

Blazing a trail in the business world using machine learning

Karl Fogelmark

Inmeta Consulting AS

2020-08-17

Table of contents

PhD Research Project (2011–2016)

- Introduction

- Gene regulatory networks

Unemployment (2016–2017)

- Trygghetsstiftelsen

- Advice for CV/Personal letter

Work at Inmeta (2017–)

- About Inmeta

- Work process

- Example projects

- Future of ML at Inmeta

Appendix

Introduction

PhD student at department of Theoretical Physics & Astronomy
2011—2016 *

- ▶ Did modelling, simulation, optimization, of genetic data
- ▶ Did not do ML[†] (although took courses)

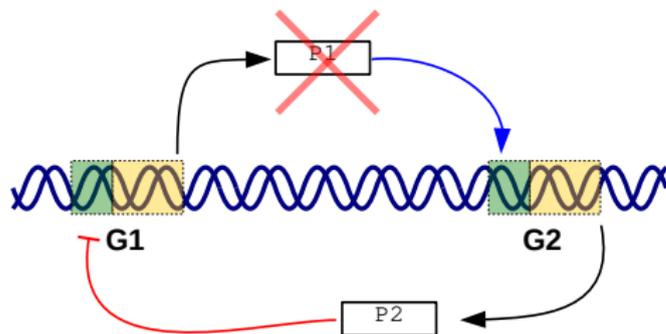
Specifically:

- ▶ Wrote papers, with supervisor Carl Troein, on gene regulatory networks
- ▶ ...and one on error estimation of function fitting when data is correlated with Tobias Ambjörnsson

*Part of the CBBP group (Computational Biology and Biological Physics)

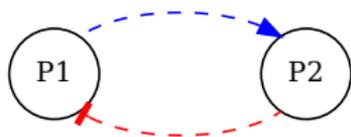
†Machine Learning

Gene regulatory networks — toy example



- ▶ The DNA segment has 2 genes: G1 and G2
- ▶ Gene G1 produces protein P1
- ▶ P1 binds to **promotor site** of G2, **activating** P2 production
- ▶ P2 **represses** gene G1 which stops producing P1

Network described by ODE



Interaction described by coupled ODEs*, with parameters a_i, b_i, c_i

- ▶ Production of protein P_2 is **activated** by protein P_1

$$\frac{dP_2}{dt} = a_2 + b_2 P_1 - c_2 P_2$$

- ▶ Production of protein P_1 is **repressed** by protein P_2

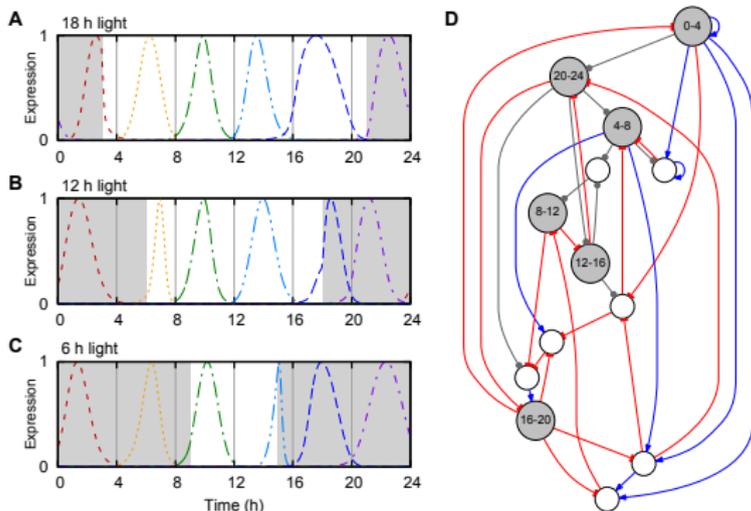
$$\frac{dP_1}{dt} = \frac{a_1}{1 + P_2} - c_1 P_1$$

- ▶ P self-decay rate \sim minutes – hours

*Ordinary differential equation

Result

Interactions form complex networks that **perform tasks**, e.g. maximize localization of **protein production** (expression) to specific time window*



* Above from simulation, where different parameter sets were used during light/dark period, respectively

