Data pipelines for ML

Gjensteforelesning for BI GRA4157

Av

Yngve Sture Christian Aalby Svalesen

BearingPoint®

Intro to Bearingpoint

5 min

Value driven design

10 min

ML usecase workshop

10 min

MLOps

20 min

Modern data platforms

15 min

Plaace Example

10 min

ML system design workshop

10 min

Practical experiences

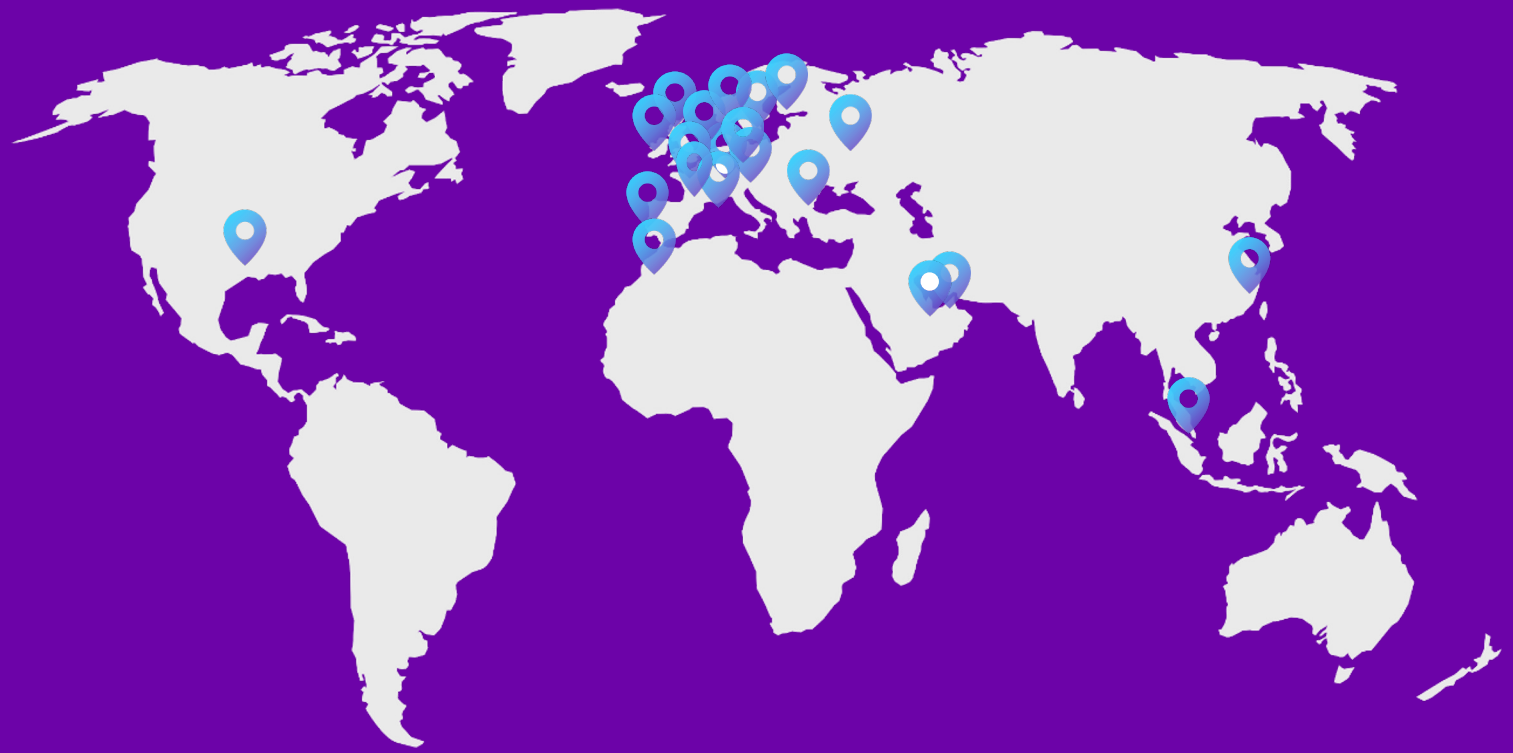
10 min

BearingPoint®

Intro to BearingPoint

BearingPoint®

We have European roots, with global reach



4 261
employees globally and 150 in Norway

50
countries were we consult clients

41
Offices in 23 countries

€738m
In revenue

BearingPoint®

DATA & ANALYTICS

DATA
ADVISORY

DATA
ENGINEERING
& INSIGHT

DATA
SCIENCE & AI

Leverandør- og teknologi-

uavhengig

rådgiver

70+

konsulenter i Oslo

20+

års erfaring

BearingPoint®

DATA & ANALYTICS

DATA ADVISORY

DATA ENGINEERING & INSIGHT

DATA SCIENCE & AI

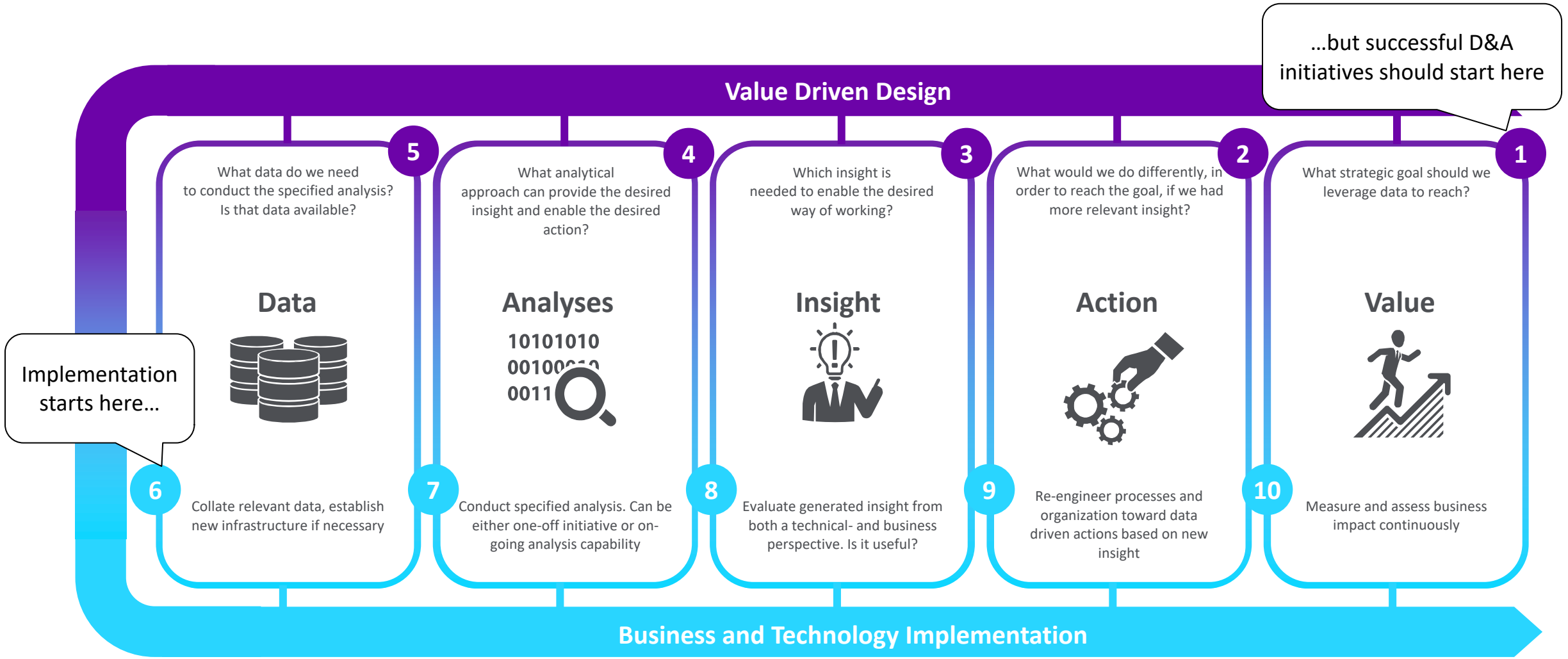
Helsedirektoratet
Mills
NorgesGruppen
ASKO – vi forsyner Norge med mat
Statnett
RiksTV
ELKJØP
Nortura
CERMAQ
Gjensidige
PLAN INTERNATIONAL
ice
Schibsted
DNB
REITAN CONVENIENCE
BANE NOR
NOKAS
NHST MEDIA GROUP
SpareBank 1
coop
FINN
DN Dagens Næringsliv
VG
Finansdepartementet
BearingPoint

Value driven design

BearingPoint®

Our approach has the desired business value in focus throughout the initiative …

BearingPoint's Value Driven Data & Analytics Framework



BearingPoint®

ML use case workshop

BearingPoint®

Cases

How to turn value into data

B2B Sales

- Large international cloud provider.
- More than 90% of all revenue comes from B2B sales.
- Security and price are important drivers in the market.
- Many customers still believe cloud computing is dangerous and unsafe.

Strategic goal: Increase revenue by 10% within 3 years.

Telecommunication

- National telecommunications provider.
- Market is fierce with many providers and low margins, making every customer very valuable.
- Retention of customers have proven to be a difficult area in the past where the company now wants to improve.

Strategic goal: Increase market share from 10% to 15% within 2 years.

How to turn your data into value?

Data Value Chain examples

	Data	Analysis	Insight	Action	Value
					
B2B Sales	Sales volume per customer categorized by product group	Demand profile analysis and identification of cross-sales potential	Identification of the product group with highest cross-sales potential per customer	Prioritize high-potential product groups in sales meetings with clients	Increase sales volumes at existing customers
Churn	Customer characterist-ics, transaction history, and behavior data	Root cause analysis aiming to identify the underlying cause of churn	Behavioral patterns in the customer journey that are strongly correlated with churn	Strategic and tactical adjustments to the customer journey	Reduce churn among subscription customers
Maintenance	Machinery failure data, sensor and production data	Develop prediction model for machine part failure using machine learning	Real-time prediction of the risk of machine parts failing	Proactively switch out parts that are about to break	Increased up-time of machine equipment

MLOps

BearingPoint®

AI and machine learning has a vast potential, and companies needs to shift from piloting to scaling up and operationalizing in order to stay competitive

English translation

2-3%

percentage points in **profit margin uplift** in the Nordics by adopting AI[1]

16%

increase in global GDP within 2030 by adopting AI[2]

71%

of companies respond that AI is considered '**an important topic**' on the executive management level[3]

> By the end of 2024, 75% of enterprises will shift from piloting to operationalizing AI [4]

1) how-artificial-intelligence-will-transform-nordic-businesses.pdf (mckinsey.com)
2) The Impact of Artificial Intelligence on the World Economy – WSJ
3) How 277 Major Companies Benefit from AI (Microsoft/EY)
4) Gartner - Trends for data and analytics 2020

BearingPoint®

English translation

The reality for many is still…

85% Of All Data Science Initiatives Fail

- Gartner

BearingPoint®

After a successful start where data driven improvements has provided tangible results, companies often experience a slow down in speed and an inability to scale the initiatives

English translation

The **technology mirage** is avoided …

Employing a **technology driven approach**, procuring data tools and technologies, gathering "all data" and selecting specific machine learning algorithms without first evaluating the business needs.

Leads to expensive, technology-driven projects that lack business buy in, and a lack of value creation.

… and the POC trap is avoided.

A bottom-up approach where passionate sub-groups in the organization identify promising prototypes which can then be rolled out in the business, with a lack of **prioritization** by **business needs** or value potential.

Leads to many POC-er that are not fully launched to or adopted by the business.

Instead, a data strategy that identifies the use cases with the highest potential is used to adapt and select the technology to the business needs.

But are not able to scale due to **a lack of processes and tools to be data driven at scale**. This could be due to:

Lead time and dependency of IT-resources outside of core team to launch new models to production

Changes in input data occur unexpected and is only recognized after models fail or produce unexpected results

Maintenance of existing models demand an increasing share of the data scientist time. *Studies show that 25-50% of time is spent on low value creating tasks.*

Difficult or time consuming to compare experiments when re-training models.

BearingPoint®

MLOps is a way of working and a set of tools that enable scaling advanced analytics by reducing the time from idea to value created, and maintain this speed over time

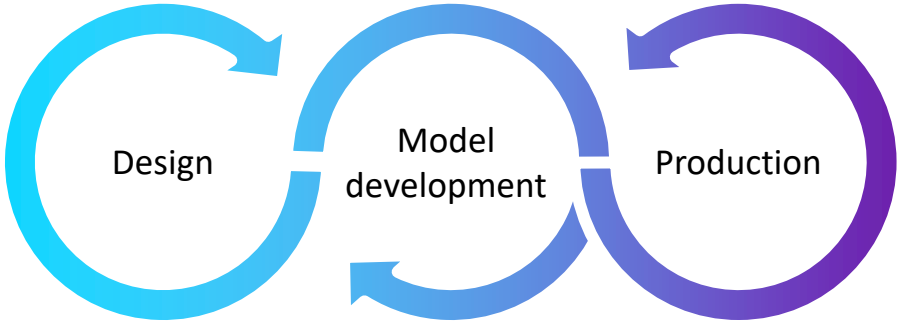
MLOps is a **way of working** and a **set of tools** that decrease the time from idea to value created

The **two main components** of MLOps

Version control of code/workflows, models and data sets

Automation of verification, testing and deployment

The MLOps process consists of **three iterative steps**



- Value->Action->Insight
- Functional and technical requirements
- Prioritization of use-cases

- Data Engineering
- Model development
- Testing and validation of models

- Deployment and integration of models (APIs)
- Monitoring accuracy and measuring value creation
- Automation

The **purpose** of MLOps

Avoid building **technical dept** by focusing on reusable and modularized technical structures

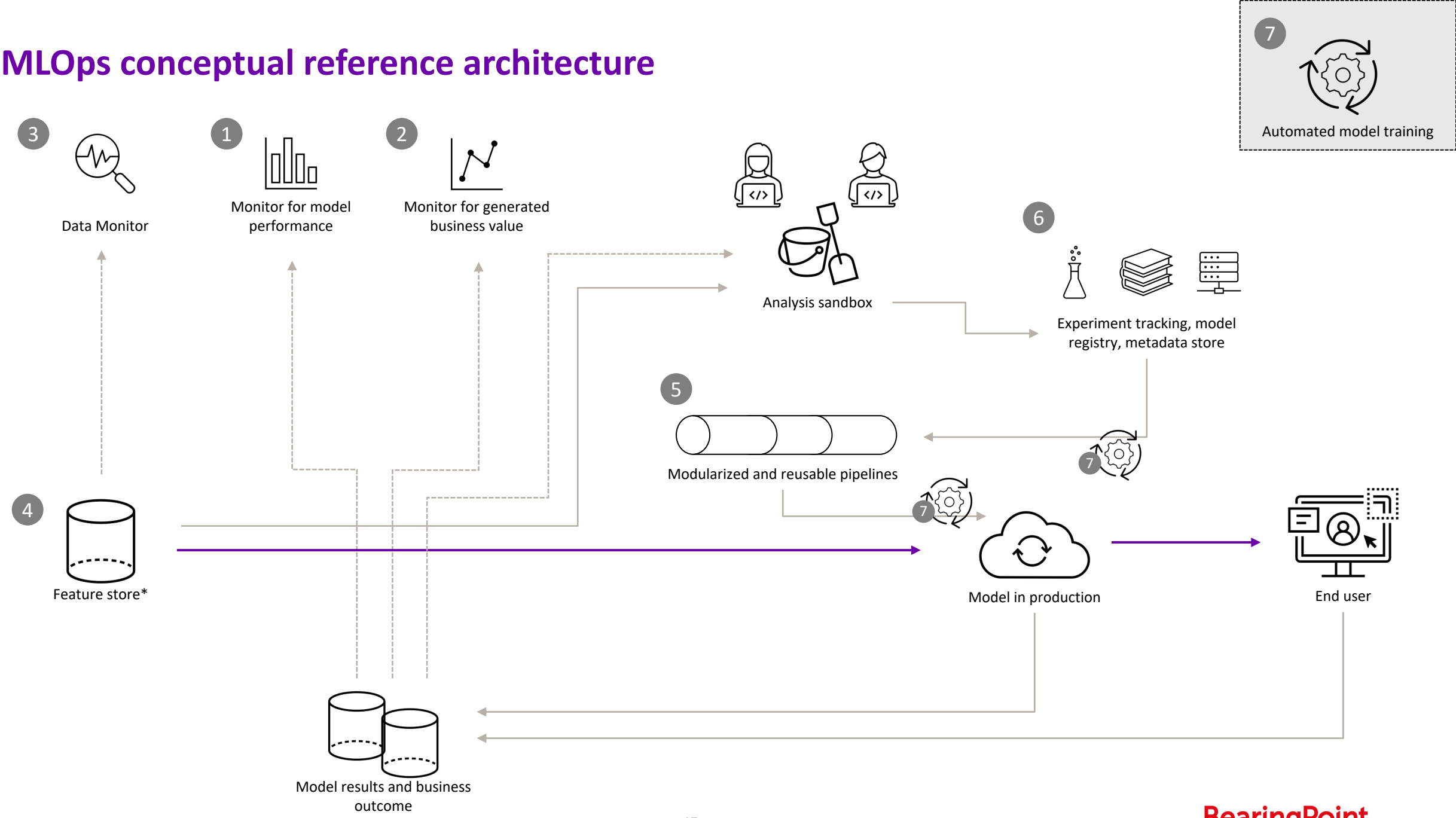
Streamline processes, **esnure quality and reduce risk of errors** and increase efficiency of development and operations

Ensure **governance** by defining clear roles and responsibilities

Reduce ramp up time: Make it easier to get new employees up to speed by having established practices

BearingPoint®

MLOps conceptual reference architecture



BearingPoint®

Technical components and processes in an MLOps-system

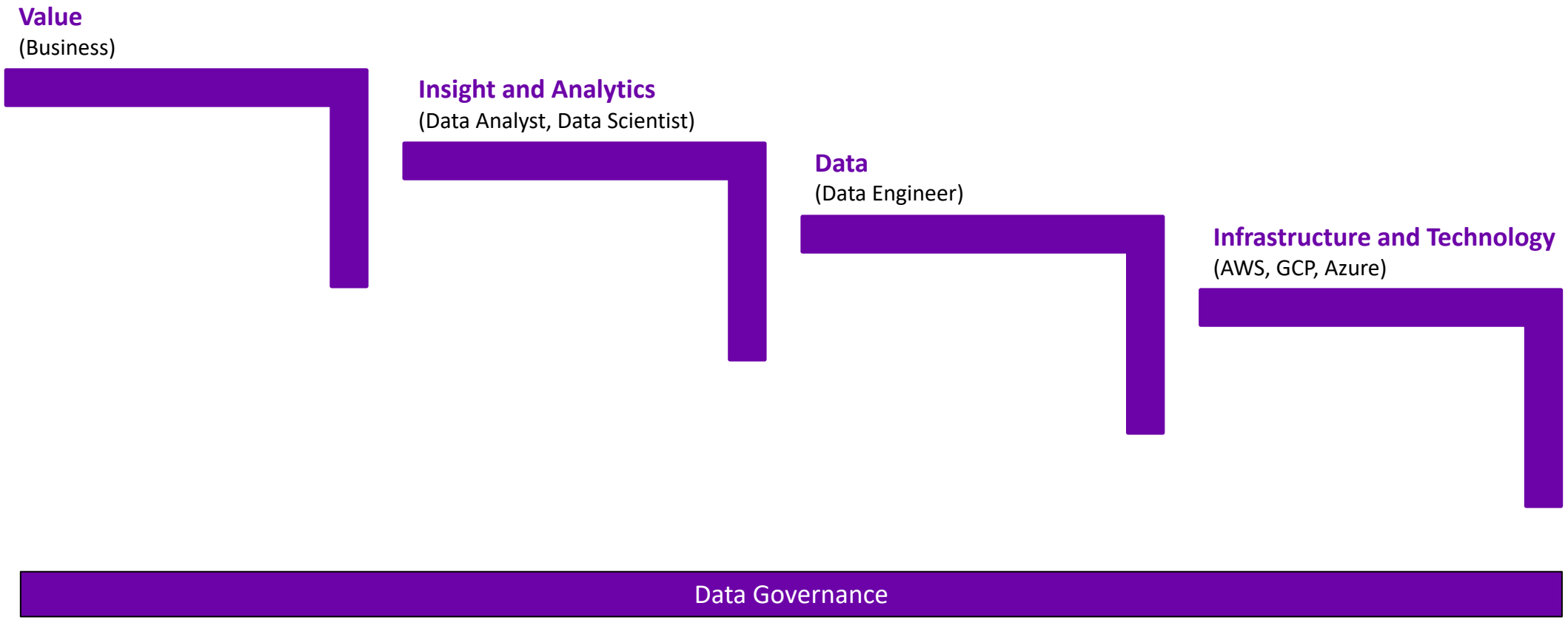
	What	Why	How
1	**Monitor for model performance**	Automated monitoring of model results in production over time.	**Know when it's time to re-train the model** (model decay) due to changes in the domain we are modeling and continuously track accuracy as this is a key parameter on the model's business case.	Dashboard to track model KPIs over time. Should include capabilities to compare competing models for the same target (where only one is actually in use). Might require control groups.
2	**Monitor for generated business value**	Automated monitoring of generated business value. E.g., monitoring if business actions based on model results generate the expected value.	**Generated business value** is a key parameter to track for the cross functional team when **following up business side actions** and creating a positive engagement behind the necessary changes in ways of working.	Dashboard for generated business value. Will typically require experiment/control groups and A/B-testing of business side actions.
3	**Data monitor**	Automated monitoring of errors and data drift in model input data, e.g., monitoring characteristics like mean, standard deviation, NULL's, min, max, etc.	Early **detection of drift in characteristics** that will require model re-training. **Detect errors and deviations** as early as possible, e.g., when changes in data sources are not announced to all teams.	Dashboard on input data (feature store). Should be prepared to connect this as a trigger to CI/CD-tools for automated model re-training. There are also dedicated tools for this (DataOps).
4	**Feature store***	One common, well documented sett of features that are used as data input to the models. Also known as analytical base table (ABT).	Avoid developing the same features multiple times for different models will **save time in data development and maintenance** and processing capacity. Re-use can also **increase quality**.	Implementation options will depend on requirements for latency. For many use cases it's a good option to implement this as (a) wide table(s) with one row per subject for model prediction.
5	**Modularized and reusable pipelines**	Common naming conventions and structures in code. Separation of concerns in code, e.g., data split into training, test and validation could be one separate and reusable code block.	**Improve and increase cooperation** between data scientists. Save time when getting new hires up to speed. Easier maintenance and improvements and **significant time saving** by avoiding duplicate work.	Common naming conventions, common conventions for code documentation. Use of version control systems (like git).
6	**Experiment tracking, model registry, metadata store**	Experiment tracking, model registry, metadata store are separate functional components, but often delivered in one tool.	Enabling **reproducibility in experiments** and modelling across environments (personal sandboxes, dev/test/prod). This enables learning and transparency in the model selection and development process.	The tool should have a graphical user interface and a programmatic interface for conducting analysis and for use with an automation tool (CI/CD). Example technologies are MLFlow or Azure Machine Learing.
7	**Automated model training**	Automated model training based on a trigger from *data monitor* or *monitor for model performance.*	**Save resources / time for data scientists** (avoid making them all ML engineers). Increased frequency of model re-training ensures that the model better represents the current reality.	Utilize CI/CD-tools to automate the steps above that are prepared for automation.

*Feature store can have two complexity levels. Level 1: Batch data where data warehouse speed is sufficient. Level 2: Streaming data (for use cases where this is a requirement).

Modern data platforms

BearingPoint®

What should a data platform enable?



BearingPoint®

Key principles when building a modern data platform

Value driven implementation

- Starts with business objectives and needs
- Identify the insight needed to create business value
- Avoid common pitfalls as starting with the technology and building complex solutions which does not create business value or is needed.

Modular Architecture

- Modularity provides flexibility as single components can be replaced or changed without affecting the rest of the architecture.
- Domain driven architecture where domain teams provide their capabilities as building blocks to other teams
- Increase scalability and reduce dependencies

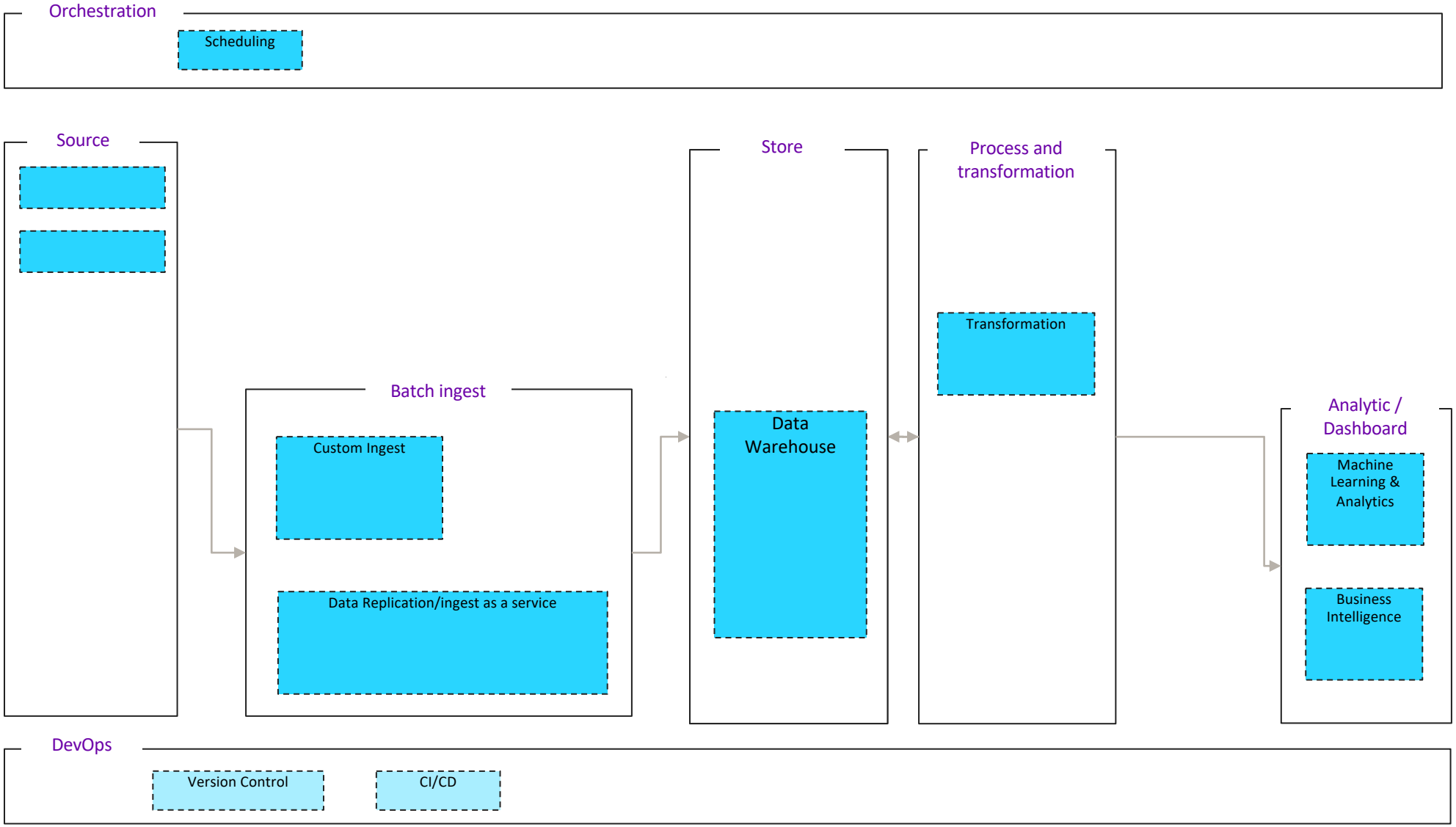
Automated and collaborative development

- Improve collaboration and version control
- Automate deployment through CI/CD
- Test logic and pipeline before pushing into production
- Enable short-cycles and incremental development

Flexible and scalable data storing and processing

- Full flexibility in use of source data
- The ELT approach enables faster implementation than the ETL process
- Decoupling of loading and transformation does not halt the load when having errors in transformations.
- Loading raw data directly enables more control of the data and debugging is easier

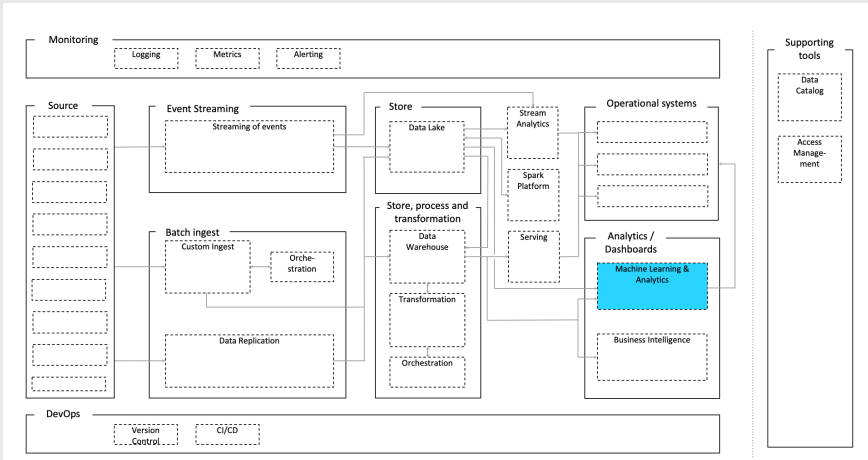
Basic components in a modern data platform



BearingPoint®

Analytics/Machine learning

- Machine learning “gives computers the ability to learn without being explicitly programmed”. This is possible today as data is more accessible than before.
- There are five important steps in a common ML workflow, get data, clean, prepare and manipulate data, train model, test model and improve. There are several tools that help you in handling the infrastructure around your ML code.
- This is a large topic, and we are only touching upon it in this material where we are looking at tools that provide infrastructure around you ML-code, for more information on how to operationalize you ML code see our MLOps framework – An introduction to MLOps by BearingPoint.



Cloud native

- Serverless compute services that provides updated infrastructure and all the resources needed to run your application.
- You write the (model) code yourself, and then the service handles the rest.

Tools:

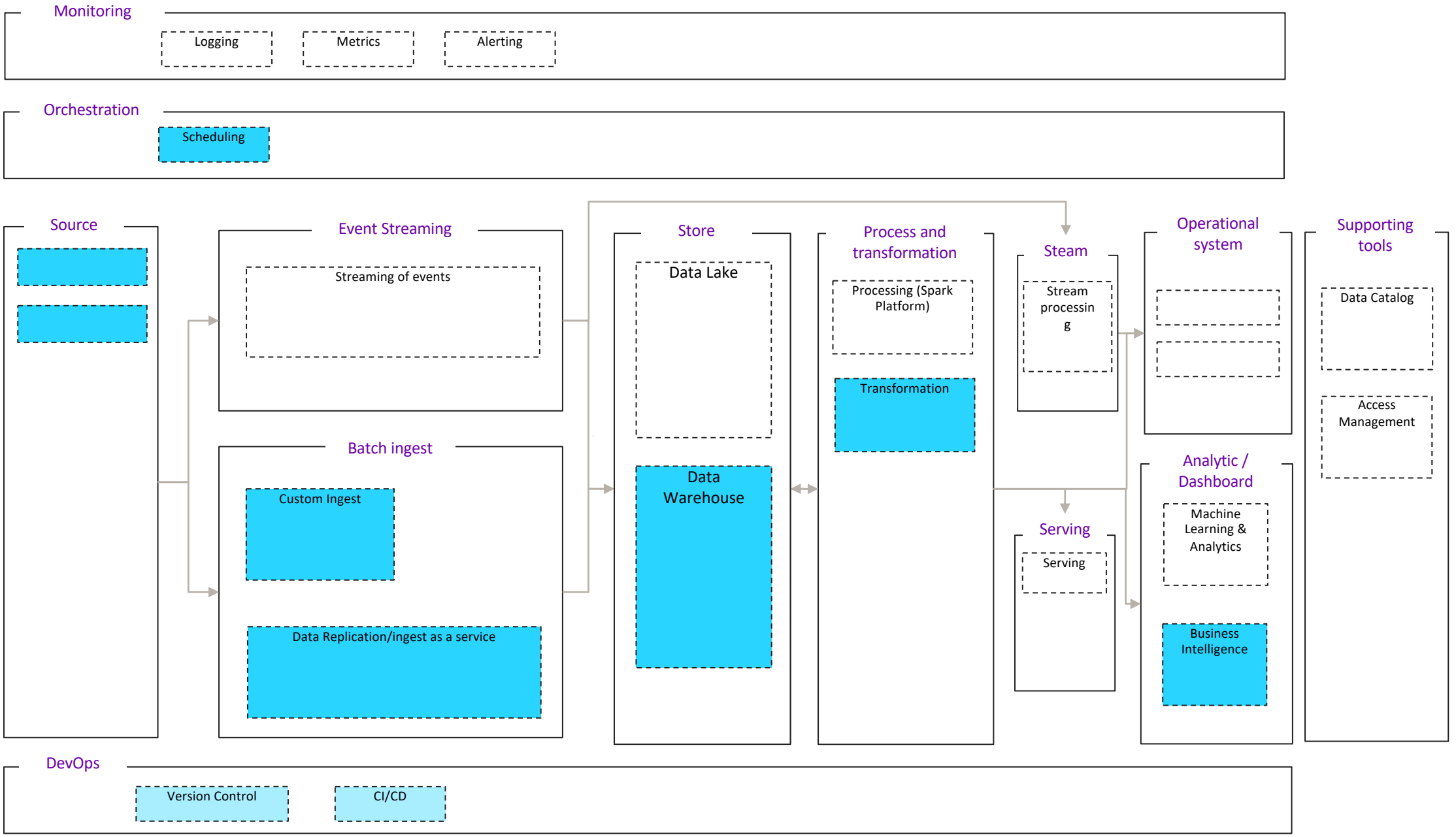
- Azure Functions
- AWS Lambda Functions

(Databricks +) mlflow

- mlflow open source platform to manage the ML lifecycle
- Designed to work with any ML library, algorithm, deployment tool or language
- Can run a hosted version of mlflow on Databricks

Other options

- Self-deploy on PaaS (Kubernetes/ECS)

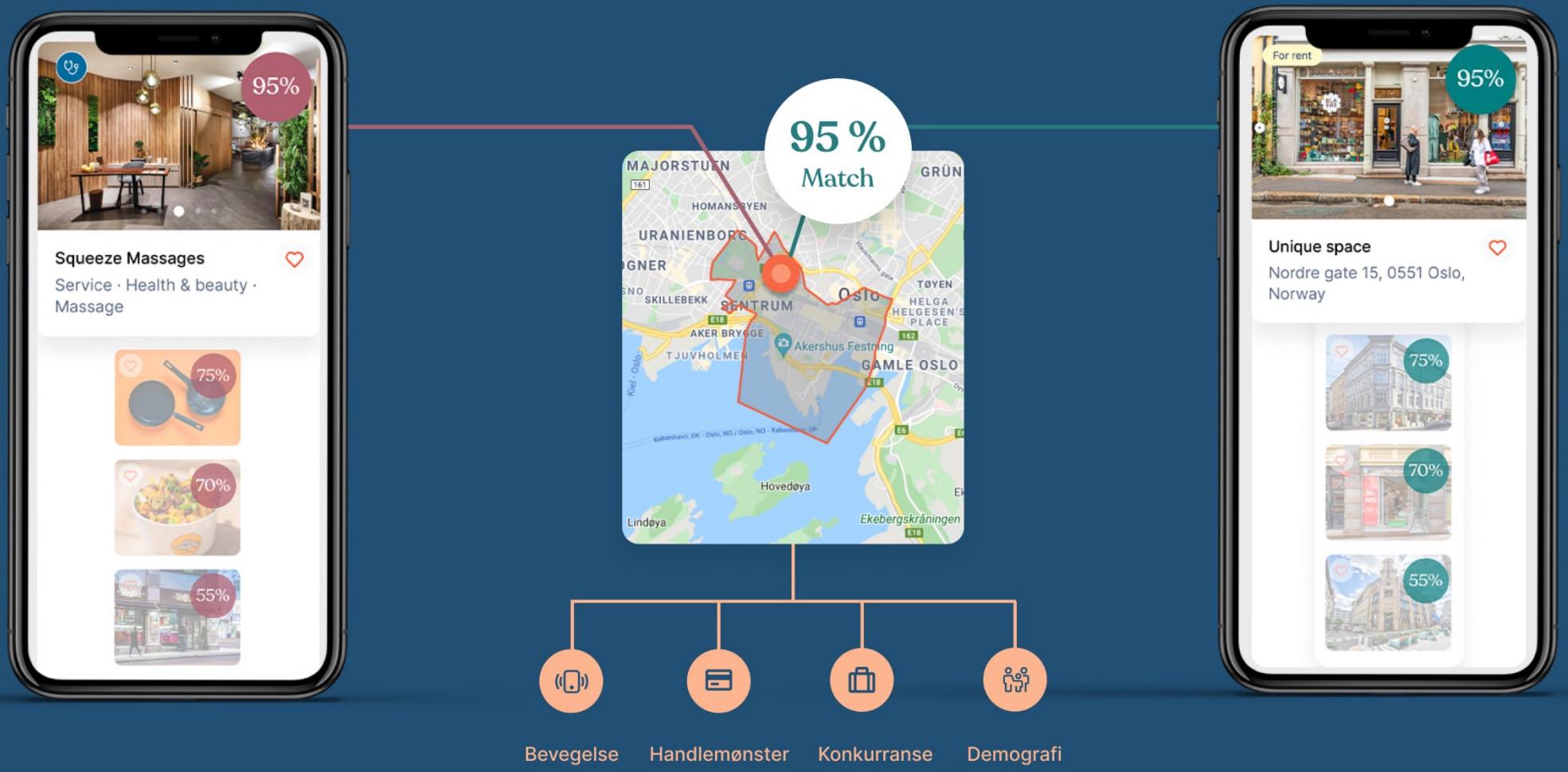


BearingPoint®

MLOps & Data platform at Plaace

BearingPoint®

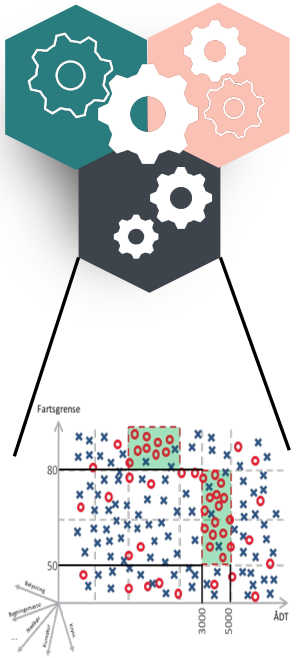
Plaace: neste generasjon matching platform for retail property



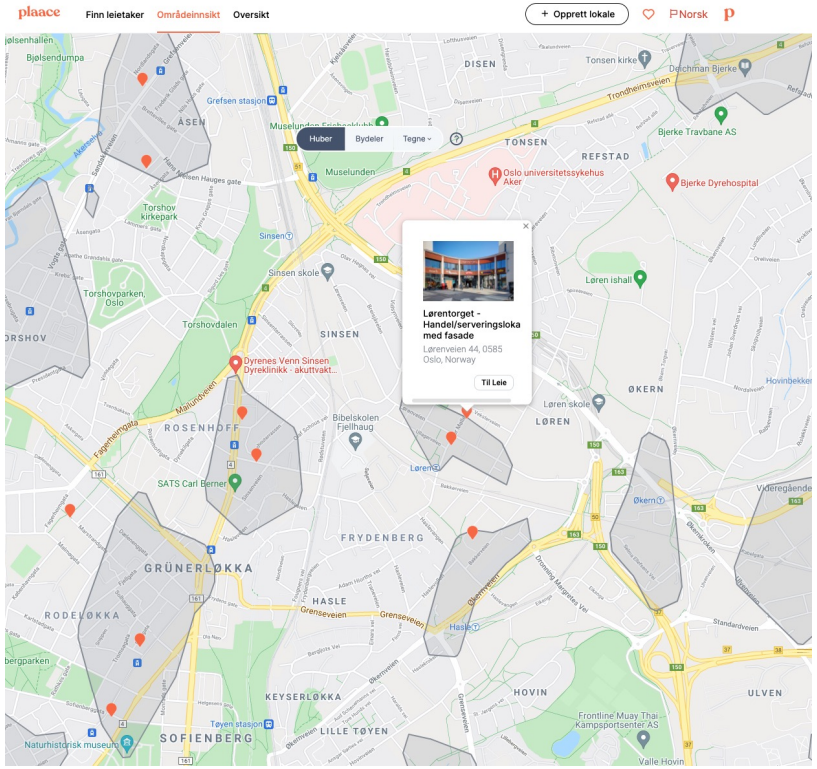
Plaace develops automatic area recommendations with machine learning

Input

Recommendation model

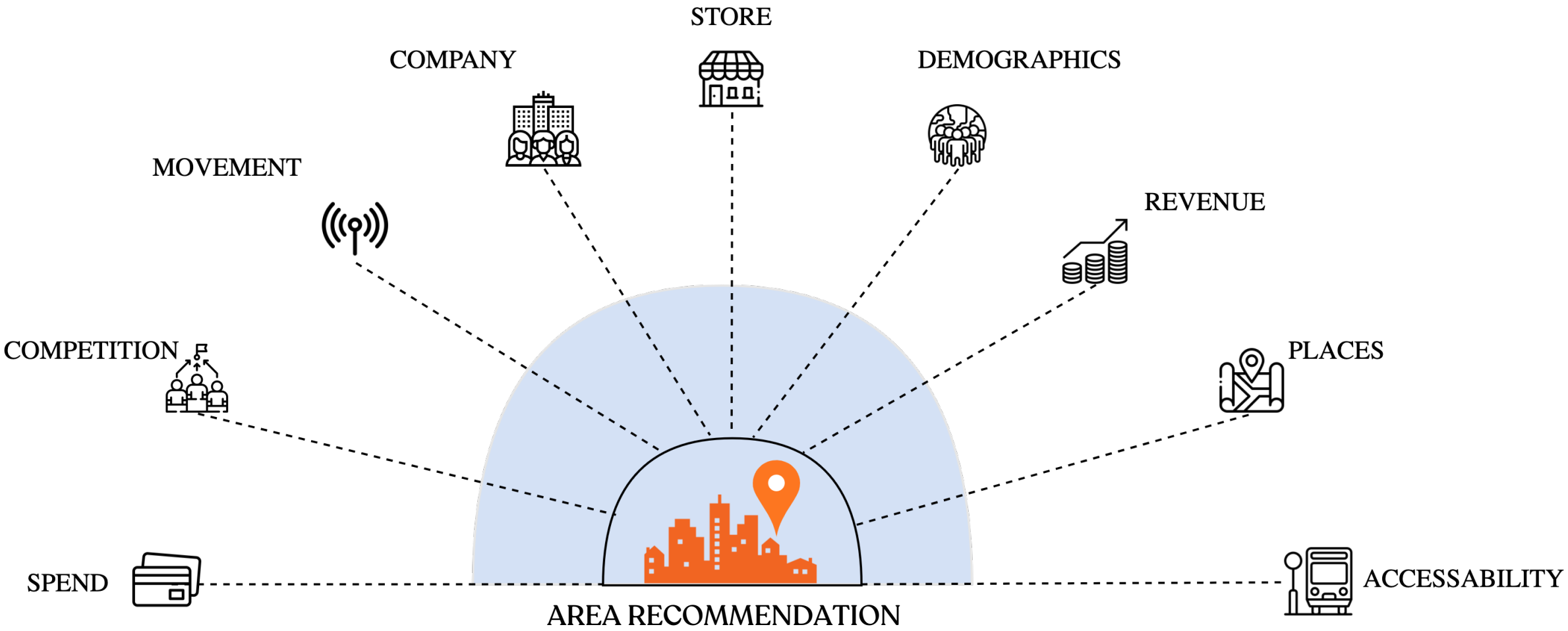


Recommends grunnkretser with highest "match" score



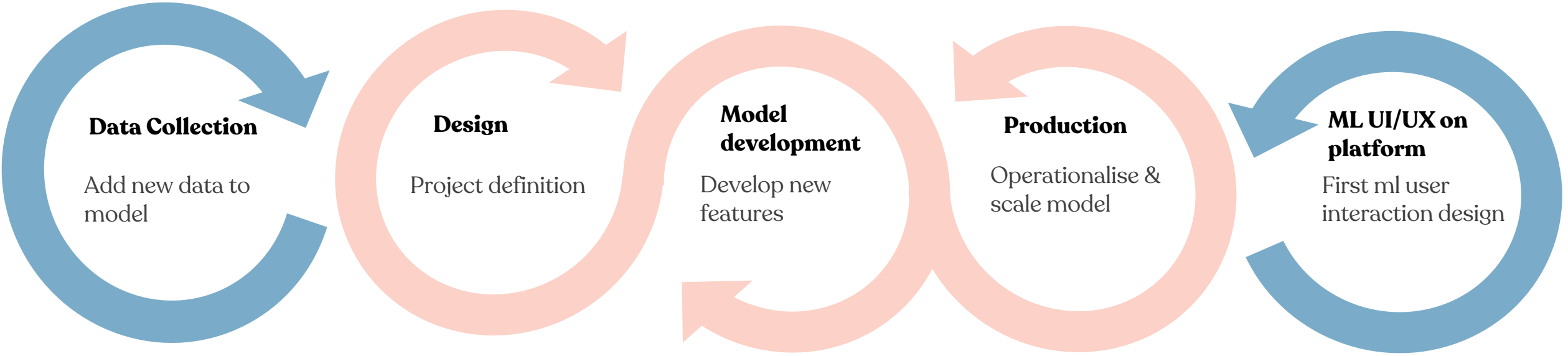
BearingPoint®

A recommendation is based on several factors



BearingPoint®

Vi use MLOps as work method



Organizational Data & Decision Support

Data Pipelines & Quality

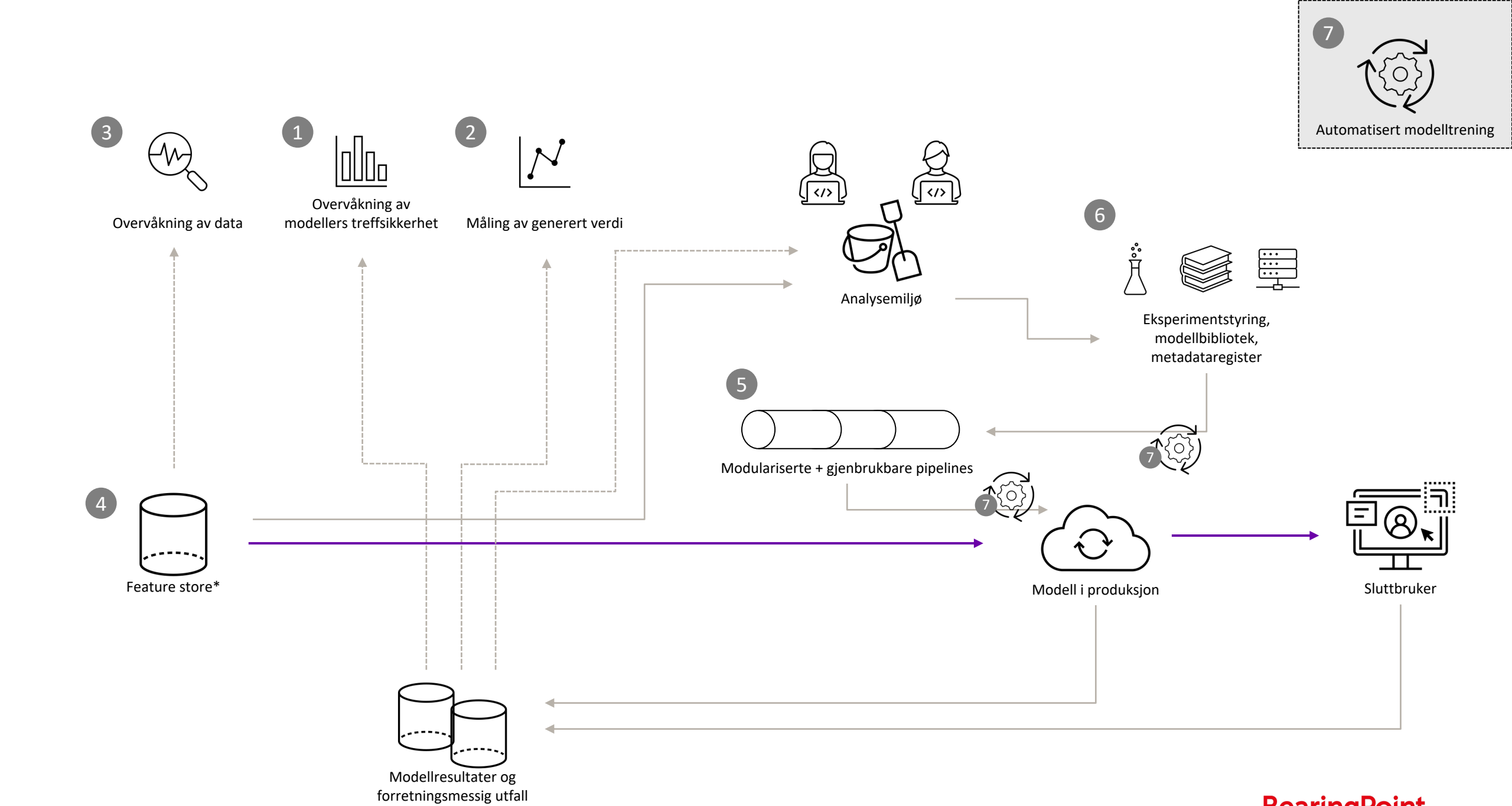
Infrastructure & Tooling

Team & Processes

BearingPoint®

ML system design workshop

BearingPoint®



BearingPoint®

Practical experiences -

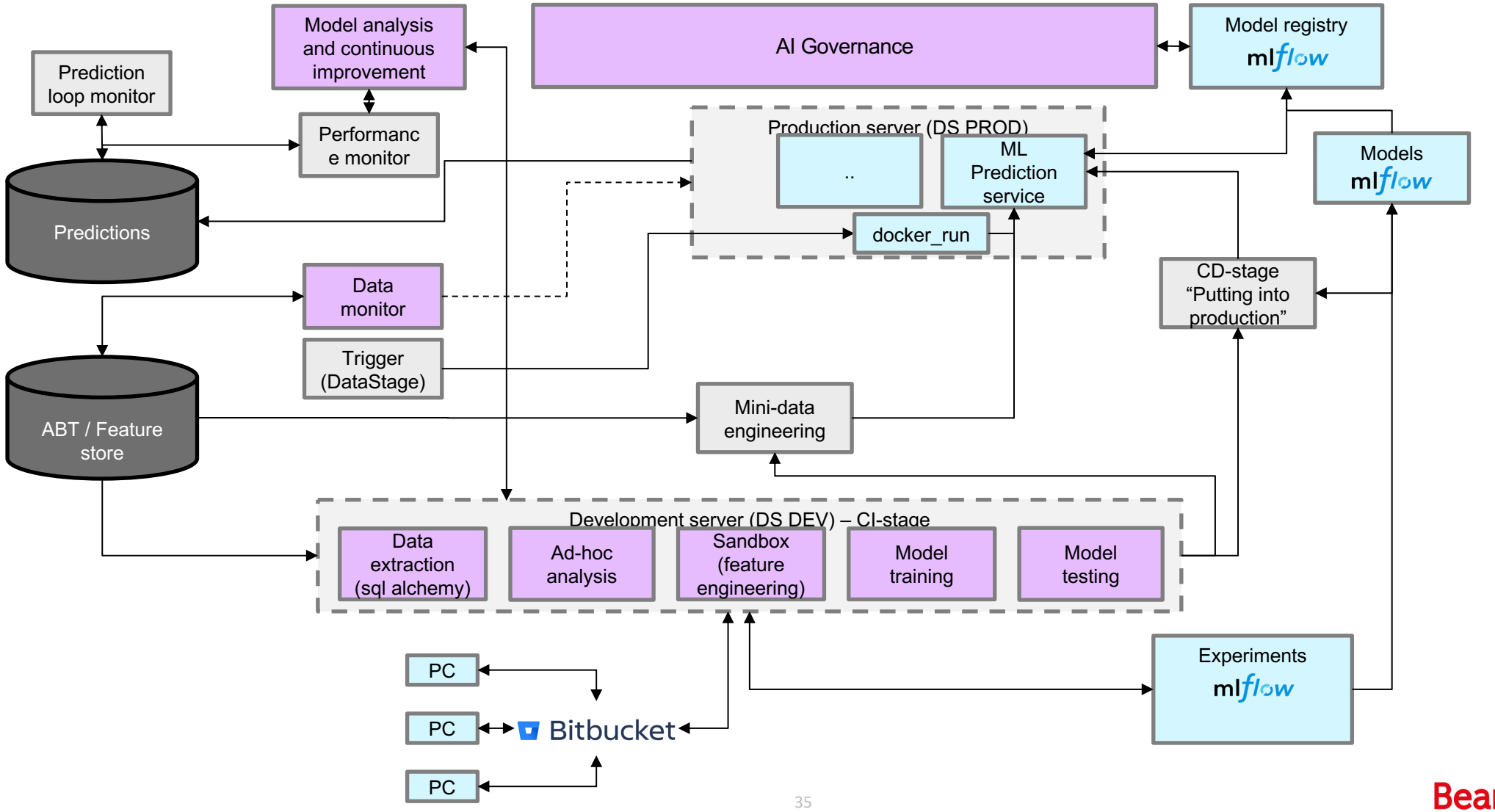
What do you wonder about?

BearingPoint®

BearingPoint®

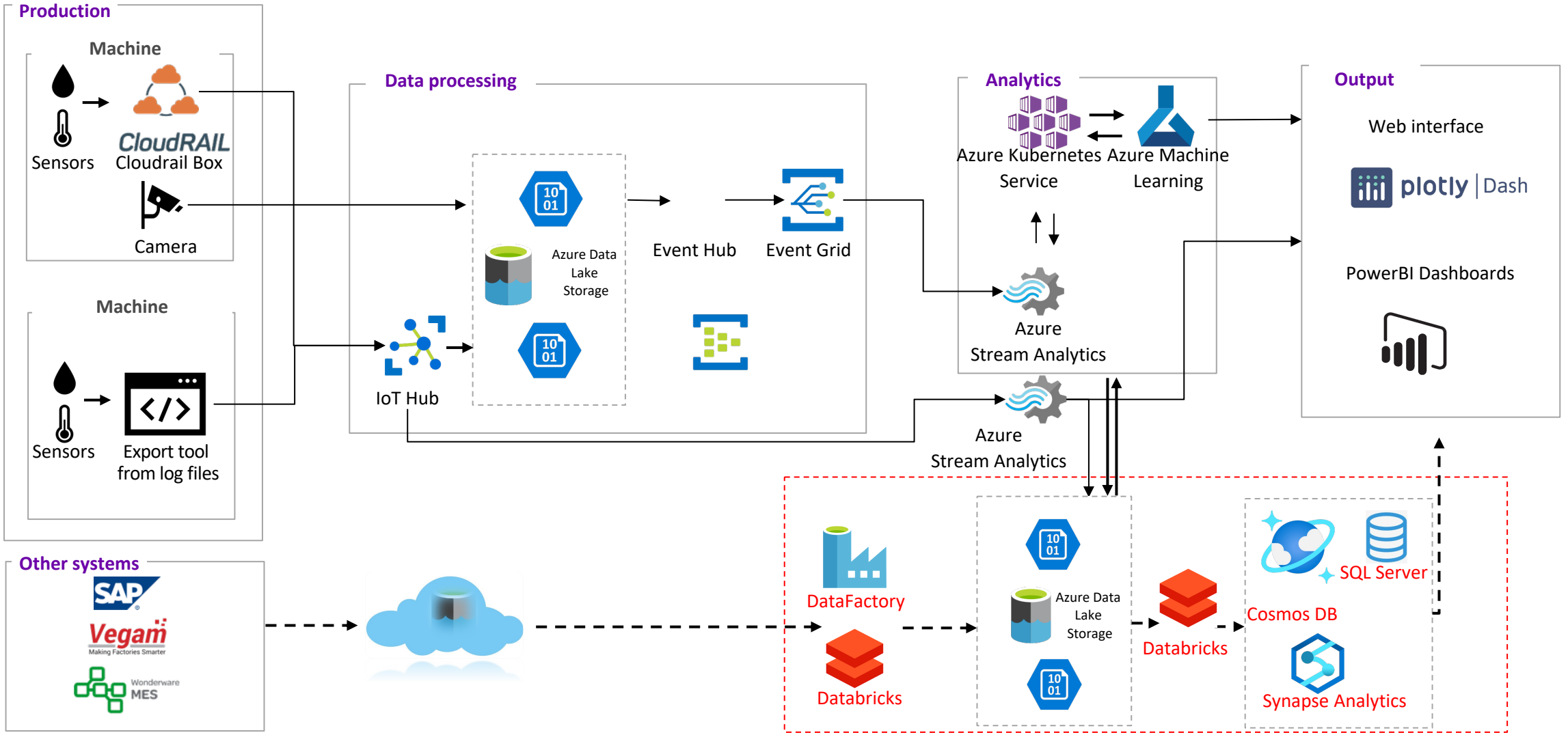
English translation

Example architecture (OSS)



BearingPoint®

End to end Azure architecture for IoT ML Ops



BearingPoint®