01 – Data Science and Management

- Data Science
- Data Management

Data Science and Management

Corso di Laurea Magistrale in Ingegneria Gestionale

Marco Mamei, Natalia Hadjidimitriou

{marco.mamei, selini}@unimore.it

What is Data Science?

"Data science is the application of **computational** and **statistical** techniques to address or gain insight into some problem in the **real world**"

[J. Zico Kolter, Carnegie Mellon University]

What is Data Science?

2024 – Marco Mamei, Full Professor (marco.mamei@unimore.it)

What is Data Science?

Image from popsci.com

What is Data Science?

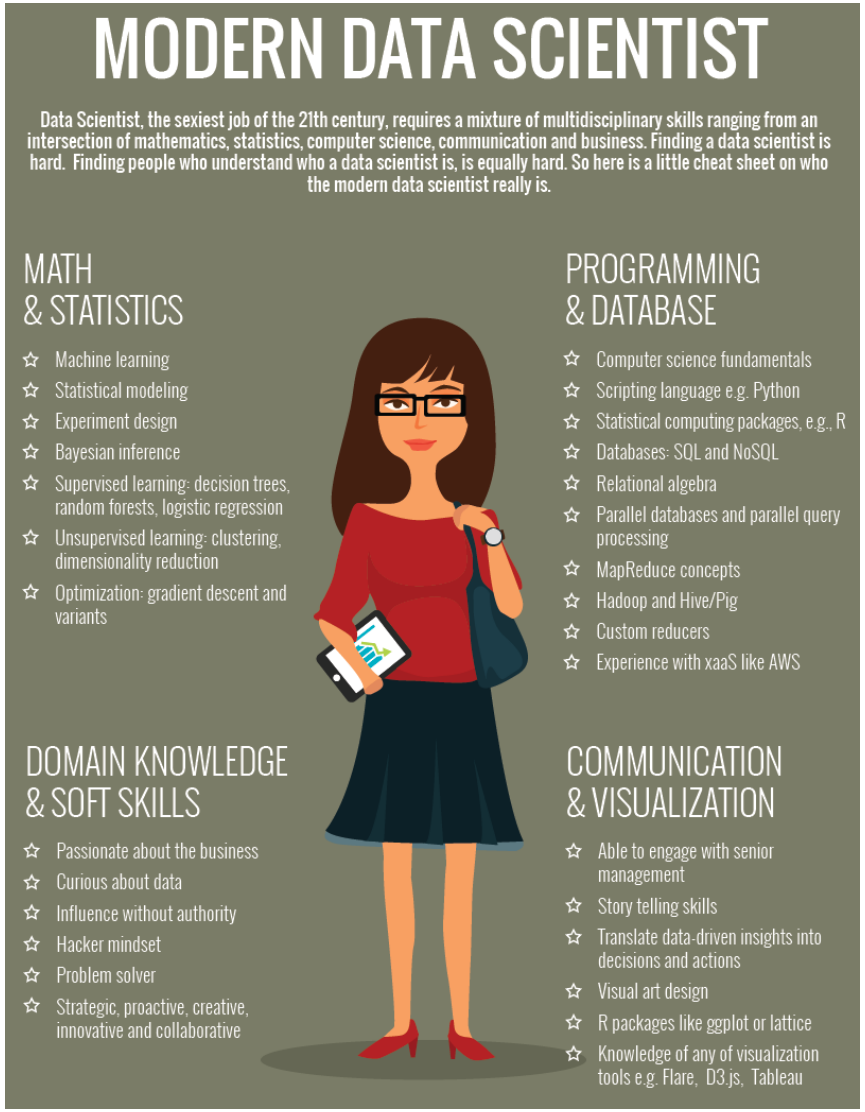


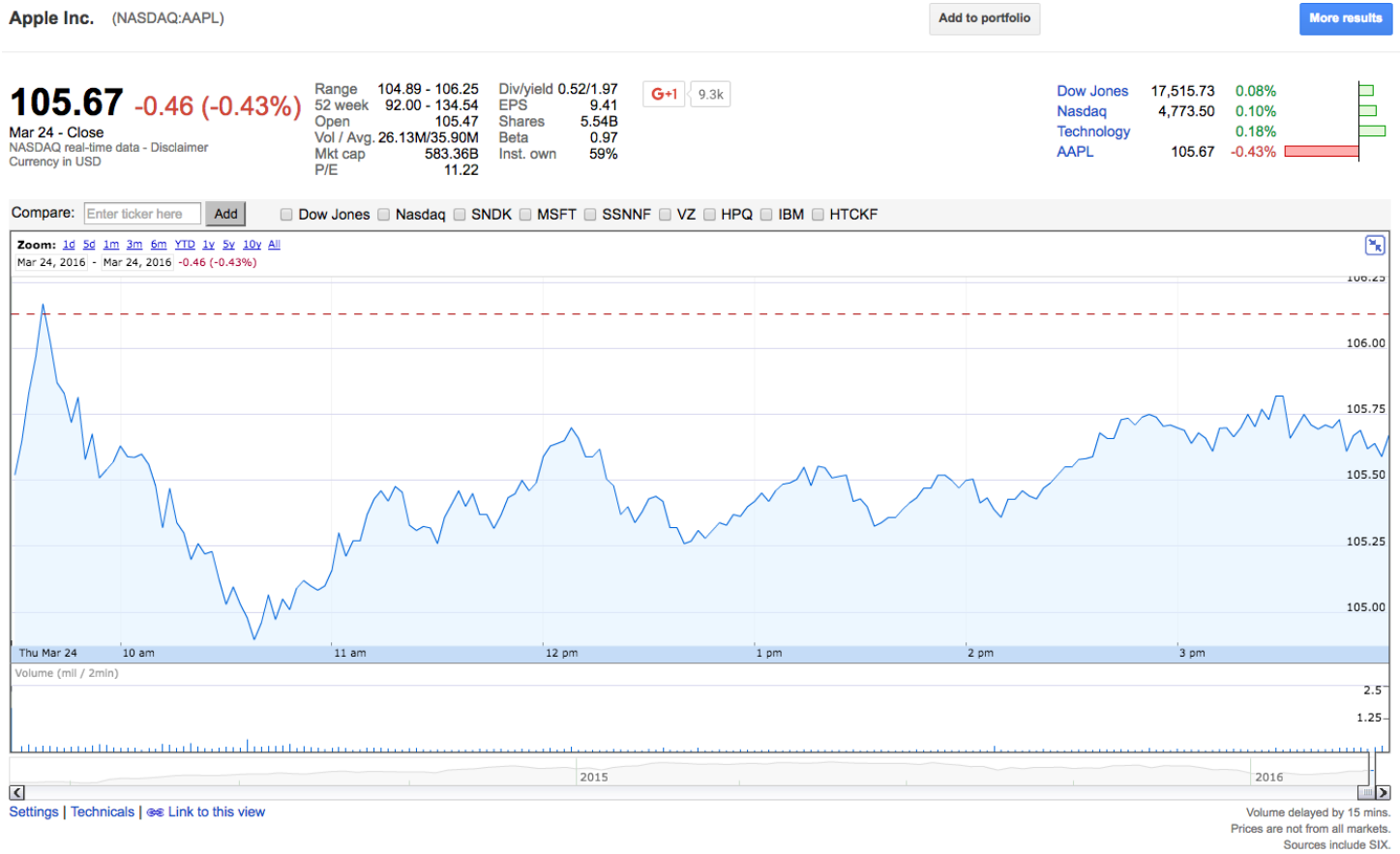
Image from marketingdistillery.com

What is Data Science?

Given **raw data** and a **problem statement**, choose a model to address the problem so that the **performance** on some evaluation **metric** is maximized.
Rigorous **testing** and **experimentation** has to be performed in order to **validate/refute hypotheses** and results.

