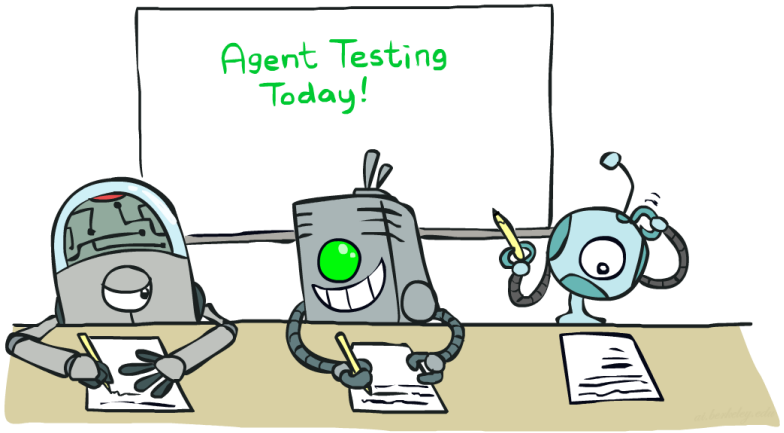
Machine Learning

CSE 473: Introduction to Artificial Intelligence



James Weichert (jpw@cs)
cs.uw.edu/473
acknowledgements

Welcome!
What's your favorite summertime activity?
sli.do **#cse473**

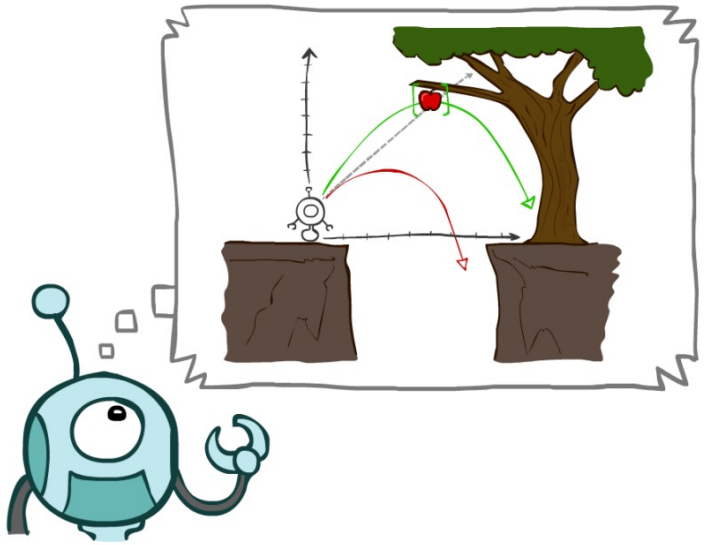
Administrivia

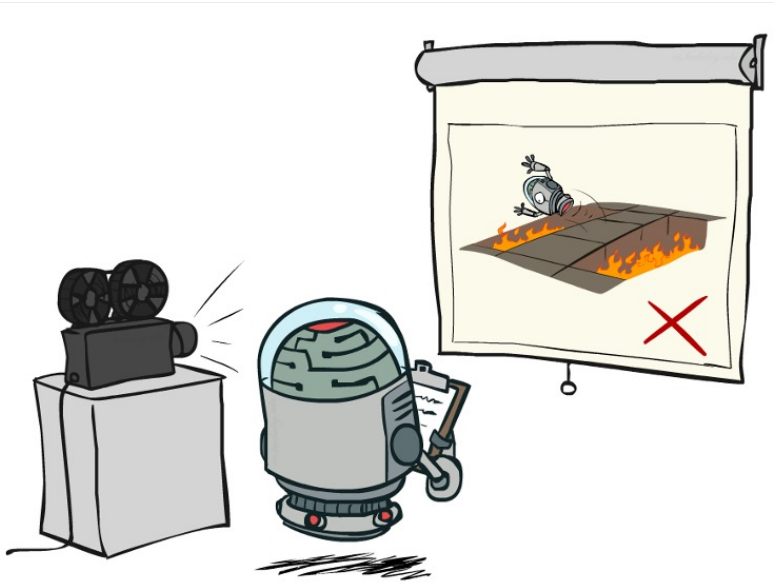
ANNOUNCEMENTS

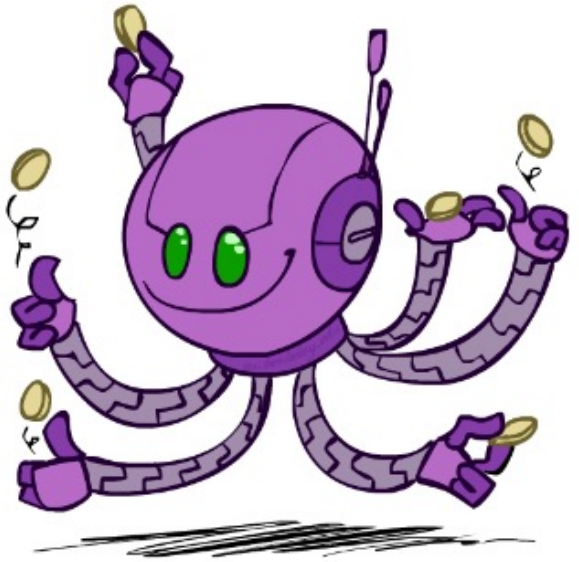
- **Test 2** next Friday
  - One double-sided 8.5x11" note sheet allowed

ASSIGNMENTS

- **Homework 3 due** next Thursday (8.6)
- **Project 4 due** in two weeks (8.13)







Taking Stock

Machine Learning?

CONTEXT

How to *construct* a model of the world from data or experience

Learn Parameters

Learn Structure

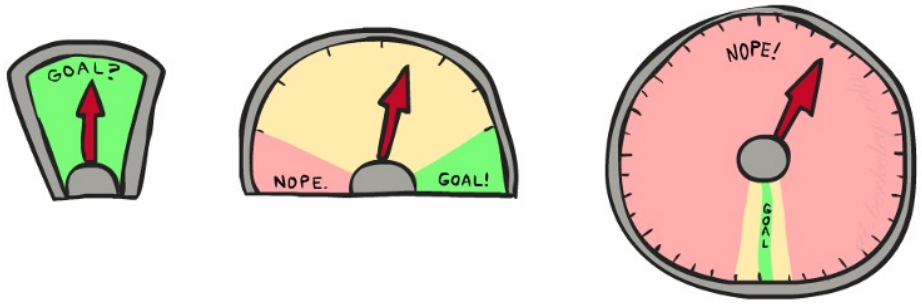
Learn Patterns

Learning

FOUNDATIONS

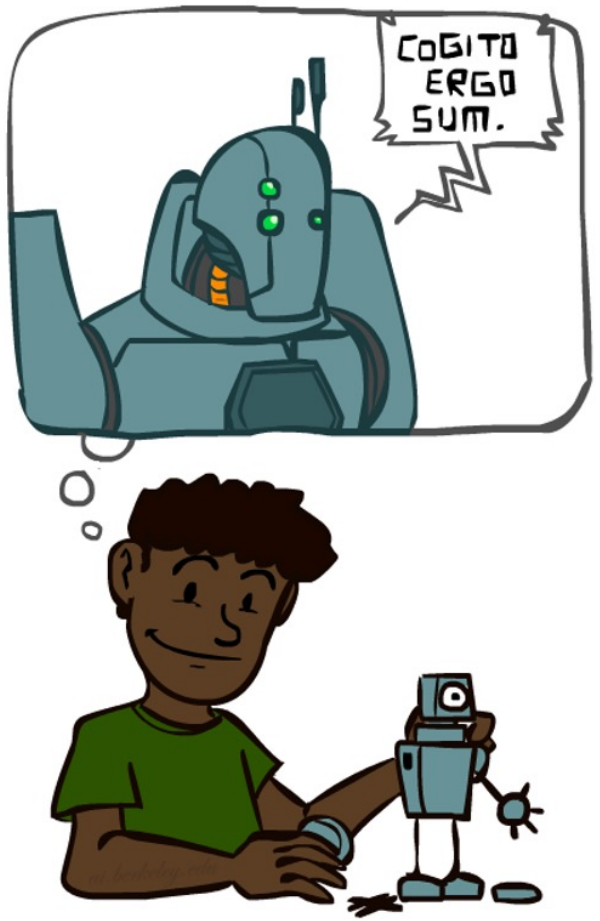
Learning involves **improving** an agent's **performance** through **experience**

- *How to define **performance**?*
- *How to **improve**? How to tell if the agent is improving?*
- *What constitutes **experience**? How to relate to new experiences?*



Why "Learning"?

