CMU Portugal Academy

CMU Portugal
Advanced Training Program

Artificial Intelligence, Machine Learning and Data Science Bootcamp

JOÃO MAGALHÃES

NOVA SCHOOL OF SCIENCE AND TECHNOLOGY

email: jmag@fct.unl.pt

João Magalhães – Short bio

jmag@fct.unl.pt

Full Professor, NOVA FCT

Head of the Multimodal Systems Group, NOVA LINCS

PhD, Imperial College London, 2008

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

I am keen to solve real-world problems with data-driven NLP, CV and DL methods in different domains.

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

Vision and Language AI

amazon | science

vision-box

RTP

FARFETCH

Google

BBC R&D

DeepNeuronic
DEEP NEURAL SYSTEMS FOR AUTOMATIC VISION

ARQUIVO.PT

Carnegie Mellon Portugal

VITEC
VIDEO INNOVATIONS

Topics

1. PROGRAM STRUCTURE

2. WHAT IS AI + ML + DS?

3. BRAINSTORM: APPLICATIONS AND USE CASES

4. LABORATORY SETUP

Visit to download materials.

01

Program Structure

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

- 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 & Rafael Ferreira	30	2
Machine Learning	Chryssa Zerva & Sweta Agrawal	30	2
Data Collection and Pre-Processing	Cátia Pesquita & Tiago Guerreiro	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

Eng. Rafael Ferreira

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. Sweta Agrawal

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
Deep Learning	Chryssa Zerva & David Semedo	18	1
Vision and Language	Bruno Martins & João Magalhães	18	1
Complex Data Analysis	André Falcão	18	1
Cloud-based Data Processing	Nuno Preguiça & Rodrigo Rodrigues	18	1
Data analytics and visualization	João Moura Pires & Manuel Fonseca	18	1
CapStone	Todos	48	3

Deep Learning

- Introduction to Deep Learning
- Neural Network Fundamentals
- Convolutional Neural Networks (CNNs)
- Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM)
- Generative Adversarial Networks (GANs)
- Attention mechanisms and transformer models
- Large Language Models (LLMs)
- Model Interpretability and Explainability
- Safe deployment, robustness, bias and sustainability

Prof. Chryssa Zerva

Prof. David Semedo

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

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

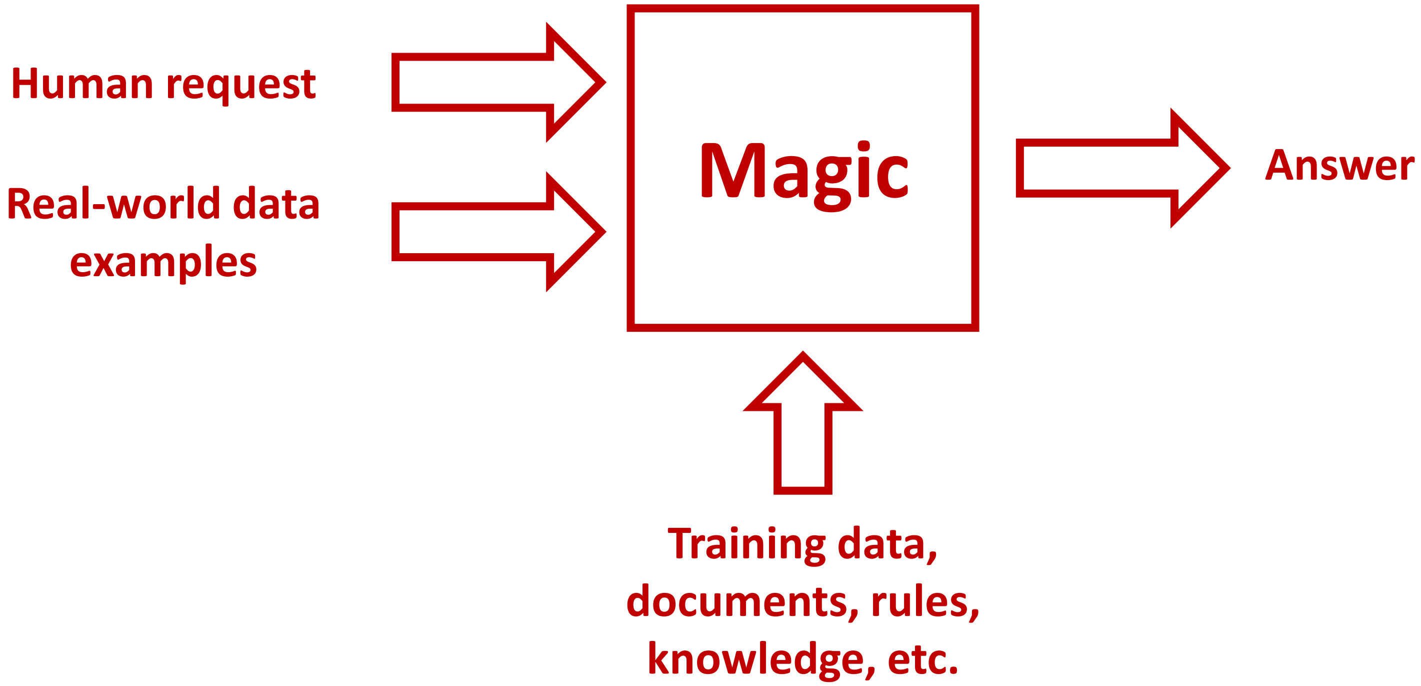
- Graph data
- Social Network Analysis
- Time Series Analysis
- Geospatial Data Analysis

Prof. André Falcão

02

What is AI + ML + DS?

What is “the” algorithm?



