MMA/MMAI 869
Machine Learning and AI

Overview of ML

Stephen Thomas

Updated: September 13, 2022

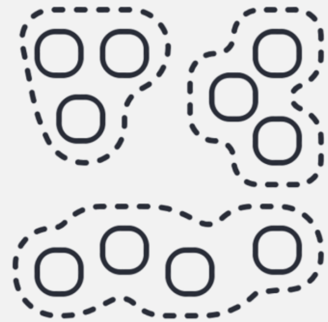
Smith
SCHOOL OF BUSINESS
Queen's University

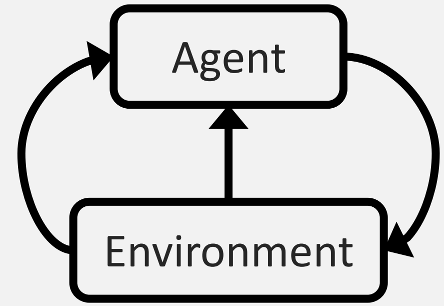
Outline

- Three Types of ML
- Supervised Machine Learning
  - Data
  - Algorithms
  - Model
  - Predictions
- Unsupervised Learning
- Reinforcement Learning

MACHINE LEARNING

Three Types of Machine Learning





	Supervised	Unsupervised	Reinforcement
What	Predict something in the future	Find relationships	Learn through trial and error
How	Algorithm builds model from past data	Algorithms finds patterns in data	Algorithm takes actions, gets rewards
Data	Labeled	Unlabeled	None
Tasks/ Algorithms	• Classification – Decision Tree, SVM, Naïve Bayes • Regression – Linear, Polynomial, Lasso • Recommenders – Collaborative filtering, matrix decomposition	• Clustering – K-Means, DBSCAN, Hierarchical • Association rules – Apriori, Eclat, FP-Growth • Dimensionality Reduction – PCA, NMF, LDA, GDA, t-SNE	• Q-learning • SARSA • Deep Q Network

Machine Learning Terminology

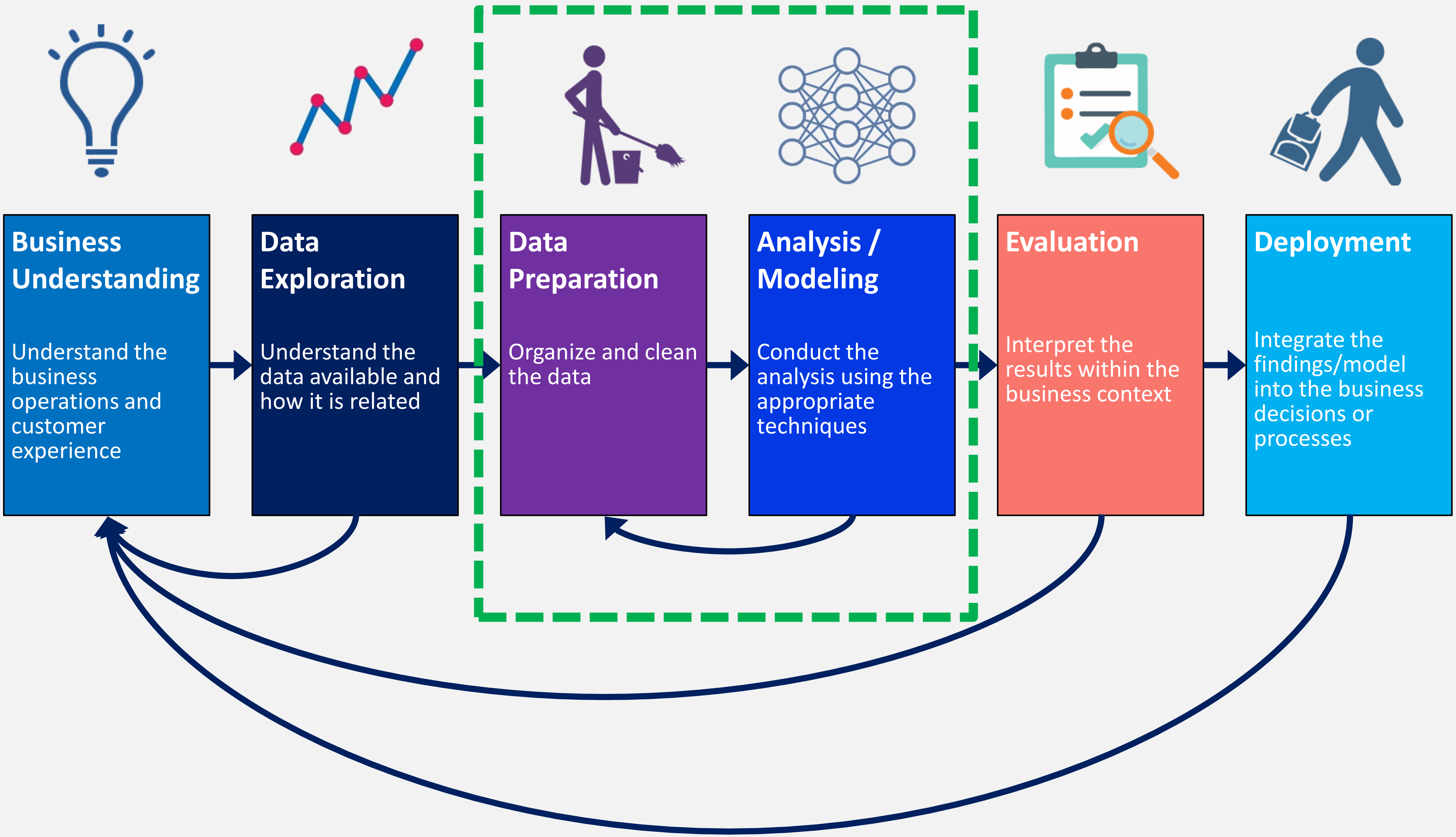
Features
(inputs, independent variables, X, columns)

Target or **Label**
(response, output, dependent variable, Y)

Instances
(rows, cases, records)

Age	Income	Married	Citizenship	Default
55	36,765	True	Canada	True
66	87,983	True	Canada	True
21	24,354	False	USA	False
24	56,654	True	Canada	False
34	98,324	False	UK	False
36	132,229	False	Germany	True
28	35,000	True	Canada	False
49	50,334	True	Canada	False

The Analytics Process: CRISP-DM



More Detail

Data Preparation

Organize and clean the data

Cleaning

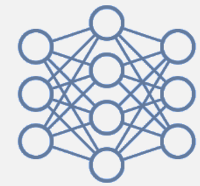
Outliers
Missing data
Data types
Inconsistencies

