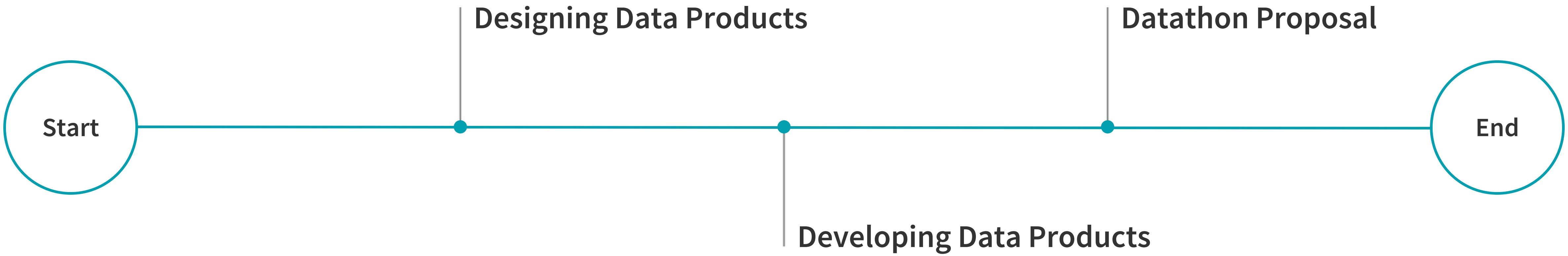
Building Data Products

ANTONIO MOLNER DOMENECH

Seminar Outline



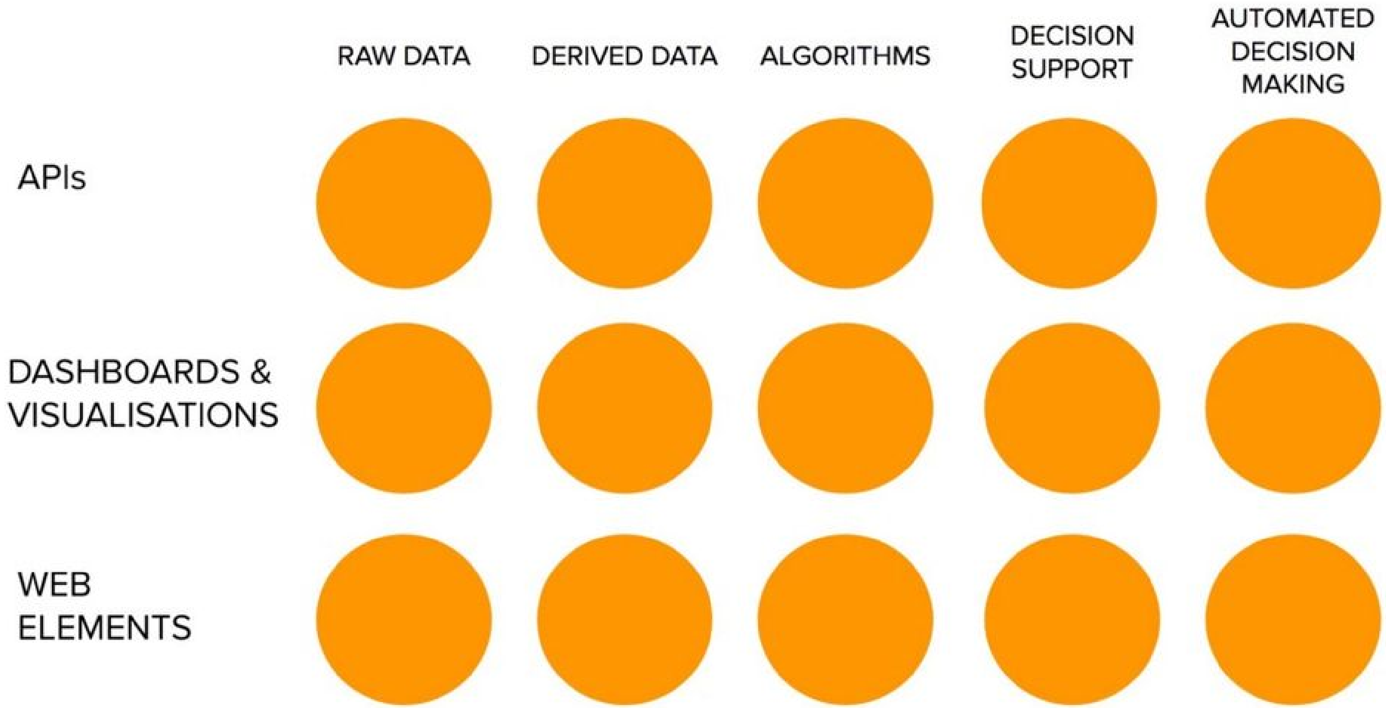
Designing Data Products

What is a Data Product?