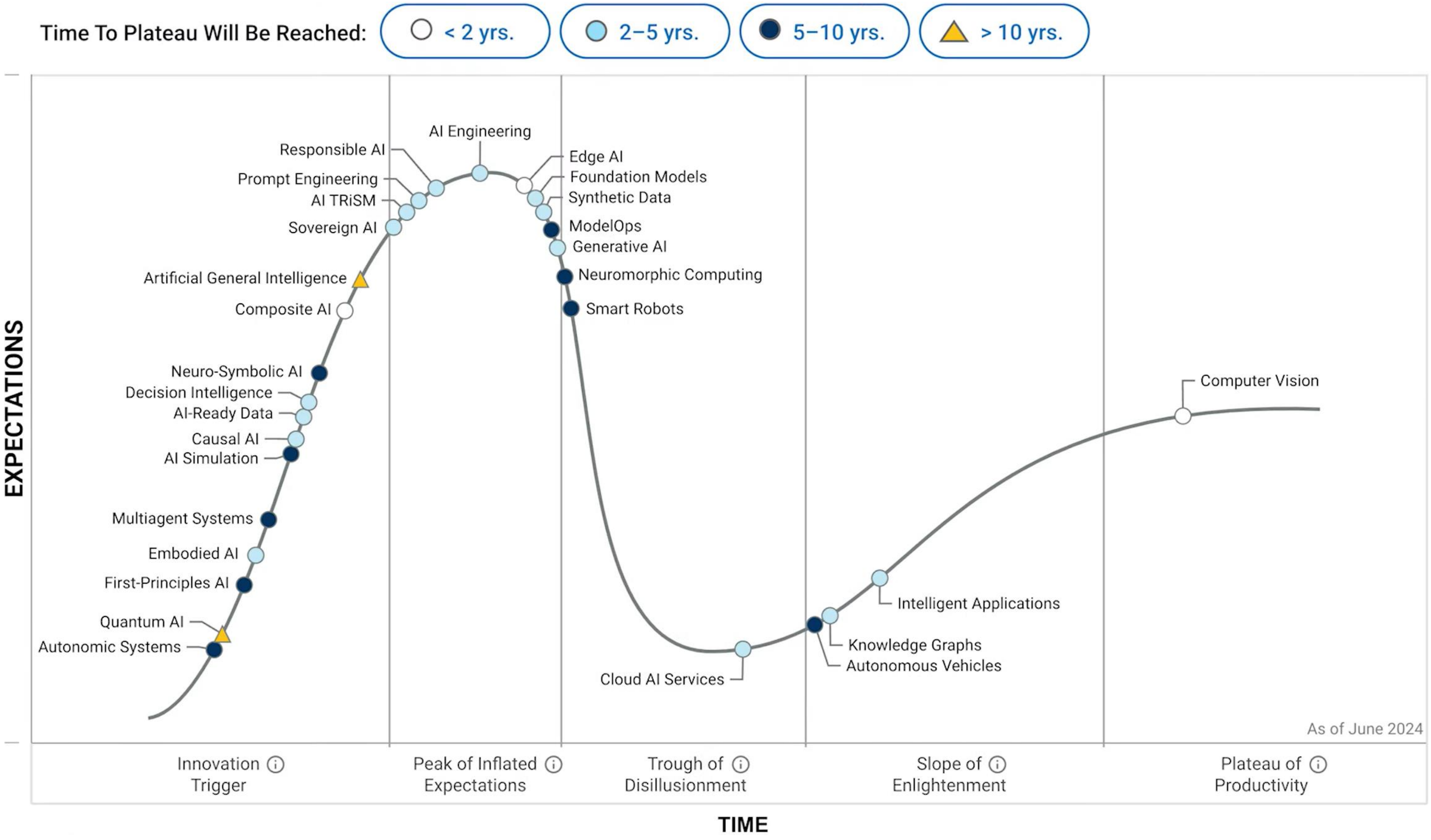
Human request

Real-world data examples

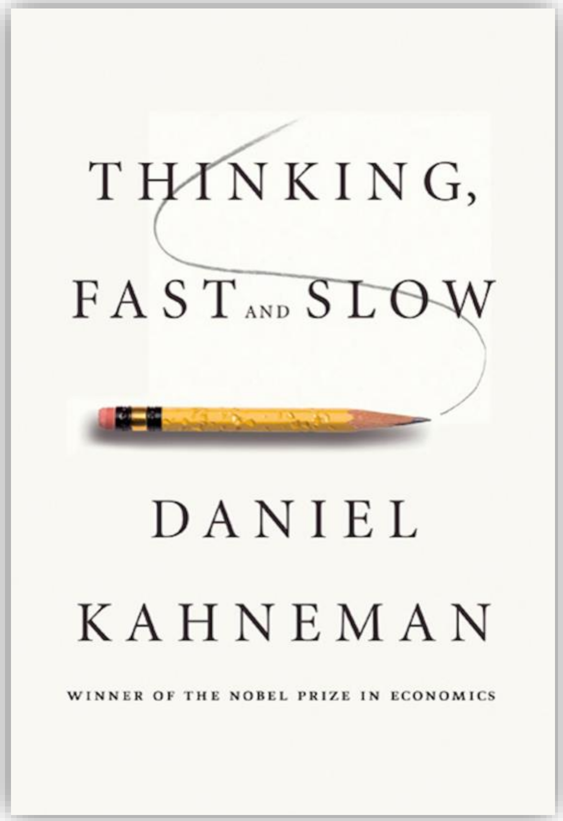
Magic

Answer

Training data, documents, rules, knowledge, etc.