Feature Engineering

Normalization
Discretization
Coding
Temporal, text, image

Feature Selection

Filter
Wrapper



Analysis / Modeling

Conduct the analysis using the appropriate techniques

Model Training

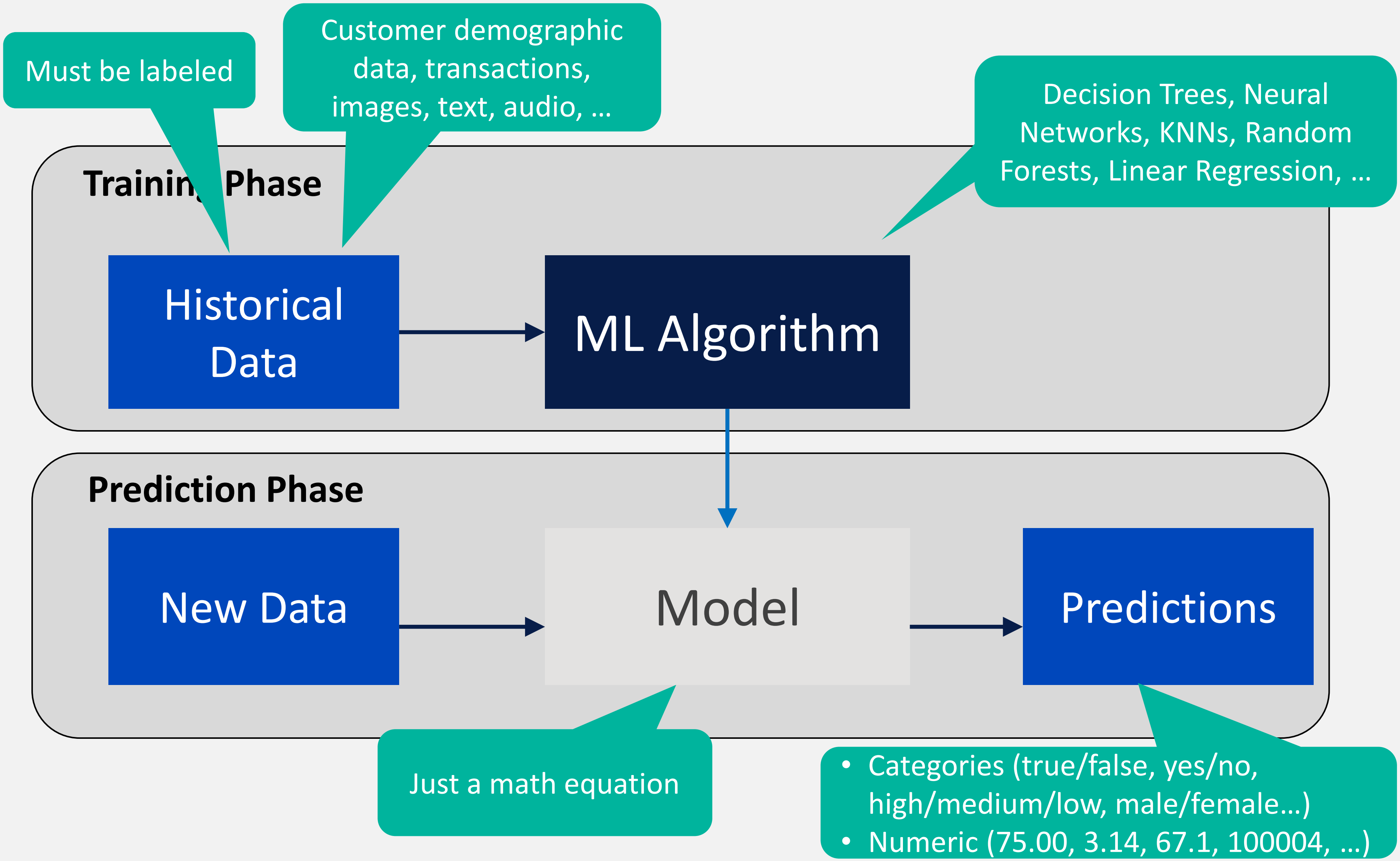
Algorithms
Ensembles
Hyperparameter tuning

