Advanced Business Data Mining

MSIS 522 – Lesson 1

Foster School of Business

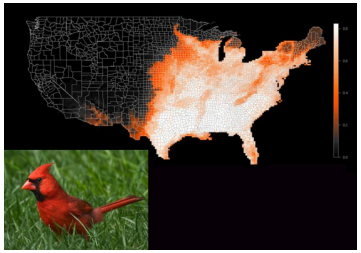
Basic Info

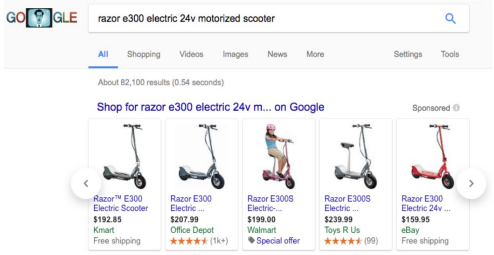
Time	Mon Night, Wed Night & Sat
Location	PACCAR 292 (Gold) & PACCAR 294 (Purple)
Instructor	Jun Yu (jnyu@uw.edu)
TA	Xiyuan Ge (xiyuange@uw.edu)
Textbook	• An Introduction to Statistical Learning (free online) • A Course in Machine Learning (free online) • Mining of Massive Datasets (free online) • Python Data Science Handbook (free online)
Software	Python on Google Colab
Grading	• 3 Individual assignments (90%) [20% per 24 hours applied to late submissions] • Participation (10%)

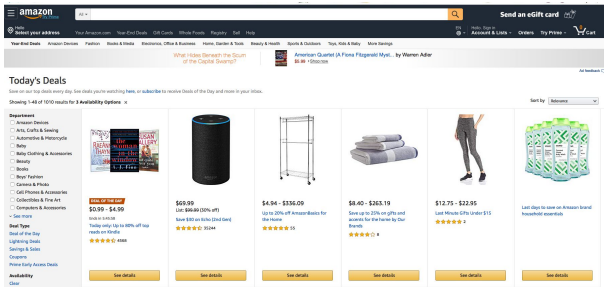
Foster School of Business

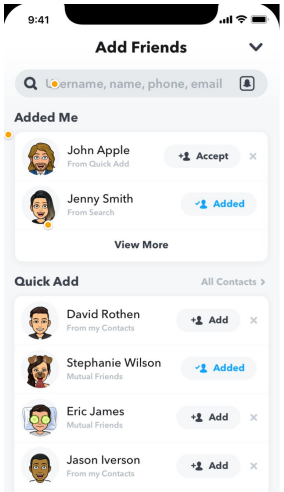
About Me

- Ph.D. in CS @ Oregon State University
  - Species Distribution Modeling for the eBird project
- Applied Researcher @ eBay
  - Paid Internet Marketing
  - Seller Risk Management
- Senior Applied Scientist @ Amazon
  - Deals Forecasting and Scheduling
  - Seller Recommendation System
- Engineering Manager @ Snapchat
  - Friend Recommendation







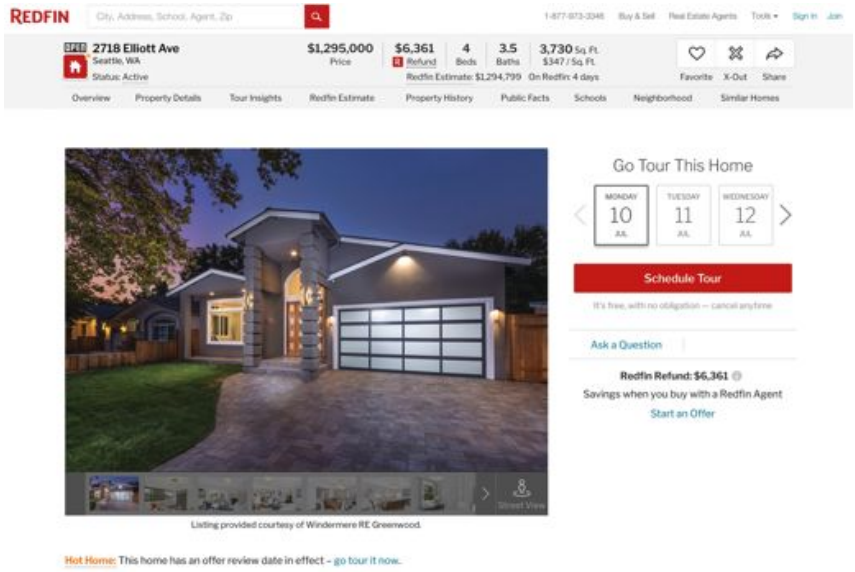


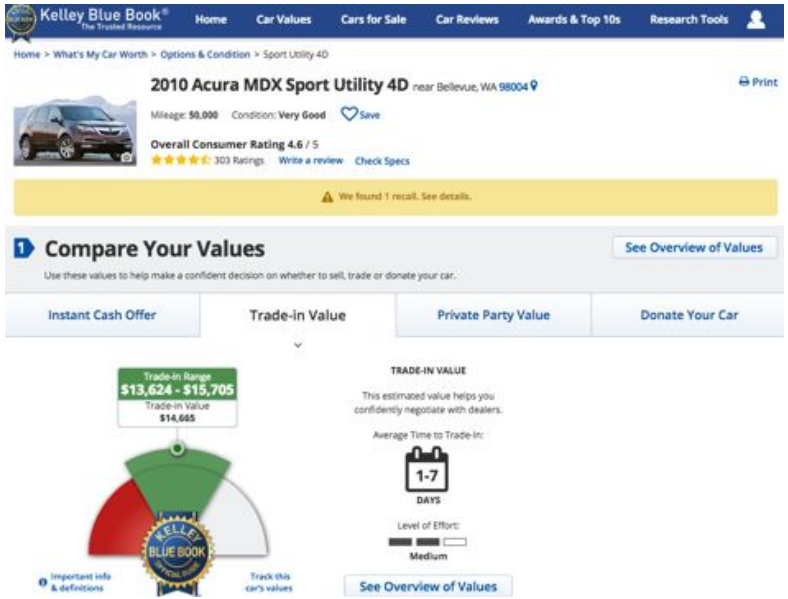
Foster School of Business

Jeff Bezos: The Golden Age of Artificial Intelligence

“Machine Learning and AI is a horizontal enabling layer. It will empower and improve every business, every government organization, every philanthropy, basically there is no institution in the world that cannot be improved with Machine Learning.” – Jeff Bezos