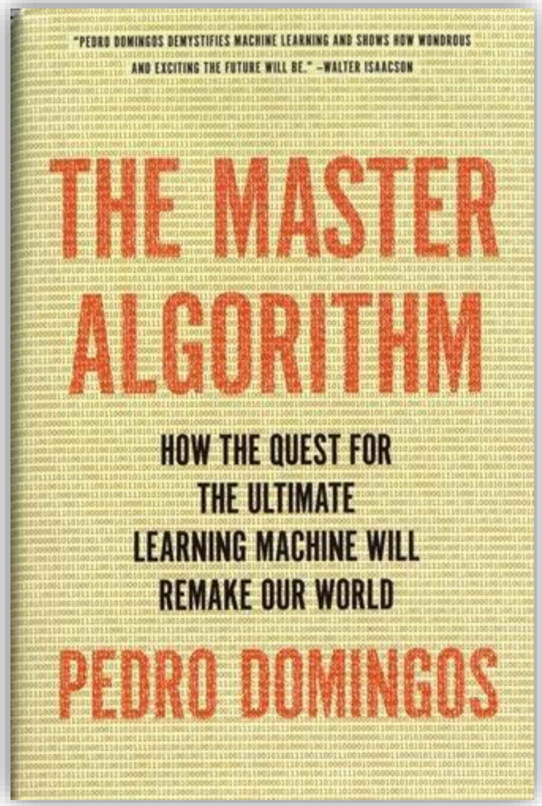

Time To Plateau Will Be Reached:
○ < 2 yrs.
● 2–5 yrs.
● 5–10 yrs.
△ > 10 yrs.
EXPECTATIONS
AI Engineering
Responsible AI
Prompt Engineering
AI TRiSM
Sovereign AI
Artificial General Intelligence
Composite AI
Neuro-Symbolic AI
Decision Intelligence
AI-Ready Data
Causal AI
AI Simulation
Multiagent Systems
Embodied AI
First-Principles AI
Quantum AI
Autonomic Systems
Edge AI
Foundation Models
Synthetic Data
ModelOps
Generative AI
Neuromorphic Computing
Smart Robots
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


"PEDRO DOMINGOS DEMYSTIFIES MACHINE LEARNING AND SHOWS HOW WONDROUS AND EXCITING THE FUTURE WILL BE." —WALTER ISAACSON
THE MASTER ALGORITHM
HOW THE QUEST FOR THE ULTIMATE LEARNING MACHINE WILL REMAKE OUR WORLD
PEDRO DOMINGOS

03

Brainstorm: Applications and Use Cases

Brainstorm

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

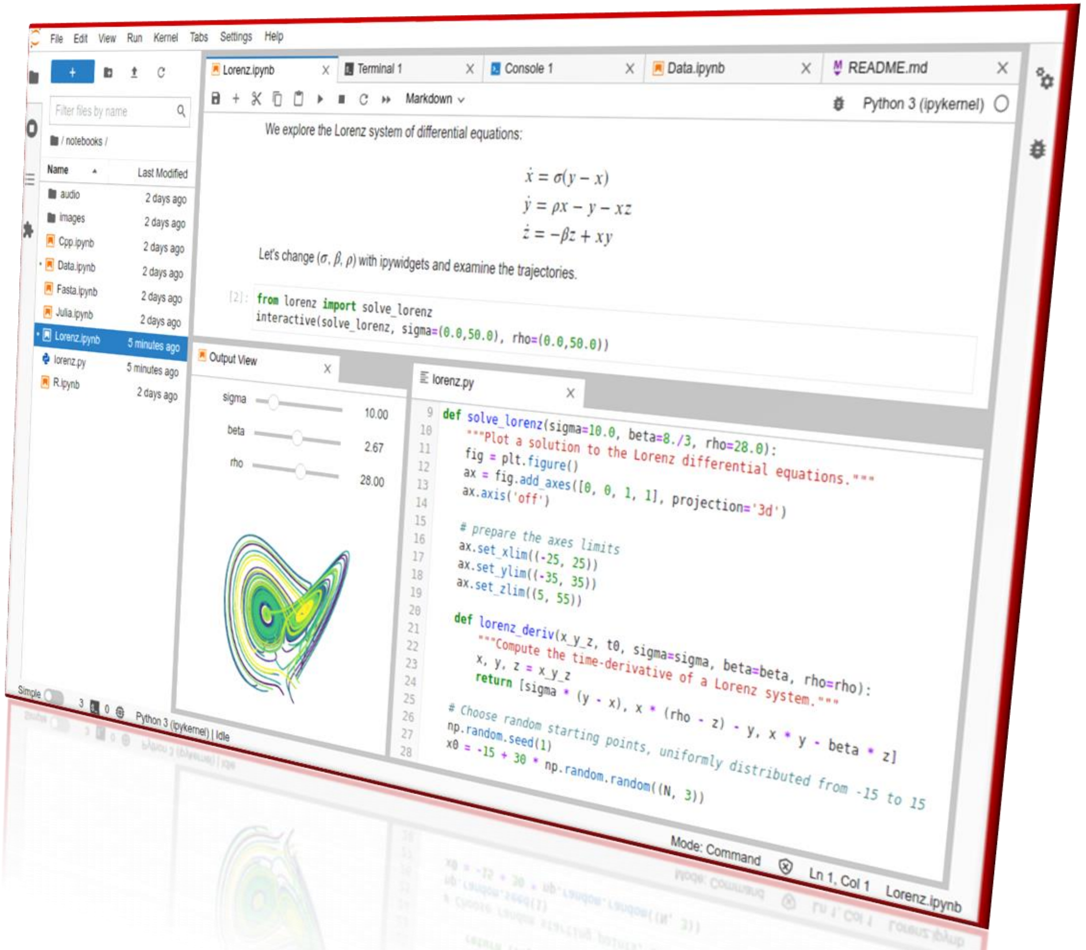
1. Identify industry and domain
2. Characterize data
3. Select end-user
4. Identify objectives and write example queries/use cases
5. List required AI + ML + DS components
6. Measure success and value

04

Laboratory Setup

Laboratory setup

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



We explore the Lorenz system of differential equations:

$$\dot{x} = \sigma(y - x)$$
$$\dot{y} = \rho x - y - xz$$
$$\dot{z} = -\beta z + xy$$

Let's change (σ, β, ρ) with ipywidgets and examine the trajectories.