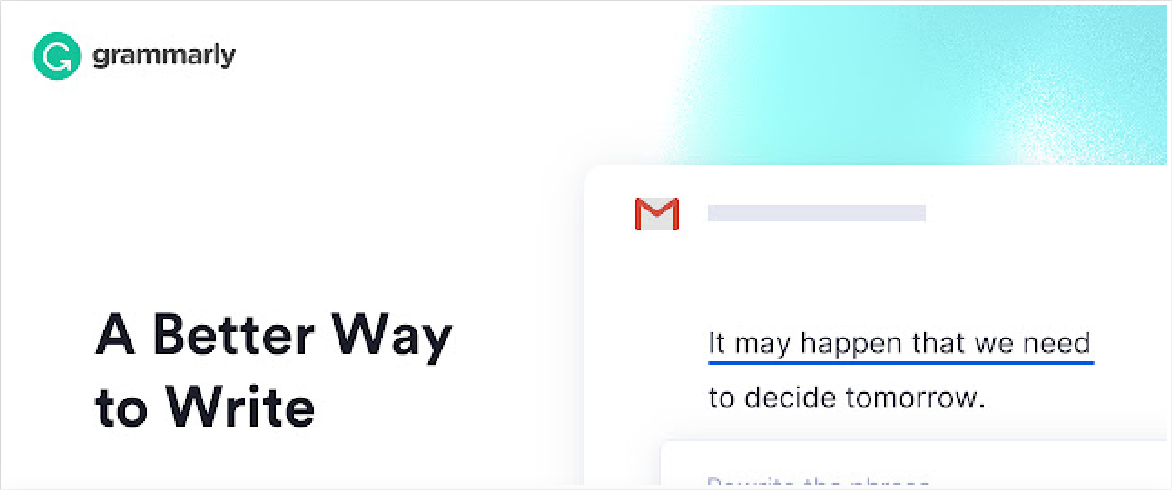
“A data product is a product that facilitates an end goal through the use of data”

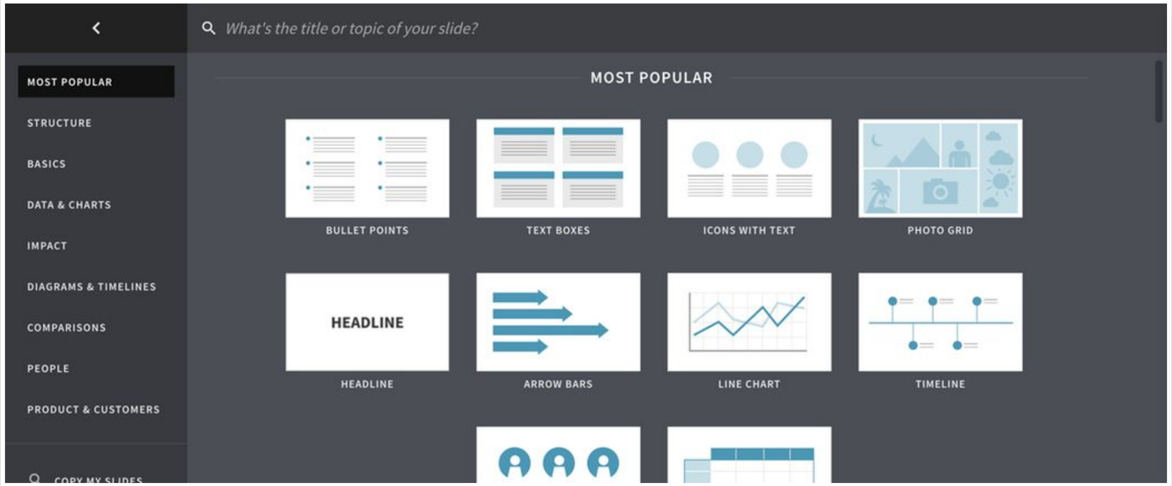
DJ Patil, Former U.S Chief Data Scientist - Data Jujitsu

