CMU Portugal Academy

CMU Portugal
Advanced Training Program

Artificial Intelligence, Machine Learning and Data Science Bootcamp

JOÃO MAGALHÃES

NOVA SCHOOL OF SCIENCE AND TECHNOLOGY

Carnegie Mellon Portugal · TÉCNICO LISBOA · Ciências ULisboa · NOVA School of Science & Technology

email: jmag@fct.unl.pt

João Magalhães – Short bio

jmag@fct.unl.pt

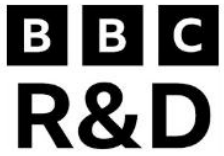
Full Professor, NOVA FCT
Co-Director of the CMU Portugal Partnership
Head of the Multimodal Systems Group, NOVA LINCS

My research is focused on text and image understanding AI algorithms and making information accessible through search and conversational systems.

I am keen to explore the limits of AI algorithms in real-world problems across different domains.

Throughout the years, my group has collaborated with world leading research institutions, e.g., Amazon, Google, BBC, Farfetch, VisionBox, CMU, Queen Mary.

Vision and Language AI



CMU Portugal Academy

Topics

1. PROGRAM STRUCTURE

2. WHAT IS AI + ML + DS?

3. BRAINSTORM: APPLICATIONS AND USE CASES

4. LABORATORY SETUP

CMU Portugal Academy

01

Program Structure

Carnegie Mellon Portugal · TÉCNICO LISBOA · Ciências ULisboa · NOVA SCHOOL OF SCIENCE & TECHNOLOGY

Learning outcomes: Knowledge

- Understand the nature and types of data problems.
- Understand the different challenges in an AI/ML/DS project.
- Understand the capacity and limits of the different family of algorithms.

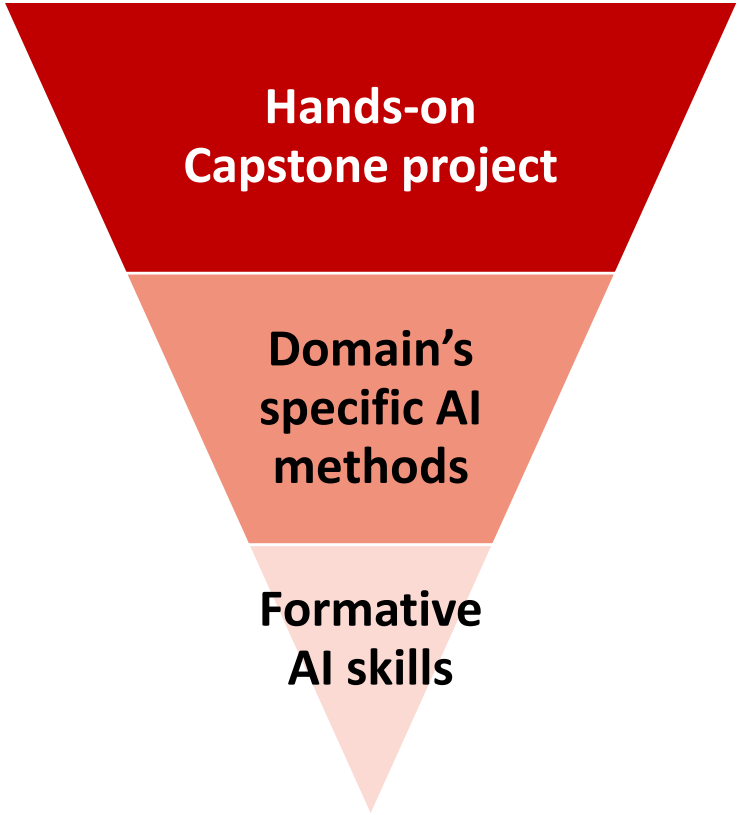
Learning outcomes: Know-how

- Identify the challenges of different data types
  - Numerical data / Dates / Categorical data / Text / Natural Language / Geographical data / Vision dat
- Design a data-driven end-to-end solution
- Integrate different algorithms
- Measure progress

Learning outcomes: Soft-skills

- Understand user needs
  - Set expectations
  - Identify required data
  - Recognize “impossible missions”
- Build a team based on the required skills
- Estimate a project’s implementation time, computing budgets and data requirements

Program structure



Program structure

- There are 3 core courses.
- These provide you with the basic concepts and tools.
- Each course will have an invited talk by a CMU faculty or industry expert.

Course	Lecturer	Teaching hours	ECTS
Foundations of Data Science	David Semedo	30	2
Machine Learning	Chryssa Zerva & David Semedo	30	2
Data Collection and Pre-Processing	Cátia Pesquita	30	2

Foundations of Data Science

- Introduction to Data Science
- Python Programming for Data Science
- Statistics and Probability
- Data Preparation and Processing with Pandas
- Machine Learning Fundamentals
- Model Evaluation and Selection
- Data Visualization

Prof. David Semedo

Machine Learning

- Introduction to Machine Learning
- Supervised Learning
- Unsupervised Learning
- Feature Engineering and Selection
- Model Evaluation and Validation
- Regression problems (linear regression)
- Support Vector Machines
- Decision Trees and Random Forests
- Association Rules
- Neural Networks

Prof. Chryssa Zerva

Prof. David Semedo

Data Collection and Pre-Processing

- Types of Data and Sources
- Data Collection Techniques
- Data Quality and Validation
- Data Pre-processing Techniques
- Handling Categorical and Numerical Data
- Data Integration and Fusion
- Data Sampling and Imputation
- Best Practices and Case Studies

Prof. Cátia Pesquita

Program structure

- The optional courses lets you specialize on:
  - Text and language data
  - Complex data
  - AI system engineering

- Capstone Project:
  - Lectured by all lecturers
  - Create and test a real system

Course	Lecturer	Teaching hours	ECTS
Multimodal Generative AI	João Magalhães	18	1
Vision and Language	Bruno Martins	18	1
Complex Data Analysis	André Falcão	18	1
Cloud-based Data Processing	Nuno Preguiça & Rodrigo Rodrigues	18	1
Information visualization	Sandra Gama	18	1
CapStone	Todos	48	3

Language and Vision

- Natural Language Processing (NLP) Fundamentals
- Text Classification and Sentiment Analysis
- Large Language Models
- Computer Vision (CV) fundamentals
- Image Classification and Object Detection
- Large Multimodal Models
- Image Captioning, Visual Question Answering and Multimodal Search
- Advanced Topics and Emerging Trends

Prof. Bruno Martins

Prof. João Magalhães

Multimodal Generative AI

- LLMs
- Prompt engineering
- RAG
- Image Captioning
- Visual Question Answering
- Image generation
- Advanced Topics and Emerging Trends

Prof. Bruno Martins

Prof. João Magalhães

Information Visualization

- Explore data types and appropriate visualization techniques.
- Interactive aspects and visual perception principles.
- Master visualization tools and techniques for storytelling.
- Handle data complexity and visualization uncertainty.
- Learn to evaluate the effectiveness and ethics of visualizations,
- User experience assessments and future trends.

Prof. Sandra Gama

Cloud-based Data Processing

- High performance data analytics.
- Large scale data analytics.
- Stream Processing and Real-time Analytics
- Cloud-based Data Storage.
- Machine Learning and AI on the Cloud
- Security, Scalability, Performance, and Cost Optimization

Prof. Rodrigo Rodrigues

Prof. Nuno Preguiça

Complex Data Analysis

- Graph data
- Social Network Analysis
- Time Series Analysis
- Geospatial Data Analysis