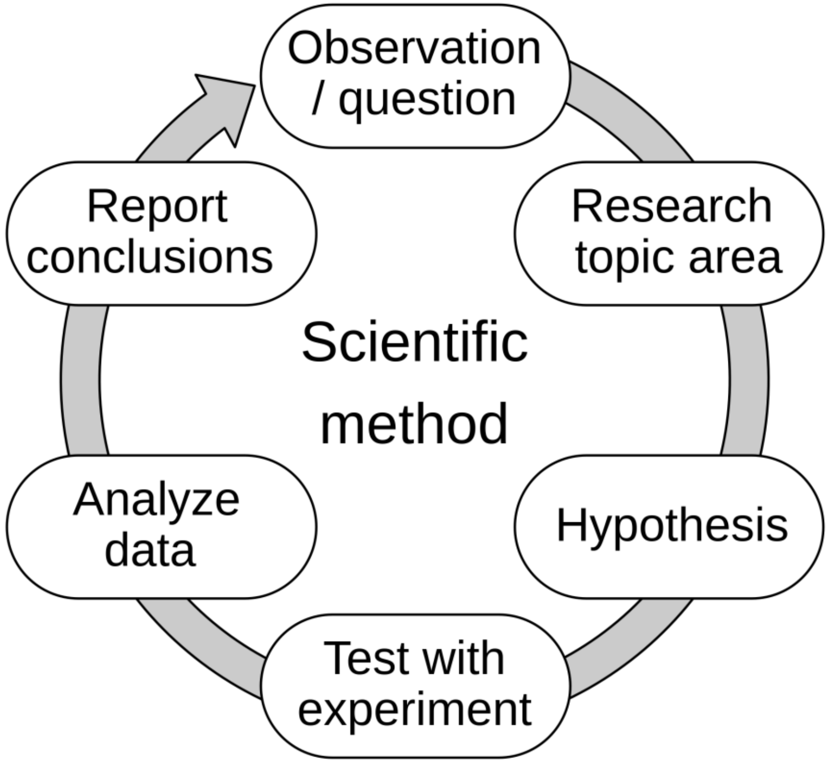
```
[2]: from lorenz import solve_lorenz
     interactive(solve_lorenz, sigma=(0.0,50.0), rho=(0.0,50.0))
```

```
def solve_lorenz(sigma=10.0, beta=8./3, rho=28.0):
    """Plot a solution to the Lorenz differential equations."""
    fig = plt.figure()
    ax = fig.add_axes([0, 0, 1, 1], projection='3d')
    ax.axis('off')

    # prepare the axes limits
    ax.set_xlim((-25, 25))
    ax.set_ylim((-35, 35))
    ax.set_zlim((5, 55))

    def lorenz_deriv(x_y_z, t0, sigma=sigma, beta=beta, rho=rho):
        """Compute the time-derivative of a Lorenz system."""
        x, y, z = x_y_z
        return [sigma * (y - x), x * (rho - z) - y, x * y - beta * z]

    # Choose random starting points, uniformly distributed from -15 to 15
    np.random.seed(1)
    x0 = -15 + 30 * np.random.random((N, 3))
```

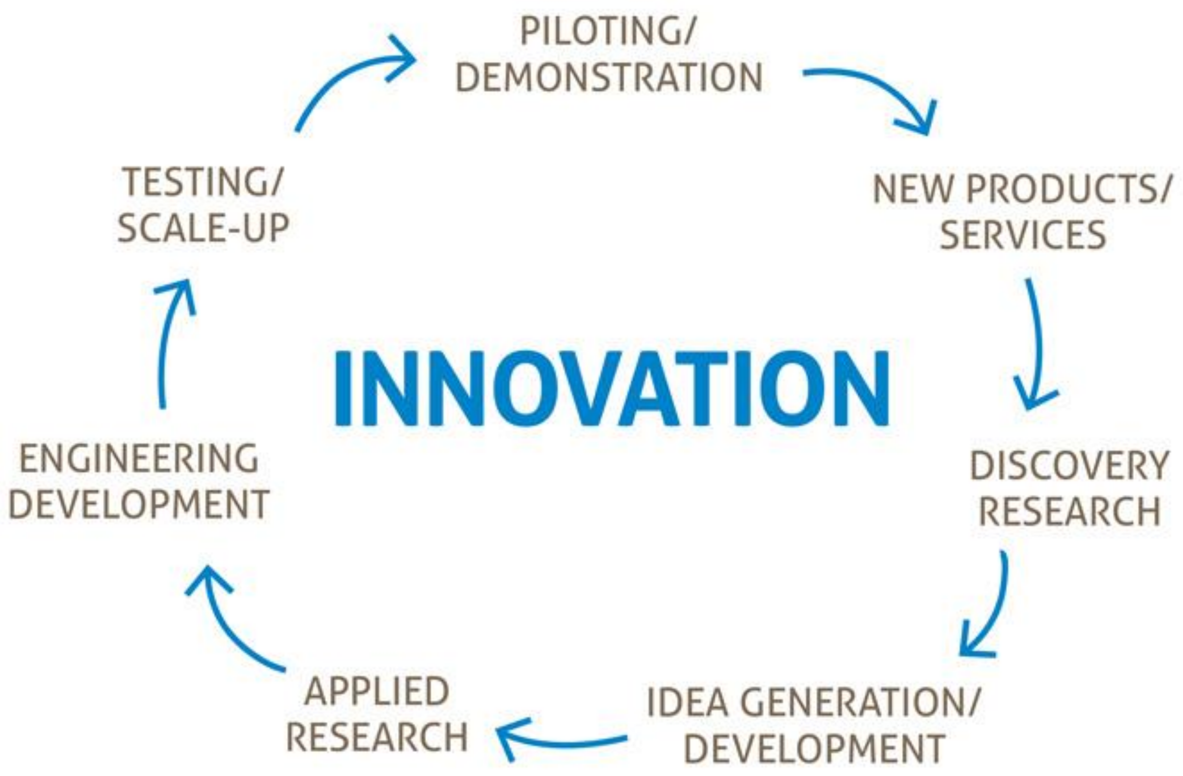
05

Measuring Success

Scientific method



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.

