1 Data Science Overview

Definitions

Data science	• Machine learning on big data • Extraction of knowledge from data through a complete data lifecycle process • Use of statistical and machine learning techniques on a huge data to interpret its meaning • Everything that has to do with data (collecting, wrangling, analyzing, modeling, etc.)
Big data	Large data sets such that old data processing method does not work

Machine Learning
- Allowing computers to learn and improve by themselves
- Through algorithms and special techniques
- No need any human intervention

Reasons for Machine Learning
- Human expertise not available / cannot explain their expertise
- User personalization
- Wide variety of situations, humans unable to handle
- Large amount of data
- Humans are expensive to work on

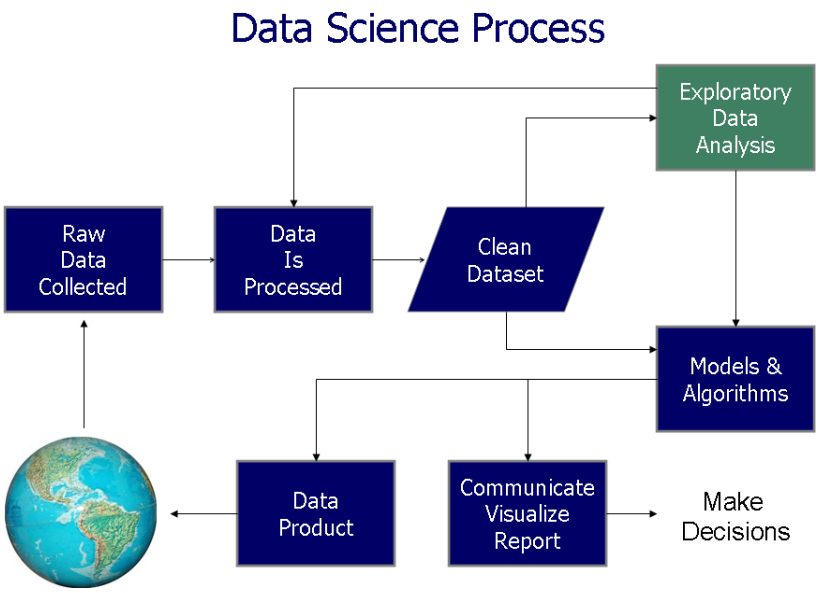
Other Reasons for Machine Learning
1. Information society
2. Information warfare
3. Information overload
4. Information access

Rise of Big Data

1. Data collection	• There is huge amount of data collected • Humans are not efficient in managing big data • Use machine learning to look for patterns and categorize them • E.g.: Google language translation where it learns better to translate many languages as time rolls on
2. Datafication	• Collecting the public's data in all fields • If a disease is circulating some parts of the society, government could use the statistics and make more investments as to reduce the risks of spreading • Use machine learning to predict the rise of the next disease
3. Information society	• **similar to data collection** • When a society is run by information, data brokers collect vast amount of data • This makes humans unable to manage them effectively

The Hype Cycle: It quantifies the level of maturity in various technologies

Overall Data Science Process

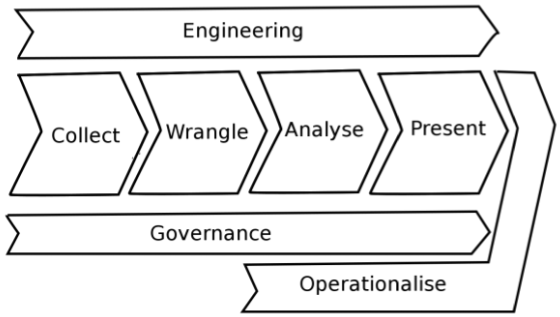


Data Science Process

Process	Description
1. Pitching Ideas	Pitching data science projects to others
2. Collecting Data	Collect data from customers/patients/people
3. Monitoring	Monitors data collected by various instruments
4. Integration	Data come from different sources
5. Interpretation	Obtain meaning / analysis from: • Linked Open Data (LOD) • Database schema
6. Governance	Roles: • Caring for data and its subjects • Managing data standards and formats
7. Engineering	Data engineers work at the back-end to manage data/servers
8. Wrangling	Inspecting and cleaning the data
9. Modelling	Methods: • Propose a functional model • Build models with different tools • Analyzing the statistics • Use machine learning to work on data
10. Visualization	Methods: • Visualizing data to interpret it • Choosing appropriate visualizations for the data
11. Operationalization	Putting the results to work

Standard Value Chain (for Data Science Project)

Parts	Description
Collection	Getting the data
Engineering	Storing and computing the data over a full lifecycle
Governance	Managing the overall data across full lifecycle
Wrangling	Cleaning the data
Analysis	Interpreting the data (through visualizations, graphs, etc.)
Presentation	Arguing the case for results are useful
Operationalization	Putting the results into work to gain value



Roles in Data Science Project

	Collection	Engineering	Wrangling	Analysis
Business Analyst	Copy and paste into excel	Use excel to store and retrieve	Use excel functions	Charts
Programmer	Web APIs, scraping database queries	Flat files	Python and Perl	Matplotlib in Python, R
Enterprise	Application database, intranet files, server logs	Teradata, oracle, MS SQL Server	Talend, informatica	Cognos, business objects, SAS, SPSS
Web Company	Application database, server logs, crawl data	Hadoop/Hive, Flume, HBase	Pig, Oozie	Dashboards, R

Relationship with Data Science

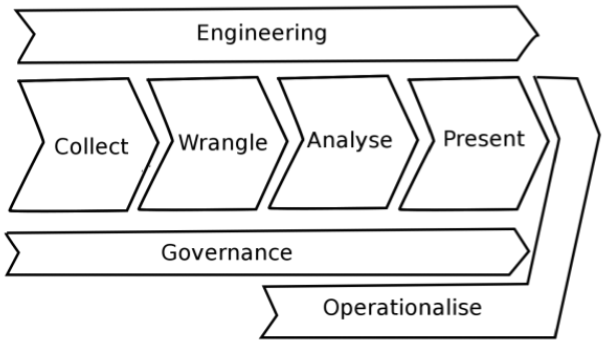
1. Data Engineering	Building a scalable data system for storing and processing data <u>Systems like:</u> • Databases • Distributed processing • Cloud computing
2. Data Analysis	Performs analysis and present results <u>Results can be represented via:</u> • Machine learning • Visualization • Computational statistics
3. Data Management	Managing data through its lifecycle which is according to regulations Data usually addresses issue like: • Ethics, privacy, security • Curation (organize and integrate data), governance