Prof. André Falcão

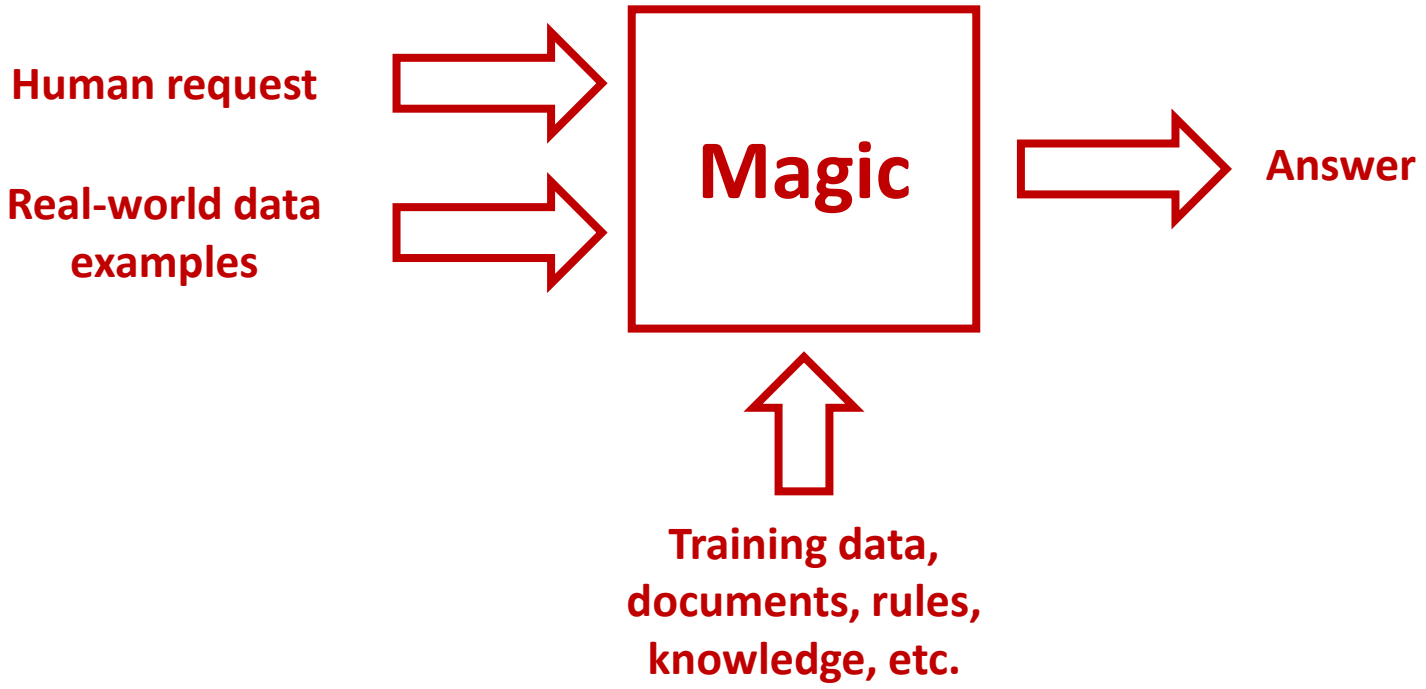
CMU Portugal Academy

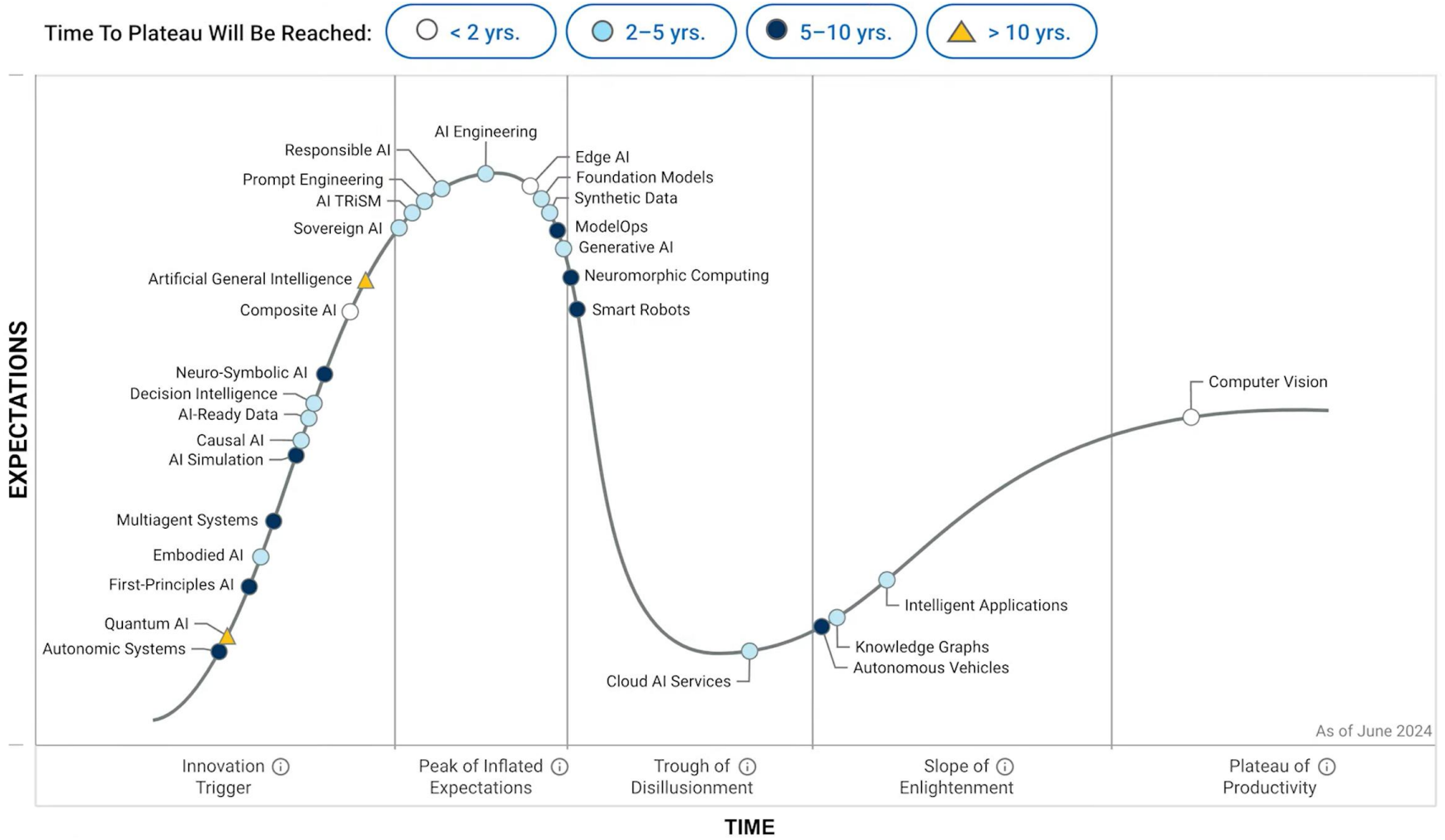
02

What is AI + ML + DS?

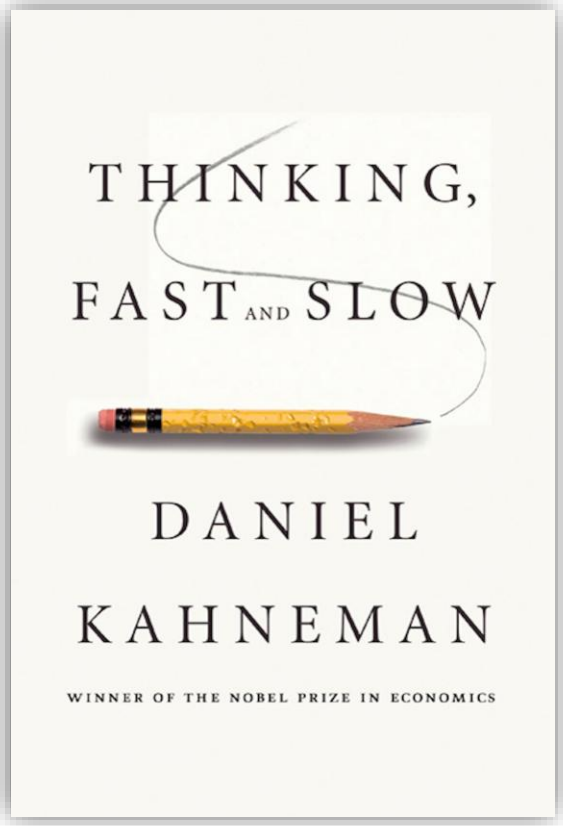
Carnegie Mellon Portugal · TÉCNICO LISBOA · Ciências ULisboa · NOVA SCHOOL OF SCIENCE & TECHNOLOGY

What is “the” algorithm?



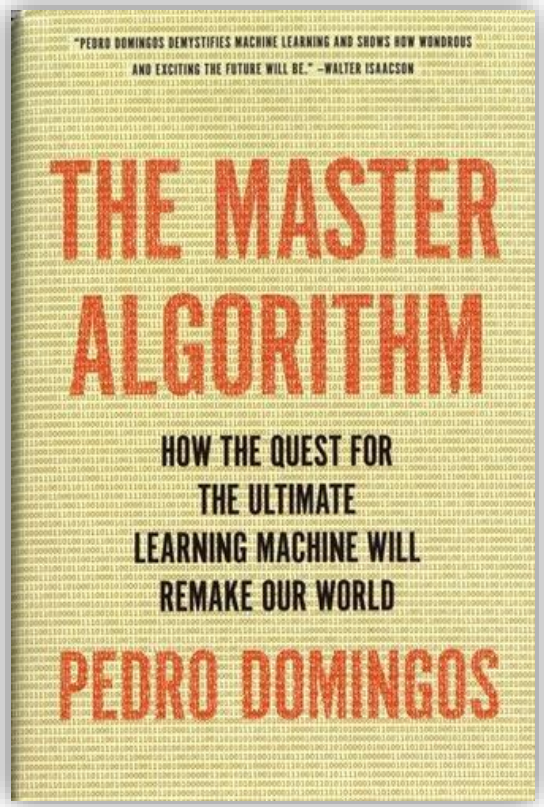

Time To Plateau Will Be Reached:
< 2 yrs.
2–5 yrs.
5–10 yrs.
> 10 yrs.
EXPECTATIONS
Responsible AI
Prompt Engineering
AI TRiSM
Sovereign AI
AI Engineering
Edge AI
Foundation Models
Synthetic Data
ModelOps
Generative AI
Neuromorphic Computing
Smart Robots
Artificial General Intelligence
Composite AI
Neuro-Symbolic AI
Decision Intelligence
AI-Ready Data
Causal AI
AI Simulation
Multiagent Systems
Embodied AI
First-Principles AI
Quantum AI
Autonomic Systems
Cloud AI Services
Intelligent Applications
Knowledge Graphs
Autonomous Vehicles
Computer Vision
As of June 2024
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
TIME

Readings



Thinking, Fast and Slow in AI

https://arxiv.org/pdf/2110.01834.pdf
https://arxiv.org/pdf/2010.06002.pdf



CMU Portugal Academy

03

AI Project Planning

Carnegie Mellon Portugal · TÉCNICO LISBOA · Ciências ULisboa · NOVA SCHOOL OF SCIENCE & TECHNOLOGY

What is data project planification?

- “Question first approach”:
  - A well-defined problem should be the priority.
  - Considerations about data, method, etc. come later.

- Focus on the key issues:
  - Look for key issues first and then expand.