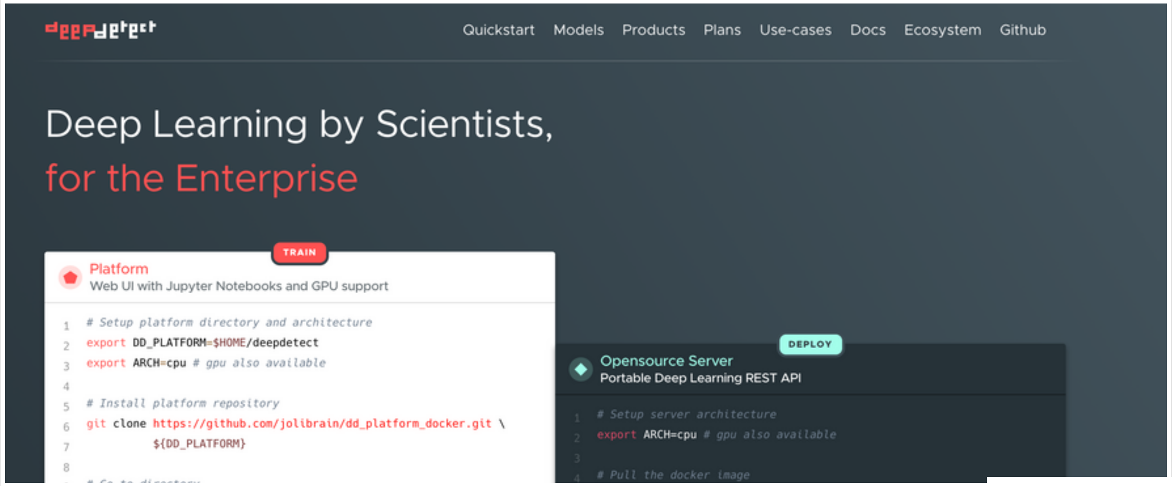
Types of Data Products



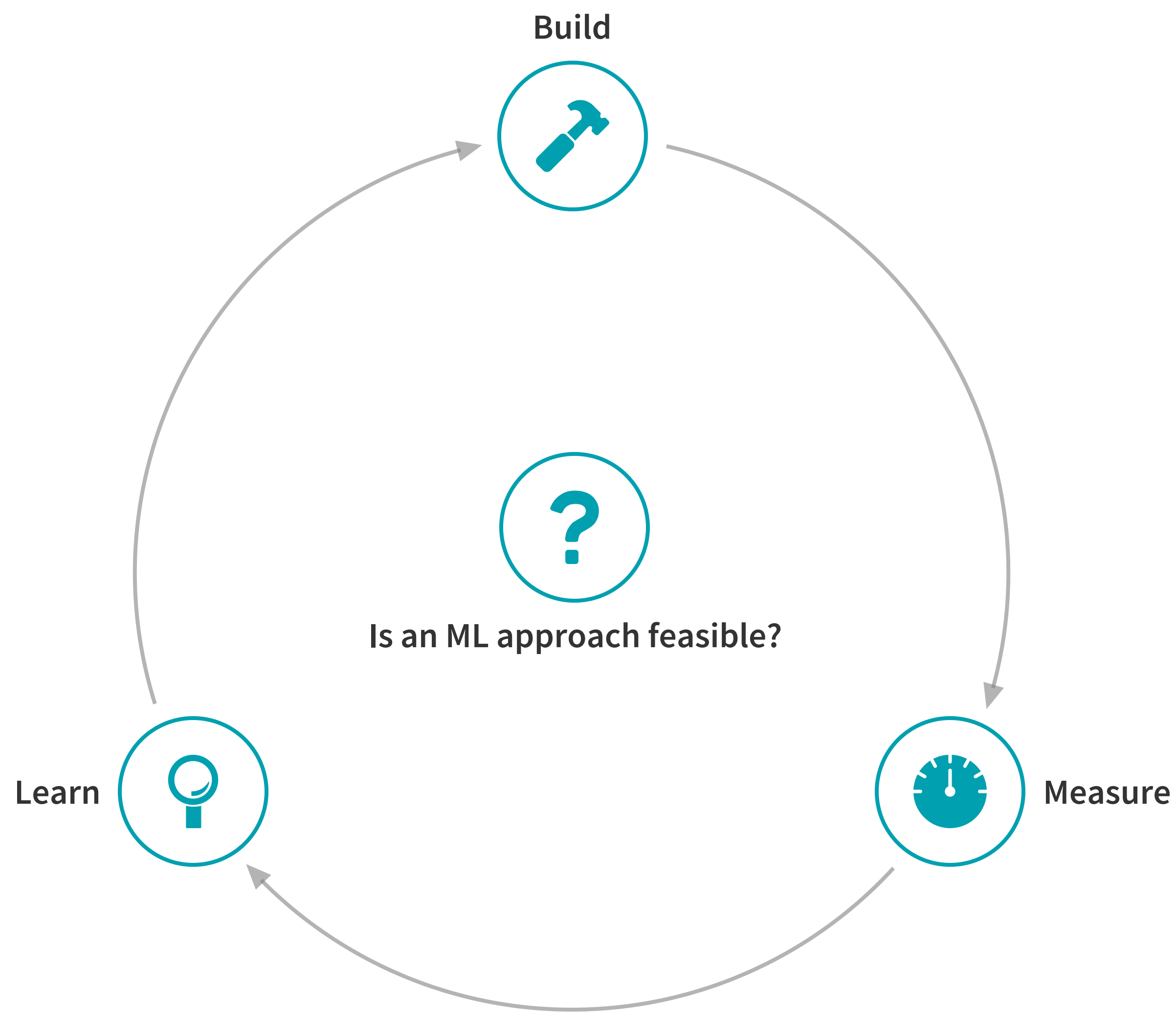
Examples of Data Products







MVP Cycle



Minimum Viable Data Product (MVDP)

Minimum Viable Implementation

Follow a customer-first approach.

Should support monitoring, feedback, and governance.

Minimum Viable Model

Forget about fancy papers and SOTA models.

Implement and **baseline** and iterate.

