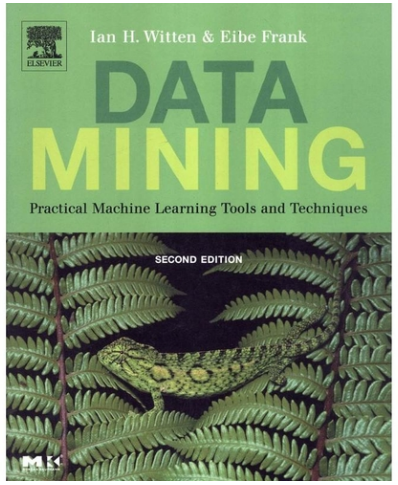
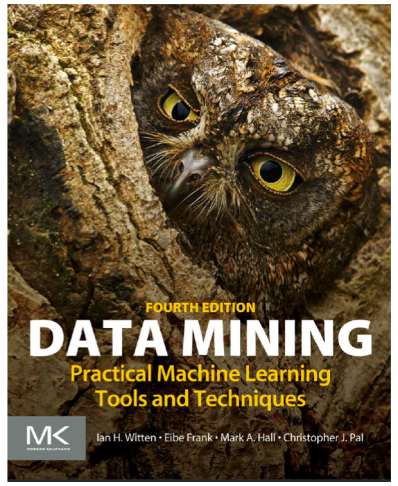
Machine Learning for Cyber-Security & Artificial Intelligence

Part 1 - Introduction

Hermes Senger





Adapted from:

Chapter 1, What's it all about?
Data Mining: Practical Machine Learning Tools and Techniques,
4th Edition. By Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher J. Pal.
Morgan Kauffman, 2017.

Introduction to Machine learning

- Before speaking on how to use machine learning for cyber security …

- Introduce **data mining** and **machine learning**

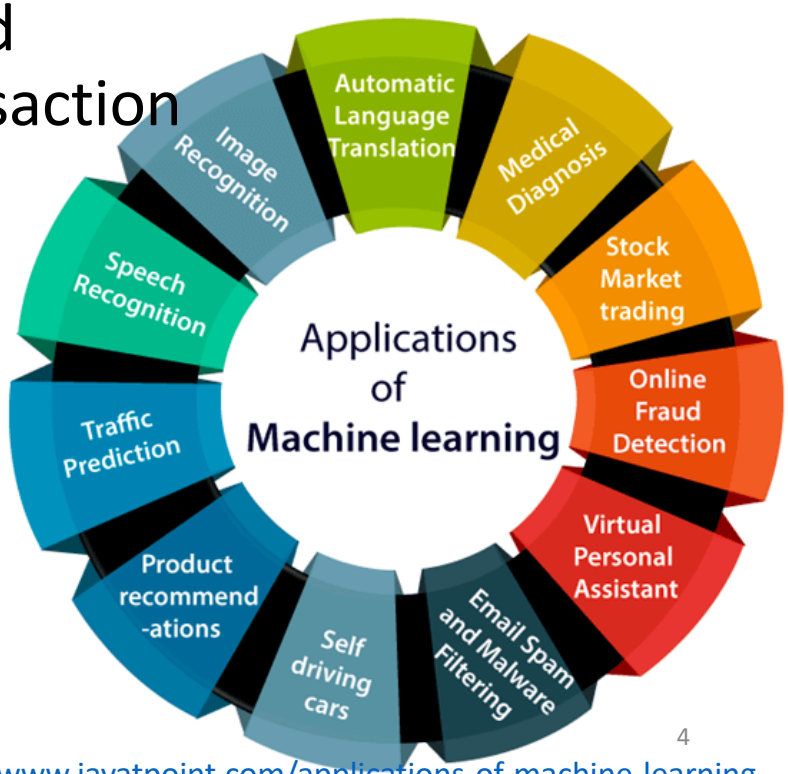
Chapter 1: What's it all about?

- Data mining and machine learning
- Simple examples: the weather problem, attack detection
- Some applications
- The data mining process

Information is crucial

Some (few) examples:

- Web mining: apply algorithms (e.g. the famous PageRank) to decide which pages are “the most relevant” in a web query
- Decisions involving judgement:
  - Banks analyze historical data to “accept” or “reject” a loan request
  - Online fraud detection (e.g. Credit card companies decide to “approve” a transaction
  - Product recommendations
  - Image recognition
  - Email spam&malware filtering
  - Self-driving cars
  - Virtual personal assistant
  - …



https://www.javatpoint.com/applications-of-machine-learning

From data to information

- Society produces huge amounts of data
  - Sources: business, science, medicine, economics, geography, environment, sports, …
  - Cyber security: operation logs from firewalls, IDS, anti-virus, application servers, etc
- This data is a potentially valuable resource
- Raw data is useless: need techniques to automatically extract information from it
  - Data: recorded facts
  - Information: **patterns** underlying the data
- We are concerned with machine learning techniques for automatically finding patterns in data
- **Patterns** that are found may be represented as *structural descriptions* or as black-box models

Structural descriptions

- Example: **if-then rules**

```
If tear production rate = reduced
   then recommendation = none
Otherwise, if age = young and astigmatic = no
   then recommendation = soft
```

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
…	…	…	…	…

Structural descriptions

- Example: if-then rules

```
If duration = short
   then traffic = normal
Otherwise, if service = http and logged_in =
   no
   then traffic = rootkit
```

Service	protocol_type	logged_in	duration	Traffic
http	tcp	No	short	Normal
http	tcp	No	long	rootkit
http	tcp	Yes	short	Normal
http	tcp	Yes	long	buffer_overflow
…	…	…	…	…

Machine learning

- Definitions of “learning” from dictionary:

To get knowledge of by study, experience, or being taught

To become aware by information or from observation

To commit to memory

To be informed of, ascertain; to receive instruction

} Difficult to measure

} Trivial for computers

- Operational definition:

Things learn when they change their behavior in a way that makes them perform better in the future.