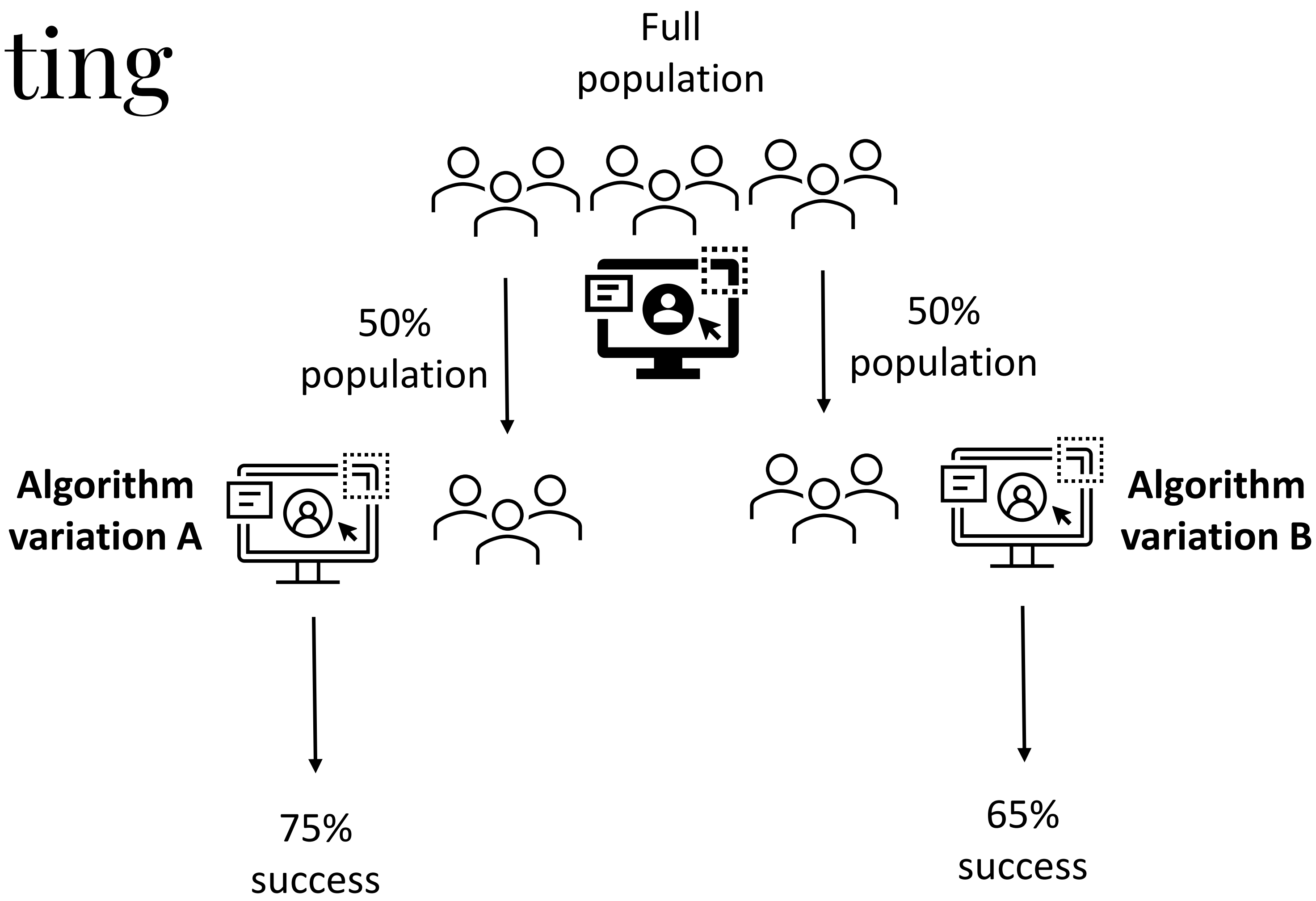
- Pros: Very close to real-word.
- Cons: Expensive. Maybe difficult to reproduce.

A/B Testing

- Assign a percentage of users to go to one system, while the other percentage uses an alternate system.
- After the appropriate period of time, 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.

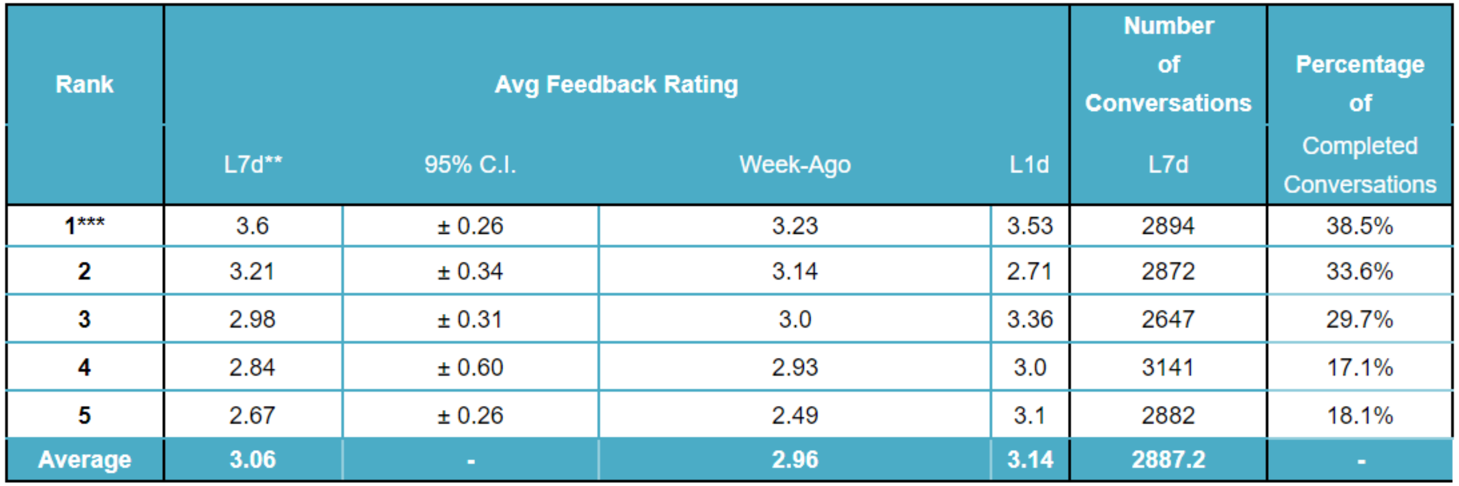
- Pros: Combine with log analysis to get a good picture of what people prefer.
- Cons: Requires setting up and maintaining two systems in production.

