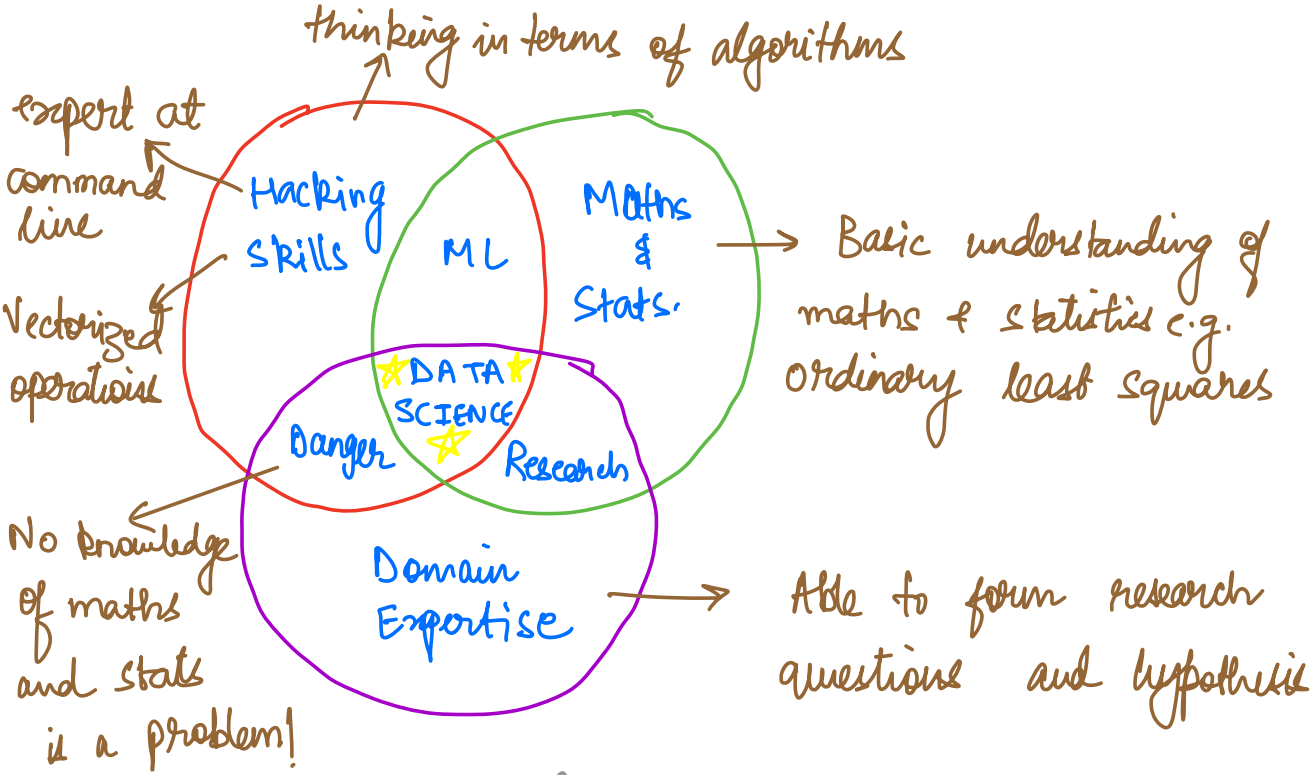
Data Science Venn Diagram



Drew Conway Data Consulting, 2015

DATA

- User clicks on websites
- location info. from smartphones
- Smart watches, smart cars, smart homes
- movies, music, sports...

SCIENCE

- Discovery
- Building knowledge
- Extracting information from data.

Data Scientists

Someone who extracts insights from data.

What is the recipe to do DS?

1. What is the problem you are trying to solve?
   * How to understand the sentiments and opinions of people on social media?
   * How to recommend music to listeners?
   * Predict future stock prices
   * Determining emotions from audio data (speech emotion recognition)
   * Solve climate change problem using physical datasets.
   .
   .
   .