Model Selection/ Evaluation

Performance Metrics
Cross validation

Supervised Machine Learning

Algorithm **learns** a model that can make predictions

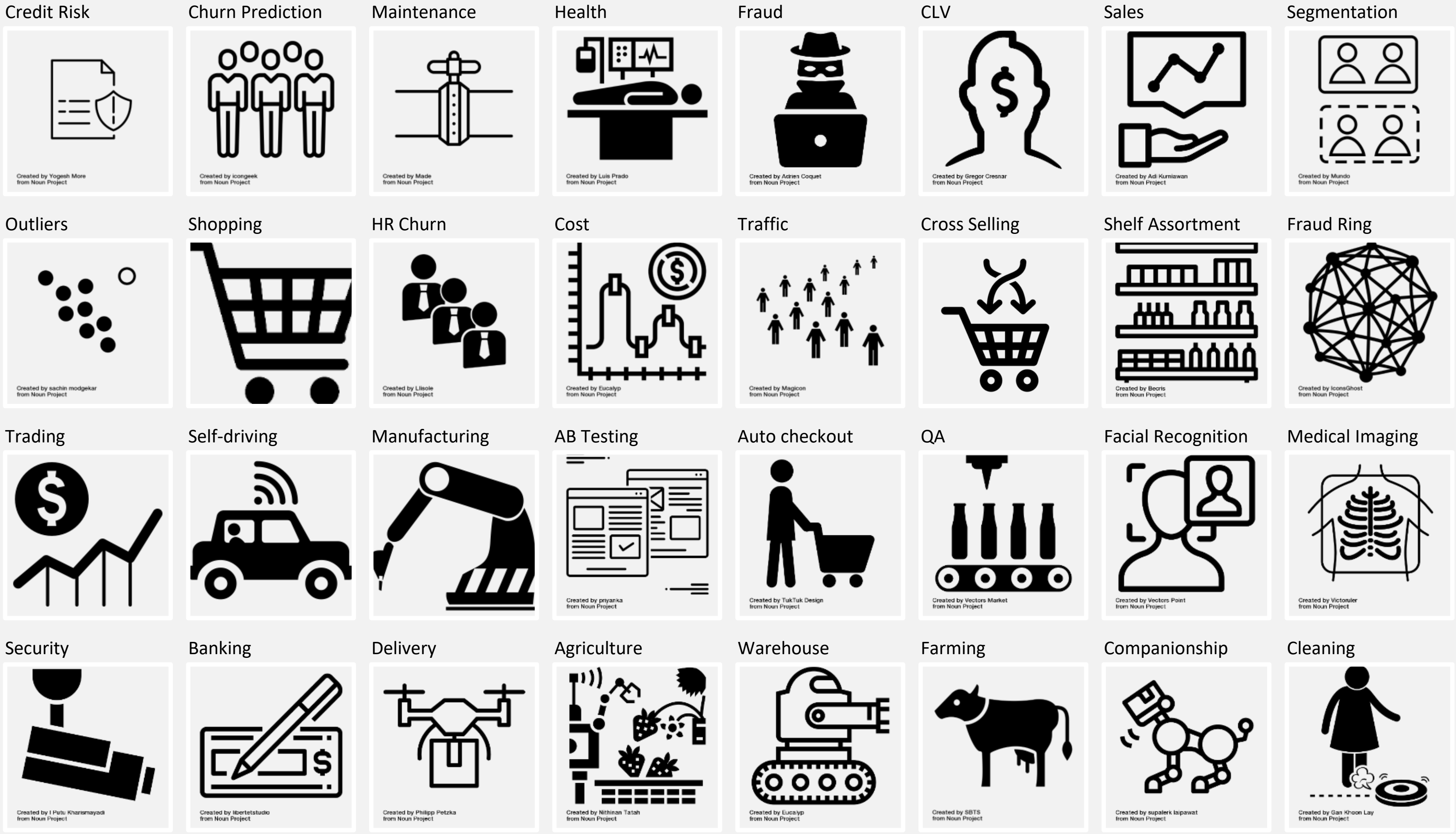


Types of Predictions

- Models can be learned to predict:

Numbers	Categories	Recommendations
• House Price • Energy demand • CLV • Foot traffic • Stock price • Sales • …	• Churn • Risk • Click • Images • Fraud • Health • …	• Movie • Product • Music • News • Books • Restaurants • …
"Regression"	*"Classification"*	*"Recommendation"*
• Ordinary Least Squares • Gradient Descent • Random Forest • …	• Decision Trees • Neural Networks • Naïve Bayes • SVM • Random Forest • ….	• Collaborative Filtering • Matrix decomposition • Ordinary Least Squares

So Many ML Applications



Data Must Be Labeled

- Usually comes from history or humans

Age	Income	Gender	Debt	Ed Level	Purpose	Default?
34	56786	M	0	3	I want to remodel my kitchen to…	Yes
54	68091	M	10000	2	Refinance credit card debt becau…	Yes
71	31287	F	25000	1	My husband bought a new tracto…	No
21	19807	M	325	1	TESLA BABY YAHOO!	No
44	79876	F	8976	3	My doctor says I need another pu…	No

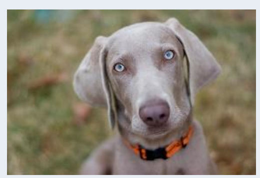
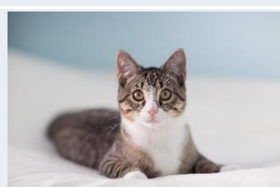
Data Must Be Labeled

- Usually comes from history or humans

Year Built	Beds	Baths	Garage	Price
2003	3	2	Attached	208,500
1976	3	2	Attached	181,500
2001	3	2	Attached	223,500
1915	3	1	Detached	140,000
2000	4	2	Attached	250,000
1993	4	1	Attached	143,000
2004	1	2	Attached	307,000
1973	3	2	Attached	200,000

Data Must Be Labeled

- Usually comes from history or humans

Image	Category
	Cat
	Dog
	Dog
	Cat
	Dog

Data Must Be Labeled

- Usually comes from history or humans

Date	Time	Amount	Merchant ID	Fraud
09/01/20	01:02:54	12.73	58447	False
09/01/20	01:03:09	58.00	63544	False
09/01/20	01:03:44	1.54	11440	False
09/01/20	01:03:51	500.87	07454	False
09/01/20	01:04:17	365.23	54784	False
09/01/20	01:04:20	412.00	22254	True
09/01/20	01:04:24	78.02	98630	False
09/01/20	01:04:24	8074.19	00744	False

Data Cleaning and Prep

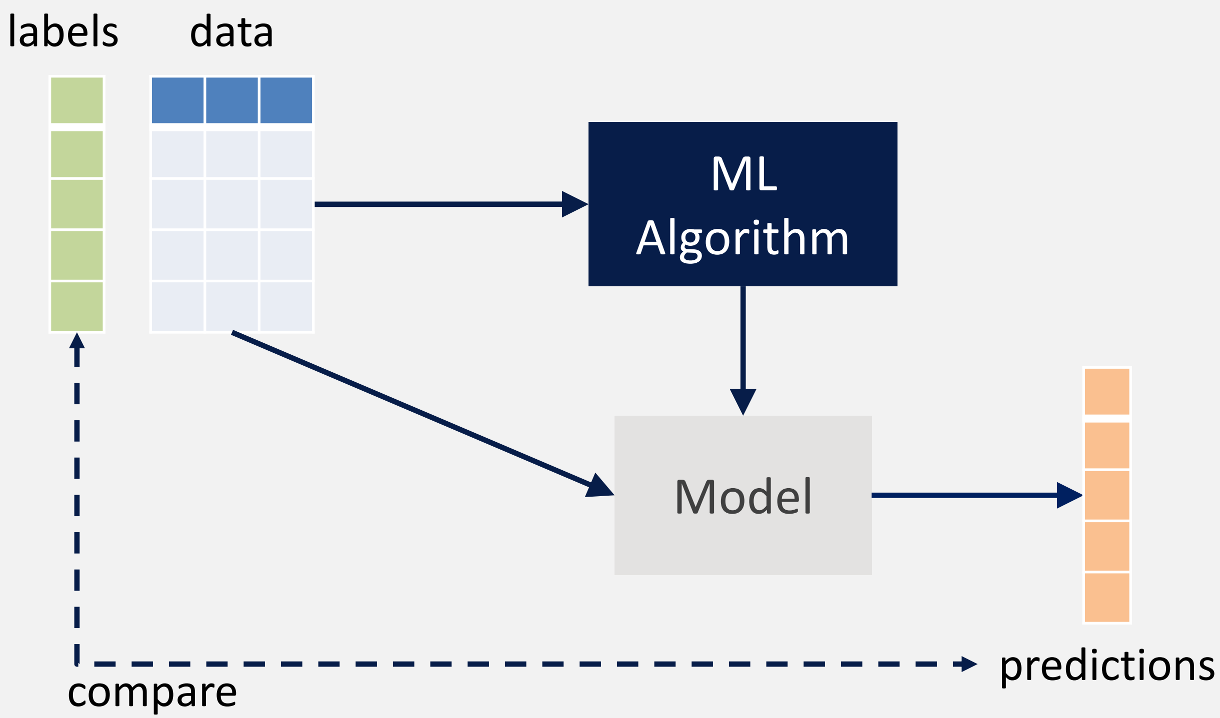
- "Garbage in, garbage out"
- "Quality data beats fancy algorithms"
- More than 50% of project time spent on data cleaning

