GENERAL ASSEMBLY

COMMUNICATING RESULTS

Abbas Chokor, Ph.D.

Staff Data Scientist, Seagate Technology

OUR PROGRESS SO FAR

UNIT 1: RESEARCH DESIGN AND EXPLORATORY DATA ANALYSIS

- ~~What is Data Science~~ ~~Lesson 1~~
- ~~Research Design and Pandas~~ ~~Lesson 2~~
- ~~Statistics Fundamentals I~~ ~~Lesson 3~~
- ~~Statistics Fundamentals II~~ ~~Lesson 4~~
- ~~Flexible Class Session~~ ~~Lesson 5~~

UNIT 2: FOUNDATIONS OF DATA MODELING

- ~~Introduction to Regression~~ ~~Lesson 6~~
- ~~Evaluating Model Fit~~ ~~Lesson 7~~
- ~~Introduction to Classification~~ ~~Lesson 8~~
- ~~Introduction to Logistic Regression~~ ~~Lesson 9~~
- Communicating Logistic Regression Results Lesson 10
- Flexible Class Session Lesson 11

Today's Class

UNIT 3: DATA SCIENCE IN THE REAL WORLD

- Decision Trees and Random Forests Lesson 12
- Natural Language Processing Lesson 13
- Dimensionality Reduction Lesson 14
- Time Series Data I Lesson 15
- Time Series Data II Lesson 16
- Database Technologies Lesson 17
- Where to Go Next Lesson 18
- Flexible Class Session Lesson 19
- Final Project Presentations Lesson 20

LAST CLASS

WHAT DID WE LEARN?

- Build a Logistic regression classification model using the statsmodels library
- Describe a sigmoid function, odds, and the odds ratio as well as how they relate to logistic regression
- Evaluate a model using metrics such as classification accuracy/error, confusion matrix, ROC/AUC curves, and loss functions

You got all objectives?

Not all of them…

Let's form groups of 1's and 2's …

LAST CLASS

ANNOUNCEMENTS

- Mid-class survey
- Moving to Rino Station starting next Thursday December 14th
- Parking is free on the streets behind the Rino Station building, and all of your key fobs should work now at Rino Station as well
- You will need to return your parking garage passes by next week.

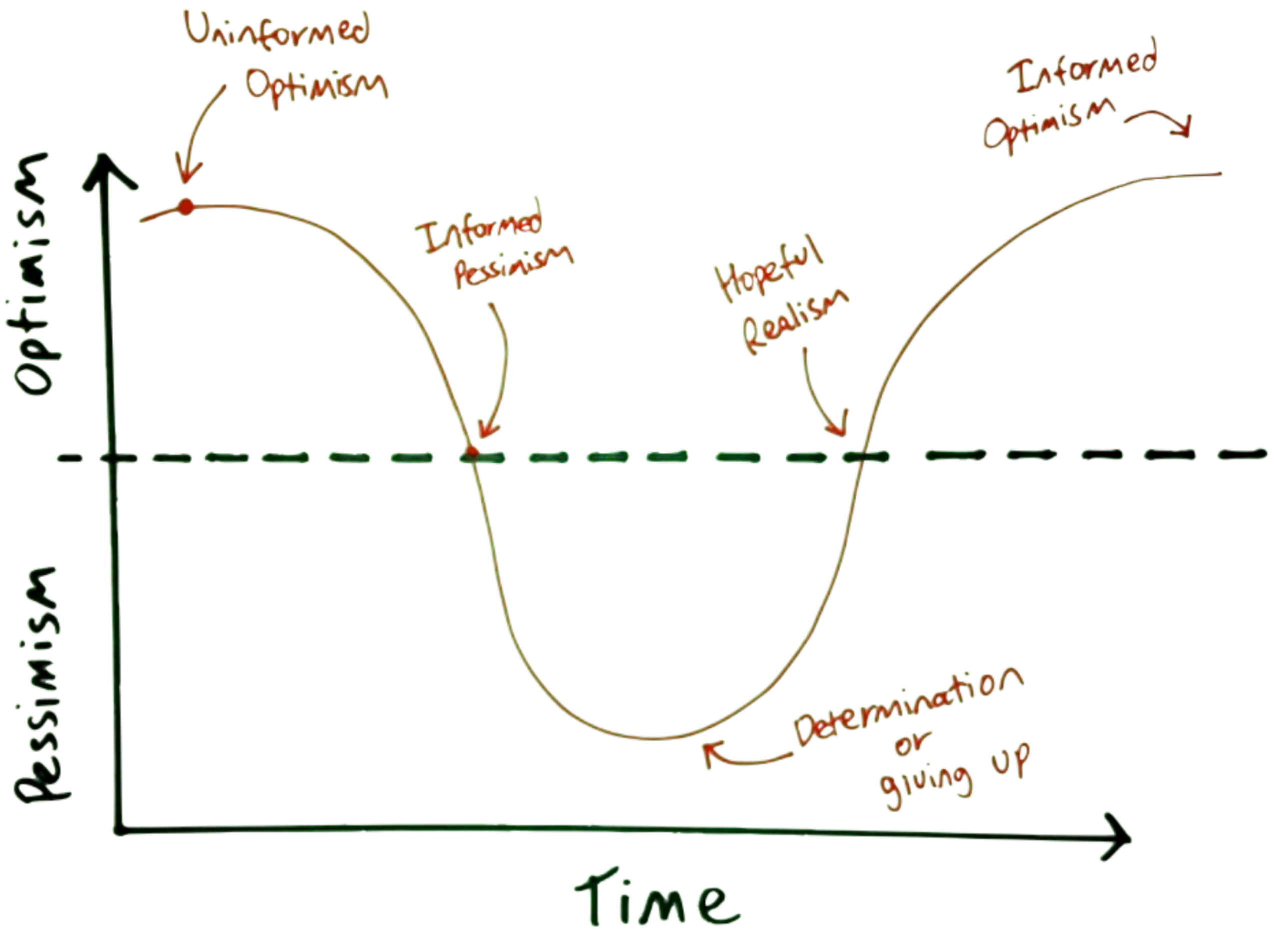
I still need to practice getting the data prepared for plotting and practicing the plotting syntax

