Course Introduction

Sprint 1 - Week 1

INFO9023 - Machine Learning Systems Design

Thomas Vrancken (t.vrancken@uliege.be)
Matthias Pirlet (matthias.pirlet@uliege.be)

2025 Spring

Agenda

What will we talk about today.

Lecture

1. Introduction to the staff
2. Introduction to ML Systems Designs & MLOps
3. Key concepts of MLOps
4. Roles & organisation of ML projects
5. Real world use cases
6. Course organisation
7. Use case definition framework

Introduction to the staff

LIÈGE université

Introduction to the staff

Thomas Vrancken

(Instructor)

t.vrancken@uliege.be

Matthias Pirlet

(Teaching assistant)

matthias.pirlet@uliege.be

Our experience & expertise make us leading AI specialists.

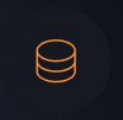
EXCEPTIONAL TALENT & SKILLS

110+ experts spread over **3** different EU locations.

Known for technical expertise
Loved for our business results

Talent magnet: 16 applications each day

STATE-OF-THE-ART TECH KNOWLEDGE & ASSETS



6 Mio downloads per month of our open source packages

150+ clients, **300+** projects, **3** spin-offs
17% of time in R&D, **250+** publications,
15 awards, avg time to value **11.5** weeks

Security, Legal and **Ethical** AI experts

We work with customers across industries and geographies.

ML6 - your partner in AI.

We accompany organisations through their entire AI journey

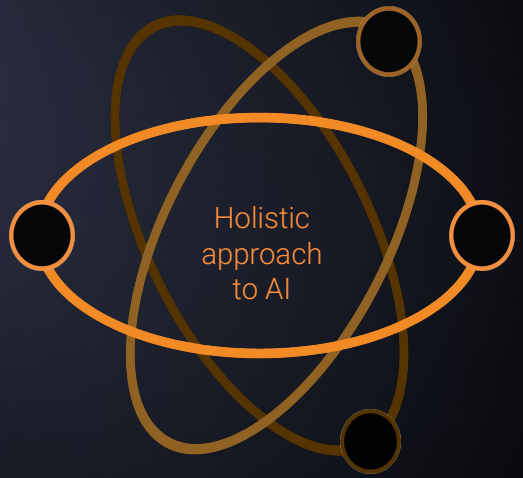
- Use case definition & assessment
- AI solution design, engineering and deployment
- Managed support and maintenance services
- Production-level scaling & evolution of AI solutions

We help remove barriers to technology adoption

- Security
- Ethics & Regulation
- Business case building
- Selecting the right tech stack
- Facilitating user adoption

We cover all AI domains

- Machine Vision
- NLP
- Structured Data
- Reinforcement Learning & Generative AI
- MLOps & Engineering best-practices



We engineer bespoke AI solutions

- Tailored to complex client needs
- Agile development & use of boiler plates where relevant
- Reliable, robust & maintainable solutions

We deliver end-to-end

- Data labelling
- Sourcing of internal and external data
- Hardware selection and/or integration (incl. IOT & edge devices)
- Front-end development

We are technology-agnostic

- Cloud agnostic: AWS, Azure, GCP
- Open source minded
- Tech radar for stack selection
- Hybrid cloud - on premise; and edge deployment

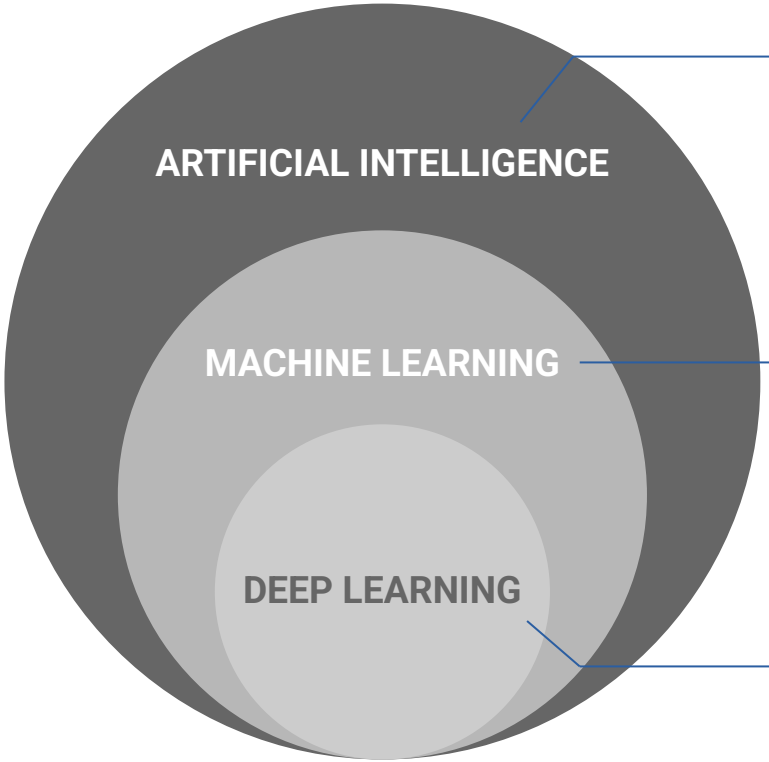
We are recognized as leaders by the industry.

Don’t just take our word for it

FT 1000 Europe’s Fastest Growing Companies	Data News	Trends Gazellen 2022	Deloitte. Technology Fast 50	Google Cloud	C	EY
#386 (EU) | #4 BE	Nominated in 2023, 2022, 2021	Multiple nominations & one award win	Nominated in 2022, 2021, 2020, 2019, 2018	Multiple nominations & award wins in 2022, 2015	Multiple nominations & award win in 2023	Scale-up of the year finalists in 2023

General introduction to ML Systems Design & MLOps

AI vs ML vs DL



ARTIFICIAL INTELLIGENCE

Ability of a machine to perform cognitive functions we associate with human minds, such as perceiving, reasoning, learning, problem solving, and even creativity

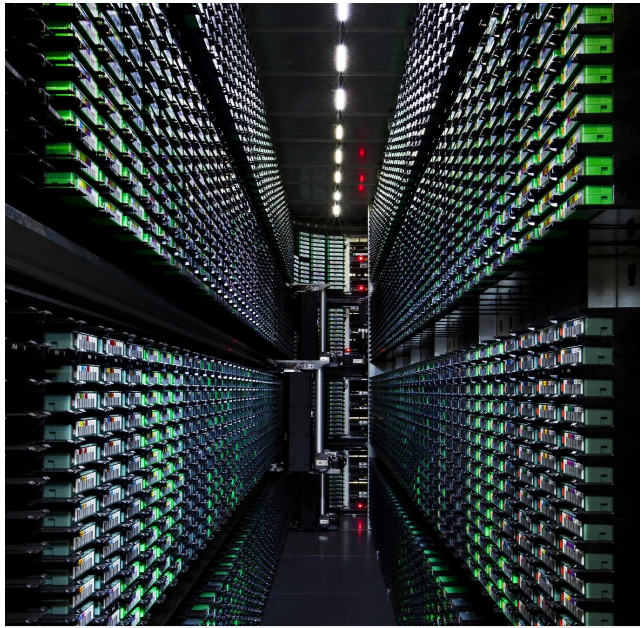
MACHINE LEARNING

AI techniques that give machines the ability to learn from data without being explicitly programmed, i.e. to automatically improve through experience

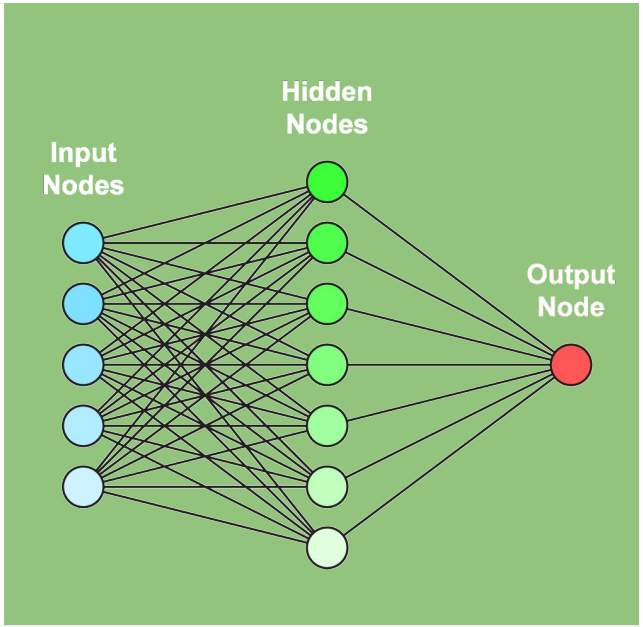
DEEP LEARNING

Type of Machine Learning built upon the concept of interconnected layers known as “neurons” that form a neural network.

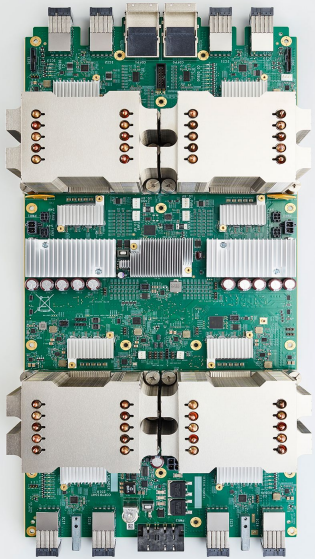
Why now ?



Large Datasets

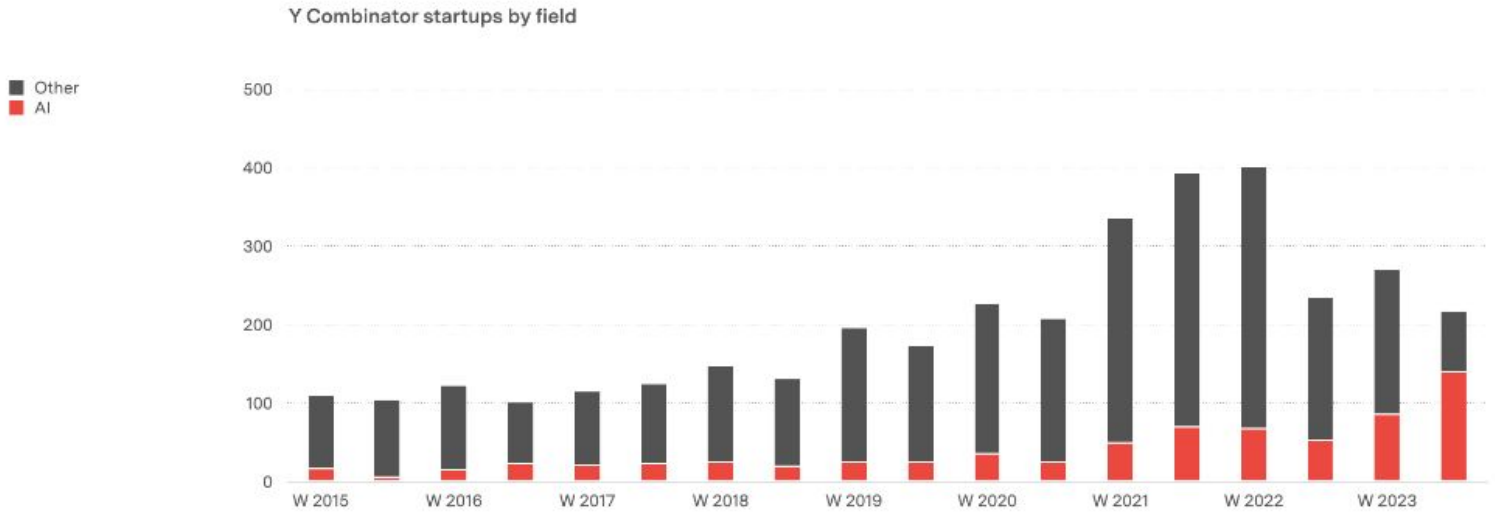


Better Models



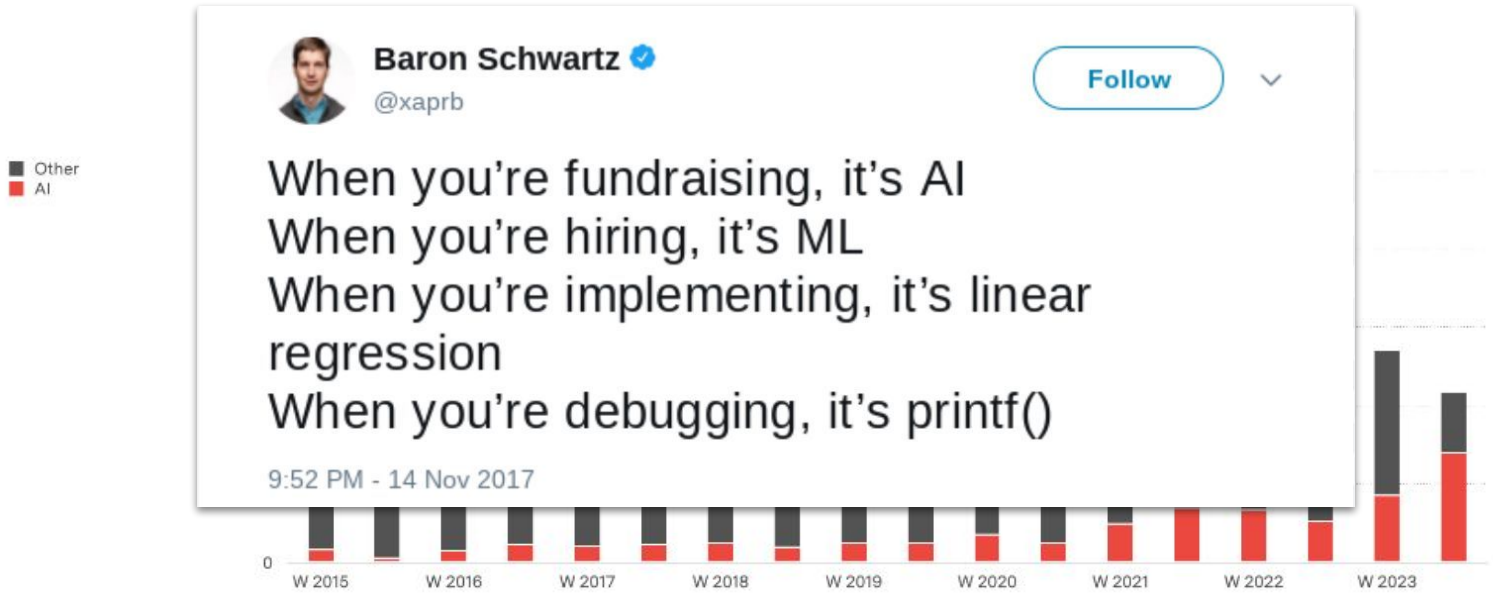
Lots of Computation

Investment in AI ventures is skyrocketing.



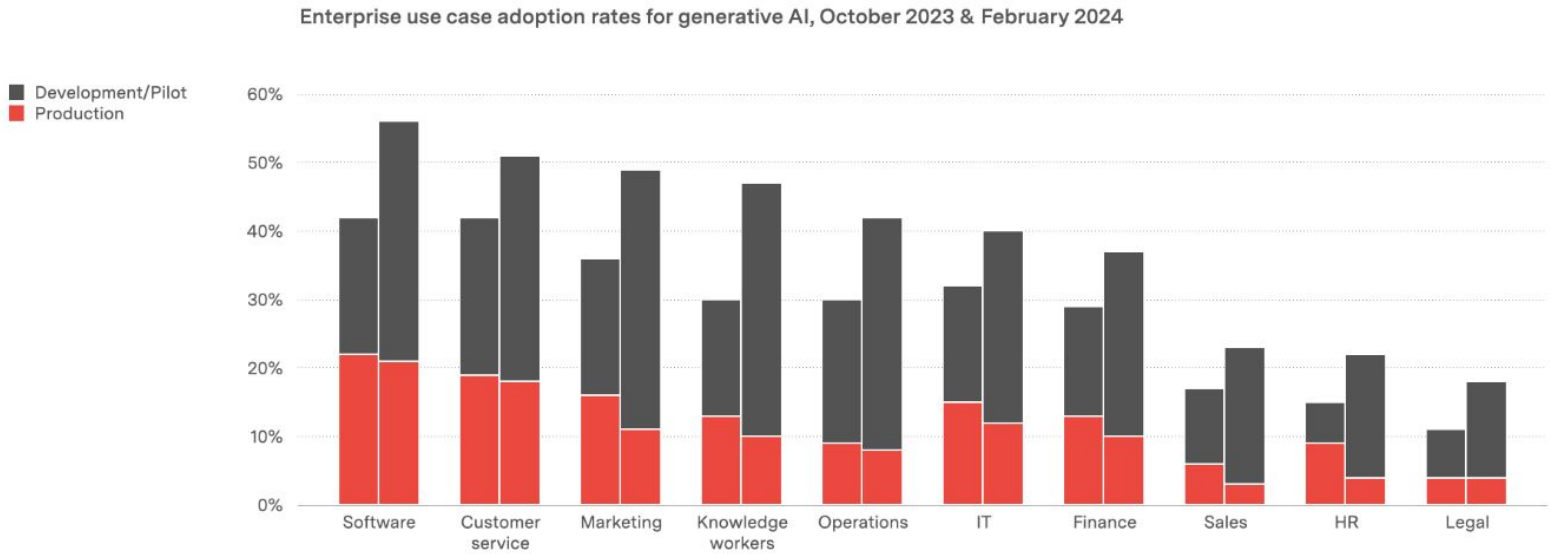
https://www.ben-evans.com/presentations [2023] - Referenced from Y Combinator.

Investment in AI ventures is skyrocketing.



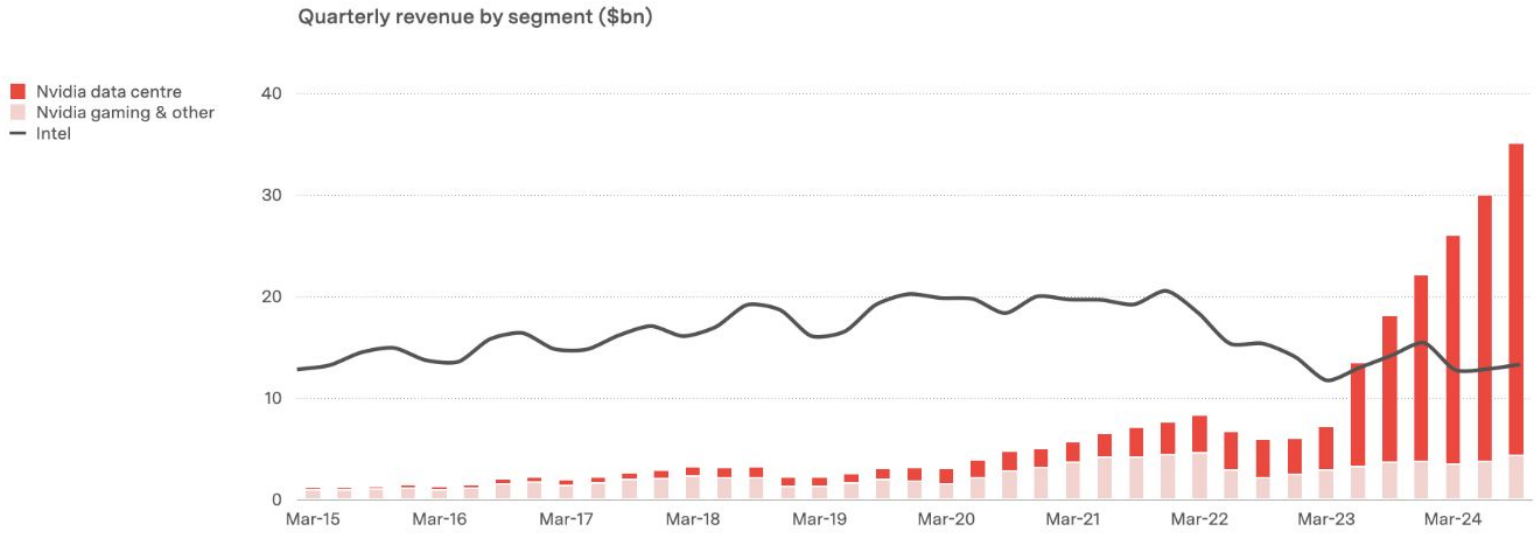
https://www.ben-evans.com/presentations [2023] - Referenced from Y Combinator.

Companies have a large interest for AI, but often struggle with getting their use cases to production.



https://www.ben-evans.com/presentations [2024] - Referenced from Bain.

Demand for AI riples down to other segments, such as Cloud infrastructure.



https://www.ben-evans.com/presentations [2024]

Future innovations & impact with AI might not be in training better models but rather in applying them and making them more efficient.

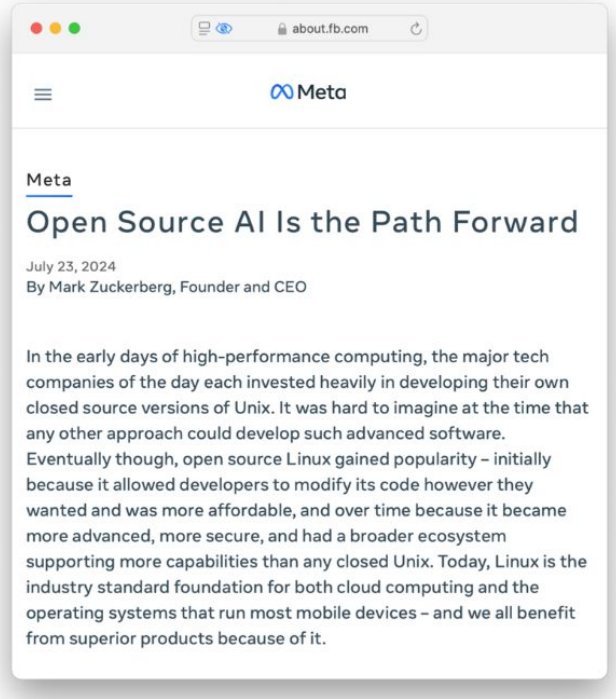
https://www.ben-evans.com/presentations [2024]

The “API model” might not last forever.

“Everyone in tech is giving someone else’s business model away for free”

Meta’s open source

Turn models into commodity infrastructure!



https://www.ben-evans.com/presentations [2024]

AI is everywhere!

A few example of ML applications.

Facial recognition

Product recommendation

Email spam filtering

Autocomplete

Finance predictions

Healthcare imaging

Weather forecast

...

Why do we need ML Systems Design?

Building a ML application means implementing much more than just your ML model.

INFO 9023 - Machine Learning Systems Design

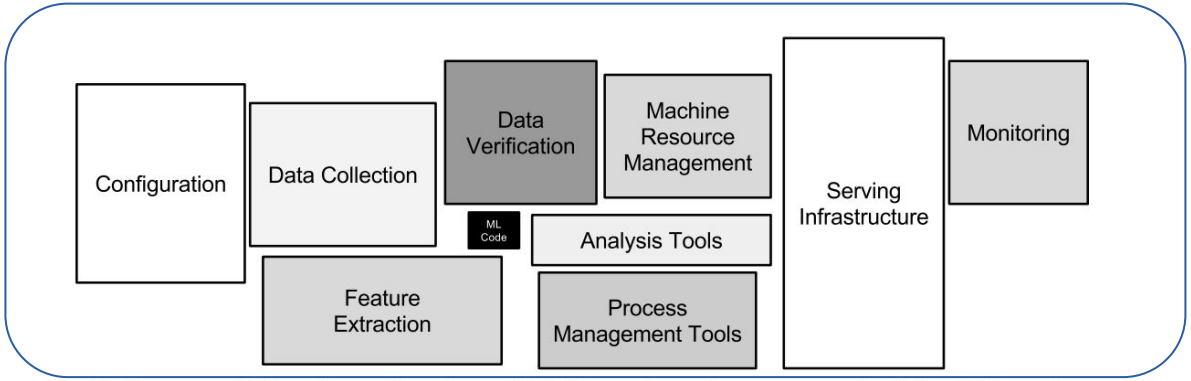
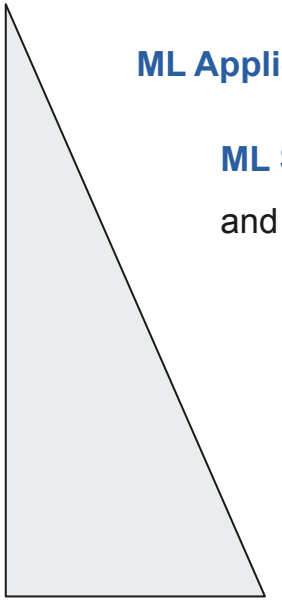


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Sculley, D. et al. (2015). Hidden technical debt in machine learning systems. https://papers.nips.cc/paper_files/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html

Important definitions



ML Application: The final solution or program powered by a Machine Learning model.

ML System: All the components responsible for the implementation and management of the data and models powering an ML application.

ML Systems Design: The act of designing the architecture and implementing an ML System.

MLOps: Set of practices that aim at implementing and maintaining ML systems in production reliably and efficiently.

Going beyond the notebook

You were tasked with implementing a model to predict electricity productions of a wind turbine farm.

What you did:

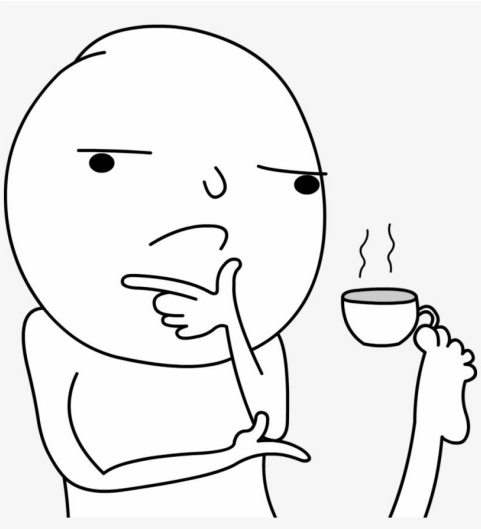
- **Export** key data for the last 10 years (electricity production, weather, …)
- **Analyse** the data and build **features**
- **Train** and **optimise** a ML models
- Make and visualise **predictions**

Going beyond the notebook

You were tasked with implementing a model to predict electricity productions of a wind turbine farm.

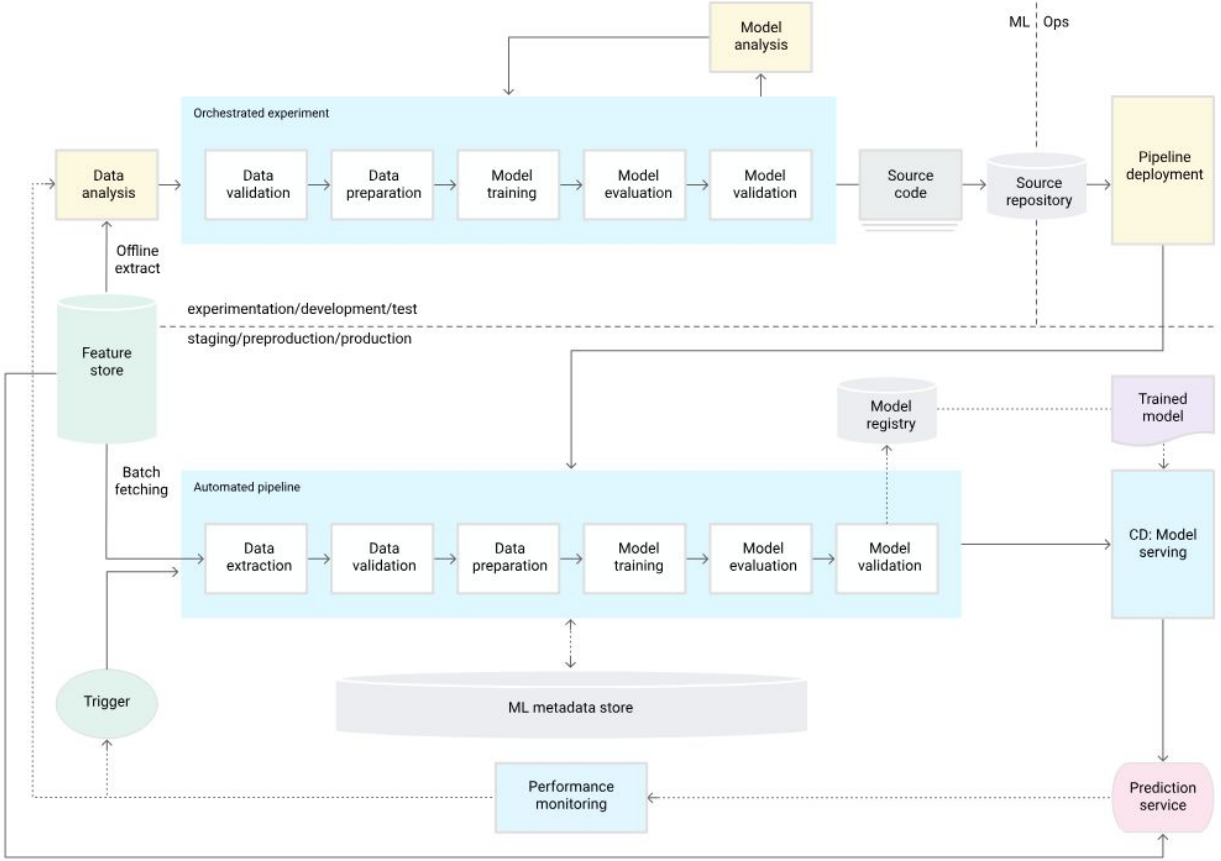
Now what?... How will you:

- Automatically do predictions every X minutes with new data
- Automatically retrain the model with new data
- Collaborate with a larger team
- Continuously integrate new changes to your application
- Not run your model locally on your own laptop
- Optimise your model so it does not consume too many resources
- Monitor your models
- …



Key concepts of ML Systems Design

Typical architecture of an ML system



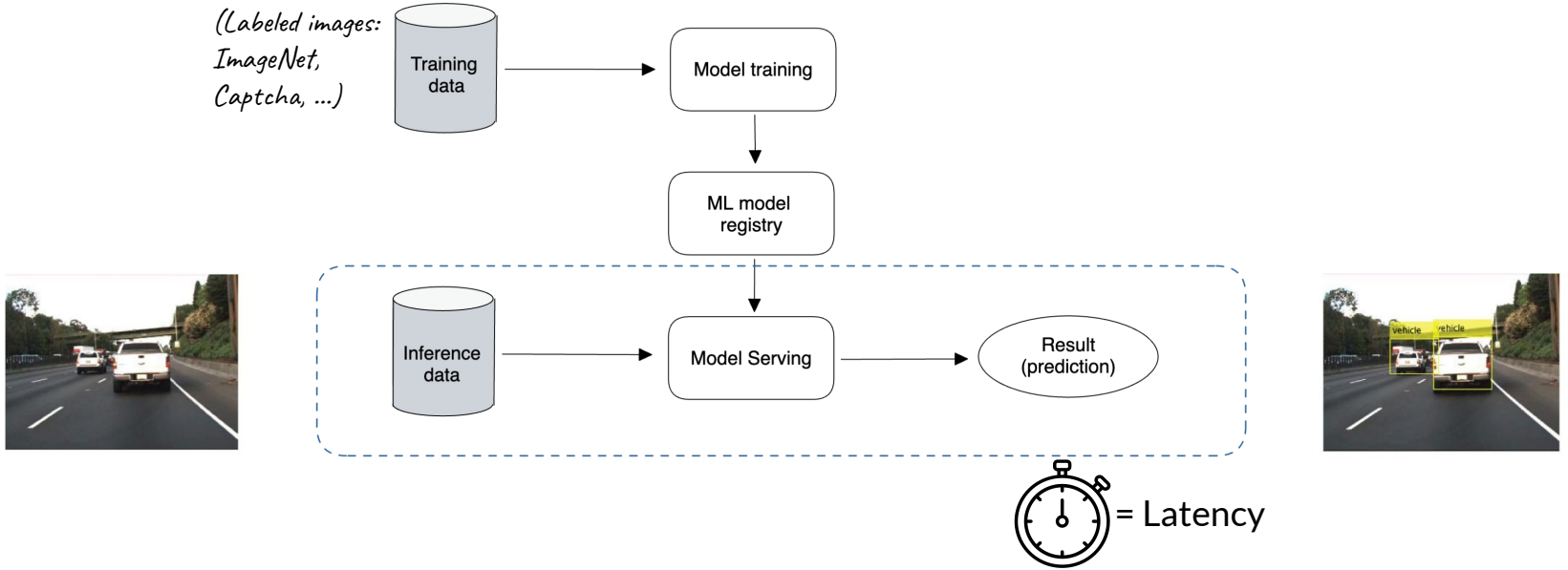
https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

Key concept: Data preparation

It all starts with data. How to go through all these steps efficiently and effectively.

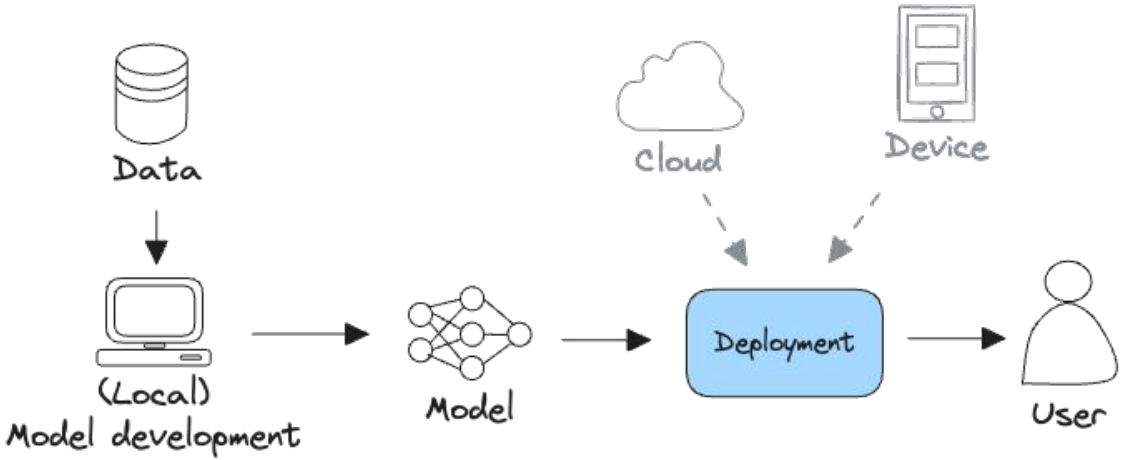
Key concept: ML model serving

How to efficiently serve ML model to client.



Key concept: ML model deployment

How to efficiently deploy your model for serving.



Key concept: Containerisation

Containers encapsulate an application as a **single executable package** that contains all the information to **run it on any hardware**:

- Application code
- configuration files
- libraries
- dependencies

Abstracts the application from its **host operating system.**

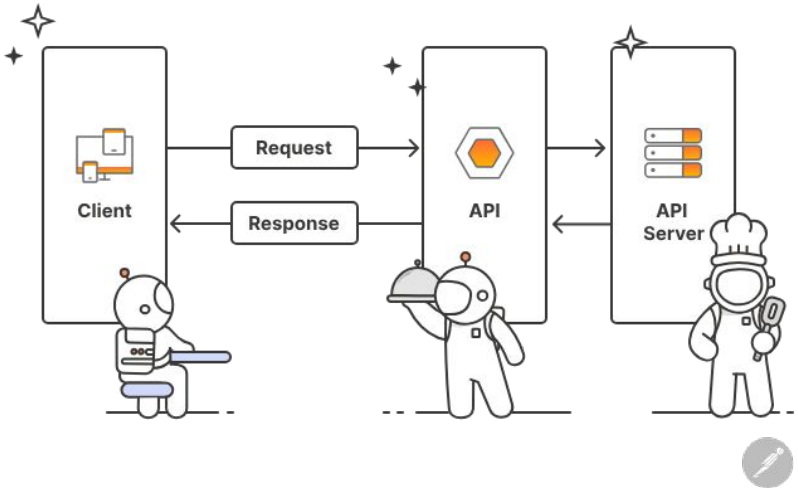
Containers can be easily transported from a desktop computer to a virtual machine (VM) or from a Linux to a Windows operating system, and they will run consistently on virtualized infrastructures or on traditional “bare metal” servers, either on-premise or in the cloud.

LIÈGE université

https://www.ibm.com/topics/containerization

Key concept: APIs

Allow other services to call your model or application.



An **Application Programming Interface (API)** is a set of protocols that enable different software components to communicate and transfer data.

Developers use APIs to bridge the gaps between small, discrete chunks of code in order to create applications that are powerful, resilient, secure, and able to meet user needs.

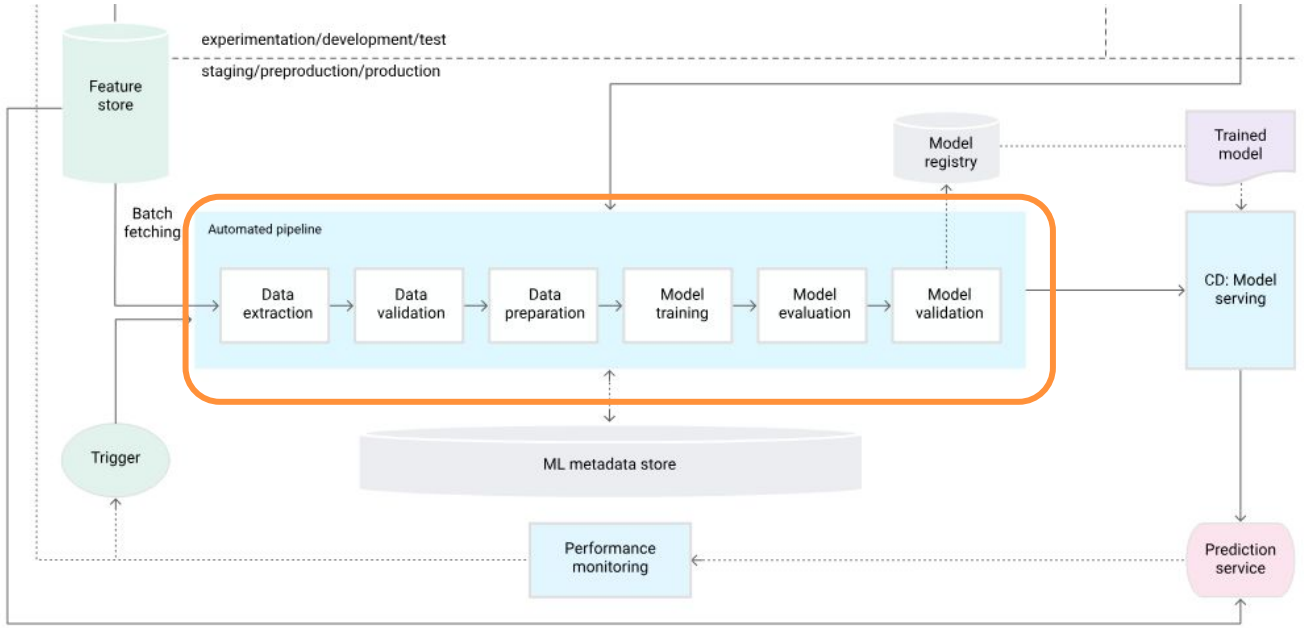
https://www.postman.com/what-is-an-api/

Key concept: Cloud infrastructure

Cloud infrastructure allow for data storage, compute allocation, training and deploying model, monitoring, …

Key concept: ML Pipeline

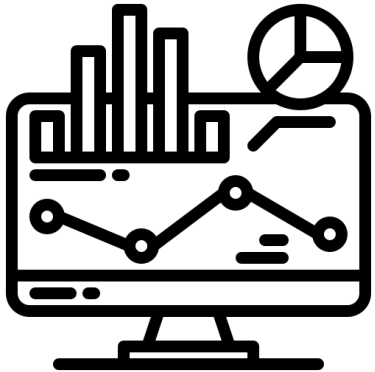
Orchestrates components to prepare data, train, evaluate and deploy ML models (among other things)



https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

Key concept: Monitoring

Ensuring that models in production are performing well.



Resource level (performance and usage of resources used by the model serving)

- How much is it being used by users?
- Are the CPU, RAM, network usage, and disk space as expected?
- What are the Cloud costs?
- Are requests being processed at the expected rate?
- What is the system uptime? Some maintenance contract depend on it.

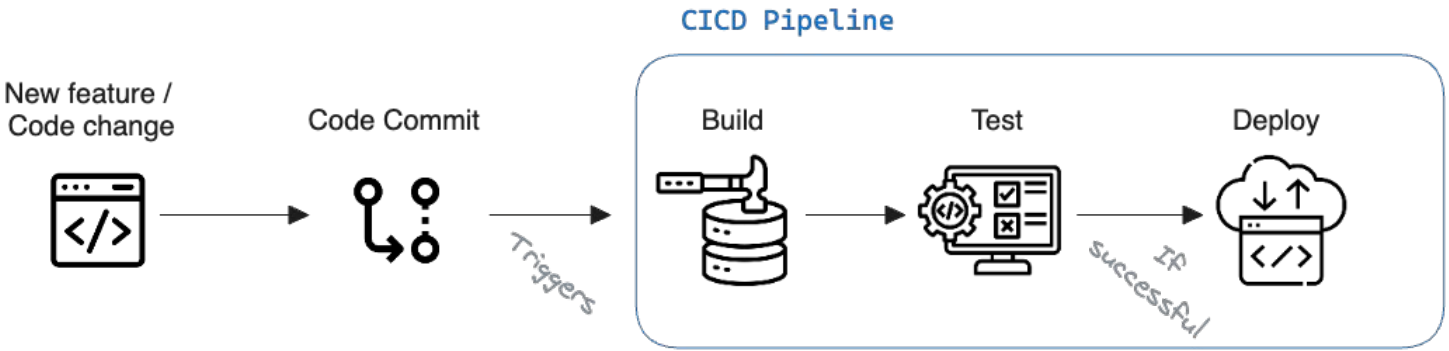
Performance level (performance/accuracy of the model over time)

- Is the model still doing accurate predictions with the new data coming in?
- Is the data distribution changing?
- Is the target variable changing?
- Are concepts around the model changing?

Key concept: CICD

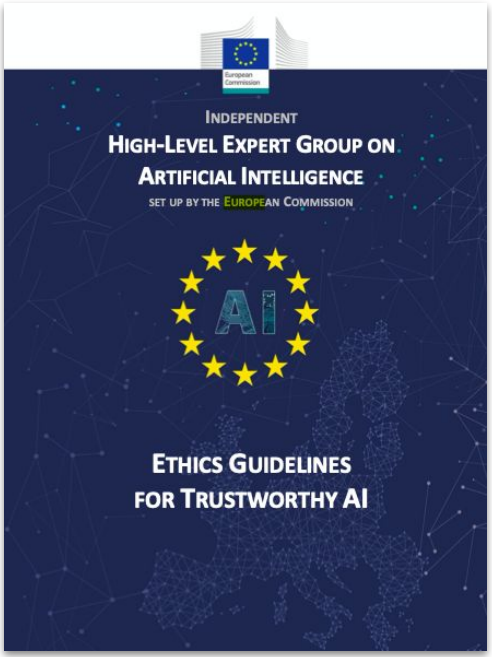
Allows you to continuously work on your application and efficiently deploy new changes to it.

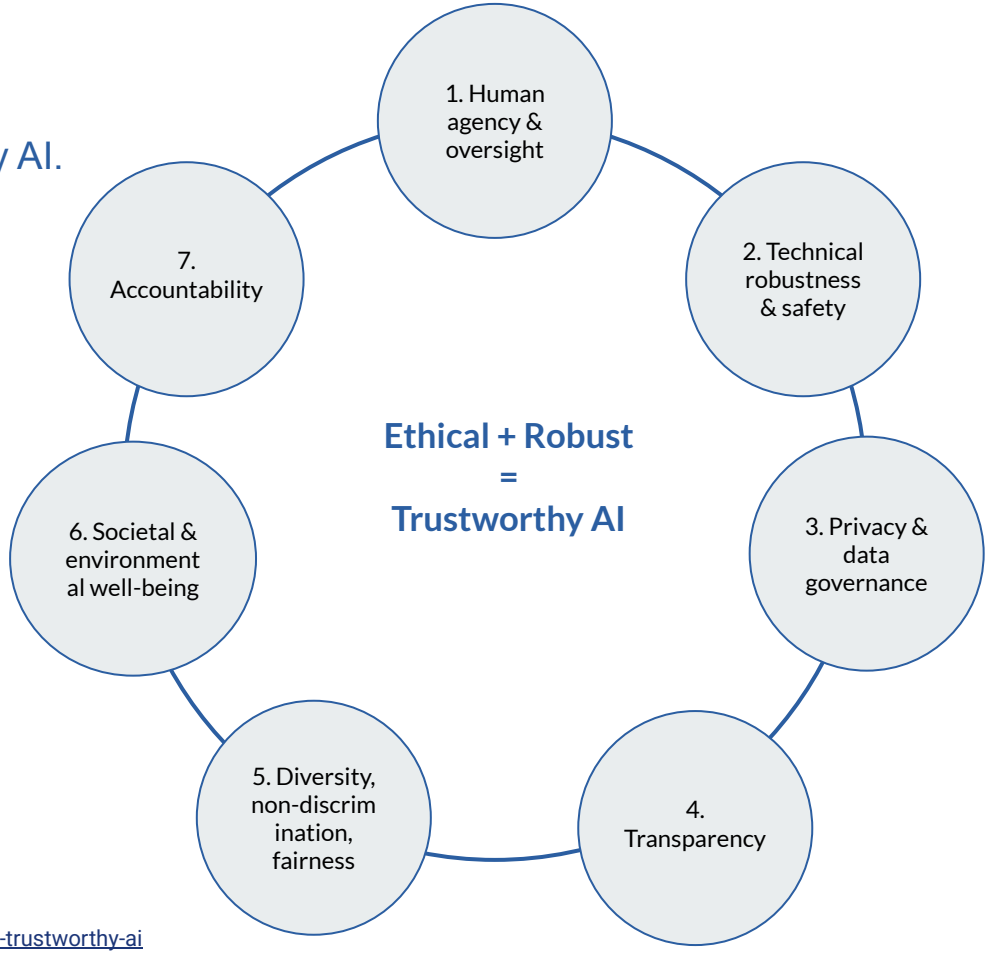
Continuous Integration and Continuous Delivery.



Key concept: Ethical AI

Guidelines & legislation on building trustworthy AI.

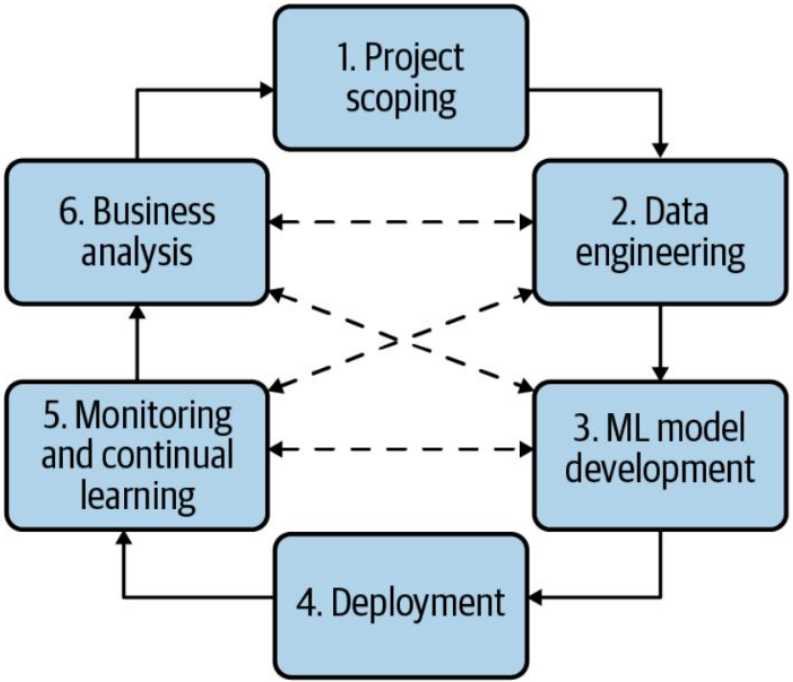




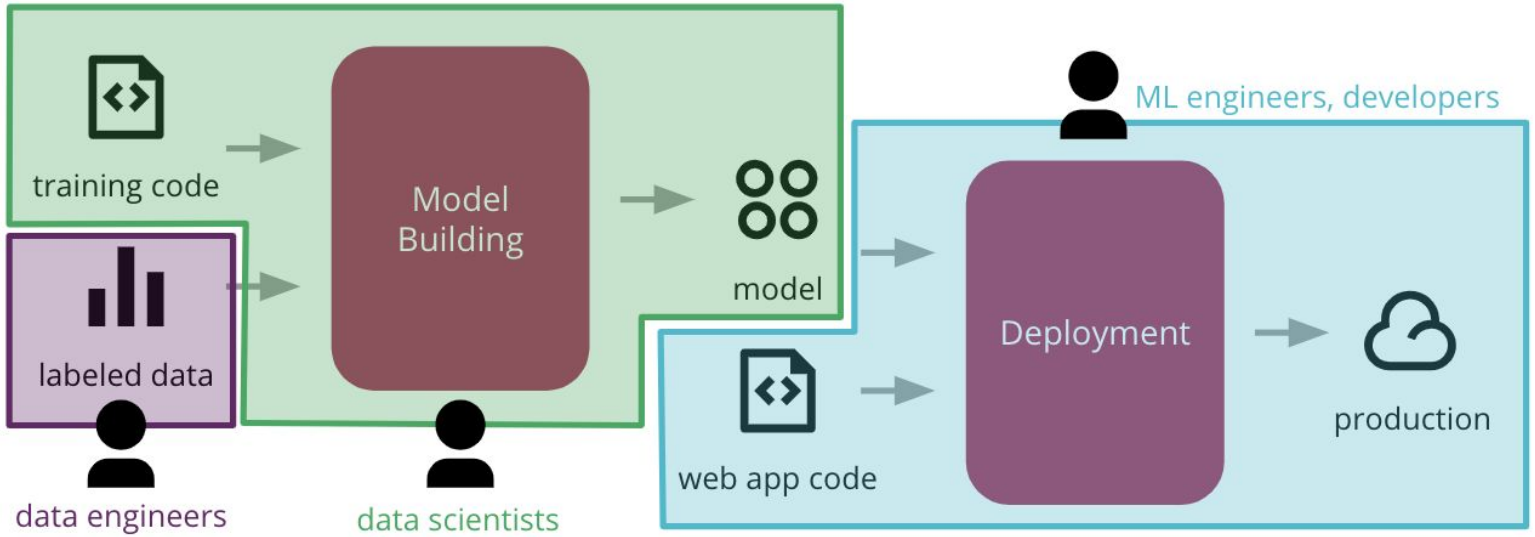
https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Roles & organisation of ML projects

Typical ML project lifecycle

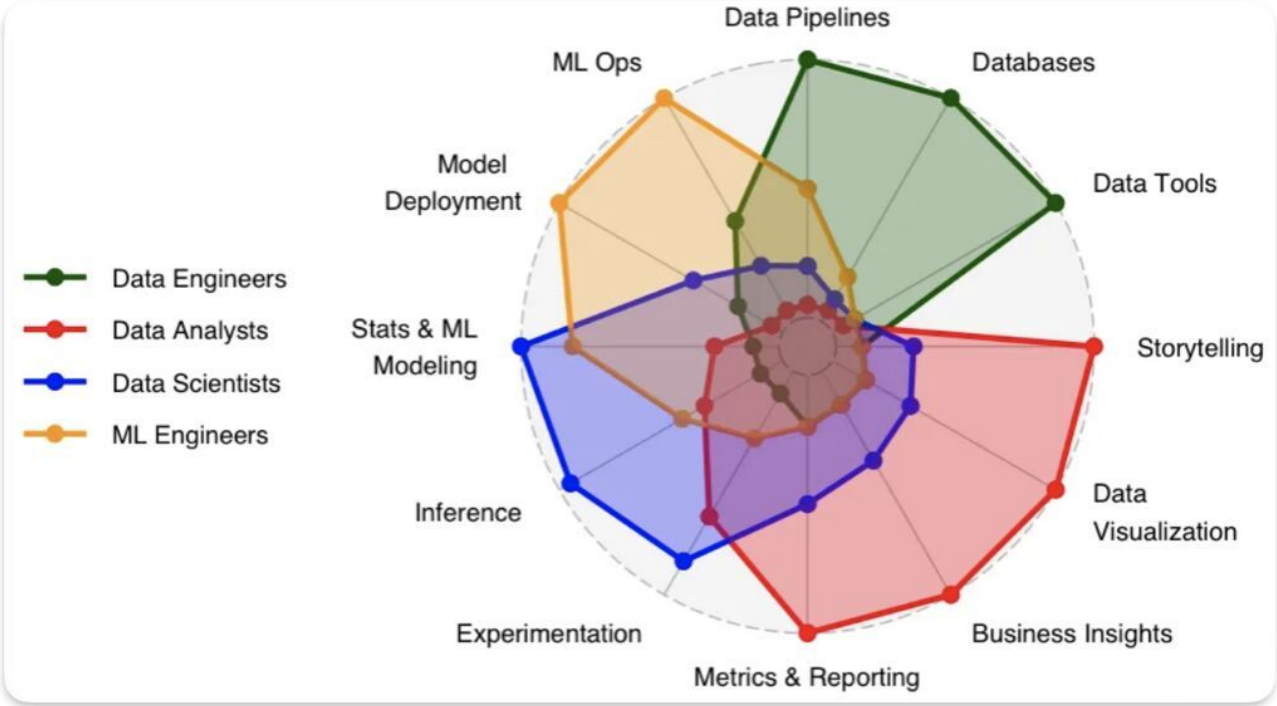


Roles around a ML system implementation.



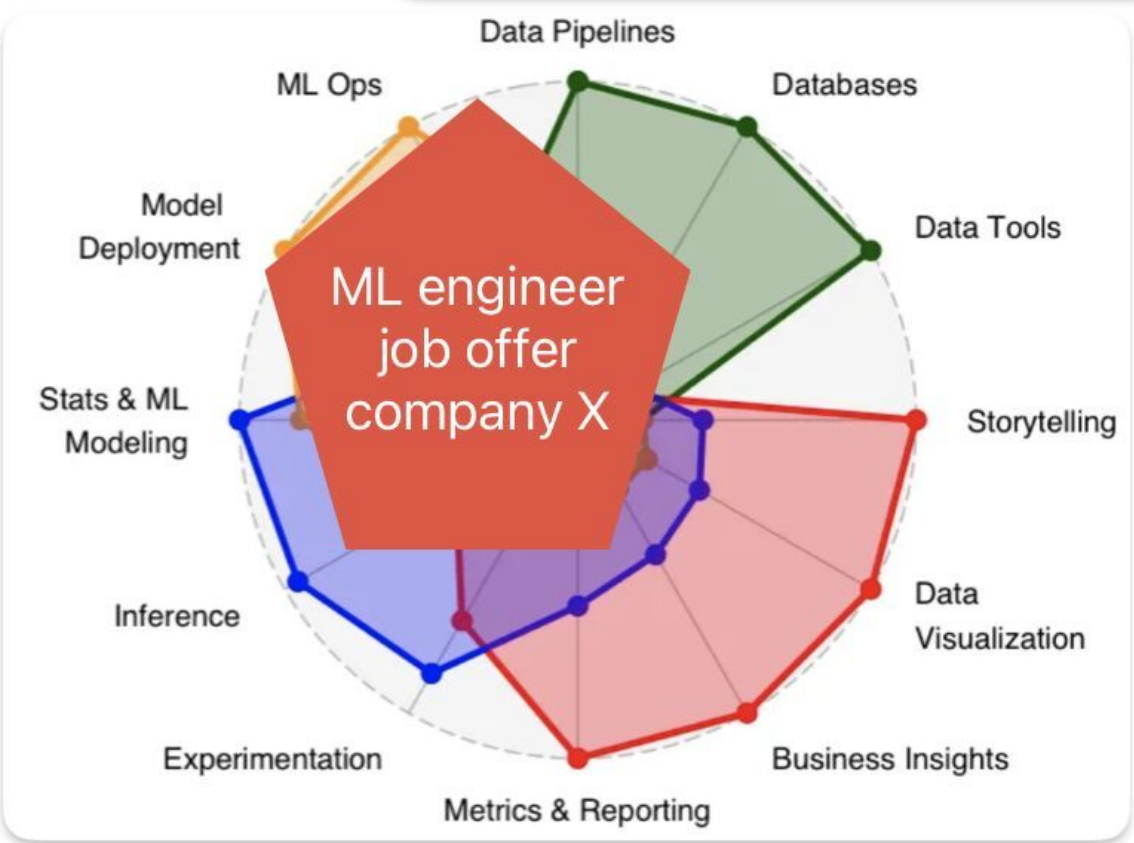
https://martinfowler.com/articles/cd4ml.html

Different set of skills per roles



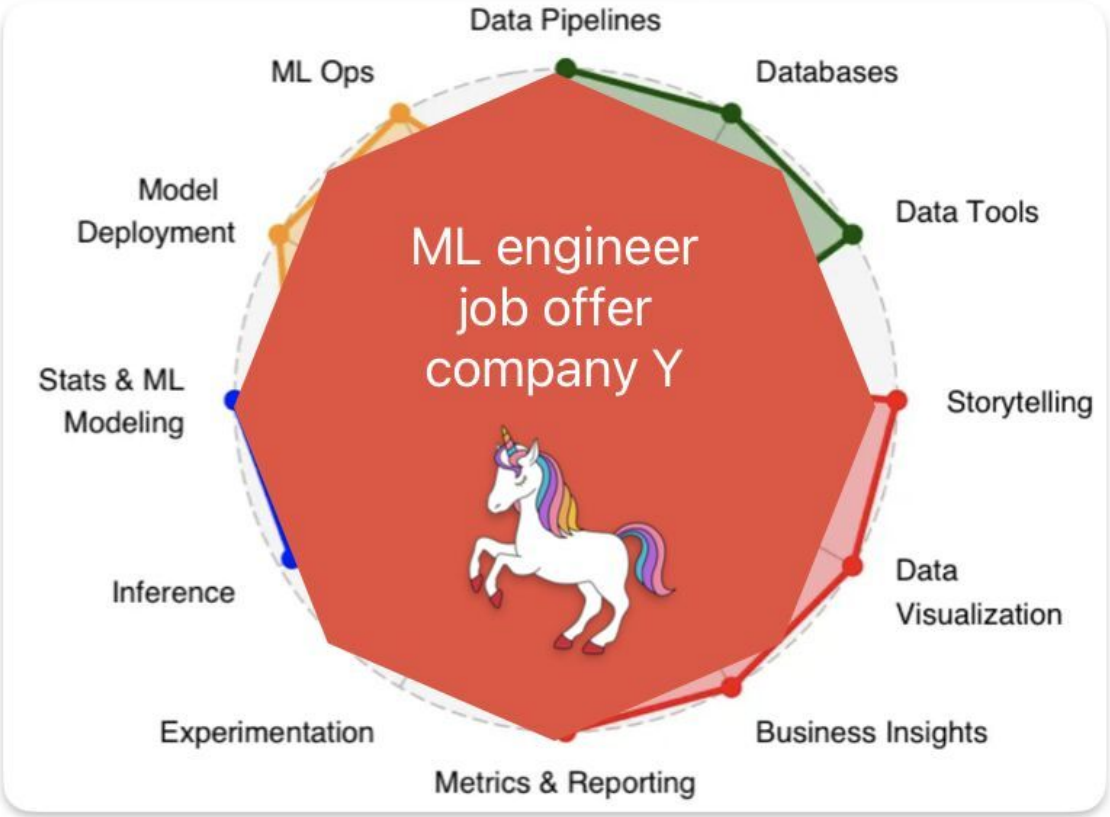
https://www.linkedin.com/posts/maria-vechtomova_data-ai-mlops-activity-7125095266403143680-9X-U/

In reality it's a bit blurry



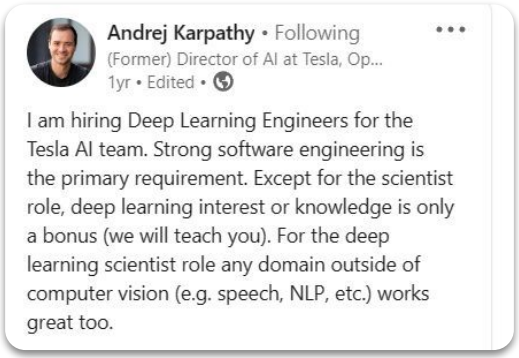
https://www.linkedin.com/posts/maria-vechtomova_data-ai-mlops-activity-7125095266403143680-9X-U/

In reality it’s a bit blurry



https://www.linkedin.com/posts/maria-vechtomova_data-ai-mlops-activity-7125095266403143680-9X-U/

ML Engineering skills are in high demand



Teams can adopt different MLOps maturity levels

Level	Highlights	Technology
Level 0 **No MLOps**	• Difficult to manage full ML model lifecycle • Teams are disparate and releases are painful • "black boxes," little feedback during/post deployment	• Manual training, builds and deployments • Manual testing of model and application • No centralized tracking of model performance
Level 1 **DevOps but no MLOps**	• Releases are less painful than No MLOps • Limited feedback on how well a model performs in production • Difficult to trace/reproduce results	• Automated builds • Automated tests for application cod
Level 2 **Automated Training**	• Training environment is fully managed and traceable • Easy to reproduce model • Releases are manual, but low friction	• Automated model training • Centralized tracking of model training performance • Model management
Level 3 **Automated Deployment**	• Releases are low friction and automatic • Full traceability from deployment back to original data • Entire environment managed: dev > test > production	• Integrated A/B testing of model performance • Automated tests for all code • Centralized tracking of model training performance
Level 4 **Full MLOps**	• Full system automated and easily monitored • Automated feedback collection and retraining • Close to zero-downtime	• Automated model training and testing • Verbose, centralized metrics from deployed model

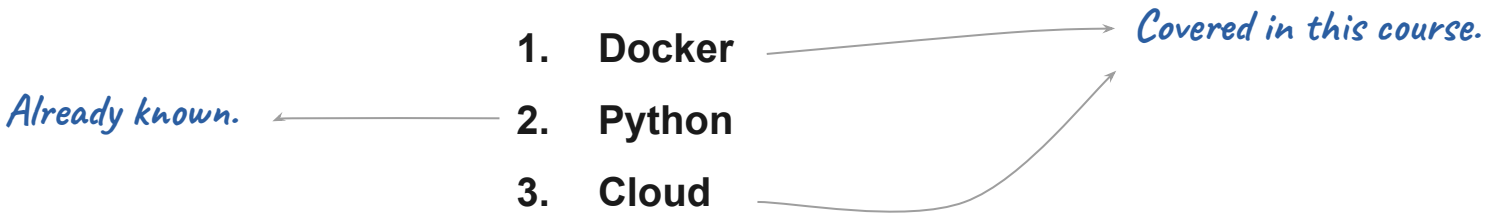
LIÈGE université

https://learn.microsoft.com/en-us/azure/architecture/ai-ml/guide/mlops-maturity-model#level-0-no-mlops

Study on demanded skills for MLOps engineers.

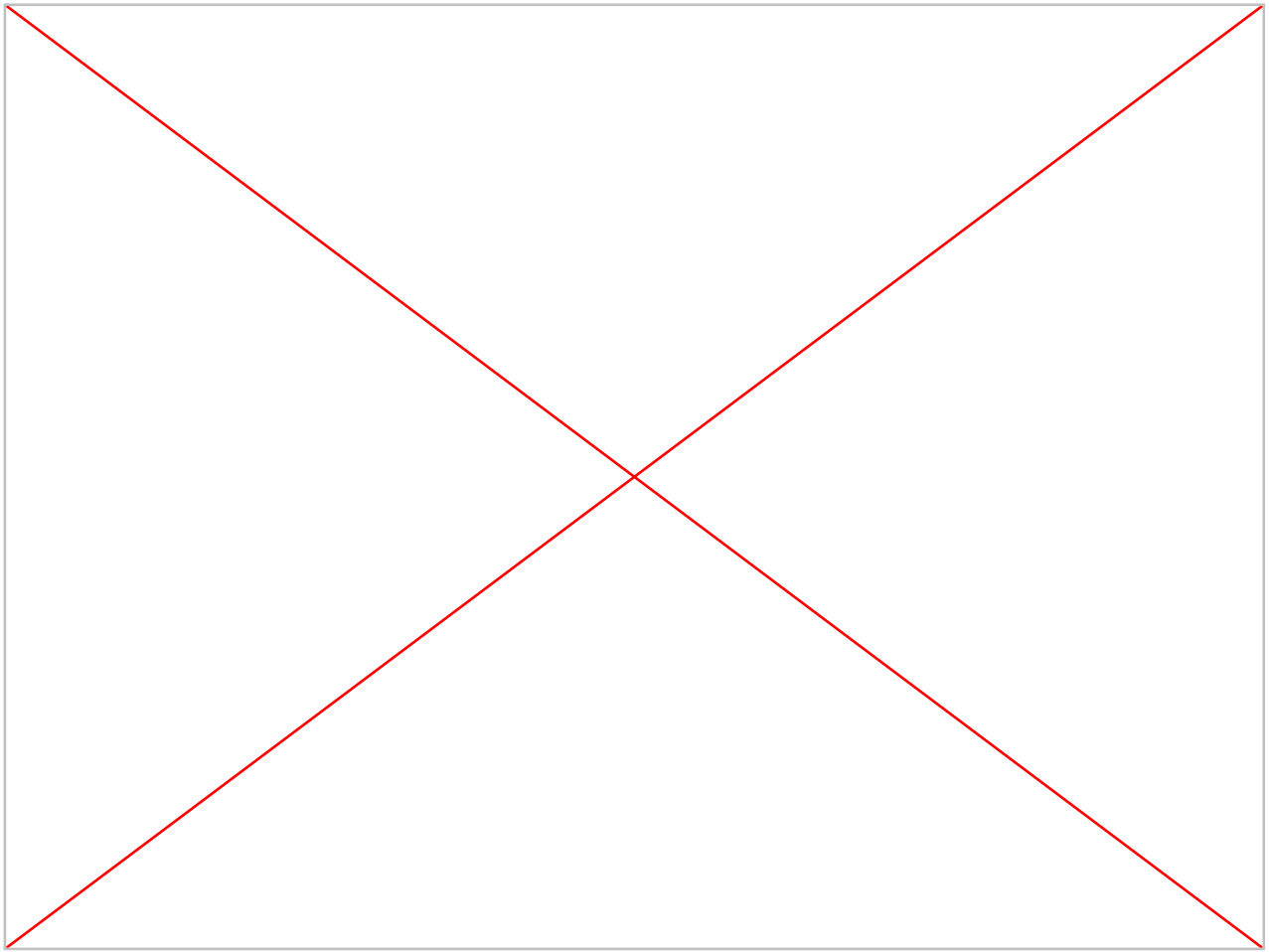
Looking at 310 job offers on MLOps in Q4 2023.

Top 3 highest demanded skills:



https://marvelousmlops.substack.com/p/what-do-teams-really-want-from-an?utm_source=profile&utm_medium=reader2

Going from standard ML Engineer to MLOps master…



Project phases & challenges

Build different stages of your solution

Proof of Concept

Use easily available data to show that your model or solution can work.
Low efforts.
Prove the feasibility and value.
Iterate fast.

Minimum Viable Product

Just enough features for a small set of users to start using it.
Gather feedback and make sure that it is designed in an optimal way.

Productionisation / scaling

Build the infrastructure to finally deploy your solution and let users use it.
Gradual roll-out to more and more users in more and more markets.
Deploy better models, attract more users, go to new markets, maintain the solution, …

Maintenance

Keep the solution up and running.
Monitor resources and performance.
Update packages and dependencies (software around solution change).
Security and up-time.

Build different stages of your solution

Use eas
to show
or solut
Low eff
Prove t
value.
Iterate

POC

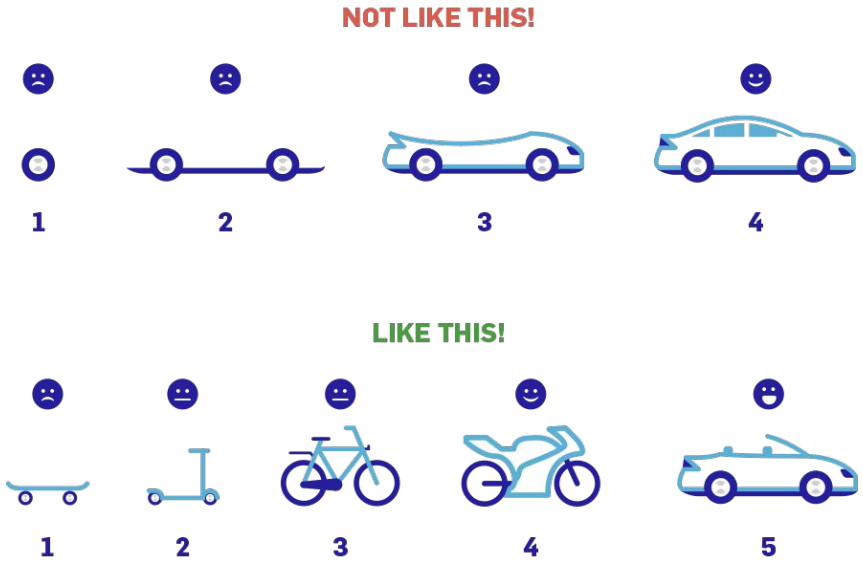
2 weeks

tenance

...

it's up...

At each stage, your product should be usable



Data science projects are challenging to bring to production

Breaking the myth

Forbes

VentureBeat

*"**87%** of data science projects never make it into production…"*

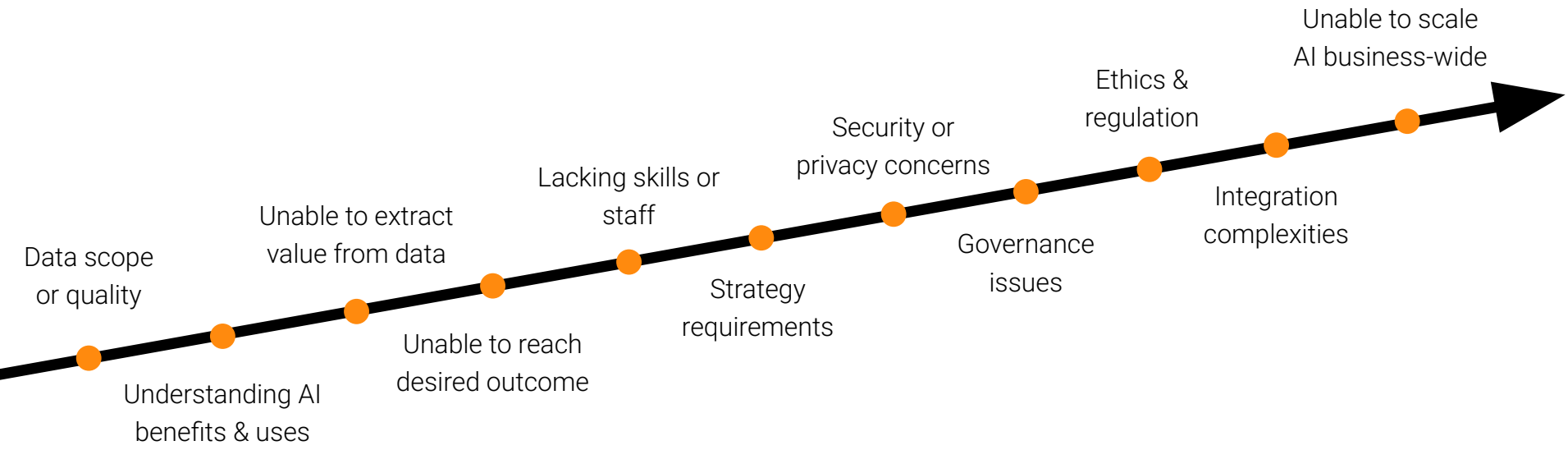
(Might not be a factual number...)

But data science project are still challenging to actually roll-out to the real world!

https://mtszkw.medium.com/why-do-87-of-data-science-projects-fail-and-are-we-sure-that-it-is-true-fe8b5ba1404c

AI Journey Challenges.

While AI is an enabler for strategic priorities, it doesn't come without its challenges.



LIÈGE université

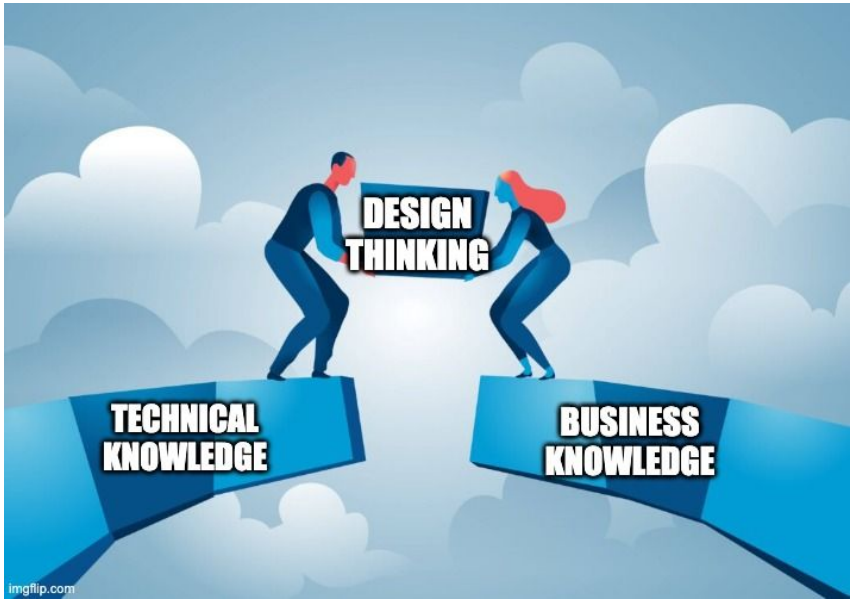
Project definition framework

LIÈGE université

Process to define new use cases.

How to get started?

- **New ideas** do not come spontaneously
- Proactively organise **workshops** to identify how ML can create value in an organisation.
- Use **design thinking** techniques.
- Make sure to have the **right people around the table** (decision makers, stakeholders, users and (ofc) engineers).
- Spend enough time in it - **starting in the right direction** is key.



Other concept: Design thinking

Same ideas, different framework (coming from front-end engineering)

1. Empathize

Engage in qualitative research methods such as interviews and workshops to deeply understand the users, their needs, and their pain points.

2. Define

Clearly articulate the user's needs and challenges based on the insights gathered during the empathize phase. Map out the user's interaction with the solution.

3. Ideate

Engage in collaborative sessions to generate a wide range of solutions and ideas.

4. Prototype

Build a **mock** application to validate whether it fits your users needs.

5. Test

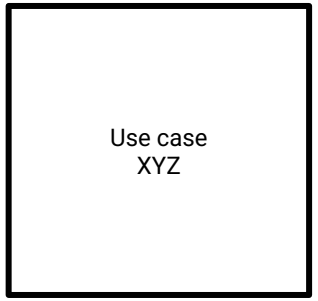
Monitor user interactions and gather data to measure the application's success. Maintain an ongoing feedback loop with users to continually refine and improve the application.

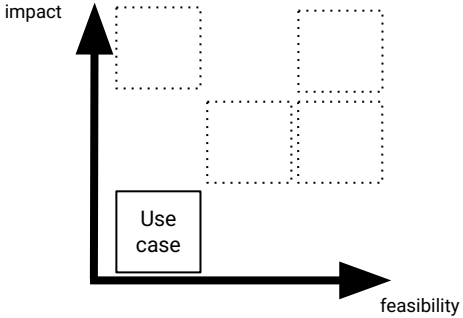
6. Implement

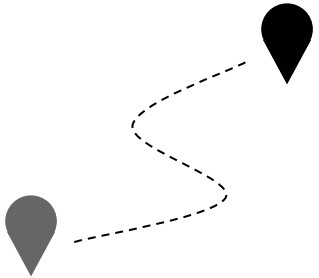
Once the design is finalized, begin the development process using appropriate technologies and frameworks.

LIÈGE université

Framework to define an AI use case.







1 Identify AI opportunities

2 Evaluate and refine selected use cases and their feasibility

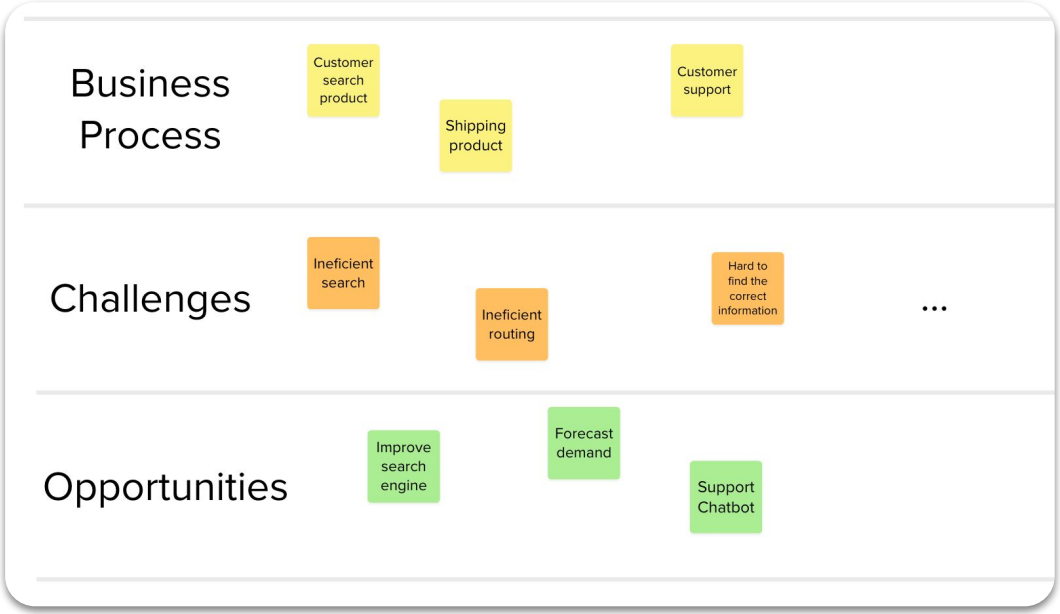
3 Prioritize top use cases to kickstart AI

4 Define the roadmap towards this AI use case

LIÈGE université

Identify opportunities

- Ideate and map user process
  - Identification of **business opportunities**
  - Identification of **challenges**
  - **Opportunities**: where can AI help?
- Cluster opportunities
- Name AI use cases



Use case template.

- How to quickly iterate over a few use cases?
- How to efficiently capture the point of view of different people?
- How to set the vision on a specific use case?

Use Case: [Cool Name]

What?

[Describe the use case in 2 sentences]

Value
[Score out 5 - flash vote]

☆ ☆ ☆ ☆ ☆

Feasibility
[Score out 5 - flash vote]

☆ ☆ ☆ ☆ ☆

Why?
[Purpose of the solution - e.g. reducing costs, helping users, climate, …]

Who?
[Stakeholders benefiting from the solution (e.g. customers, users, role X, …)

How?
[Approach, simplified]

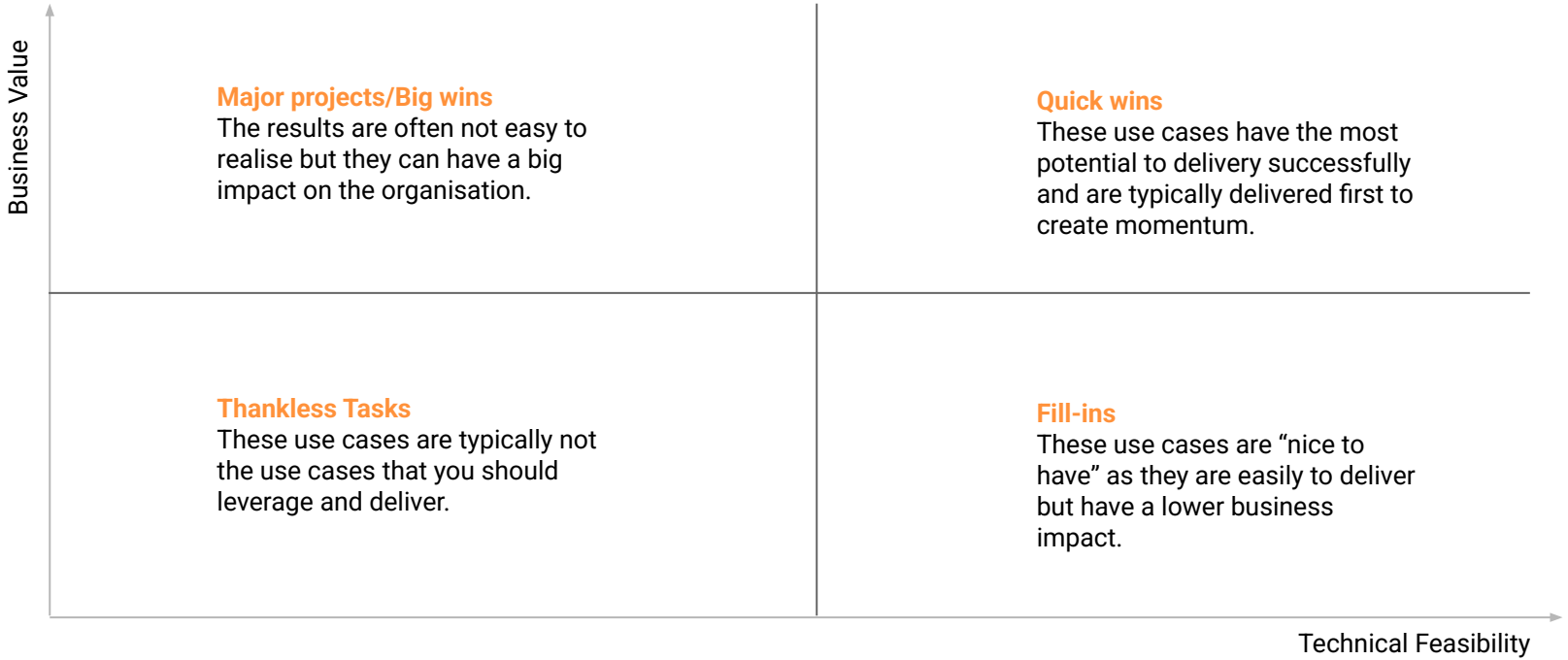
Challenges?

- …
- …
- …

Evaluation?
[Metrics and success criteria]

Prioritisation matrix.

How to evaluate the different use cases?



Define and scope your project.

Which questions to answer before getting started with the selected project? (Often done offline, after the workshop)

Define value drivers

Set success criteria

Identify challenges

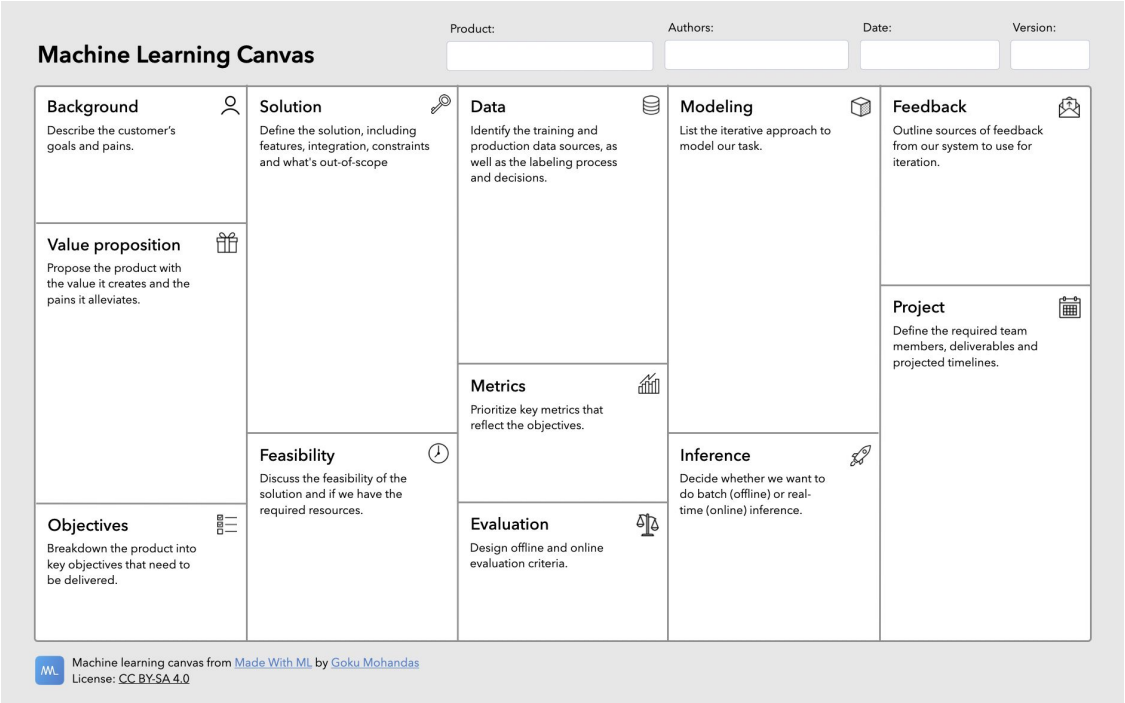
Think about intermediate milestones that show value

Define building blocks

Estimate time & budget

Define and scope your project.

Product design template

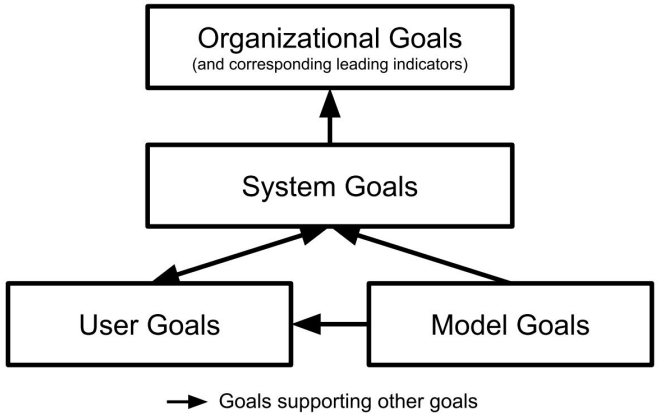


https://madewithml.com/courses/mlops/product-design/

Aligning your solution with goals on different levels.

- **Organizational goals**: Innate/overall goals of the organization.
- **System goals**: Goals of the software system/feature to be built.
- **User goals**: How well the system is serving its users, from the user's perspective.
- **Model goals**: Quality of the model used in a system, such as performance.

These goals <u>should</u> be aligned with each other



User adoption

"You can have the best model with the best data, success always depends on how users will adopt it."

Ways to ensure user adoptions:

- **Power users**: Work with users since day 1. Throughout the use case ideation and during development. You receive critical feedback and can get champions who fully understand the solution to spread its usage once developed.
- **Change management strategy**: From executives and process experts.
- **Integration**: Make sure if works with users favorite tools (a new board in existing platform has much higher chances of being utilised than a new program/website).
- **Documentation**: Clear explanation of *how the model works, performs* and *should be used*. Training program, videos, tutorials, FAQs, support line, …
- **Monitor usage**: … and improve the solution from it.

When not to use Machine Learning?

It’s not always the right solution...

- Clear specifications are available
- Simple heuristics are good enough
- Cost of building and maintaining the ML system outweighs its benefits
- Correctness is of utmost importance
- ML is used only for the hype (e.g., to attract funding)

Examples of these?

(Really) accurate predictions might not even be that important

The over-optimizing paradox

- "Good enough" may be good enough
- Prediction critical for system success or just an gimmick?
- Better predictions may come at excessive costs
  - Data is often the bottleneck
  - Cost of producing more data (labeling, infra, collection, …)
- Better user interface ("experience") may mitigate many problems
  - Explain decisions to users with Explainable AI (XAI)
- Use only high-confidence predictions?

Critical thinking when doing the project definition

Ask the right questions - make sure you have a solid use case before you start building anything.

- **Baseline**: What is the performance of an alternative to ML? How do simple heuristics or human guess-predictions perform?
- **Probabilistic**: ML is by definition not deterministic. Are probabilities/ranges fine for this use case? E.g. for demand forecasting the model can make errors, for self-driving cars not…
- **Precision / recall**: Are both important? If not, can I make it a success by sacrificing one? E.g. for fraud detection we can raise a warning on false positive, but cannot have false negative…
- **Interpretability**: Do we need to explain why the model makes specific decisions? If yes, can we?
- **Do not reinvent the wheel:** Are there existing open source or 3rd party solutions? Did anybody in my organisation work on something like this?

Real-life example of a MLOps organisation (or AI Platform)

Linkedin case study

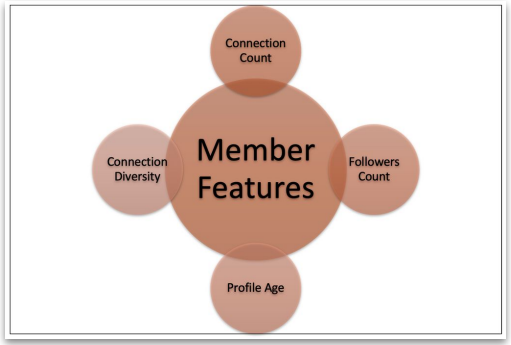
Linkedin integrates many ML applications

Viral spam content detection

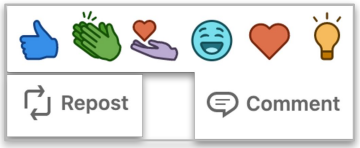
Detecting spam content…

… Using boosted tree algorithm on the following features:

Post features



Member features



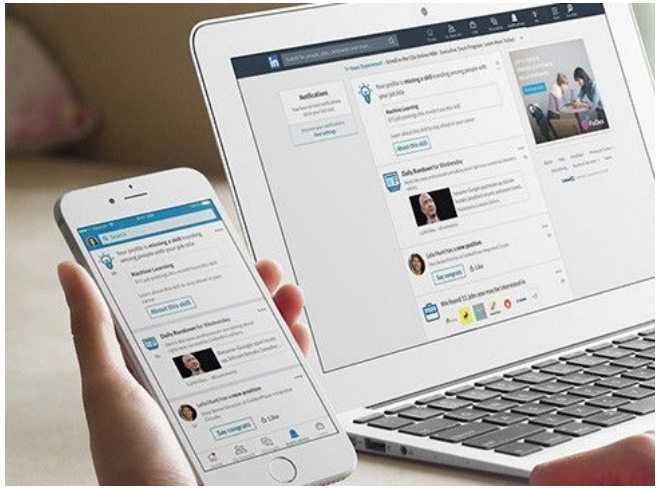
Engagement features

https://www.linkedin.com/blog/engineering/trust-and-safety/viral-spam-content-detection-at-linkedin

Linkedin integrates many ML applications

Personalised LinkedIn News Feed

Select personalised content for users…



… Using boosted tree algorithm on the following features:

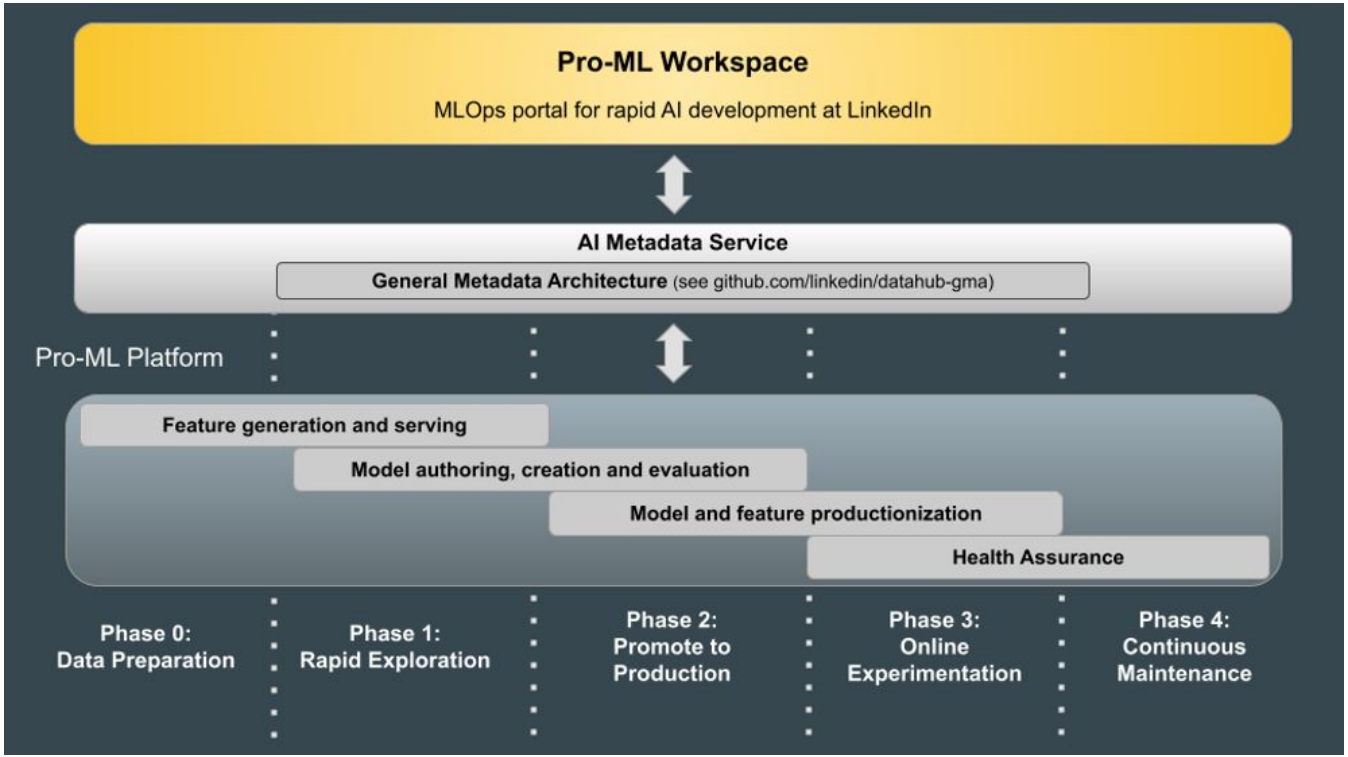
Identity: Who are you? Where do you work? What are your skills? Who are you connected with?

Content: How many times was the update viewed? How many times was it “liked”? What is the update about? How old is it? What language is it written in? What companies, people, or topics are mentioned in the update?

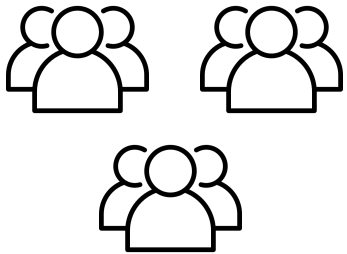
Behavior: What have you liked and shared in the past? Who do you interact with most frequently? Where do you spend the most time in your news feed?

https://www.linkedin.com/blog/engineering/feed/a-look-behind-the-ai-that-powers-linkedins-feed-sifting-through

Linkedin's Productivity Machine Learning (Pro-ML) platform.



Teams of data scientists

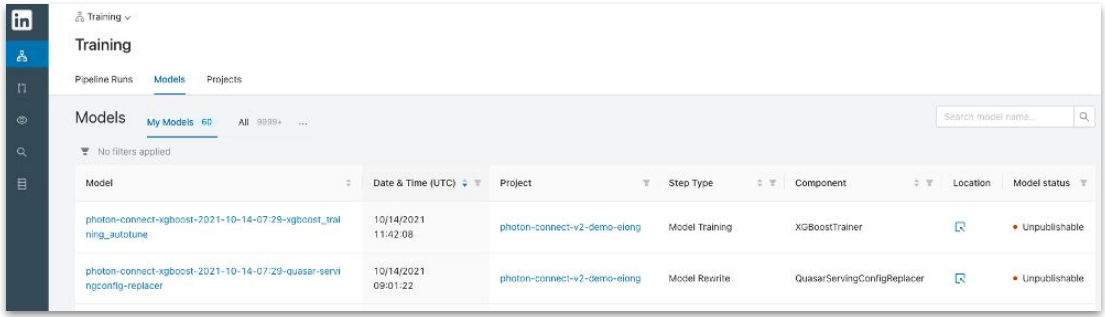


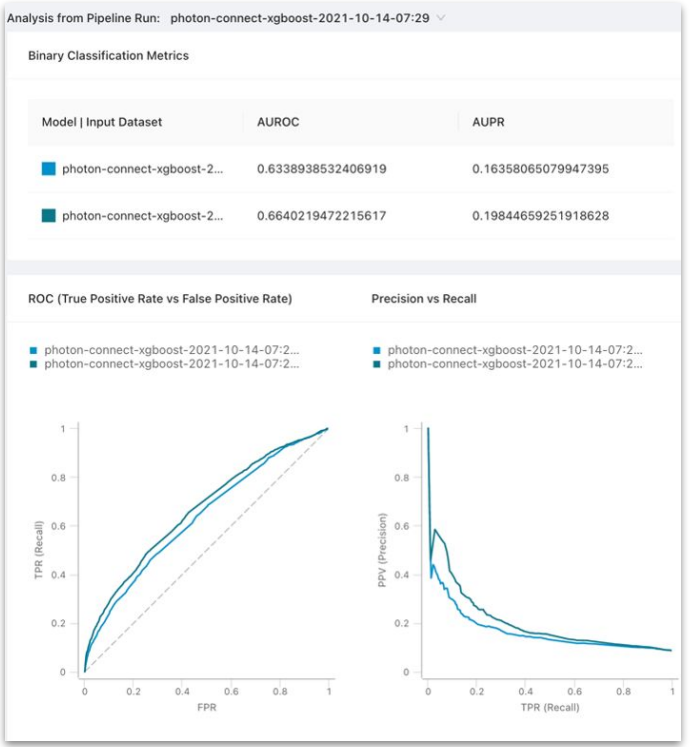
https://www.linkedin.com/blog/engineering/machine-learning/one-stop-mlops-portal-at-linkedin

Linkedin Pro-ML platform.

Step: Model authoring, creation, and evaluation

Model tracking and experimentation platform
Similar to **MLFlow** or **Weights & Biases** (which we will cover in this course).

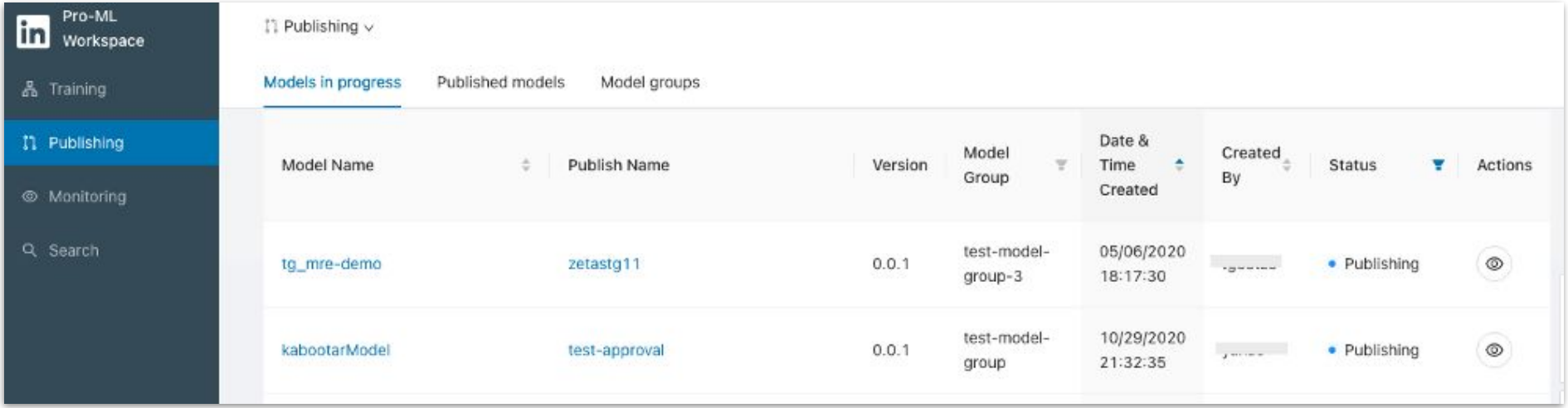




https://www.linkedin.com/blog/engineering/machine-learning/one-stop-mlops-portal-at-linkedin

Linkedin Pro-ML platform.

Step: Model productionisation



Workflows to **publish** or **deprecate** models.

https://www.linkedin.com/blog/engineering/machine-learning/one-stop-mlops-portal-at-linkedin

Linkedin Pro-ML platform.

Step: “Health insurance” (aka monitoring)

https://www.linkedin.com/blog/engineering/machine-learning/one-stop-mlops-portal-at-linkedin

Real-world use case deep dive

Real-estate valuation assistant

LIÈGE université

Context & Problem Statement

…heard of Fednot?

Fednot

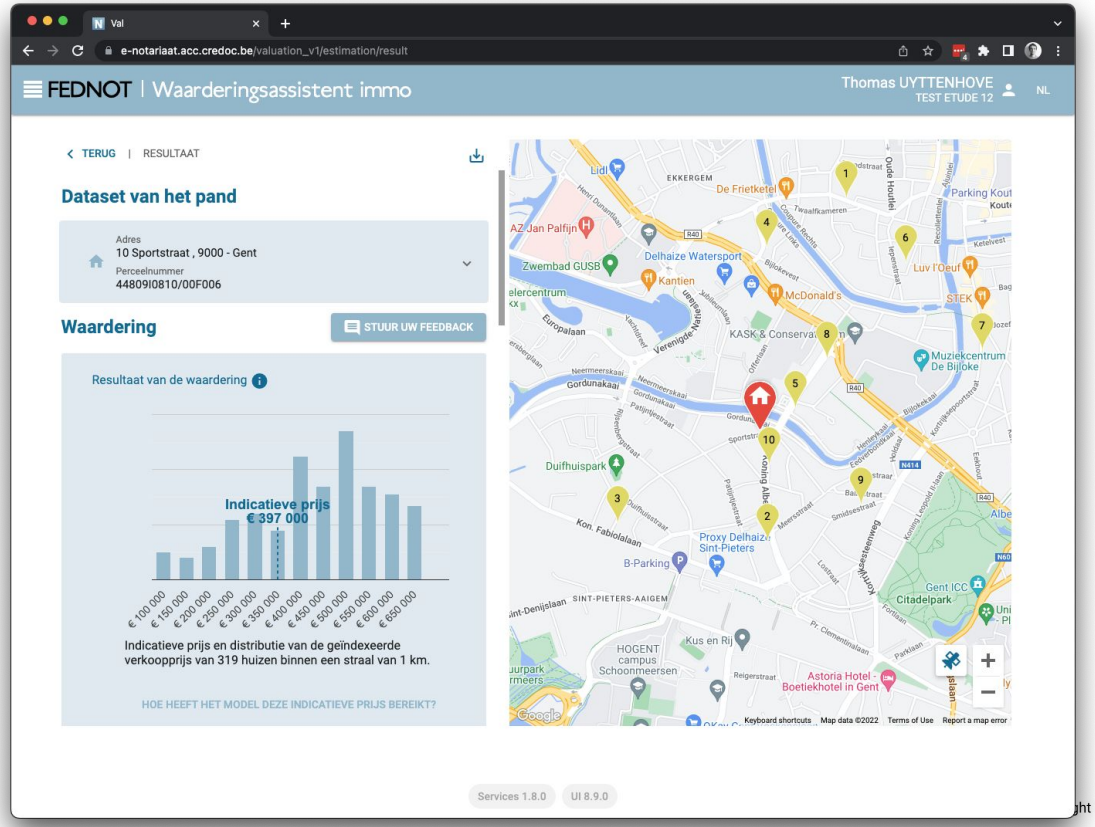
= Koninklijke **Fed**eratie van het Belgisch **Not**ariaat

= **Féd**ération Royale du **Not**ariat belge

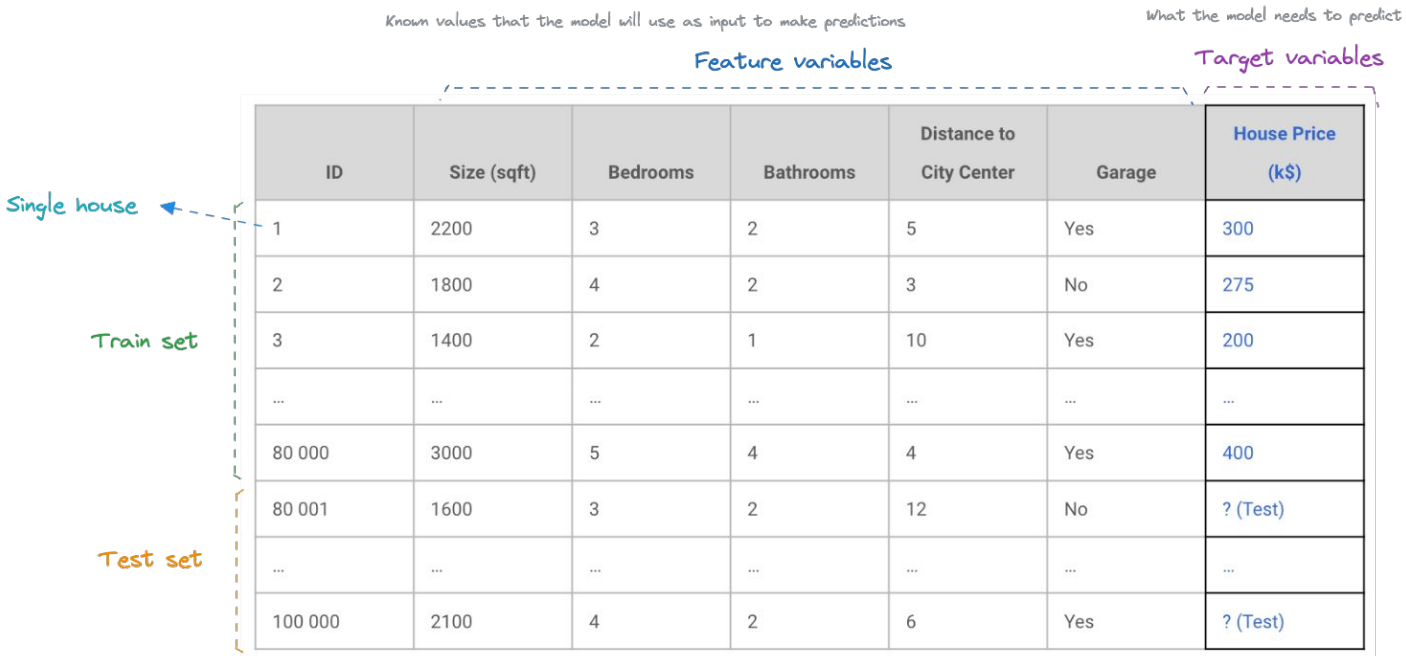
= Royal **Fed**eration of the Belgian **Not**aryship

Fednot supports the notary studies with juridical advice, office management, IT solutions, trainings, and information for the general public.

Valuation assistant.



How the ML model conceptually works



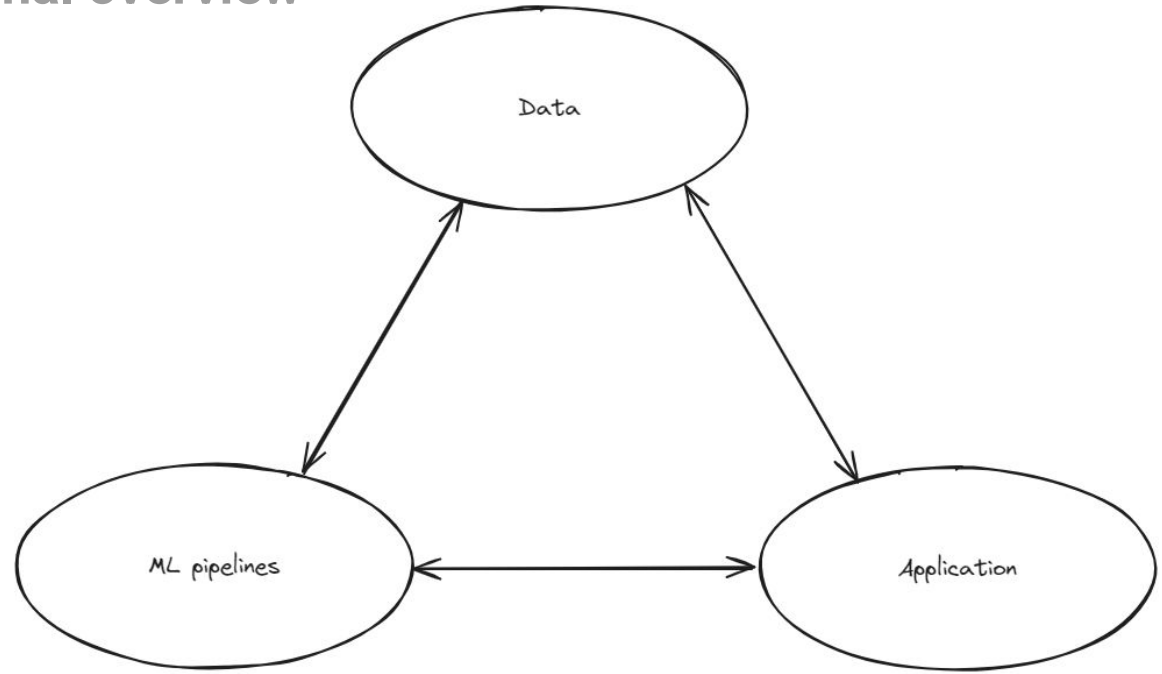
ID	Size (sqft)	Bedrooms	Bathrooms	Distance to City Center	Garage	House Price (k$)
1	2200	3	2	5	Yes	300
2	1800	4	2	3	No	275
3	1400	2	1	10	Yes	200
...	...	...	...	...	...	...
80 000	3000	5	4	4	Yes	400
80 001	1600	3	2	12	No	? (Test)
...	...	...	...	...	...	...
100 000	2100	4	2	6	Yes	? (Test)

The ML model will see many **observations** (houses) defined as a set of **features** (information, variables).
From it the model will learn patterns and what impacts **target variable** (house prices).

If given new observations, the model can **predict** the target variable based on the input features.

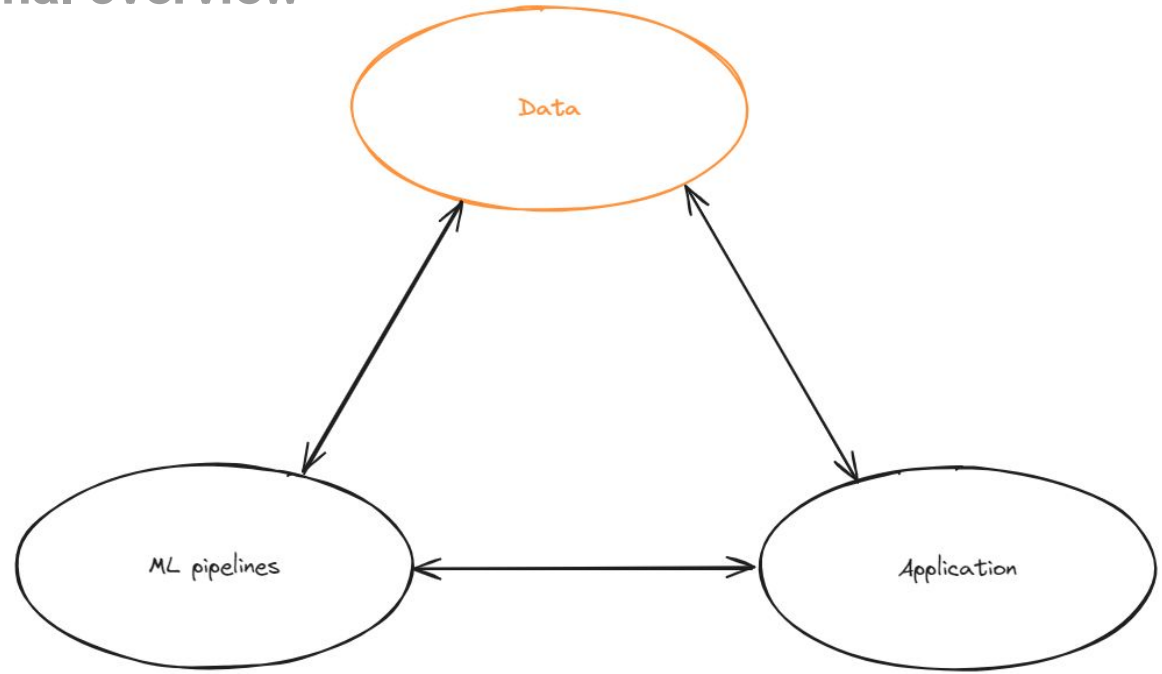
Solution architecture.

A functional overview



Solution architecture.

A functional overview



Valuation features.

Legal real-estate data

processed document | legal platforms

transaction

good *garage*

good *house*

transaction

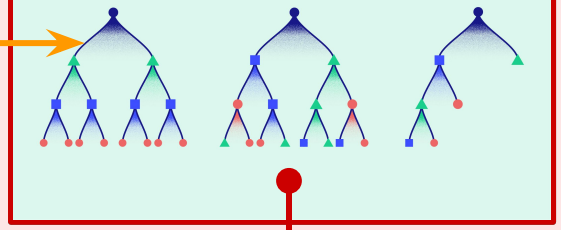
good *land*

transaction

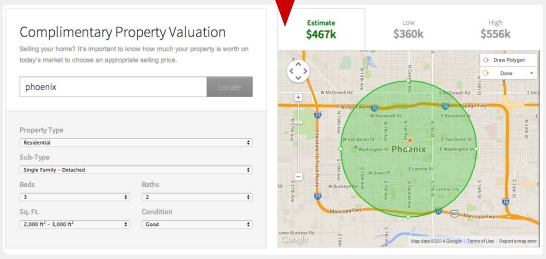
Legal ('AI') features

Legal features alone do not capture sufficient information to accurately predict real estate prices...

valuation model



valuation tool



ML6 × FEDNOT

Valuation features.

Open real-estate data

processed document | **legal** platforms

transaction

good *garage*

good *house*

transaction

good *land*

transaction

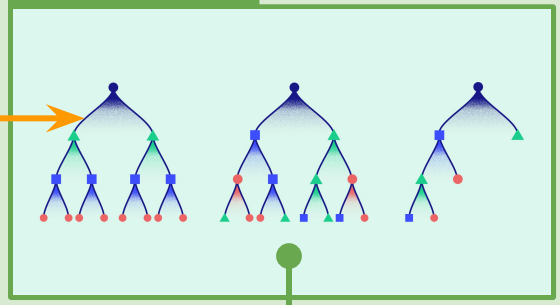
Legal ('AI') features

+

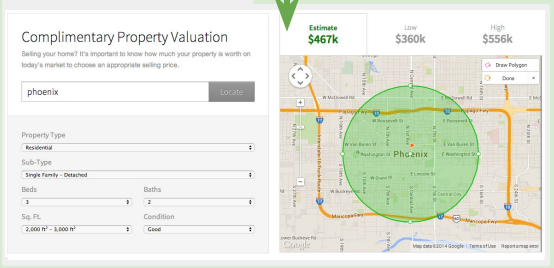
Leverage **open data** to expand limited feature set

- **Various types** of data sources provided by the government or community
- (Mostly) **freely accessible**
- Opportunities for **complex, more informative features**!

valuation model



valuation tool



ML6 × FEDNOT

Valuation features.

Open real-estate data

- **Cadastral information**
  - Parcel area, street width, ratio, and orientation;
  - Building area, type (“open”, “half-open”, “closed”), facade width, and orientation

Valuation features.

Open real-estate data

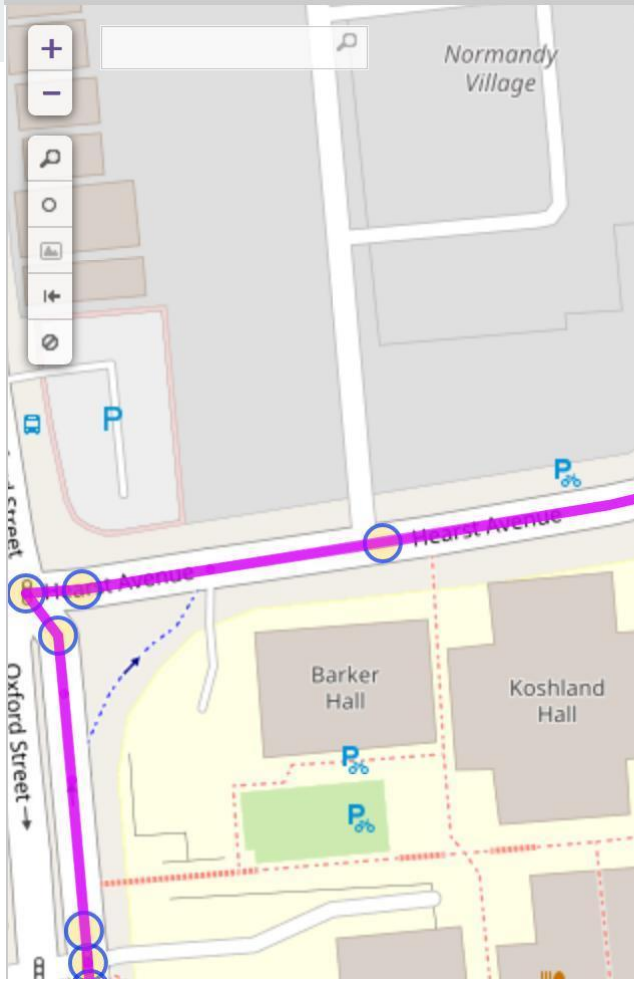
- Cadastral information
- **Height information**
  - Building height, volume, number of stories.

LIÈGE université

Valuation features.

Open real-estate data

- Cadastral information
- Height information
- **Location information**
  - Distance to major cities and to nearest city center, highway (entry), primary road, railway, station, bus stop, etc.

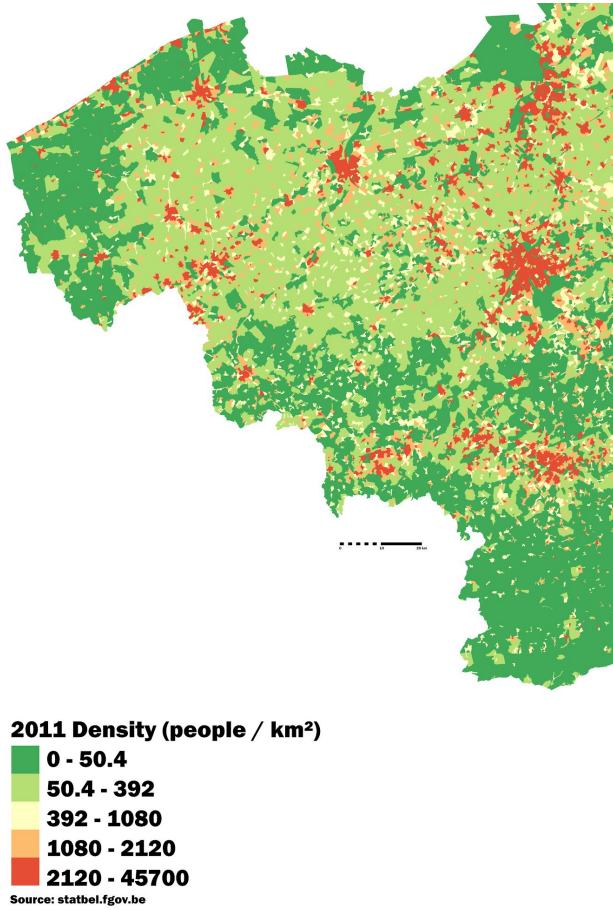


LIÈGE université

Valuation features.

Open real-estate data

- Cadastral information
- Height information
- Location information
- **Local socio-economic and demographic statistics**
  - Municipality population size, tax percentage, prosperity index, avg. income;
  - Statistical sector cadastral income percentiles



LIÈGE université

Open Data

Sources

BOSA

Federale Overheidsdienst FINANCIEN

.be

RÉGION DE BRUXELLES-CAPITALE

Vlaanderen verbeelding werkt

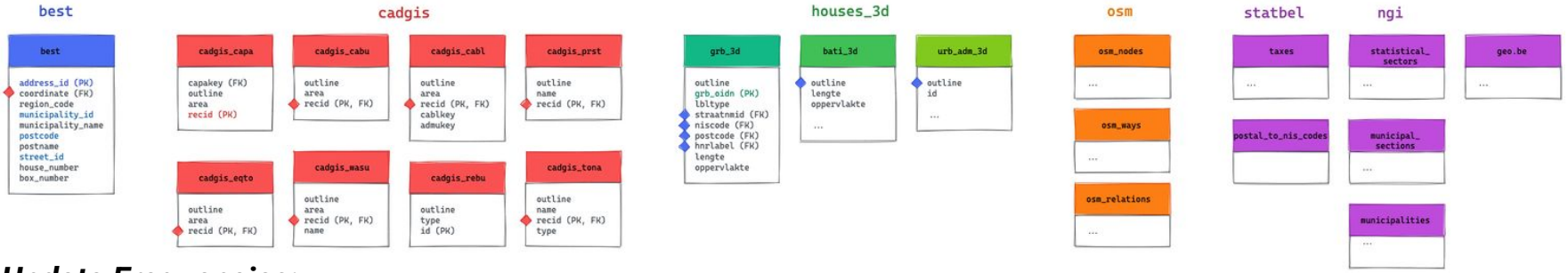
Géoportail Wallonie geoportail.wallonie.be

STATBEL België in cijfers

geo.be

NGI Nationaal Geografisch Instituut

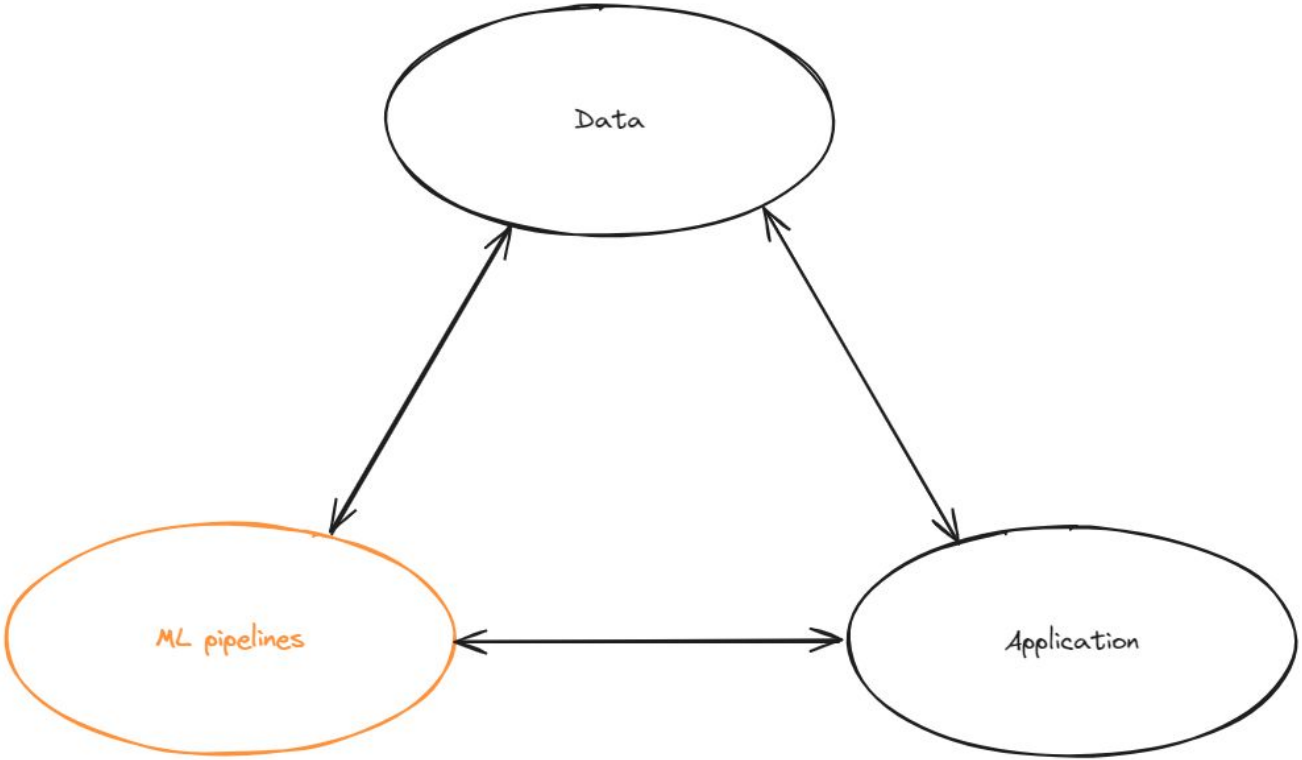
IGN Institut Géographique National



Update Frequencies:

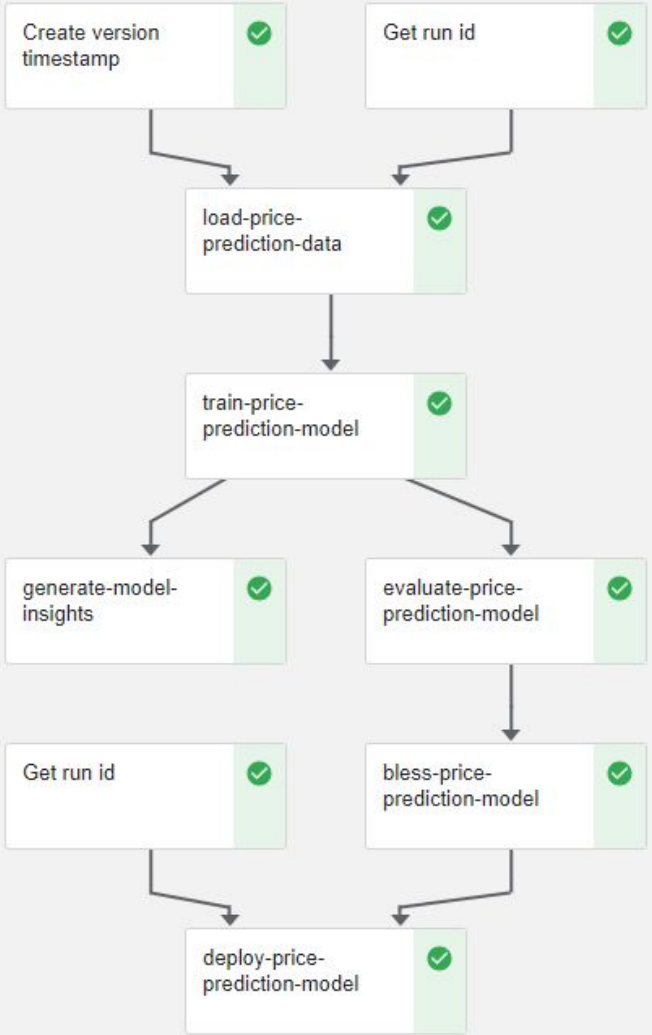
Weekly	Yearly	Wallonia - Not (2016) Flanders - Not (2016) Brussels - Yearly	Weekly	Yearly	Yearly

LIEGE université

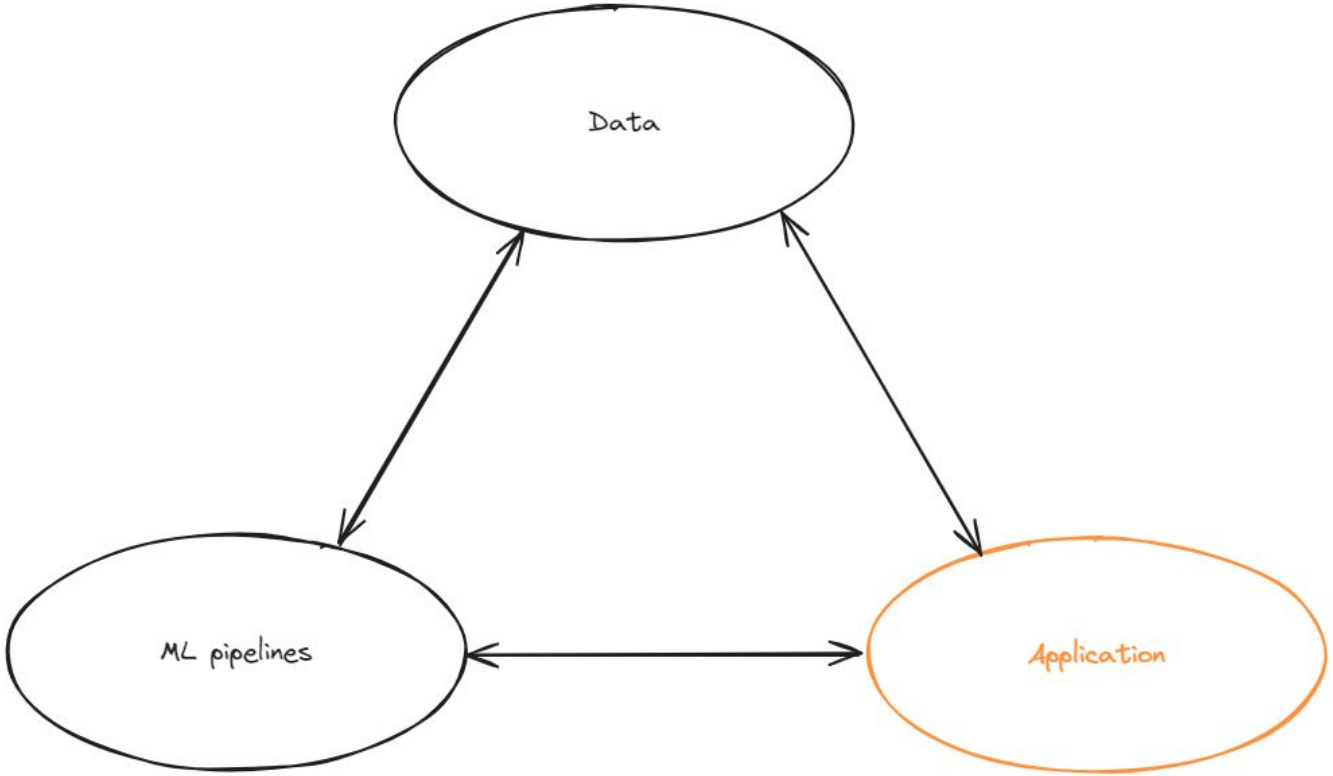


Automated pipeline to train and deploy new price prediction models

- Allows you to implement a **ML pipeline** made of different **components**, usually ran sequentially.
- Each component can be a **Docker image**
- Hosted on a **kubernetes cluster** (set of node machines for running containerized applications). Can be on the **Cloud**.
- Benefits
  - Modularized
  - Reproducible
  - Efficient
  - Scalable
  - Deployments
  - Collaboration
  - Version control and documentation

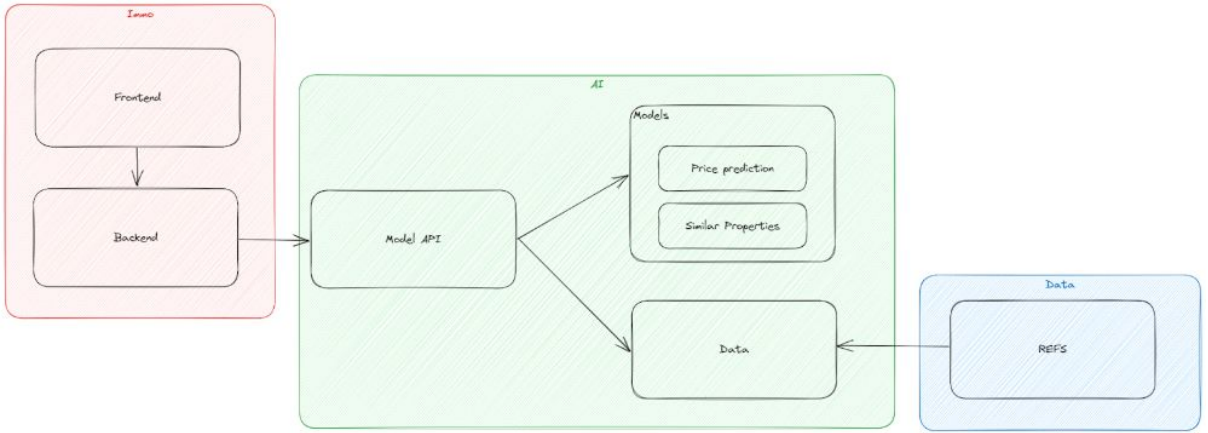


LIÈGE université



Model API

- **Cloud Run**
  - Hosting the front-end
  - Hosting APIs to connect components
- **GKE**
  - Hosting the models
- **Bigquery**
  - Hosting the data
- **Storage**
  - Hosting artifacts



Course organisation

Objective for this course.

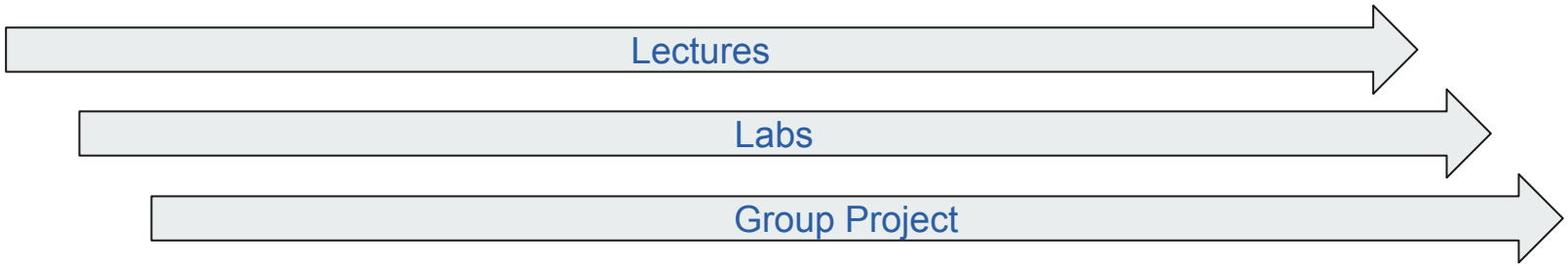
We want to enable you with skills to **design** and **build ML application** 🛠️

We selected core **topics** of MLSD to be tackled in this course. Tools are selected based on usability, performance, popularity and accessibility.

Goal is to provide

- **Theoretical** concepts
- **Technical** tools & skills
- Practical real world **practices**

Structure of the course



Relevant

Focused on core concepts of building ML applications. Tailored choice of current best practices.

Practical

Concrete Labs, resources, real life examples, time to experiment, support line, …

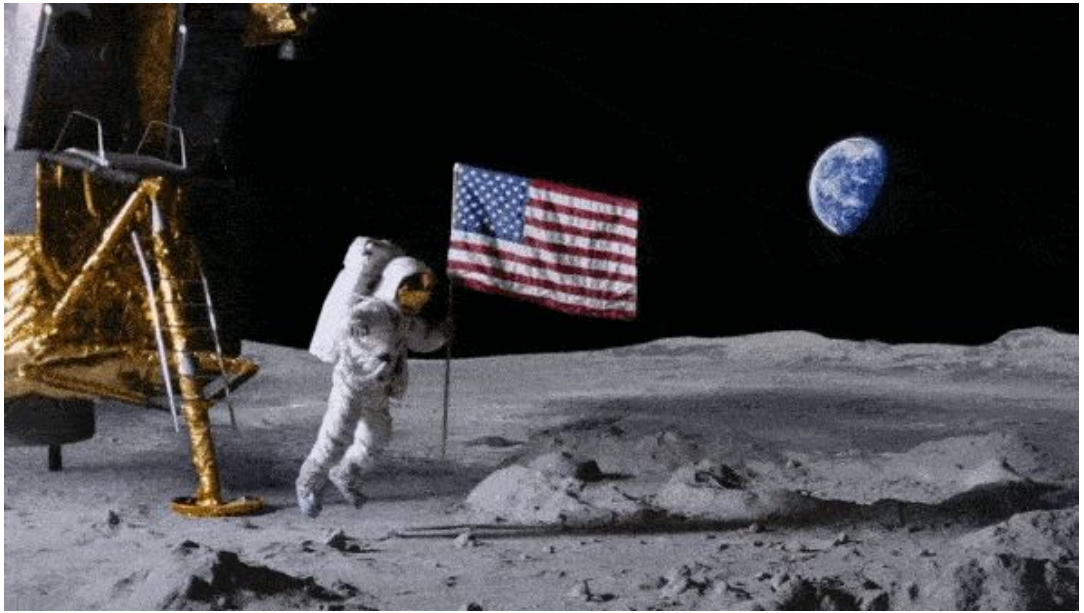
Engaging

Interactive class session.
Healthy tempo (break out exercises, QA, …).

We’re (again) making history.

This is the second edition of this class

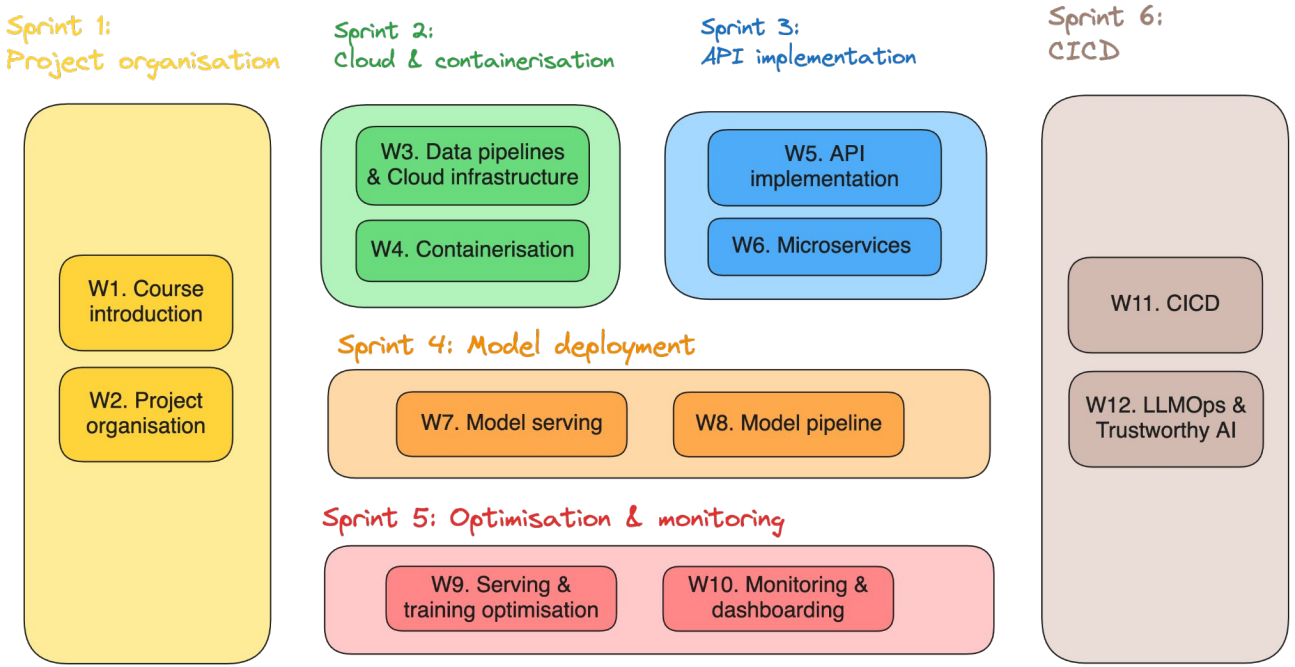
- Quick feedback cycles
- Open communication
- Enthusiasm for trying new things 🧪
- Active support from teaching staff



LIÈGE université

Course outline

Overview of sprints & classes



Overall organisation & communication

Class organisation

- We meet every Monday from **9:00** to **12:30**
- Typically you'll have about 2h of lecture + labs. Remaining of the time can be spent working on your project.

Useful links:

- All info on the Github page: https://github.com/ThomasVrancken/info9023-mlops
  - Project info
  - Sample exam
  - Lecture & labs (before the class)
- Discord: https://discord.gg/kY6B3cchkr
- Open office hours on **Monday afternoons** (office Number I 77 B in Montefiore)

Labs are split in two kinds: Demos & Directed work

Directed Work (DW)

You are given an assignment to do at the end of the course, in class.
The teaching staff is there to support.
At the end, you need to upload your codes & results to XXX
You will get a pass/fail grade for it.

There are 5 DWs, which sum up to 20% of your final grade (4% each).

Demos / hands-on

We want to introduce you to more than 5 tools.
For lower requirement tools, the teaching staff will give demos or you will get non-graded hands-on exercises.

Grade

Practice exam available on Github.

Your final grade is divided in:

- **Exam (30%)**: Oral exam on the topics covered during the lectures
- **Directed work (20%)**: Pass/fail on practical exercises given at the end of the courses. 5 DWs, 4% grade each.
- **Group project (50%)**: One large group project, see next slides.

Project

All project information is also on Github.

Organisation

Build one ML system throughout the course. The application is picked by yourself.

- **Teams:** 2 - 4 students
  - Form group by next week!
  - Let the teaching staff know if you don't have a group and you'll be assigned one
- **Structure**
  - The building blocks to be implemented in the project follow the course's **6 sprints**.
- **Handovers**
  - There will be **3 milestone meetings** where you can present your results
  - **Code submission** - make sure to document clearly anything you want the teaching staff to read
- **Support**
  - Often lectures/labs will be shorter than the time slot for this course. You can spend the extra time working with your team. Teaching staff will be in the room to provide support.
  - Open office hours on Monday afternoon in office Number I 77 B in Montefiore
  - Feel free to reach out by email if you have any question/struggle
- **You're in the driving seat!**
  - Many building blocks are optional. You are free to choose the overall design and tools used for your project. Experiment and ask questions if you have any.

Project

Guiding principles

- Learn, learn and learn!
- Find an interesting project to work on - ideally with a real world usage
- Come up with your own design and toolstack
- Focus on relevant parts of your specific system
- Motivate your choices

- … And pick a cool name for it

Example projects from last year

- Hessian: https://github.com/alexandre-eymael/HESSIAN
- ClipMorph: https://github.com/iSach/clipmorph
- Triple-P: https://github.com/lambi702/MLOps-TripleP
- …

Project objective for sprint 1

Project guidelines on github!

Week	Work package	Requirement
W01	Pick a **team** ● Try to mix skills and experience ● If you didn't find one let one of the teachers know and we'll allocate you to one	Required
W02	Select a **use case** ● Previous course ● Kaggle Datasets ● … Make sure to pick a use case where **data is available**. Ideally pick something with interesting data and a real world application.	Required
W02	**Define** your use case. Fill in a ML Canvas template page (You can skip the *Inference* part as we will tackle that in a later sprint.)	Required
W02	Find a **cool name** for your project ✨	Required
W02	Submit your project by sending a filled in project card to the teaching staff with basic information about your project. We might give you some feedback and ask for parts to be changed.	Required
W02	Setup **communication channel** (Discord, trello)	Required
W02	Setup a **code** versioning **repository** ● We recommend Github as we will cover Github Actions during this course	Required

LIÈGE université

Resources

Similar courses

- University of Bari
  - Paper: "Teaching MLOps in Higher Education through Project-Based Learning." arXiv preprint arXiv:2302.01048 (2023)
  - Lanubile, Filippo, Silverio Martínez-Fernández, and Luigi Quaranta
- Stanford University
  - CS 329S: Machine Learning Systems Design (link)
  - Chip Huyen
- Carnegie-Mellon University
  - Machine Learning in Production / AI Engineering (link)
  - Christian Kästner

Interesting resources

- Machine Learning Engineering for Production (MLOps) Specialization (Coursera, Andrew Ng)
  - GitHub, Youtube
- Made with ML (link)
- Marvelous MLOps (link)

Books

- Designing Machine Learning Systems: An Iterative Process for Production-Ready Applications (Chip Huyen)
- Building Machine Learning Powered Applications: Going from Idea to Product (Emmanuel Ameisen)
- Introducing MLOps (Mark Treveil, Nicolas Omont, Clément Stenac et al.)
- Machine Learning Design Patterns (Valliappa Lakshmanan, Sara Robinson, Michael Munn)

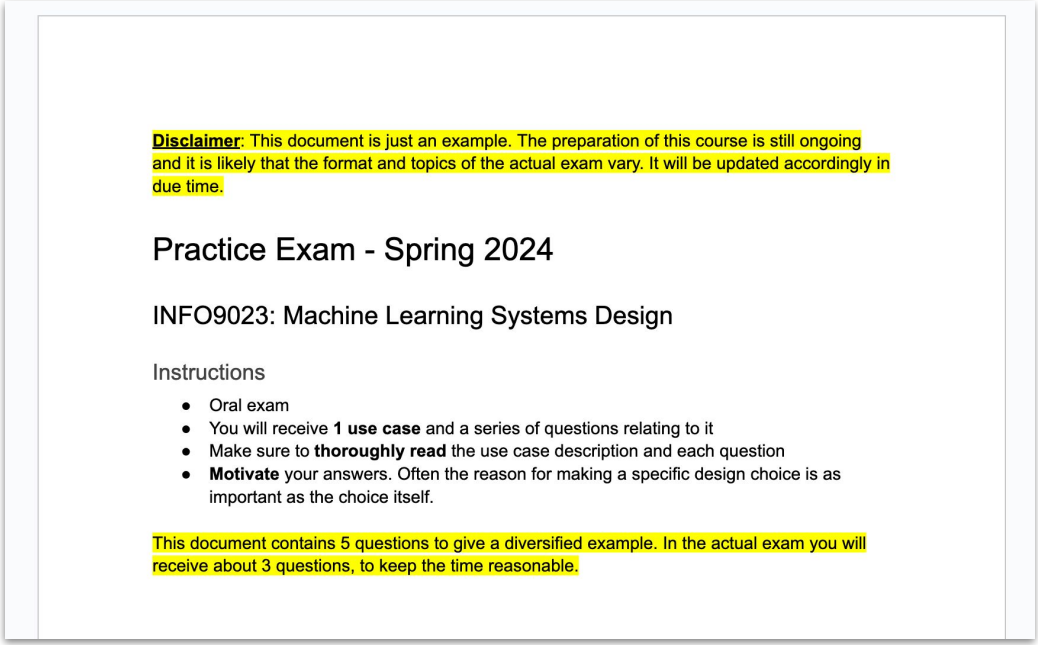
That’s it for today!

LIÈGE université

Grading

Exam & Project

1. Oral exam (30% of the final grade)
    - Find practice exam on Github
2. Project (70% of the final grade)



Disclaimer: This document is just an example. The preparation of this course is still ongoing and it is likely that the format and topics of the actual exam vary. It will be updated accordingly in due time.

Practice Exam - Spring 2024

INFO9023: Machine Learning Systems Design

Instructions

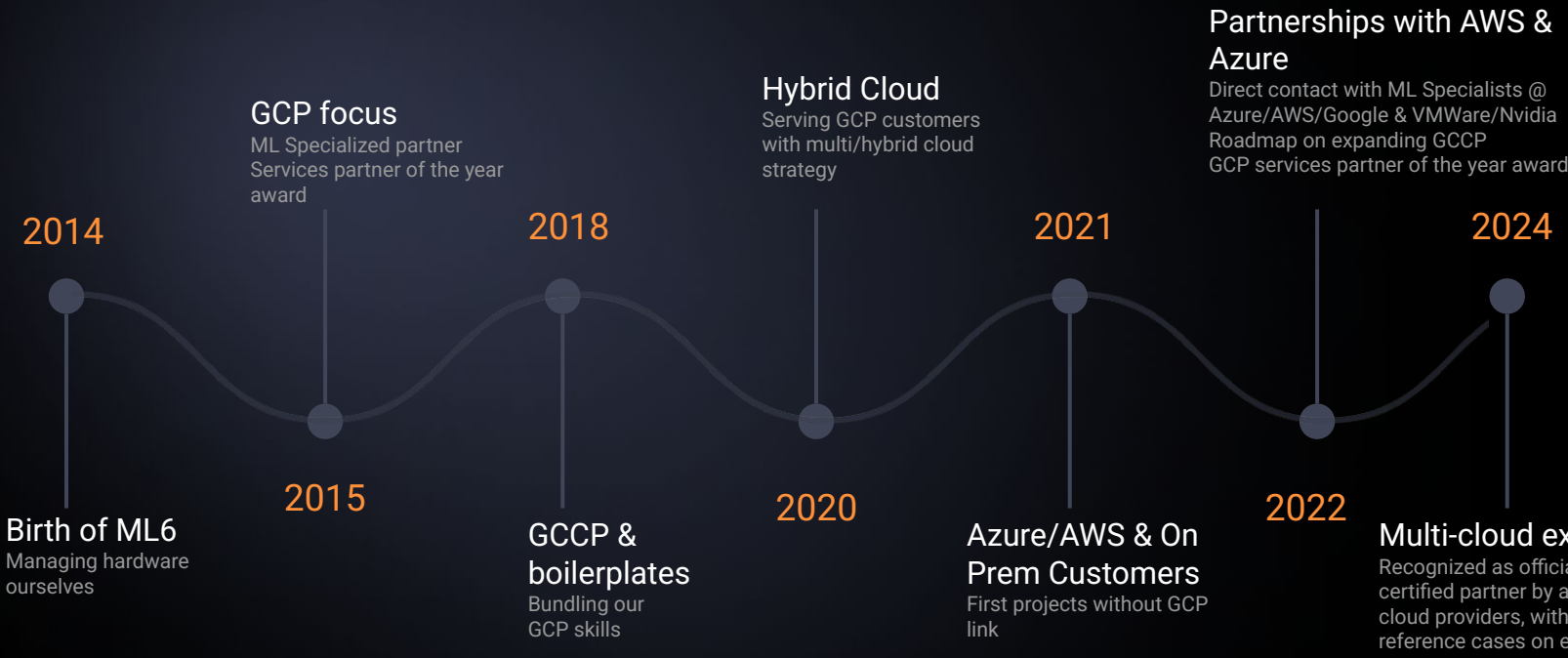
- Oral exam
- You will receive **1 use case** and a series of questions relating to it
- Make sure to **thoroughly read** the use case description and each question
- **Motivate** your answers. Often the reason for making a specific design choice is as important as the choice itself.

This document contains 5 questions to give a diversified example. In the actual exam you will receive about 3 questions, to keep the time reasonable.

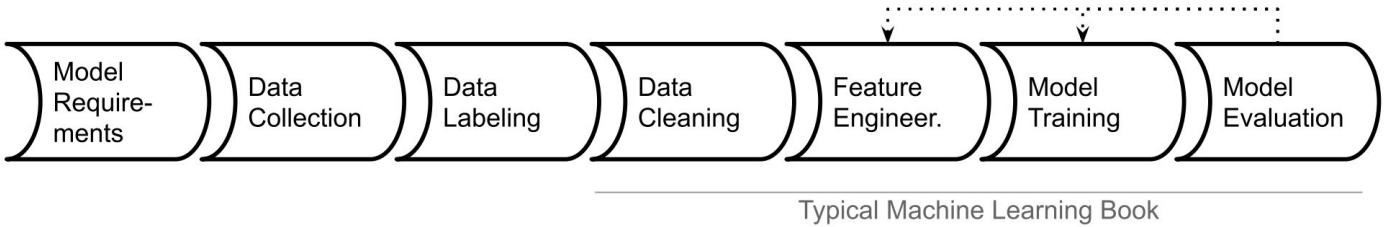
APPENDIX

We partner closely with Google, Amazon and Microsoft.

Official and certified experts



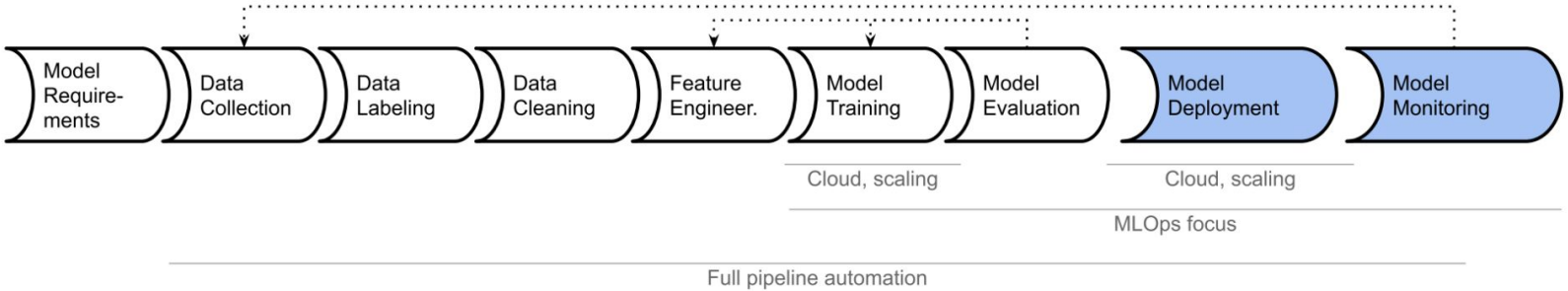
Traditional data science



Output:

- Jupyter Notebook
- Single model working on static dataset

MLOps - Fully automated pipeline



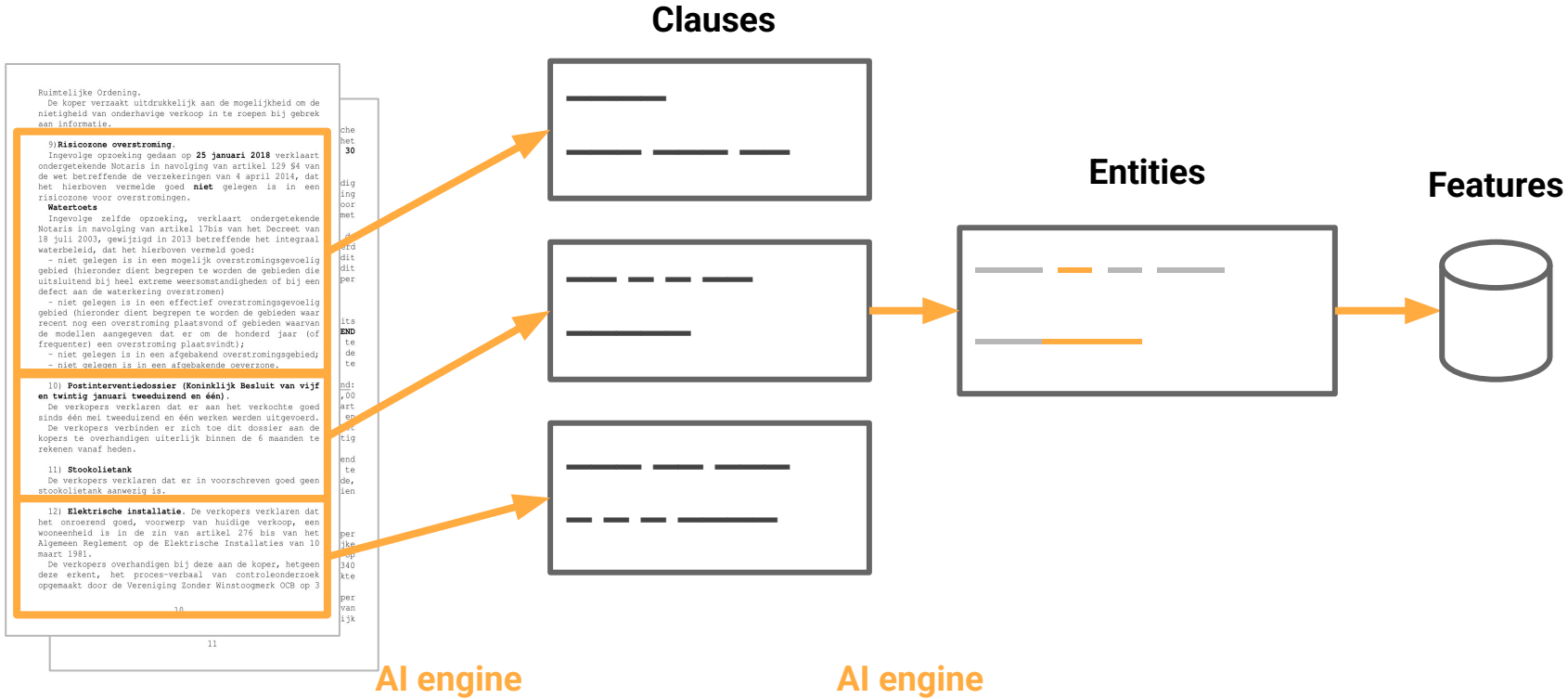
Output:

- Deployed model (e.g. API in the Cloud)
- Monitor live model performance
- Directly connected to data source
- Fully automated pipeline to train and deploy new models
- ...

Machine Learning in Production / AI Engineering - Carnegie Mellon University (Christian Kästner)
https://github.com/ckaestne/seai/tree/F2022/lectures/02_systems

Valuation features.

Starting from unstructured PDF documents



Valuation features.

Processing the texts from deeds

We want to extract named entities!

- Persons
- Dates
- Addresses
- Locations
- ...

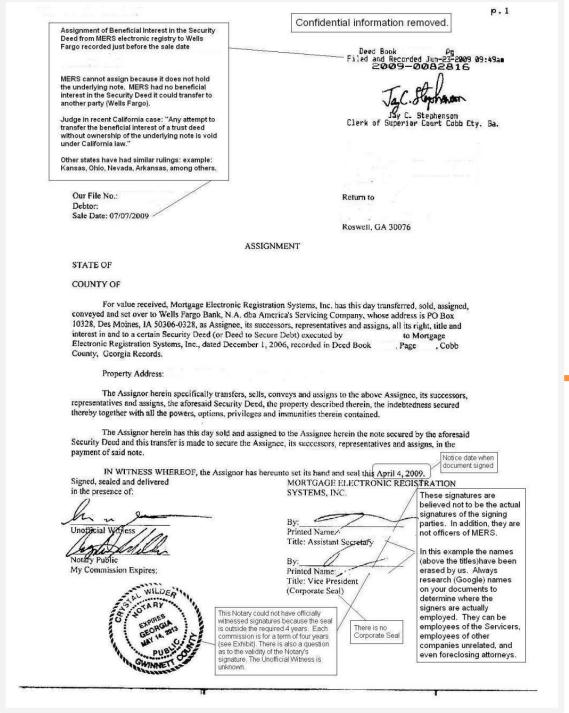
Erfdienstbaarheden
Het goed wordt verkocht met al zijn gekende en ongekende, zichtbare en onzichtbare, voortdurende en niet-voortdurende erfdienstbaarheden en zakelijke rechten en verplichtingen.
De verkoper verklaart geen weet te hebben dat het goed is bezwaard met onzichtbare erfdienstbaarheden behoudens de erfdienstbaarheid vermeld in de eerder aangehaalde eigendomsakte verleden voor notaris Hans Deprez op 14 oktober 2012 waarin letterlijk het volgende wordt vermeld:
“Ten titel van inlichting en zonder de bedoeling exhaustief te zijn, worden de volgende erfdienstbaarheden aangehaald, zoals deze vermeld staan in de akte verleden door notaris Hendrickx te Antwerpen op 7 april 2014 overgeschreven inhoudende verkoop door de vennootschap aan de heer Peeters Tom:
“Over het verkochte goed te Stationsstraat 148 wordt onvergeld en eeuwigdurend een erfdienstbaarheid van doorgang behouden in het voordeel van het erf van de bewoner.
Deze doorgang zal mogen gebruikt worden zowel te voet als met gemotoriseerde voertuigen, om de terreinen achter de bestaande gebouwen van de verkoper en deze gebouwen zelf langs hun achterzijde te kunnen bereiken of verlaten, doch enkel in noodgevallen (bijvoorbeeld door de brandweer of andere hulpdiensten).
Om in voorkomend geval te allen tijde en ten gerieve van de aanpalende erven van verkoper de vrije doorvaart te verzekeren, zal vijf meter achter de bestaande gebouwen van verkoper, nooit enige constructie, aanplanting of hindernis mogen opgesteld worden, incluis enige obstructie ten belope van de bedrijfsuitvoering van Dynamo BVBA die hieronder omschreven staat als zijnde de primaire en op het moment van schrijven enige...”

LIÈGE université

Valuation features.

Legal real-estate data

legal document



processed document

transaction

good *garage*

good *house*

transaction

good *land*

legal platforms

good *house*

# of bedrooms
# of bathrooms
terrace
...

construction yr.
cadastral income
epc
...

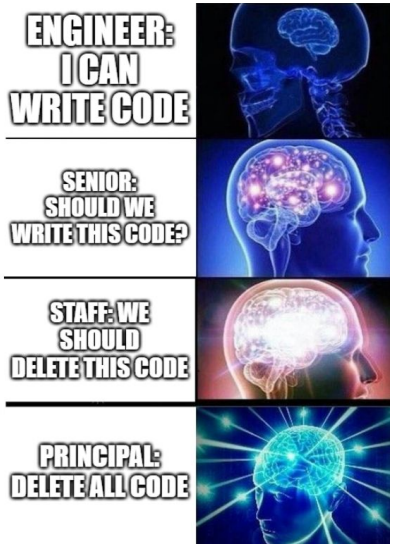
Legal ‘AI’ features

- Transaction date
- Good (sub)type
- Price(s)
- Energy Performance Certificate (EPC)
- Newly built/existing
- Cadastral income
- ...

LIÈGE université

Learnings

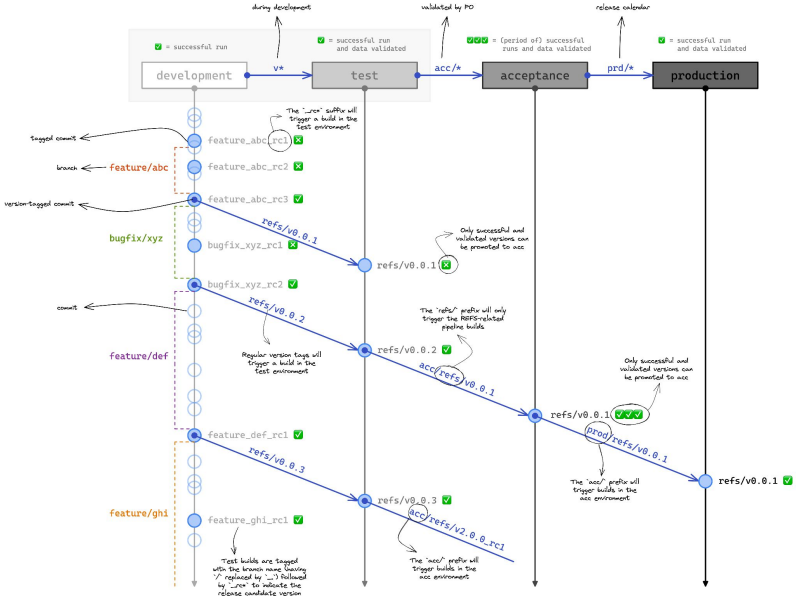
- **“Software engineering is programming integrated over time”** - Titus Winters
  - Think enough about the time aspect
- **Documentation is not just writing stuff down**
  - Hardest part of old code is not figuring out what it IS doing, but what it is INTENDED (or SUPPOSED) to do
  - What was decided is “trivial” (you see the end result), the CONTEXT and WHY is the interesting (and hard) part
  - cfr. Chesterton’s fence
  - Keep handovers in mind → Think about your bus factor
- **Document your decisions**
  - Plan for the *future* and not for the *present*
    - Road to production should be clear
    - For example: postponing caching implementation
  - Visibility towards stakeholders!
    - “Why did *you* not do this before?” becomes “*We* decided to do other things”
- **Dare to delete stuff**



LIÈGE université

Technical Learnings

- **Build once, deploy anywhere**
  - Manual deployments to tagged deployments to ML pipelines
  - Scheduled execution
  - Balance between model performance and freshness
- **“Garbage in, garbage out”**
  - Time as a crucial feature
  - “Good” (house, apartment, cabin, …) type is hard to define
  - Notaries are too lazy/busy to provide accurate information
- **We should have started caching earlier**
  - Huge amount of deeds flowing through regularly
  - Dataflow pipelines took a looooong time and regularly went OOM




ML APPLICATIONS
ML APPLICATIONS EVERYWHERE
imgflip.com


ML APPLICATIONS
WE SEE
ML BEING APPLIED
TO EVERY INDUSTRY
imgflip.com

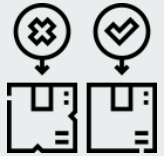
A few (other) example of ML applications.

Document AI

Pricing

Video games

Quality control



Robotic

Customer clustering

Customer support

...

LIÈGE université

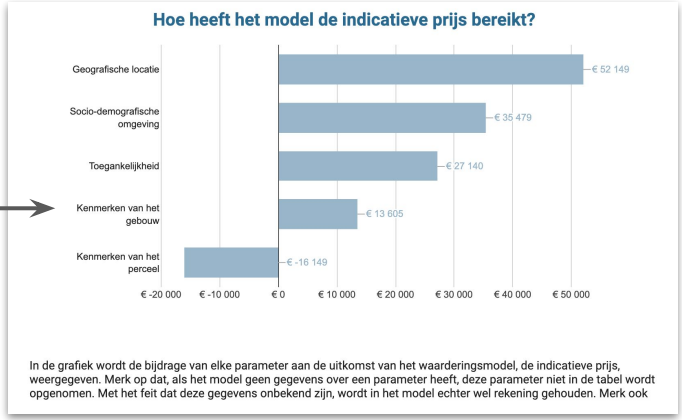
... 300 ML use cases: https://www.evidentlyai.com/ml-system-design

Valuation model.

Price prediction model

Model is not an “oracle”, but supports notaries

- Important to express uncertainty.
- Shows indicative price compared to price distribution of similar properties.
- Explains how a prediction was obtained based on feature importances.



LIÈGE université

Deed data

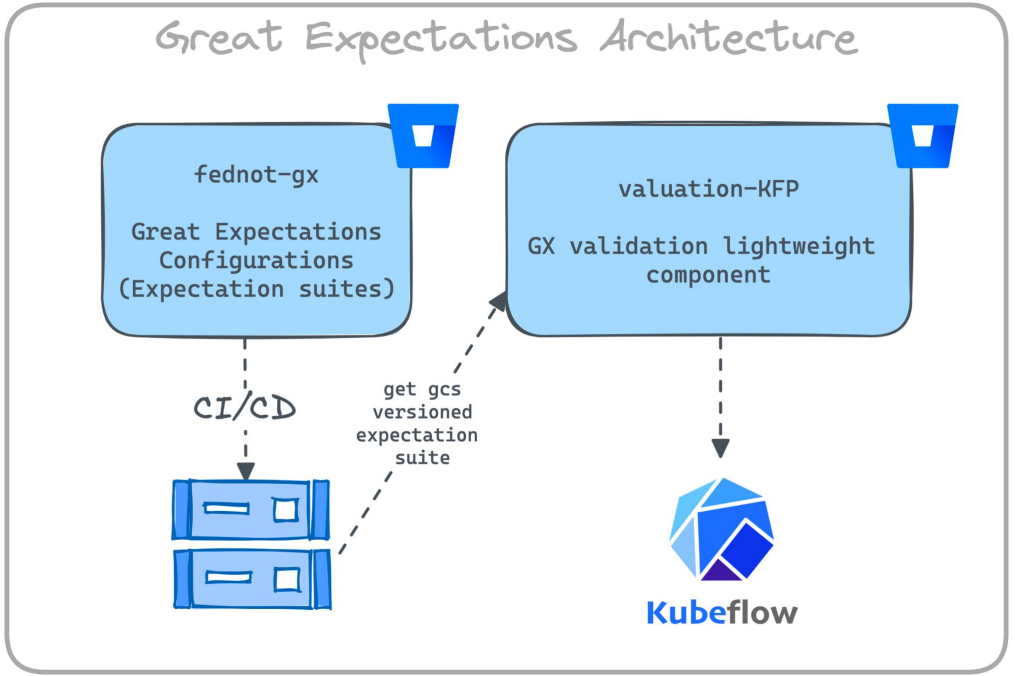
Data validation with great expectations

https://greatexpectations.io/

Deed data

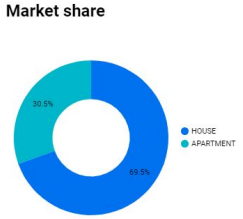
Data validation with great expectations

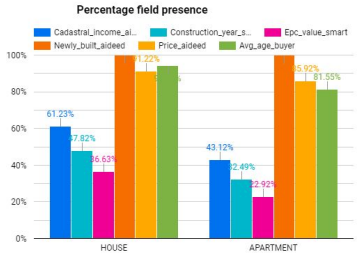
- Enable non-technical people to define expectations
  - **Expectations**: A JSON file that defines rules for validation. Eg: *feature_x should be 75% not null*
- Make the GX component reusable for multiple validation needs.



Deed data

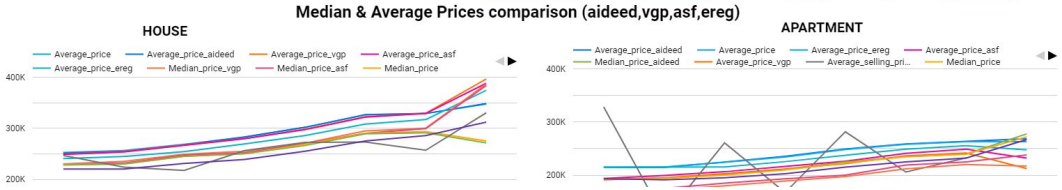
Data quality dashboard





good_type / geo_level_value / Average_price / Median_price

	HOUSE		APARTMENT	
	België		België	
Year	Average_price	Median_price	Average_price	Median_price
2017	252,705 €	229,341 €	214,948 €	192,664 €
2018	256,400 €	230,000 €	215,498 €	193,992 €
2019	268,476 €	245,000 €	224,142 €	200,000 €
2020	282,622 €	250,000 €	234,026 €	209,963 €
2021	301,602 €	268,576 €	248,156 €	222,958 €
2022	326,414 €	290,000 €	258,193 €	234,811 €
2023	328,846 €	293,463 €	262,929 €	239,296 €
2024	347,345 €	275,000 €	262,875 €	272,000 €



Visualisations

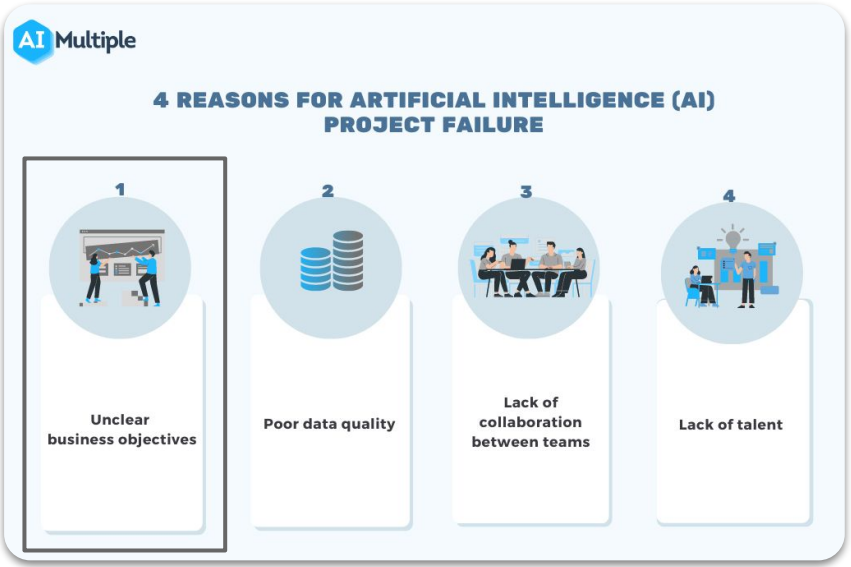
Explainability

		VAN			WHT			
		nr_observations	mae	mape	nr_observations	mae	mape	nr_observ
quantile	key_features							
Q0 - Q10	0 or 1	1547	80006.176115	0.565104	870	47111.933130	0.505811	1115
	2 or 3	1312	59277.557874	0.401544	1245	37980.680478	0.409488	751
Q10 - Q25	0 or 1	1983	38821.581166	0.161699	1238	26893.919732	0.220017	1523
	2 or 3	2299	29467.034478	0.121787	1950	22944.052560	0.187566	1301
Q25 - Q50	0 or 1	3225	34544.096558	0.111164	2039	23278.130844	0.142917	2281
	2 or 3	3892	30662.822366	0.098830	3279	21559.459244	0.132382	2390
Q50 - Q75	0 or 1	3335	52391.982770	0.132128	2077	36009.678320	0.164627	2279
	2 or 3	3782	43711.812300	0.110839	3296	32903.844697	0.150859	2407
Q75 - Q90	0 or 1	1945	90583.684565	0.177171	1328	49428.070282	0.169335	1326
	2 or 3	2329	69172.979267	0.134752	1858	44220.675478	0.150880	1484
Q90 - Q100	0 or 1	1311	229875.407221	0.280143	912	116302.847821	0.245235	937
	2 or 3	1539	175906.829581	0.213262	1200	92134.481717	0.194406	937

Static HTML

Let’s look at reasons for project failures

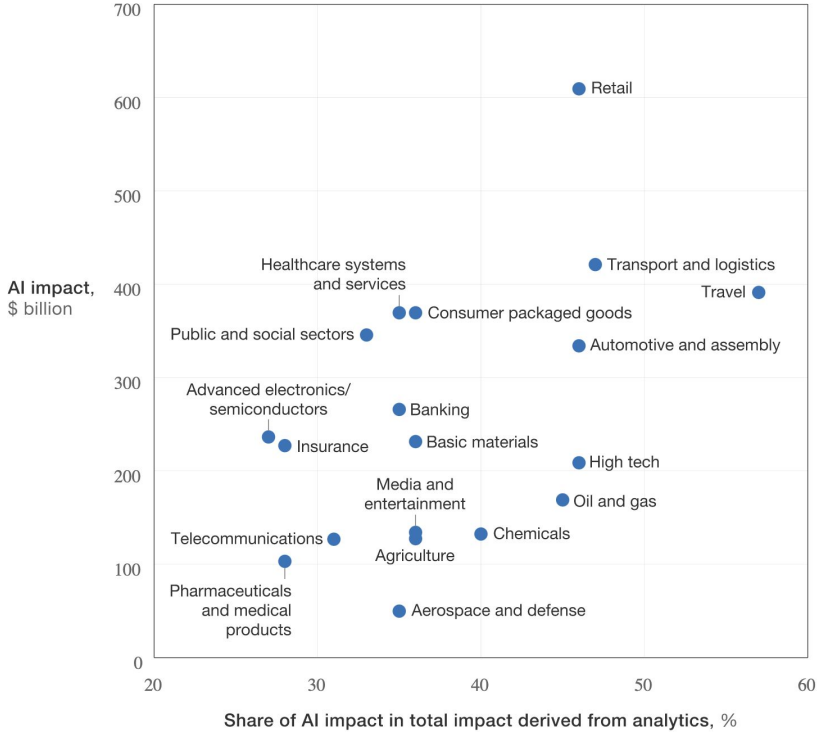
How can we prevent this from happening?



https://research.aimultiple.com/ai-fail/

AI is bringing a revolution in many different industries

Artificial intelligence (AI) has the potential to create value across sectors.



AI value creation by 2030

13 trillion USD

Most of it will be outside the consumer internet industry

CS 329S: Machine Learning Systems Design - Stanford (Chip Huey)
https://stanford-cs329s.github.io/syllabus.html

McKinsey&Company | **Source:** McKinsey Global Institute analysis