2 Roles of Data Scientists

Type of Data Jobs

Type	Description
Data Analyst	Primarily people who develop insights with data (excel, R)
Data Scientists	Primarily people who develop data models and products and turn them into insights (python, R)
Data Engineers	Primarily people who manage data infrastructure, automate data processing, and deploy models at scale (SQL)

Standard Value Chain



	Data Scientist (Does ✓ and knows about ✗)	Chief Data Scientist (manages ✓)
Engineering	✗	✓
Collect	✓	✓
Wrangle	✓	✓
Analyze	✓	✓
Present	✓	✓
Governance	✗	✓
Operationalize	✗	✓ (evaluates)

Types of Data Scientists

Data Scientist	Address the data science process to extract meaning from data
Chief Data Scientist	A form of chief scientist who addresses data management data engineering and data science goals
Chief Scientist	A corporate position that is responsible for science related aspects of a company

Skills of Data Scientists

Field	Description
Business	Product development, business
Machine learning / big data	Unstructured data, structured data, machine learning, big and distributed data
Mathematics / operations research	Optimization, mathematics, graphical models, algorithms
Programming	System administration, back end programming, front end programming
Statistics	Visualization, temporal statistics, surveys and marketing, spatial statistics, science, data manipulation

Data Scientist vs Data Engineers

	Data Scientist	Data Engineers
Differences	• Visualization • Modelling • Story-telling	• System implementation • DB administration • Data storage
Similarities	• Statistics • Math • Programming	

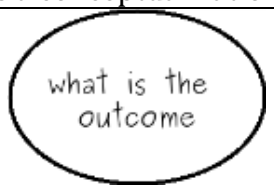
3 Data Business Models & Application

Value Chains
- A sequential process to create value
- Refers to the Standard Value Chain

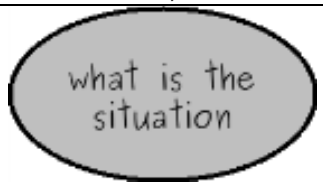
Analytic Levels
- Broad classification of different analysis
- Levels are classified into three type of analysis as below

Type	Description	Examples
Descriptive Analysis	Gain insight from historical data	• Relate sales revenue by region or by product category • Correlate advertising revenue per region
Predictive Analysis	Making predictions using statistical and machine learning techniques	• Predict next quarter's sales result using economic projection (understanding on statistics)
Prescriptive Analysis	Recommending decisions using optimizations, simulations, etc.	• Recommend which region to advertise based on budget

Influence Diagrams
- Method for modelling data and decision making
- It's a conceptualization aid to think about actions, values and unknowns