Others?

REFRESH YOUR KNOWLEDGE

BIG IMAGE

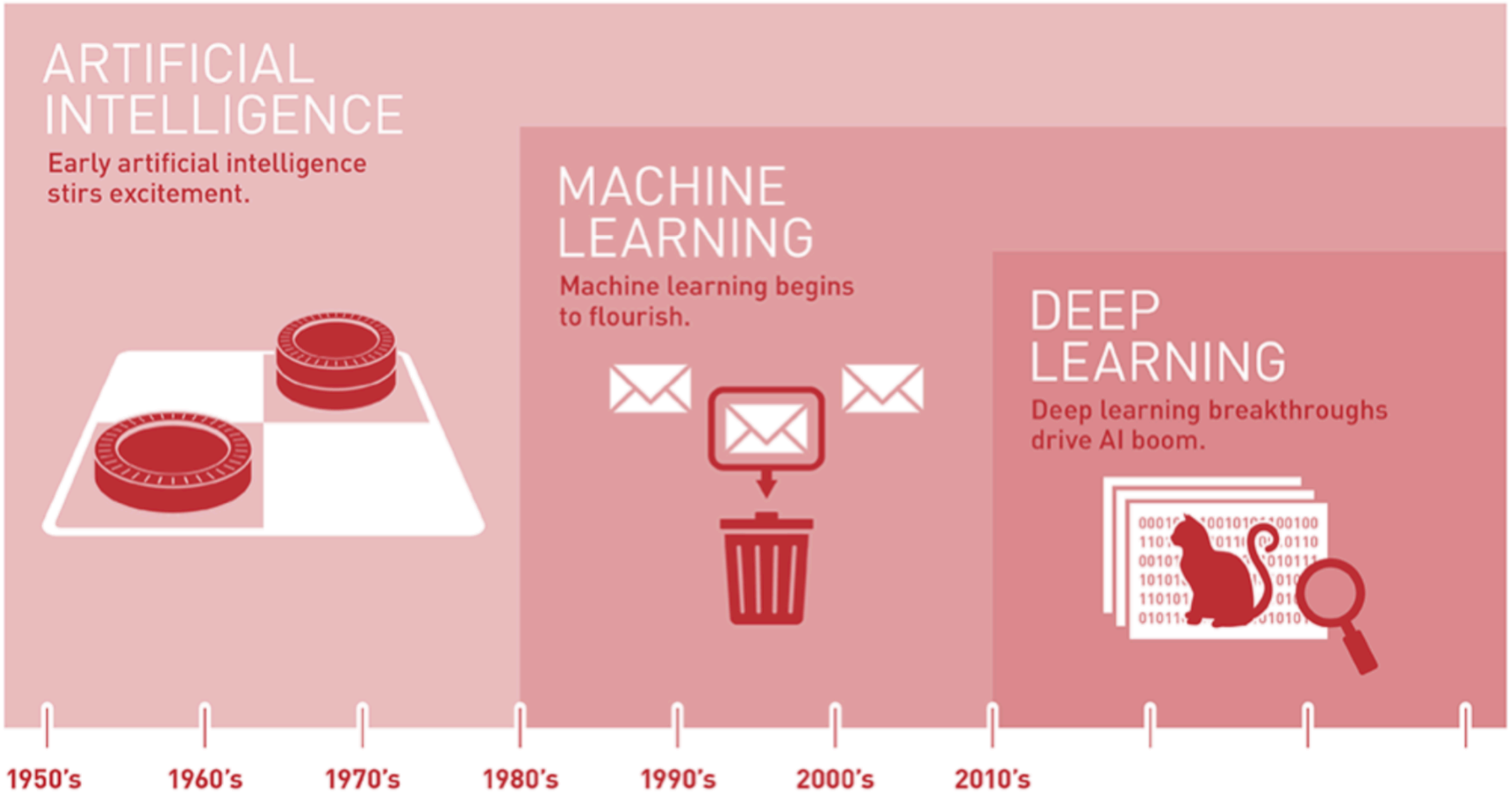
Where Are You Now?



