

Replicability challenges in applied data science: insights from the industry

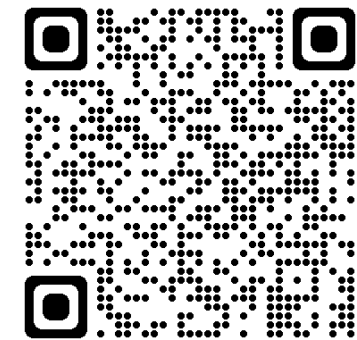
Davide Posillipo

Summer School «Replicability Crisis in Science?»

20th September 2023, Padua

About me

- Davide Posillipo
- Statistician, Freelancer Data Scientist, based in Milan
- Co-Founder of Deep Learning & Big Data Department (DL&BD) at [Alkemy](#)
- Industrial Partner of [BRIO](#) (Bias, Risk and Opacity in AI), PRIN MUR
- External Advisor of [PHILTECH](#) (Center of Philosophy of Technology, University of Milan)
- Let's get in touch for collaborations!



<https://davideposillipo.com/>

Goals for today

- Provide an overview of the industry of *applied data science*
- Discuss if and how the concepts of reproducibility and replicability are relevant for this industry
- Show how the industry faces threats to reproducibility and replicability
- Raise your interest for foundational studies that could be done in this field (potential collaborations!)



Motivations

- Issues in reproducibility and replicability are an economic burden for companies
- Lack of reproducibility and replicability increases the *opacity* of data applications, which affect the reliability of decisions made upon them
- Data applications are ubiquitous and affect our daily lives: opaque, unreliable results are a danger for our society



Applied Data Science: the context

What is applied data science?

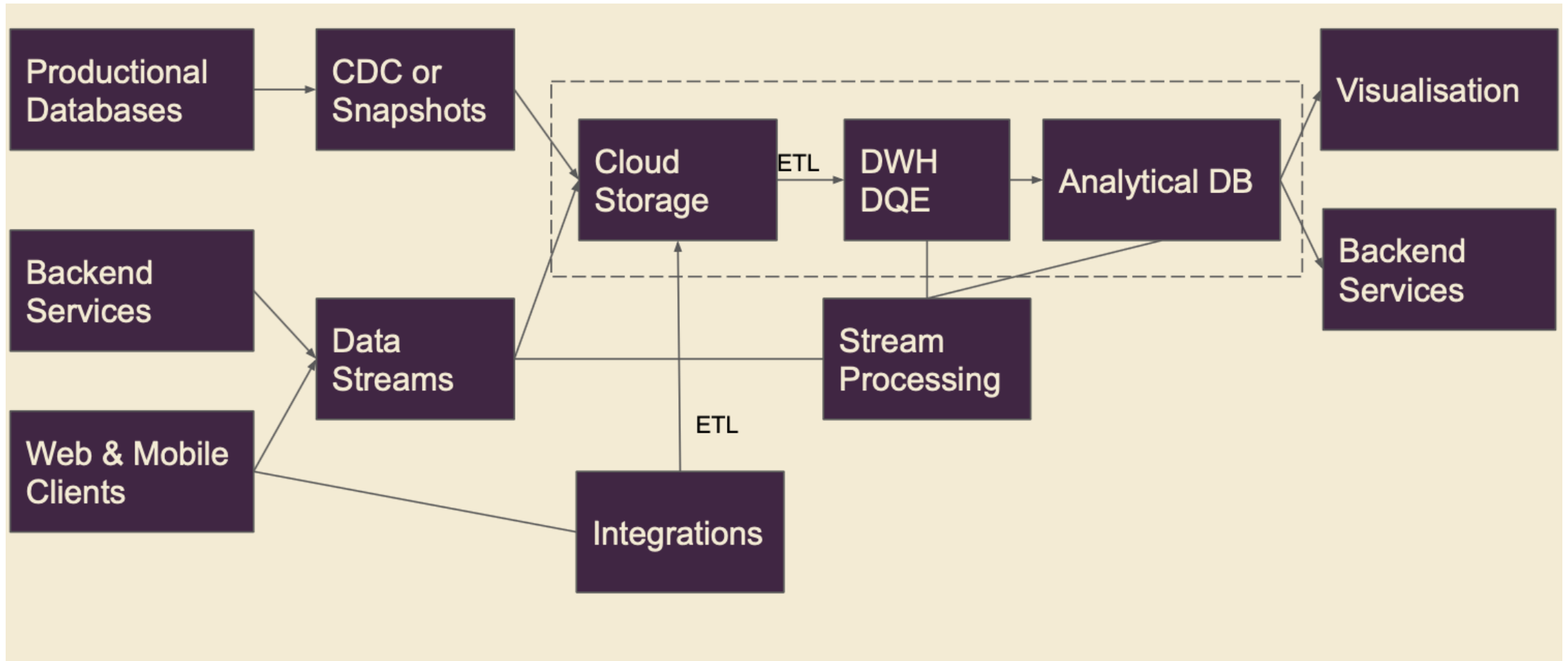
- With *applied data science* I refer to all the applications where data is continuously gathered from:
 - people
 - software applications
 - IoT devices
- After being gathered, data is stored and processed for different purposes.
- Occasionally, data is gathered through ad hoc surveys, but *the most of the times data is a secondary product of the companies' operations.*
- Here I **don't** consider applications like clinical trials in the pharma industry, where sample design plays a central role.



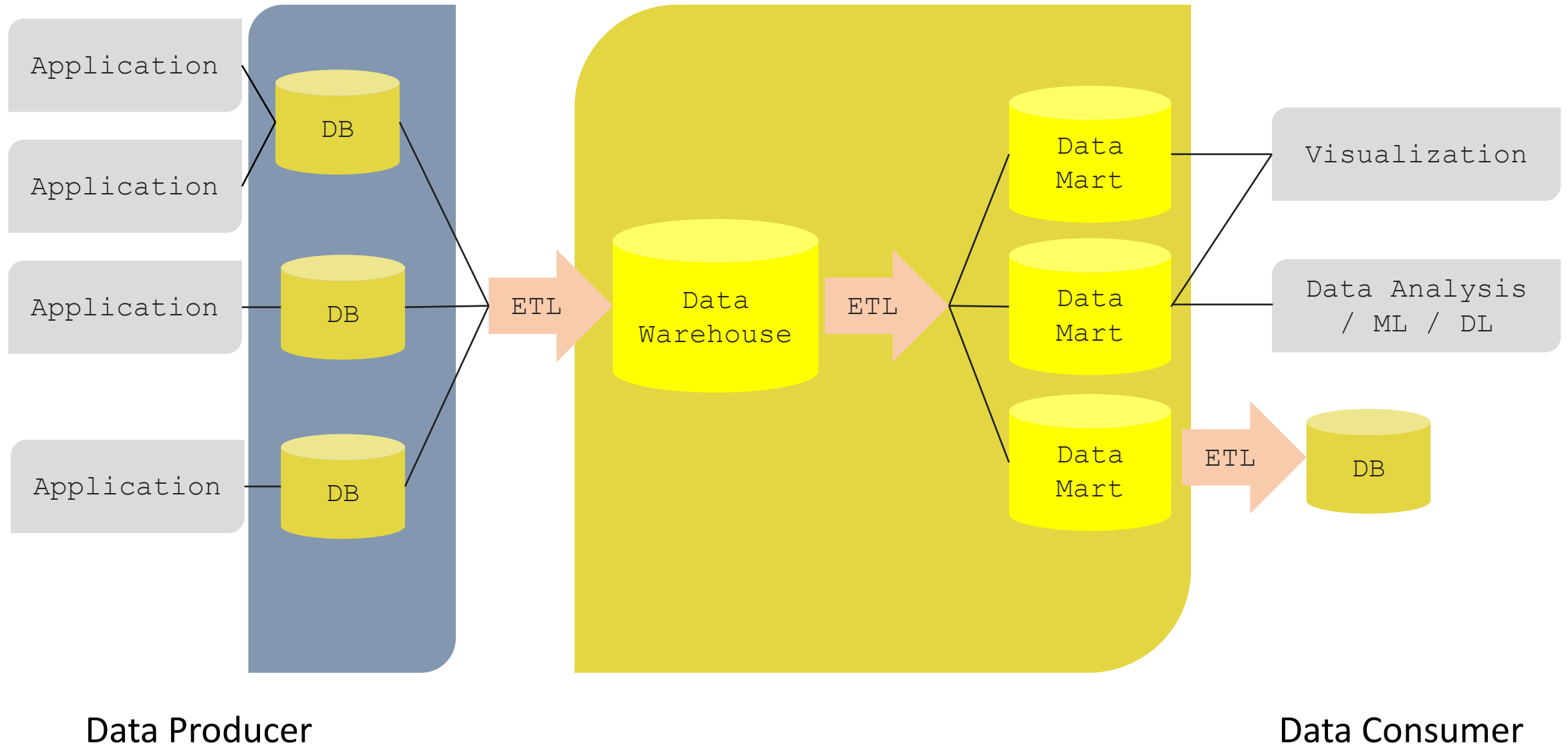
Data landscape in applied data science

- The issue of replicability is tied with the conceptualization of the materials and tools we work in, and with the way we work.
- Understanding how data architectures are imagined and thought about helps in understanding where the most critical issues are found.
- The most common data infrastructure found in companies follows the metaphor of a *data pipeline*: the data *flows* from sources to some destinations. This concept is connected to data architecture and technologies such as *data warehouses*, *data lakes*, *data marts*, *databases*.
- More modern architectures such as *data fabric*, *data mesh* and *data space* are still not really adopted, despite solving many of the issues we will see in this presentation.

Data landscape: the big picture



Data landscape: zooming in



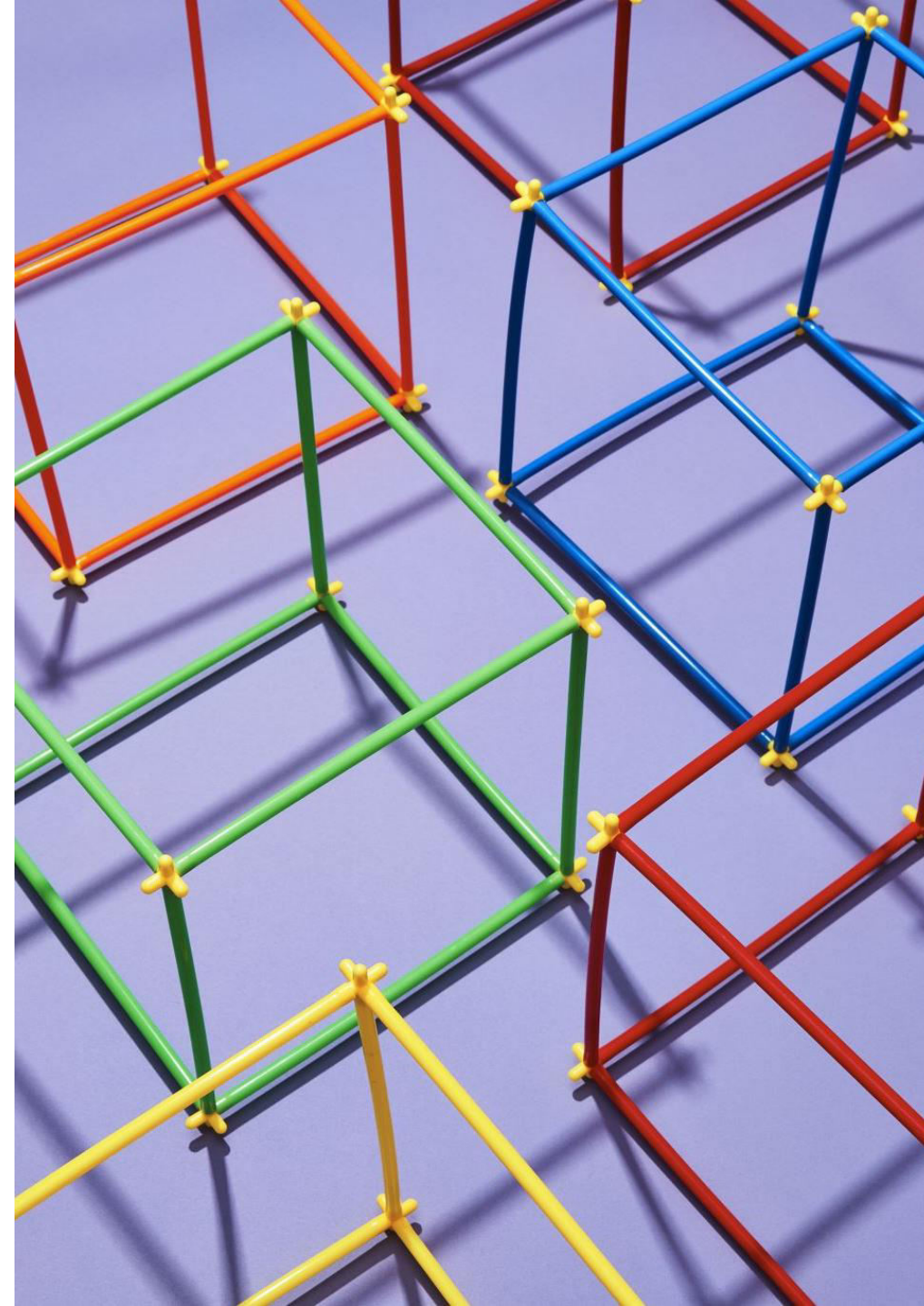
The tip of the iceberg: data analysis, ML, DL

- Once the data has flowed all the way through the data pipeline, it can be finally used by data scientists and machine learning engineers
- It is important to notice that at this point usually the content of the data, object of analysis, must be accepted as reliable
- In other words, all the transformations that produced the data are considered as given, and a whole new series of data manipulation is used. This data manipulation takes different forms:
 - data pre-processing
 - aggregations, filtering, summaries
 - features engineering



Different sections of the pipeline, different roles

- Despite an abuse of the term *data scientist*, more and more the roles in the data industry evolve and specialize.
- We can at least distinguish between:
 - *Data Engineers/Analytical Engineers*: overseeing the activities that move the data from the sources to the data consumers
 - *Data Scientists/Data Analysts/Machine Learning Engineers*: use the data resulting from the different steps of the pipeline for creating predictive models or ad hoc analyses






A dynamic setting

- All the challenges we encounter in industry can be understood only realizing that each component of the pipeline is constantly subject to changes, updates, and maintenance activities.
- At the same time, data must continuously flow from the sources to the destination; we refer to this scenario as *being in production*.
- This double effect (pipeline components' updates + new data coming) makes the issues of reproducibility and replicability particularly complex and entangled.

Theoretical Concepts



Replicability vs Reproducibility

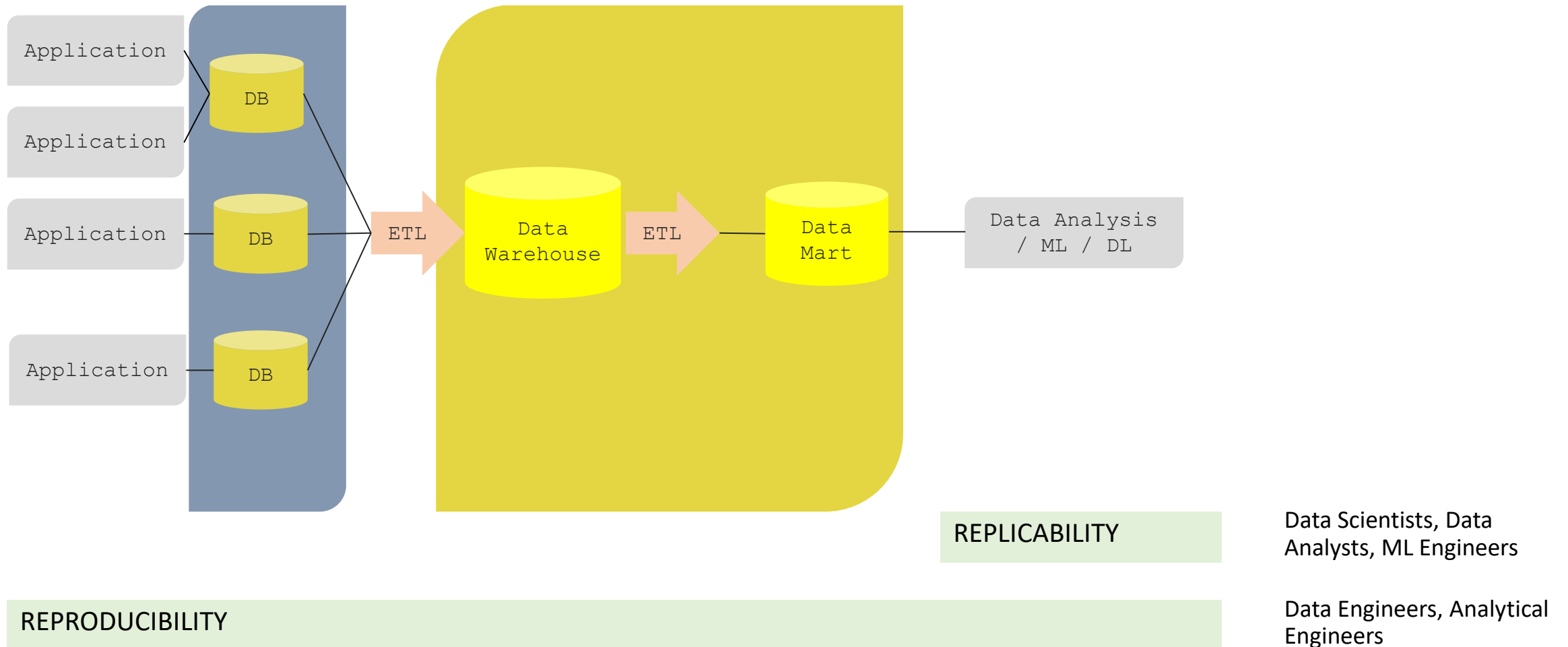
- During the following discussion, I will use the definition “B1” from Barba (2018):
 - “Reproducibility” refers to instances in which the original researcher’s data and computer codes are used to regenerate the results, while “replicability” refers to instances in which a researcher collects new data to arrive at the same scientific findings as a previous study.
- Both concepts are applicable and relevant for our industry, with a different weight given by context, domain and stakeholders involved.

Different roles, different concerns

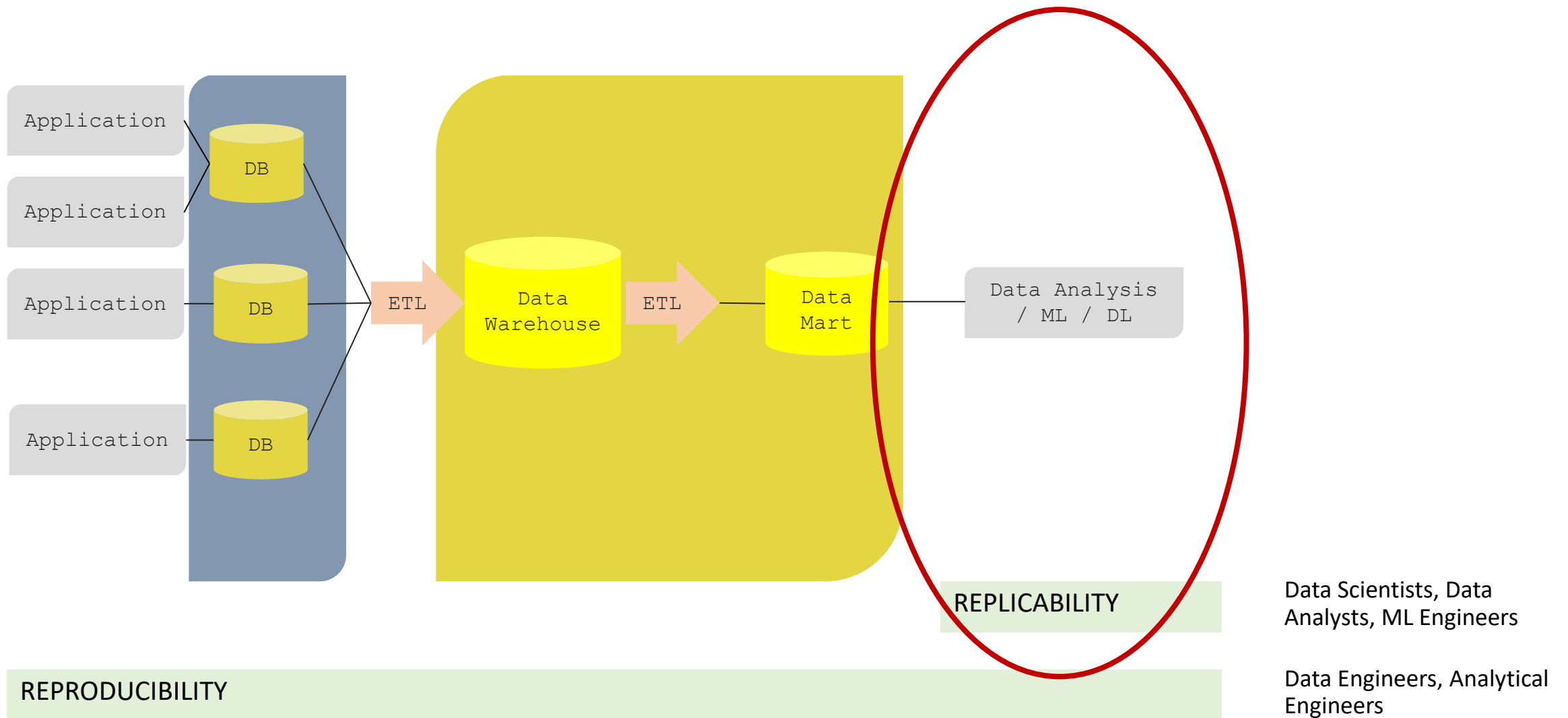
“To clarify the notion of reproducibility we need to address the following question: reproducibility of what and by whom?”, (Radder 1996)

- Data Engineers are (or better, should be) concerned with *reproducibility*: given the sources, is the pipeline producing the same output?
- Data Scientists need to address reproducibility as well, but also *replicability*: given the data produced by the pipeline, do the results of models and analyses “generalize”?

Reproducibility and Replicability: a first application to the pipeline

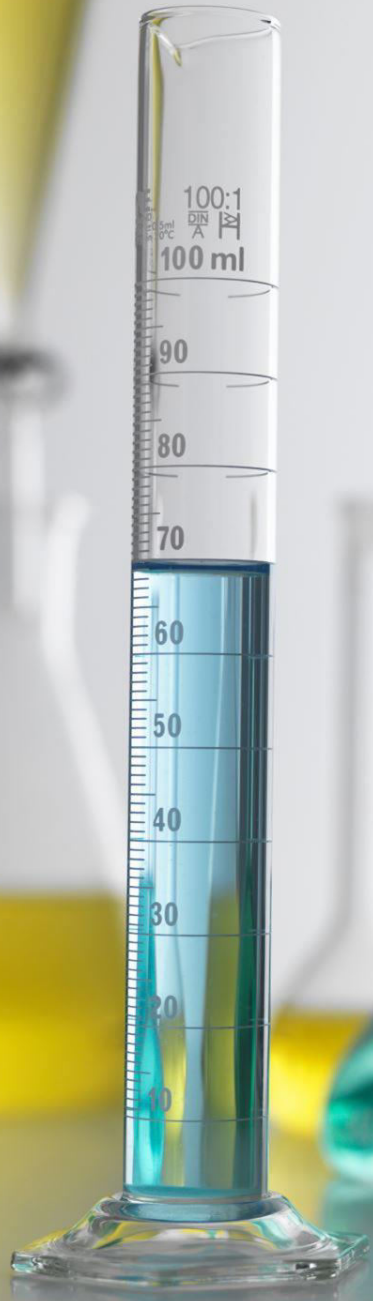


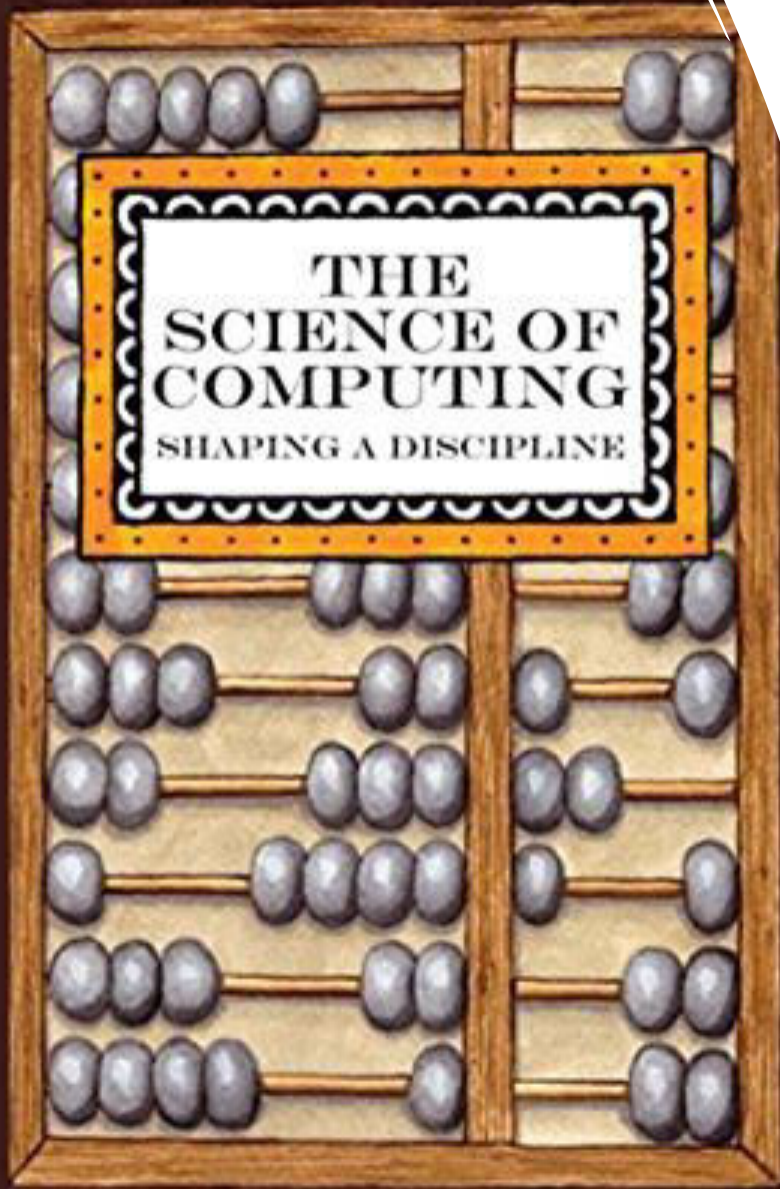
Replicability?



Does applied data science use the scientific method?

- The title of this Summer School is “Replicability Crisis in Science?”. Science is equipped with the scientific method, with experimentation as its core.
- Experimentation and its properties are crucial part of the (possible) replicability crisis discussed in this setting (putting apart reproducibility, which should be easier to “fix”).
- So these questions arise: does applied data science use the scientific method? Is experimentation to be found in industry, and in which forms?





Different types of experimentation

“At least **five views** are somewhat prevalent: experiment as a **demonstration of feasibility**, experiment as a **trial run**, experiment as a **field test**, experiment as a **comparison between competitors**, and the **controlled experiment**. Many would object against calling, for instance, feasibility demonstrations ‘experiments,’ arguing that the term ‘experiment’ has a special meaning in science. They are right. But if one looks at how authors in computing have used the term—not how it should be used—those five uses are easily found.” (Tedre 2015)

Controlled experiment

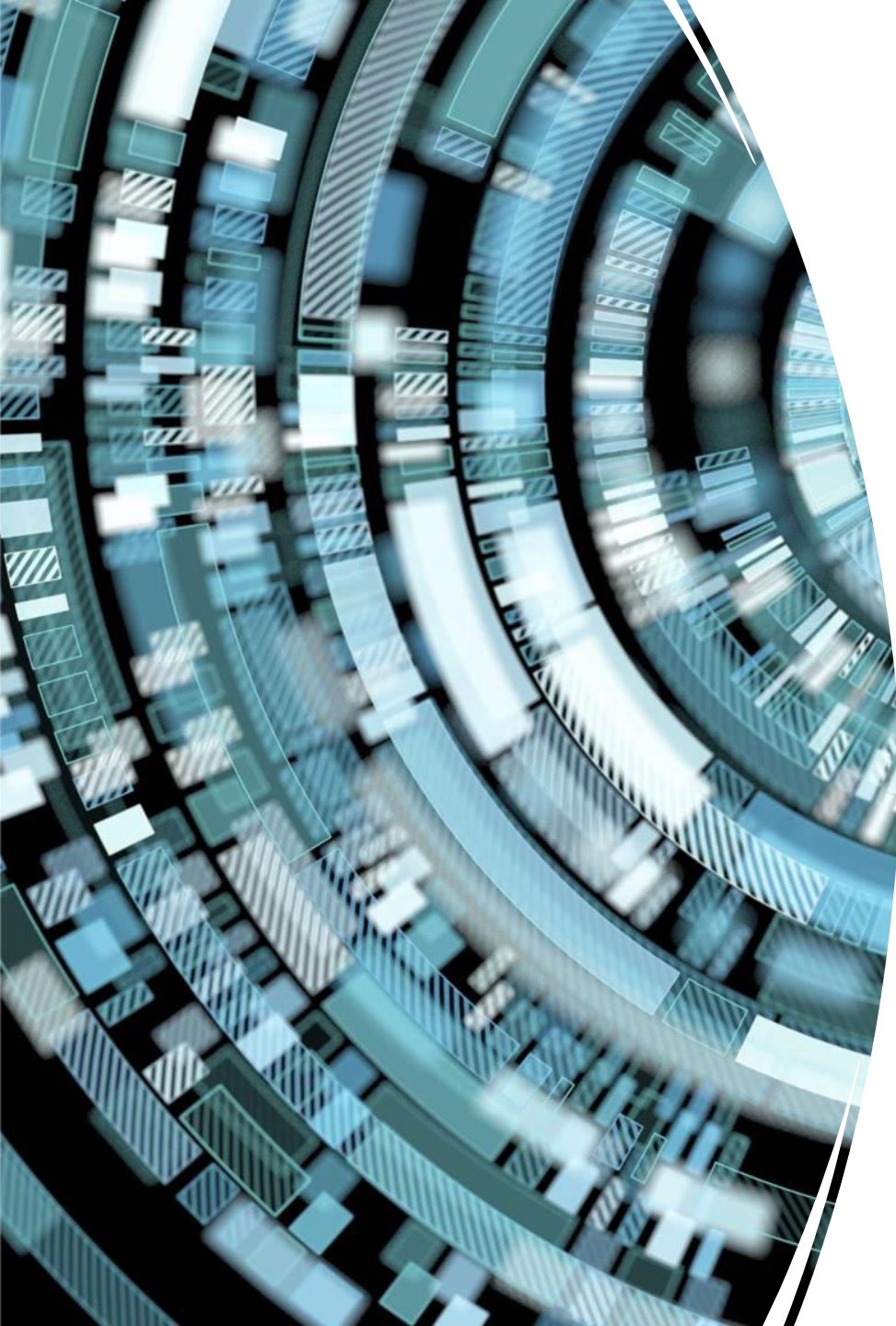
- *“Positing of hypotheses and the rigorous testing of these hypotheses under controlled conditions”* (Schorr 1984)
- This experiment allows **generalization** and **prediction**
- This kind of experiment is the core of the scientific method and, thus, of replicability
- Is it relevant in industry?



Controlled experiment in applied data science

- As mentioned at the beginning, in applied data science a proper controlled experiment is performed only for some applications
- Usually, in this context a controlled experiment takes the name of *A/B testing* and involves some *sample design*, both for digital (e.g., experiments on the web) and physical scenarios (e.g., surveys)





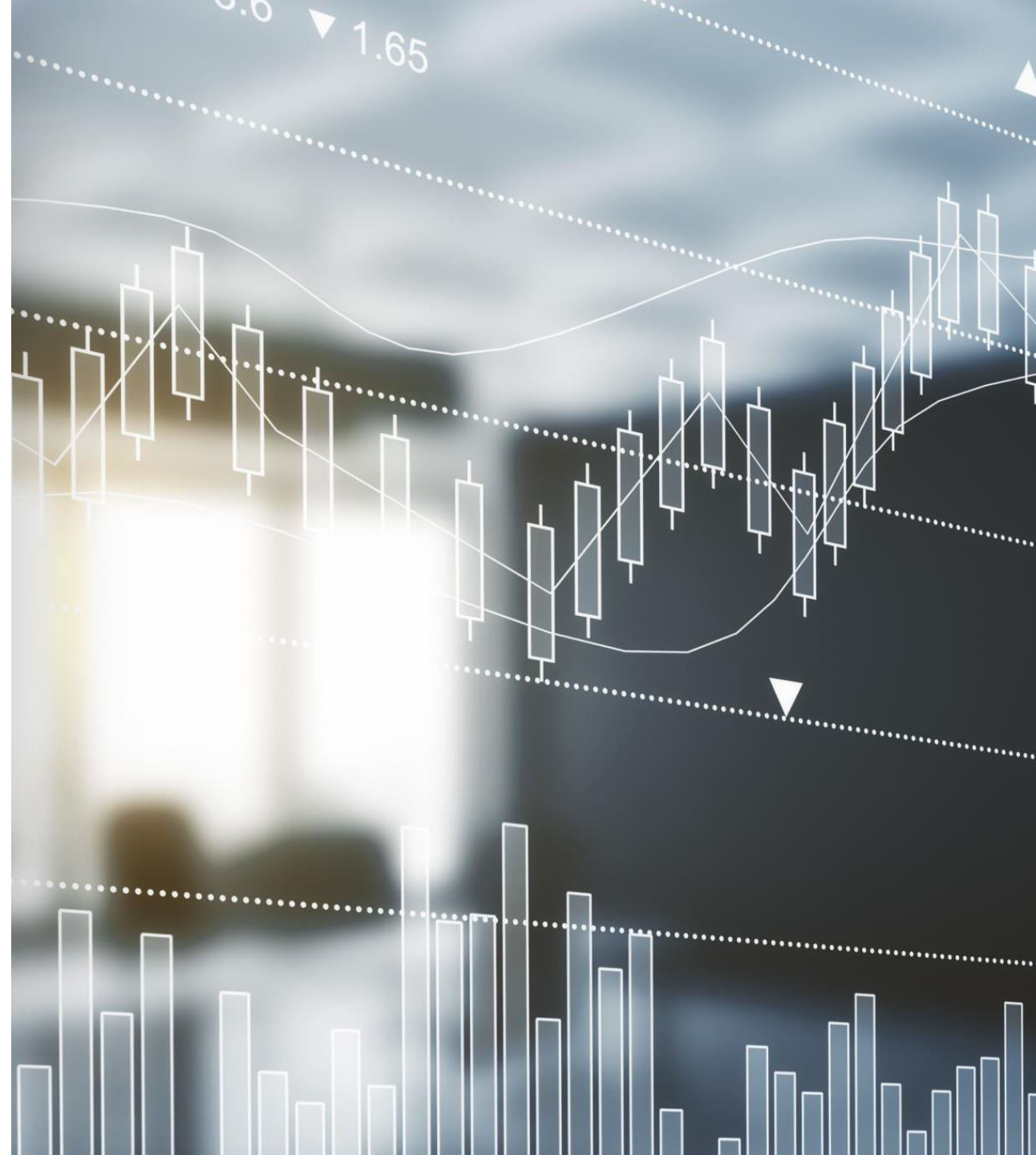
Analyses that are not controlled experiments

- Apart from Controlled Experiments, a Data Scientist performs different kinds of analyses, whose replicability is uncertain.
- These analyses can be classified as:
 - *ad-hoc “data queries”*: find a specific information in an observed dataset X (“what is the most sold product this month?”)
 - *supervised machine learning*: predict/classify Y given an observed dataset X (“train a model to detect spam email”)
 - *unsupervised machine learning*: find interesting patterns in an observed dataset X (“find clusters of customers in our datawarehouse”)
 - *statistical models on observed data*: model Y given an observed dataset X in order to interpret the effect of X on Y (“what is the effect of this kind of marketing campaign on my customers’ satisfaction?”)
- I ignore here the applications of generative deep learning models and reinforcement learning, that are fundamentally different.

“Experiments” that are not controlled experiments

In applied data science, especially in machine learning, the term *experiment* can refer to different things:

- comparing different architectures of the same model on the same data, to find the best performing architecture (*hyperparameters optimization*)
- comparing different models on the same data, to find the best performing model (*model selection*)
- comparing different model transformations, to find the most performing one for a given model (*features engineering*)



Generalizability as a more nuanced concept

- As we have seen, most of the analyses in applied data science are not examples of Controlled Experiments.
- A finding (interpretative or predictive), result of a data science project, is usually tied to the *observed* data used to produce it.
- This means that another Data Scientist cannot re-create the conditions that produced the data used for the analysis (i.e., replicate the result)
- In this setting, practitioners aim to make their result *generalizable* in the sense of getting “similar” results when using new data coming from the same distribution.





Object

Professionals involved

Reproducibility



Replicability

ETL/ELT

- Data Engineers
- Analytical Engineers

- Possible
- Very Difficult
- Not always addressed

- Does not apply

Ad-hoc data queries

- Data Analysts
- Analytical Engineers

- Possible
- Difficult
- Rarely addressed

- Does not apply

Machine Learning Models

- Machine Learning Engineers
- Data Scientists

- Possible
- Reasonably easy
- Addressed

- Possible with limitations (generalizability)
- Pursued

Statistical Analyses on observed data

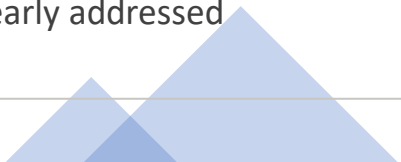
- Data Scientists

- Possible
- Reasonably easy
- Not always clearly addressed

- Possible with limitations (generalizability)
- Not pursued as main goal

Statistical Analyses with sample design (controlled experiments)

- Data Scientists

- Possible
 - Reasonably easy
 - Not always clearly addressed
- 

- Possible
- Pursued

Issues and remedies

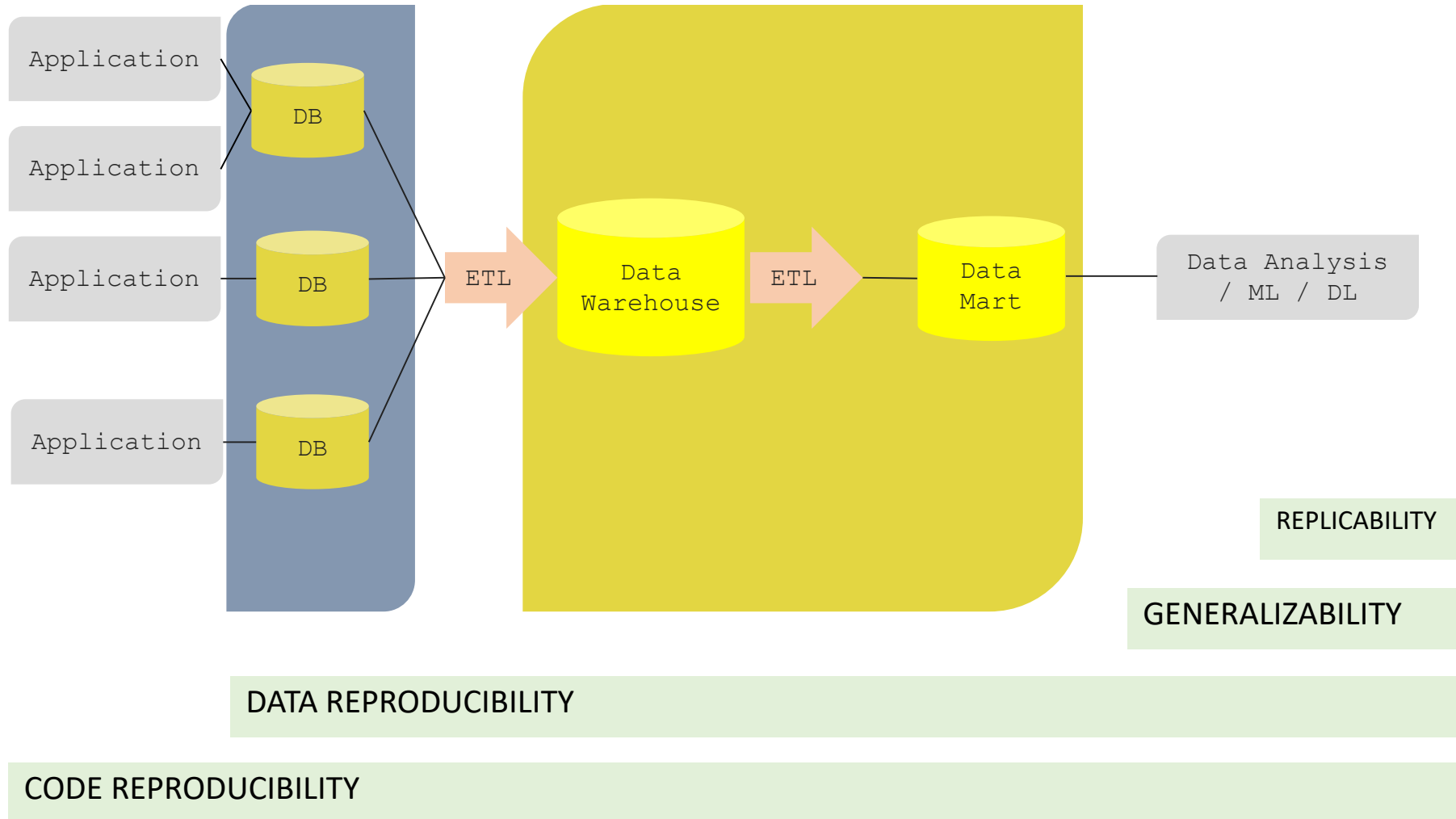
A personal concepts taxonomy

Keeping in mind the dynamic setting of a production system, we can differentiate the concepts analyzed so far as follows:

- *Code reproducibility*: is the codebase used for my project producing the same results, given the input data?
- *Data reproducibility*: are the data transformations producing the expected output?
- *Generalizability*: do my results hold also when new data from the same distribution is used?
- *Replicability*: can my result be obtained by another team in another company using the same experimental setting?



Where can we achieve what?



Threats for reproducibility

- Data Sources change structure and/or meaning or new Data Sources are added
- Components of the ETL/ELT pipeline are updated or new components are added
- Proliferation of untracked ML «experiments»
- Proliferation of data transformations (pre-processing; features engineering)
- Poor collaboration workflows in big teams



Threats for generalizability/replicability

- Overfitting
- Wrong application of cross validation
- Information leakage during models training
- Data preprocessing and features engineering performed in the wrong way
- Hyperparameters optimization performed in the wrong way
- Wrong sample design



DevOps

WHAT

Rigorous management of codebase, through systems that make possible to track changes, coordinate collaboration, favour portability.

MAIN CONCEPTS AND TOOLS

- Code Versioning (git)
- Git workflows
- Issues Tracking
- OOP
- Dockerization

WHICH PROBLEMS ADDRESSES

Code Reproducibility

ROLES INVOLVED

- Software Engineers
- Data Engineers

Test Driven Development (TDD)

WHAT

Approach to software development that prioritize the creation of precise tests, in order to assess the correct application behavior when changes are introduced.

WHICH PROBLEMS ADDRESSES

Code Reproducibility

MAIN CONCEPTS AND TOOLS

- Unit tests
- Integration tests
- CI/CD

ROLES INVOLVED

- Software Engineers
- Machine Learning Engineers
- Data Engineers

DataOps

WHAT

Application of DevOps principles to ETL/ELT pipelines

WHICH PROBLEMS ADDRESSES

Data reproducibility

MAIN CONCEPTS AND TOOLS

- Dbt
- SQL pipelines testing
- Data Lineage
- Data versioning

ROLES INVOLVED

- Data Engineers
- Analytical Engineers

MLOps

WHAT

Application of DevOps principles to Machine Learning «experiments»

WHICH PROBLEMS ADDRESSES

- Data Reproducibility
- Generalizability

MAIN CONCEPTS AND TOOLS

- Experiments Tracking
- MLFlow
- Feature Stores

ROLES INVOLVED

- Machine Learning Engineers
- Data Scientists

Training-Validation-Test and Cross Validation

WHAT

Workflows and techniques used to train a supervised model ensuring its generalizability to unseen data.

WHICH PROBLEMS ADDRESSES

Generalizability

MAIN CONCEPTS AND TOOLS

- Model assessment and selection
- Bias-Variance Decomposition
- Overfitting and underfitting
- Model complexity

ROLES INVOLVED

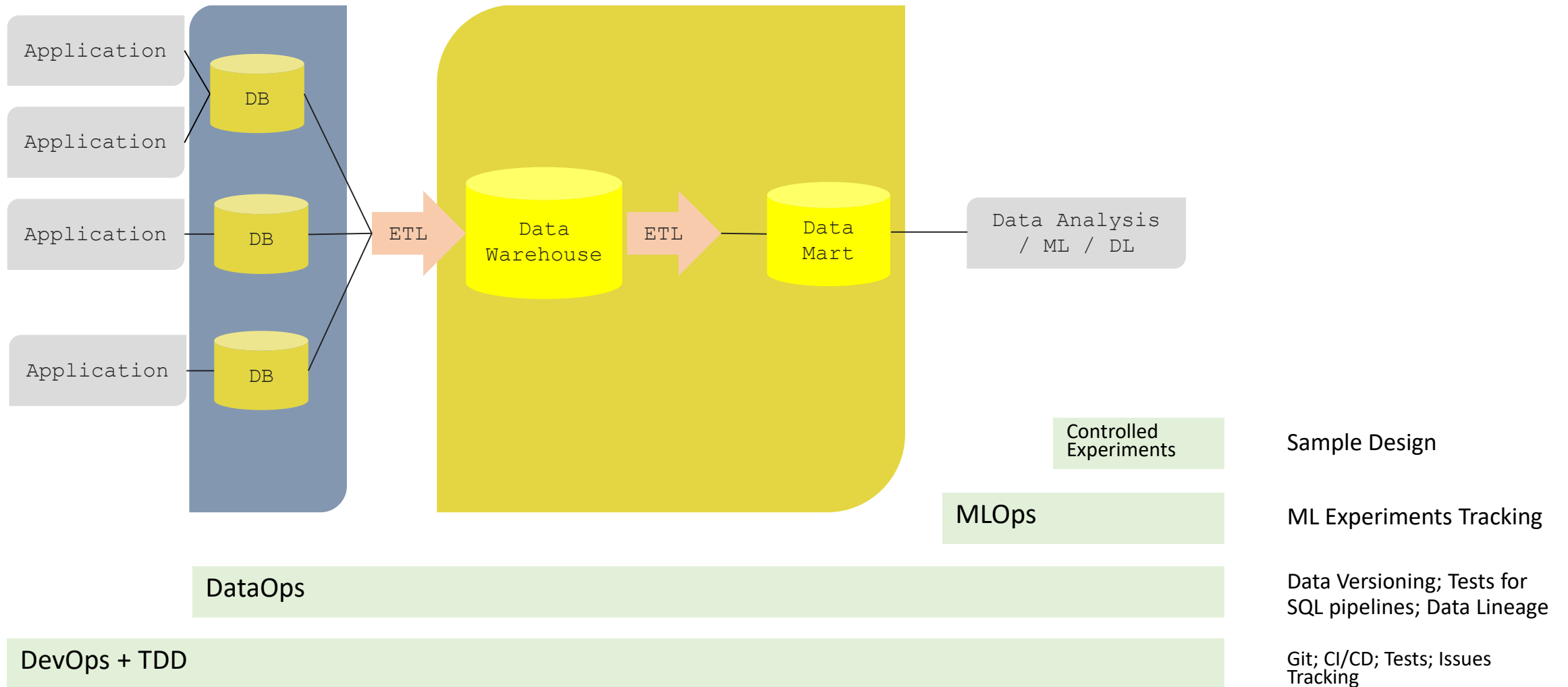
- Data Scientists
- Machine Learning Engineers

	Code Reproducibility	Data Reproducibility	Generalizability	Replicability
DevOps	✓			
Test Driven Development (TDD)	✓			
DataOps		✓		
MLOps		✓	✓	
Training-Validation- Test split			✓	
Cross Validation			✓	
Sample design for controlled experiments				✓

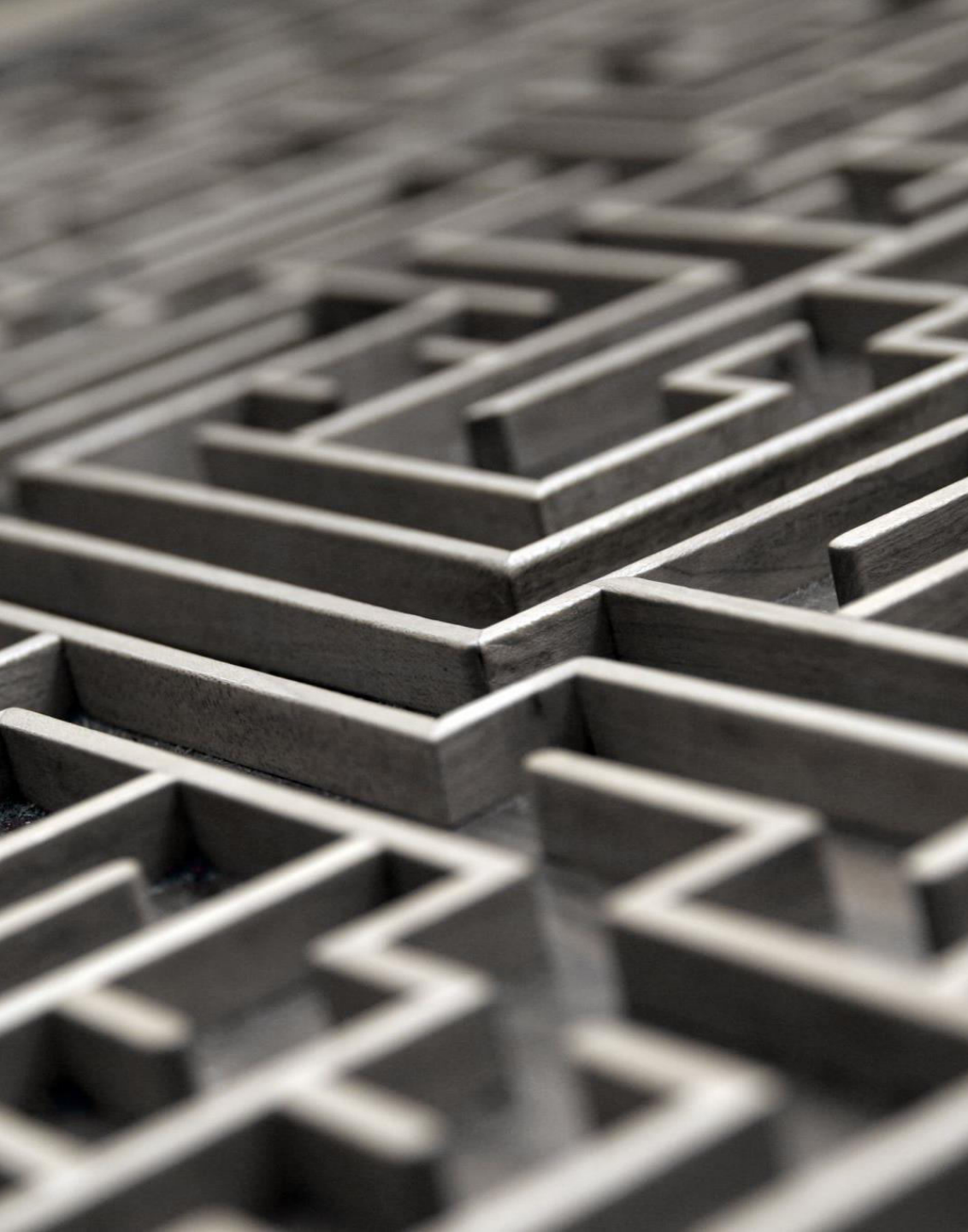
Is all of this actually used?

- Despite an abundance of tools and approaches, many teams in many companies still lack even the fundamentals, posing a serious threat to generalizability but often to reproducibility too.
- This situation is the result of company cultures that don't see data as a primary asset: teams don't have the right incentives that would make reproducibility and replicability a "KPI" to aim for.
- Resistance to innovation is also common: adopting new ideas, workflows and tools is demanding.

Making the pipeline reproducible and replicable



Wrapping up...



Conclusions

- Applied Data Science is a dynamic and complex scenario where it is hard to apply concepts like reproducibility and replicability without further considerations
- Foundational work has to be done to better understand issues and propose good remedies
- Resistance to adoption of innovation and misaligned incentives remain the biggest threats to reproducibility and replicability in applied data science
- Given the ubiquitous influence of data on our lives and our society, improving this situation is important and urgent

Thank you for your attention!

References

- Barba, L.A. (2018). Terminologies for Reproducible Research. arXiv, 1802.03311. Available: <https://arxiv.org/pdf/1802.03311> [December 2018].
- Radder, H. (2009), “The Philosophy of Scientific Experimentation: A Review.” Automated Experimentation, 1(2) doi:10.1186/1759-4499-1-2.
- Schiaffonati, V. (2023), “Computers, robots, and experiments. Reproducibility and beyond”, presentation from 1st BRIO Meeting, 17th September 2022, Milan.
- Schorr, H. (1984). Experimental computer science. Annals of the New York Academy of Sciences, 426(1):31–46.
- Tedre, M. (2015). The Science of Computing. Boca Raton: CRC Press, Taylor & Francis Group.

Backup

1. Demonstration of feasibility

- We find this first type of experimentation in Data Science under the terms *prototyping* and *Minimum Viable Product (MVP)*
- Here the aim is only showing, often to stakeholders that have budget to “buy” the idea, that it is possible to accomplish a specific result through data.
- “Experiment” is synonymous of untested, novel innovation.
- It seems more connected to reproducibility than to replicability.

2. Trial run

- The goal is to find out how well a prototype or complete system works; not hypothesis-driven
- In data science the concept of *test environment* is often used: it is the corresponding of a lab setting in science. In this environment, models are tested mainly to check bugs and to assess their *reproducibility*

3. Field experiment

- Evaluation of those requirements relative to the system's surroundings outside of the lab
- The system is tested in a live environment and is measured for performance, usability attributes, robustness
- In Data Science, the corresponding concept is the *staging environment*, an environment that mimics the *production environment*, the real-world scenario where the data product will operate

4. Comparative experiment

- A comparison between solutions and set-up to measure and compare one solution with a competing solution with the same data set and parameters
- In Data Science this can be seen when *competitors analyses* are performed (sometimes upon management's request) in order to justify an investment in in-house development
- On a more academic side, machine learning papers spend plenty of space in comparing novel model architectures against *benchmarks*