ID	Name	Gender	Address	Phone	Salary
7474	Steve	m		123456789	48K
6532	Dr. Stephen Thomas	Male	451 Shallow Rd.	613-453-6969	100,214
2144	Jessica Smith	fem.	LONDON		17,856
231		FemalE		1547854587	
	Bob Doe	Female	Edmnton		35,748
6532	Dr. Stephen Thomas	Male	451 Shallow Rd.	613-453-6969	1
471	April Garcia	Female	547 Main St.	520-854-9658	41,012
7488	John Harris	Male	11 One Ave.	471-774-0000	68,745

Exercise: What's wrong here?

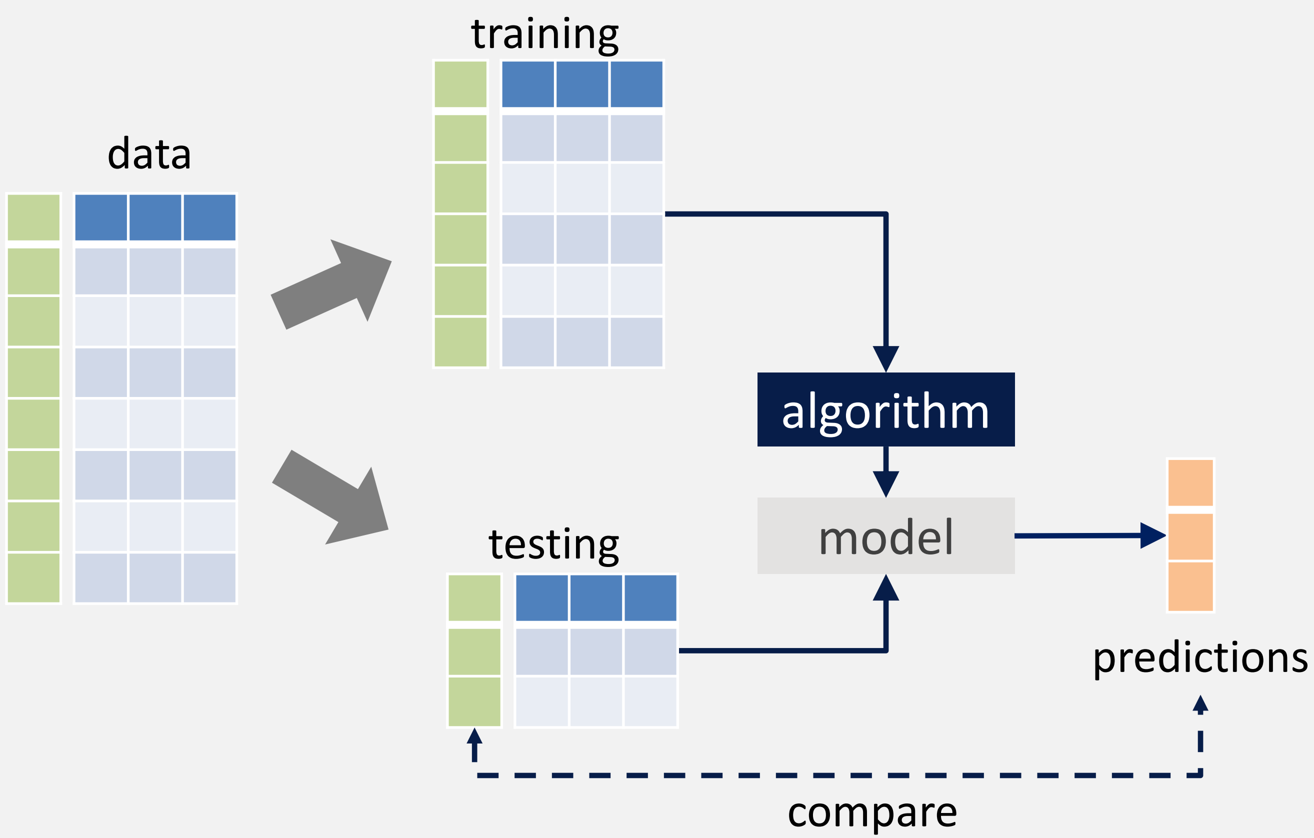
How to Assess Prediction Accuracy?

- Naïve way:
  - Use all the data to train the model
  - Ask model to create predictions on same data
  - Compare predictions with labels
- A horrible idea! Why?



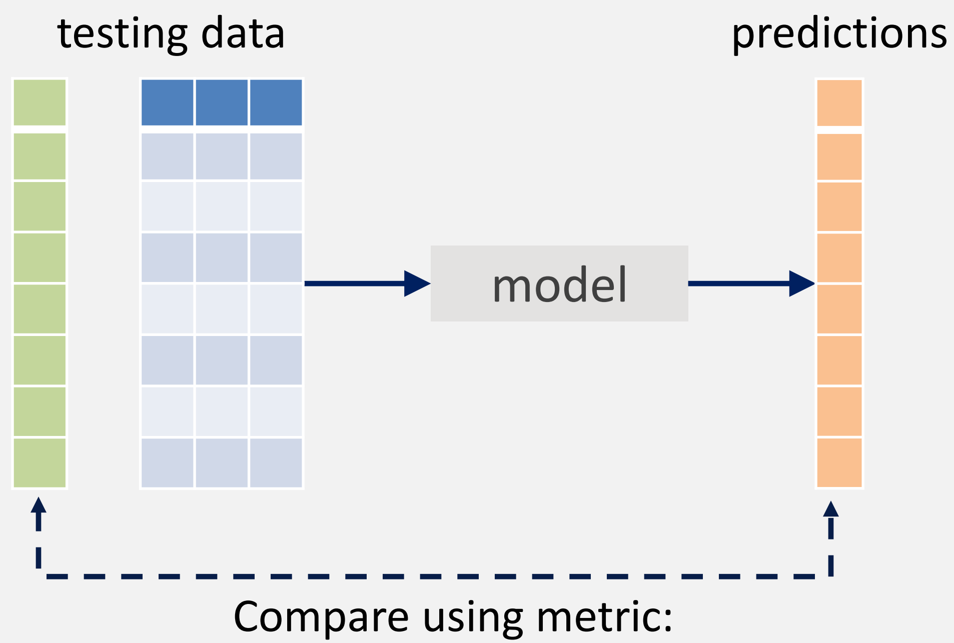
A Better Way

- ***Holdout Method***: Randomly divide training data into two subsets
  - *Training set:* Used to train the model
  - *Test set*: used to evaluate model's predictions
    - Pretend that the test data is "future data"



Performance Metrics

- How good are a model's predictions?