WHAT IS MACHINE LEARNING

[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

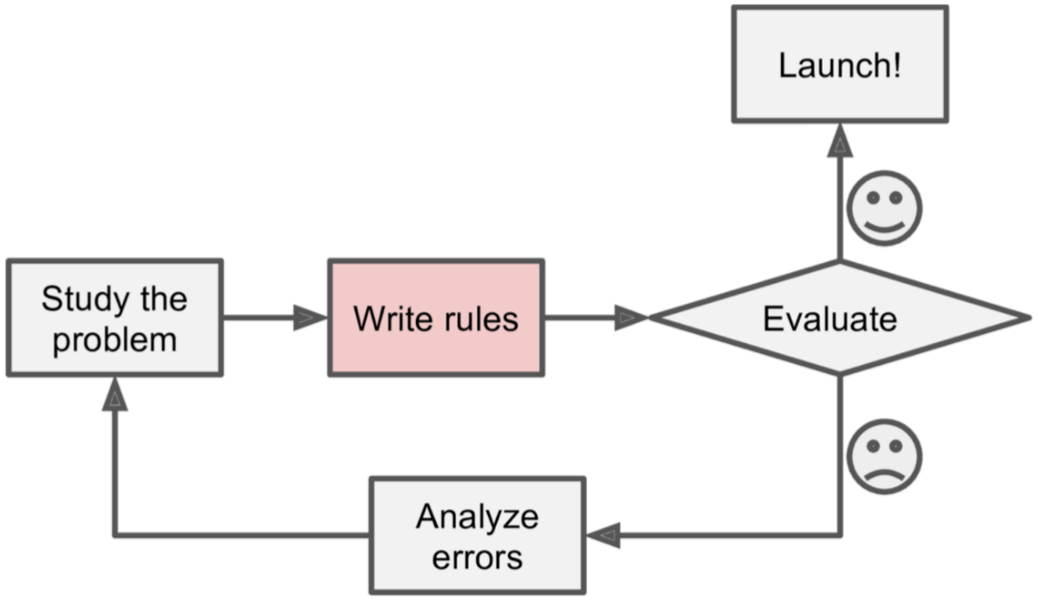
Arthur Samuel, *1959*

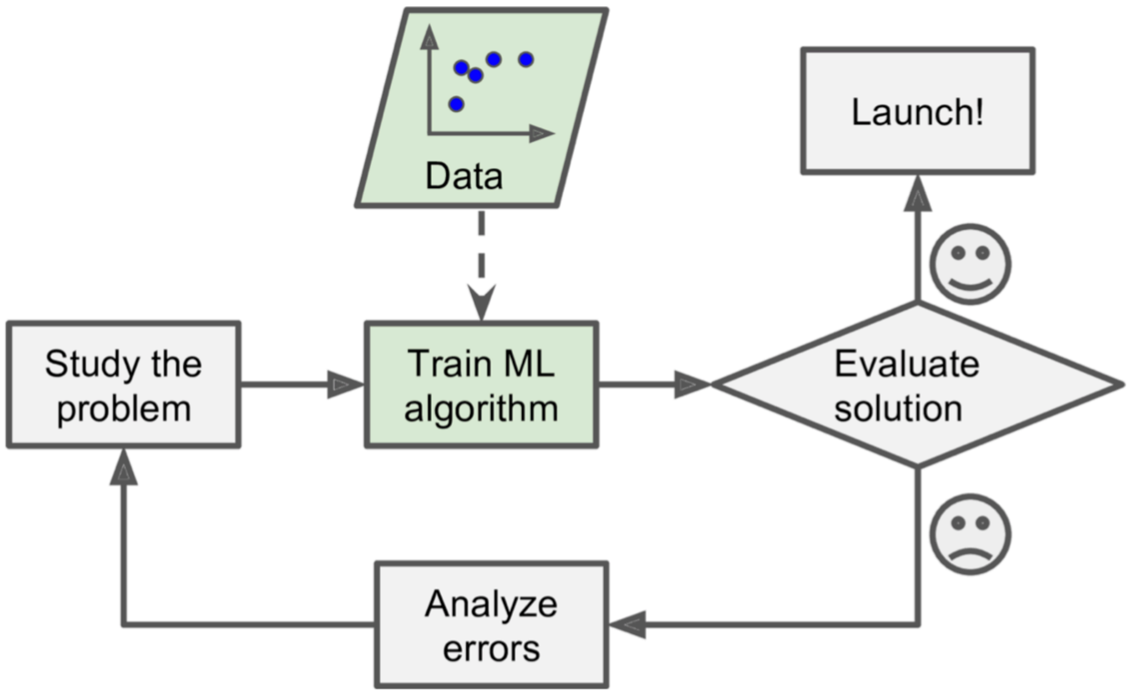


Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

<u>Machine Learning</u>: is the science (and art) of programming computers so they can *learn from data*.

WHY TO USE MACHINE LEARNING?