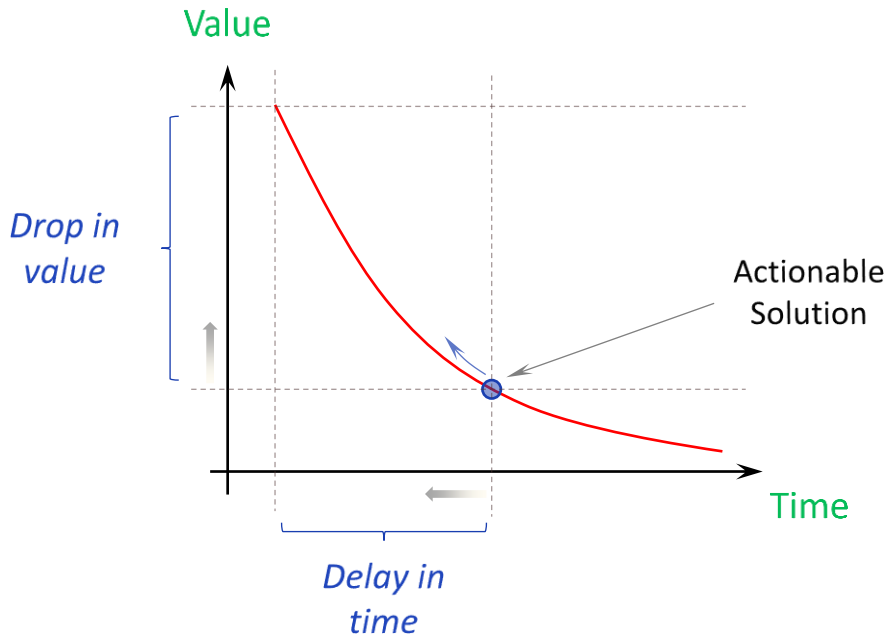
Four project types

What is the problem?

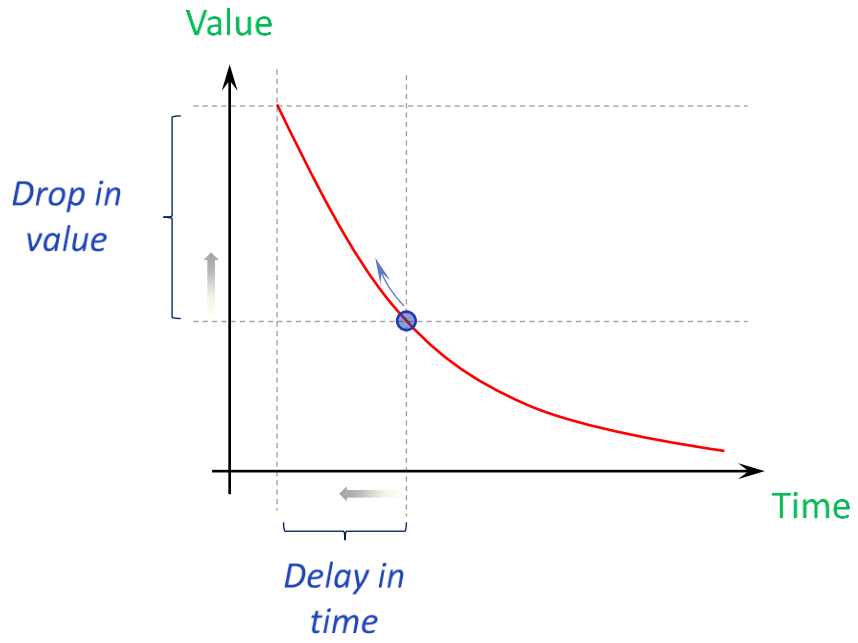
Which method to use?

	Known	UnKnown
Known	Optimization	Insight
UnKnown	Solution	Discovery

Value vs Time

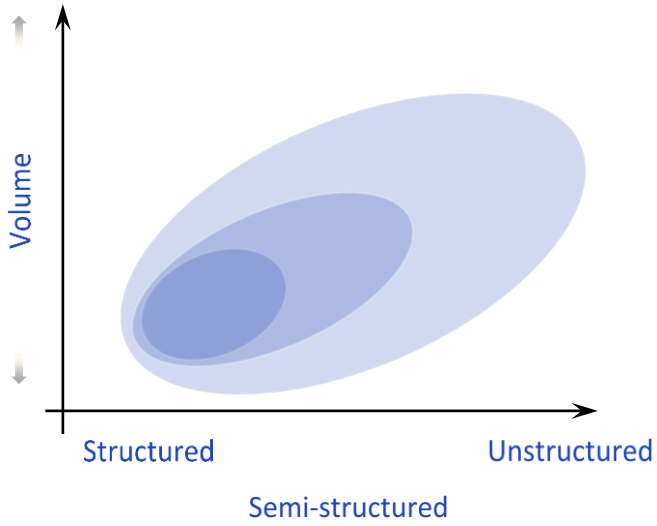


Value vs Time

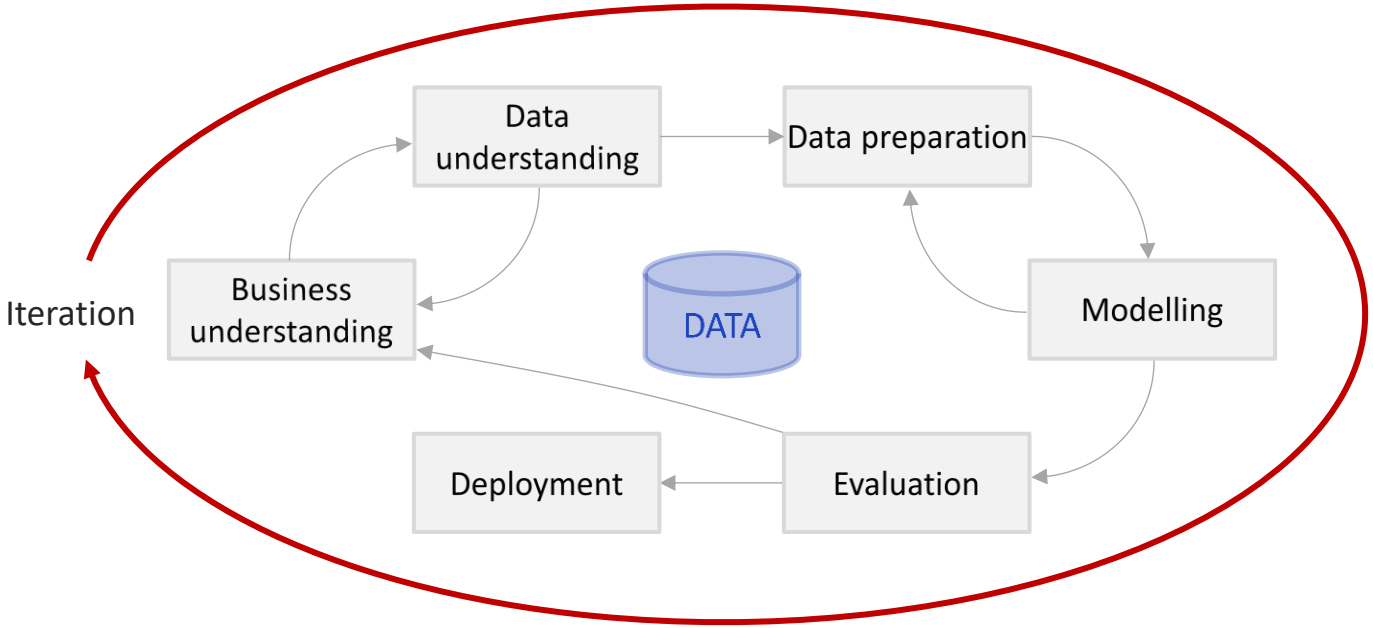


Considerations at the planification phase

1. How to acquire the data and which method to use.
2. Explore the use cases and preexisting solutions.
3. Identify potential hurdles and plan ahead.
4. Management of updates and actualizations.