Chance Variable
- A chance variable
- An uncertain quantity

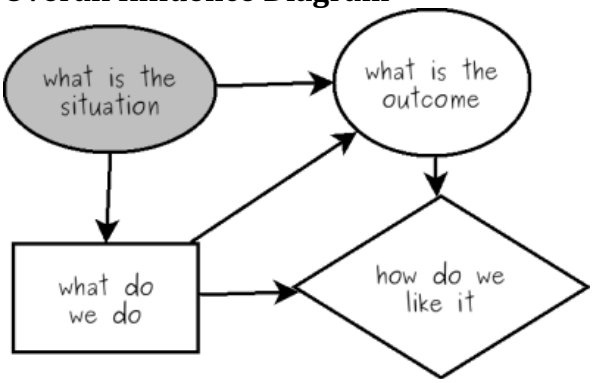


Known variable
- A known variable
- An uncertain quantity

Decision or action
- A decision or action

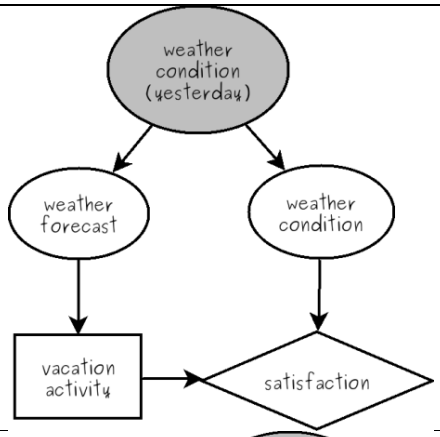
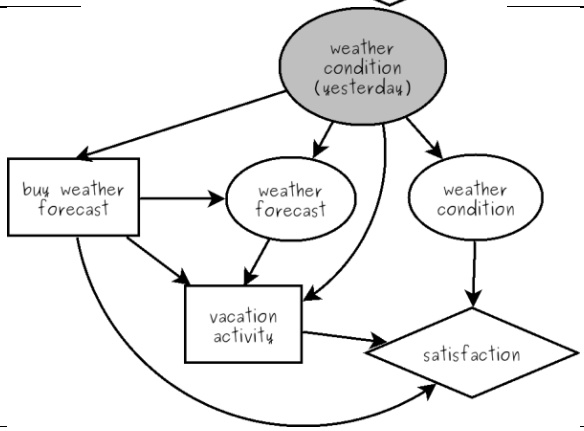
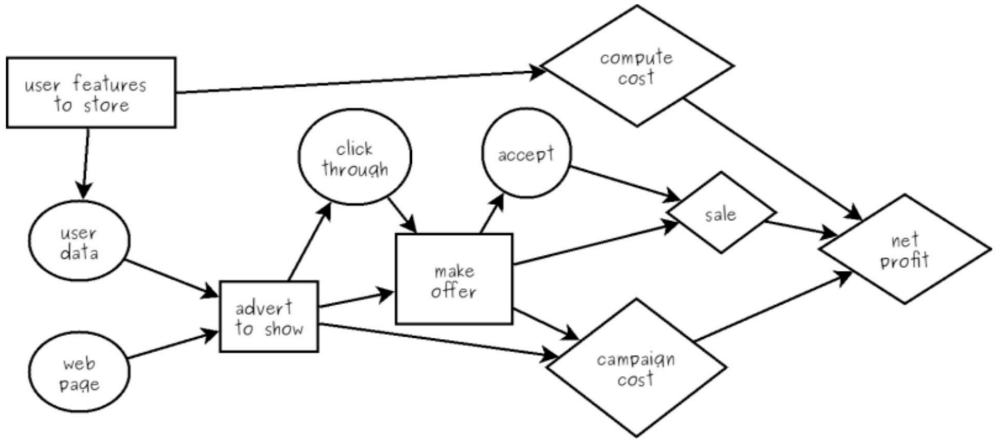
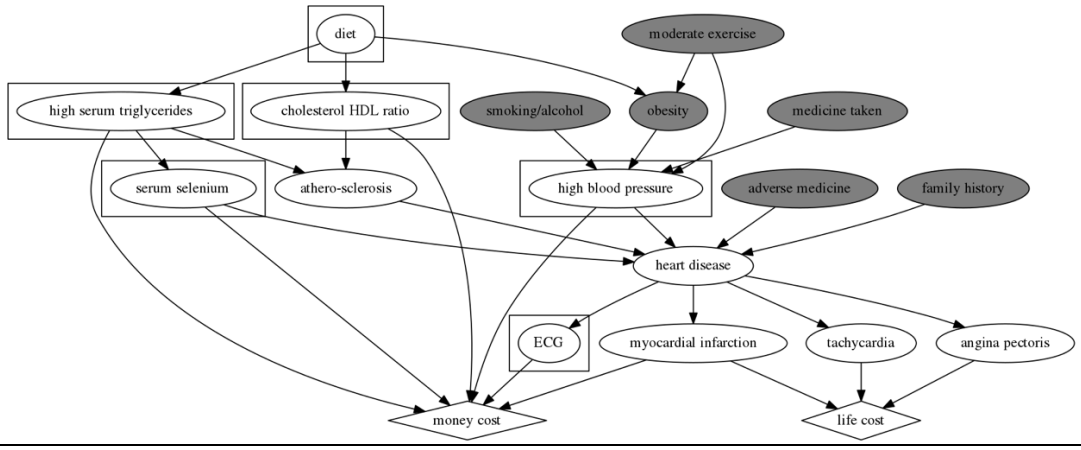
Objective
- An objective is a measure of value

Overall Influence Diagram



directed graphs (arrows) are used to convey influence

Examples of Influence Diagrams

Last Minute Vacation	
Last Minute Vacation with Forecast	
Internet Advertising	
Heart Disease	

Business Models

- Business models describe how an organization creates, delivers and captures value in economic, social, cultural or other contexts
- General classes of business are retail wholesaler, software vendor, service provider
- Two types of business models – normal (traditional) and intelligent systems (modern)
- Software and a Service (SaaS) is both a software vendor and consultancy, running under traditional IT business models

Data Business Models

Type of Model	Description	Examples
Information brokering service	Buy and sells data for others (middleman-broker)	Bloomberg terminal • A computer system used to monitor and analyze real-time financial market data • Places trades on the electronic trading platform
Information-based differentiation	Satisfies customers by providing a differentiated service built on the data	Amazon • Superior info (like reviews) about products
Information-based delivery service	Deliver data to others	Amazon • Merchants in marketplace directly get customers
Information provider	Selling the data that it collects	LexisNexis • Provide legal and public records data • Like traditional business model, selling data • Known as data economy (as it sells data only)

Intelligent Systems Business Models

Type	Description	Examples
Data providers	Collect datasets	Satellite data Google maps
Alchemists	Promise to turn data into gold	Self-service APIs
Gateways	Create new cases from specific data types	Image Audio Video Genomic data (genetics data)
Magic wands	Fix a workflow using software as a service tool (SaaS)	Help recruiters write better job descriptions
Navigators	Autonomous system for the physical world	Self-driving cars
Agents	Create cyborgs (smart robot) and bots to help with virtual tasks	Customer service real-time chat

4 Data Characterization and Big Data

The Four V's of Big Data

Type	Description	Examples
Volume	Scale of data	• 1.2 trillion google searches per year • 6 billion smartphones worldwide • 100 terabytes of data stored on average for every company
Variety	Different forms of data	• 4 billion hours YouTube videos • 30 billion pieces of Facebook content
Velocity	Analysis on streaming data	• New York Stock Exchange has 1 TB of trade info • Moderns cars have almost 100 sensors for pressure, fuel, etc.
Veracity	Uncertainty of data	• Fake news spread worldwide • Leaders don't trust information to make decisions

Metadata
Definition: Structured data that describes, explains, locates, makes it easier to retrieve, use or manage an information resource

Key Concepts of Metadata

Concepts	Description	Examples
Machine-readable data	Data that can be understood by a computer	JPG XML JSON
Markup language	System of annotating a document in a way that is distinguishable from the text	Markdown HTML Javadoc
Digital container	File format that have characteristics that describe how different type of data coexist in a computer file	MPEG (compressing video audio)

Advantages of Metadata
- Easier to discover data
- Easily determine the applicability of the data
- Enable interpretation and reuse
- Clarify ownership and restrictions on reuse

Examples of Metadata
1. EXIF images
2. Book (author, copyright, etc.)
3. Media (author, title, location, date, license, etc.)
4. Call data record
5. Javadoc

Kinds of Data
1. Geospatial data (geographical shape data)
2. Linked open data (XML)
3. IP connection data
4. Transactional data (banking)
5. Twitter data
6. Internet of Things data

Growth Laws

→ Describes the growth of computing power

Type of Law	Description
Moore's law	• Number of transistors double every two years • Smaller chips but faster performance • Current pace is slowing
Koomey's law	• Followed after Moore's law • Number of batteries needed will fall by a factor of 100 every decade • Resulted in computers everywhere
Bell's law	• Followed after Moore's law and Koomey's law • A new computing class will emerge every decade • Examples: PCs, mobile computing, cloud, internet of things
Zimmerman's law	• Computers' ability to track us double every two years • Surveillance increase and privacy decrease

