UNIVERSITAT DE BARCELONA

Why Agile Data Science

Eloi Puertas Prats

Data

We are rendering into **data** many aspects of the world that have never been quantified before.

Digital Transformation in many aspects in our lifes and work places enhances this process of **datification**.

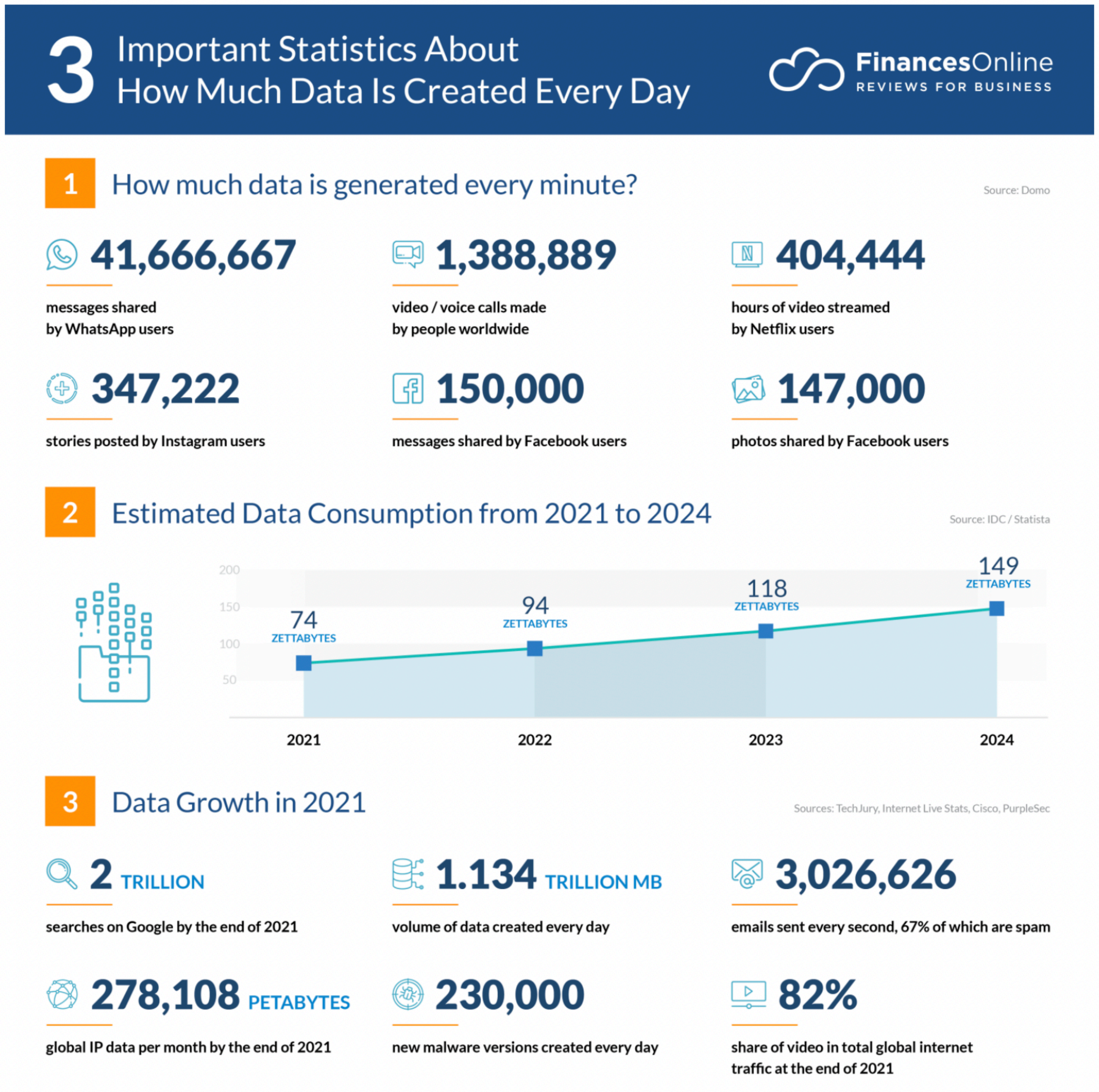
Data everywhere

Where **Information** comes from?