2. Best ways to clean your data.

* Data Discovery and assimilation
* How to account for missing data.
* spelling mistakes
* Different schema in different datasets

3. Exploring cleaned data.

[Exploratory Data Analysis]

* Data types
* Categorical and Numeric Data.
* Visualization of cleaned data.
* statistical treatment of data.
* Check whether different datasets are correlated

4. Preprocessing of data.

* Conversion of categorical data to numeric data.
* Conversion of data to a specific format that a particular machine learning method understands e.g. One-Hot Encoding.
* Feature Selection:

Row↓	Name	Gender	Age	City	How Fast You clap
1					
2					
3					
4					

20

Implies that not all features are important for us to solve problem. We can select the most important features. Domain Expertise, PCA, ...

5. Model Selection and Model Development

Model Selection

* Which model is best suited to your data science problem?
  - is it prediction
  - is it classification
  - do you have small or large data
  - what computing resource you have

Model Development

* Once you select a model(s), it is necessary to make the best version of the selected model
* Take care of bias & variance, overfitting & underfitting
* Hyper-parameter tuning

6. Evaluation of the developed model

* Report certain metrics that evaluates the performance of the developed model

Choosing the correct metric for communication of evaluation of the developed model is critical

Metrics :-
- Confusion Matrix
- F1 Score
- Receiver Operating Curves (ROC)
- Area under Curve (AUC)
- Intersection Over Union (IOU)

Let us workout a data science problem!

① Problem:

How to predict weather in agriculture sector

② Data Cleaning:

Data Discovery

* Cloud Coverage
* Precipitation
* Temperature profile
* Humidity
* Wind direction and speed

Data Assimilation

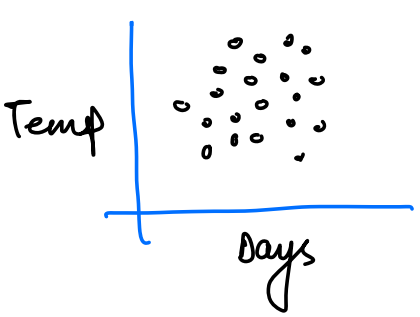
Days	wind speed (m/s)	Ppt (mm)	Temp (°C)	Humidity	Cloud
1	10 %	20	20°C	60 %	Yes
2	12	100	21°C	80 %	Yes
3	20 %	80		40 %	NO
4	18	10	22°C	40 %	NO
5	18 %	2	21°C	45 %	Yes
6	:	:	:	:	:
:					

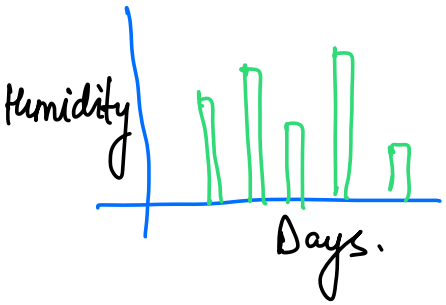
Data Cleaning

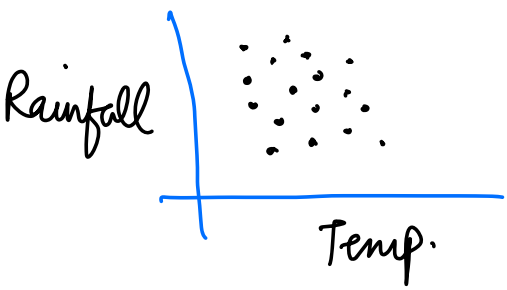
Days	wind speed (m/s)	PPt (mm)	Temp (°C)	Humidity	Cloud
1	10 %	20	20°C	60 %	Yes
2	12	100	21°C	60 %	Yes
3	20 %	80		40 %	NO
4	18	10	22°C	40 %	NO
5	18 %	2	21°C	45 %	Yes
6	⋮	⋮	⋮	⋮	⋮
⋮		⋮	⋮	⋮	⋮

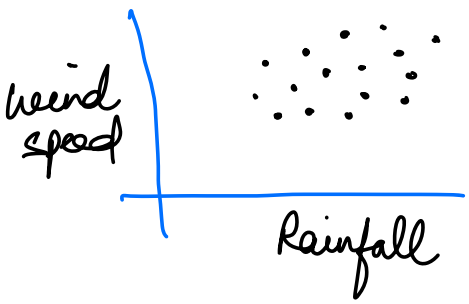
* % sign in second column
* °C in temperature column
* % sign in humidity column.
* Missing data in temperature column.