Difference between Python and R

Python	R
Developed by computer scientists	Developed by statisticians, more towards analysis
Easy to integrate with other systems	More on stand-alone analysis and exploration
Easier to learn	Harder to learn
Great libraries Good visualization tools for basis analysis	Less scalable

5 Data Sources and Case Studies

Databases
Definition: Storing and accessing data

Database Methods

Method	Description
SQL	• A larger scale of Excel with better index and retrieval
JSON	• No format • Semi-structured key values pairs • Easier way of storing data than SQL • Friendly alternative to XML
Graph	• Stores graph as triples (subject, verb, object) • Used to store Linked-Open Data

Difference between SQL and NoSQL

SQL	NoSQL
• Structured data • Data does not change	• Unstructured data • Data constantly changing • Storing large data • Database offer wide variety of features

Distributed Processing
Definition: Breaking up computation to scale the power up

Process Overview

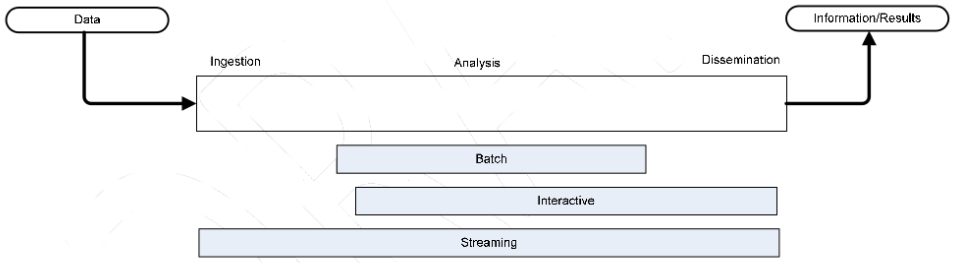


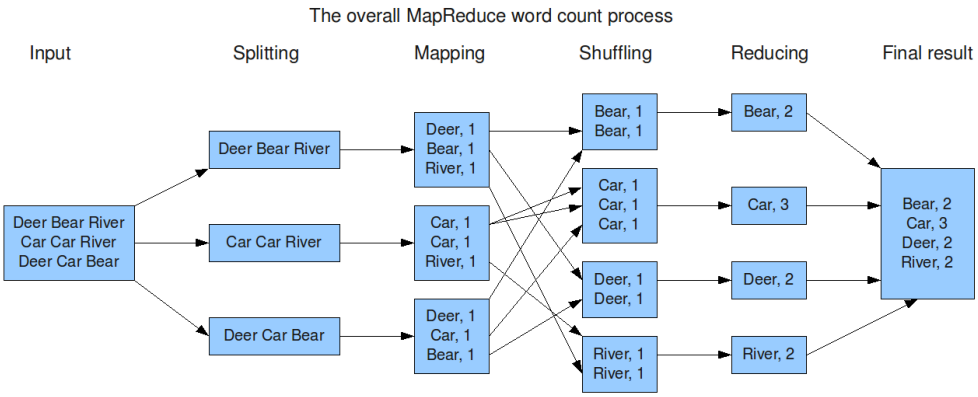
Figure 5: Information Flow

Part	Function
Interactive	Bringing humans into loop
Streaming	Massive data streaming through system with little storage
Batch	Data stored in large blocks, easier to be developed and analyzed

Process Concepts

Concept	Description
In-memory	Process is executed in memory (faster than hard disk)
Distributed computing	Process executed across multiple machines
Scalability	The ability to handle sophisticated work
Data parallel	The algorithm distributes the process into multiple blocks of data that runs independently

Map-Reduce Processing



- It performs filter, sorting (map) and summarizing (counting instances – reduce) the cluttered data
- Requires data parallelism (to run process independently, like shuffling)
- For simple word count task:

⇒ Data is divided across machines
⇒ map( ) to key value pairs (calculate number of occurrences)
⇒ sort and merge ( ) identical keys (merge is to total up the counts)

Types of Map-Reduce

	Hadoop (old)	**Spark (new)**
Similarities	Based on map-reduce architecture	
Type	Open-source Java	Interfaces in Python, Java, R
Features	• Can access Java libraries, tools • Not suited to streaming • Good for offline processing	• Works with some Hadoop ecosystem • Provides real-time in-memory processing (execute in memory instead of hard disks) • Faster than hadoop

Both Hadoop and Spark are a distributed storage and processing of big data. Such processing data is done by using map-reduce programming model to manage the big data in an efficient way.

6 Resources and Standards

Data Resources

Type	Description
Open data	• Data is publicly available • Usually provided by the government, or IT sectors • Data are machine readable but not always usable (need specific skills) • Must be picked by skilled people • Example of open data format: Linked Open Data (LOD), commonly used in graph database
Data wrangling	• Generate clean data from raw data so that it can be analyzed to generate powerful insights and results • Data comes in: ⇒ different shapes, sizes, formatting ⇒ wrong data entries • Data wrangling process like: ⇒ Data pre-processing ⇒ Data preparation ⇒ Data cleansing ⇒ Data transformation

7 Resources Case Studies

Semi-Structured Data

- A form of structured data but does not conform to the structure of the relational databases
- Examples:

⇒ XML: A form of markup language that is readable for both humans and machines
⇒ JSON: Open standard file format that uses human readable format to transmit data objects
⇒ YAML: **similar to JSON** but with extra indentation so that it's easier to read

XML

```
<person>
  <firstName>John</firstName>
  <lastName>Smith</lastName>
  <age>25</age>
  <address>
    <streetAddress>21 2nd Street</streetAddress>
    <city>New York</city>
    <state>NY</state>
    <postalCode>10021</postalCode>
  </address>
  <phoneNumber>
    <type>home</type>
    <number>212 555-1234</number>
  </phoneNumber>
  <phoneNumber>
    <type>fax</type>
    <number>646 555-4567</number>
  </phoneNumber>
  <gender>
    <type>male</type>
  </gender>
</person>
```