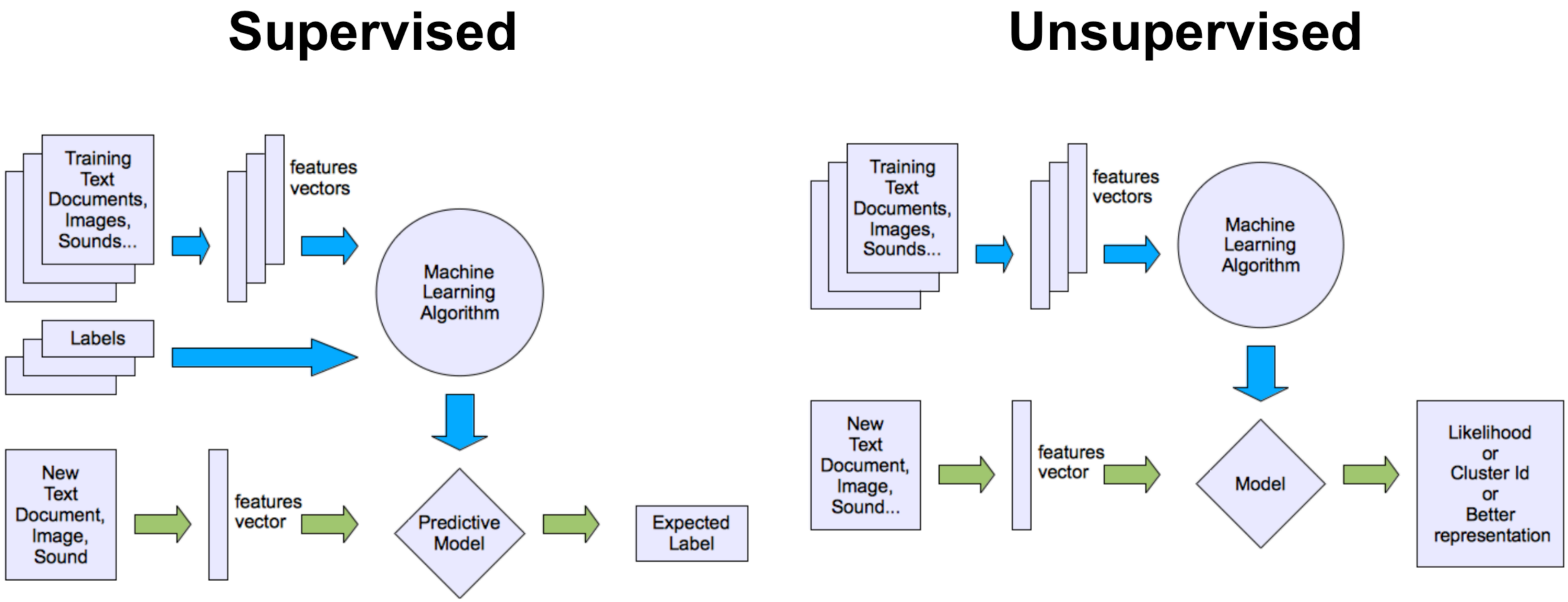


Traditional approach **Vs.** ***Machine Learning approach***

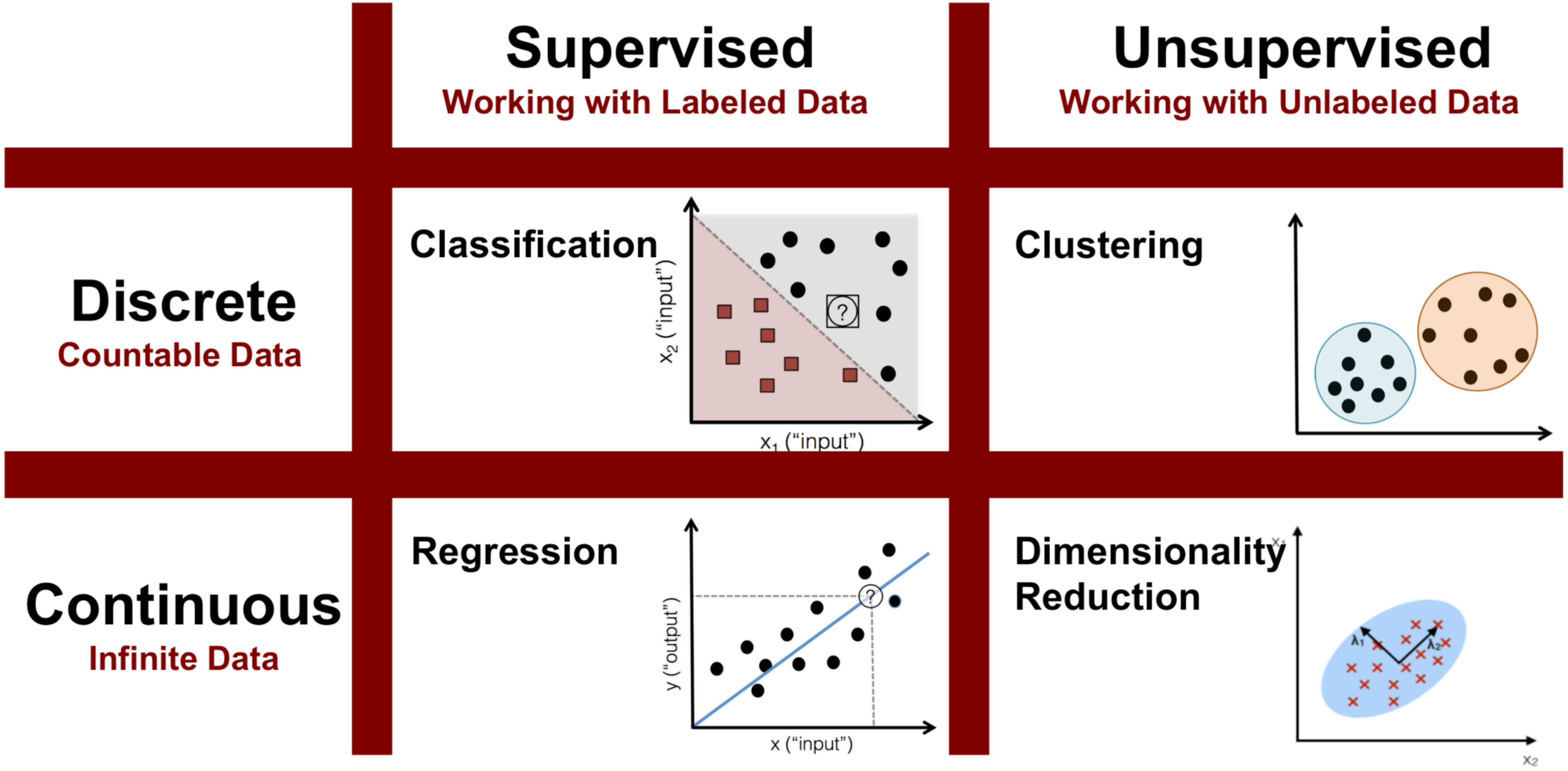
Did you know?

Machine learning algorithms are expected to replace 25% of the jobs across the world, in the next 10 years.