A/B testing


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

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

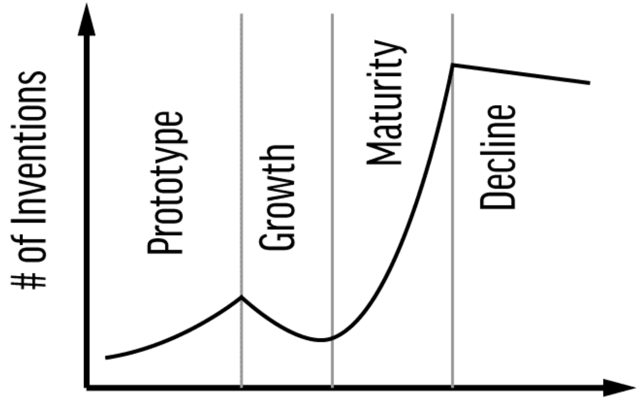
Deployment

- 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

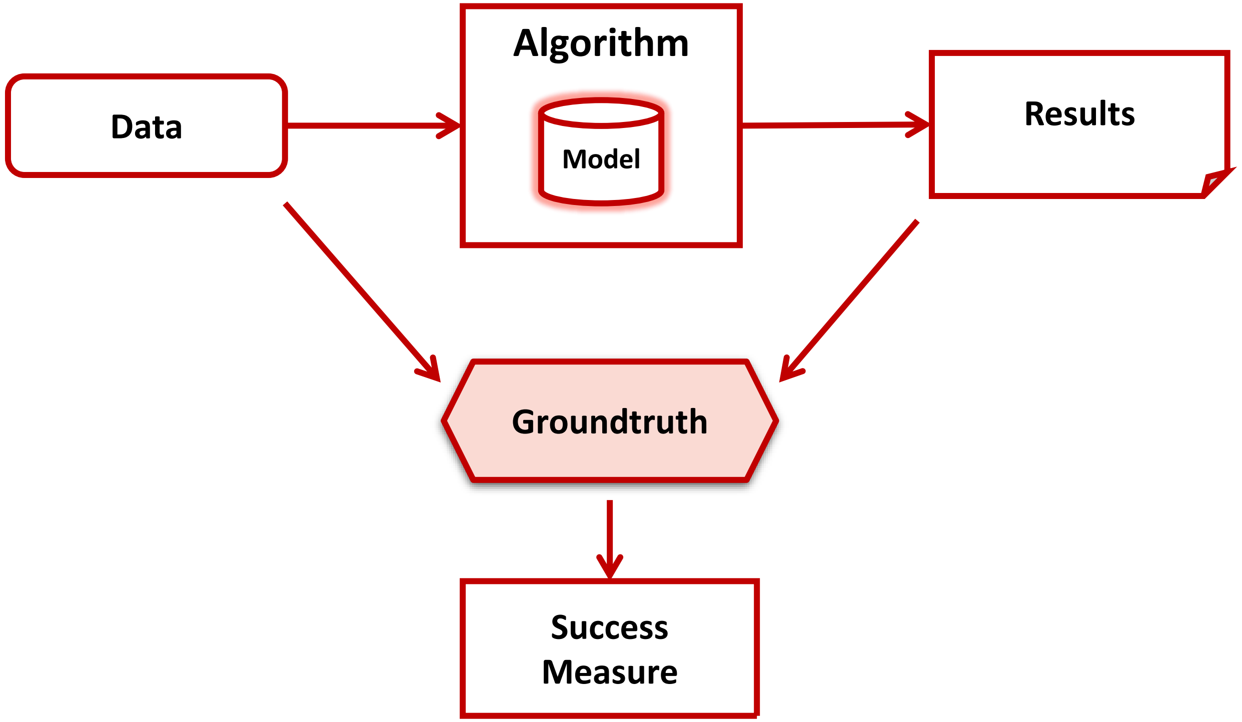
- Prototype phase:
  - Lab experiments are the main drivers
  - Groundbreaking invention unlocks innovation
  - Leading tech companies and academic research
- Growth phase:
  - Real-world experiments are the main drivers
  - Problem is well understood
  - Initial ideas generate high-gains
- Maturity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



06

Metrics

Evaluation setup



Ground-truth

- The theoretical “ultimate goal” is to devise a method that produces exactly the same output as the ground-truth.

- The practical “ultimate goal” can be very different:
    - ground-truth is incomplete, incorrect and only mirrors a small portion of reality.