JSON

```
{
  "firstName": "John",
  "lastName": "Smith",
  "age": 25,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021"
  },
  "phoneNumber": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "fax",
      "number": "646 555-4567"
    }
  ],
  "gender": {
    "type": "male"
  }
}
```

YAML

```
firstName: John
lastName: Smith
age: 25
address:
  streetAddress: 21 2nd Street
  city: New York
  state: NY
  postalCode: '10021'
phoneNumber:
- type: home
  number: 212 555-1234
- type: fax
  number: 646 555-4567
gender:
  type: male
```

Markup Language

- Markup language is a method to distinguish between different texts
- Example: Predictive Model Markup Language (PMML)

⇒ A standard language that describes a predictive model that can be passes between analytic software (like R to SAS)

⇒ It acts like a translator / bridge that connects analytical software and a software that is used to generate a predictive model (most common example would be R – clustering)

Software Usage

Operating System	Programming Language	Relational Database	Management and Big Data	Visualization
• Windows • Linux • Mac OS • Unix • iOS • Android	• SQL • R • Python • Bash • Java, JavaScript • Visual Basic • C++ • Matlab	• MySQL • SQL • Oracle • SQLite • Teradata • Vertica • IBM DM2	• Spark • Hive • MongoDB • Amazon redshift • HBase • Pig • Impala • Cassandra	• Ggplot • Tableau • Matplotlib • Shiny • D3 • Google Charts • Bokeh

API and SaaS

	API	REST	SaaS
Full Name	Application Programmer Interface	Representational State Transfer	Software as a Service
Function	Routinely provides pragmatic access to an application	A stateless API running over HTTP (simplier) *standard rules on structuring API via web*	Provides the software in a web browser and/or via an API over the web as a subscription service
Examples	• Twitter API ⇒ Followers, locations, hashtags ⇒ Retweet counts ⇒ User info • Google Maps API ⇒ Latitude, longitude ⇒ Estimated arrival time ⇒ Distance ⇒ Real time traffic situation		• Emailing service ⇒ Gmail ⇒ Microsoft Office 365 • File sharing service ⇒ Dropbox ⇒ Google drive ⇒ Box • Business service ⇒ Salesforce ⇒ Servicenow
Features	-		• Pay as you go • Low maintenance • Better performance Cons: data privacy

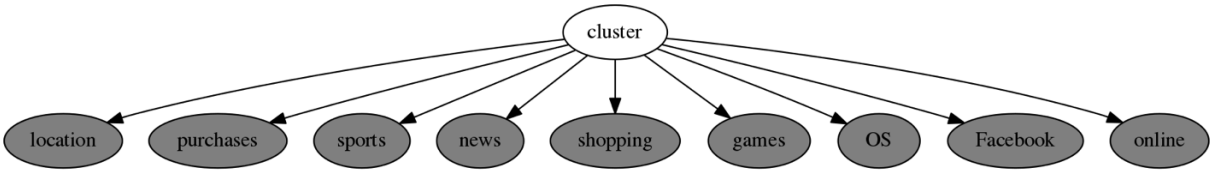
8 Data Analysis Theory

Prediction Task

	Simple Prediction Task	Complicated Prediction Task
Properties	• Model contains known values • Using those known values to reach certain goals (e.g. achieve lesser loss on defaulting a loan) • Known variables are not related to each other	• Model contains many known and unknown variables (known as Bayesian Network) • Different datasets have different knowns and unknowns • Known and unknown variables might have connection to one another
Training Set	Use dataset of cases where it has known variables that to predict the outcome (e.g. predict if a person's loan will be defaulted based on historical known values)	Use dataset of cases where it has both known and unknown variables to predict the outcome (e.g. predict if a person will get heart disease based on prior patients' known (like high blood pressure) and unknown values (like serum selenium)

Segmentation Task

- Clustering is done by grouping data that have similar characteristics
- By doing so, it makes us easier to analyze and predict because the data is now in segments
- At first, we do not know the clusters to group, hence the cluster variable is unknown
- Then, we group the points that have similar characteristics (e.g. cluster to 10 groups)
- Based on the grouping that we have done (clustering), we use that to create our segments
- From here, we could make better analysis and prediction in the form of groups
- An example of clustering is as below, where the cluster variable is unknown and then it then identifies the segments based on the number of clusters that we've chosen

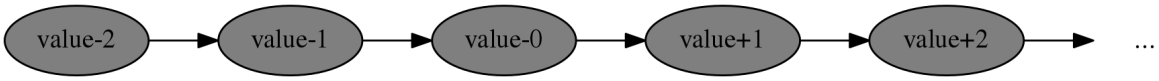


Time Series Forecasting
1st Order
→ Predicts the next value based on only the previous value in the same series

Prediction

- The prediction (outcome) is done solely by using the previous value only
- For example, the series above shows that the "loss" is solely based on the "value+1"

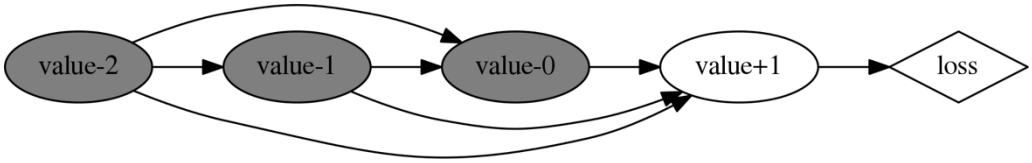
Training Set



- To train a predictive 1st order time series model, we use the dataset that has one or more series of values
- More values increase the accuracy of the prediction

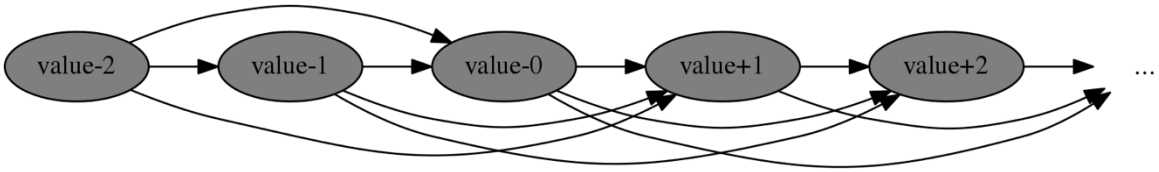
3rd Order
→ Predicts the next value based on <u>all previous values</u> in the same series