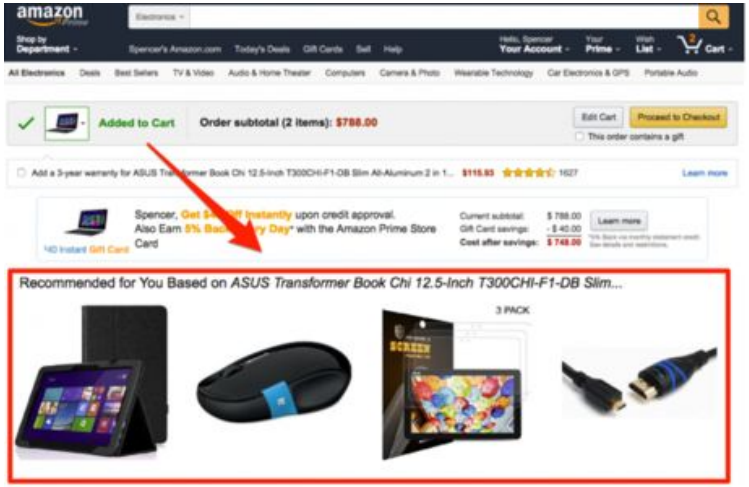
House & Vehicle Value Estimates





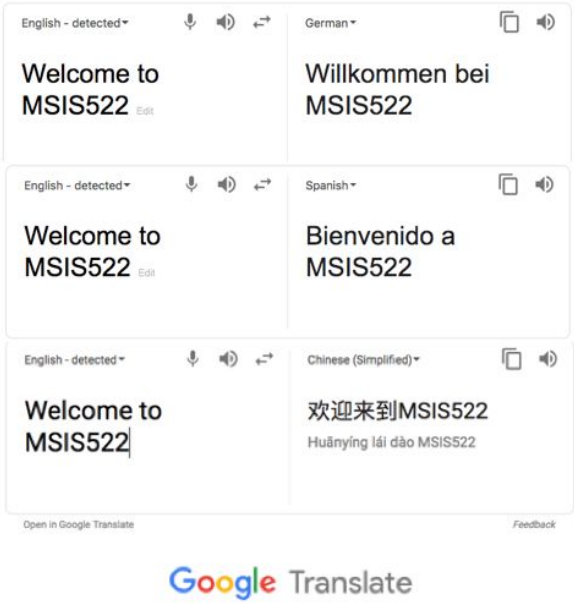
Foster School of Business

Product & Movie Recommendation



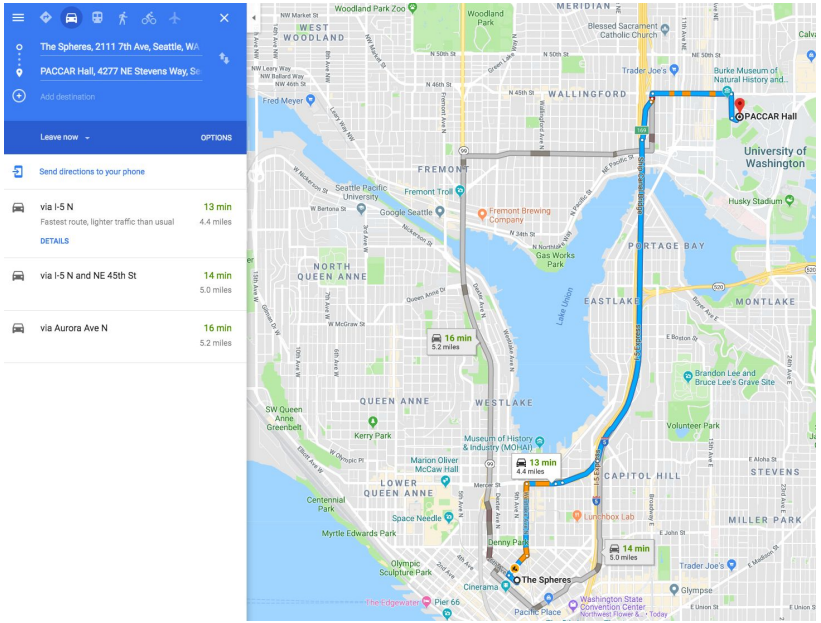
Foster School of Business

Voice Recognition & Machine Translation



Foster School of Business

Play Game of GO & Route Planning



Foster School of Business

Machine Learning Everywhere!

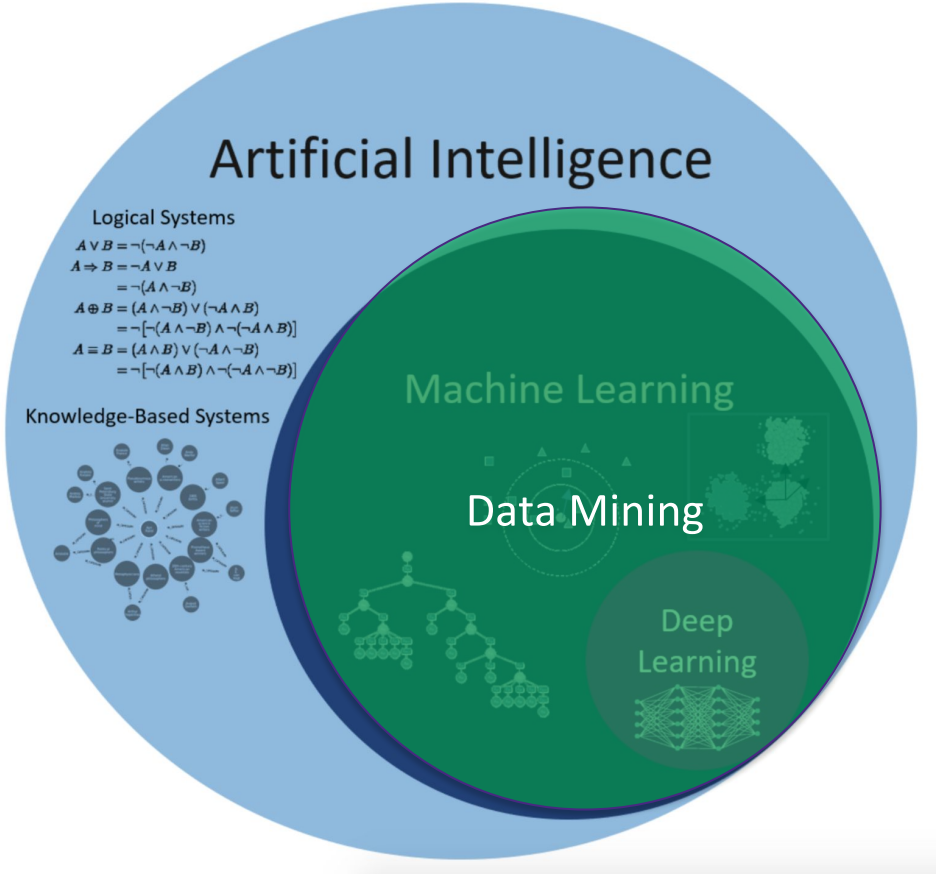
A real life example of Machine Translation