} Does a slipper learn?

- Does learning imply intention?

Data mining

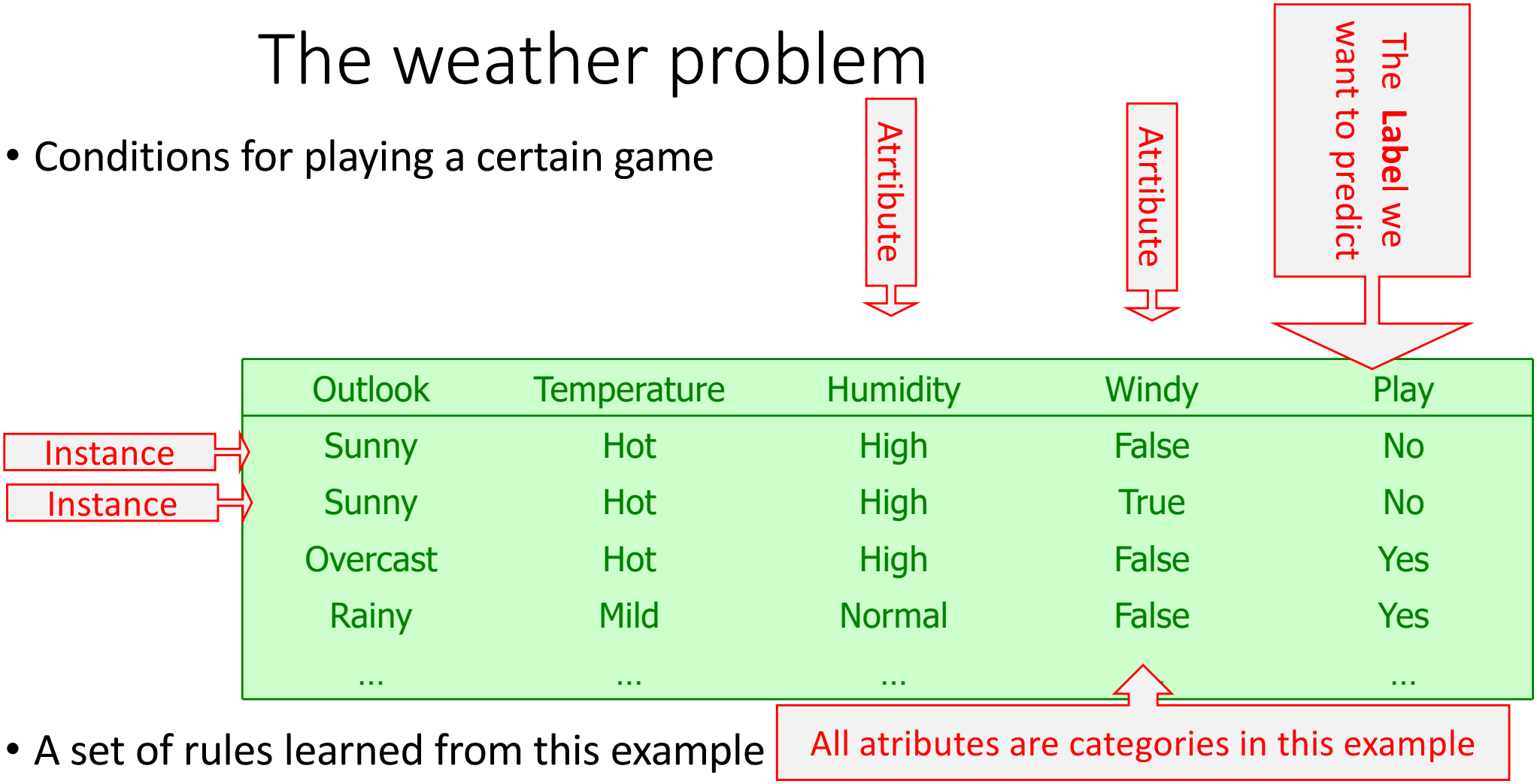
- Finding **patterns** in data that provide insight or enable fast and accurate decision making

- Strong, accurate patterns are needed to make decisions
  - Problem 1: most patterns are not interesting
  - Problem 2: patterns may be inexact (or spurious)
  - Problem 3: data may be garbled or missing

- Of primary interest are machine learning techniques that provide structural descriptions

The weather problem

- Conditions for playing a certain game



- A set of rules learned from this example

```
If outlook = sunny and humidity = high then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity = normal then play = yes
If none of the above then play = yes
```

Classification vs. association rules

- **Classification rule**:
  predicts value of a given attribute (the classification of an example)

```
If outlook = sunny and humidity = high
   then play = no
```

- **Association rule**:
  predicts value of arbitrary attribute (or combination)

```
If temperature = cool then humidity = normal
If humidity = normal and windy = false
   then play = yes
If outlook = sunny and play = no
   then humidity = high
If windy = false and play = no
   then outlook = sunny and humidity = high
```

Weather data with mixed attributes

- Some attributes have **numeric values**

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
…	…	…	…	…

```
If outlook = sunny and humidity > 83 then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity < 85 then play = yes
If none of the above then play = yes
```