<u>Prediction</u>



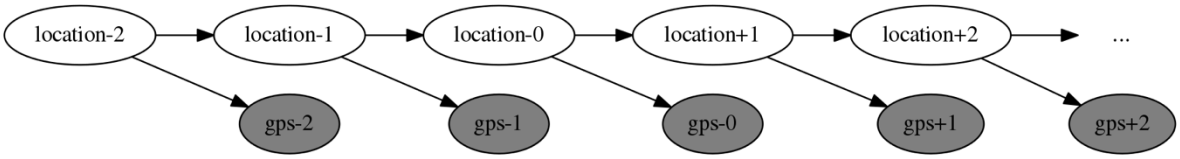
- The prediction (outcome) is done based on all the prior values
- For example, the "value+1" that produces the outcome is determined by all the previous values

<u>Training Set</u>



- To train a predictive 3rd order time series model, we use the dataset that have a sequences of data
- More values increase the accuracy of the prediction

Sequential Learning: GPS Tracking



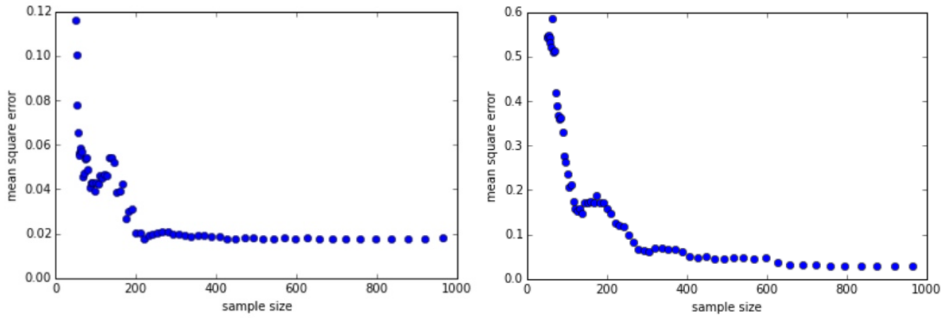
- Our exact location (or coordinates) is unknown at all times
- Time series prediction is used to estimate a person's approximate location based on factors like GPS signal, movement speed, signal noise, etc.
- The current location is usually estimated by taking in account of the previous location and aspects like movement speed that indicates how fast a person moves from one place to another

Learning Theory
- A subfield of artificial intelligence that studies the design and analysis of machine learning
- Deals with the concepts of how machine learns, how it checks the accuracy of its predictions, etc.

Truth and Quality (of prediction)

Truth	Quality
• A true model can be obtained by collect massive data • This true model must be able to make as accurate predictions as possible	• Quality is measured through a function of loss • 3 types of measurements: ⇒ Loss: quality of prediction is bad ⇒ Gain: quality of prediction is good • Error is measured by considering the difference between the prediction and truth value (Function of error)

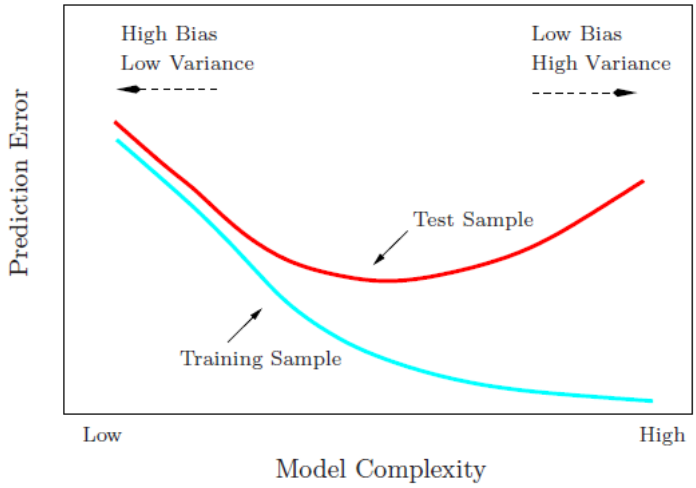
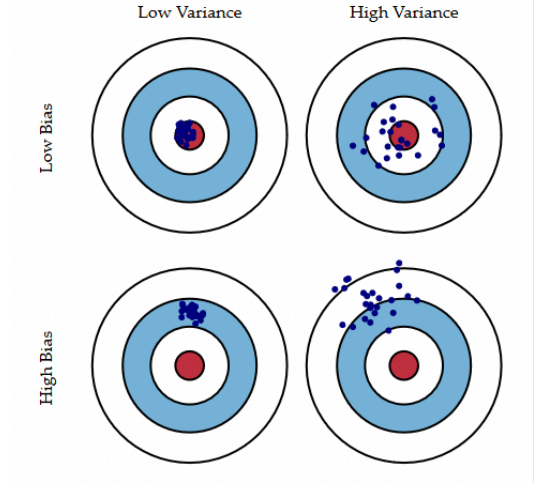
Loss and Training Data



- Mean square error (MSE) can be calculated according to the size of the training data
- The "learning curve" above shows that the greater the data size, the lower the mean square error
- In general, the accuracy of a prediction is said to be much better with bigger sample size

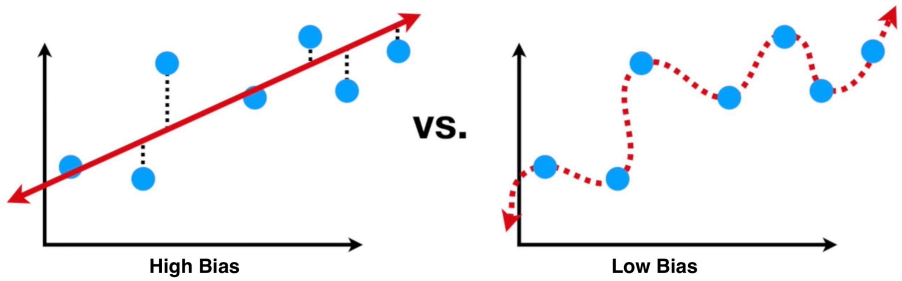
Bias and Variance

Bias	How much the prediction differs from the desired regression function Simplified: how much difference between the prediction values and the actual values
Variance	How much the predictions for individual sets vary around their average Simplified: how much the prediction values are spread out