Foster School of Business

Quiz

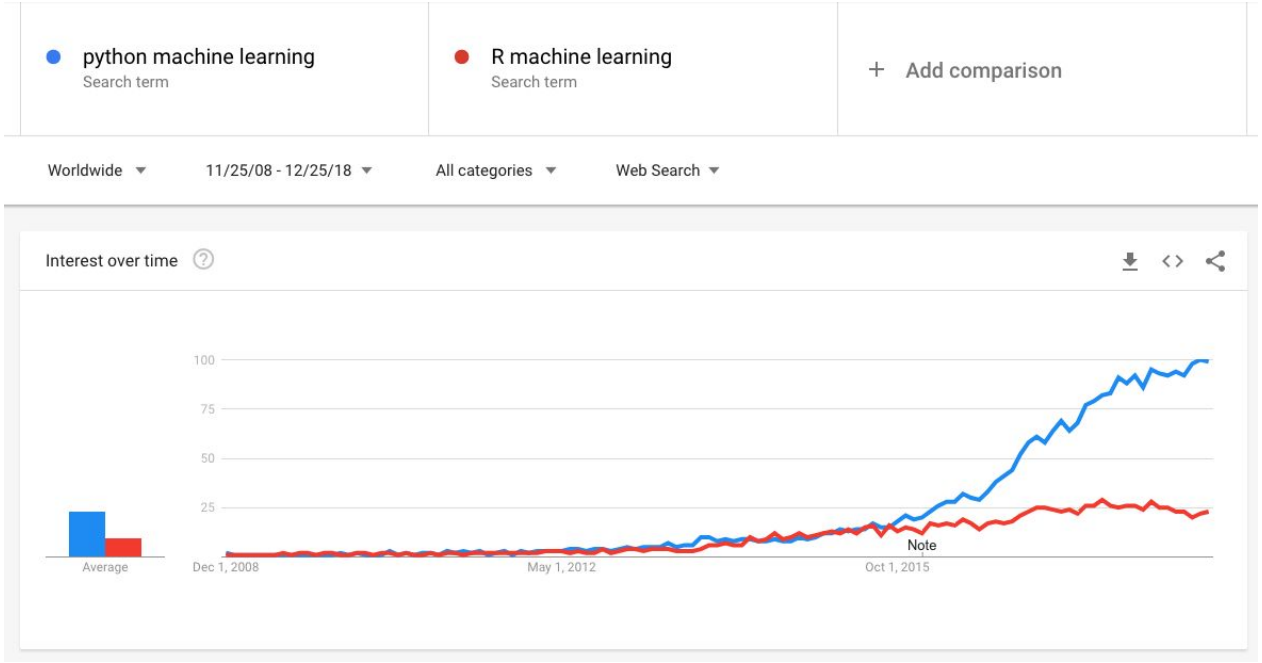
Can you think of any real world applications driven by ML?

AI vs. ML vs. DL



Foster School of Business

Why Python (vs. R) for ML?



Data Science Wars: R vs. Python

Why Python (vs. rest)?

PYPL PopularitY of Programming Language

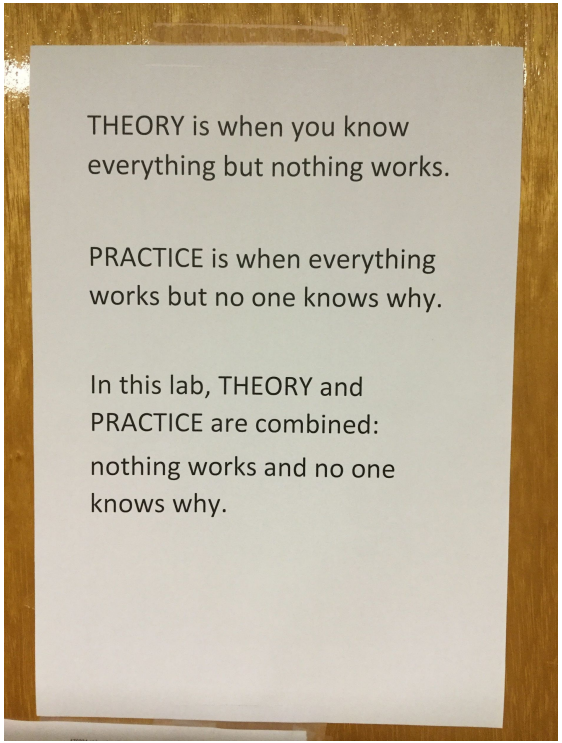
Worldwide, Jan 2019 compared to a year ago:

Rank	Change	Language	Share	Trend
1	↑	Python	25.95 %	+5.2 %
2	↓	Java	21.42 %	-1.3 %
3	↑	Javascript	8.26 %	-0.2 %
4	↑	C#	7.62 %	-0.4 %
5	↓↓	PHP	7.37 %	-1.3 %
6		C/C++	6.31 %	-0.3 %
7		R	4.04 %	-0.2 %

Life is short, use **Python**

Foster School of Business

Theory vs. Practice



- 50% Theory + 50% Practice (lab).
- Work hard on the homework.
- Explore and participate in online ML competitions.