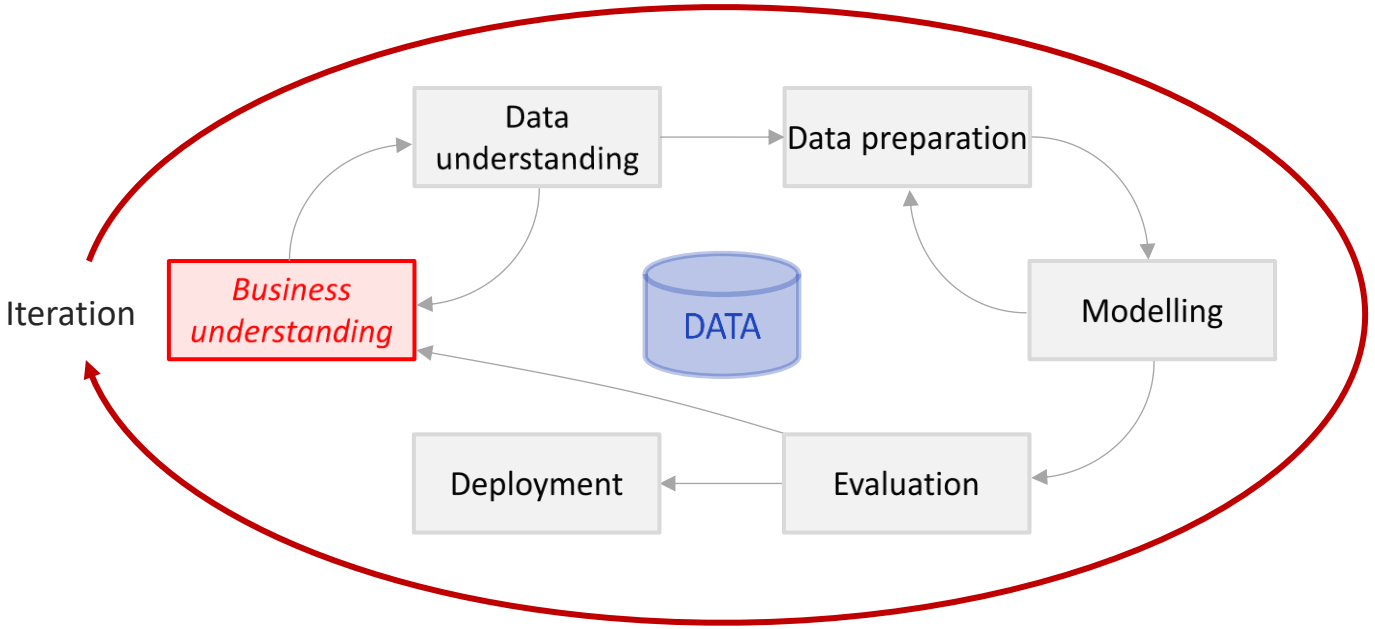


CRISP-DM
Cross Industry Standard Process for Data Mining



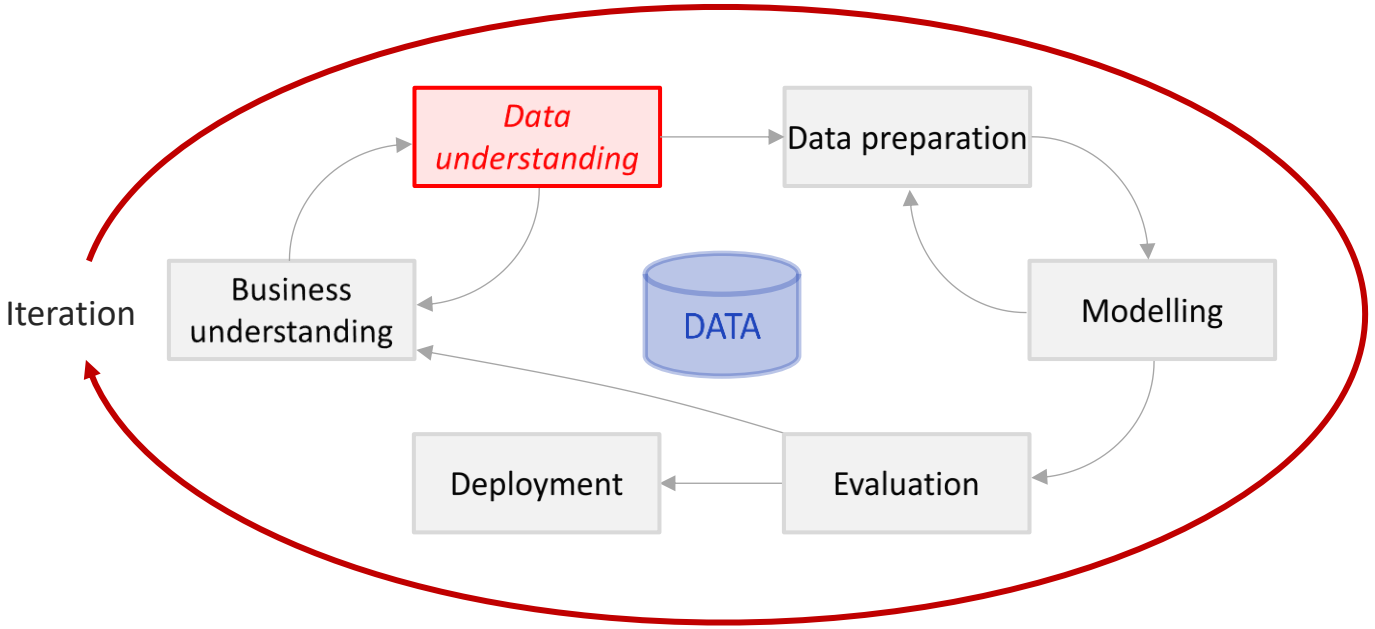
- Structured approach to planning a data mining project.

CRISP-DM
Cross Industry Standard Process for Data Mining



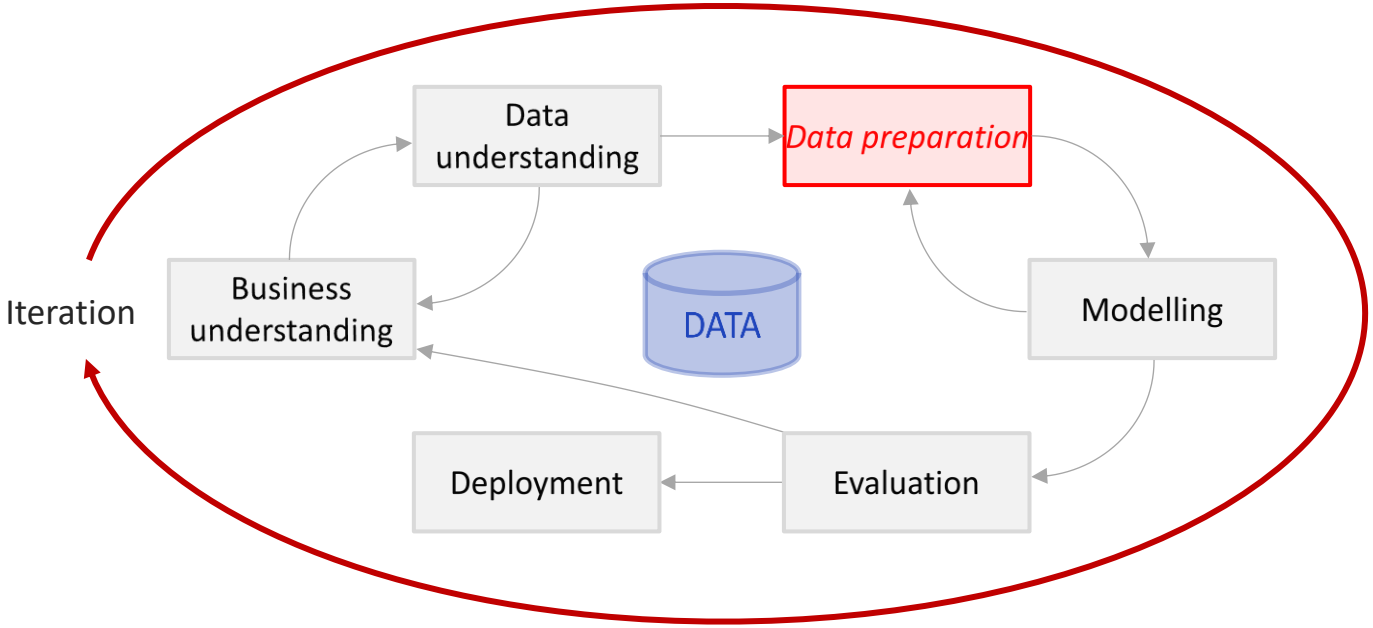
- *Business understanding*: analysis of business objectives and needs.

CRISP-DM
Cross Industry Standard Process for Data Mining



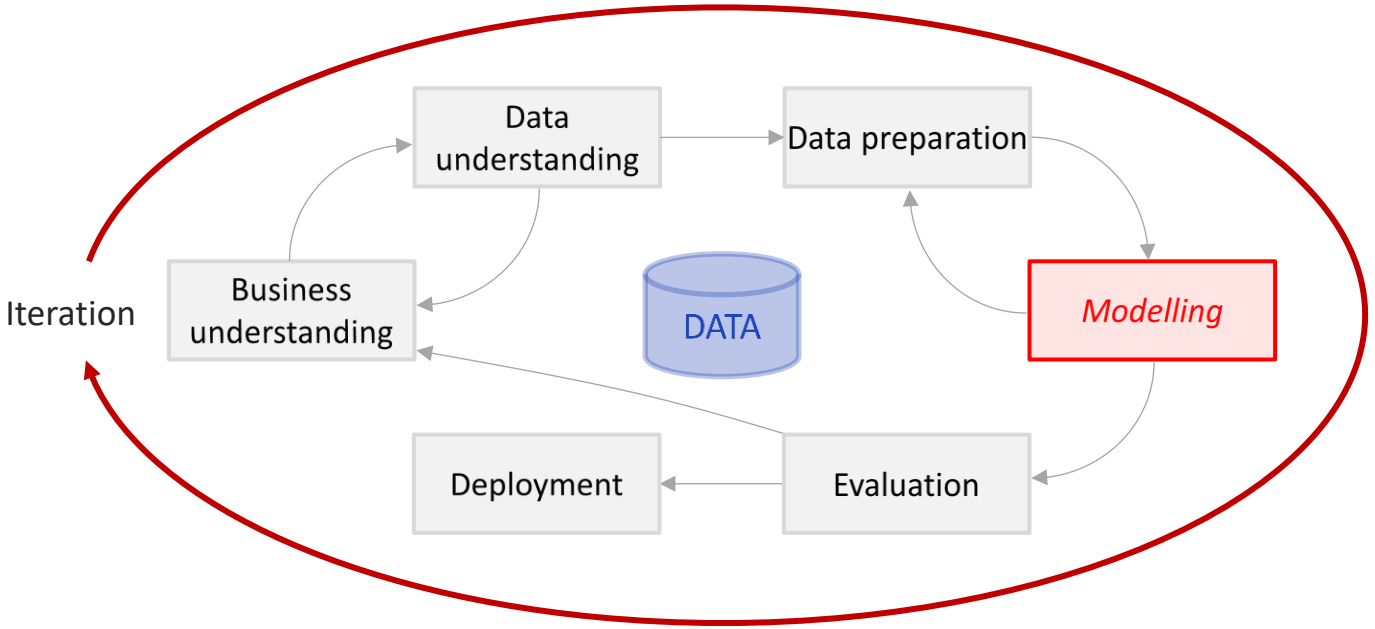
- *Data understanding*: exploratory analysis to gain further understanding of the gathered data.