CONTEXT

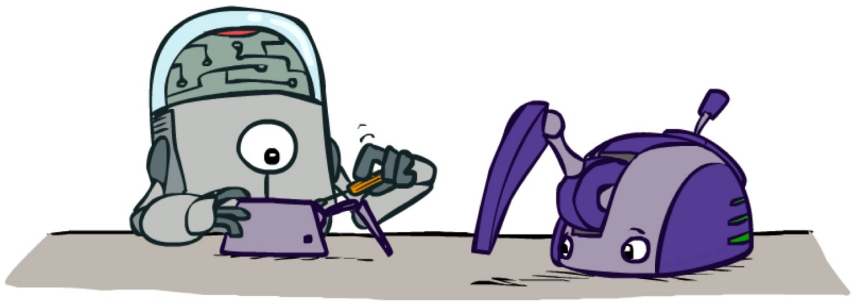


Learning Agents

FOUNDATIONS

Learning involves **improving** an agent's **performance** through **experience**

- *How does **performance** relate to the agent's design?*
- *How does **improvement** in performance manifest?*
- *What **data** are relevant to that part of the system?*
- *What knowledge is available or assumed?*



Contextualizing CSE 473 So Far

EXAMPLE

Agent	Component	Representation	Feedback	Knowledge
Adversarial Search	Evaluation function	Linear polynomial	Win/loss utility	Rules of game
Planning Agent	Transition model	Successor transitions	Action outcomes	Available actions
Weather Model	Graph model	Bayesian network	Observations	Topology assumptions
Localization	Sensor model	Hidden Markov model	Sensor readings	Sensor information
Image Classifier	Classification model	Naïve Bayes classifier	Training labels	Network architecture
Language Model	N-gram model	Markov Chain	Word frequencies	Value of n

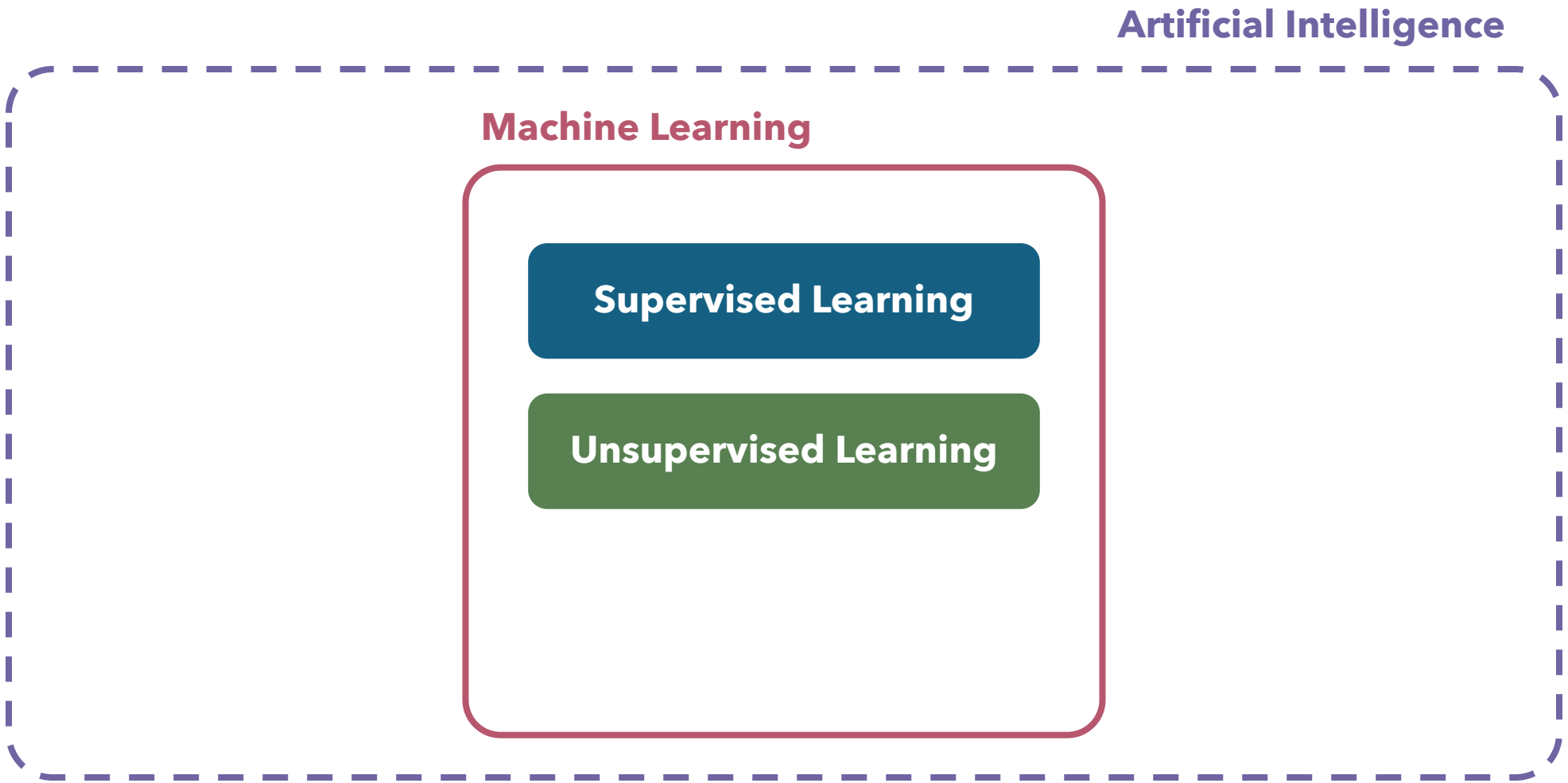
QUESTIONS

Questions?

live and on sli.do #cse473

A Broad Segmentation

CONTEXT



Supervised Learning

FOUNDATIONS

Goal: Learn unknown **target function** f

- Input: **Training data** with *labeled* examples (x_i, y_i), where $f(x_i) = y_i$
- Output: **Hypothesis** h that is "close" to f
  - $h \approx f$ implies that h predicts well on unseen (test set) data
- Knowledge: Family of hypotheses H to choose h from
  - Linear models, logistic regression, neural networks, decision trees, non-parametric models, etc.
- Classification: Learn f with discrete output value
- Regression: Learn f with continuous output value