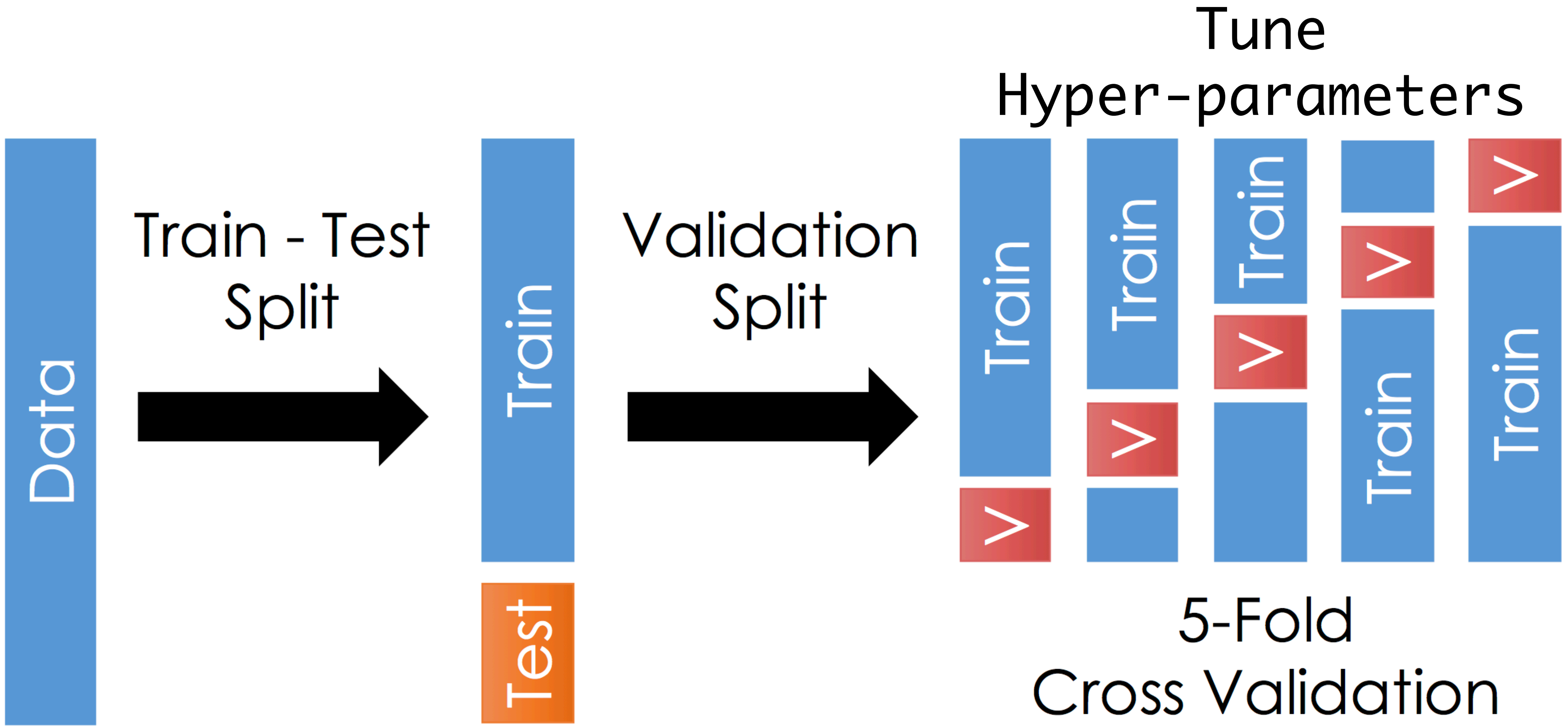
TYPES OF MACHINE LEARNING



TYPES OF MACHINE LEARNING



TRAINING – VALIDATING - TESTING



LEARNING OBJECTIVES

- Explain the trade-offs between the precision and recall of a model while articulating the cost of false positives vs. false negatives
- Describe the difference between visualization for presentations vs. exploratory data analysis
- Practice, practice, and practice!

OPENING

COMMUNICATING RESULT

WE BUILT A MODEL! NOW WHAT?

- We've built our model, but there is still a gap between your Notebook with plots/figures and a slideshow needed to present your results.

- Classes so far have focused on two core concepts:
    - developing consistent practices
    - interpreting metrics to evaluate and improve model performance

- But what does that mean to your audience?

WE BUILT A MODEL! NOW WHAT?

- Imagine how a non-technical audience might respond to the following statements:
  - The predictive model I built has an accuracy of 80%.
  - Logistic regression was optimized with L2 regularization.
  - Gender was more important than age in the predictive model because it has a larger coefficient.
  - Here’s the AUC chart that shows how well the model did.