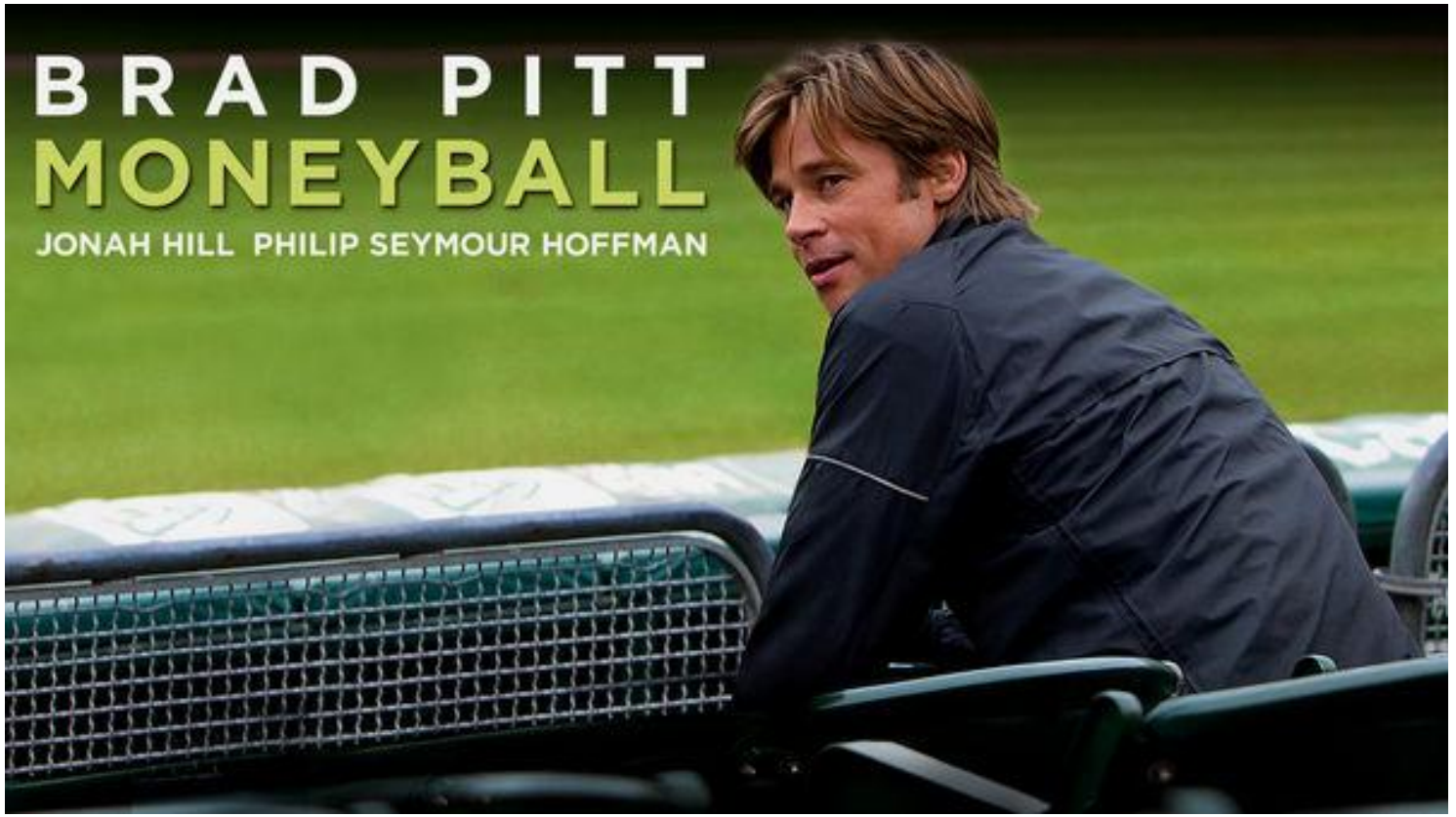
Data Science Applications

Will stock X raise in the next few hours?



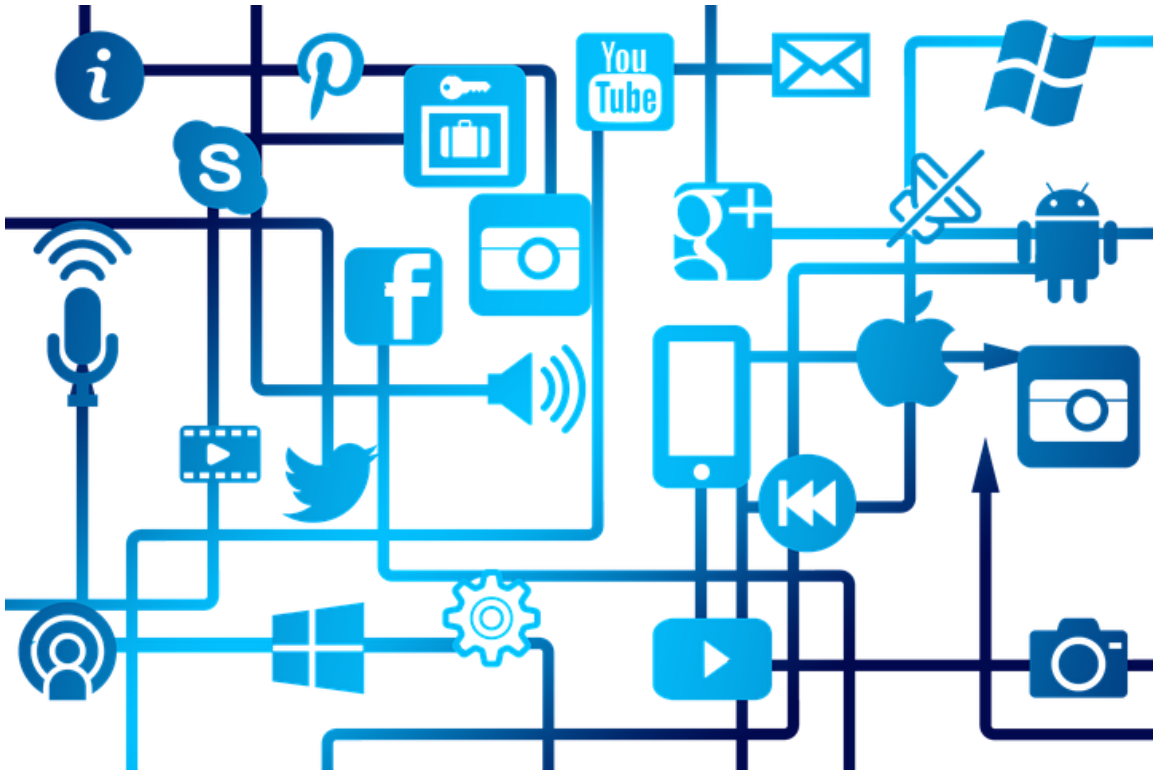
Data Science Applications

Should team T sign player X or Y?



Data Science Applications

Is customer X more similar to customer Y or customer Z?