The contact lenses data

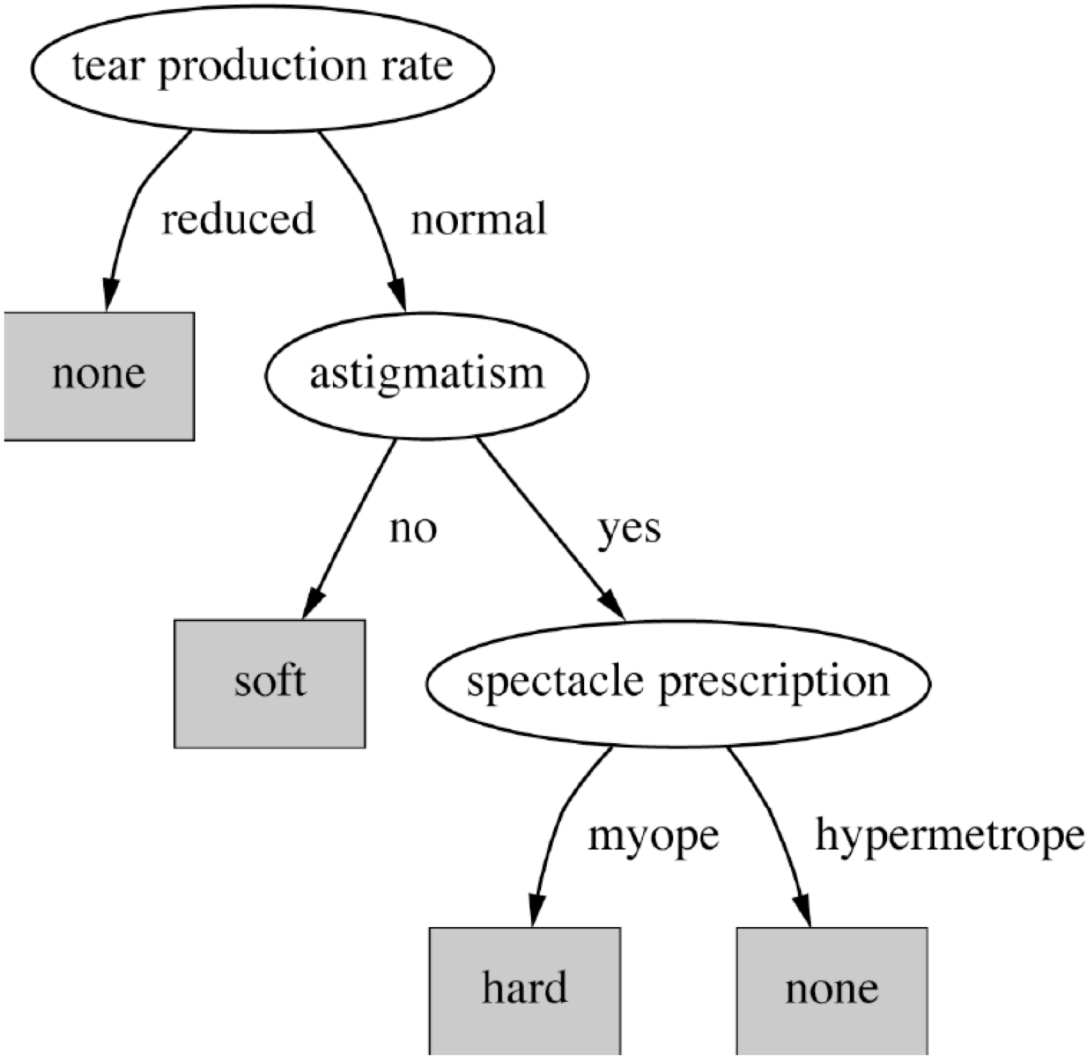
Which kind of contact lens to prescribe, given certain information about a patient

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Myope	No	Normal	Soft
Young	Myope	Yes	Reduced	None
Young	Myope	Yes	Normal	Hard
Young	Hypermetrope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Young	Hypermetrope	Yes	Reduced	None
Young	Hypermetrope	Yes	Normal	hard
Pre-presbyopic	Myope	No	Reduced	None
Pre-presbyopic	Myope	No	Normal	Soft
Pre-presbyopic	Myope	Yes	Reduced	None
Pre-presbyopic	Myope	Yes	Normal	Hard
Pre-presbyopic	Hypermetrope	No	Reduced	None
Pre-presbyopic	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	Yes	Reduced	None
Pre-presbyopic	Hypermetrope	Yes	Normal	None
Presbyopic	Myope	No	Reduced	None
Presbyopic	Myope	No	Normal	None
Presbyopic	Myope	Yes	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
Presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Hypermetrope	No	Normal	Soft
Presbyopic	Hypermetrope	Yes	Reduced	None
Presbyopic	Hypermetrope	Yes	Normal	None

A complete and correct rule set

```
If tear production rate = reduced then recommendation = none
If age = young and astigmatic = no
    and tear production rate = normal then recommendation = soft
If age = pre-presbyopic and astigmatic = no
    and tear production rate = normal then recommendation = soft
If age = presbyopic and spectacle prescription = myope
    and astigmatic = no  then recommendation = none
If spectacle prescription = hypermetrope and astigmatic = no
    and tear production rate = normal then recommendation = soft
If spectacle prescription = myope and astigmatic = yes
    and tear production rate = normal then recommendation = hard
If age young and astigmatic = yes
    and tear production rate = normal then recommendation = hard
If age = pre-presbyopic
    and spectacle prescription = hypermetrope
    and astigmatic = yes then recommendation = none
If age = presbyopic and spectacle prescription = hypermetrope
    and astigmatic = yes then recommendation = none
```

A decision tree for this problem



Classifying iris flowers

Iris Versicolor | Iris Setosa | Iris Virginica

	Sepal length	Sepal width	Petal length	Petal width	Type
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
…					
51	7.0	3.2	4.7	1.4	Iris versicolor
52	6.4	3.2	4.5	1.5	Iris versicolor
…					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
…					

```
If petal length < 2.45 then Iris setosa
If sepal width < 2.10 then Iris versicolor
...
```