WE BUILT A MODEL! NOW WHAT?

- Who is your audience? Are they technical? What are their concerns?

- Remember: in a business setting, you may be *the only person* who can interpret what you've built.

- Some people may be familiar with basic visualization.

- You need to be able to efficiently explain your results in a way that makes sense to **all** stakeholders (technical or not).

WE BUILT A MODEL! NOW WHAT?

- Today, we'll focus on communicating results for "simpler" problems, but this applies to any type of model you may work with.

- First, let's review classification metrics, review our knowledge, and talk about how we might communicate what we know.

REVIEW

BACK TO THE CONFUSION MATRIX

BACK TO THE CONFUSION MATRIX

- Confusion matrices allow for the interpretation of correct and incorrect predictions for *each class label*.

- It is the first step for the majority of classification metrics and goes deeper than just accuracy.

BACK TO THE CONFUSION MATRIX

- Let's recall our confusion matrix.

		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

condition positive (P)
 the number of real positive cases in the data

condition negatives (N)
 the number of real negative cases in the data

---

true positive (TP)
 eqv. with hit

true negative (TN)
 eqv. with correct rejection

false positive (FP)
 eqv. with false alarm, Type I error

false negative (FN)
 eqv. with miss, Type II error

sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

specificity or true negative rate (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{N} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$$

precision or positive predictive value (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

negative predictive value (NPV)

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}}$$

accuracy (ACC)

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{P+N} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$

F1 score
 is the harmonic mean of precision and sensitivity

$$F_1 = 2\cdot\frac{\mathrm{PPV}\cdot\mathrm{TPR}}{\mathrm{PPV}+\mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$$

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

1. Without looking at the previous slide, how do we calculate the following?
   a. Accuracy
   b. True positive rate
   c. False positive rate

DELIVERABLE

Answers to the above questions

INTRODUCTION

PRECISION AND RECALL

PRECISION AND RECALL

- Our previous metrics were primarily designed for less biased data problems: we could be interested in both outcomes, so it was important to generalize our approach.

- For example, we may be interested if a person will vote for a Republican or Democrat. This is a binary problem, but we're interested in both outcomes.

PRECISION AND RECALL

- Precision and recall, metrics built from the confusion matrix, focus on *information retrieval*, particularly when one class is more interesting than the other.

- For example, we may want to predict if a person will be a customer. We care much more about people who will be a customer of ours than people who won't.

PRECISION AND RECALL

- Precision asks, “Out of all of our positive predictions (both true positive and false positive), how many were correct?”
- Recall asks, “Out of all of our positive class labels, how many were correct?”

THE DIFFERENCE BETWEEN PRECISION AND RECALL

- The key difference between the two is the attribution and value of error.
- Should our model be more pick in avoiding false positives (precision)?
- Or should it be more pick in avoiding false negatives (recall)?
- The answer should be determined by the problem you’re trying to solve.

GUIDED PRACTICE

COST BENEFIT ANALYSIS

ACTIVITY: COST BENEFIT ANALYSIS

EXERCISE

DIRECTIONS

One tool that complements the confusion matrix is cost-benefit analysis, where you attach a *value* to correctly and incorrectly predicted data.

Like the Precision-Recall trade off, there is a balancing point to the *probabilities* of a given position in the confusion matrix, and the *cost* or *benefit* to that position. This approach allows you to not only add a weighting system to your confusion matrix, but also to speak the language of your business stakeholders (i.e. communicate your values in dollars!).

ACTIVITY: COST BENEFIT ANALYSIS

EXERCISE

DIRECTIONS

Consider the following marketing problem:

As a data scientist working on marketing spend, you've build a model that reduces user churn--the number of users who decide to stop paying for a product--through a marketing campaign. Your model generates a confusion matrix with the following probabilities (these probabilities are calculated as the value in that position over the sum of the sample):

```
| TP: 0.2 | FP: 0.2 |
---------------------
| FN: 0.1 | TN: 0.5 |
```

ACTIVITY: COST BENEFIT ANALYSIS

EXERCISE

DIRECTIONS (15 minutes)

In this case:

- The *benefit* of a true positive is the retention of a user ($10 for the month)
- The *cost* of a false positive is the spend of the campaign per user ($0.05)
- The *cost* of a false negative (someone who could have retained if sent the campaign) is, effectively, 0 (we didn't send it... but we certainly didn't benefit!)
- The *benefit* of a true negative is 0: No spend on users who would have never retained.