Numbers
- Mean Squared Error
- Mean Absolute Error
- Root MSE

Categories
- Accuracy and Error
- Precision, Recall
- F1 score
- Sensitivity/Specificity
- ROC Curve and AUC
- Log Loss

Recommendations
- Mean Average Precision @ K
- Coverage
- Personalization
- Intra-list similarity

UNSUPERVISED LEARNING

Clustering

Clustering, Cluster analysis

noun

- Putting instances in clusters/groups so that:
  - Instances in the same group are "similar" to each other
  - Instances in different groups are "not similar" to each other

Training Phase

ID	Age	Salary
1	44	$68K
2	23	$563K
3	70	$24K
4	36	$58K
5	81	$33K
…		

unlabeled training data

clustering algorithm

K-means, DBSCAN, Hierarchical, GMM, …

ID	Age	Salary
1	44	$68K
4	36	$58K
17	37	$60K
…		

Cluster 0

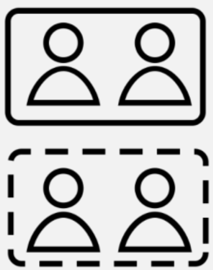
ID	Age	Salary
3	70	$24K
5	81	$33K
8	74	$29K
…		

Cluster 1

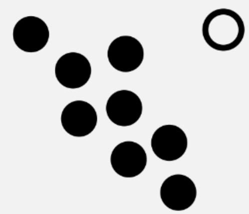
ID	Age	Salary
2	23	$563K
11	27	$600K
96	26	$412K
…		

Cluster 2

Clustering Applications



Customer Segmentation
(Marketing, Behavior analysis)

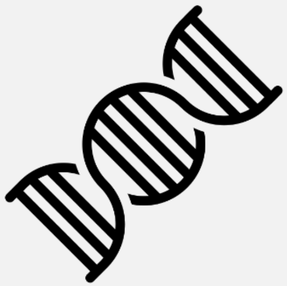


Anomaly Detection
(Insurance Fraud, etc.)

Document Clustering
(Survey analysis, EDA)

Feature Engineering

Educational Data Mining



Science/Biology

Association Rule Learning

Association Rule Learning

noun

- Discovering interesting relationships ("rules") in data.

Training Phase

unlabeled training data

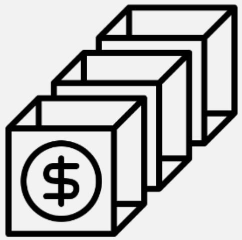
association rule learning algorithm

Just counting and math

Who	What	When		What	Confidence	Lift
Men	Diapers	Friday	→	Beer	40%	2.5x
-	Sausage	-	→	Mustard	31%	1.9x
Women	Candy, Pickles	-	→	Pregnancy Test	37%	40.4x
-	Lobsters	-	→	Butter	80%	6.6x