Minimum Viable Platform

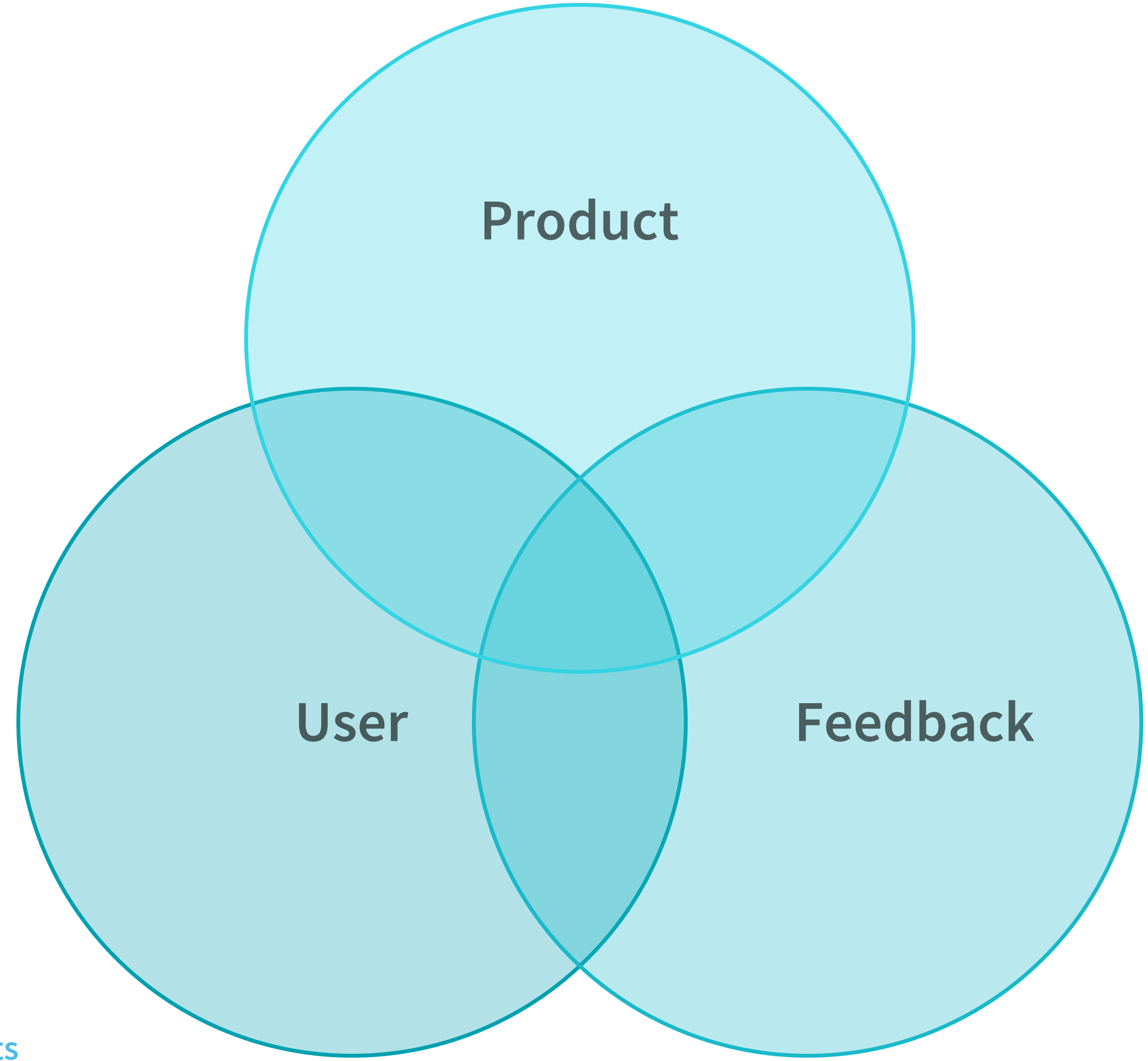
Amazon S3

amazon EMR

Paperspace

Amazon SageMaker

Principles for Data Product Design



Principle for building great data products

Product Intrinsic Principles

- Build Trust with Transparency
- Invoke Discovery and Delight
- Visualize the Complex
- Blend in

Feedback-aware infrastructure

- Collect Data Passively
- Constantly Validate with Data

https://www.frogdesign.com/designmind/nine-principles-for-designing-great-data-products

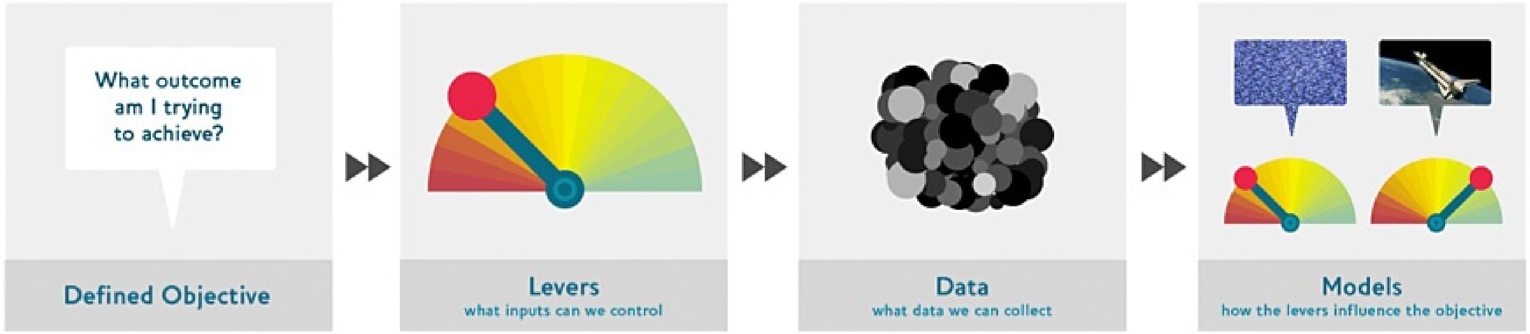
User-focused Principles

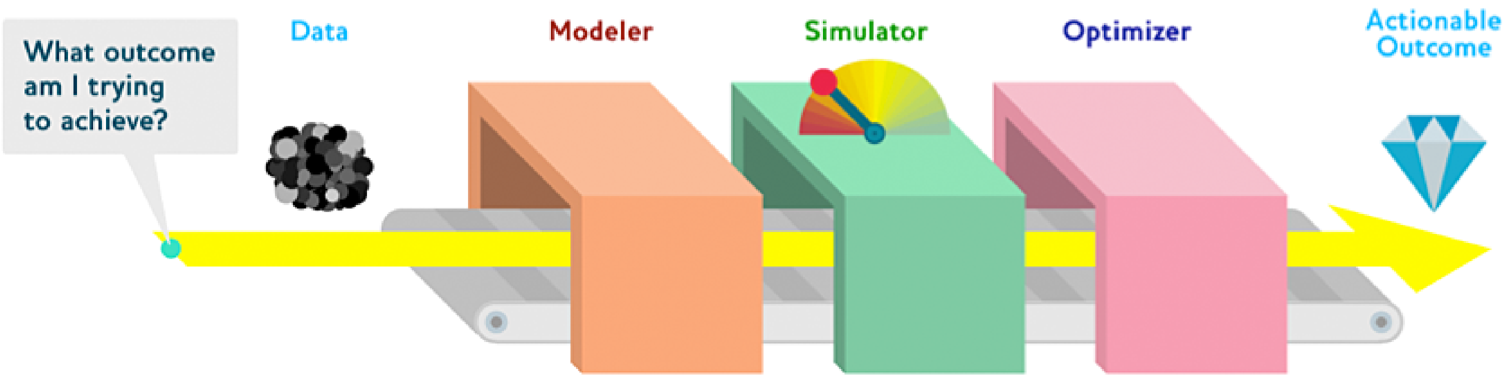
- Meet Unexpressed Needs
- Don’t Exhaust the User
- Give Users Control

https://www.frogdesign.com/designmind/nine-principles-for-designing-great-data-products