CRISP-DM
Cross Industry Standard Process for Data Mining



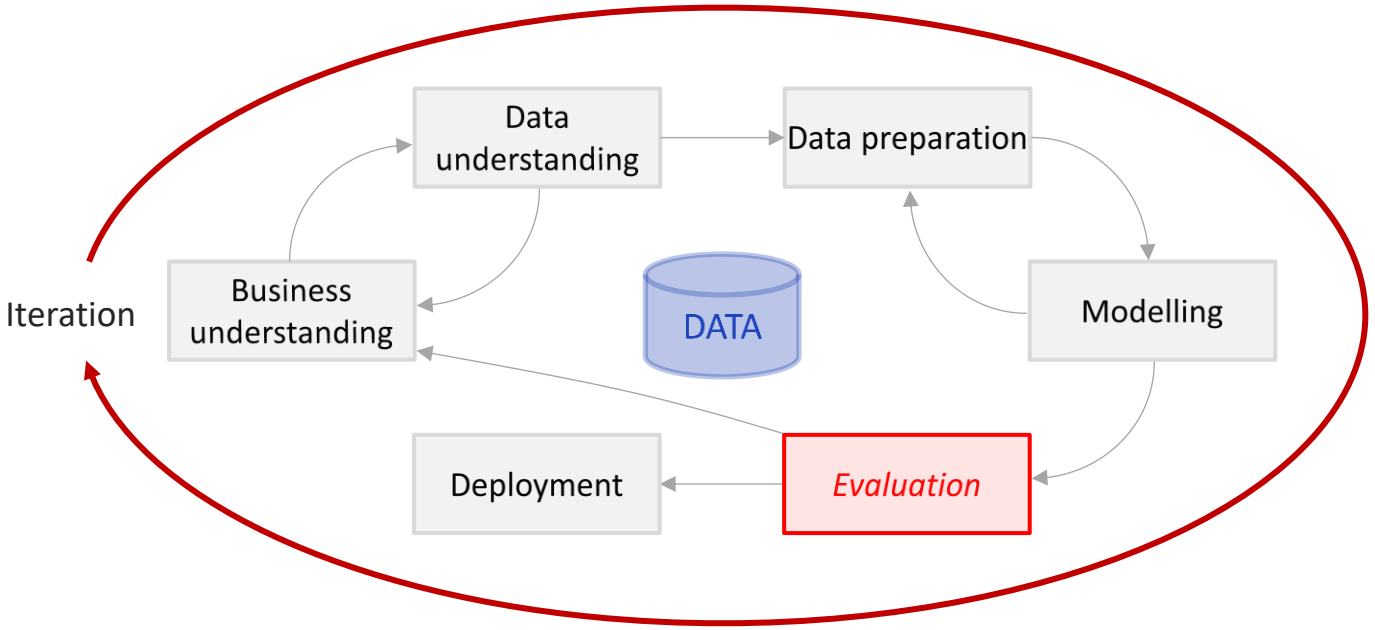
- *Data preparation*: cleaning, formatting, etc. of the data using the insight gained from the previous steps.

CRISP-DM
Cross Industry Standard Process for Data Mining



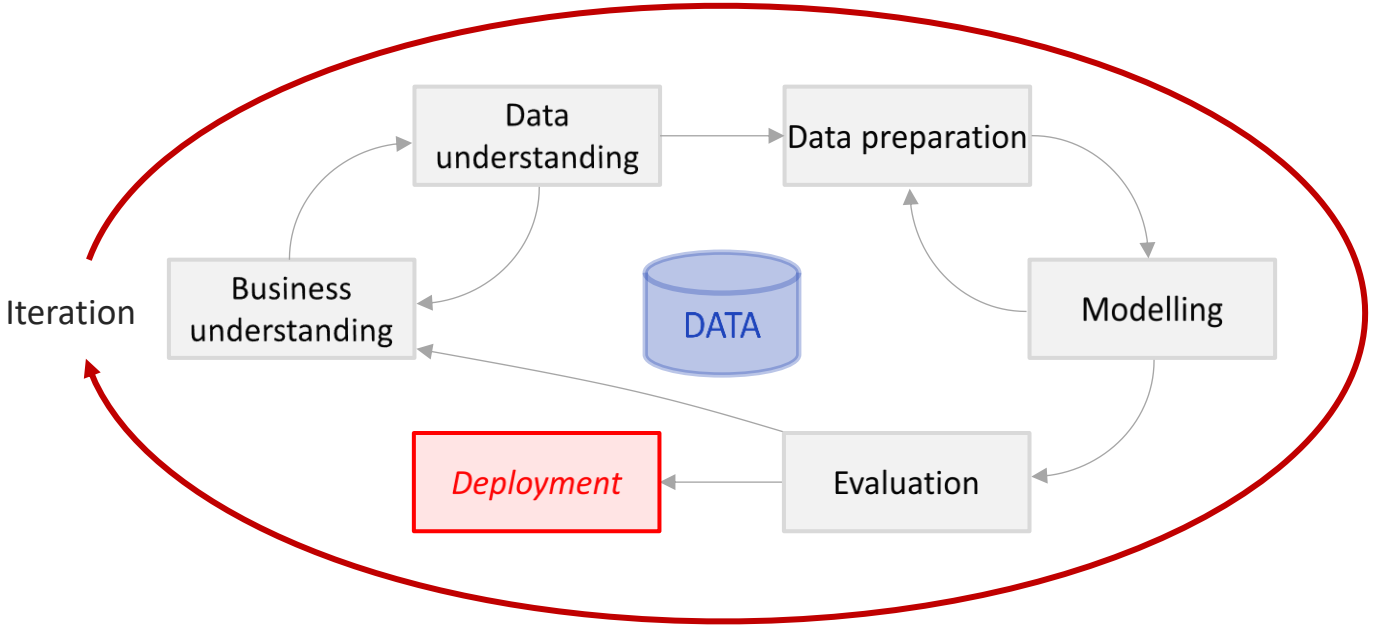
- *Modelling*: involves selection, optimization and streamlining a predictive model.

CRISP-DM
Cross Industry Standard Process for Data Mining



- *Evaluation*: the model from the previous step is put to test in a realistic scenario.

CRISP-DM
Cross Industry Standard Process for Data Mining



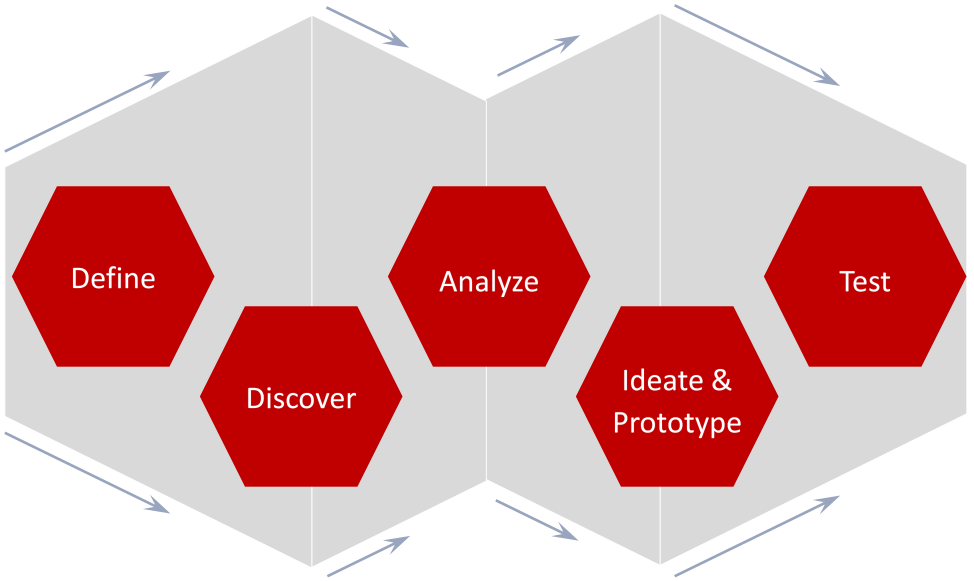
- *Deployment*: the system implemented in a usable format conforming with the business requirements.