Business Applications

Shelf Assortment



Price Bundling

Consumer Profiles

Cross Selling

Shopping Patterns

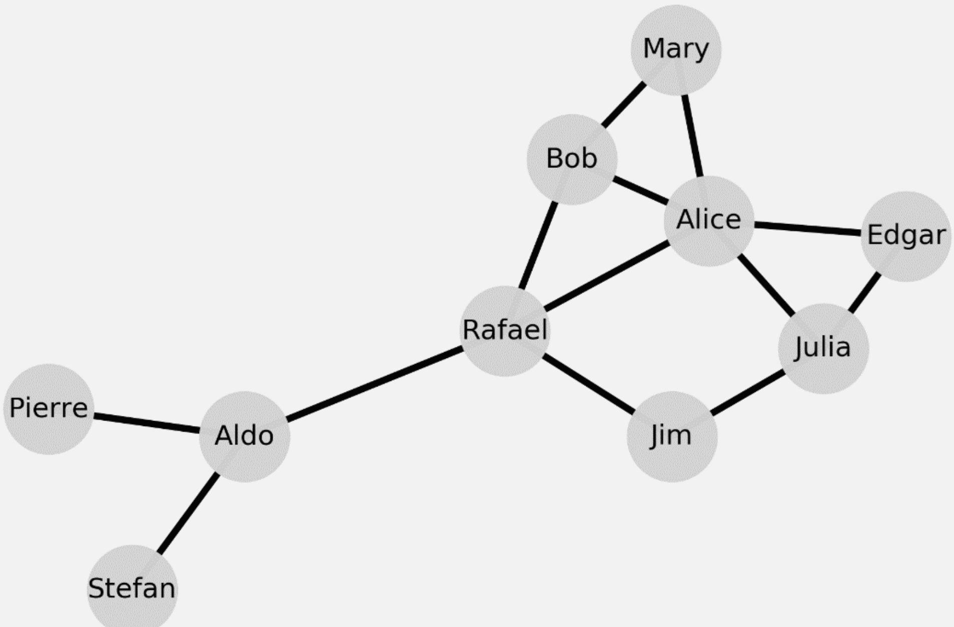
Web Analytics

Graph Analytics

graph analytics

noun

- Analyzing relationships and interactions amongst entities in a graph

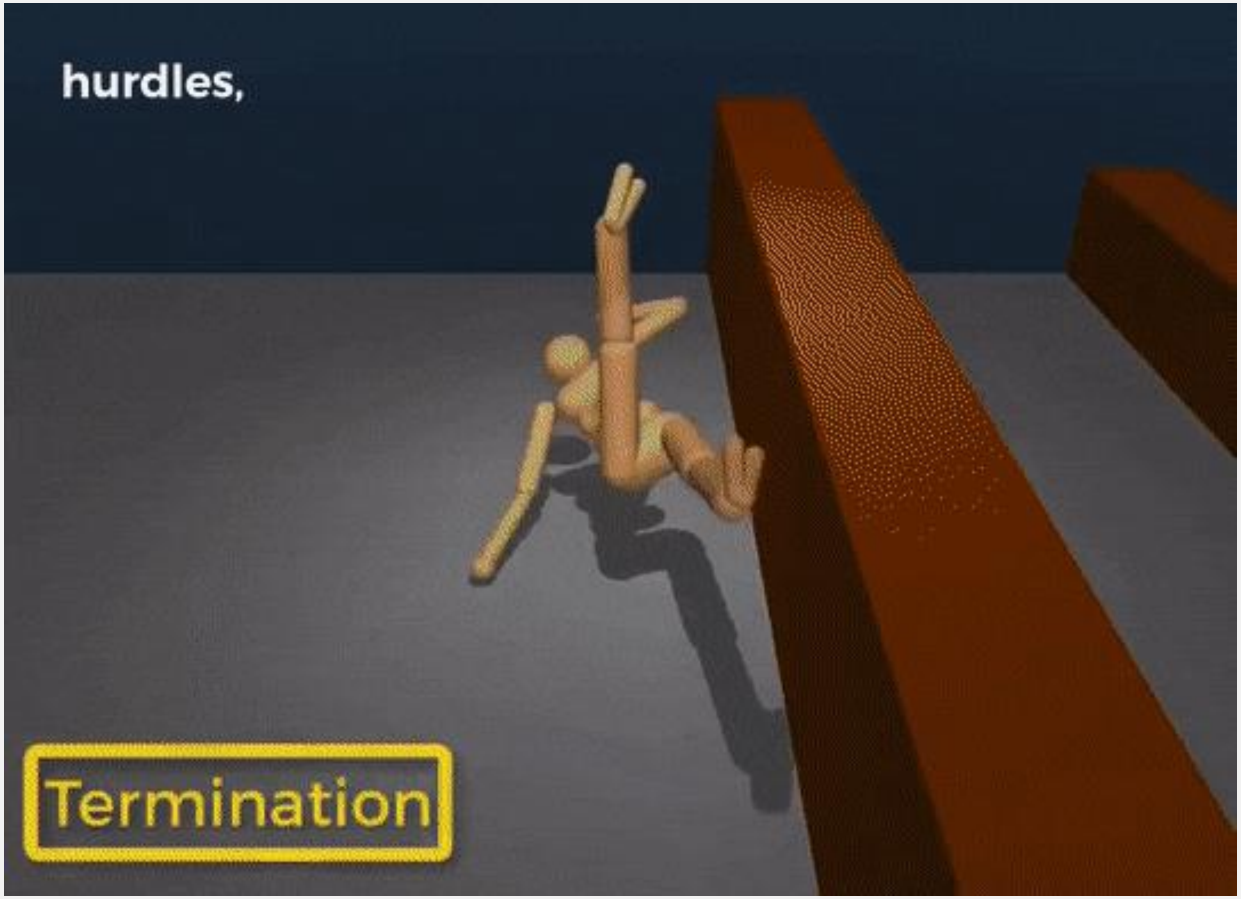


- Find most important nodes
- Find shortest paths
- Find bridges/hubs
- Find communities
- …

