Week 1

Concepts in Machine Learning

fredhutch.io

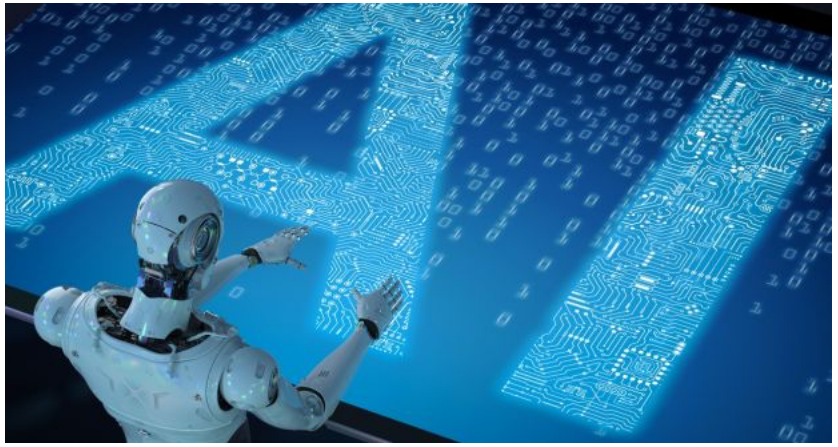
Fred Hutchinson Cancer Research Center

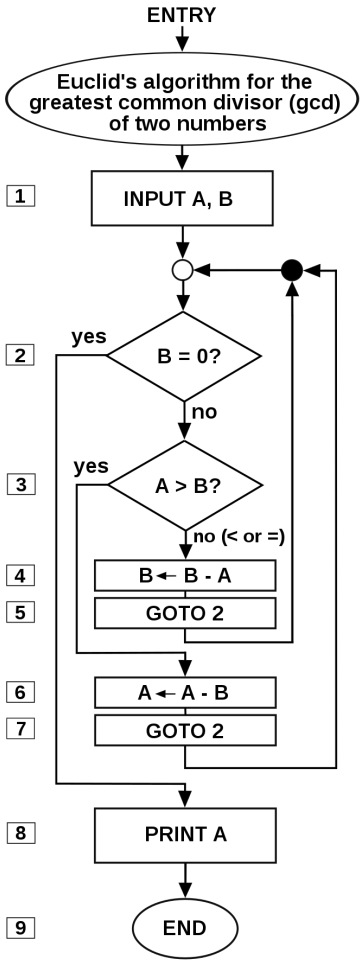
Week 1 Learning Objectives

- Machine Learning and Data Science
  - Working definitions & Connection
- CRISP-DM
  - Connection with Experimental Design
  - Steps & Cyclical nature
- Machine Learning
  - What it is and isn’t
  - Key paradigms
    - Supervised learning
    - Unsupervised learning

What is Machine Learning?

- A branch of AI?
- Algorithms?
- Magic/Special Sauce?