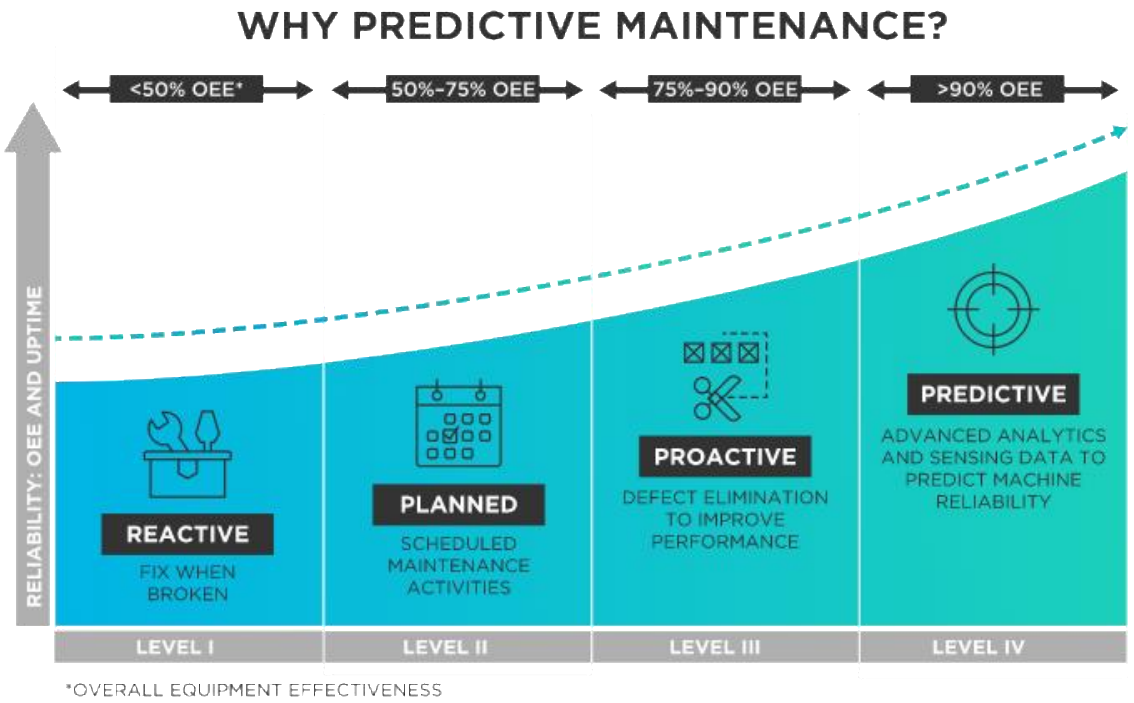
Data Science Applications

Should we grant a loan to X?

2024 – Marco Mamei, Full Professor (marco.mamei@unimore.it)

Data Science Applications

When/where should we perform equipment maintenance?



[Immagine: Tibco]

Data Science Applications

Lead time prediction for supply chain optimization

2024 – Marco Mamei, Full Professor (marco.mamei@unimore.it)

Data Science Applications

Has procedure X improved company Y's production?

2024 – Marco Mamei, Full Professor (marco.mamei@unimore.it)

Data Science Applications

Does drug X cure disease Y?

2024 – Marco Mamei, Full Professor (marco.mamei@unimore.it)

Data Science Applications

Does gene X cause pathology Y?

2024 – Marco Mamei, Full Professor (marco.mamei@unimore.it)

Caveat

Observation, reason and experiment make up what we call the scientific method.

[Richard Feynman]

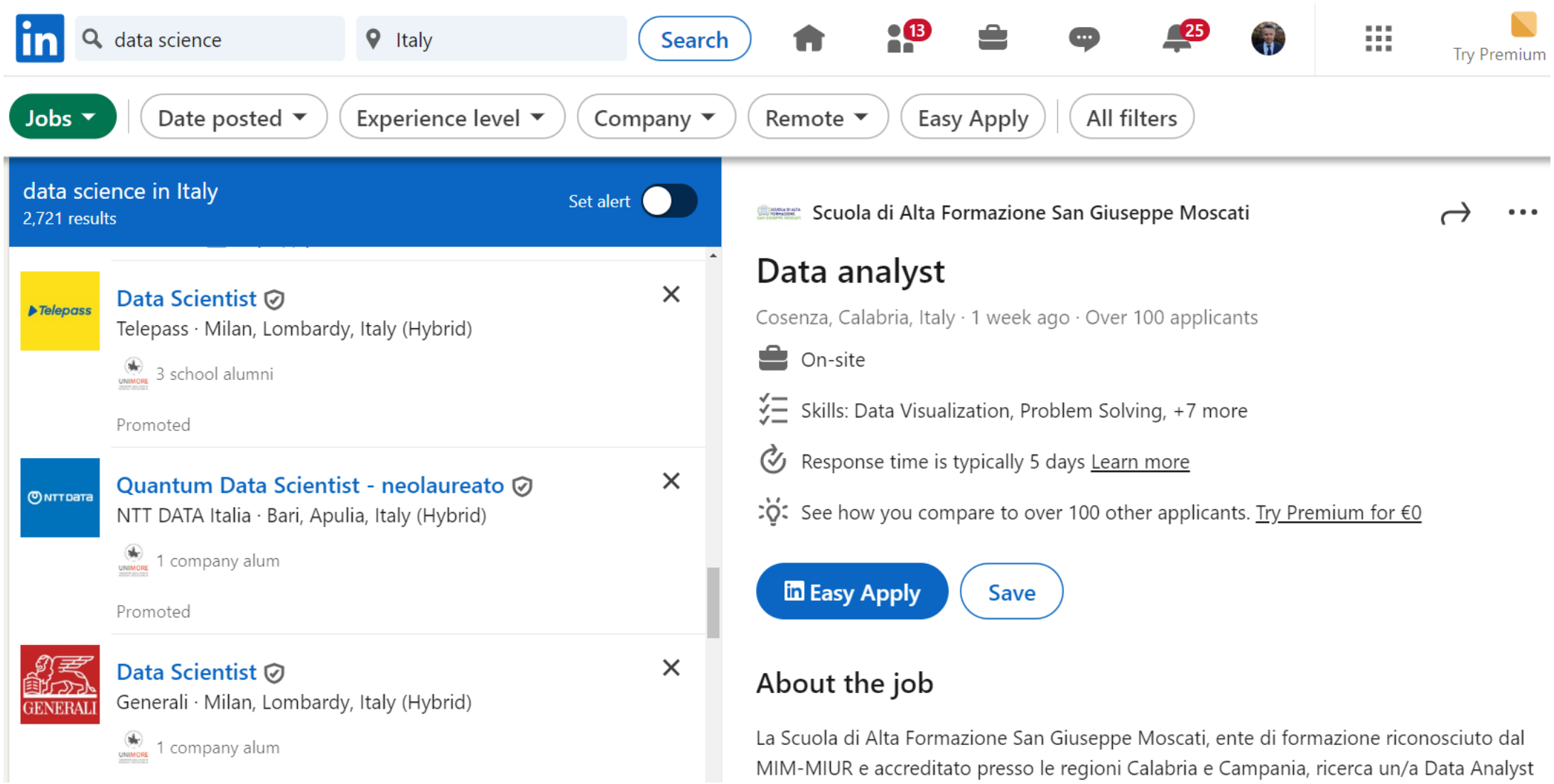
You do not get discoveries in the sciences by taking huge amounts of data, throwing them into a computer and doing statistical analysis of them… That's not the way you understand things… You have to have theoretical insights.