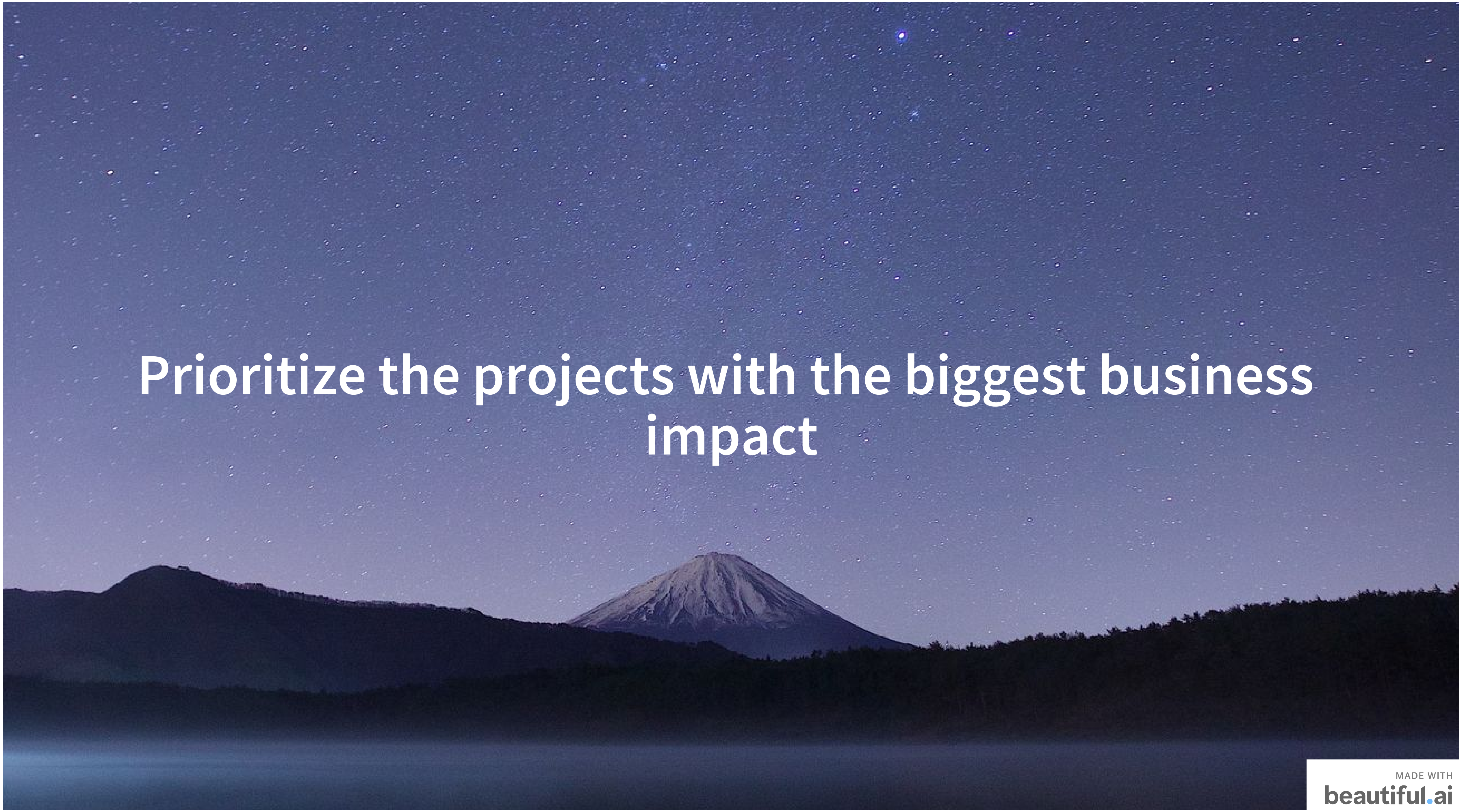
Drivetrain Approach

Designing Great Data Products



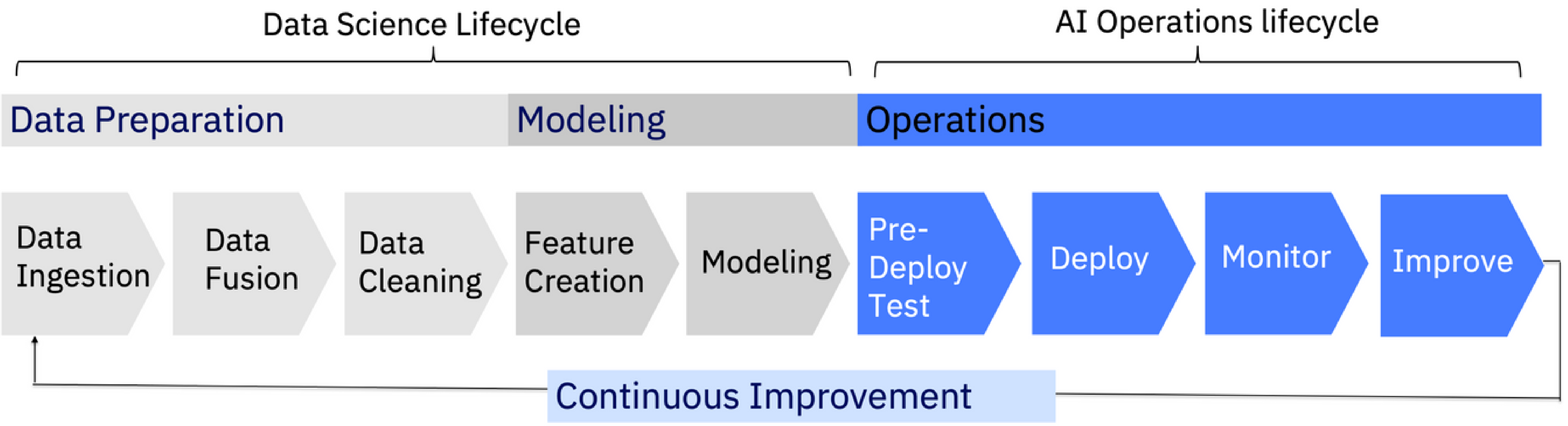


Prioritize the projects with the biggest business impact



Developing Data Products

Data Science Process



Data Preparation

Master Data Management and Governance — Stewardship, Data quality, cleansing, etc.