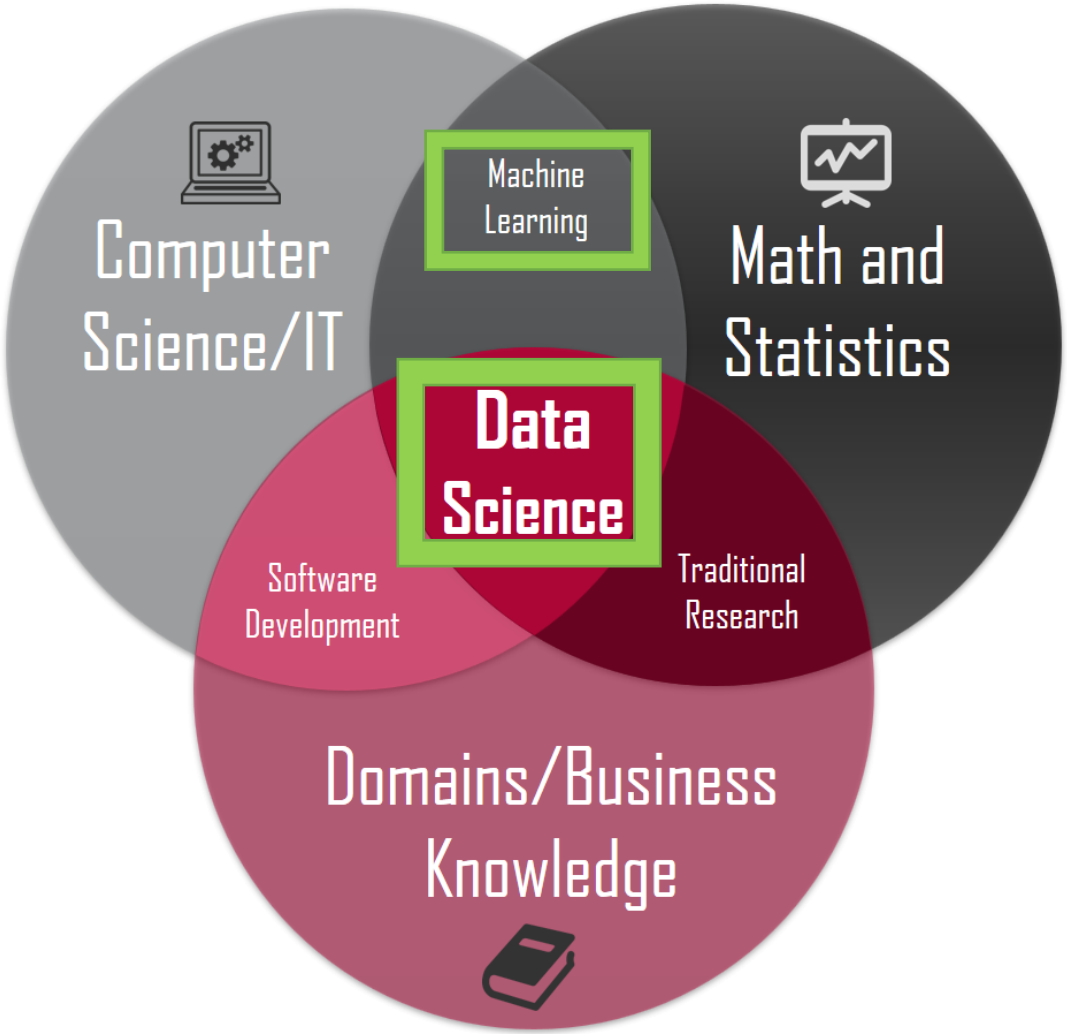




Definitions

- Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. (https://expertsystem.com/machine-learning-definition/)
- Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. (https://en.wikipedia.org/wiki/Machine_learning)
- Problem + Data + Algorithm(self-adjusting) + Compute ==> Insight

Data Science



Cabinet Making?

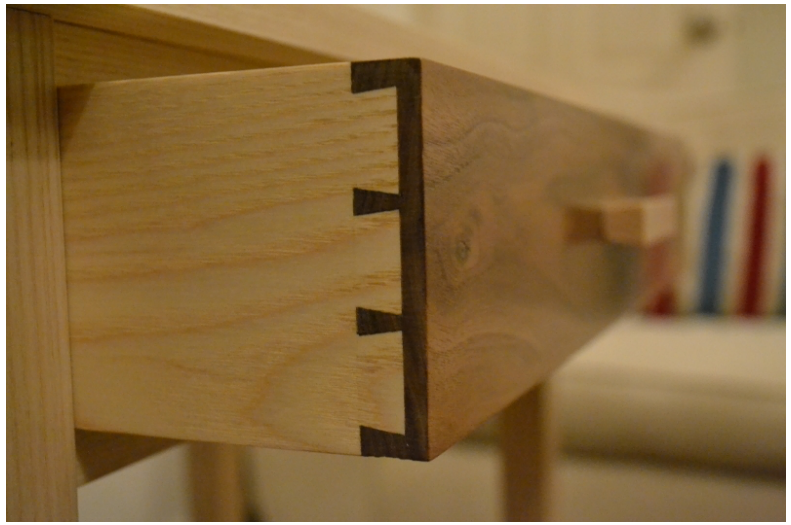
- Storage Need + Raw Materials + Tools + Work ==> Cabinet

- SAT analogy format

  Tools : Cabinet Making as Algorithms : Machine Learning

Capable Cabinet Maker

- Inspects and understands raw materials
- Uses the tools thoughtfully to shape and join materials
- Chooses approach and tools based on materials and goals
- Thoughtfulness born of experience



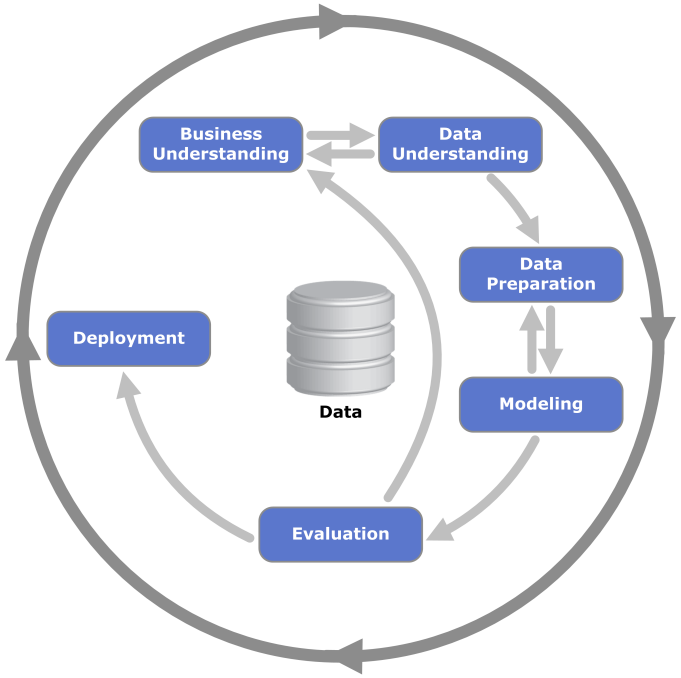
Experimental Design

- Difficult to master or even do well
- Close interplay between
  - Goals
  - Methods
  - Data
  - Execution
- Requires thoughtful approach and broad understanding

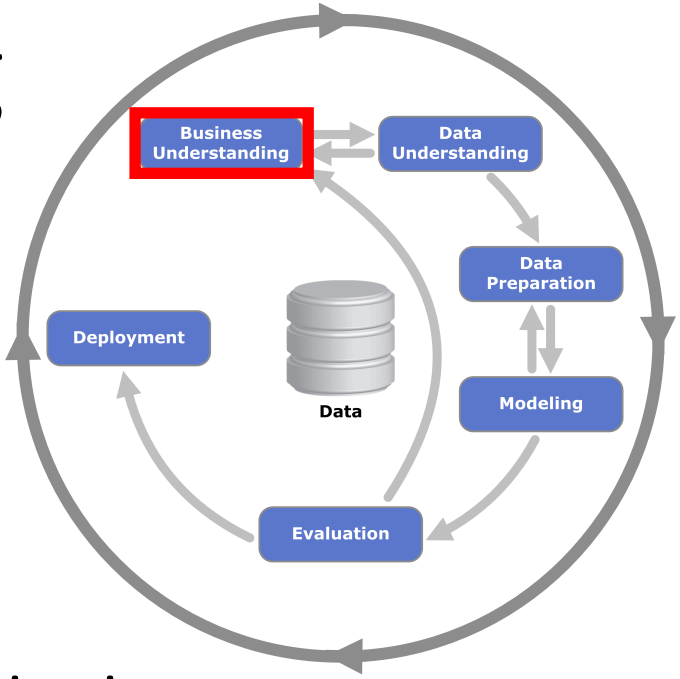
CRISP-DM
Cross-industry standard process for data mining

CRISP-DM
Cross-industry standard process for data mining

- Conceived in 1996
- Newer variations
- Still most widely used analytics model
- Six major phases
  - Canonical image
    - Cyclical and iterative
  - This course focuses on the first 5 phases
- Requires thoughtful approach and broad understanding



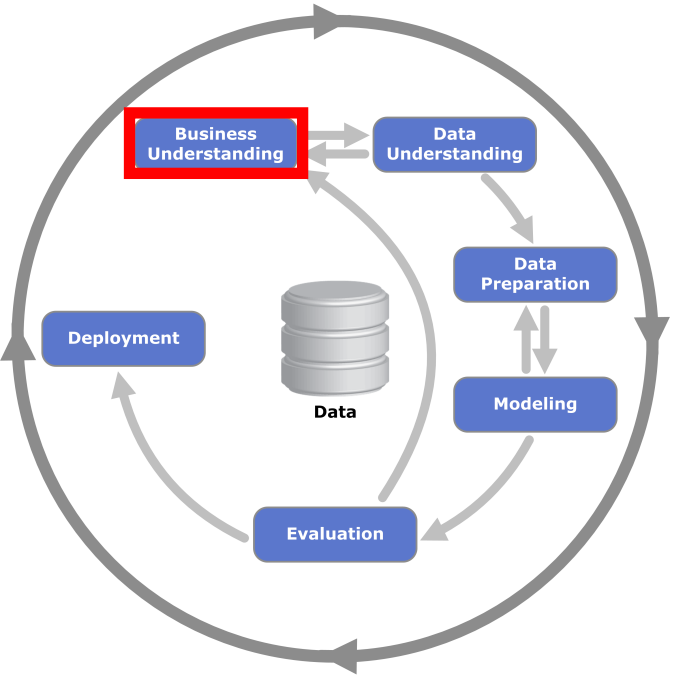
Business/Scientific Understanding



- Understanding the business/research goal
  - Scientific domain knowledge
- Situation assessment
  - What do we already know or believe?
- Translating the business goal in a data mining objective
  - What are our research aims?
- Development of a project plan
  - How do we think we should proceed initially?

Business/Scientific Understanding II

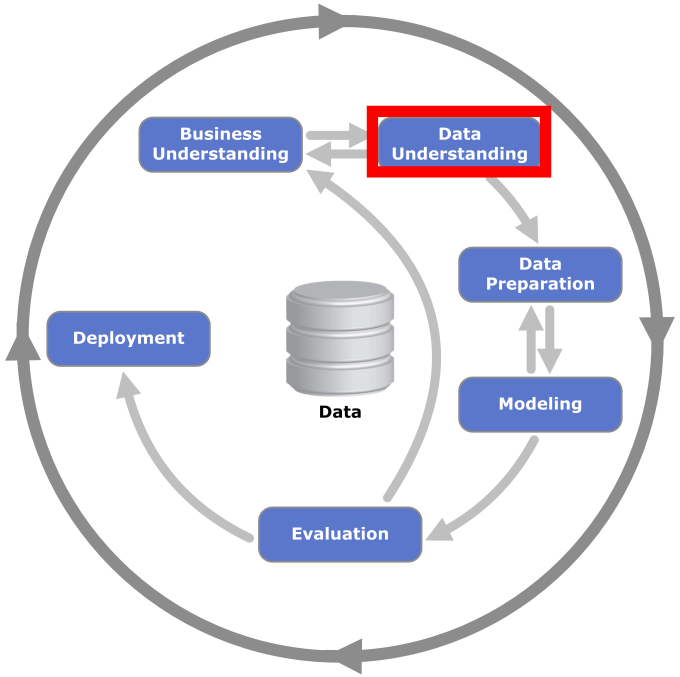
- Important DS Note:
  - Prediction vs Inference
    - Different standards and assumptions
      - Classic source of disagreement between statisticians and data scientists
    - Interested in “Truth” or just predictive performance?



Data Understanding I

- Important DS Note:
  - Observational Data vs Experimental Data
    - Are we creating/collecting data or just using it?
    - ML arose in a largely observation context
    - Don’t draw causal inferences from observational data
- Considering data requirements
  - Do promising approaches have specific data needs?

Data Understanding II



- Initial data collection
  - If necessary/applicable
- Exploratory Data Analysis (EDA)
  - Data structure
  - Data quality
  - First insights
  - Interesting subsets
  - Form hypotheses for hidden information

Data Preparation

- Selection of required data
- Data acquisition
  - If necessary/applicable
- Data integration and formatting
  - Diverse sources
  - Feature format
- Data cleaning
  - Missing values
  - Imputation
- Data transformation and enrichment
  - Feature engineering
  - Alignment with candidate model input specifications

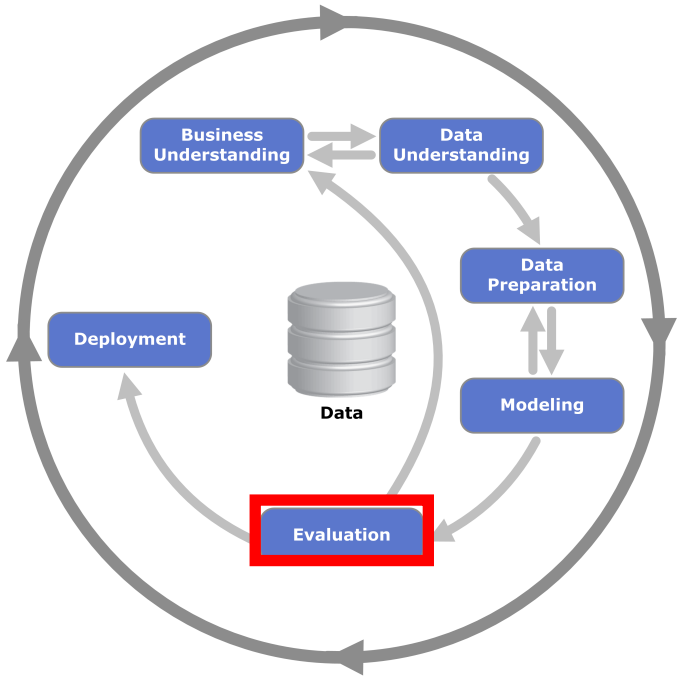
Modeling I

- Selection of appropriate modeling technique
  - Start with a simpler base model
    - If it's good enough, you're done
    - Important DS Note:
      - Don't create/fit a wildly complex model before you know you need it
- […] Splitting of the dataset into training and testing subsets for evaluation purposes
  - Training subset used for fitting model
  - Testing subset used for evaluation purposes
  - Important to train and test on separate data

Modeling II

- Development and examination of alternative modeling algorithms and parameter settings
  - Loss function
    - Method of evaluating how well a specific algorithm models the given data
    - Algorithm “learns” by optimizing the chosen loss function
    - Many to choose from
      - Not one-size fits all
- Fine tuning of the model settings according to an initial assessment of the model’s performance

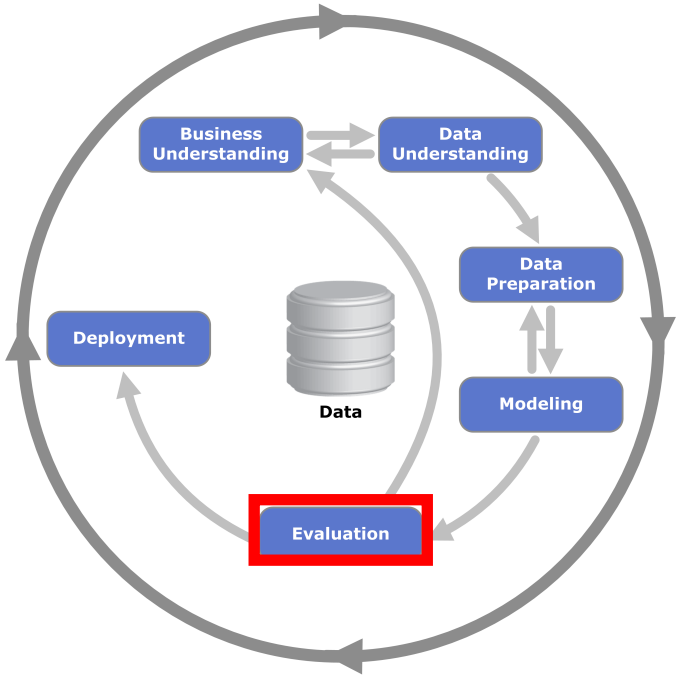
Evaluation I



- Evaluation of the model in the context of the business success criteria
  - Important DS Note:
    - Accuracy Fallacy
      - 99% Accurate HIV test
      - Only works with balanced data sets
- Loss function
  - Method of evaluating how well a specific algorithm models the given data
  - Algorithm to “learns” by optimizing the chosen loss function
  - Many to choose from
    - Not one-size fits all

Evaluation II

- Important DS Note
  - Bias-variance tradeoff
    - Fitted model should ideally:
      - Capture all of the “signal” in the data
      - Ignore all of the “noise” in the data
    - It’s generally hard to do both

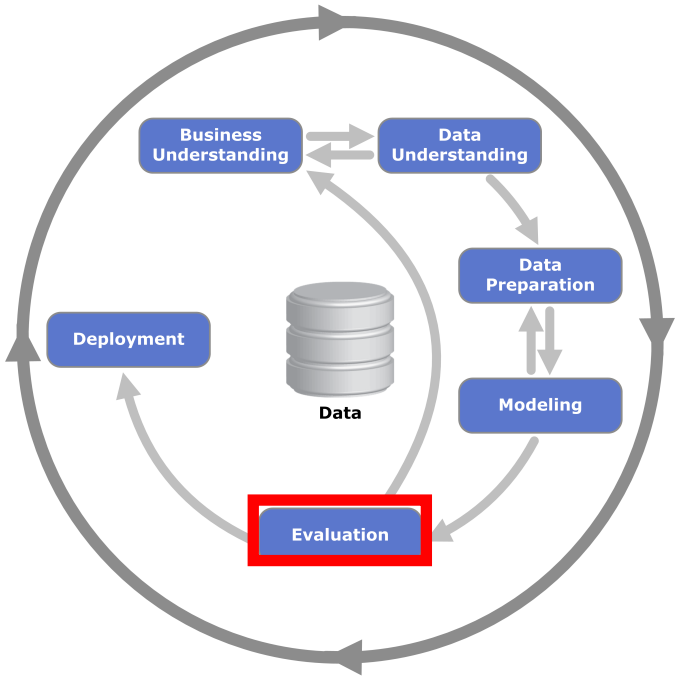


- High variance models
  - Can “overfit” to training data
    - Capture signal...
    - ...but also capture noise
  - Generalize poorly to unseen/test data

- High bias models
  - Can “underfit” to training data
    - Avoid capturing noise...
    - ...but also fail to capture all signal
  - Perform poorly on unseen/test data

Evaluation III

- Model approval
  - “All models are wrong but some are useful”
    --George Box



- Ruth Etzioni’s question:
  - When is an algorithm acceptable?

Deployment

- Will be specific to each problem space
- May include:
  - Create a report of findings
  - Planning and development of the deployment procedure
  - Deployment of the model
  - Development of a maintenance / update plan
  - Review of the project
  - Planning the next steps
- At FHCRC could relate to grants or publications

Machine Learning Paradigms

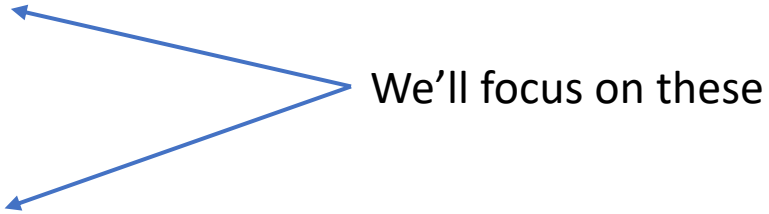
- A way to group or predict new information, based on information we have seen before

- “Given Age, and Height, what is someone’s Weight?”
- “Given Petal Length and Width, what kind of flower is this?”

- “Given a Patient’s clinical history*, what is the likelihood* they will have to enter the Emergency Department soon*?

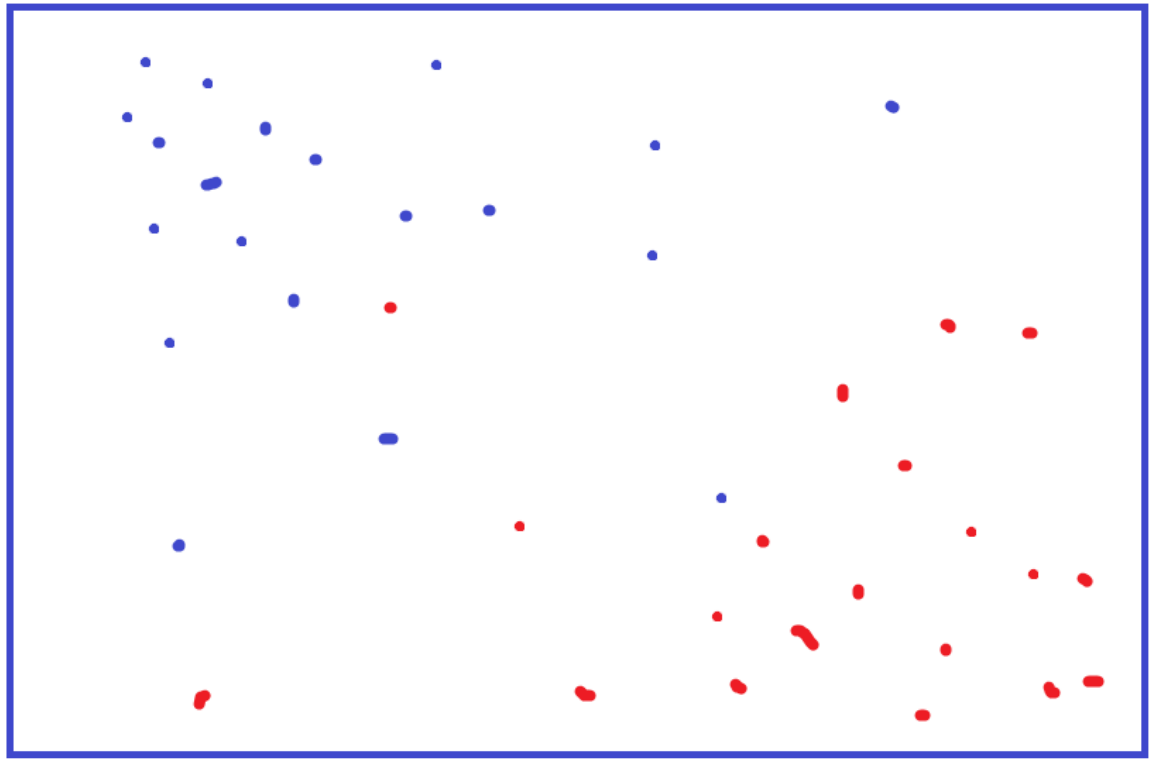
3 or 4 Machine Learning Paradigms

- Supervised Learning
- Unsupervised Learning

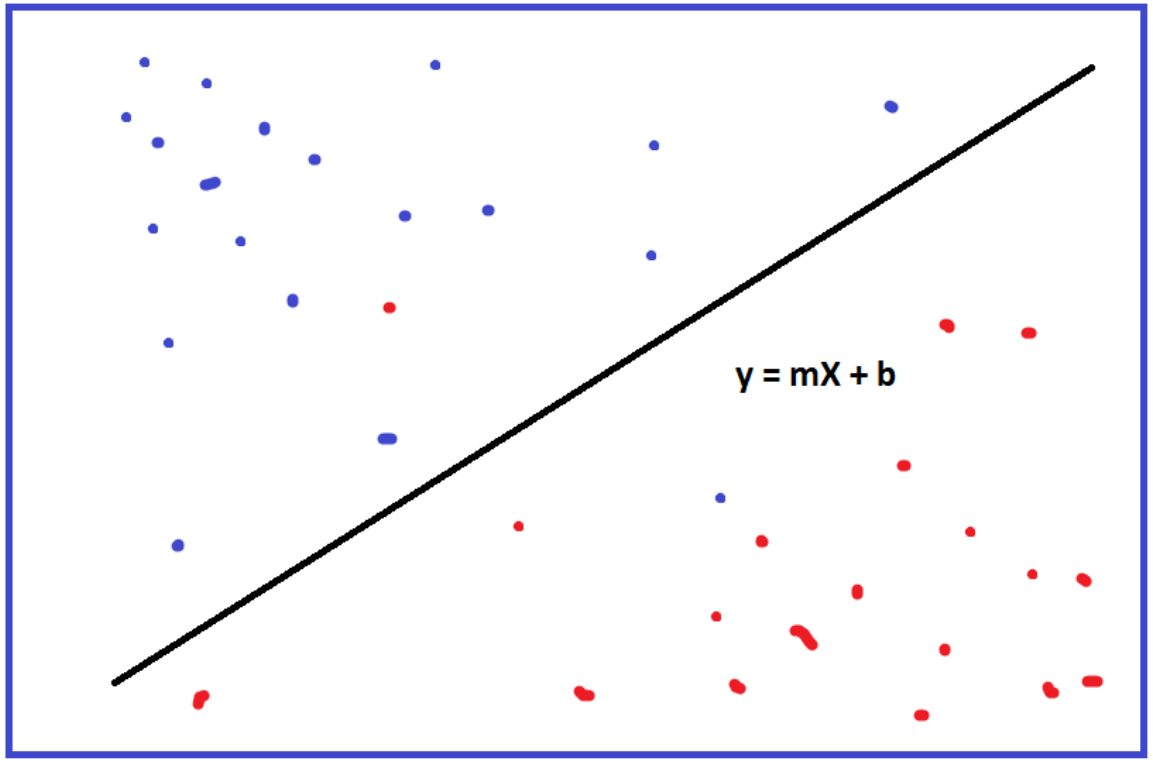


- Reinforcement Learning
- Semi-Supervised Learning

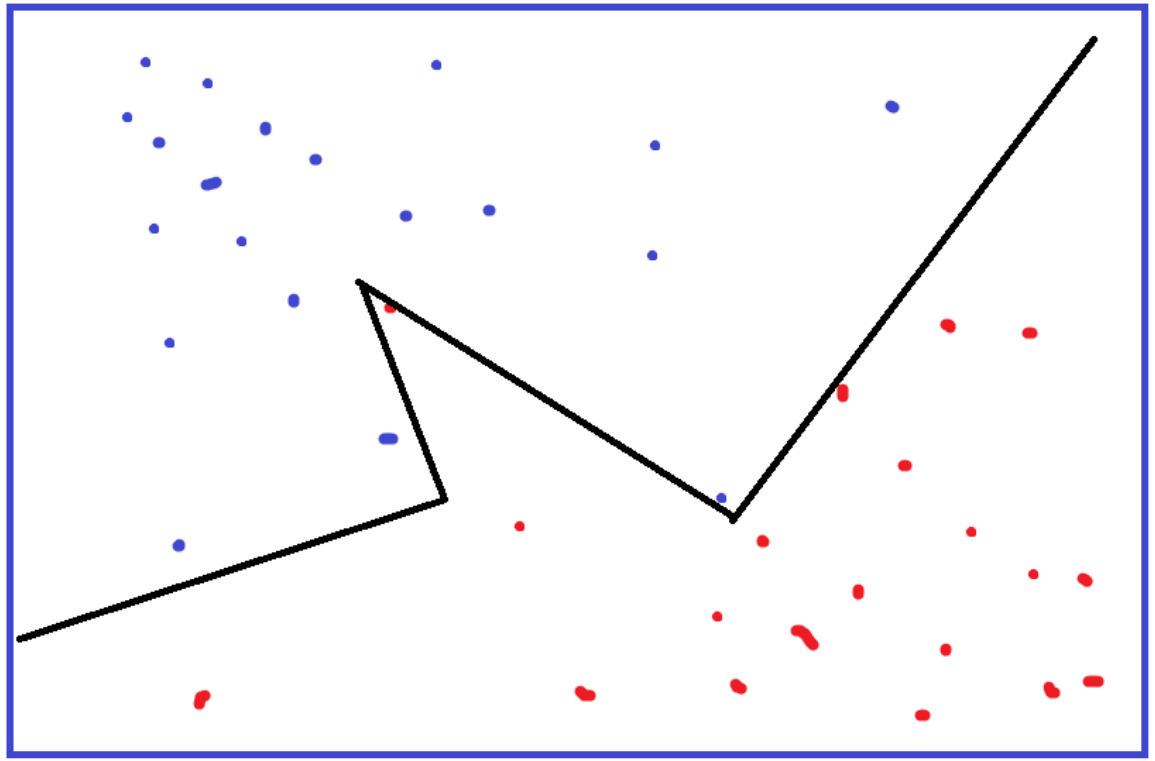
How can we divide this into Red vs. Blue?



How can we divide this into Red vs. Blue?



How can we divide this into Red vs. Blue?