③ Exploratory Data Analysis









→ mean, standard deviation, inter-quartile range, for different features

④ Preprocessing of Data

* Conversion of categorial to numeric data

Cloud
Yes
Yes
NO
NO
Yes

⇒

Cloud_Yes	Cloud_No.
1	0
1	0
0	1
0	1
1	0

→ Also depending on domain expertise, some features may be ignored. Or use more sophisticated technique of principal component analysis.

Days	wind speed (m/s)	Ppt (mm)	Temp (°C)	Humidity	Cloud
1	10 %	20	20°C	60 %	Yes
2	12	100	21°C	60 %	Yes
3	20 %	80		40 %	No
4	18	10	22°C	40 %	No
5	18 %	2	21°C	45 %	Yes
6	⋮	⋮	⋮	⋮	⋮
⋮					

(5) (5.1) Model selection

Is it a prediction or classification problem?

Prediction: Prediction can happen if there is some fundamental equation governing the weather

e.g. weather_severity = $\beta_0 + \beta_1 \times$ wind speed $+ \beta_2 \times$ rainfall + ...

Classification : Severity of weather e.g.
Most Severe, Severe, less severe, normal, ...
On a scale of 10, 1 being the least severe &
10 being the most severe.

It can be considered a classification problem.

What data are we missing?

Days	wind speed (m/s)	Ppt (mm)	Temp (°C)	Humidity	Cloud	Weather Severity
1	10 %	20	20°C	60 %	Yes	High
2	12	100	21°C	60 %	Yes	High
3	20 %	80		40 %	NO	Medium
4	18	10	22°C	40 %	NO	low
5	18 %	2	21°C	45 %	Yes	low
6	:	:	:	:	:	:
:						

Weather_Severity is the output data.
We need output data to train our machine learning models.

Is our dataset small or big?

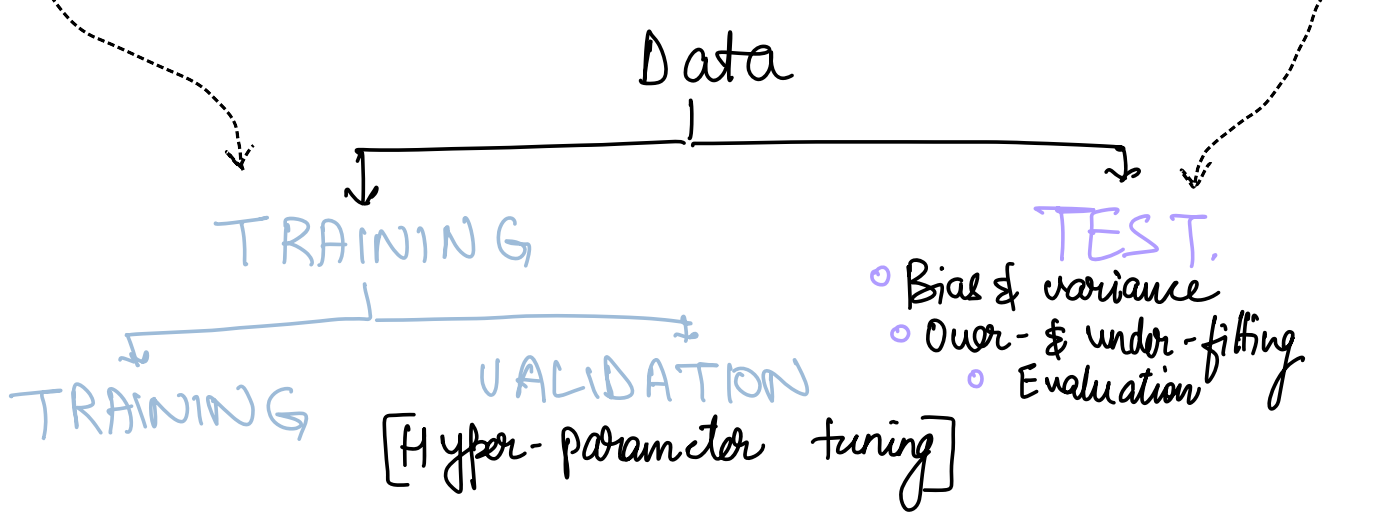
* How many data points we have?
  hundreds, thousands, million ...
* What computing resource we have?
  laptop, workstation, high-performance cluster, (HPC)
  cloud computing.

Depending on your answers, we can choose either less parameter classification models like logistic regression, support vector machines, random forest or high parameter classification models like artificial neural networks.