Data Engineering — Pipeline design, ETL, etc.

Kubeflow

Apache Airflow

Data Acquisition

Data science pyramid

Modeling

TensorFlow

PYTORCH

fast.ai

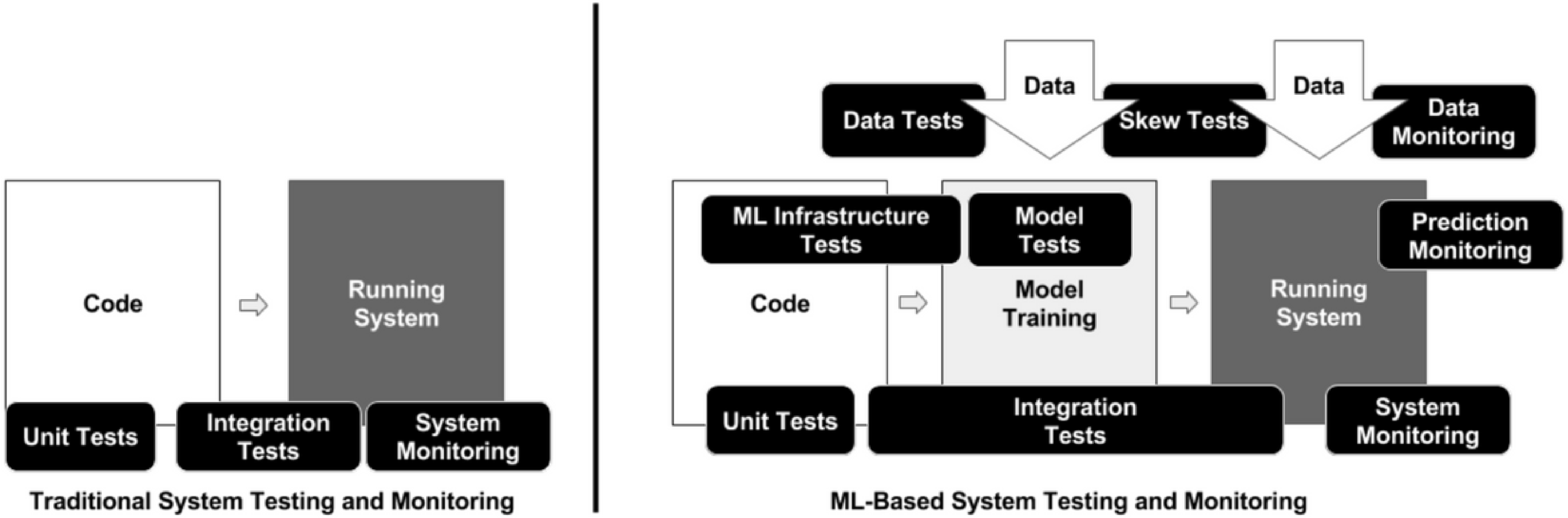
PROPHET

Keras

scikit learn

Testing

The ML Test Score:A Rubric for ML Production Readiness and Technical Debt Reduction