Searching for Project Ideas

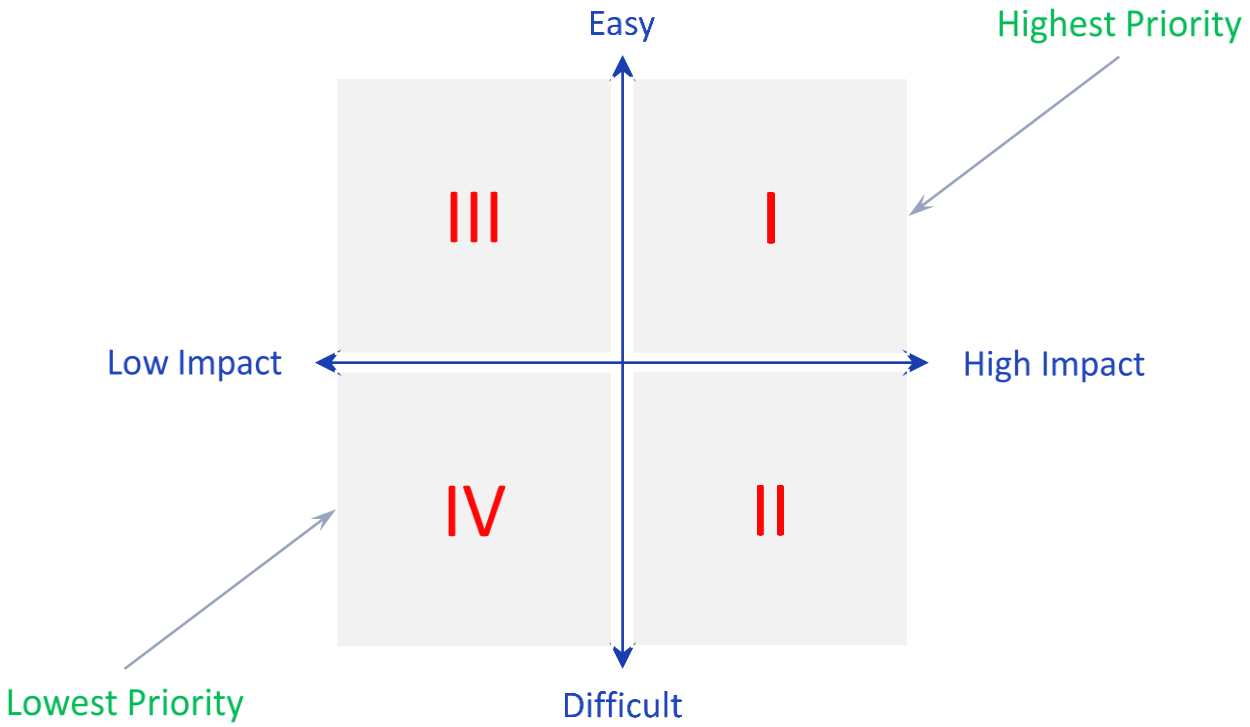
- Top-Down approach
  - When the problem is already well defined
- Bottom-Up approach
  - The problem needs to be defined
  - through exploration of data

Top-Down approach - Design Thinking

- Five steps: Define, Discover, Analyze, Ideate & Prototype, and Test.
- Iterations of convergence and divergence.

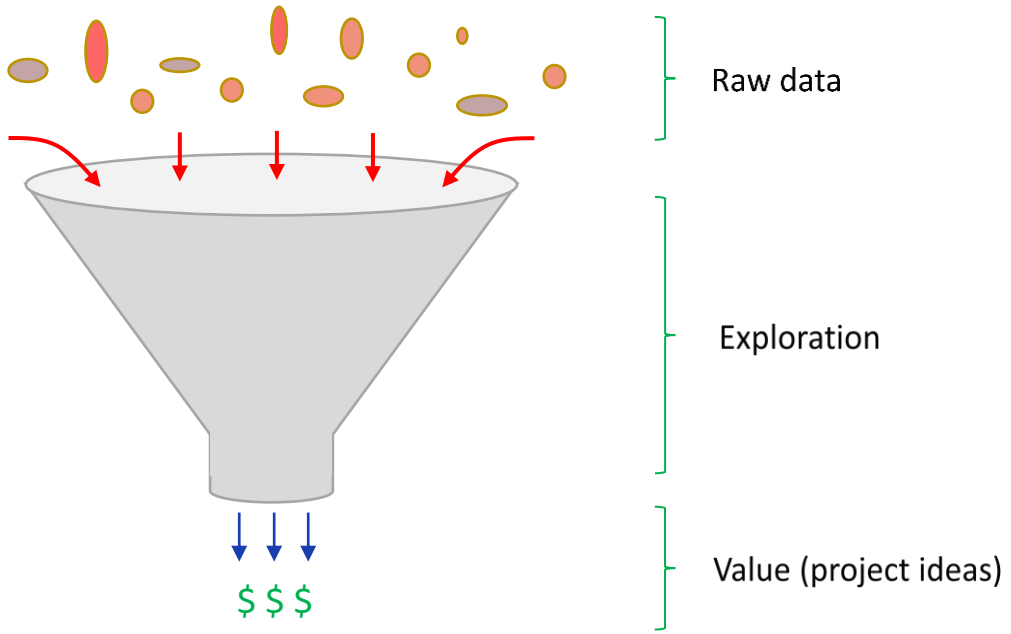


Top-down approach: Prioritizing ideas



Bottom-up approach

- Data Exploration:
  - Explore the raw data to find valuable problems and project ideas.



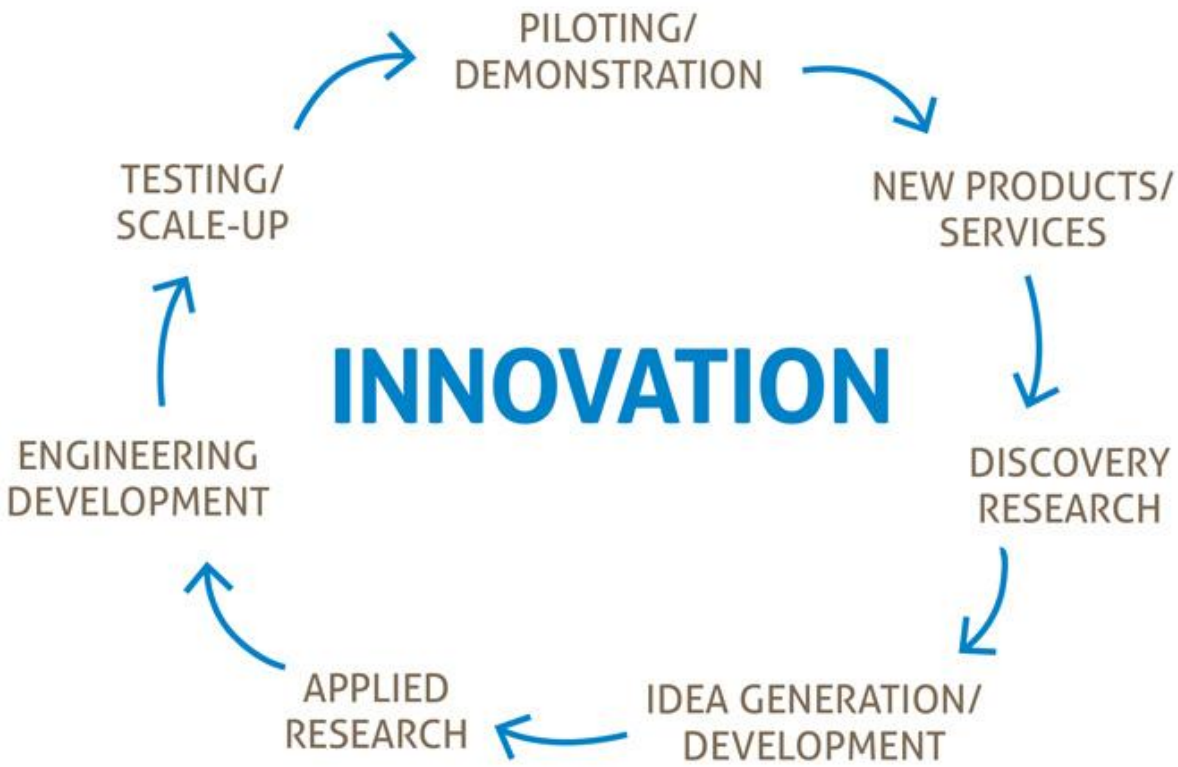
CMU Portugal Academy

04

Measuring Success

Carnegie Mellon Portugal | TÉCNICO LISBOA | Ciências ULisboa | NOVA SCHOOL OF SCIENCE & TECHNOLOGY

Innovation Lifecycle



From theoretical models to experimental models

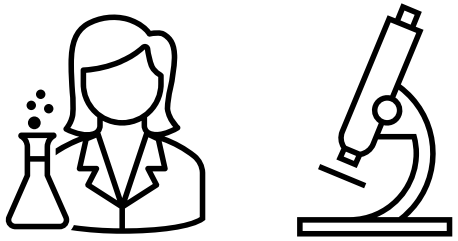
- Is the fundamental law known?
  - Yes: Electromagnetism, transistor, have a sound theoretical model.
  - No: shopping patterns, vision, language, etc. have no fundamental laws.

From experimental model to the deployed model

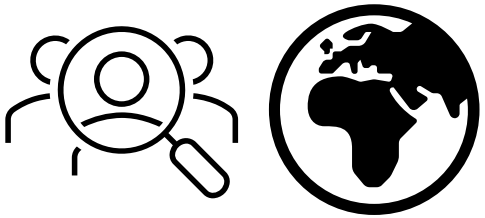
- Every model is flawed by design.
- The true model is unknown.
- The best possible model is the one that best approximates the observed data.

Validating the model

Laboratory



Real-world



Lab evaluation

- Ad-hoc
- Focus Group

} Initial evaluations

- Toy Datasets
- Real World Datasets

} Reproducible evaluations

Ad-hoc

- Developers, Scientists, Quality Assurers and other people dream up data, enter them into the system and eyeball the results.
- Feedback is usually broad and nonspecific, as in “this sample the result was good” or “the results suck”.

- Pros: Low initial cost, low startup costs, gives a general overall sense of the system.
- Cons: Not repeatable and not reliable. Doesn’t produce measures.