		Ground-truth	
		True	False
Method	True	True positive	False positive
	False	False negative	True negative

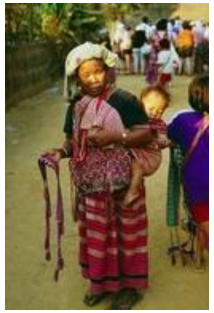
Type I error

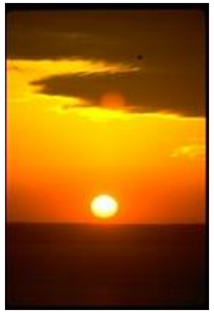
Type II error

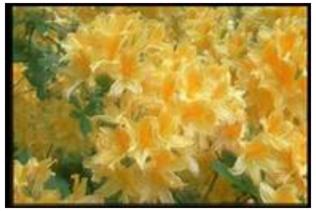
Obtaining groundtruth annotations

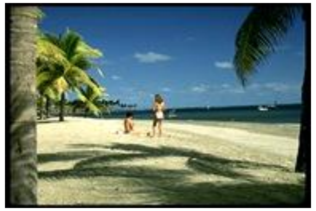
- Crowdsourcing system
  - DefinedCrowd, Amazon Mechanical Turk, …
  - Limesurvey: https://github.com/LimeSurvey/LimeSurvey
  - Relevation: https://github.com/ielab/relevation
- Quality annotations
  - Redundant annotations
- Cost reduction strategies
  - Convergence
  - Pooling strategies

Annotate these pictures with keywords:









Groundtruth by annotation

People
Nepal
Mother
Baby
Colorful dress
Fence

Sunset
Horizon
Clouds
Orange
Desert

Flowers
Yellow
Nature

Beach
Sea
Palm tree
White-sand
Clear sky

- Groundtruth is incomplete
- Not all groundtruth is of equal importance/relevance.

Evaluation metrics

- Utility metrics are focused in evaluating the results that are presented to the user
  - Common metrics for binary relevance judgments: Top Precision and Recall

- Stability metrics are focused in evaluating the robustness of the system results.
  - Common metrics: AUC, AP, Precision-Recall curves, ROC curves

Em PT: exatidão, precisão e abragência.

Binary relevance judgments

$$Accuracy = \frac{truePos + trueNeg}{truePos + falsePos + trueNeg + falseNeg}$$

$$Precision = \frac{truePos}{truePos + falsePos}$$

$$Recall = \frac{truePos}{truePos + falseNeg}$$

$$F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$

		Ground-truth	
		True	False
Method	True	True positive	False positive
	False	False negative	True negative

https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall

Beware of Accuracy

You easily get 99.999999% by not retrieving non-relevant results!!!

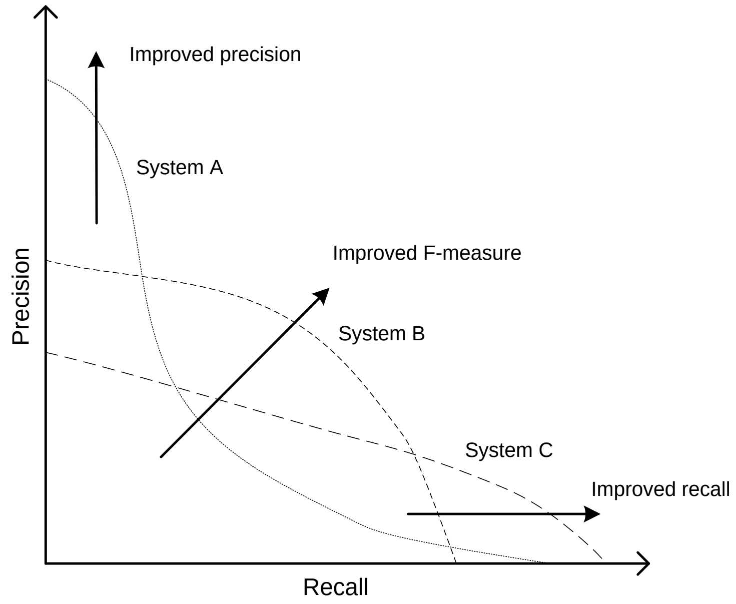
$$Accuracy = \frac{truePos + trueNeg}{truePos + falsePos + trueNeg + falseNeg}$$

Precision-recall graphs for ranked results

- The precision-recall curve shows the tradeoff between precision and recall for different thresholds.

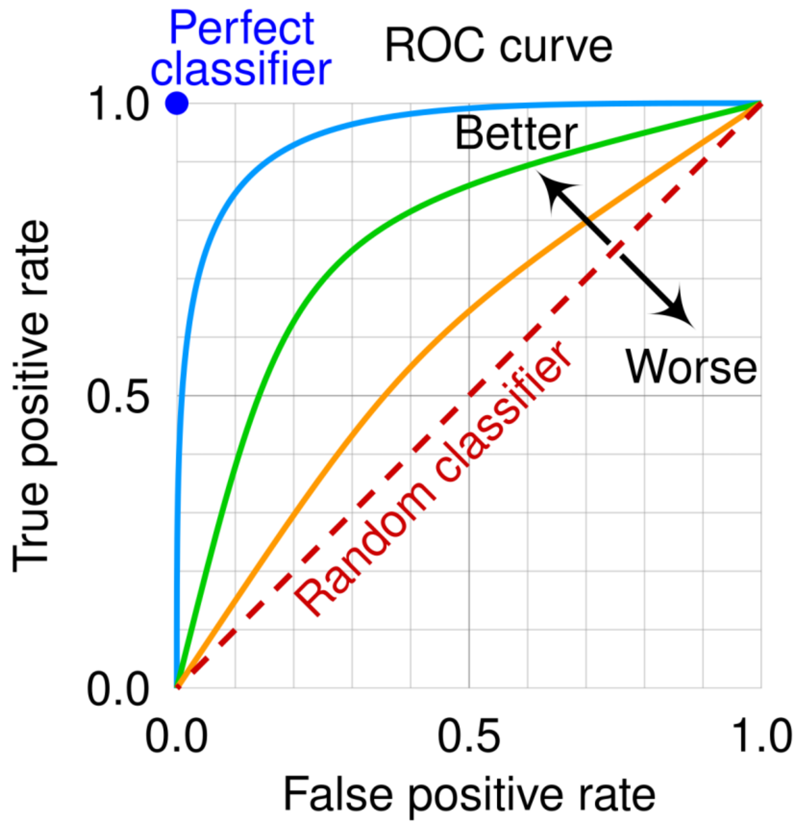
- Consider three examples:

S_A	S_B	S_C
A	…	…
B	A	…
…	…	…
…	B	A
…	…	B
…	C	C
…	…	D
…	…	…



Receiver operating Curve

- The ROC curve is the plot of the true positive rate (TPR) against the false positive rate (FPR) at each threshold setting.