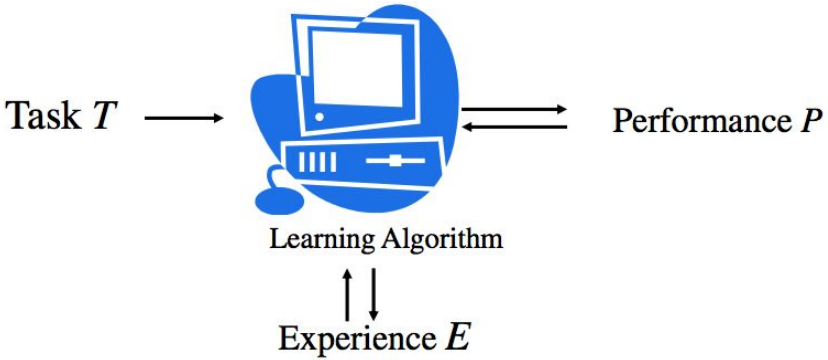
Syllabus

- Lecture 1 - Fundamentals of Machine Learning
- Lecture 2 - Decision Tree
- Lecture 3 - Ensemble Learning
- Lecture 4 - Clustering
- Lecture 5 - Recommendation System

Fundamentals of Machine Learning

What is Machine Learning?

A computer program is said to learn from **experience *E*** with respect to some class of **tasks *T*** and **performance measure *P*** if its performance at tasks in *T*, as measured by *P*, improves with experience *E*. -- Tom Mitchell



Improving ***performance P*** with ***experience E*** at some ***task T.***

Foster School of Business

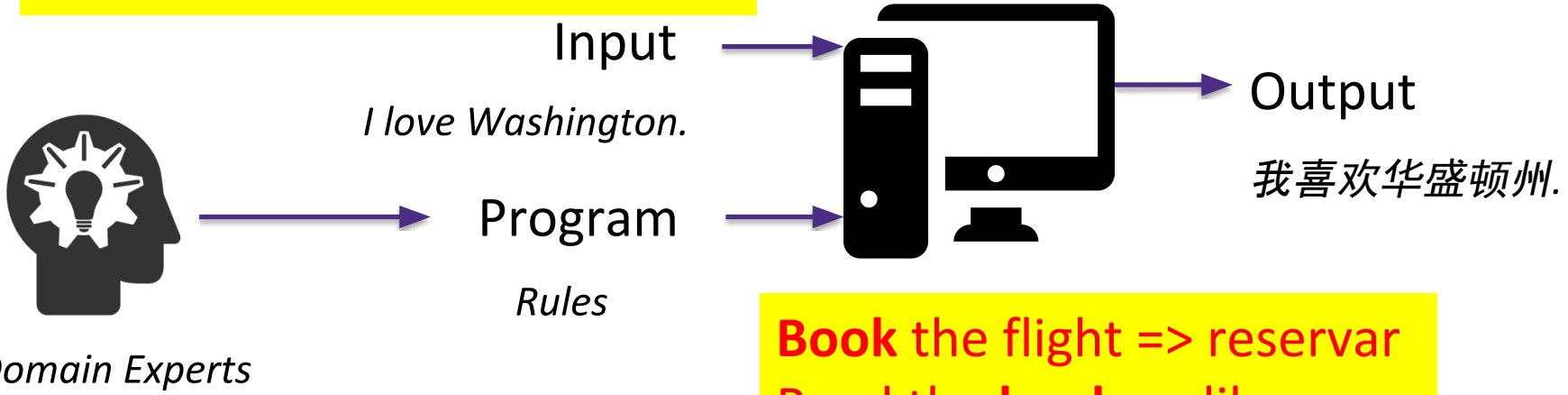
Examples

Machine Learning: Improving ***performance P*** with ***experience E*** at some ***task T.***

Task	Experience	Performance
Predict house price	Historical house sales	Error in price estimate
Recommend products	Customer's transactions	Click through rate
Voice recognition	Human voice clips	Accuracy of transcription
Play Go	Practice games of Go	% of games won

Machine Translation: Traditional Programming

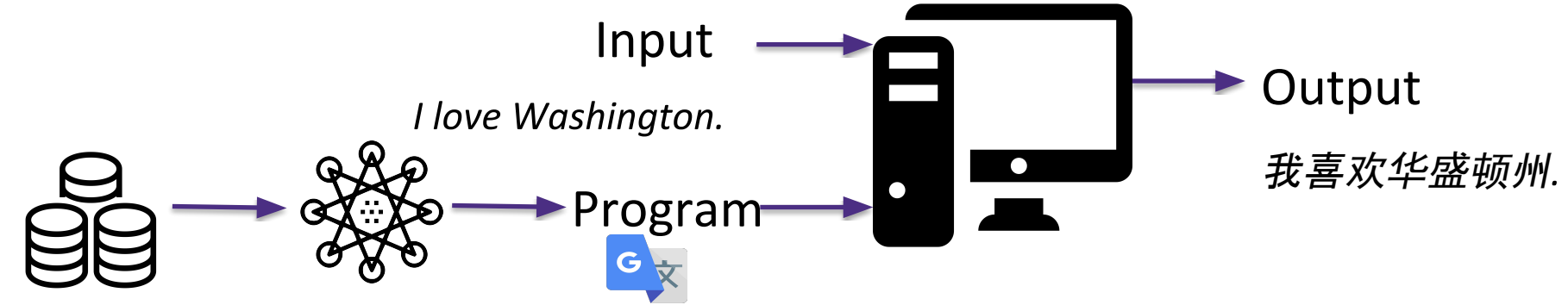
English: subject-verb-object
Japanese: subject-object-verb