To calculate Cost-Benefit, we'll use this following function:

```
(P(TP) * B(TP)) + (P(TN) * B(TN)) + (P(FP) * C(FP)) + (C(FN) * C(FN))
```

which for our marketing problem, comes out to this:

```
(.2 * 10) + (.5 * 0) - (.2 * .05) - (.1 * 0)
```

or $1.99 per user targeted.

ACTIVITY: COST BENEFIT ANALYSIS

EXERCISE

FOLLOW UP QUESTIONS

Think about precision, recall, and cost benefit analysis to answer the following questions:

1. How would you rephrase the business problem if your model was optimizing toward *precision*? i.e., How might the model behave differently, and what effect would if have?

2. How would you rephrase the business problem if your model was optimizing toward *recall*?

3. What would the most ideal model look like in this case?

DELIVERABLE

Answers to the above questions

INTRODUCTION

SHOWING WORK

SHOWING WORK

- We’ve spent a lot of time exploring our data and building a reasonable model that performs well.

- However, if we look at our visuals, they are most likely:

    - Statistically heavy: Most people don’t understand histograms.

    - Overly complicated: Scatter matrices produce too much information.

    - Poorly labeled: Code doesn’t require adding labels, so you may not have added them.

SHOWING WORK

- In order to convey important information to our audience, make sure our charts are:
    - Simplified
    - Easily interpretable
    - Clearly labeled

SIMPLIFIED

- At most, you’ll want to include figures that either explain a variable on its own or explain that variable’s relationship with a target.

- If your model used a data transformation (like natural log), just visualize the original data.

- Try to remove any unnecessary complexity.

EASILY INTERPRETABLE

- Any stakeholder looking at a figure should be seeing the exact same thing you're seeing.
- A good test for this is to share the visual with others less familiar with the data and see if they come to the same conclusion.
- How long did it take them?

CLEARLY LABELED

- Take the time to clearly label your axis, title your plot, and double check your scales - especially if the figures should be comparable.

- If you’re showing two graphs side by side, they should follow the same Y axis.

QUESTION TO ASK

- When building visuals for another audience, ask yourself these questions:
  - **Who**: Who is my target audience for the visual?
  - **What**: What do they already know about this project? What do they need to know?
  - **How**: How does my project affect this audience? How might they interpret (or misinterpret) the data?

INDEPENDENT PRACTICE

PROJECT PRACTICE – IRIS DATASET

INDEPENDENT PRACTICE

PROJECT PRACTICE - MARKETING

INDEPENDENT PRACTICE

PROJECT PRACTICE - AFFAIR

CONCLUSION

TOPIC REVIEW

REVIEW AND NEXT STEPS

- What do precision and recall mean? How are they similar and different to True Positive Rate and False Positive Rate?

- How does cost benefit analysis play a role in building models?

- What are at least two very important details to consider when creating visuals for a project's stakeholders?

COURSE

BEFORE NEXT CLASS

OUR PROGRESS SO FAR

UNIT 1: RESEARCH DESIGN AND EXPLORATORY DATA ANALYSIS

- ~~What is Data Science~~ ~~Lesson 1~~
- ~~Research Design and Pandas~~ ~~Lesson 2~~
- ~~Statistics Fundamentals I~~ ~~Lesson 3~~
- ~~Statistics Fundamentals II~~ ~~Lesson 4~~
- ~~Flexible Class Session~~ ~~Lesson 5~~

UNIT 2: FOUNDATIONS OF DATA MODELING

- ~~Introduction to Regression~~ ~~Lesson 6~~
- ~~Evaluating Model Fit~~ ~~Lesson 7~~
- ~~Introduction to Classification~~ ~~Lesson 8~~
- ~~Introduction to Logistic Regression~~ ~~Lesson 9~~
- ~~Communicating Logistic Regression Results~~ ~~Lesson 10~~
- Flexible Class Session Lesson 11

Next Class

UNIT 3: DATA SCIENCE IN THE REAL WORLD

- Decision Trees and Random Forests Lesson 12
- Natural Language Processing Lesson 13
- Dimensionality Reduction Lesson 14
- Time Series Data I Lesson 15
- Time Series Data II Lesson 16
- Database Technologies Lesson 17
- Where to Go Next Lesson 18
- Flexible Class Session Lesson 19
- Final Project Presentations Lesson 20

BEFORE NEXT CLASS

UPCOMING

- Project: Unit Project 3 due Thursday Dec 14^{th}

LESSON

Q & A

LESSON

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET