Machine Learning

What is Data Science

Mostafa S. Ibrahim
Teaching, Training and Coaching for more than a decade!

Artificial Intelligence & Computer Vision Researcher
PhD from Simon Fraser University - Canada
Bachelor / MSc from Cairo University - Egypt
Ex-(Software Engineer / ICPC World Finalist)

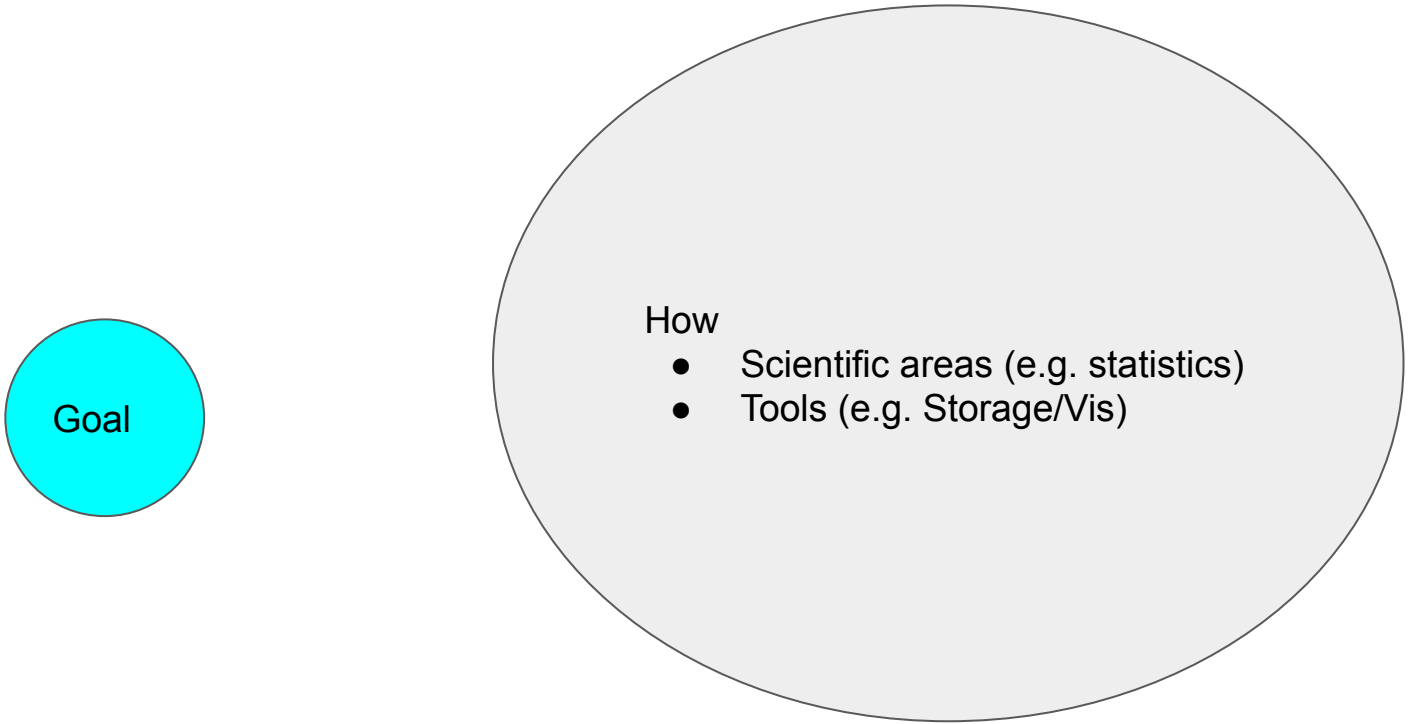
Data Science Field

- Motivation: How can we unlock real value from our (large) data?
  - Can we turn these data into product? Revenue? Business ideas?
- Historically ‘Data Science’ is an old term, but the modern proper definitions are around **finding insights** in the data
- A field of **deep** study of data that includes extracting useful **data insights** data, and **processing** that information using different tools, statistical models, and Machine learning algorithms
  - There is a wide diversity of definitions of this field in the last decade
  - There is a huge set of tools and topics to learn in this domain
    - You don’t have to learn all to find a job. You learn a lot during the journey


01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.
DATA SCIENCE LIFECYCLE
sudeep.co