Book the flight => reservar
Read the **book** => libro

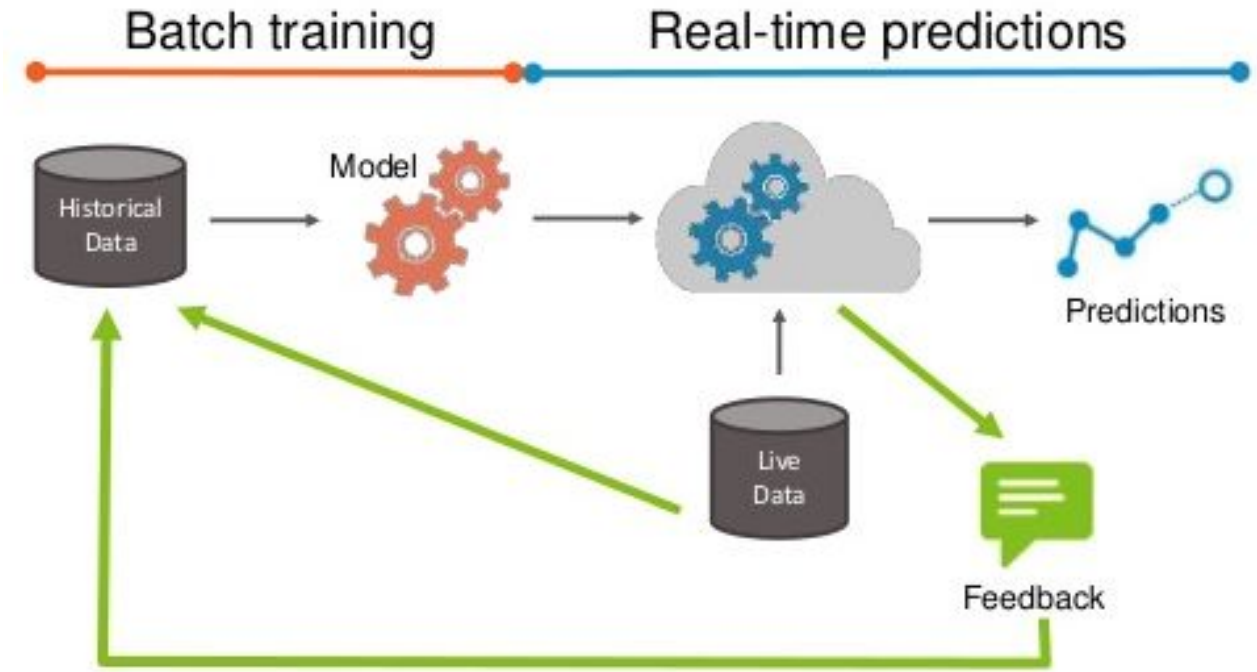
Rule-based machine translation. (1950 – 2000)

Machine Translation: Machine Learning



Google Machine Translation. (2013 – now)

Feedback Loop on Continuous Learning



Quiz

Can you think of any scenarios when ML is preferred to human?

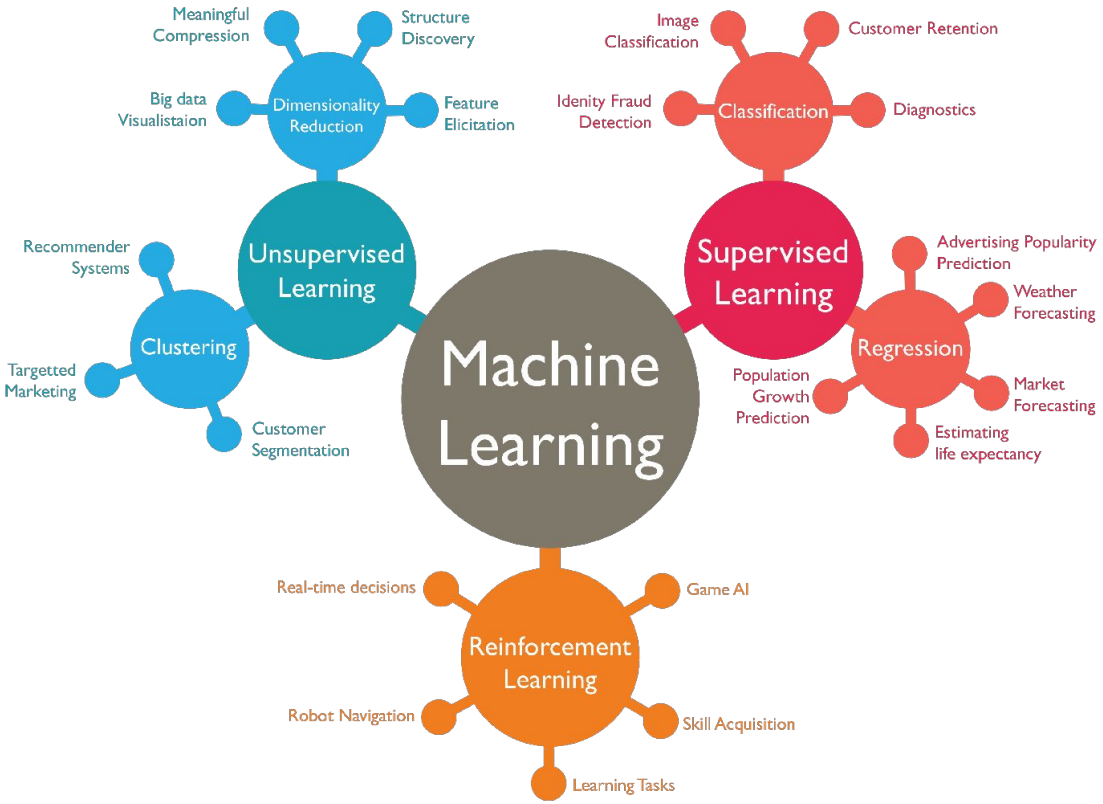
When do we need Machine Learning?