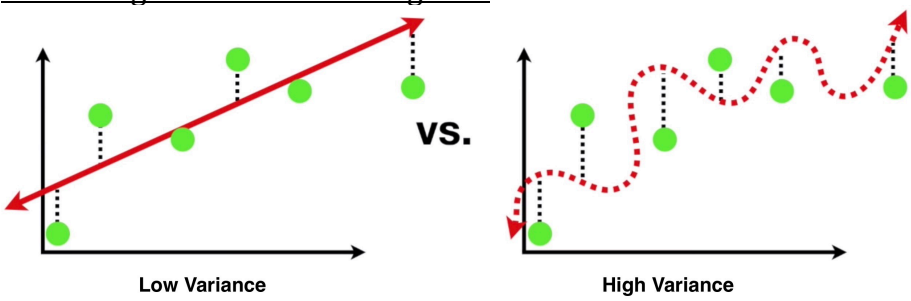
Bias-Variance Tradeoff	**Comparison**
	

Simple Model and Complex Model

Building models with training set:



Evaluating models with testing test:



Underfitting (simple model)	Overfitting (complex model)	Optimum Model Complexity
When the regression line is fitted on the training set, it's seen as high bias as the line differs a lot than the curve. When the line is plotted on the testing test, it has low variance because the values spread out lesser.	When a regression curve is fitted on a training set, it is seen as low bias because the curve is similar to the true curve. When the curve is plotted on the testing test, it has high variance because the values spread out a lot.	Low bias and low variance are considered as the optimal model. It finds the sweet spot between a simple model (regression line) and a complex model (regression curve)

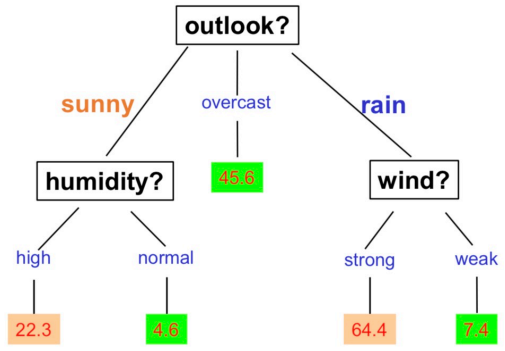
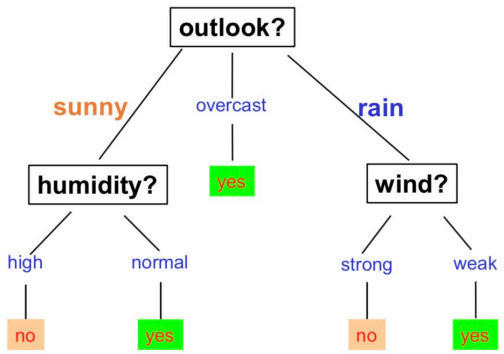
No Free Lunch Theorem

- No optimization algorithm is expected to perform better than any other optimization algorithm
- Any two optimization algorithms are equivalent when their performance is averaged across all possible problems

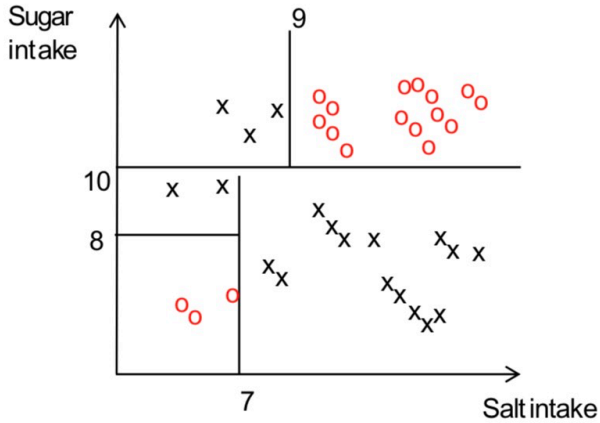
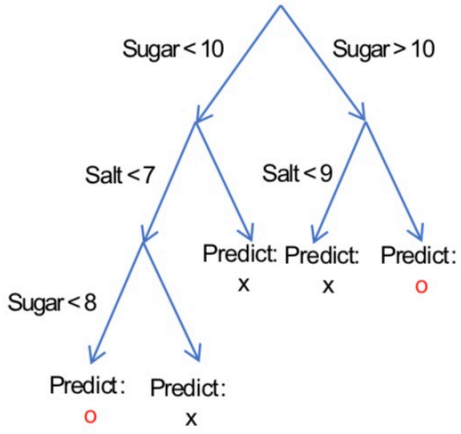
Ensembles

- It uses several different machine learning algorithms to improve the predictive performance
- Through the collection of possible models, we could understand the range of realistic predictions
- The accuracy of the prediction is obtained by averaging the predictions of a few models

9 Regression and Decision Trees

	Regression Tree	Decision Tree
Function	To predict continuous values (i.e. real values)	To predict binary/categorical outcomes (i.e. yes-no question)
Example		
Outcome	Predicts most <u>common values</u> in each region	Predicts <u>average values</u> in each region

Recursive Partitioning
1. Recursively partition the space into regions
2. Group similar instances together

Partitioning	Prediction Tree
	

Analysis of Prediction Tree
- This prediction tree predicts the classification of a value
- If the sugar is less than 10, it checks if the salt is less than 7
- If the salt is less than 7, it checks if the sugar is less than 8. Or else, it's classified under X
- If sugar is less than 8, then it is classified under o. Or else, it's under X
- This is decision tree as the prediction by using the common values in a region

Random Forest: A form of ensembles learning method that constructs multiple decision trees based on different criteria and selects the best tree