Predicting CPU performance

- Example: 209 different computer configurations

	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performance
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
…							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

- Linear regression function

```
PRP = -55.9 + 0.0489 MYCT + 0.0153 MMIN + 0.0056 MMAX
      + 0.6410 CACH - 0.2700 CHMIN + 1.480 CHMAX
```

Data from labor negotiations

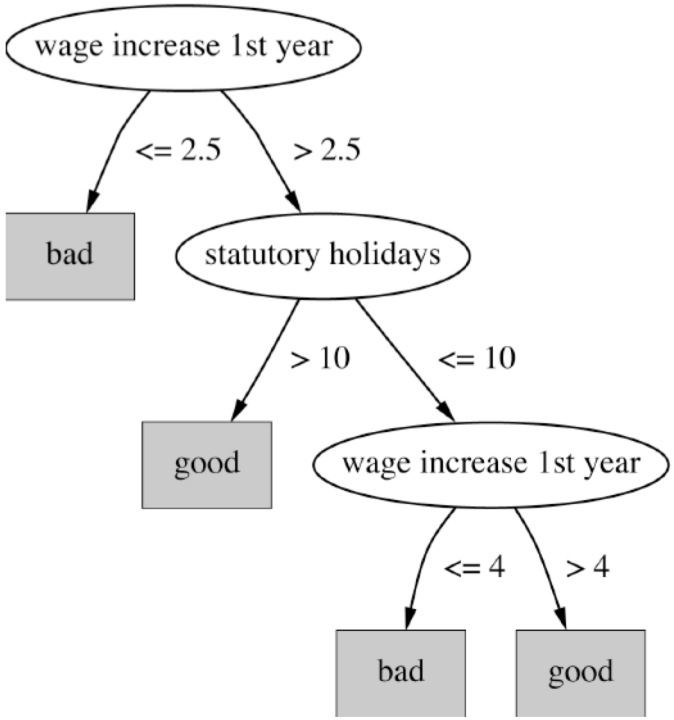
Canadian labor contract negotiations in 1987 and 1988.

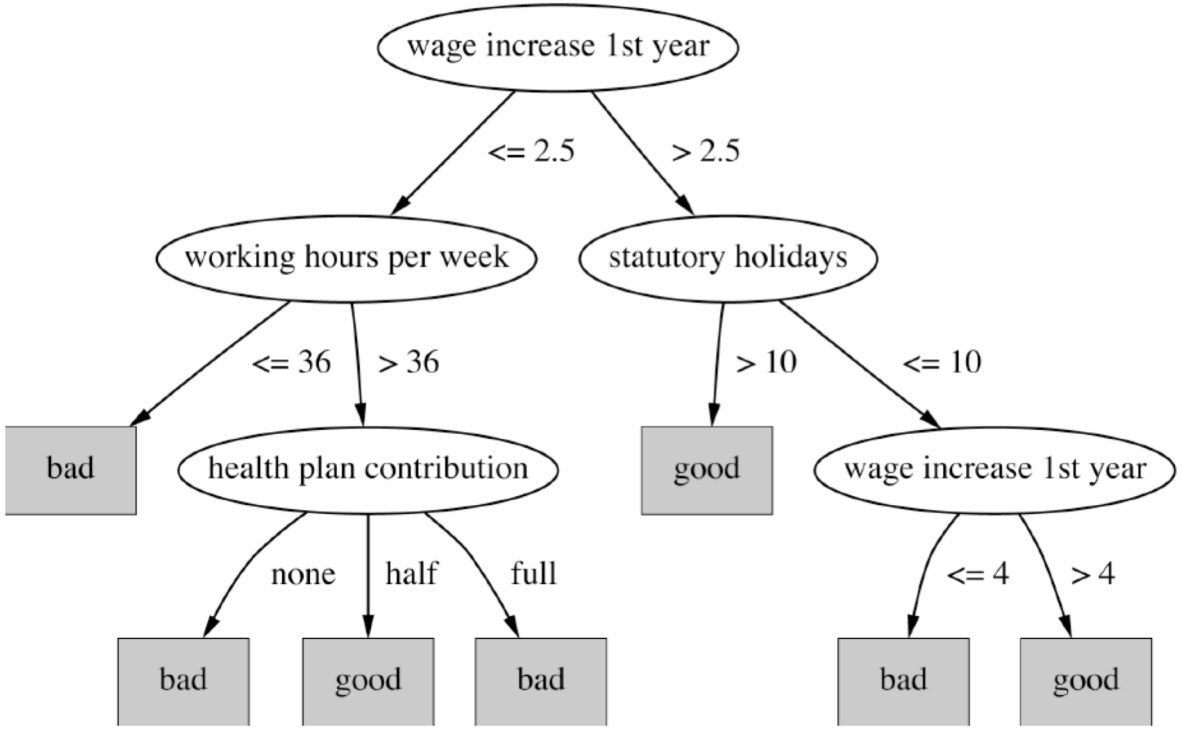
This is a much more realistic dataset than the others we have seen. It contains many missing values, and it seems unlikely that an exact classification can be obtained.