REINFORCEMENT LEARNING

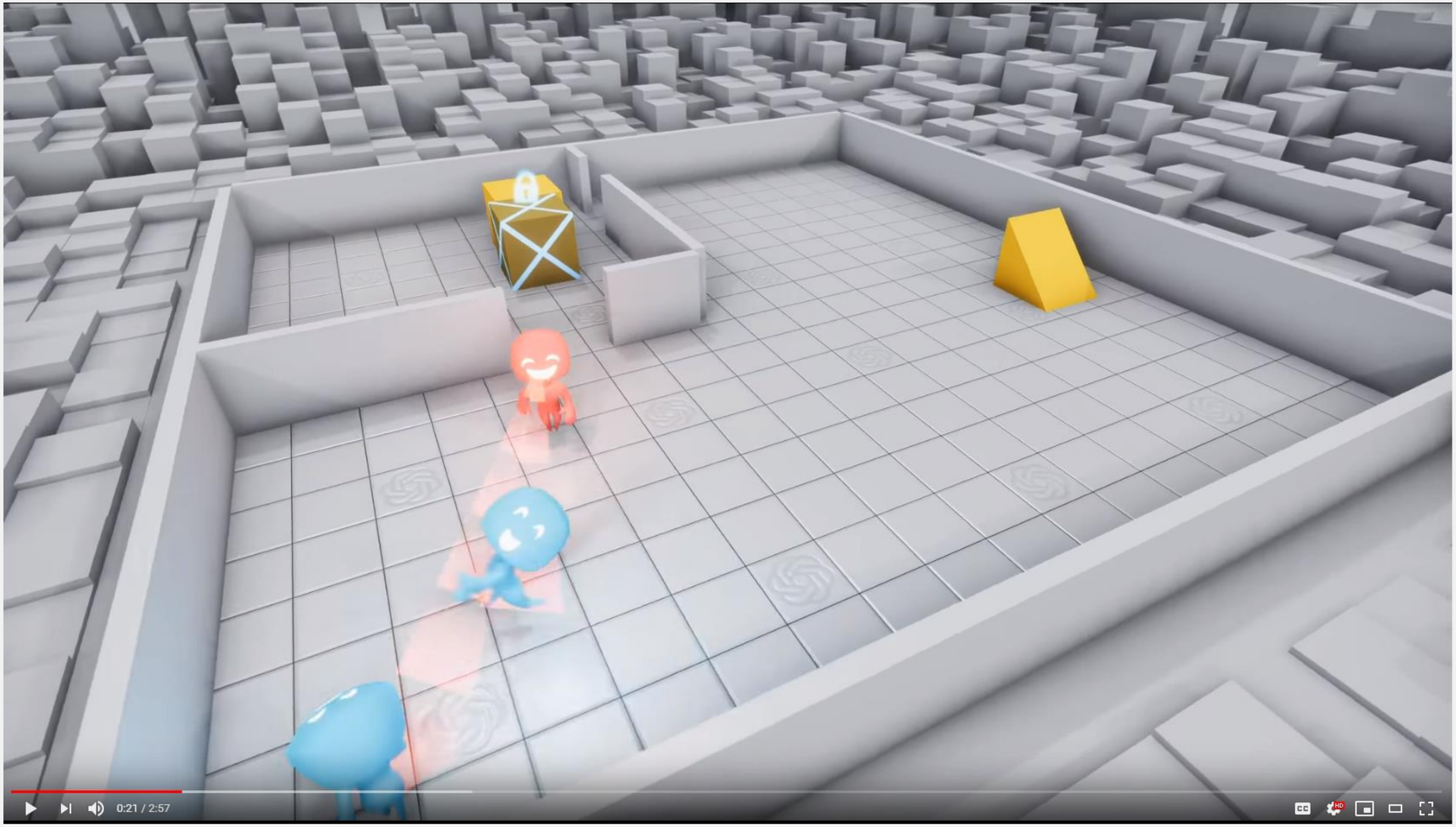
Reinforcement Learning

Through trial and error, algorithm finds actions that lead to reward



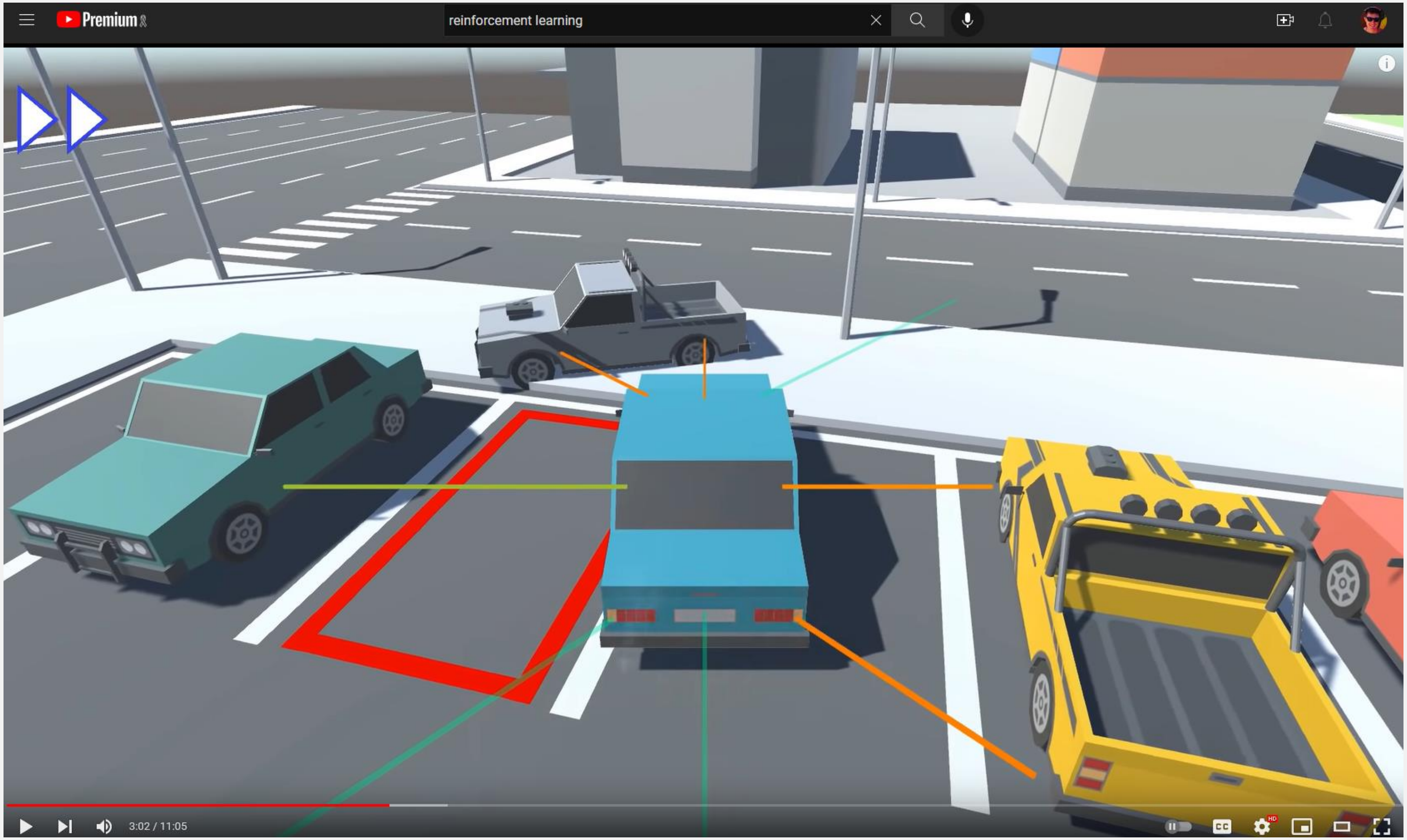
https://www.youtube.com/watch?v=gn4nRCC9TwQ