Unsupervised Learning

FOUNDATIONS

Goal: Learn unknown **target function** f

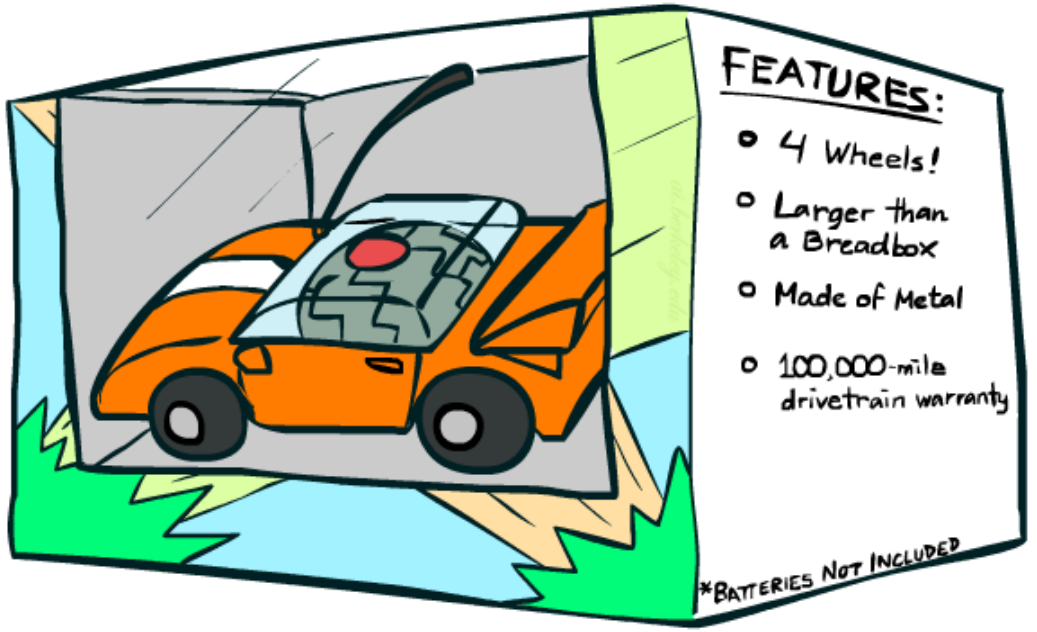
- Input: **Training data** with *unlabeled* examples x_i, where y_i is unknown
- Output: **Hypothesis** h that "makes sense"
  - h provides useful information about the data
- Knowledge: Unsupervised learning task
  - Clustering, outlier detection, latent variable modeling

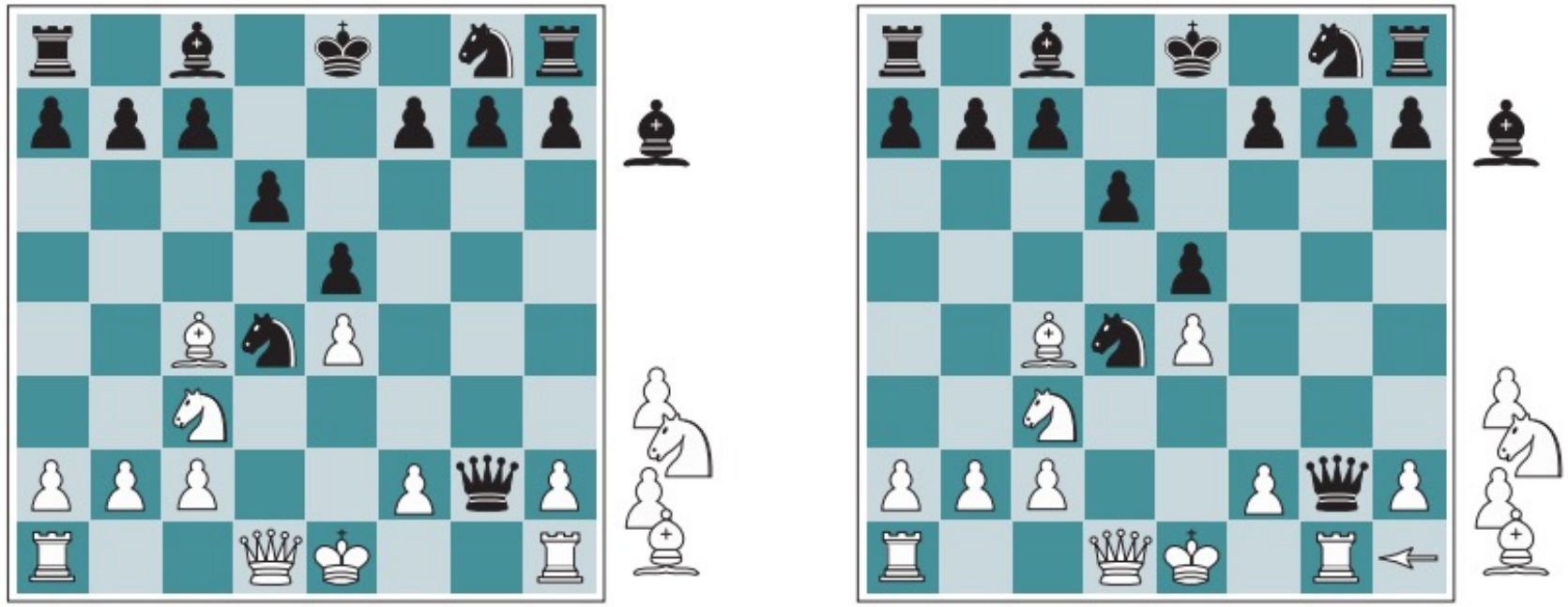
Features

DEFINITION

Features $\vec{x}_i$ describe item (observation) i

What should the features be?





(a) White to move (b) White to move

Figure 5.8 Two chess positions that differ only in the position of the rook at lower right. In (a), Black has an advantage of a knight and two pawns, which should be enough to win the game. In (b), White will capture the queen, giving it an advantage that should be strong enough to win.

Feature Extraction

QUESTIONS

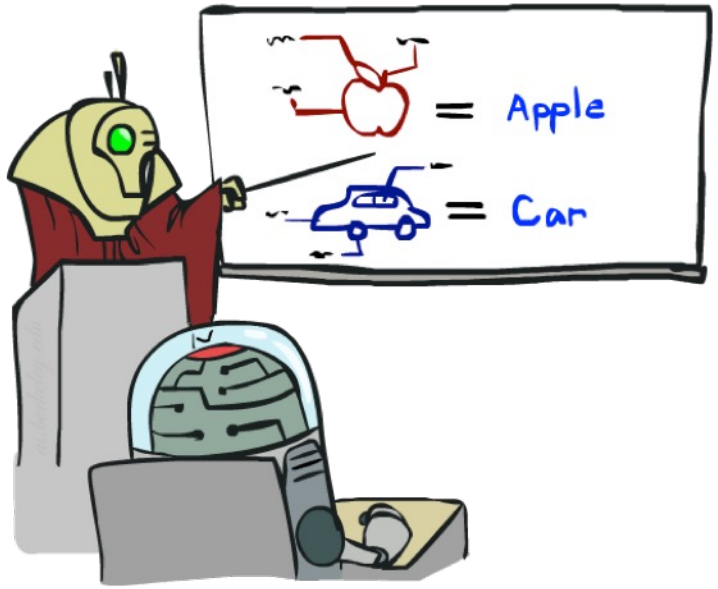
Questions?

live and on sli.do #cse473

The Learning Process

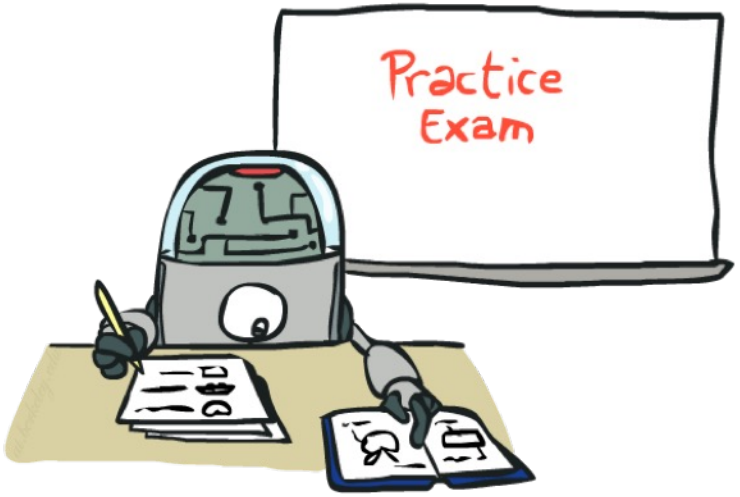
FOUNDATIONS

Learn



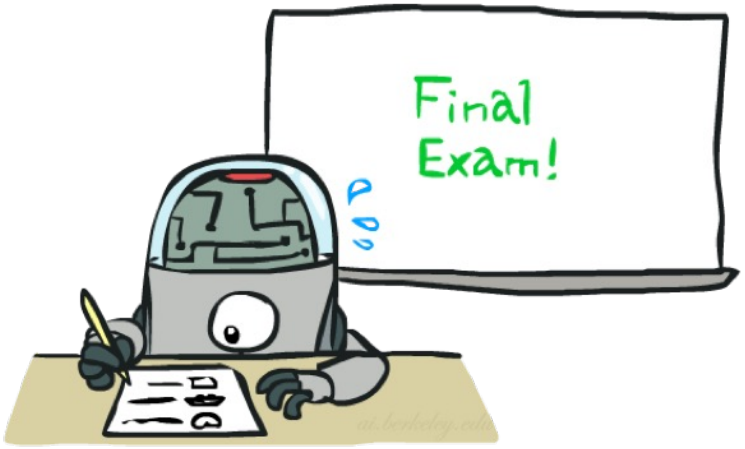
Training

Double Check



Validation

'Real Deal'