Attribute	Type	1	2	3	...	40
Duration	(Number of years)	1	2	3		2
Wage increase first year	Percentage	2%	4%	4.3%		4.5
Wage increase second year	Percentage	?	5%	4.4%		4.0
Wage increase third year	Percentage	?	?	?		?
Cost of living adjustment	{none,tcf,tc}	none	tcf	?		none
Working hours per week	(Number of hours)	28	35	38		40
Pension	{none,ret-allw, empl-cntr}	none	?	?		?
Standby pay	Percentage	?	13%	?		?
Shift-work supplement	Percentage	?	5%	4%		4
Education allowance	{yes,no}	yes	?	?		?
Statutory holidays	(Number of days)	11	15	12		12
Vacation	{below-avg,avg,gen}	avg	gen	gen		avg
Long-term disability assistance	{yes,no}	no	?	?		yes
Dental plan contribution	{none,half,full}	none	?	full		full
Bereavement assistance	{yes,no}	no	?	?		yes
Health plan contribution	{none,half,full}	none	?	full		half
Acceptability of contract	{good,bad}	bad	good	good		good

Missing

Decision trees for the labor data

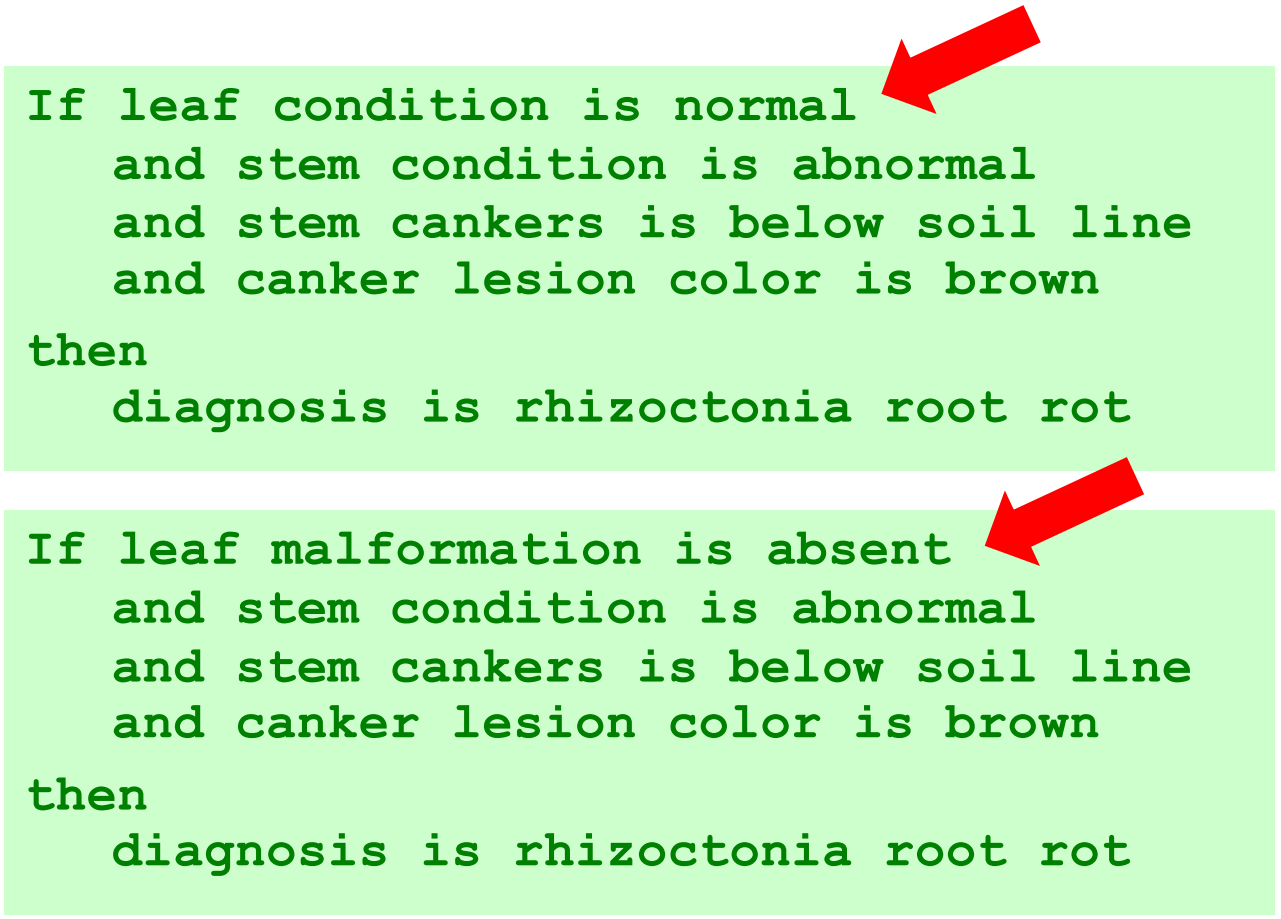




Soybean classification

	Attribute	Number of values	Sample value
Environment	Time of occurrence	7	July
	Precipitation	3	Above normal
...			
Seed	Condition	2	Normal
	Mold growth	2	Absent
...			
Fruit	Condition of fruit pods	4	Normal
	Fruit spots	5	?
Leaf	Condition	2	Abnormal
	Leaf spot size	3	?
...			
Stem	Condition	2	Abnormal
	Stem lodging	2	Yes
...			
Root	Condition	3	Normal
Diagnosis		19	Diaporthe stem canker