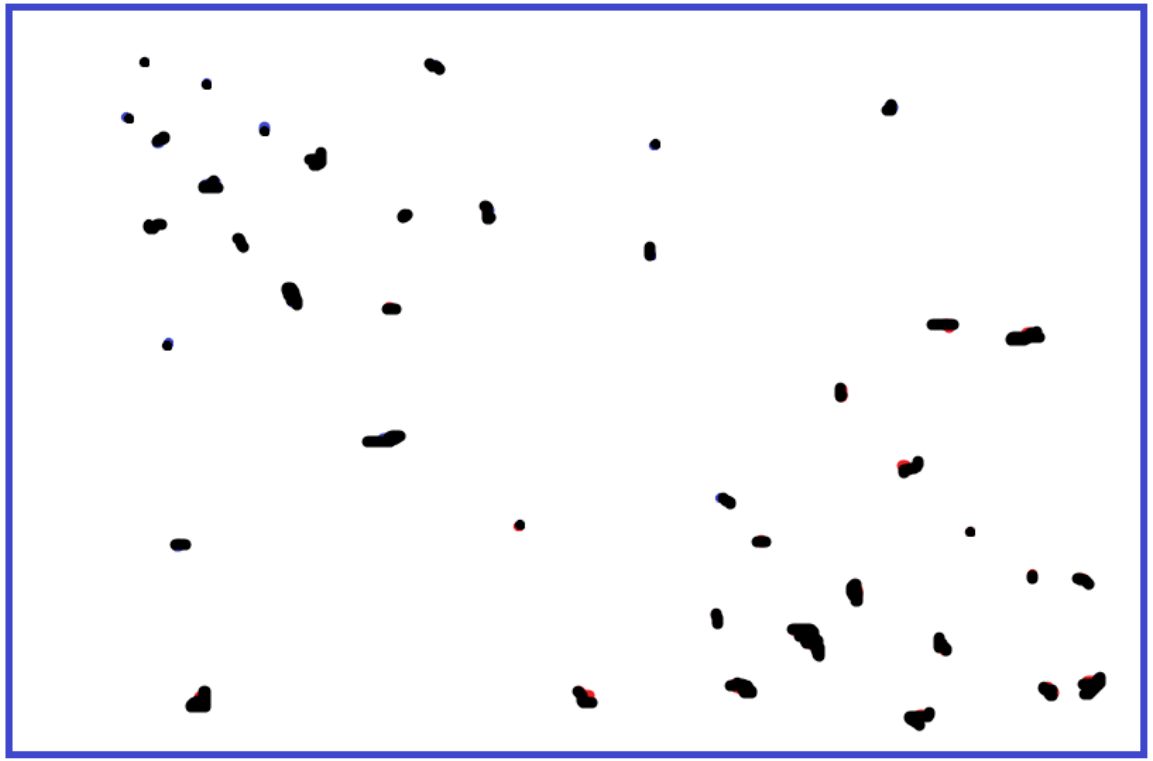


if 0 < x < 1 then:

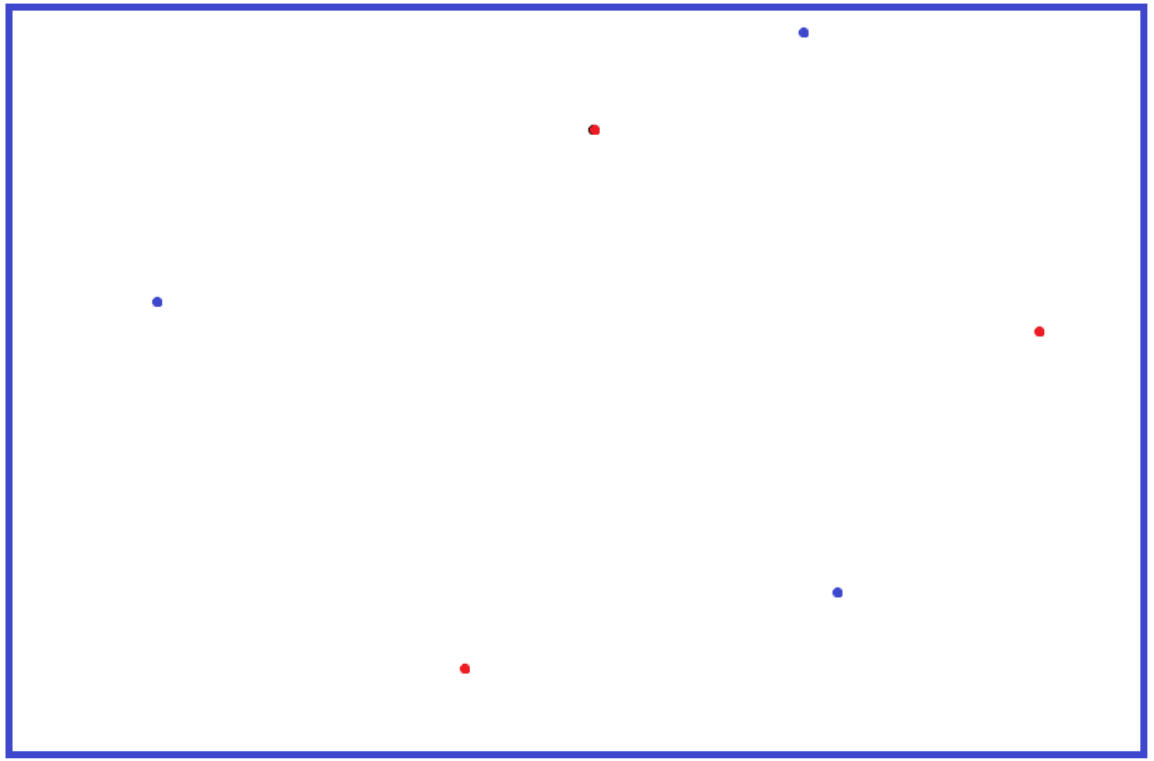
if 1<= x < 2 then:

...

What about now?



What about now?



What it isn’t

- Magic
- A Silver Bullet
- Old Bay Seasoning, to be sprinkled liberally on an otherwise mundane application to give it a zesty new flavor

Machine Learning is not a substitute for good understanding of your problem!

Defining our problem in terms of our data

3 or 4 Machine Learning Paradigms

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Semi-Supervised Learning

Is it a Supervised Problem?