[Noam Chomsky, April 2014]

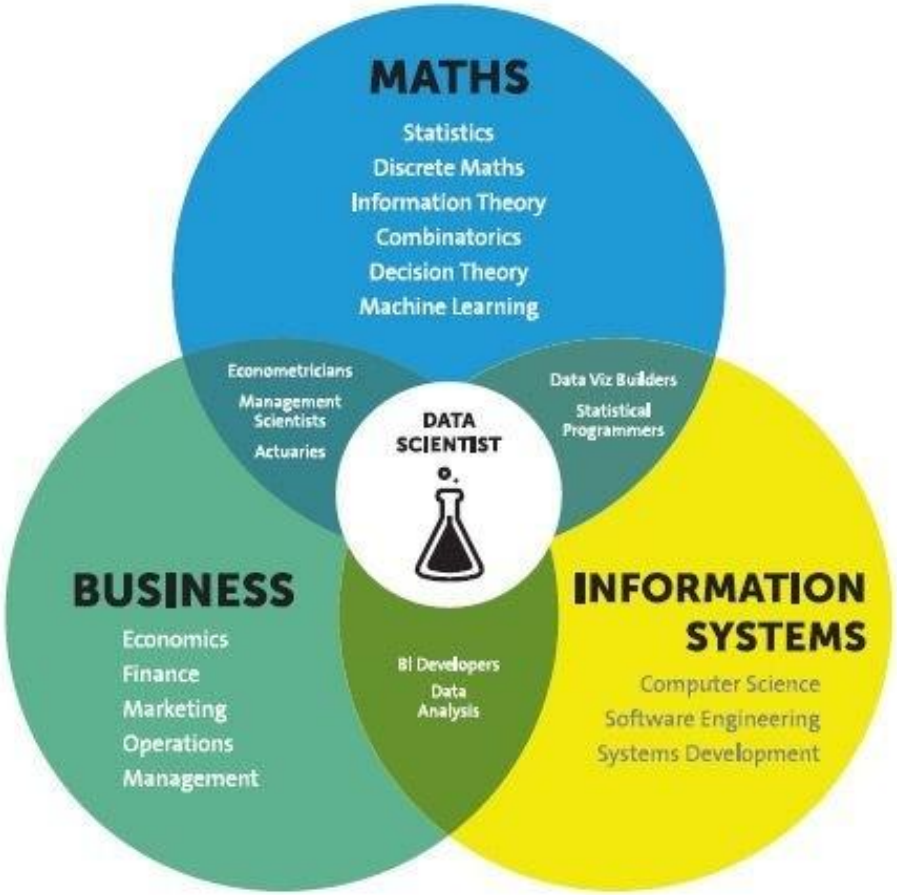
Data science requires critical thinking

Putting the science back in data science

Who Needs Data Science?



Related Fields



Business analytics requires a combination of expertise across business, information systems and mathematics.