- Corporate Data Bases (structured information).
- Unstructured information in documents, Wikipedia, textbooks, journals, blogs, tweets, etc.
- Images in the web, public cameras, phones, TV, YouTube, etc.
- Public APIs: smart cities, government, search engines, etc.
- Sensor Data: GPS, accelerometer, physicochemical sensors, sociometric sensors, supercolliders, telescopes, etc.

How to handle such amount of data?

Big Data



3 Important Statistics About How Much Data Is Created Every Day

FinancesOnline
REVIEWS FOR BUSINESS

1 How much data is generated every minute?

Source: Domo

41,666,667
messages shared by WhatsApp users

1,388,889
video / voice calls made by people worldwide

404,444
hours of video streamed by Netflix users

347,222
stories posted by Instagram users

150,000
messages shared by Facebook users

147,000
photos shared by Facebook users

2 Estimated Data Consumption from 2021 to 2024

Source: IDC / Statista

3 Data Growth in 2021

Sources: TechJury, Internet Live Stats, Cisco, PurpleSec

2 TRILLION
searches on Google by the end of 2021

1.134 TRILLION MB
volume of data created every day

3,026,626
emails sent every second, 67% of which are spam

278,108 PETABYTES
global IP data per month by the end of 2021

230,000
new malware versions created every day

82%
share of video in total global internet traffic at the end of 2021

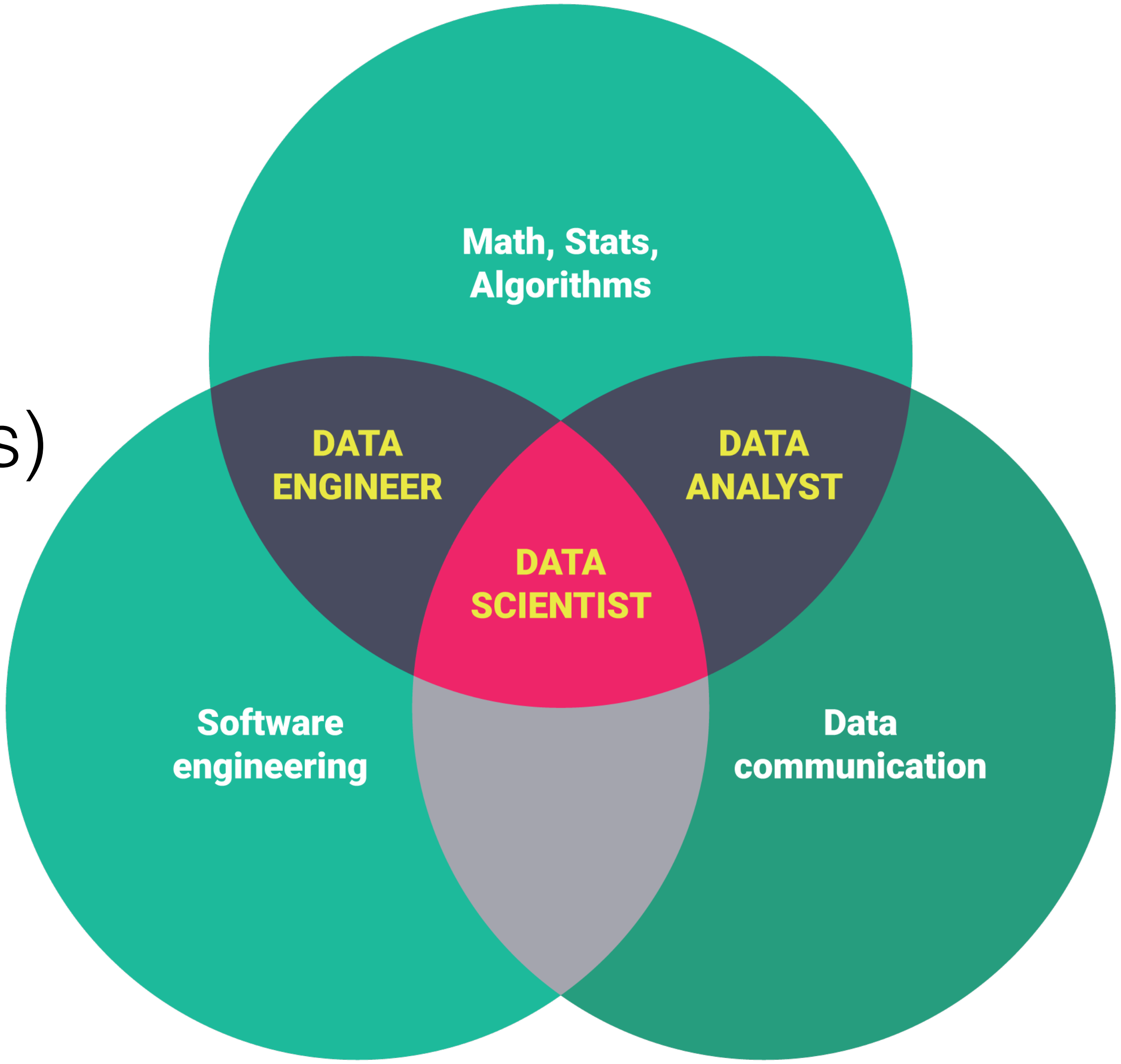
Data Science

Data Science is a **methodology** to define:

- what we want to do with data,
- how do we evaluate our actions,
- what decisions can be grounded on data,
- how do we combine evidences from several sources.