The role of domain knowledge

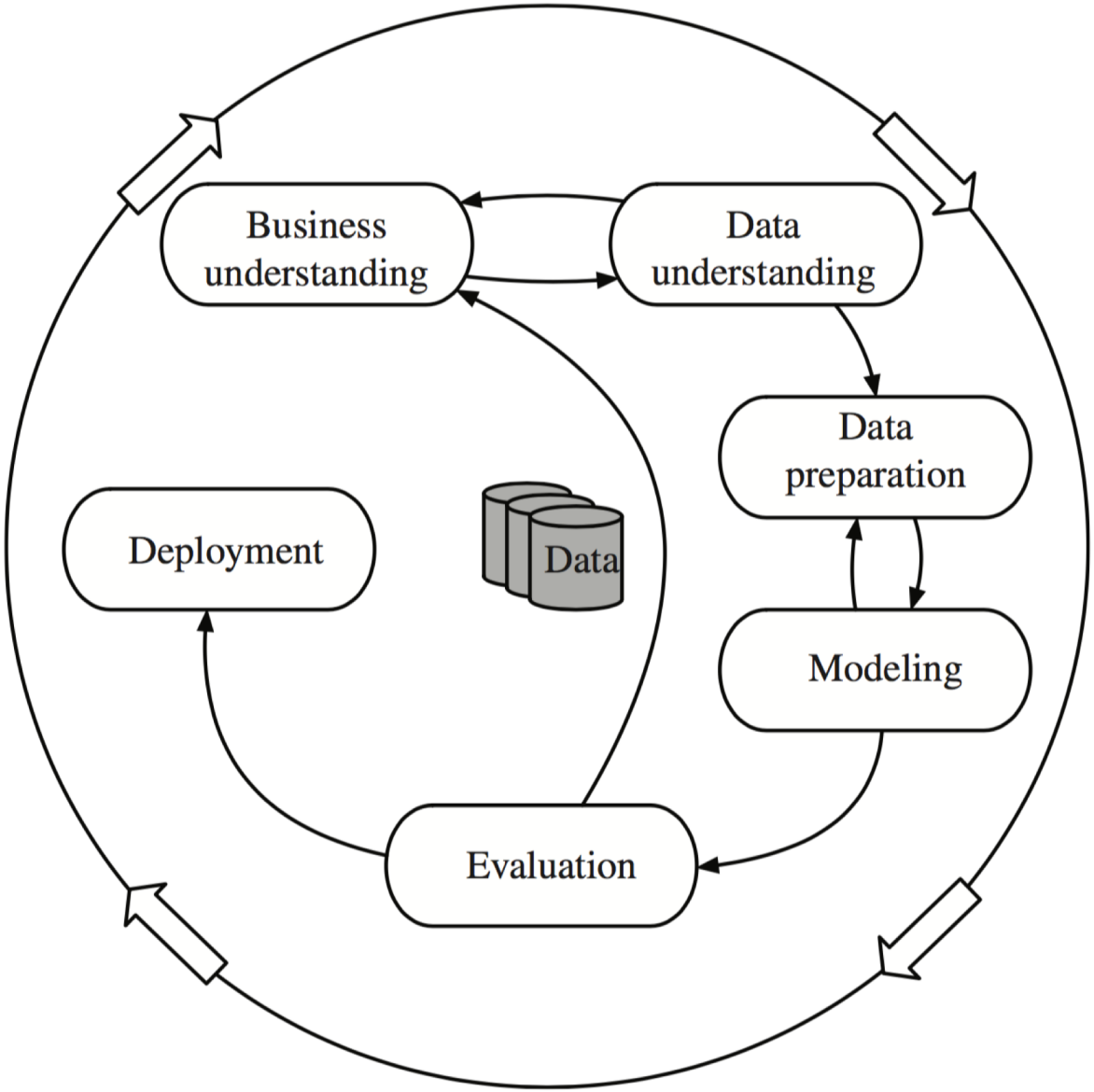


But in this domain, “leaf condition is normal” implies “leaf malformation is absent”!

Cybersecurity applications

- There are many applications in cybersecurity
  - Malicious network traffic detection
  - Intrusion detection
  - Malware detection
  - Detection of “abnormal” behavior
  - Fake news detection
  - …

The data mining process



Classincal problems in machine learning

- ***Regression:*** trying to predict a real value.
- ***Binary Classification****:* trying to predict a simple yes/no response.
- ***Multiclass Classification:*** trying to put an example into one of a number of classes.
- ***Ranking:*** trying to put a set of objects in order of relevance.

NEW YORK UNIVERSITY

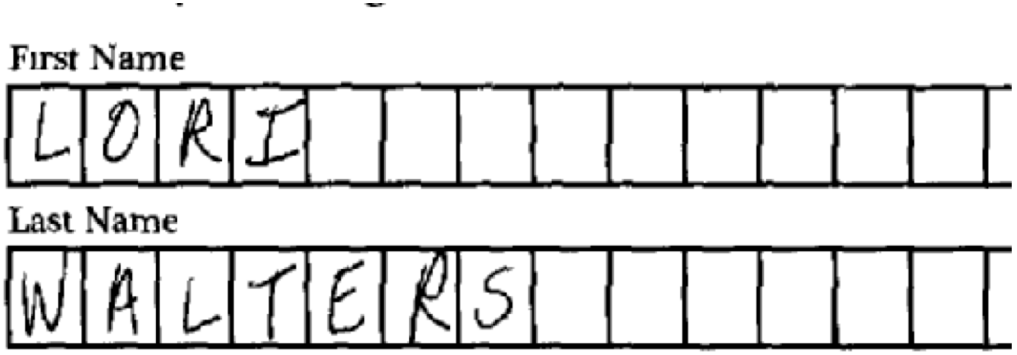
DLI Teaching Kit

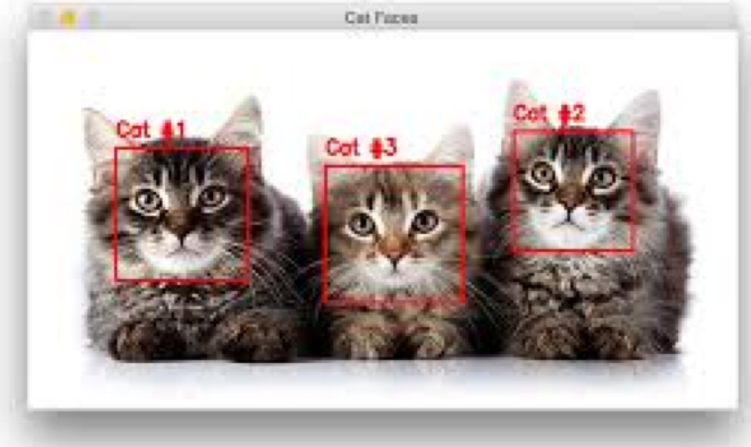
Lecture 1.2 - Introduction to Machine Learning