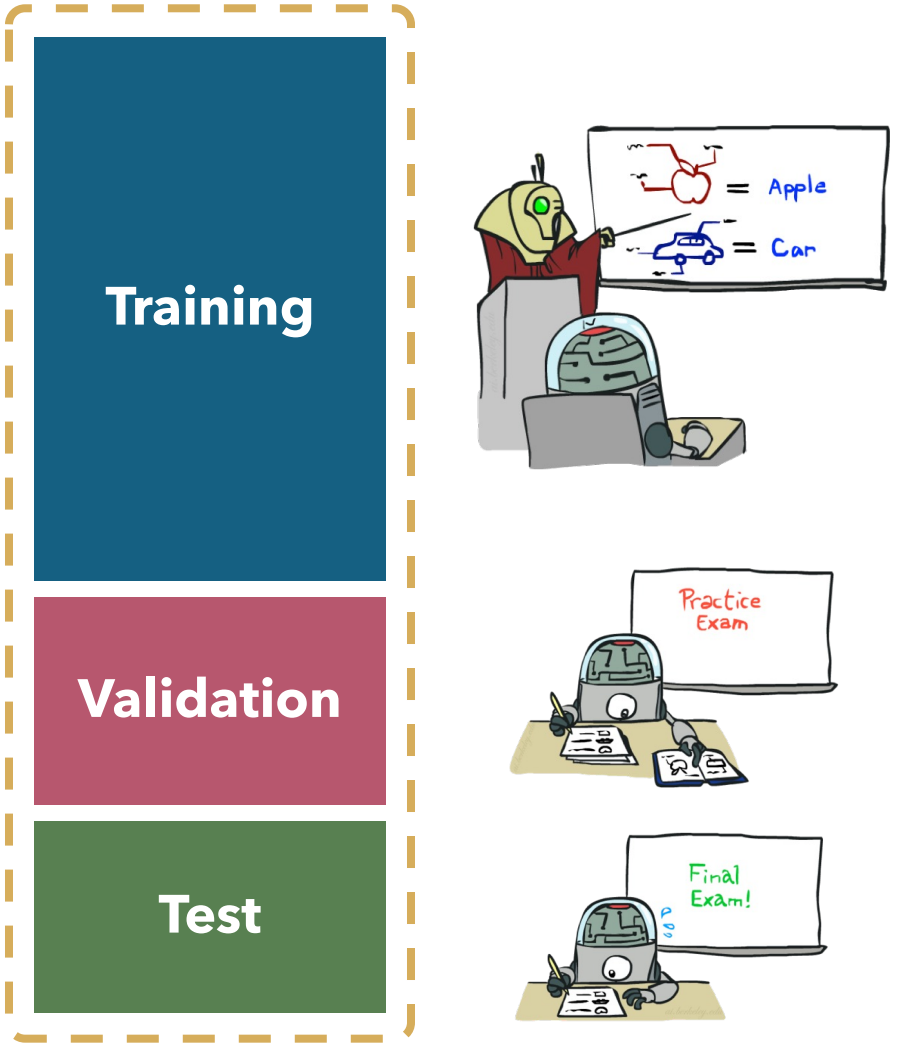


Testing

Model Training

FOUNDATIONS

- Data: labeled instances
  - Split into **training**, **validation**, and **test** sets
- Features: attribute-value pairs
- Experimentation Cycle:
  - Learn parameters on training set
  - Tune hyperparameters on validation set
  - *Only* when done, open test set "vault"
- *How to evaluate performance?*

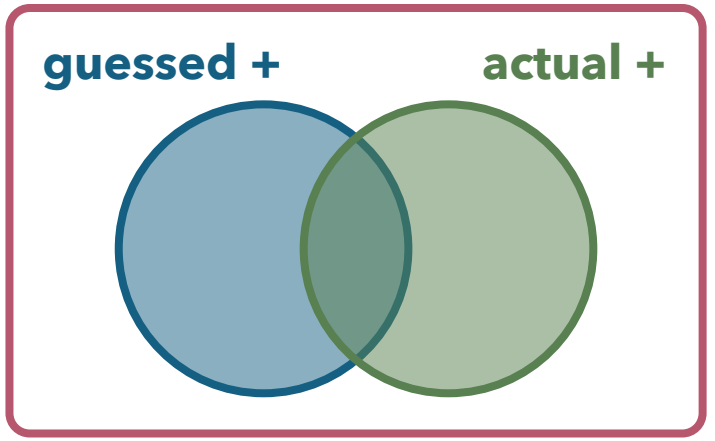


Evaluating Performance

FOUNDATIONS

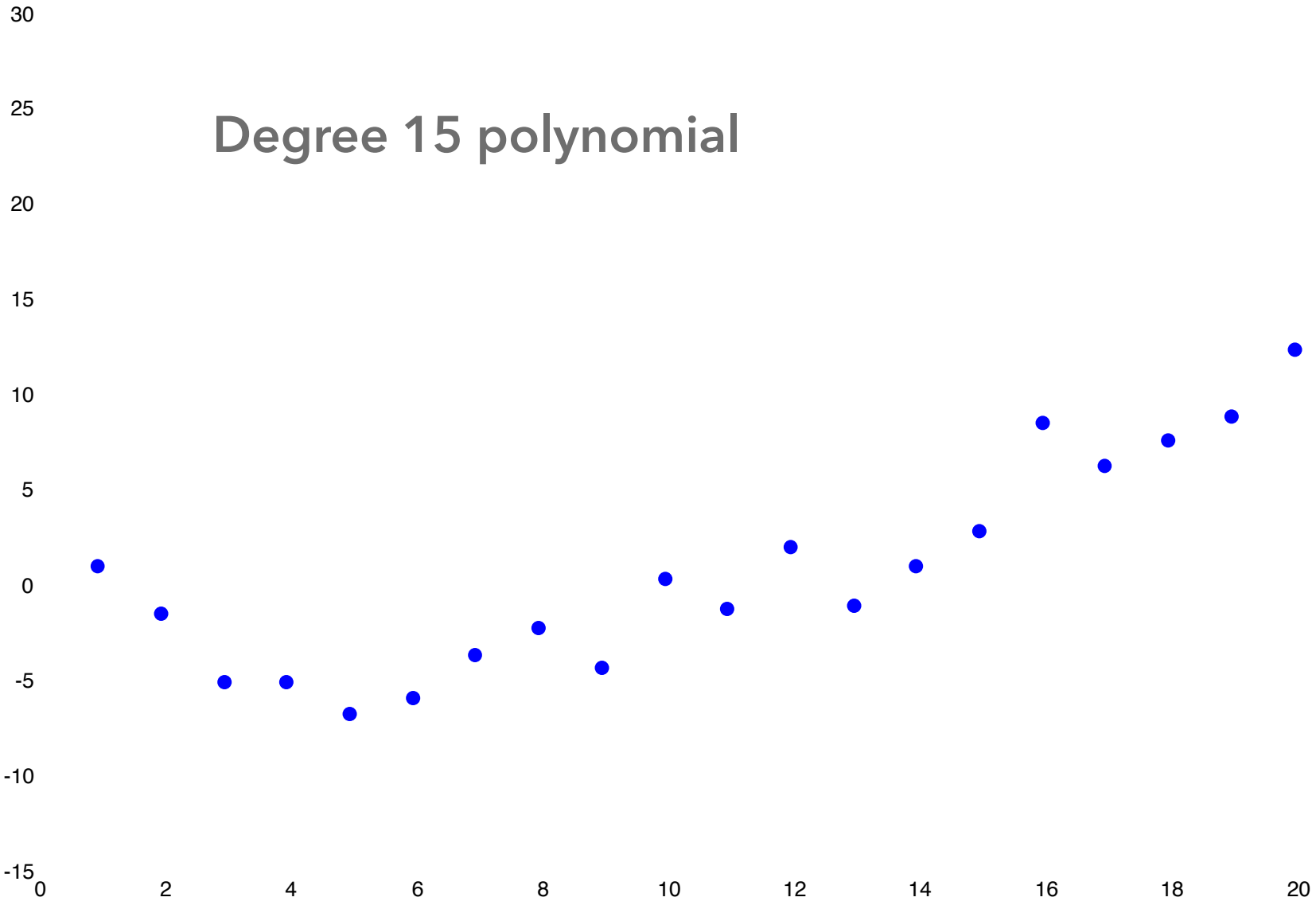
*How should we measure the **performance** of the model?*