Data Science as a Team

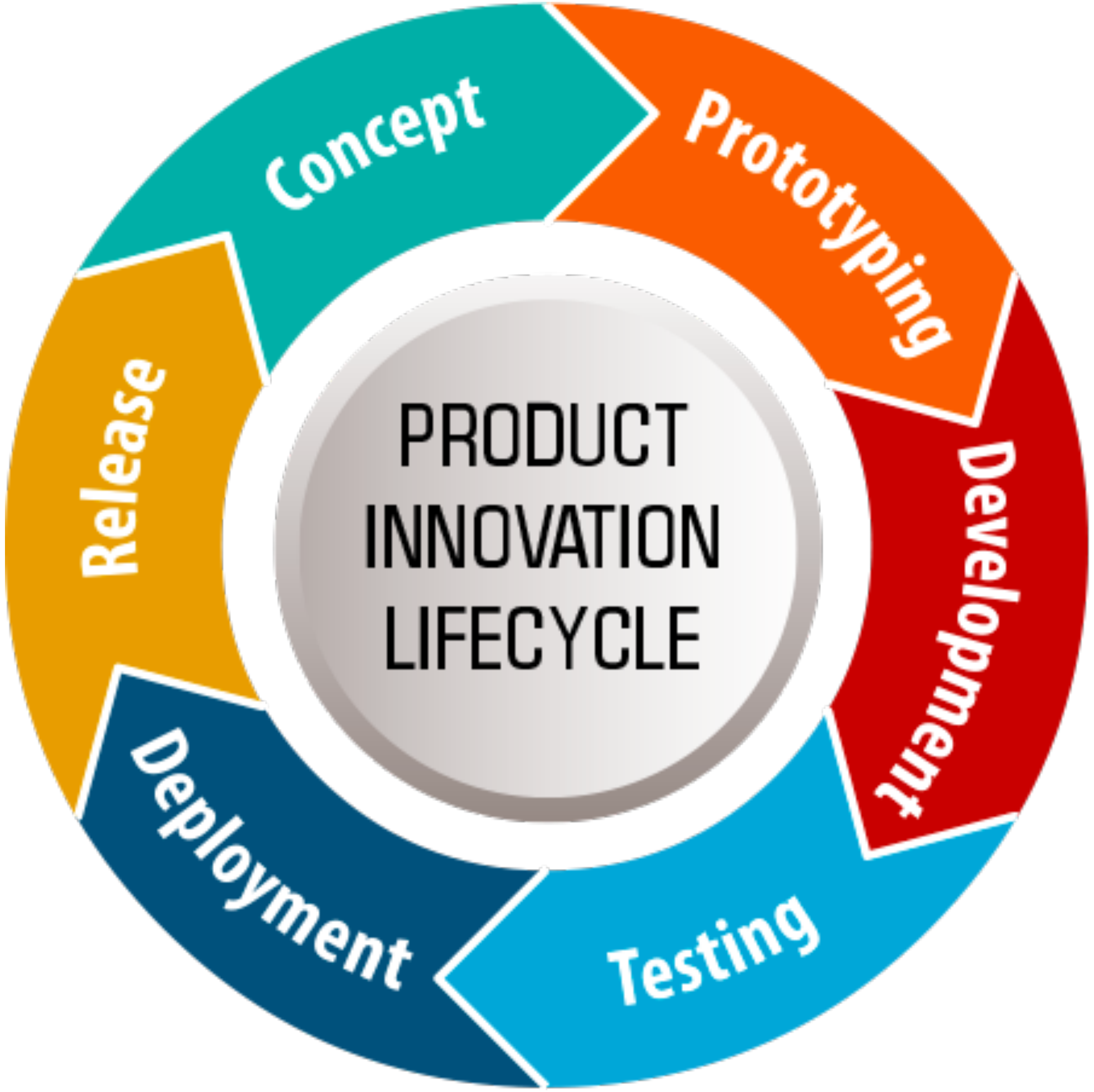
- **Data Analyst** (A.k.a Bussiness analysts)
- **Data Scientist** (A.k.a Statisticians, Data Managers)
- **Data Engineer** (A.k.a Data Managers, Database Administrators)
- **ML Engineer**
- **ML-ops**



https://www.springboard.com/blog/data-science-career-paths-different-roles-industry/

The Data Product Lifecycle

Build data products is not any more just to run a Notebook to train a model in python… there are much more steps.



Data Science Project Steps

Project Steps:

1. Ask a question.

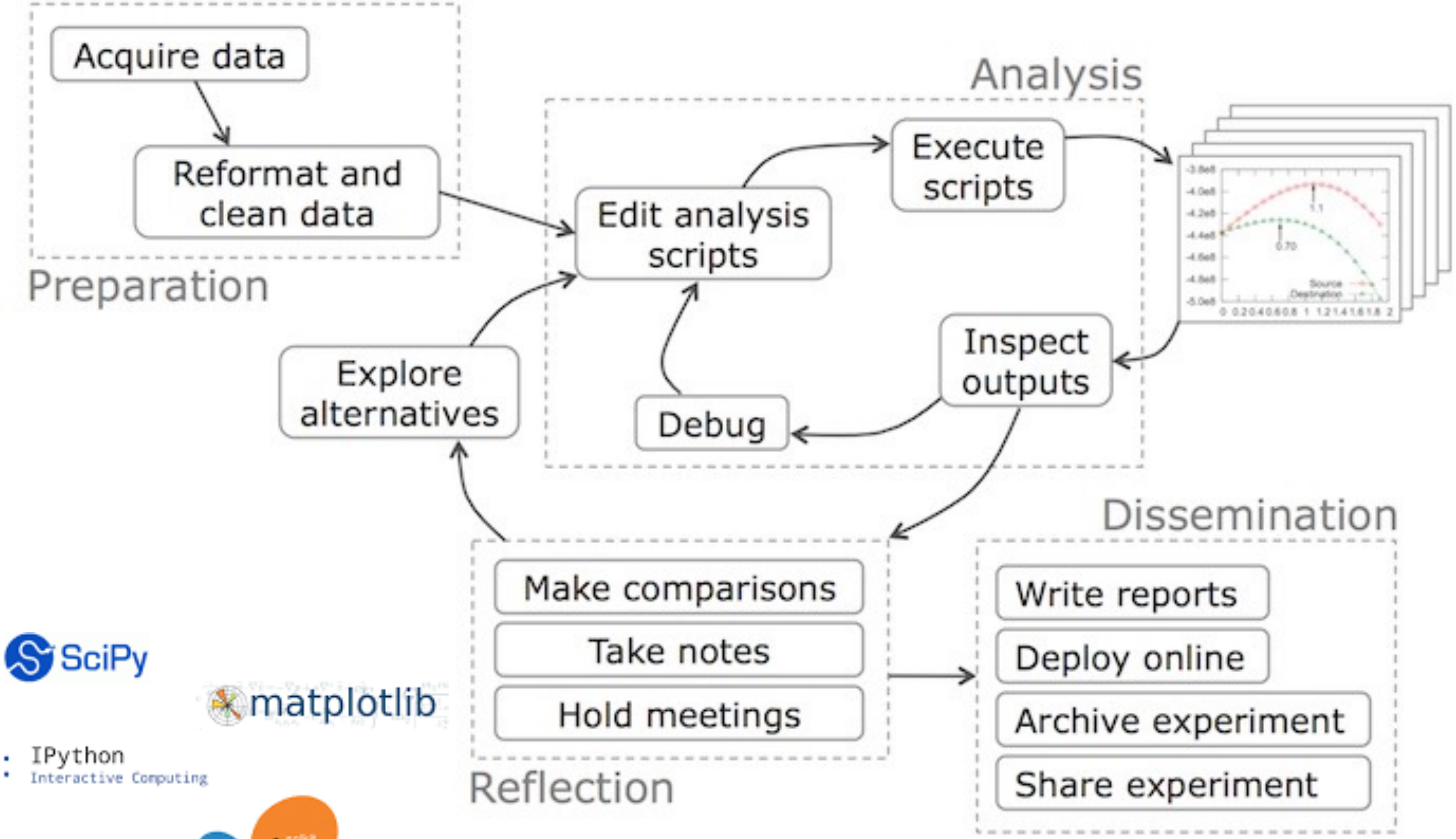
2.Get the data from a source. Data can be heterogeneous and non structured

3.Data Processing (cleaning, ETL.).

4.Data Analysis (machine learning, statistics…).

5.Take a decision and act.