Hide and Seek



https://www.youtube.com/watch?v=Lu56xVlZ40M

RL Learns to Park

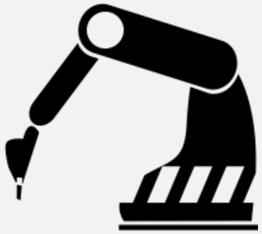


https://www.youtube.com/watch?v=VMp6pq6_QjI

RL Use Cases

Trading Optimization

Self-driving Cars



Inventory Robots

Electricity Grid Load Balancing

SUMMARY

Summary

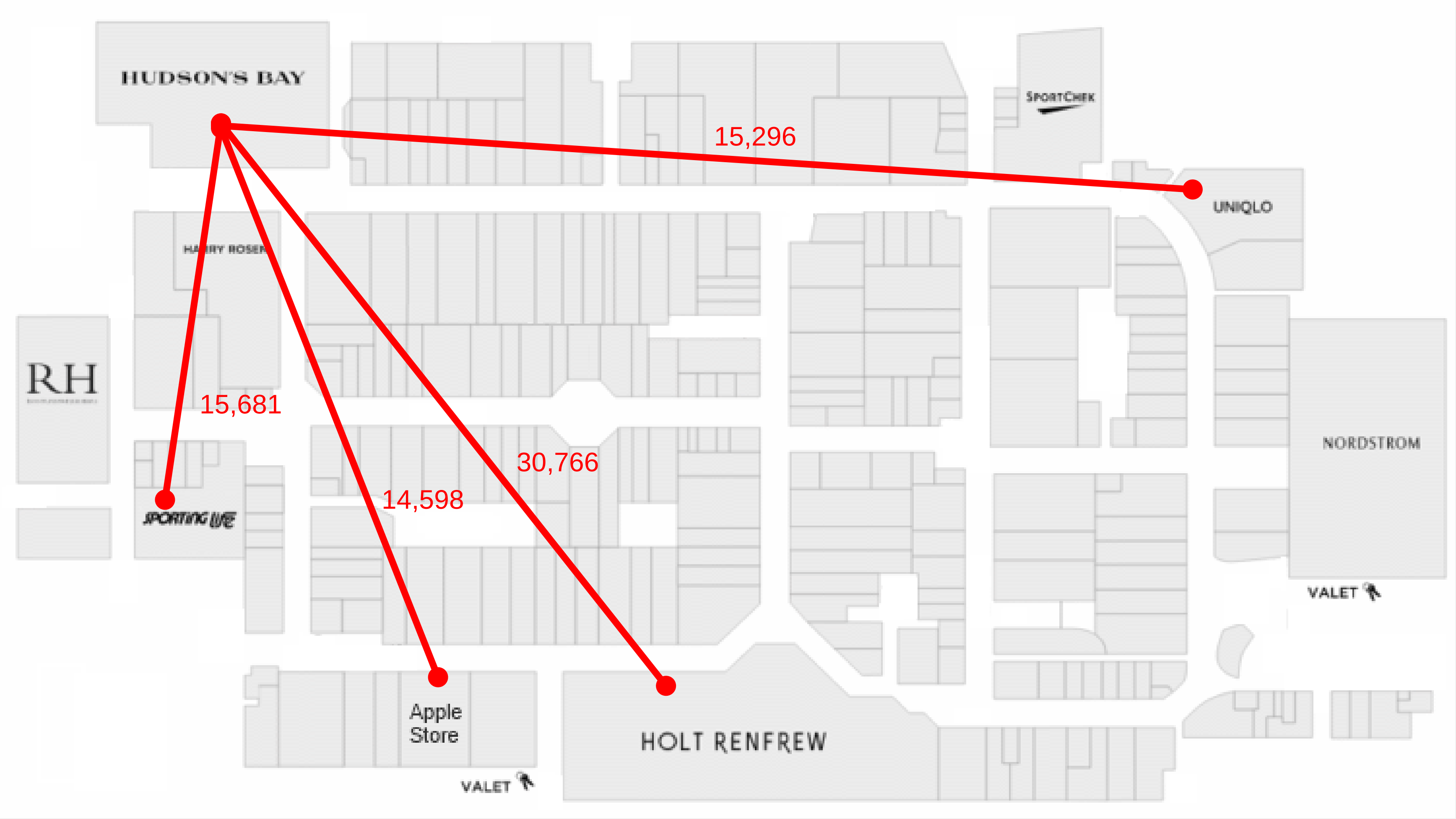
Three types of Machine Learning

- Supervised
  - Training data is labeled
  - Learn/Build model to make predictions about future
- Unsupervised
  - Training data is unlabeled
  - Find groupings, patterns
- Reinforcement
  - No training data, but you can easily inform the algorithm when it is "right" or "wrong"
  - Through trial and error, algorithm finds actions that lead to "right" answer

APPENDIX

Success Story: Yorkdale Mall

- Built rules from 1.8M customer journeys
- Looking for rules that showed a ***long customer journey***
- Red: Rules containing HB on LHS is greater than 200 meters away
- Results: were able to identify a store that should be moved



Success Story: Fraud Detection

Business Understanding → Data Exploration → Data Preparation → Analysis / Modeling → Evaluation → Deployment

Business Understanding

- Fraud costs $80B/year
- Traditional detection techniques are costly and prone to false positives

Data Exploration

- Transaction history per customer
- Look-alike customers
- Fraudulent transaction histories

Analysis / Modeling

Built supervised ML model to predict, in real-time, whether a transaction is fraudulent

Evaluation

- 15% increase in fraud detection
- 50% reduction of false alarms
- 60% in total savings

Success Story: Singapore Healthcare

Business Understanding → Data Exploration → Data Preparation → Analysis / Modeling → Evaluation → Deployment

- Historical biometric data, labeled with flu or not

- Can predict flu **24 hours before** any physical symptoms appear
- Save hundreds of lives per year

Built supervised ML model to predict flu

- **Flu** in NICU is **devastating**
- By the time flu is detected, it's often too late

Success Story: Steward Health Care

Business Understanding → Data Exploration → Data Preparation → Analysis / Modeling → Evaluation → Deployment

- Historical patient volume
- School vacation schedules
- New England sports
- Moon phases!

- Models are > 95% accurate
- More accurate in some units (like ICU)

- 60% of hospital operations expenses come from **staffing**
- Current staffing model uses averages; not efficient

Built supervised ML model to predict patient volume

- Achieved 1% reduction in nurse hours → $2M savings per year

Case Study by Data Robot

Success Story: Avis

Business Understanding → **Data Exploration** → **Data Preparation** → **Analysis / Modeling** → **Evaluation** → **Deployment**

- Data on 40 million customers
  - Rental transactions
  - Website visits

- Offer tiered incentives to improve customer loyalty
- **$100M+ increased revenues**

- Auto rental business is very competitive
- Need a way to increase customer loyalty

Built regression model to predict customer lifetime value

Success Story: Kroger

Business Understanding → Data Exploration → Data Preparation → Analysis / Modeling → Evaluation → Deployment

- Years of purchase history per customer (97% of customers use loyalty card)

- Created personalized direct mailers
- **Coupon return rate of 70%**
- **$10B+ additional revenue**

- Kroger wants to **maintain customer loyalty** and broaden shopping
- Coupons work, but only have an industry average return rate of 4%
  - Coupons based on averages, or demographics

Built an association rules model to learn what products are often purchased together

Success Story: Amazon

amazon

Business Understanding → Data Exploration → Data Preparation → Analysis / Modeling → Evaluation → Deployment

- Amazon has 600M+ products
- Hard for customers to discover

- 20 years of purchase history for millions of customers

Built a recommender system to find products that are often purchased together

- Use model to display related products dynamically on webpage
- **$35B/year additional revenue**

Success Story: McDonald's

Business Understanding → **Data Exploration** → **Data Preparation** → **Analysis / Modeling** → **Evaluation** → **Deployment**

- Customer purchase history
- Customer response to offer history
- Location, time of day

- **700% increase in redemption**
- Customers using app spent 47% more than non-app users

- Netherlands, Sweden, and Japan make up 60% of McDonald's locations worldwide
- Want to increase customer engagement with personalization

Built a recommender system to push personalized offers on mobile app

Source: TechEmergence

Success Story: Hyatt

Business Understanding → Data Exploration → Data Preparation → Analysis / Modeling → Evaluation → Deployment

Business Understanding

- Hyatt wanted to increase cross sell and upsell success to guests
- Increase incremental room revenue

Data Exploration

- Guest demographics
- Guest transactions

Analysis / Modeling

- Built a segmentation model
- Determined each segment's spending habits and preferences

Deployment

- At check-in, desk agents see cross and upsell opportunities for that customer's segment
- **Increased incremental room revenue by 60%**

Success Story: Fanuc

Business Understanding → Data Exploration → Data Preparation → Analysis / Modeling → Evaluation → Deployment

- No training data available
- However, you can easily tell robot if it is "right" or "wrong"

- After 8 hours of training, robot gets 90%+ accuracy on the task

- Traditional robots need to be programmed very carefully for every task, such as picking up an object in a box
- Difficult and time consuming

- Use reinforcement learning to train robot's algorithm
- Robot tries to pick up object, and it will learn what leads to "right" or "wrong"