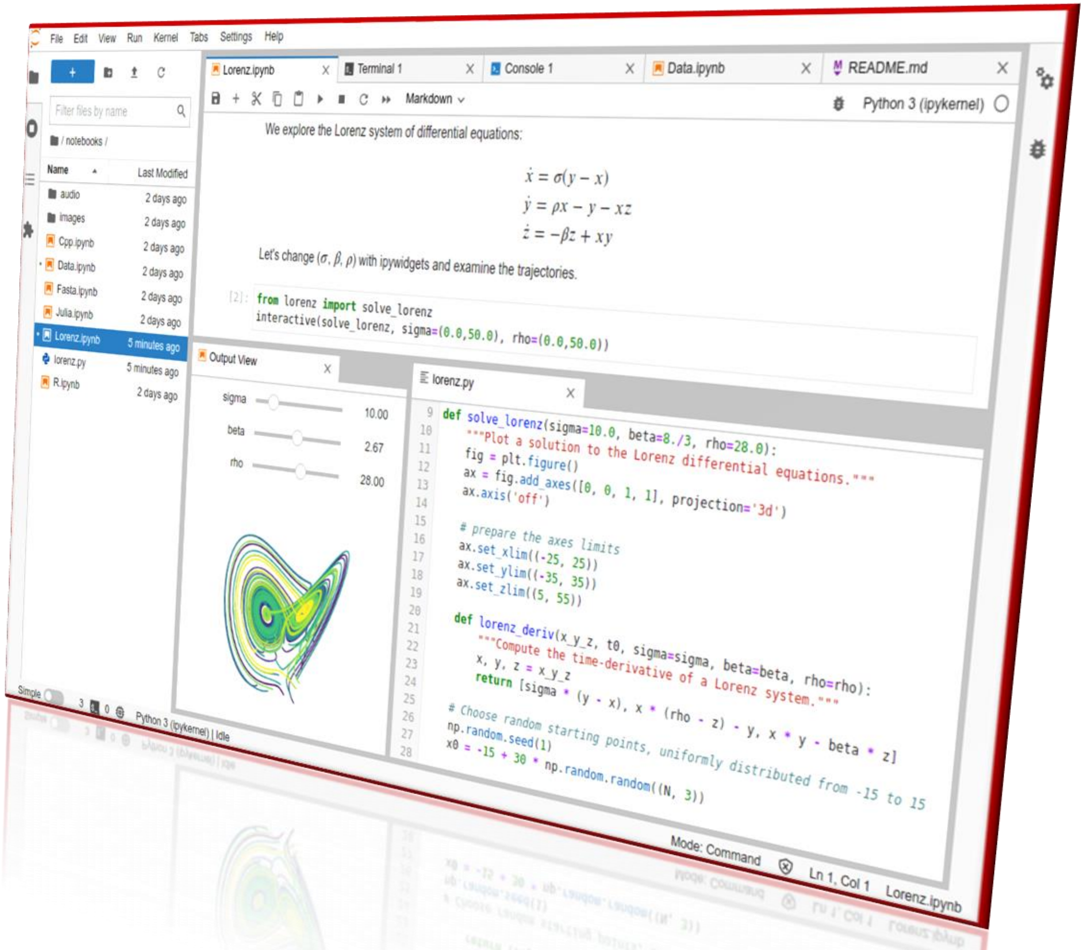
https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Summary

- Measuring success
  - Metrics (for replicable+repeatable experiments)
  - A/B testing is the best way of measuring success
- Groundtruth
  - Incomplete, incorrect, ambiguous
  - Results pooling is a balanced strategy
- Metrics
  - Accuracy, precision, recall, F1
  - Precision-Recall curves, ROC curve

Hands-on session

CMU Data Science Bootcamp



Name	Last Modified
audio	2 days ago
images	2 days ago
Cpp.ipynb	2 days ago
Data.ipynb	2 days ago
Fasta.ipynb	2 days ago
Julia.ipynb	2 days ago
Lorenz.ipynb	5 minutes ago
lorenz.py	5 minutes ago
R.ipynb	2 days ago

We explore the Lorenz system of differential equations:

$$\dot{x} = \sigma(y - x)$$
$$\dot{y} = \rho x - y - xz$$
$$\dot{z} = -\beta z + xy$$

Let's change (σ, β, ρ) with ipywidgets and examine the trajectories.

```
[2]: from lorenz import solve_lorenz
     interactive(solve_lorenz, sigma=(0.0,50.0), rho=(0.0,50.0))
```

sigma 10.00

beta 2.67

rho 28.00

```
def solve_lorenz(sigma=10.0, beta=8./3, rho=28.0):
    """Plot a solution to the Lorenz differential equations."""
    fig = plt.figure()
    ax = fig.add_axes([0, 0, 1, 1], projection='3d')
    ax.axis('off')

    # prepare the axes limits
    ax.set_xlim((-25, 25))
    ax.set_ylim((-35, 35))
    ax.set_zlim((5, 55))

    def lorenz_deriv(x_y_z, t0, sigma=sigma, beta=beta, rho=rho):
        """Compute the time-derivative of a Lorenz system."""
        x, y, z = x_y_z
        return [sigma * (y - x), x * (rho - z) - y, x * y - beta * z]
    # Choose random starting points, uniformly distributed from -15 to 15
    np.random.seed(1)
    x0 = -15 + 30 * np.random.random((N, 3))
```

Thank you!