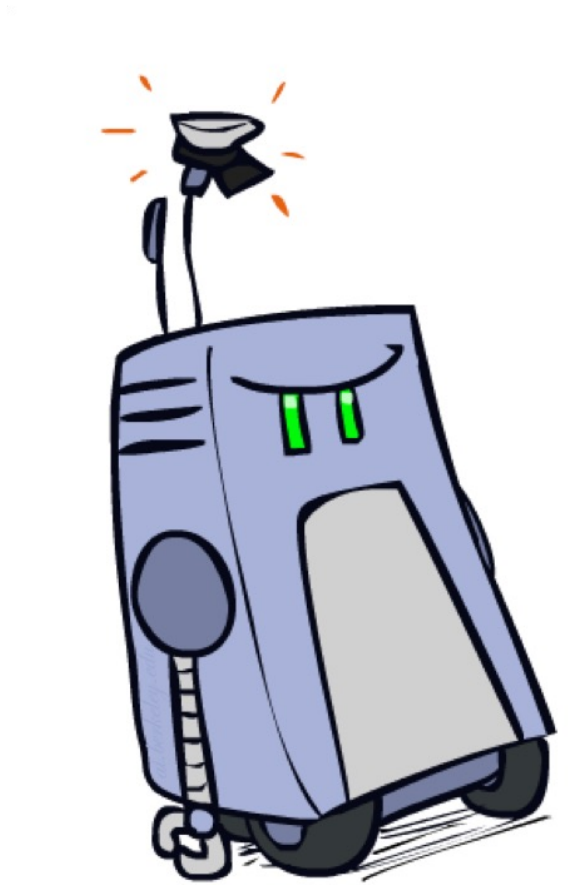
- Accuracy?
  - What happens when classes (outcomes) are unevenly likely?
  - Example:
    - Out of 1000 lung x-rays, there are 3 cases of pneumonia
    - A simple classifier easily achieves 99.7% accuracy! Why?



- Precision and Recall?
  - $\text{Precision} = \frac{TP}{TP+FP}$ (Type I Error), $\text{Recall} = \frac{TP}{TP+FN}$ (Type II Error)

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$



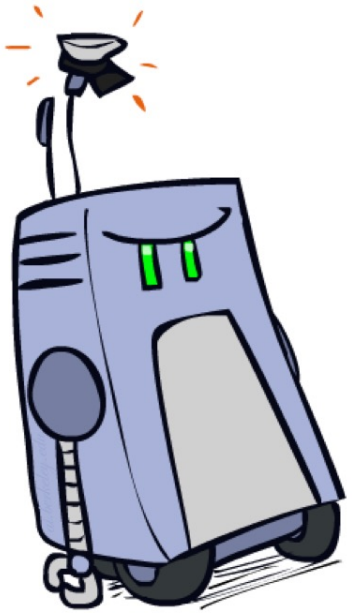


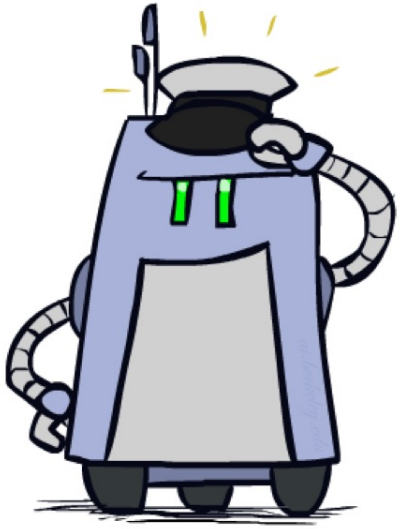
The Dark Side of Accuracy

Overfitting and Generalization

CONTEXT

- The **variance** of the model's predictions should also be evaluated
  - Models that do no *generalize* to unseen data are not "good"

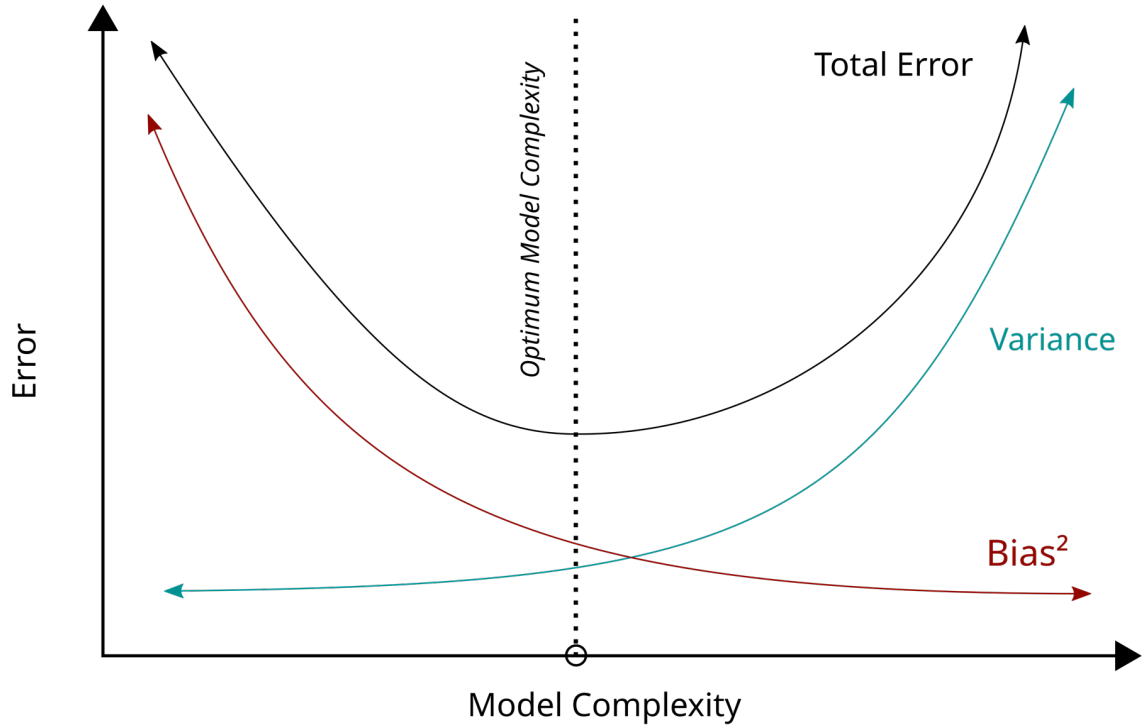




Bias-Variance Tradeoff

DEFINITION

- **Bias** (accuracy) and **variance** both constitute forms of model error
  - Minimizing the combined error is tricky, and there is some amount of error that is irreducible



Other evaluation criteria?