Deck credit: J. Seng

Machine Learning

- Machine Learning is the ability to teach a computer without explicitly programming it
- Examples are used to train computers to perform tasks that would be difficult to program





Types of Machine Learning

- Supervised Learning
  - Training data is labeled
  - Goal is correctly label new data
- Reinforcement Learning
  - Training data is unlabeled
  - System receives feedback for its actions
  - Goal is to perform better actions
- Unsupervised Learning
  - Training data is unlabeled
  - Goal is to categorize the observations

Applications of Machine Learning

- Handwriting Recognition
  - convert written letters into digital letters
- Language Translation
  - translate spoken and or written languages (e.g. Google Translate)
- Speech Recognition
  - convert voice snippets to text (e.g. Siri, Cortana, and Alexa)
- Image Classification
  - label images with appropriate categories (e.g. Google Photos)
- Autonomous Driving
  - enable cars to drive

Features in Machine Learning

- Features are the observations that are used to form predictions
  - For image classification, the pixels are the features
  - For voice recognition, the pitch and volume of the sound samples are the features
  - For autonomous cars, data from the cameras, range sensors, and GPS are features
- Extracting relevant features is important for building a model
  - Time of day is an irrelevant feature when classifying images
  - Time of day is relevant when classifying emails because SPAM often occurs at night
- Common Types of Features in Robotics
  - Pixels (RGB data)
  - Depth data (sonar, laser rangefinders)
  - Movement (encoder values)
  - Orientation or Acceleration (Gyroscope, Accelerometer, Compass)

Measuring Success for Classification

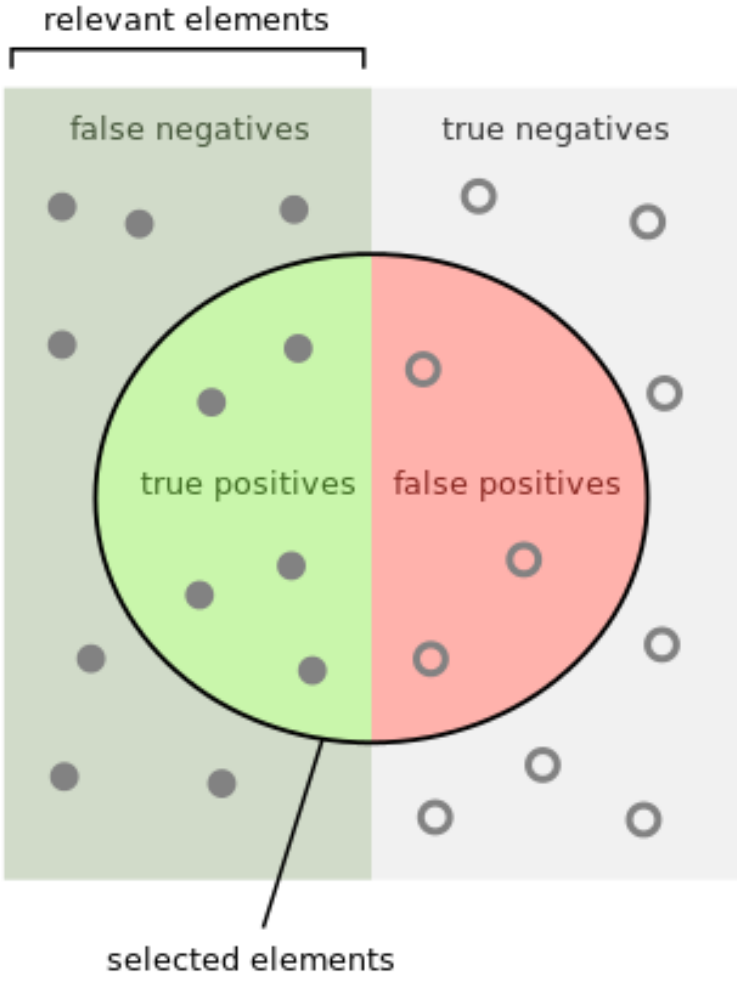
- True Positive: Correctly identified as relevant
- True Negative: Correctly identified as not relevant
- False Positive: Incorrectly labeled as relevant
- False Negative: Incorrectly labeled as not relevant