Deployment

Flask

FastAPI

cortex

Amazon SageMaker

SELDON

Kubeflow

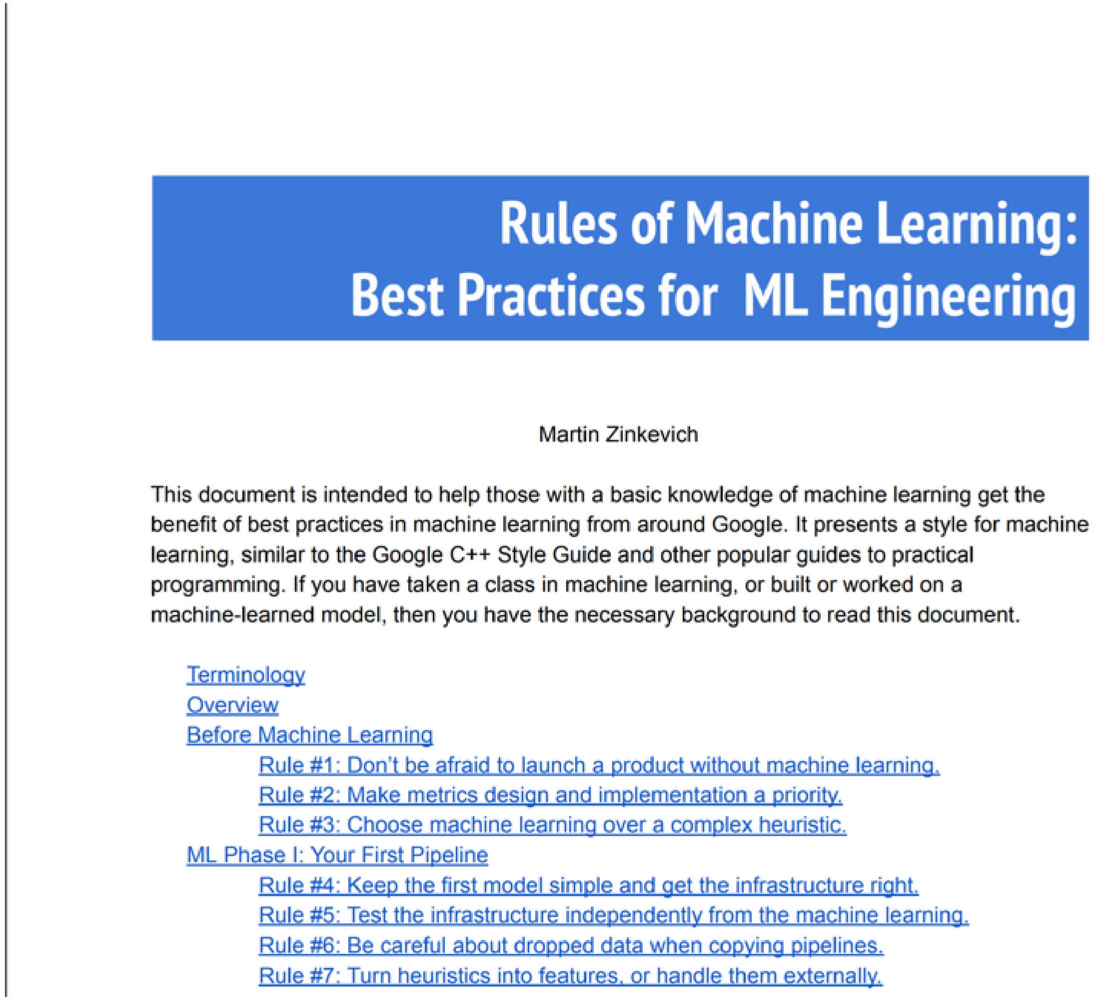
Monitoring

hydrosphere.io

cortex

Amazon SageMaker

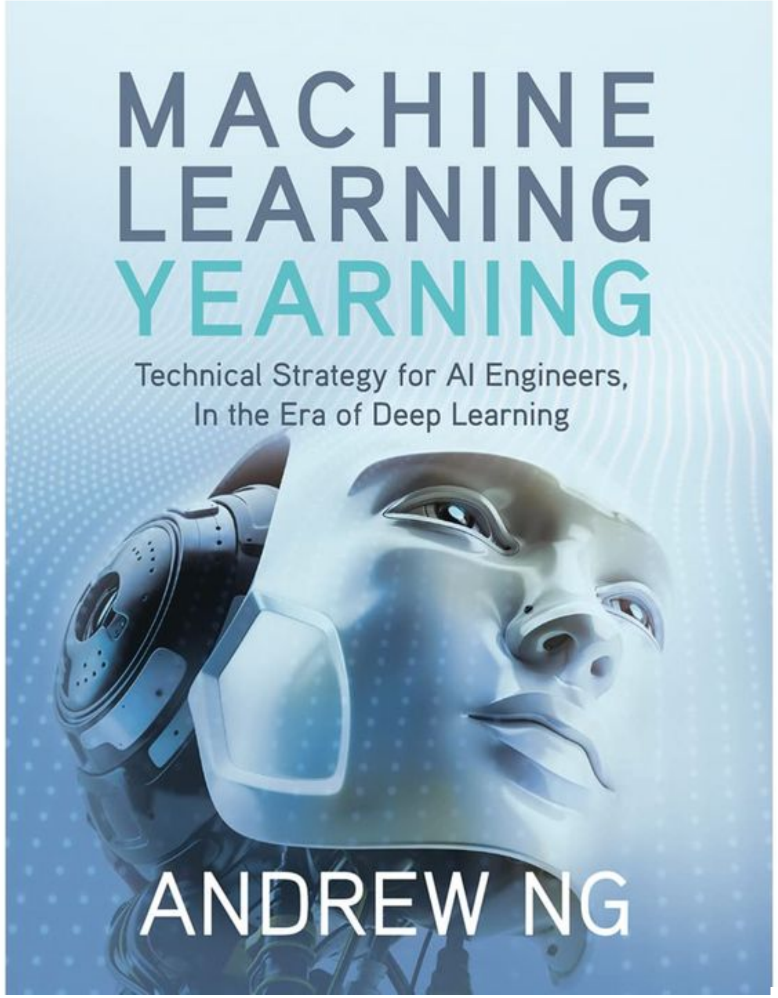
Best Practices



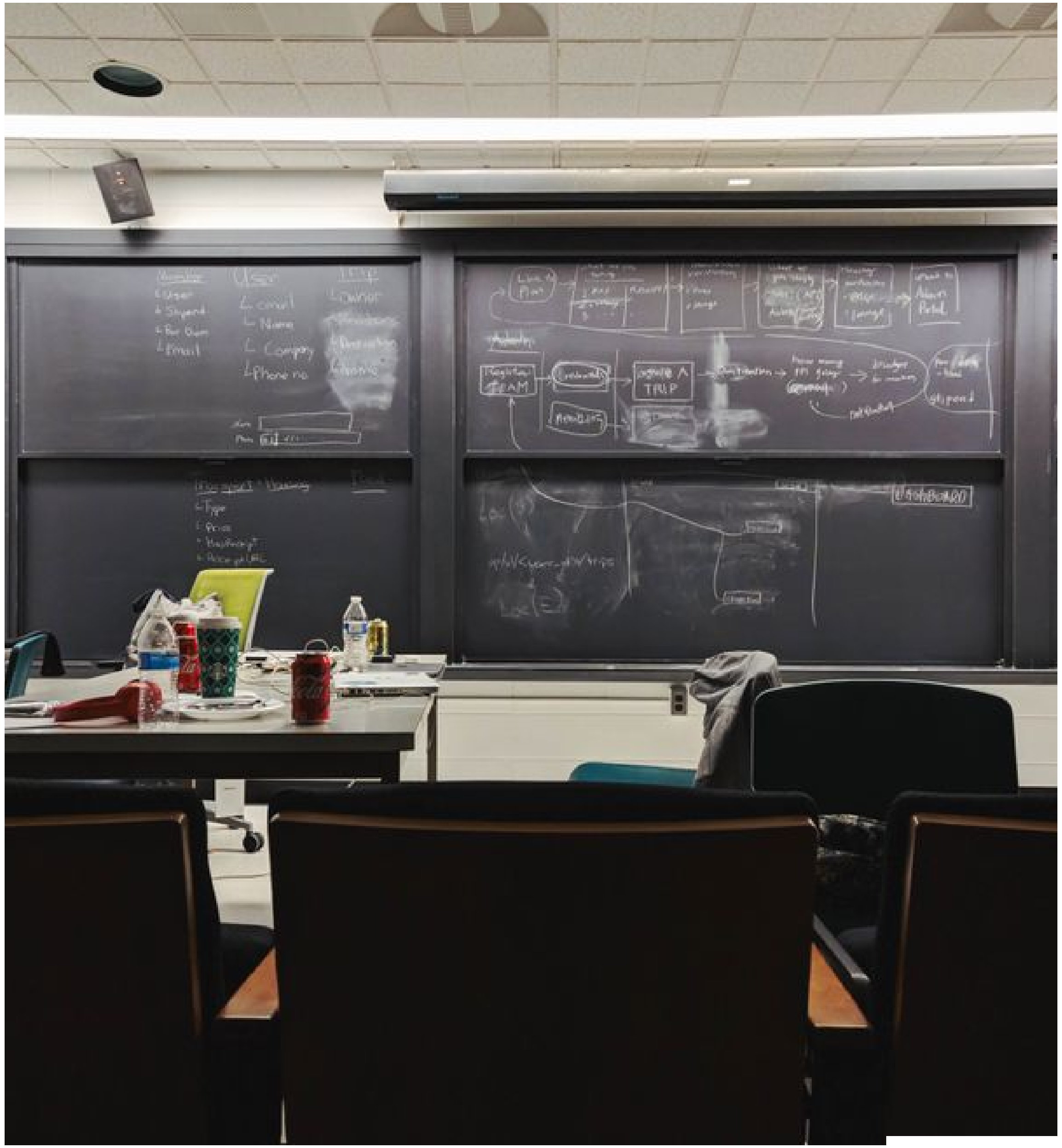
Rules of Machine Learning: Best Practices for ML Engineering

Martin Zinkevich

This document is intended to help those with a basic knowledge of machine learning get the benefit of best practices in machine learning from around Google. It presents a style for machine learning, similar to the Google C++ Style Guide and other popular guides to practical programming. If you have taken a class in machine learning, or built or worked on a machine-learned model, then you have the necessary background to read this document.



Datathon



Objetives

- Learn to design and implement a MVDP
- Understand the problems that DS can solve
- Understand the implications and problems that arises when applying DS

Proposed Projects

MNIST Online Tool

Example

Text Topic Classification

- Example 1
- Example 2

Movie Recommendation System

Example

Bitcoin Trading Bot

- Example 1
- Example 2
- Example 3

Hotdog - No hotdog App

- Example 1
- Example 2
- Example 3

MNIST Online Tool

- Data extraction: Download data from Keras datasets or similar
- Data preparation: Data augmentation
- Modelling: Recommended CNN with pretrained weights
- Evaluation: Accuracy metric
- Validation: Use a Logistic regression or a former SOTA model
- Model deployment: As microservice, as part of the backend, lambda, etc.

Movie Recommendation System

- Data extraction: MovieLens dataset. Netflix activity.