5.2 Model Development

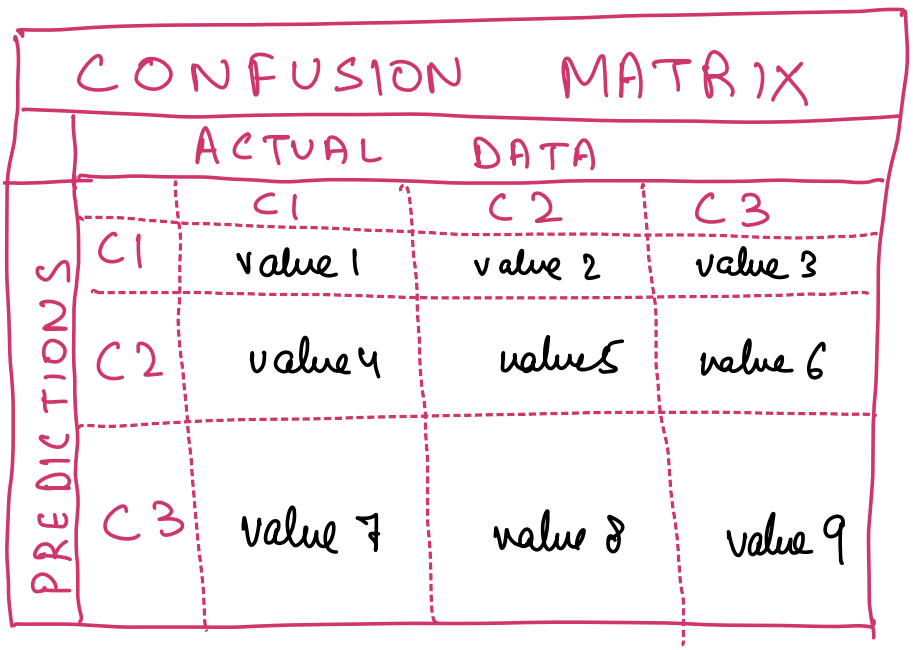
* Choose hyper-parameters using hyper-parameter tuning.
  e.g. learning rate, loss function, number of layers in ANN, ...

Days	wind speed (m/s)	PPt (mm)	Temp (°C)	Humidity	Cloud	Weather Severity
1	10 %	20	20°C	60 %	Yes	High
2	12	100	21°C	50 %	Yes	High
3	20 %	80		40 %	NO	Medium
4	18	10	22°C	40 %	NO	low
5	18 %	2	21°C	45 %	Yes	low
6	:	:	:	:	:	:
:						



6. Evaluation of the developed model

Evaluation is performed on the test dataset.



CONFUSION MATRIX

PREDICTIONS \ ACTUAL DATA	C1	C2	C3
C1	value 1	value 2	value 3
C2	value 4	value 5	value 6
C3	value 7	value 8	value 9

C1, C2, C3 → output classes

e.g. C1 High weather severity

C2 Medium " "

C3 low " "

a) what is value 1?

b) what Value 1, 5, 9 indicate?

c) what is value 4?

d) what do row sum & column sum mean?

Based on confusion matrix, you can derive other metrics.

* Precision
* Recall
* F1-score.
* Receiver Operating Curve
* Area under the Curve

.
.
.
.
.
.

* AI is concerned with understanding intellectual beings.

* AI is also concerned with building intellectual entities

* AI is relevant to any intellectual task

* To understand What is AI?, we can approach it in different ways.

(a) Is AI about acting humanly?

a.1 Natural language processing to communicate in a human language

a.2 Knowledge Representation to store what it knows

a.3 automated Reasoning to answer questions and to draw conclusions

a.4 Machine learning to adapt to new circumstances

and to detect and extrapolate patterns

a.5 Computer Vision and Speech recognition to perceive the world

a.6 Robotics to manipulate objects and move about.
move about.

(b) Is AI about thinking humanly?

b.1 Introspection – trying to catch our own thoughts as they go by

b.2 Psychological Experiments – observing a person in action.

b.3 Brain Imaging – observing a brain in action.

The field of cognitive science that brings together computer models from AI and experimental techniques from psychology for understanding the mind.

(c) Is AI about thinking rationally?