Data Science Pipeline

Data science mainly needs three steps:

1. Data **collection** (pre-processing)
2. Data **analysis** (exploration, modeling, testing)
3. Data **presentation** (communication of results)

Data Collection

The first step in almost any data science application is to **collect some data**

Data could be available in several different ways:

- Plain text, Images, Audio,…
- Database (SQL, …)
- Web (HTTP requests, API, …)

2024 – Marco Mamei, Full Professor (marco.mamei@unimore.it)

Data Collection

Most common data formats

- Comma Separated Values (CSV)
- JavaScript Object Notation (JSON)
- HyperText Markup Language (HTML)
- eXtensive Markup Language (XML)

2024 – Marco Mamei, Full Professor (marco.mamei@unimore.it)

Data Collection

More on advanced data formats…

- How to deal with **high-dimensional** data?
- How to deal with **structured data** (e.g., graphs)?
- How to deal with **continuous data streams**?

Data Analysis

Exploration

- Take a first look at the data

Modeling

- Choose a model to be used

Testing

- Perform experiments to test model and hypothesis

2024 – Marco Mamei, Full Professor (marco.mamei@unimore.it)

Data Presentation

Present results of analysis in an **appropriate** way

Again, make use of data **visualization**

Give the **simplest** possible representation that conveys the message

“Maximize information, minimize ink”

2024 – Marco Mamei, Full Professor (marco.mamei@unimore.it)

Data Presentation

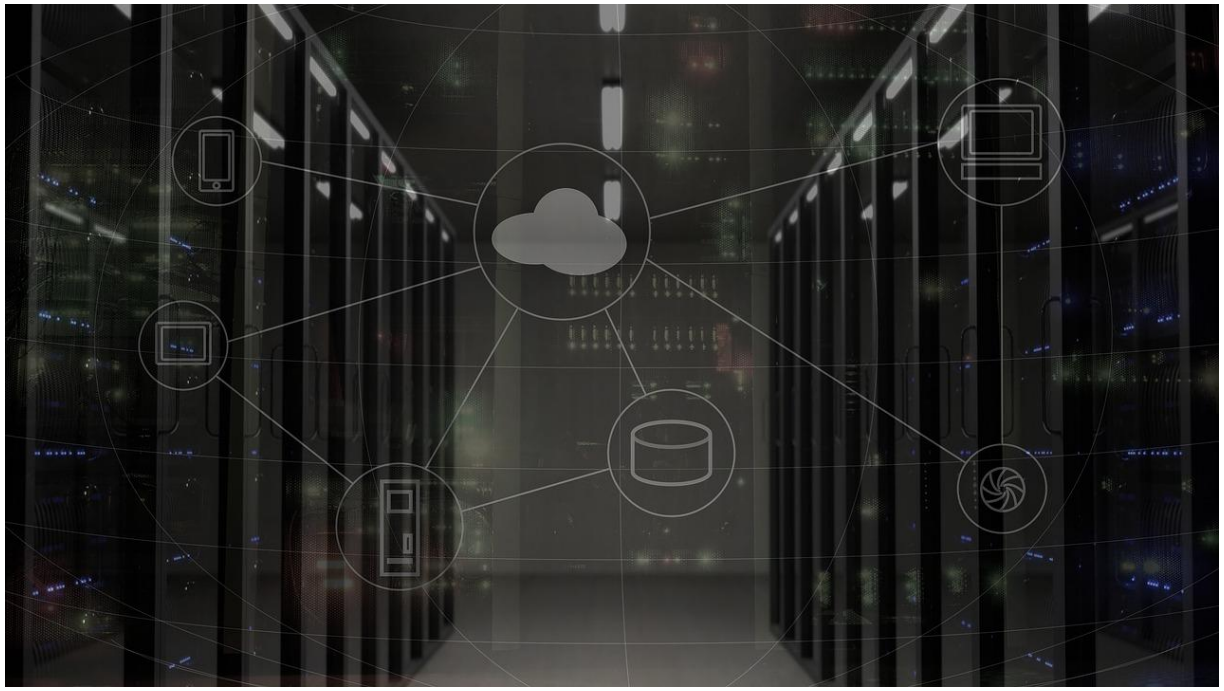
"The ability to **take data** – to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data."

[Hal Varian, Google's Chief Economist]

2024 – Marco Mamei, Full Professor (marco.mamei@unimore.it)

What is Data Management?

How to access/organize/store to all these data?



2024 – Marco Mamei, Full Professor (marco.mamei@unimore.it)

What is Data Management?

How to access to all these data?

- Efficient **storing** and **retrieval**
- Dedicated **hardware**
- Dedicated **software**

- Appropriate **infrastructures**
- **Cloud**

…It is the necessary support to data science!!!

Data Warehouse

An **information repository** that integrates and organizes data collected from **heterogeneous** sources and makes them **available** for…

