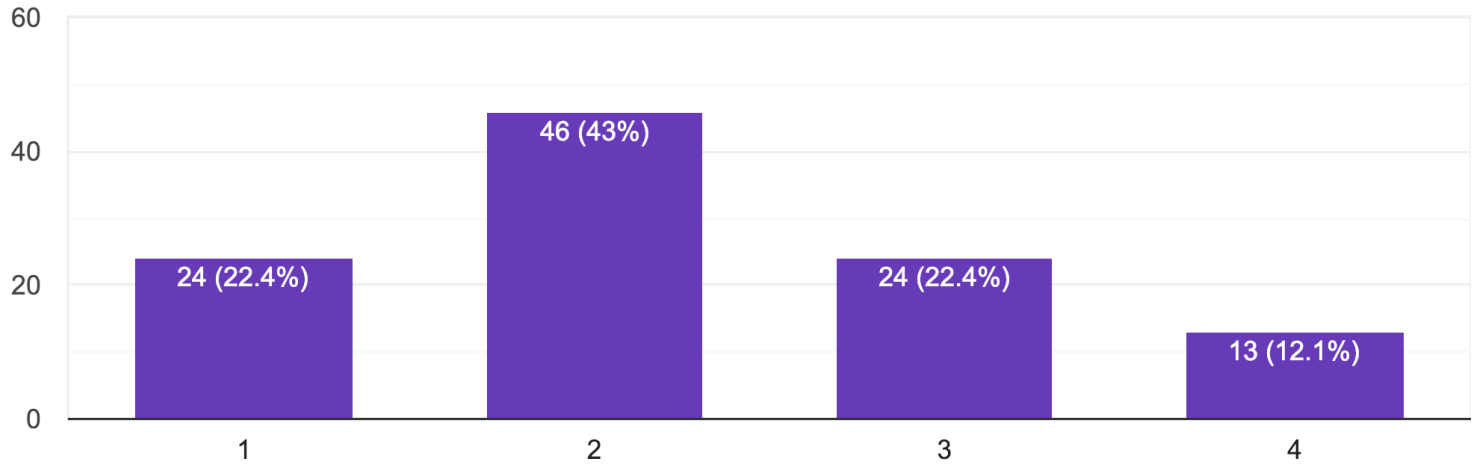
Do you have background in upper division probability? E.g., how familiar are you with concepts covered in Prob 140..., Markov and Chebyshev inequalities?

107 responses



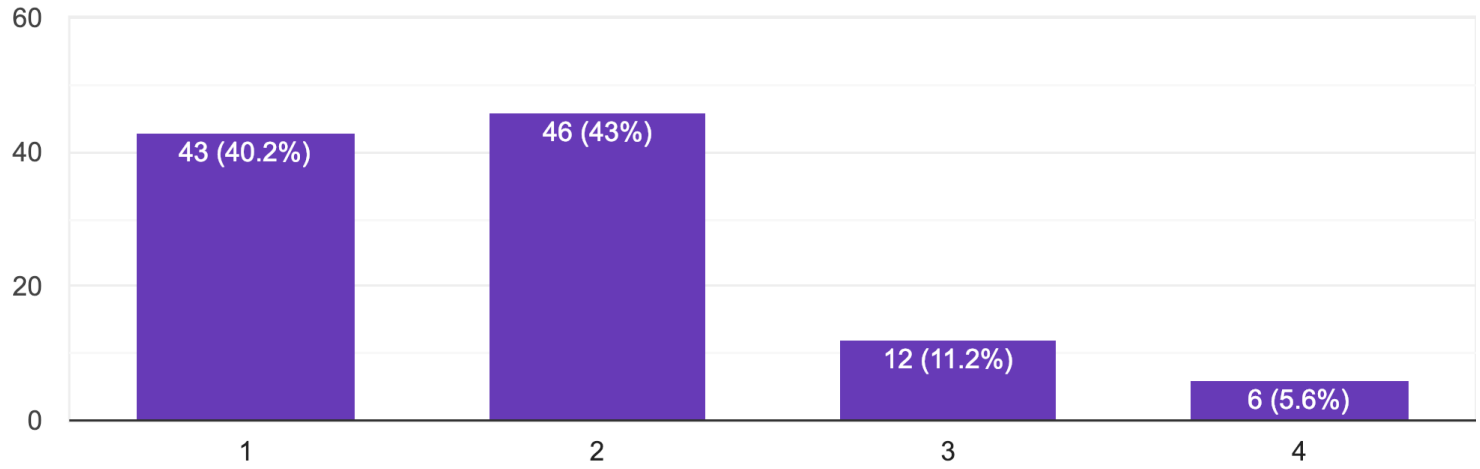
Do you have background with linear models? E.g., how familiar are you with concepts covered in STAT 135 like OLS, ridge regression?

107 responses



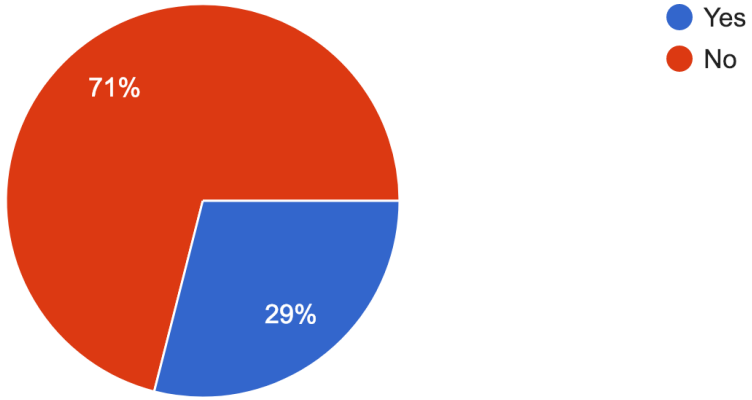
How proficient are you with programming in R?