QUESTIONS

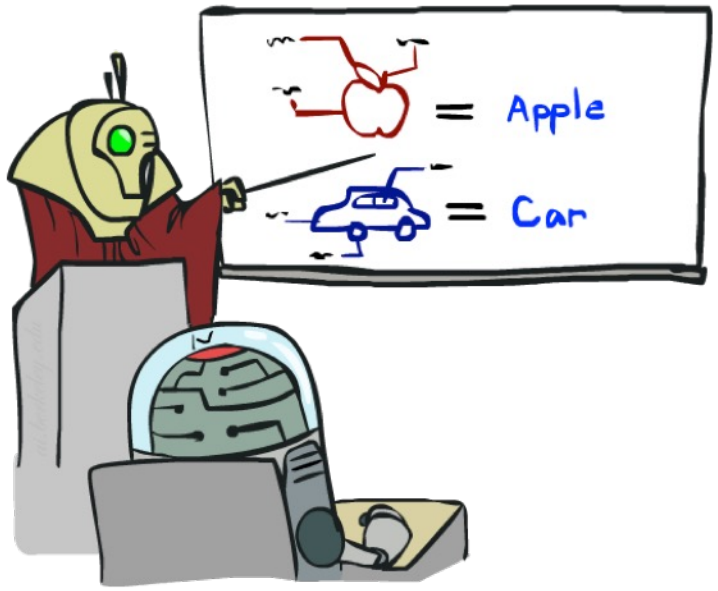
Questions?

live and on sli.do #cse473

Unsupervised Learning Process?

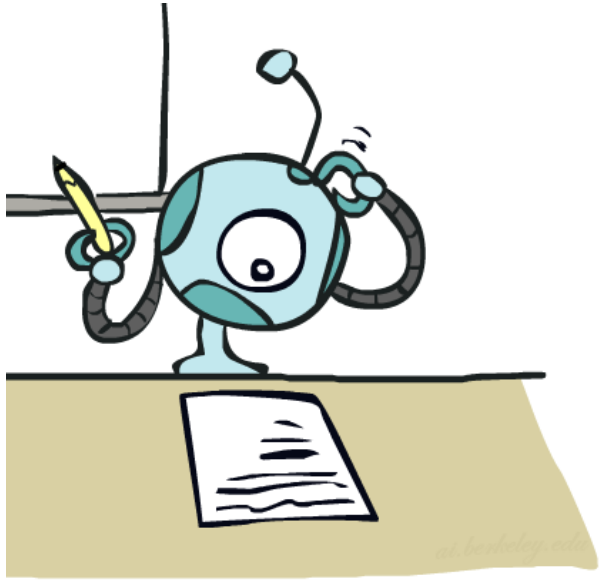
FOUNDATIONS

Learn



Training

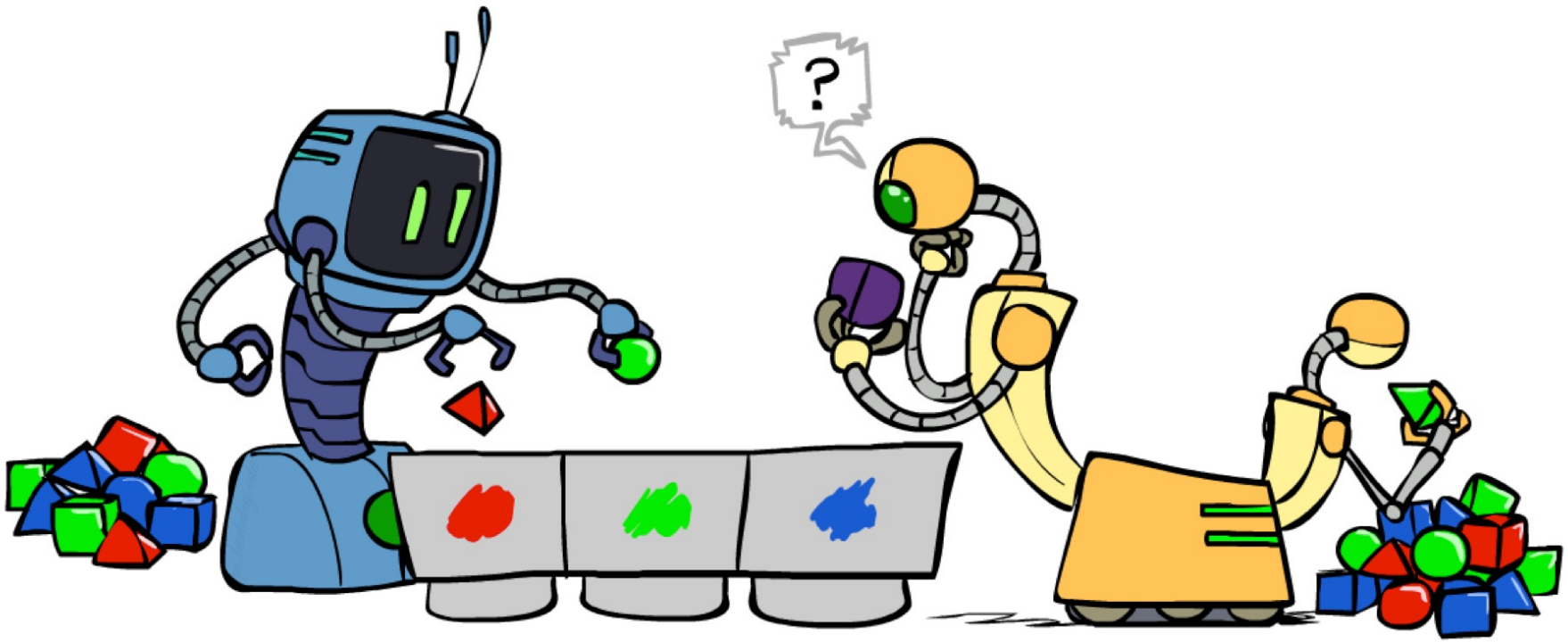
Check?