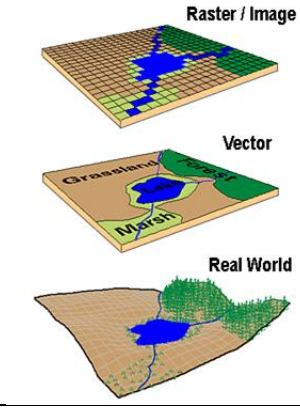
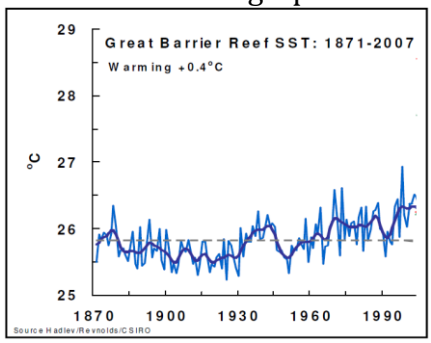
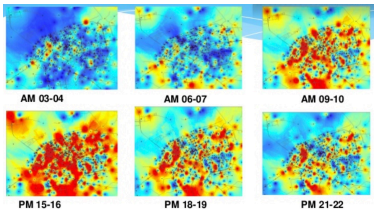
10 Data Analysis Process

Preprocessing Data

- Checks for missing data and do something on it

	Normalization	Imputation
Used on	Data with <u>no missing values</u>	Data with <u>missing values</u>
Method	Scale the data so that it falls within a specific range (like [0,1])	"fill in" the missing variables using imputation <u>Imputation</u>: • Inputting values with crude approximation • Use complex algorithms to input values

Types of Data

Types of Data	Spatial (or geospatial)	Temporal	Spatio-Temporal
Properties	Geographical data that whose numerical values are represented as coordinates	Data that varies over time	Integration of space and time (combine both spatial and temporal data)
Examples	Physical object like the geo (land) construction 	Statistical graph 	Plotting the statistics on the geographical map based on its coordinates 

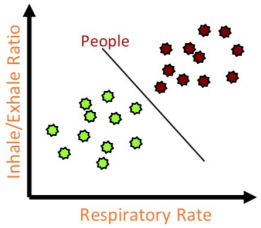
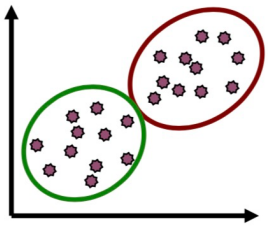
Data Analysis Tools

Type	Tools
Access	SQL, Hadoop, MS SQL Server, PIG, Spark
Wrangling	Common scripting languages (Python, Perl)
Visualization	Tableau, MATLAB, JavaScript
Statistical Analysis	Weka, SAS, R
Multi-purpose	Python, R, SAS, KNIME, RapidMiner
Cloud-based	Azure ML (MS), AWS ML (Amazon)

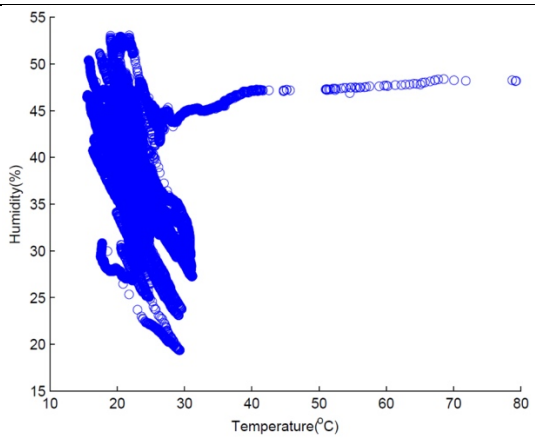
Scripting Languages

- Refers to high level programming languages
- It automates tasks that are originally done one-by-one by hand
- It also extracts information from a dataset
- Easier and faster to code as compared to traditional languages like C++
- This is typically in the command prompt or terminal that allows faster execution
- For example, some scripting language like Bash, Python, Perl, R

Difference between Classification and Clustering

	Classification	**Clustering**
Grouping	The groups are known beforehand, group some of the points by labelling them	The groups are unknown beforehand, cluster based on the suggested patterns in data
Algorithm	Using decision tree, regression tree	Using k-means
Objective	Used to classify all future observations	Used to explore and analyze different types of datasets
Example		

Anomaly Detection

Anomaly	Data points that are inconsistent to the normal data points
Anomaly Detection	Process of finding abnormal data in data sets
Example	 The temperature of the data points is not consistent, there is a section where the temperature escalated upwards until 80 Celsius.

11 Data Management

Data Management

- It organizes and manages the data processes as a valuable resource and to meet lifecycle needs.
- It controls, protects, delivers and enhances the value of the data and information assets.

Terminologies

Privacy	Having control over how one shares oneself with others Example: • only allowing personal files (like diaries) to be known to himself • closing the blinds in the room
Confidentiality	Information privacy, how information about an individual is treated and shared Example: • disallowing anyone to view your browser search history • keeping personal medical records • sharing location data to only specific apps
Security	Protection of data, prevent from improper usage Example: • prevent hackers from misusing stolen card information
Implicit data	Data is not explicitly stored but can be interfered with details from known data Examples: • things can be predicted solely based on a person's Facebook likes • online retailers predict if a woman is pregnant based on purchase history

Compliances and Regulations

Key Points	1. Ethics: Handling data morally 2. Regulations: Ensuring that confidentiality is met 3. Compliances: Ensuring that regulations are met
Example	**1. PCI (Payment Card Industry)** PCI was established to reduce credit card fraudulent transactions. This is done by placing all credit card information in an encrypted format, disallowing anyone other than the clients themselves to view their data. Such regulations must be adopted by all credit cards companies like VISA or MasterCard. On top of that, audit is also reviewed on a yearly basis to ensure that all regulations are met (compliances). **2. Medical Industries** Personal medical data that is used for predicting infections or disease must abide confidentiality and security. This is to protect the patient's medical records. **3. Internet Advertising** Internet companies must comply to rules on what data can be stored implicitly and explicitly about a user. Implicit data is data gathered unintentionally like Facebook likes, status update, user purchases where the users are unaware of sharing data. While explicit data is data gathered intentionally like surveys, membership application, where the users aware of sharing data.