- Data preparation: Algorithm-dependent
- Modelling: Collaborative filtering algorithms
  - https://github.com/NicolasHug/Surprise
  - https://github.com/NVIDIA/DeepRecommender
  - https://github.com/benfred/implicit
  - https://github.com/zhenghaoz/gors
- Evaluation: Regression or classification metrics
- Validation: None
- Model deployment: As microservice, as part of the backend, lambda, etc.

Bitcoin Trading Bot

- Data extraction: Historical data of bitcoin prices
- Data preparation: Algorithm-dependent

  - Time series features (lags, cumsums, etc) for traditional algorithms

  - None for LSTM or Prophet
- Modelling: RNN, LSTM, traditional regression methods, Prophet (Recommended)
- Evaluation: Backtesting
- Validation: Against traditional methods like mean reversion, moving average, etc.
- Model deployment: Quantopian platform

Hotdog - No hotdog App

- Data extraction: Images of hotdog and not hotdogs
- Data preparation: Image preprocessing + Data augmentation
- Modelling: Pretrained CNN
- Evaluation: Accuracy
- Model deployment: As microservice, Tensorflow lite, CoreML, etc.

Resources

- Designing great data products
- The Fundamentals of Building Better Data Products
- Data Jujitsu - Ebook
- Rules of Machine Learning: Best Practices for ML Engineering
- Best MLOps Tools
- Monitoring Machine Learning Models in Production
- What is hardcore Data Science in practice?
- Applied Artificial Intelligence: An Introduction For Business Leaders - Book
- Machine Learning Yearning - Book
- The AI Organization - Book
- Agile Data Science 2.0 - Book
- Introducing MLOps - Book


Thanks!