Focus Groups

- Gather a set of real users and have them interact with the system over some period of time. Log everything they do and explicitly ask them for feedback.

- Pros: Feedback and logs are quite useful, especially if users feel invested in the process.
- Cons: The results may not be extrapolated to broader audience, depending on how well the users represent your target audience.

Toy Datasets

- Toy datasets usually have a well behaved sample of the real problem.

- Pros: Good for sanity checks and proof of concepts
- Cons: Requires other testing methodologies. Doing well in the dataset doesn't necessarily translate to doing well in real-life.

Real-world Datasets

- Run a relevance study using a set of queries, documents and relevance judgments created by your group or a third-party group.
- Pros: Relatively easy to use, test and compare to previous runs. Completely repeatable. Good as a part of a larger evaluation.
- Cons: Doing well in the dataset doesn't necessarily translate to doing well in real-life.

Deployed evaluations

- Log Analysis of a Beta System
- A/B testing

} Initial evaluations

- Empirical Testing a Live System
- Monitoring a Live System

} In live systems

Log Analysis on a Beta System

- Deploy a live system to a large audience. You should only do this once you are reasonably confident things work well.
- It is imperative to have good logging in place first, which means thinking about the things your want to log, such as queries, results, clickthrough rates, etc.
- Finally, invite feedback from your users.

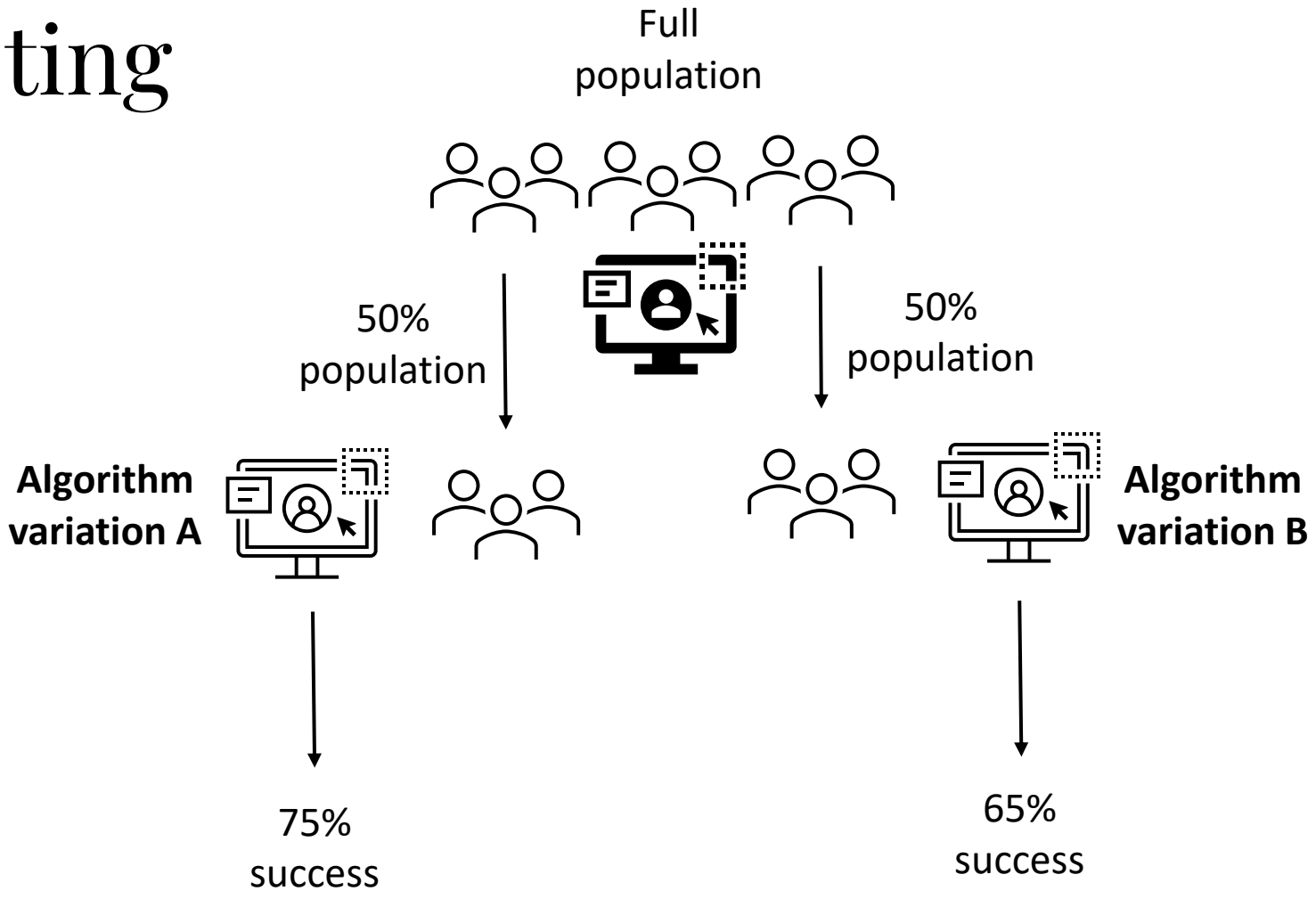
- Pros: Very close to real-word.
- Cons: Expensive. Maybe difficult to reproduce.

A/B Testing

- Assign a percentage of users to go to one system, while the other percentage uses an alternate system.
- Evaluate the choices made by those in the A group and those in the B group to see if one group had better results than the other.
- If feasible, have the users in each group rate the results.

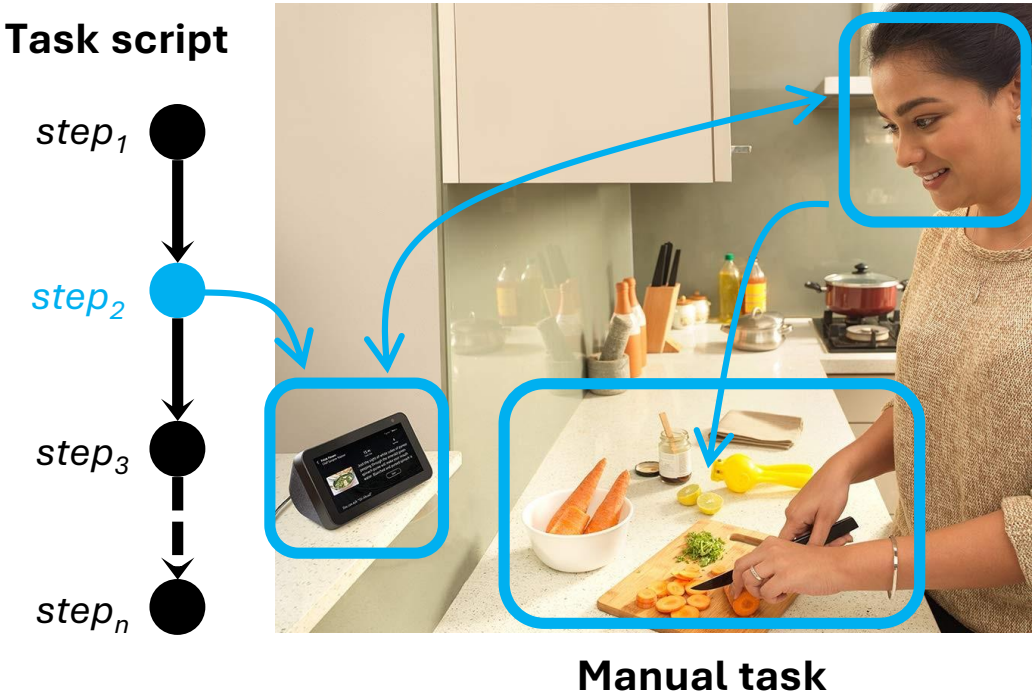
- Pros: Combine with log analysis to get a good picture of what people prefer.
- Cons: Requires setting up and maintaining two systems in production.

A/B testing


Full
population
50%
population
50%
population
Algorithm
variation A
Algorithm
variation B
75%
success
65%
success

Conversational AI for Complex Manual Tasks

- Assistant's behavior is grounded on an arbitrarily selected task.
- Each task requires a set of resources and is organized into a sequence of steps.
- Users wish to navigate the task steps.
- Users ask questions about a task action, element or video.



User ratings of 5 different Amazon Alexa Systems

5 different systems

Rank	Avg Feedback Rating				Number of Conversations	Percentage of Completed Conversations
	L7d**	95% C.I.	Week-Ago	L1d	L7d	
1***	3.6	± 0.26	3.23	3.53	2894	38.5%
2	3.21	± 0.34	3.14	2.71	2872	33.6%
3	2.98	± 0.31	3.0	3.36	2647	29.7%
4	2.84	± 0.60	2.93	3.0	3141	17.1%
5	2.67	± 0.26	2.49	3.1	2882	18.1%
Average	**3.06**	**-**	**2.96**	**3.14**	**2887.2**	**-**

Average rating over last 7 days

Number of user tests over the last 7 days

Continuous Testing of a Live System

- Given an existing system, select the top X inputs in terms of volume and Y randomly selected inputs not in the top X.
- Have your Quality Assurance team examine the input/output and rate the top five or ten results as relevant, somewhat relevant and not relevant (and a fourth option: embarrassing).

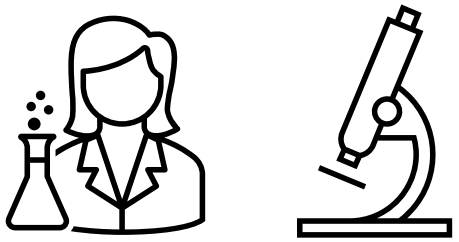
- Pros: Real queries, real documents, real results.
- Cons: Time consuming. Y becomes too large for long-tail settings.