- Humans cannot do it or can do it but cannot describe how they do it (thus cannot be programmed).
  - Effectiveness of a new compound treating some disease (cannot do)
  - Object recognition (cannot describe)

- Machines can do better than humans on certain tasks.
  - Search page ranking (more data)
  - Play the game of Go (more computing power)

- Machines are cheaper than humans.
  - Handwritten digits recognition
  - Voice recognition

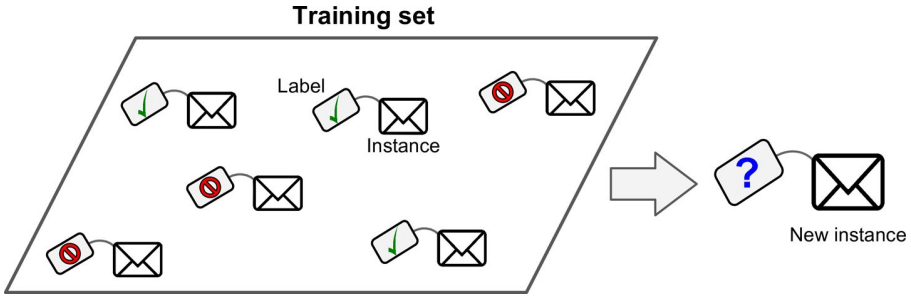
3 Types of Machine Learning Algorithms



Foster School of Business

Supervised Learning

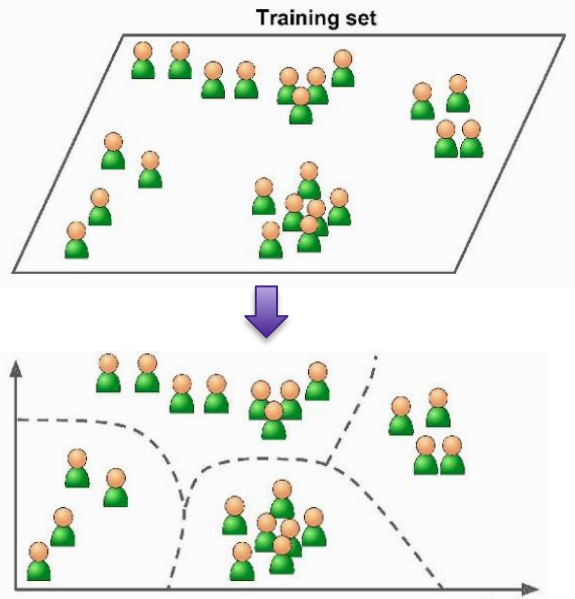
Learn to predict output from input.



- Linear/Logistic Regression
- K Nearest Neighbor
- Decision Tree
- Random Forest
- Support Vector Machine
- (Deep) Neural Network
- ...

Unsupervised Learning

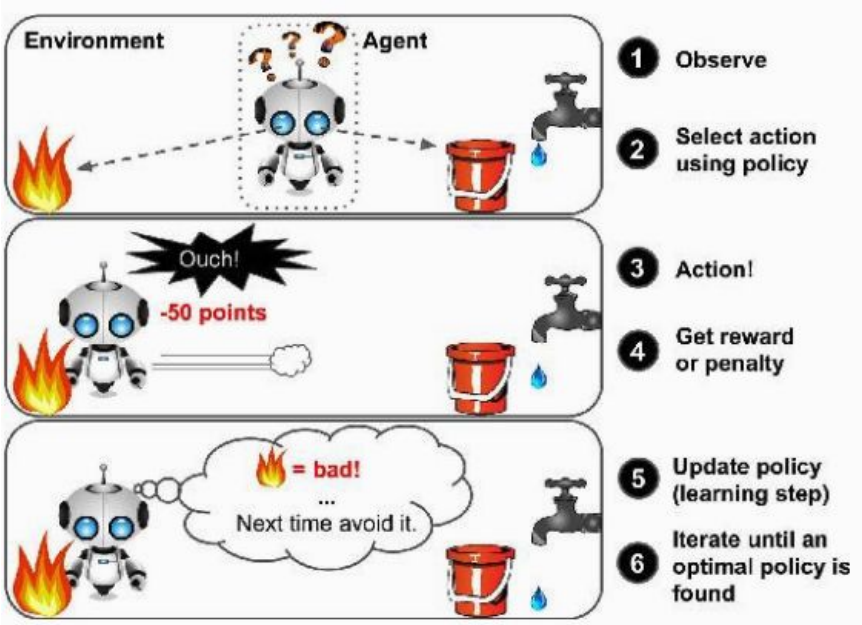
Learn to find patterns and structure in an **unlabeled** date set.



- K means
- Hierarchical clustering
- Gaussian Mixture Model
- DBSCAN
- Spectral clustering
- Mean Shift clustering
- …

Reinforcement Learning

Learn to **take actions** in an environment to maximize expected future rewards.



- Model-based RL
- Model-free RL

Quiz

Task	Learning Type
Home value estimates on Redfin	Supervised
Product recommendations on Amazon	Unsupervised
Voice recognition in Alexa	Supervised
Play the game of Go	Reinforcement
Predict stock price	Supervised
An advertising platform segments the U.S. population into smaller groups with similar demographics and purchasing habits so that advertisers can reach their target market with relevant ads.	Unsupervised
Learn to play StarCraft	Reinforcement
Airbnb groups its housing listings into neighborhoods so that users can navigate listings more easily.	Unsupervised
File tax return	Not a learning problem

Supervised Learning

- **Regression**: A regression model predicts **continuous values**. For example, regression models make predictions that answer questions like the following:
  - What is the value of a house in California?
  - What is the demand of a product on Amazon next month?

- **Classification**: A classification model predicts **discrete values**. For example, classification models make predictions that answer questions like the following:
  - Is a given email message spam or not spam?
  - Is this an image of a dog, a cat, or a hamster?