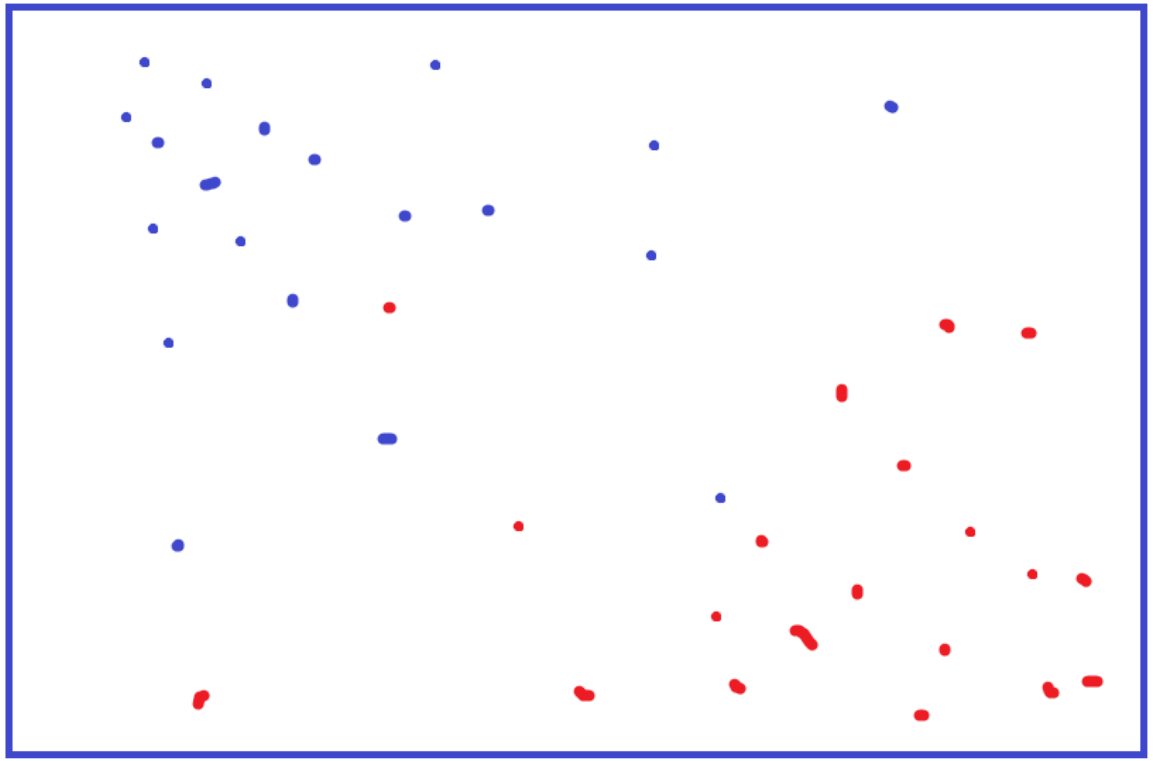
- Do we have data?
- Do we have some feature within the data that represent what we ultimately want to predict?

- If so, we can formulate it as a Supervised Problem
  1. “Train” a model by predicting the label and comparing to the correct answer. Update the model when we are wrong.
  2. “Test” the trained model by predicting the label of new data and evaluate
     - Our goal is a generalizable model- one that applies to new data well

Question Time

- Why do we want to test on new data? What would happen if we didn’t?
  - We would already know the correct answer, since we have trained on it

What Problem Statement could we make?



Given a set of coordinates, will a point be Red or Blue?

Supervised Problems, cont'd: More things to consider

- What is the best performance we expect from this predictor?
  - "If a human being made these predictions given this information, how good would they do?"
- Is the data relatively sparse?
  - How much data is missing/has been imputed?
  - How many variables of input are there relative to the total number of examples?
- How 'True' is the target?
  - Does it represent an estimate?
- Are the targets we are trying to predict skewed?
  - Eg: 95% of all participants answered 'No', and 5% answered 'Yes'

Supervised Problems: Categorical Problems

- Can we state our outcome as a choice between of A **vs.** B?

- Can be any number of categories
  - What problems could be introduced as we add more categories to an outcome to predict?
    - Target value sparsity
    - Need more data to keep the estimate solid

Quiz Time

Which of these are Supervised Learning Problems?

1. If we know someone's demographics, can we guess their favorite movie genre?
2. What demographic sub groups exist of people that like Horror movies?
3. If we know someone's demographics, can we guess their weight?

Supervised Learning: Regression Problems

- Can we state our target as a real number?
  - https://en.wikipedia.org/wiki/Real_number
- Since we don't have distinct categories, we don't have to worry how our data is binned (since there are no distinct bins!)

- What problems do we have to be aware of?
  - Is our data representative of the problem?
  - Are we fitting the wrong regression model to our data?
  - Does our data have outliers that are throwing off our model?

Let’s make a problem statement!

- From our clinical data (located at https://raw.githubusercontent.com/fredhutchio/R_intro/master/extra/clinical.csv)
- Form a problem statement like “Given X, predict Y”
  - Given a Patient’s **primary_diagnosis, tumor_stage and disease**, predict **vital_status**
    - This is a **Categorical** problem. Why?
  - Given a Patient’s **primary_diagnosis, tumor_stage and disease**, predict **days_to_death**
    - This is a **Regression** problem. Why? Can we formulate it as a categorical one?

3 or 4 Machine Learning Paradigms

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Semi-Supervised Learning

Unsupervised Problems

- Can we find out anything interesting by comparing aspects of our data to each other?

- For our purposes, think of it as grouping our data, and assigning meaning to the groups after the fact

Unsupervised Problems, cont’d

- Supervised problems can be characterized as having two phases: training and testing.
- Unsupervised problems have a single phase: fitting

Question Time

- Question: Why does a unsupervised problem only have a single phase?
  - Since there is no canonically ‘correct’ target/outcome to predict, there is no “training” to give better predictions, or “testing” to evaluate how generalizable our model is.

Closing Thoughts
The Importance of Understanding our Data

Data are Messy

- Most effort will be spent on cleaning, imputing, and transforming the data to make new or better input.
- The second most effort will be spent on analyzing the results and figuring out if they are:
  - Meaningful
  - Good enough

You Will Not Have Enough Data

- The more, and more varied the information you have, the more useful your model and predictions will be
- Having too many variables (columns) and not enough observations (rows) leads to problems of **sparsity**
- Having too little information to train over leads to **ungeneralizable models** or **over-trained models** (these are essentially the same thing)

Even when you have enough data, you may not have the complete picture

- Some problems are just hard, and even trained experts will not agree on what constitutes “correct”
- Problems can have factors that are not captured by your data
  - You cannot put together a whole puzzle with only half of the pieces

In Conclusion

- The Data predicate the model, not the other way around
- A more complicated model is not a cure for poor or incomplete data