Clustering

FOUNDATIONS

Goal: Group together "**similar**" instances

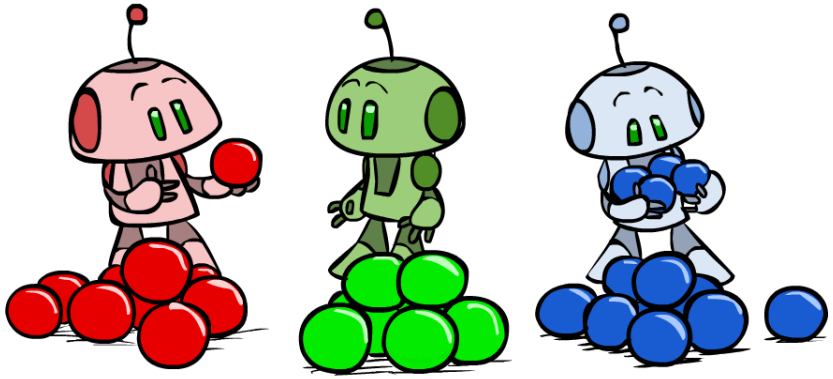


K-Means Clustering

ALGORITHM

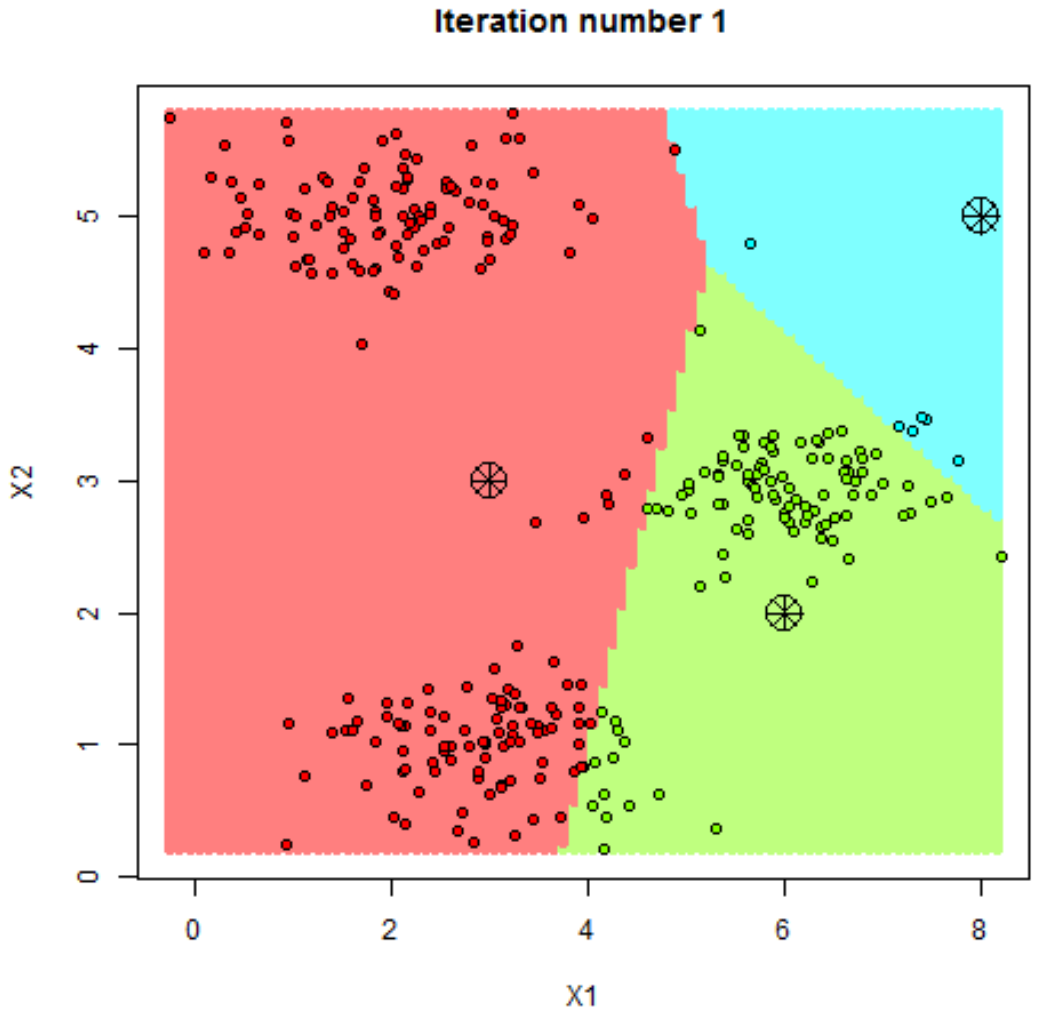
Define **similarity** as **distance** between points in feature space

- Pick *k* random points as cluster means
- Iterate until convergence (no changes):
  - Assign data points to closest cluster mean
  - Update cluster means to be the average of the data points in the cluster



K-Means Example

EXAMPLE



K-Means as Optimization

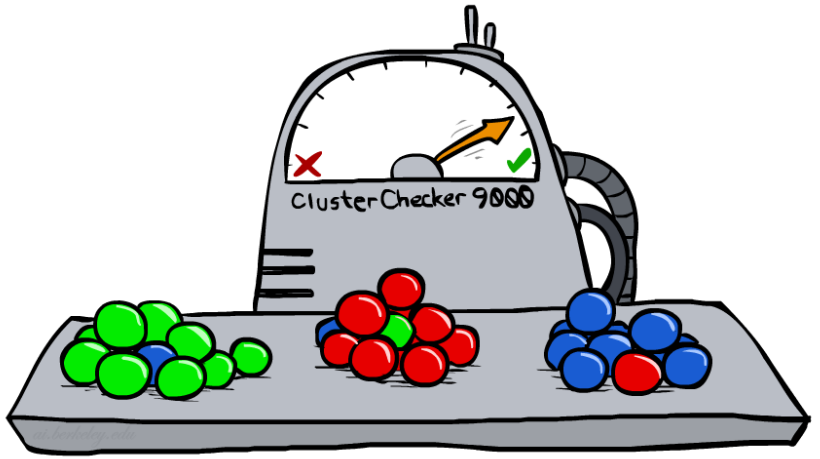
DEFINITION

Consider total distance to means

$$\phi(\{x_i\}, a, \{c_k\}) = \sum_i dist(x_i, c_{a(i)})$$

- given data x_i, cluster means c_k, and function a mapping points to cluster assignments

- Claim: Each iteration reduces ϕ
- Algorithm:
  - Phase 1: Update assignments by fixing c and updating a
  - Phase 2: Update means by fixing a and updating c



Phase 1: Update Assignments

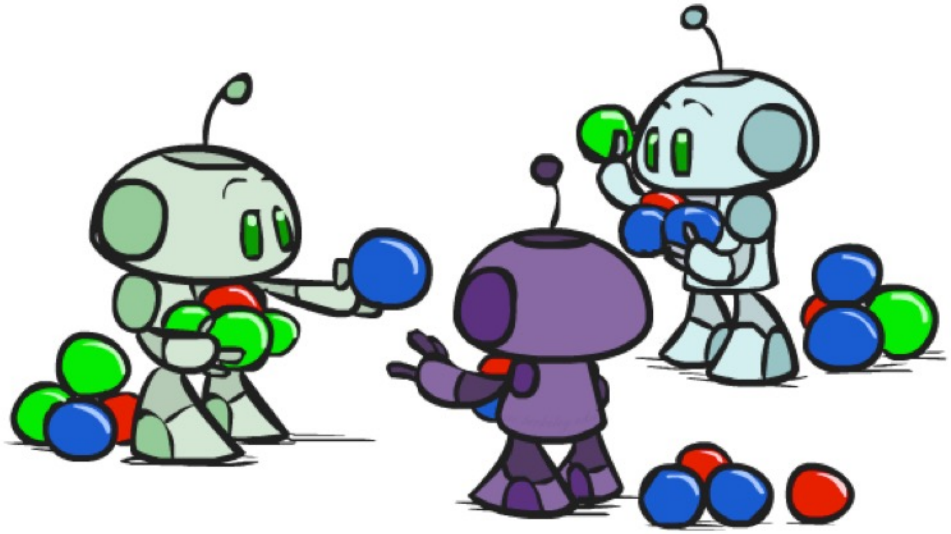
DEFINITION

- For each point, re-assign to closest mean

$$a(i) \leftarrow \text{argmin}_k \, \text{dist}(x_i, c_k)$$

This can only decrease total distance ϕ

$$\phi(\{x_i\}, a, \{c_k\}) = \sum_i dist(x_i, c_{a(i)})$$



Phase 2: Update Means

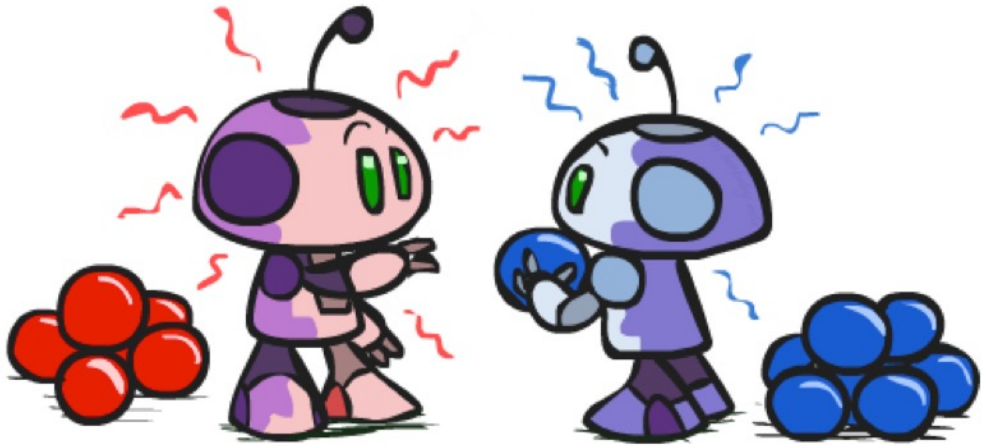
DEFINITION

- Move each cluster mean to average of its assigned points

$$c_k \leftarrow \frac{1}{|\{i: a(i) = k\}|} \sum_{i:a(i)=k} x_i$$

This also can only reduce total distance ϕ

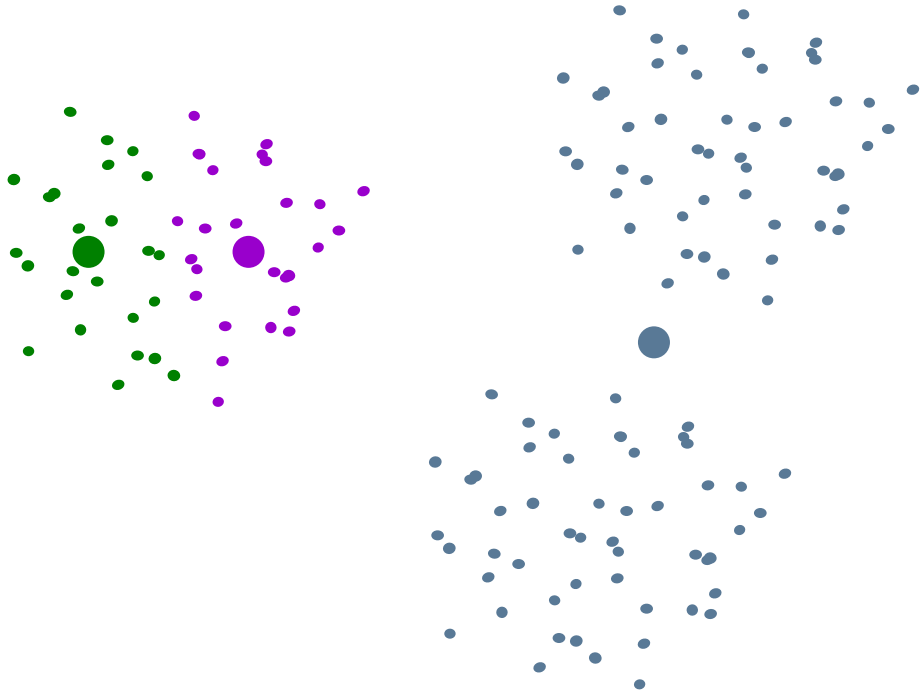
- Why? Point p with minimum squared Euclidean distance to a set of points $\{x\}$ is their mean

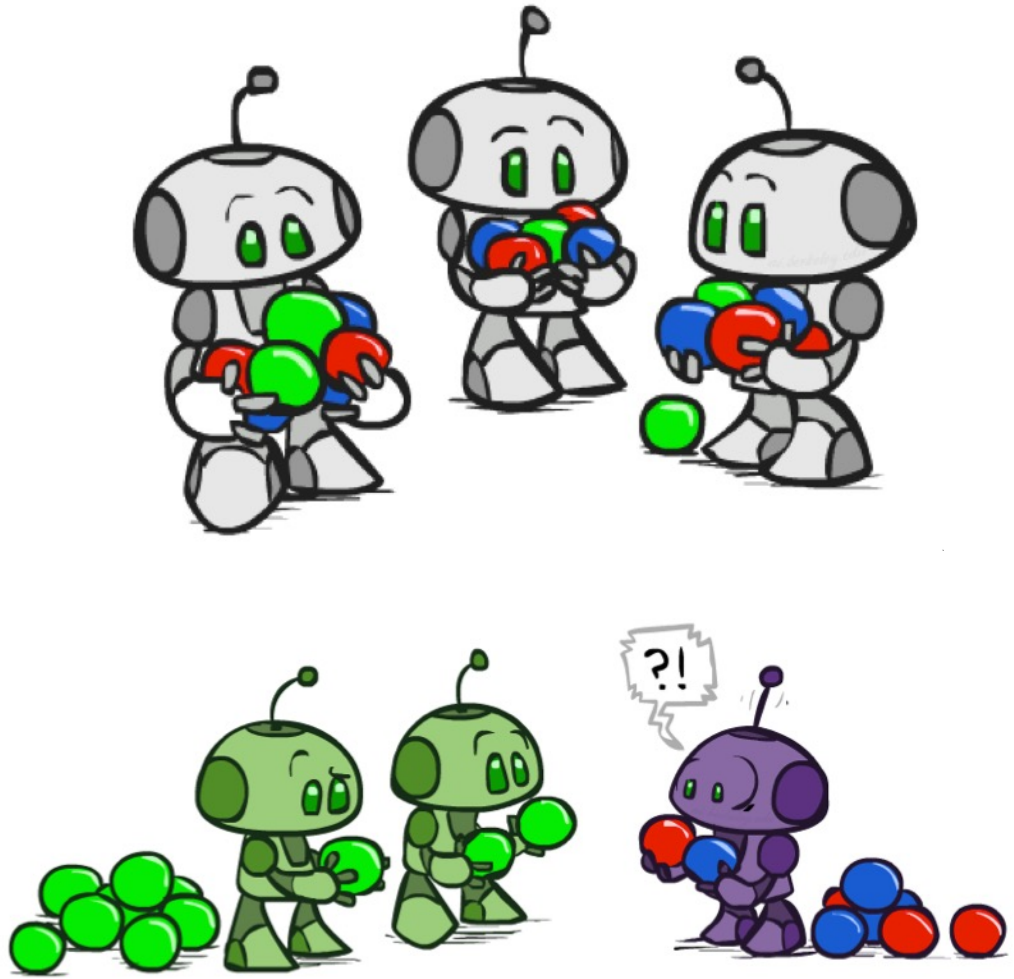


Initialization

DEFINITION

- K-Means is non-deterministic
  - Initialization of means affects outcome





Clustering Performance

DEFINITION

How good is a cluster?

- Silhouette Score
  - $S(i) = \frac{(b(i)-a(i))}{\max(a(i),b(i))}, \quad a(i) = \text{mean}(\text{dist}_{\text{own}}), \quad b(i) = \min(\text{mean}(\text{dist}_{\text{other}}))$
- Mutual information
  - $MI(y,z) = \sum\sum P(y_i, z_j) * \log\left(\frac{P(y_i,z_j)}{P(y_i)*P'(z_j)}\right), \quad y_i = \text{ground truth}, \quad z_j = \text{predicted}$

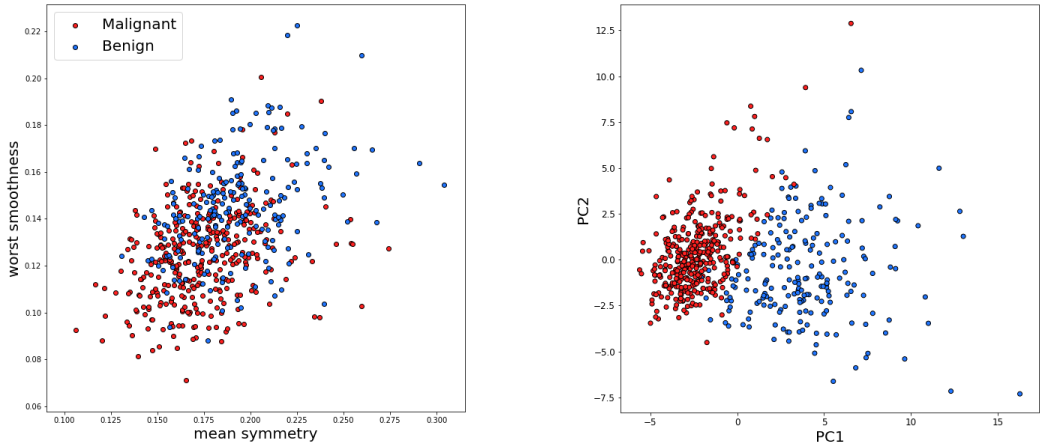
Dimensionality Reduction

FOUNDATIONS

- **Principal component analysis** (PCA) finds the principal components (basis functions) that capture the most variation in the data

$$\widehat{\boldsymbol{X}}_k = \boldsymbol{X} - \sum_{s=1}^{k-1} \boldsymbol{X}\boldsymbol{w}_{(s)}\boldsymbol{w}_{(s)}^{\top}$$

  - Singular Value Decomposition $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}$ gives principal components as columns of $\boldsymbol{V}$



QUESTIONS

Questions?

live and on sli.do #cse473

that's it for today!

SUMMARY

- Learning is *essential* in unknown environments, useful in many others
- Learning process depends on agent design, evaluation metric, data, prior knowledge
- Clustering involves finding patterns in unlabeled data

UPCOMING

- Regression
- Classification

REMINDERS

- Complete **Practice Problem 15**
- Do **Homework 3** (due 8.6)
- Do **Project 4** (due 8.13)