Supervised Learning Terminology

Target Variable	The thing we're predicting, aka label. - e.g. Future stock price, cat image, spam email
Feature	An input variable, which is used to predict the label. - e.g. Historical stock price, pixels in the image, words in the email body

	Age	Gender	Weight	Height
Observation #1	**12**	**M**	**80** lbs	**55** in
Observation #2	**11**	**M**	**85** lbs	**58** in
Observation #3	**12**	**F**	**73** lbs	**52** in
Observation #4	**10**	**F**	**71** lbs	**49** in
. . . Observation #150				

Features

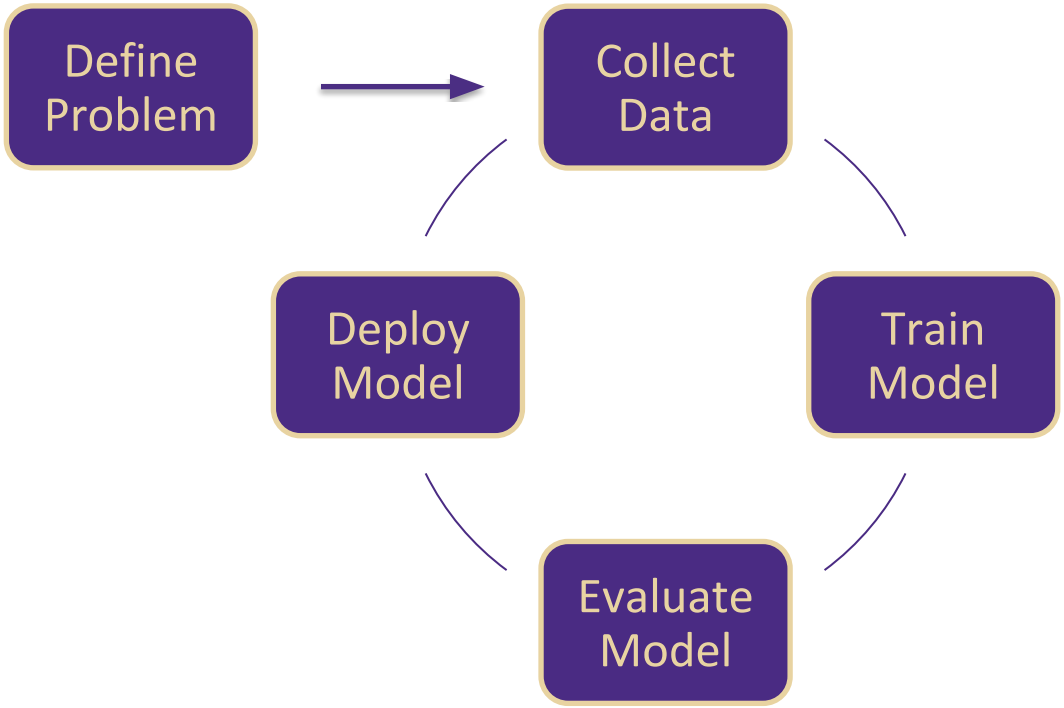
Target Variable

Foster School of Business

Supervised Learning Terminology

Model Training	Build a ML model by learning the relationships between features and label. - e.g. linear regression, decision tree, neural network
Model Scoring	Apply a ML model on unlabeled examples to make predictions. - e.g. linear regression, decision tree, neural network
Model Evaluation	Measure the model's predictive accuracy. - Mean Squared Error, Accuracy, Precision and Recall and AUC.

Supervised Learning Workflow



Confusion Matrix

Misclassification Error: classifying a record as belonging to one class when it belongs to another class.

		Actual Class	
		C1 (Positive)	C0 (Negative)
Predicted Class	C1 (Positive)	Number of records with actual class **1** and predicted class **1**	Number of records with actual class **0** and predicted class **1**
	C0 (Negative)	Number of records with actual class **1** and predicted class **0**	Number of records with actual class **0** and predicted class **0**

ERRORS !!!

Foster School of Business

Classification Performance Measures

Most classification accuracy measures are derived from the **confusion matrix**.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

		Actual Class	
		C1 (Positive)	C0 (Negative)
Predicted Class	C1 (Positive)	True Positive **(TP)**	False Positive **(FP)**
	C0 (Negative)	False Negative **(FN)**	True Negative **(TN)**

Quiz

Fill in the following confusion matrix.

		Actual Class	
		C1 (Positive)	C0 (Negative)
Predicted Class	C1 (Positive)	1	1
	C0 (Negative)	1	2

Actual class	Predicted class
C1	C0
C1	C1
C0	C0
C0	C0
C0	C1

Accuracy = ? 3 / 5

Precision = ? 1 / 2

Recall = ? 1 / 2

Regression Performance Measures

Regression Error: the difference between the predicted value (Z_i) and the actual value (Y_i).

$$\text{Root Mean-Squared Error} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - z_i)^2}$$

$$\text{Mean Absolute Error} = \frac{1}{N}\sum_{i=1}^{N}|y_i - z_i|$$

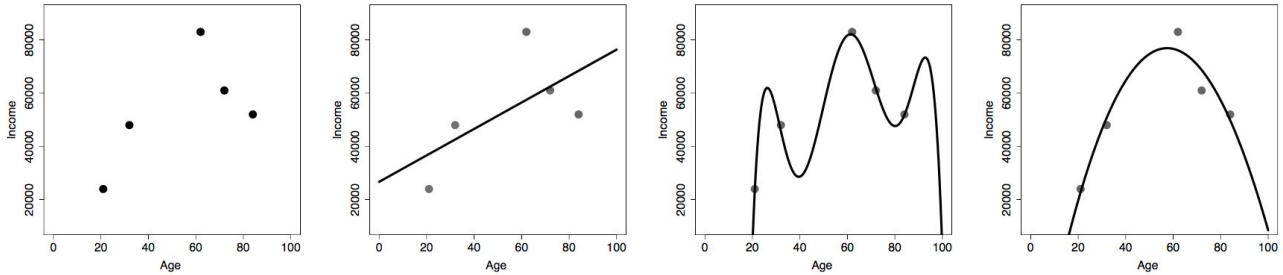
Generalization

Machine Learning is all about **generalization** to future unseen data points.

- **Underfitting** – a model is too simple and can not capture the underlying patterns within the data, thus does not perform well on new data.
- **Overfitting** – a model tries to fit the training data so closely that it does not generalize well to new data.