Data Science workflow



Preparation

Acquire data

Reformat and clean data

Analysis

Edit analysis scripts

Execute scripts

Inspect outputs

Debug

Explore alternatives

Reflection

Make comparisons

Take notes

Hold meetings

Dissemination

Write reports

Deploy online

Archive experiment

Share experiment

NumPy

SciPy

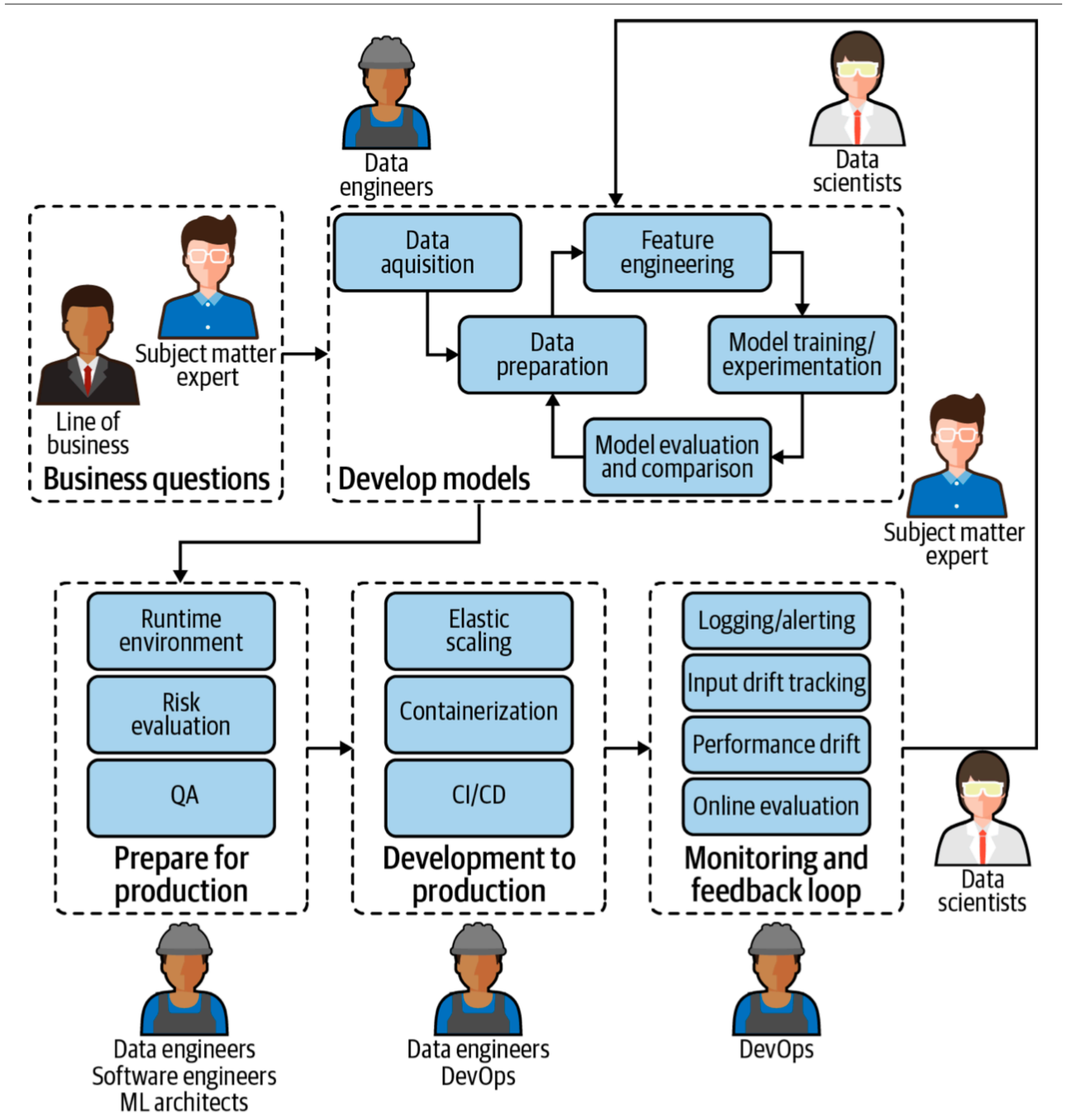
matplotlib

IP[y]: IPython Interactive Computing

pandas

scikit learn

Data Science pipeline (in real life)


Data engineers
Data scientists
Subject matter expert
Line of business
Business questions
Data aquisition
Feature engineering
Data preparation
Model training/ experimentation
Model evaluation and comparison
Develop models
Subject matter expert
Runtime environment
Risk evaluation
QA
Prepare for production
Elastic scaling
Containerization
CI/CD
Development to production
Logging/alerting
Input drift tracking
Performance drift
Online evaluation
Monitoring and feedback loop
Data scientists
Data engineers
Software engineers
ML architects
Data engineers
DevOps
DevOps