107 responses

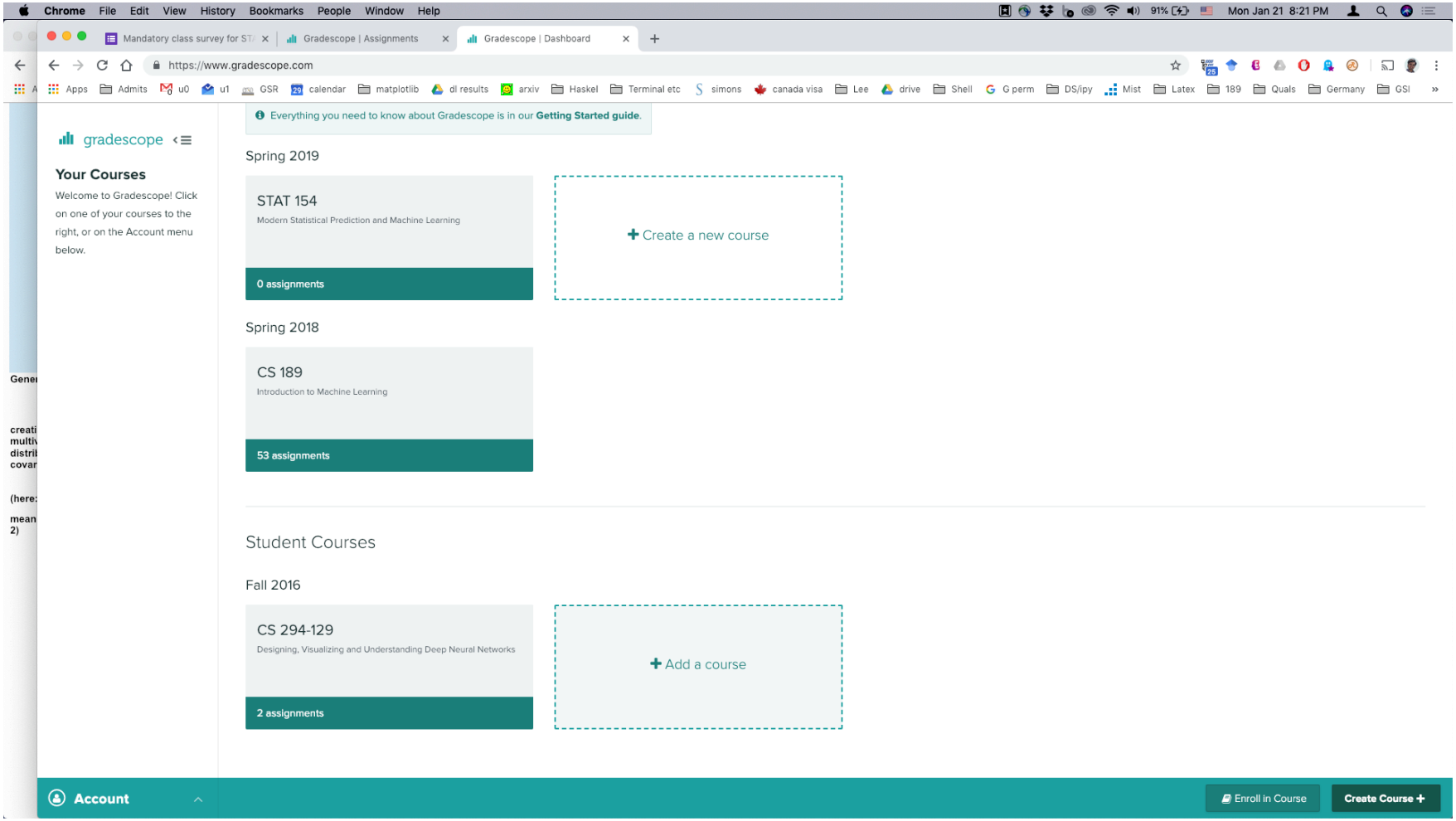


Have you taken DS 100

107 responses



Gradescope to submit homework and project



How to use lecture slides

- Lectures slides are not self-contained class notes – attendance in class is necessary to take notes
- Blackboard will also be used – notes are necessary
- They are uploaded at bcourses and provide useful review points after attending class
- If you have to miss a lecture, please ask notes from classmates

Summary

- Importance of attendance of lectures and labs
- 3-circle representation of data problem (including prediction)
- Prediction leads to action – informed predictions lead to good actions
- Math notations for a prediction problem
- New: hold-out set, cross-validation (CV)

Reading assignments

- Review math pre-requisites listed earlier
- Review R-studio
- Reading on Cross-validation (CV) (book chapter) for Thurs.

Lectures slides are not self-contained class notes – attendance in class is necessary.

They are uploaded at bcourses and provide useful review points after attending class

154 learning environment

- Our goal – a supportive, respectful, collaborative and intellectually honest environment for learning – learn how to learn
- No question is stupid – getting out of one's comfort is how we grow
- Shy? Sitting in the front helps