Img src

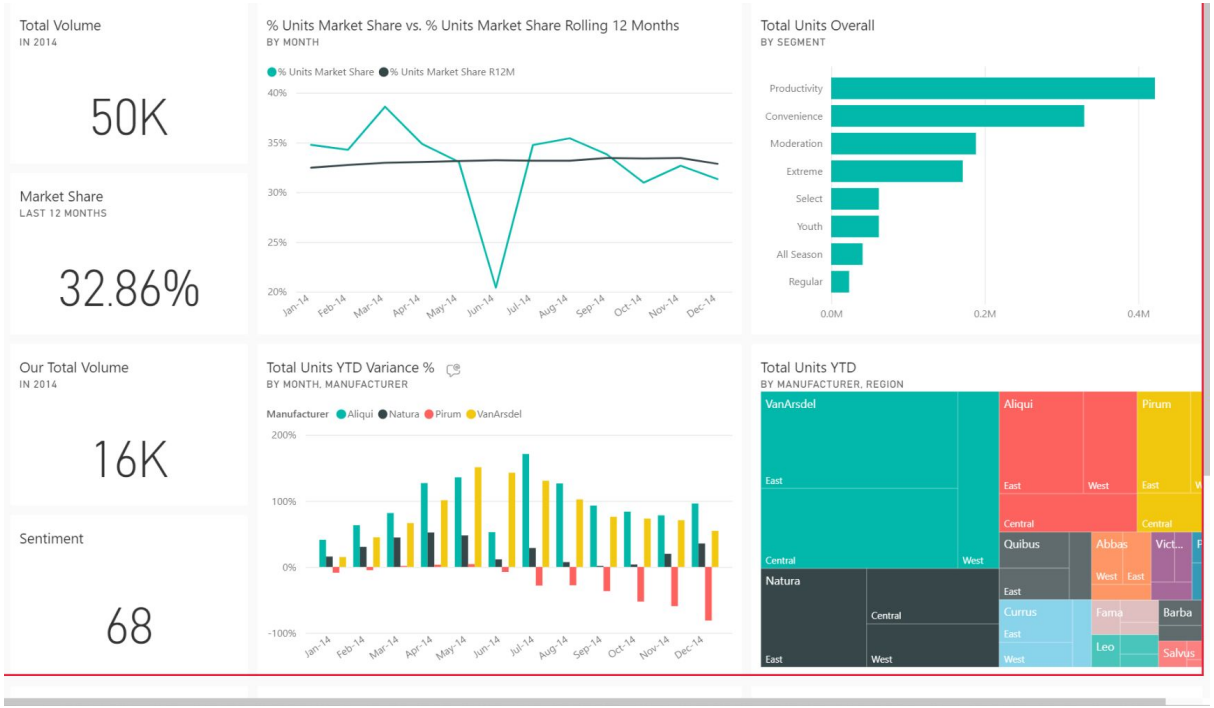
The WRONG balance!



The WRONG balance!

- **What** is the **goal** of Data Science? Finding insights
- **How** to achieve this goal? There are a lot of **areas** (programming, statistics, machine learning) and MANY **tools** (Apache Spark/Hadoop, ML frameworks, Visualization tools, *Tableau, Power bi, Excel*, SAS, Jupyter, Matplotlib, etc, SQL/NoSql)
- Many courses/roadmaps exhaust students in either **science** or **tools** and less stress on their goal (finding insights)
  - This requires **critical** thinking, **analytical** thinking and later **storytelling** skills
  - From a company to another, subset of tools are in-use
  - Data preparation takes the most of the time in the market, but we do that for the GOAL!
  - *You mainly need to apply a lot in challenging problems*
  - *You need to deeply understand the domain of the problem*

- **Visualization** is an important key
  - For your analysis, hypothesis and explorations
  - To present your insights
- Storytelling skills



Img [src]

Little Example

- Apple would like to launch a marketing campaign for their new sport device!
- They asked the data science team for **strategy** for that
  - They need to find insights, build conclusions and help sketching a strategy
- They got a lot of gathered data from **many sources** about customers
  - Then they performed cleaning, processing and analyzing for customers purchases
- After analysis and some **ML models**, they found that
  - **similar** products are used intensively by customer's **age 20 to 30**
  - countries with low-internet bandwidth rarely buy internet hungry devices
  - most-expensive cities tends to buy the most expensive product variants
  - an increase in temperature **column** correlates to # of purchase **column**

Data Science vs Machine Learning

- Data Science with 50-70% of his tasks might decide more insights can be driving by machine learning
- So, some ML models can be applied
  - Typically, extracting features and performing some prediction
- Typically does not require full awareness of many ML different models
- If the data is huge, deep learning can be advantage
- Overall all, DS people need fair classical ML skills

Data Science vs Data Analysis

- Data analysts works on data to also find insights. They includes visualization and database queries
- There are actually a lot of definitions out there making a lot of overlap
- Here is how I like to view it relevant to what might happen in the industry
- Data Science = Data Analysis + Machine Learning + Strong programming
    - That is, data science is a superset, as you can code more or apply machine learning

Data Science vs Data Engineers

- Data engineers **build** and test **scalable Big Data** systems that can be used by data scientists and ML guys to load and model data
- This requires **software engineering** skills to achieve different criteria like scalability, availability, etc

Applying for a Data Science Job

- Ask your manager explicitly what skill set you need for this job
  - You can be confused by the wide diversity of things that people do with data and call themselves ‘data scientist’!
- Tip: If a company doesn’t have DS role but have DA role, accept it
  - Then internally convince them importance of ML to unlock more insights!
- You don’t need to master all the tools before finding a job

"Acquire knowledge and impart it to the people."

"Seek knowledge from the Cradle to the Grave."