- Analysis
- Evaluation
- Planning
- Decision-Making

From Data to Information

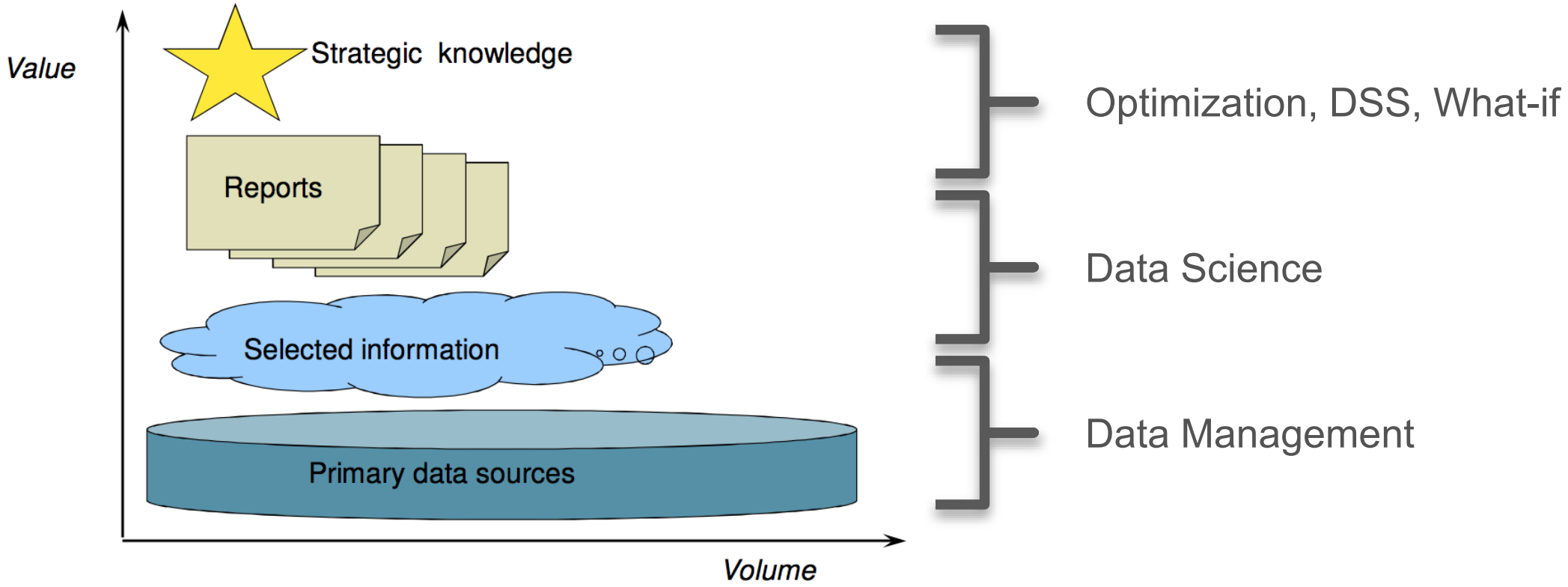


Figure by Prof. Stefano Rizzi @ UniBO

Big Data

Big Data is nowadays become a ubiquitous term…

It actually has to do with…

- Massive databases?
- Methods for the analysis of massive databases?
- Machine Learning and Data Mining?
- Scaling up algorithms?
- **All of the previous…?**

Business Intelligence

Support companies in **strategic decision making**

- Understand user **needs** and **goals**
- Understand **technological** needs to reach goals
- Transform **data** into **information**
- Transform **information** into **value**

2024 – Marco Mamei, Full Professor (marco.mamei@unimore.it)

Other Issues

- Privacy
- Ownership
- Reputation / Data quality
- Ethics
- Security and safety

2024 – Marco Mamei, Full Professor (marco.mamei@unimore.it)

Privacy

2024 – Marco Mamei, Full Professor (marco.mamei@unimore.it)

Ownership

HOME » TECHNOLOGY » SOCIAL MEDIA

Facebook terms and conditions: why you don't own your online life

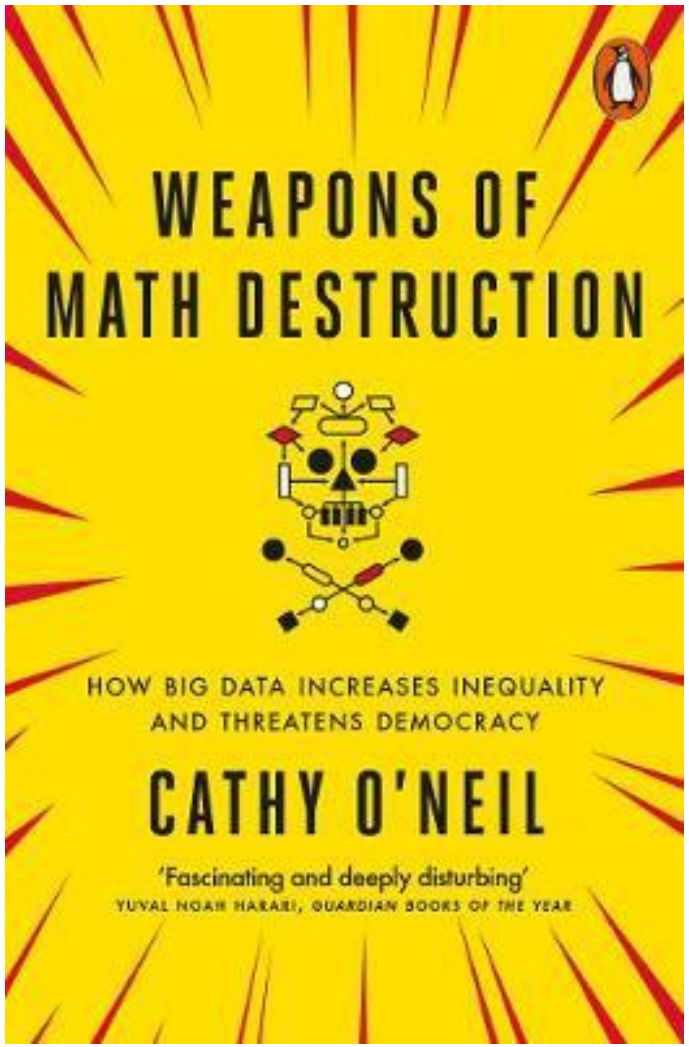
Did you read the terms when you joined Facebook, Twitter or LinkedIn? Oliver Smith explains how social networks effectively own your online content.