Monitoring a Live System

- Return rates
- Conversion rates
- Abandonment rates
- Churn prediction
- …

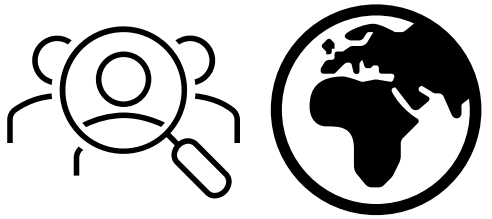
Validating the model

Development



- Ad-hoc
- Focus Group
- Toy Datasets
- Real World Datasets

Deployment

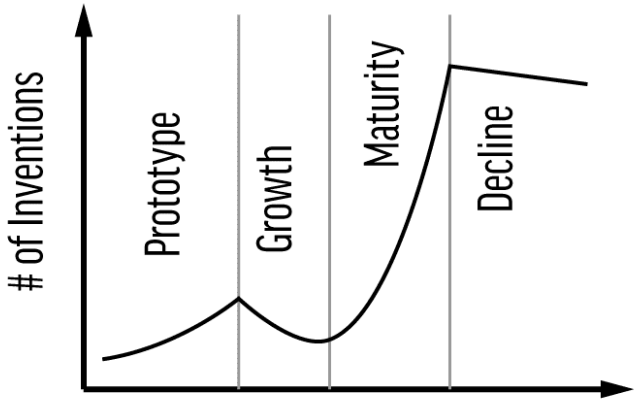


- Log Analysis of a Beta System
- A/B testing
- Empirical Testing a Live System
- Monitoring a Live System

https://blueironip.com/the-s-curve-innovation-and-intellectual-property/

Impact of an innovation technology

- Prototype phase:
  - Lab experiments are the main drivers
  - Groundbreaking invention unlocks innovation
  - Leading tech companies and academic research
- Growth phase:
  - Real-world experiments are the main drivers
  - Problem is well understood
  - Initial ideas generate high-gains
- Maturity phase:
  - New ideas generate low-gains
  - Mainly industry research
  - Operations optimizations



CMU Portugal Academy

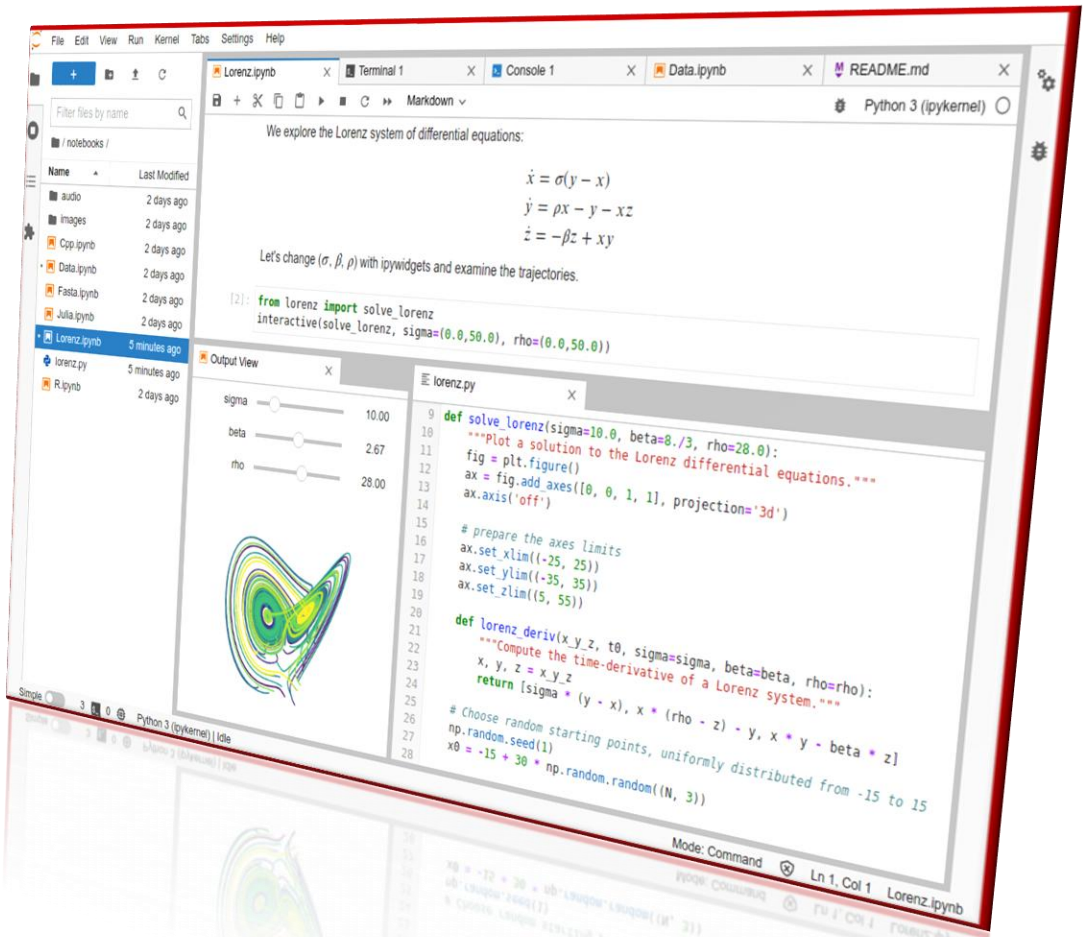
05

Hands-on practice

Carnegie Mellon Portugal · TÉCNICO LISBOA · Ciências ULisboa · NOVA SCHOOL OF SCIENCE & TECHNOLOGY

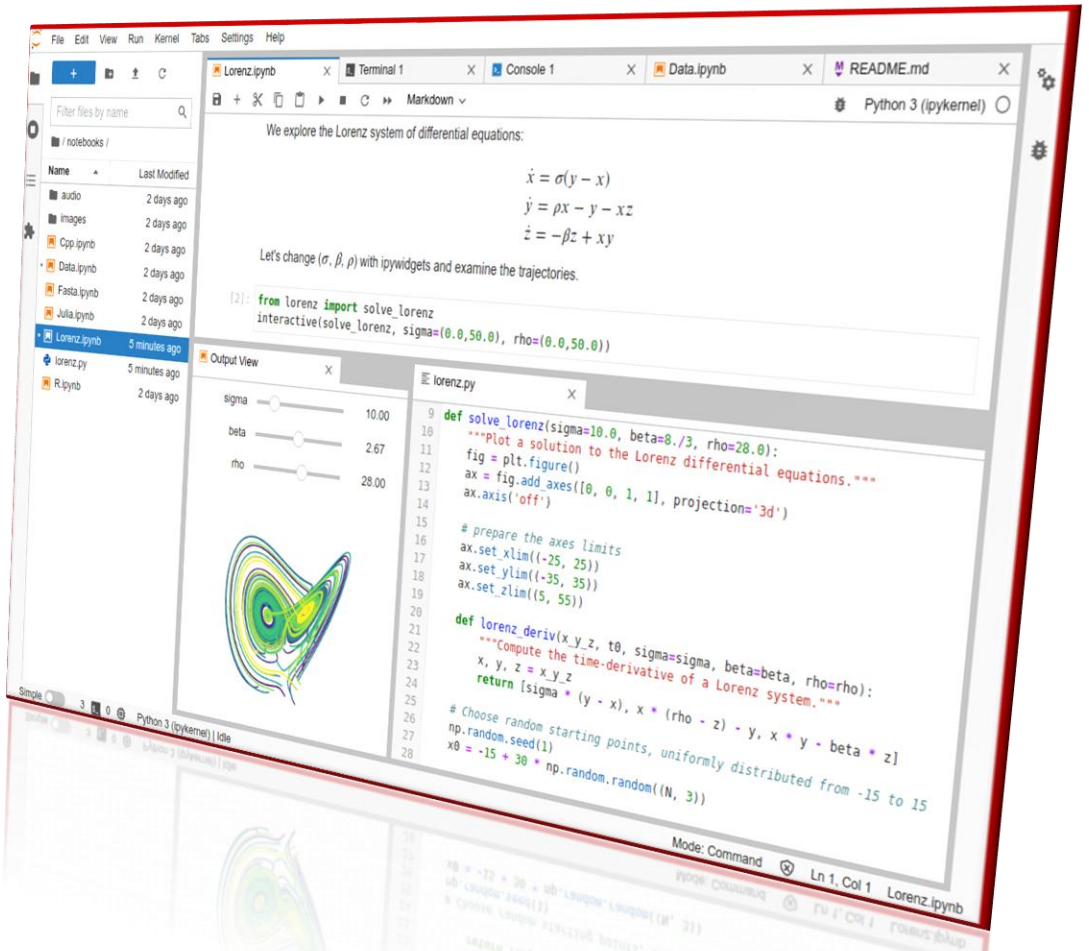
Laboratory setup

https://wiki.novasearch.org/wiki/lab_setup



Programming practice

[CMU Data Science Bootcamp]



Exercise

1. Identify industry and domain
2. Characterize data
3. Select end-user
4. Identify objectives and write example queries/use cases
5. List required AI + ML + DS components
6. Measure success and value

Send Miro chart to
jmag@fct.unl.pt
by next thursday.

CMU Portugal Academy

Carnegie Mellon Portugal · TÉCNICO LISBOA · Ciências ULisboa · NOVA SCHOOL OF SCIENCE & TECHNOLOGY