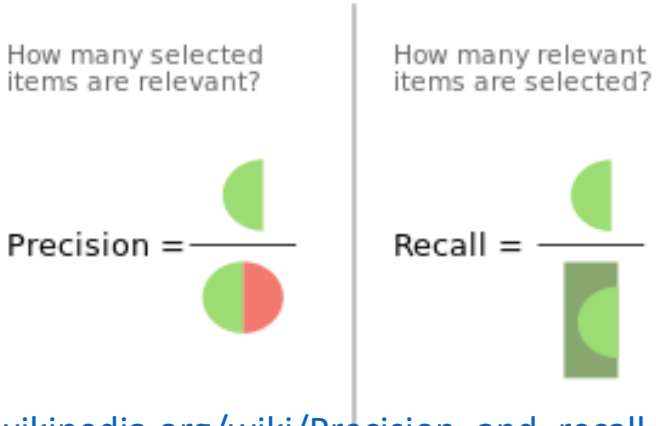
Example: Identify Cats

Images from the STL-10 dataset

Precision, Recall, and Accuracy

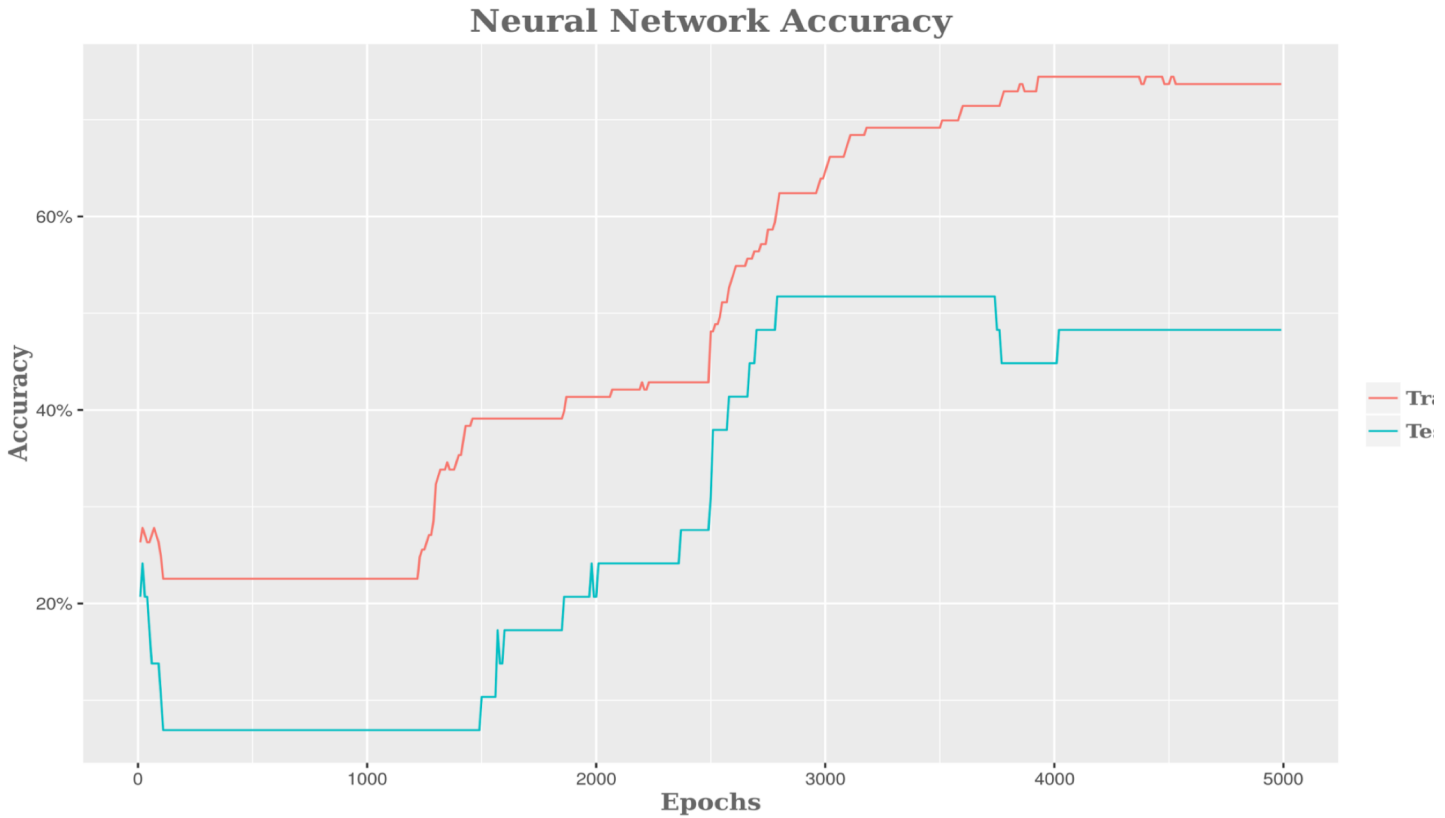
- Precision
  - Percentage of positive labels that are correct
  - Precision = (# true positives) / (# true positives + # false positives)
- Recall
  - Percentage of positive examples that are correctly labeled
  - Recall = (# true positives) / (# true positives + # false negatives)
- Accuracy
  - Percentage of correct labels
  - Accuracy = (# true positives + # true negatives) / (# of samples)





https://en.wikipedia.org/wiki/Precision_and_recall

Training and Test Data



- Training Data
  - data used to learn a model
- Test Data
  - data used to assess the accuracy of model

- Overfitting
  - Model performs well on training data but poorly on test data

Bias and Variance

- Bias: expected difference between model's prediction and truth
- Variance: how much the model differs among training sets

- Model Scenarios
  - High Bias: Model makes inaccurate predictions on training data
  - High Variance: Model does not generalize to new datasets
  - Low Bias: Model makes accurate predictions on training data
  - Low Variance: Model generalizes to new datasets

Supervised Learning Algorithms

- Linear Regression
- Decision Trees
- Support Vector Machines
- K-Nearest Neighbor
- Neural Networks

Supervised Learning Frameworks

Tool	**Uses**	**Language**
Scikit-Learn	Classification, Regression, Clustering	Python
Spark MLlib	Classification, Regression, Clustering	Scala, R, Java
Weka	Classification, Regression, Clustering	Java
Caffe	Neural Networks	C++, Python
TensorFlow	Neural Networks	Python

NEW YORK UNIVERSITY

DLI Teaching Kit

Thank you