From telegraph.co.uk

Reputation / Data Quality

From blog.acuate.com

Ethics Issues



Safety and Security

Security issues

- Data protection against **unauthorized use**
- Control data **access**
- Control data **privileges**

Safety issues

- Protecting data against **loss**
- Regular **backups** and data **redundancy**

2024 – Marco Mamei, Full Professor (marco.mamei@unimore.it)

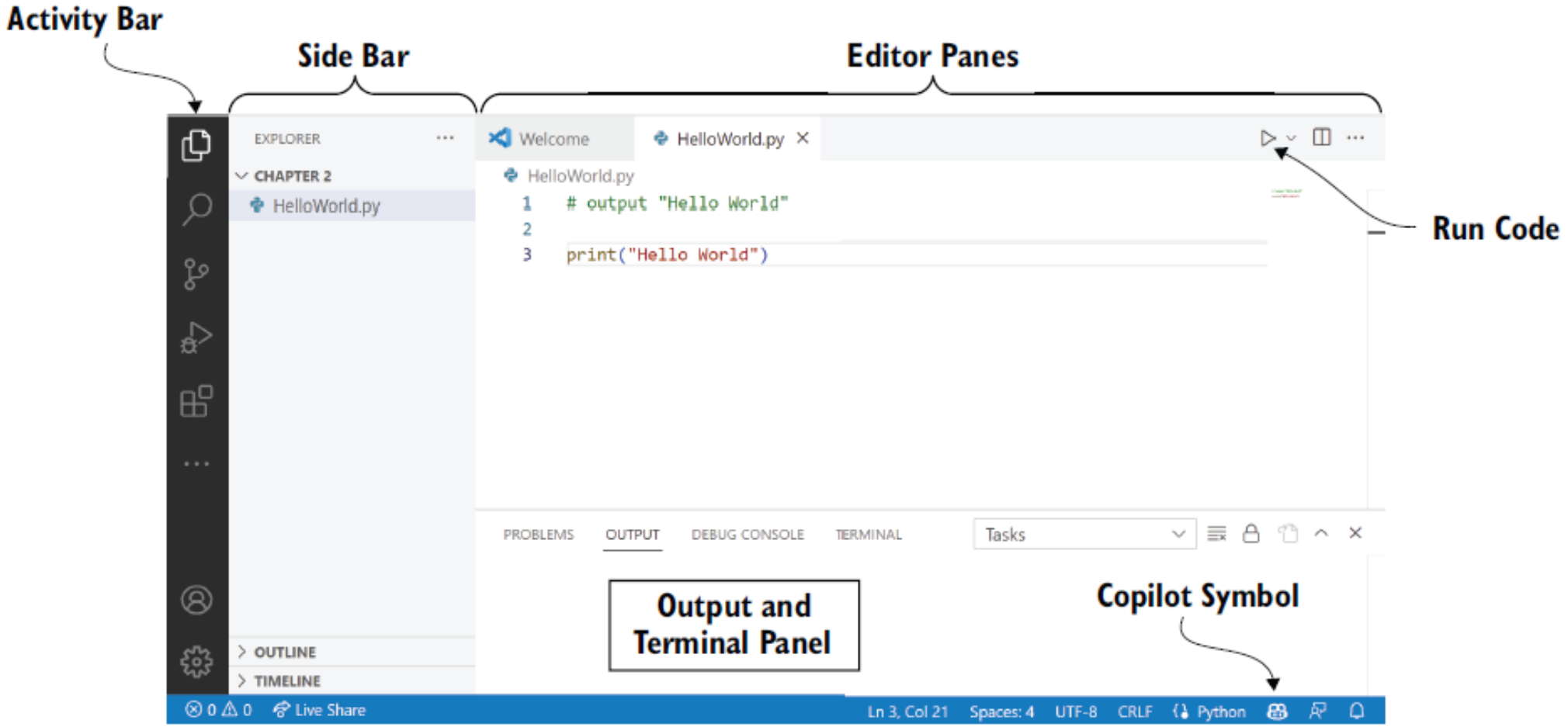
Exercise!

- **Superstore Sales**

- What is the total revenue generated by the store?
- Which category of products contributes the most to sales?
- How has the sales trend been for the past year?
- Which region has the highest sales and which one has the lowest?

Other Exercises: https://hackernoon.com/15-excel-datasets-for-data-analytics-beginners

For the next time…

Python3
IDE (VSCode)
Copilot

Script (.py)
Documenti interattivi (.ipynb)