Thinking rationally implies irrefutable reasoning process that deals with precise notation for statements about the objects in the real world and relations among them.

But, in real world, we have uncertain situations.

Certain	Uncertain
- Chess	- War
- Traffic Rules	- Politics
- Competitive exams	

Therefore, probability helps construct rigorous statements with uncertain information. It helps develop the comprehensive model of rational thought.

(d) Is AI about acting rationally?

A rational agent is one that acts so as to achieve the best outcome or best expected outcome.

This approach supersedes other approaches.

How?

This is about the study and construction of agents to complete the objective provided to the agent.

But, sometimes, we cannot specify the objective completely and correctly.

For example. win a chess game. What do you think how will a rational agent act in this situation?

If we cannot transfer our objectives perfectly to the machine, then we need a new formulation — one in which machine is pursuing our objectives, but is necessarily uncertain as to what they are. Then, machines will act cautiously, ask permission, learn more about our preferences through observation.

Disciplines that contributed to ideas, viewpoints, and techniques to AI include:

- Philosophy
  e.g. utilitarianism v/s deontological ethics
- Mathematics
  e.g. logic, probability, statistics, computable functions, tractable problems,
- Economics
  e.g. decision theory, game theory, multiagent systems, operations research, Markov decision processes

- Neuroscience

  e.g. neurons, brain-machine interfaces, singularity

- Psychology

  e.g. Behaviorism, Cognitive science,
  human-Computer Interaction, Intelligence Automation
  (i.e. computers should augment human abilities rather than automate away human tasks).

  - Computer Engineering

    e.g. Moore's law, Quantum computing

  - Control theory

    e.g. Homeostatic, cost function

  - Linguistics

    e.g. Natural Language Processing (or computational linguistics), knowledge representation

The History of Artificial Intelligence

- Warren McCulloch and Walter Pitts (1943) introduced the concept of artificial neuron.
- Donald Hebb (1949) introduced Hebbian Learning to strengthen connection strengths between neurons.
- Marvin Minsky and Dean Edmonds built the first neural network computer in 1950.
- Allen Newell and Herbert Simon from Carnegie Mellon University made a mathematical theorem solving system known as the Logic Theorist.
  They also made the General Problem Solver to imitate human problem solving protocols.
- In 1958, John McCarthy defined the high level language of LISP, which was used for creating AI related programs.

- However, early systems failed on more difficult problems.
  (a) System were more focused on how humans perform a task to solve a problem rather than a careful analysis of the task.
  (b) Not scaling up to more difficult problems since only simple problems with fewer objects were solved.

- Next, the era of expert systems began. In expert systems, expertise is derived from large numbers of special-purpose rules.
  e.g. Dendral project for inferring molecular structure from the information provided by a mass spectrometer.
  Mycin project for diagnosing blood infections
  R1 for configuring orders for new computer systems

  However, it was difficult to build and maintain expert systems for complex domains, in part because the reasoning methods used by the systems broke down in the face of uncertainty and in part because the systems could not learn from experience.

Since expert systems were not giving desired results. researchers started re-using neural networks and began applying backpropagation learning algorithm in mid-1980s.

Researchers started developing benchmark datasets like MNIST, COCO, ImageNet and applied probabilistic reasoning & machine learning concepts to solve problems.

1988 was an important year to connect AI to other fields of research such as robotics, speech recognition computer vision, natural language processing with new developments in Bayesian Networks and reinforcement learning.

Since 2001 to present times, very large datasets are created including trillions of words of text, billions of images, billions of hours of speech and video, vehicle tracking data, clickstream data, social network data and so on. This is known as Big Data.

Majority of examples in such datasets are unlabeled. For example, does 'plant' refers to flora or factory. With large datasets, suitable algorithms can identify sense in a sentence. Similarly, large image datasets were created such as ImageNet to label objects and identify objects using computer vision.

Moving on, the term deep learning refers to machine learning using multiple layers of simple, adjustable computing elements. Deep learning found some

success with the introduction of LeNet-5 in 1995. Since 2011, deep learning has entered the mainstream of AI. For example, AlexNet won the 2012 ImageNet challenge to classify images.

Deep Learning relies heavily on hardware capable of doing $10^{14} - 10^{17}$ operations/sec. Standard CPU can do $10^9 - 10^{10}$ operations/sec.

https://aiindex.stanford.edu/