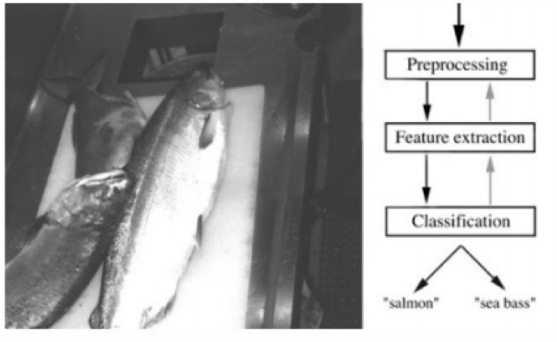
Table: The age-income dataset.

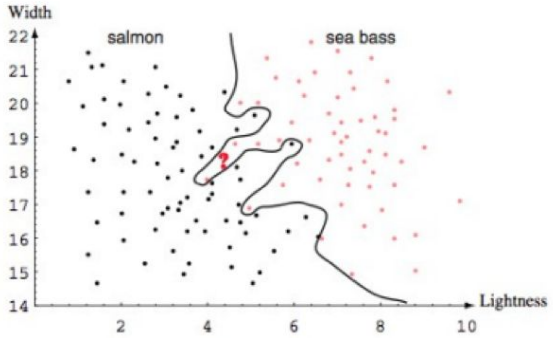
ID	AGE	INCOME
1	21	24,000
2	32	48,000
3	62	83,000
4	72	61,000
5	84	52,000

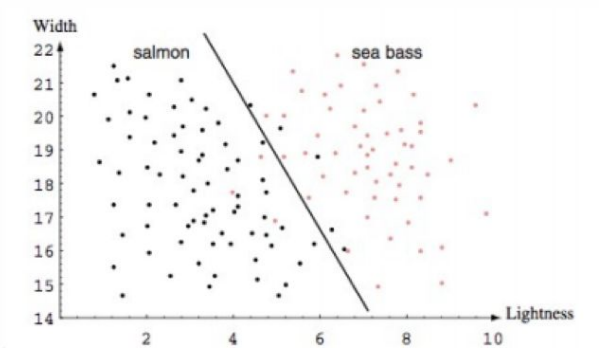


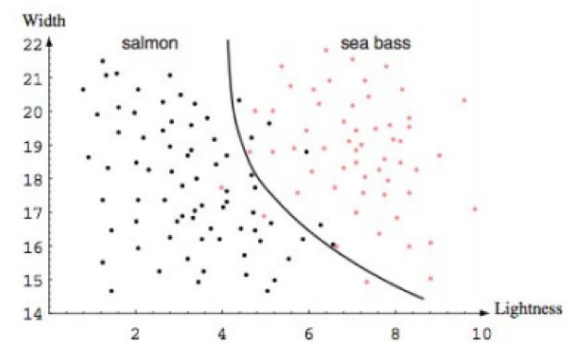
Foster School of Business

Quiz









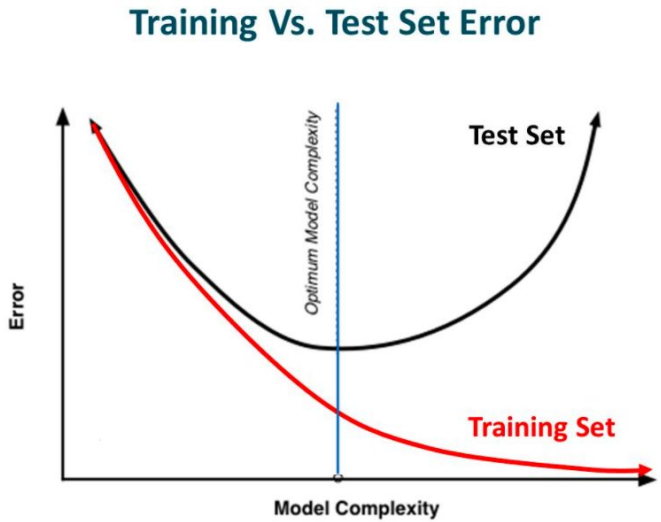
Foster School of Business

Model Selection

- Split the data into two sets:
  - **Training Set**: a subset of the data to train a model.
  - **Test Set**: a subset of the data to evaluate the performance of the trained model.

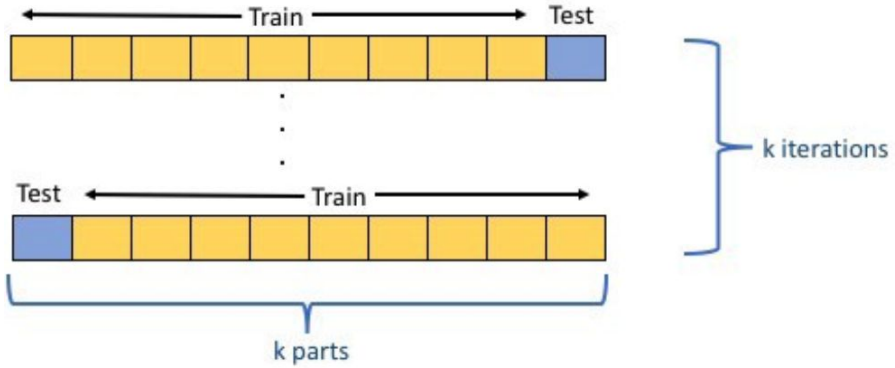
- Plot the errors of models with increasing complexity on both train and test sets.

- Choose the model which optimize the error on the test set.



K-fold Cross-validation

1. Divide the training data into K parts.
2. Use K-1 of the parts for training and 1 for testing.
3. Repeat the procedure K times, rotating the test set.
4. Determine the performance based on the results across all K iterations.



Leave-one-out Cross-validation is the extreme case of K-fold Cross-Validation where we keep only one data point in the test set.

Quiz

What are the leave-one-out cross-validation errors for the following dataset, using 1 nearest neighbor (1-NN) and 3 nearest neighbor (3-NN)?

Answers: 1-NN: 5/10. 3-NN: 1/10.

Lab

Python Machine Learning Libraries

- ***Pandas*** provides a DataFrame object along with a powerful set of methods to manipulate, filter, group, and transform data.

- ***Matplotlib*** provides a useful interface for creation of high-quality plots and figures.

- ***Seaborn*** is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

- ***Scikit-Learn*** provides a uniform toolkit for applying common machine learning algorithms to data.