Publications

- ▶ Karl Fogelmark, Carsten Peterson and Carl Troein; "Selection shapes transcriptional logic and regulatory specialization in genetic networks" *PLoS ONE* **11**, e0150340 (2016)
- ▶ Karl Fogelmark and Carl Troein; "Rethinking transcriptional activation in the Arabidopsis circadian clock" *PLoS Computational Biology*, **10**, e1003705 (2014)
- ▶ Karl Fogelmark, Michael A. Lomholt, Anders Irbäck and Tobias Ambjörnsson; "Fitting a function to time-dependent ensemble averaged data" *Scientific Reports* **8**, 6984 (2018), (arXiv:1805.03057)
- ▶ Simon Pigeon, Karl Fogelmark, Bo Söderberg, Gautam Mukhopadhyay, Tobias Ambjörnsson "Tracer particle diffusion in a system with hardcore interacting particles" *Journal of Statistical Mechanics: Theory and Experiment* (2017), (arXiv:1712.03996)
- ▶ Lloyd P. Sanders, Michael A. Lomholt, Ludvig Lizana, Karl Fogelmark, Ralf Metzler and Tobias Ambjörnsson; Severe slowing-down and universality of the dynamics in disordered interacting many-body systems: ageing and ultraslow diffusion *New Journal of Physics* **16**, 113050 (2014) (arXiv:1311.3790)
- ▶ R. Metzler, L.P. Sanders, M.A. Lomholt, L. Lizana, K. Fogelmark and T. Ambjörnsson; Ageing single file motion *European Physical Journal* **223**, 3287-3293 (2014)

PhD Research Project (2011–2016)

Introduction

Gene regulatory networks

Unemployment (2016–2017)

Trygghetsstiftelsen

Advice for CV/Personal letter

Work at Inmeta (2017–)

About Inmeta

Work process

Example projects

Future of ML at Inmeta

Appendix

Unemployment

Support from Trygghetsstiftelsen:

- ▶ Helped finance my interview trips to Oslo
- ▶ Helped finance my move to Oslo*
- ▶ Prepare for job interview / salary negotiation
- ▶ Read CV and personal letter

(Trygghetsstiftelsen is a resource to help assist finding new job of all previous government employees (e.g. PhD-students))

Advice 1: Personal letter — don't assume competency

Don't use terminology not found in the advert

- ▶ Reader **lacks domain knowledge**, is HR / external consultant
- ▶ Reader will just compare your letter/cv to words in job advert
- ▶ Don't use synonyms / alternative terminology to that in advert
- ▶ **Use terminology** found in job advert, exclusively — even if it feels restrictive to “just copy/paste” terminology

You are now the expert, not your boss! Use small words.

Advice 2: CV — Add a summary

- ▶ Add a **4-5 sentence summary** to the top of your CV, with all the phrases from the advert that fit your description
- ▶ Also **rephrase** each entry on your CV to match job advert, where possible

Karl Fogelmark

Curriculum vitae

Luthvångens 12B
SE-221 37 Lund
+46 40 12 14 00
k.fogelmark@gmail.com

Summary

As an ambitious, driven, PhD in Theoretical physics with deep knowledge of data analysis, machine learning, computational modeling and statistics, able to extract useful and meaning from data, C++ programming and my currently has shown me to learn many languages. I have worked in an international research environment for more than five years, with theoretical models and simulations.

Work experience

- 2011-2016 **PhD project**, Theoretical physics, Lund University
- Through this project and several seminars, I developed mathematical models of biological systems, as part of my PhD-thesis. In addition I also:
 - Took responsibility of 1-2 years of additional courses in mathematics, physics, machine learning, high performance computing, and programming.
 - Taught several classes, e.g. Computational Physics and Probability Statistics.
 - As administrator I was responsible for maintaining the department web page.
 - Organized conferences (e.g. "Water dual talk" and Swedish Bioinformatics Workshop 2012).
 - Participated in international conferences.
- 2010-2011 **Research assistant**, Theoretical physics, Lund University
- Investigated random (disordered) processes through modeling and computer simulation.
 - Supervised students computer exercises in Computational Physics.
 - Taught first year physics students in problem solving.

Education

- 2011-2016 **PhD**, Theoretical physics, Lund University
- Degree from Department of Mathematics and Theoretical Physics. Research project focused on developing computational models in systems biology and computational physics.
- 2009-2010 **Bachelor**, Theoretical physics, Lund University
- Degree from Department of Mathematics and Theoretical Physics.

PhD Research Project (2011–2016)

Introduction

Gene regulatory networks

Unemployment (2016–2017)

Trygghetsstiftelsen

Advice for CV/Personal letter

Work at Inmeta (2017–)

About Inmeta

Work process

Example projects

Future of ML at Inmeta

Appendix

Inmeta

- ▶ Inmeta is an Oslo based IT-consultancy with 200 employees
- ▶ AI/ML (Machine Learning) group **started in 2015**
- ▶ ML-Team (currently) consists of 17 data scientists



The ML-Team: High energy!

Most of the data scientists have **PhD-background**, typically physics



In 2019, **Inmeta sponsored** the ML-team's 3 day weekend **trip to CERN**

Similarity to PhD-research

Share and gain new knowledge

- ▶ We **present recent papers** & techniques every Monday, e.g.
 - ▶ *An introduction to separable convolutions*
 - ▶ *Probabilistic programming with PyMC3*
 - ▶ *Git Hooks to the People*
 - ▶ *Neural collaborative filtering*
 - ▶ *Lottery ticket hypothesis*
 - ▶ *Recap of ODSC conference*
 - ▶ ...
- ▶ We try to attend **conferences** ~ 1 week / year (pre-corona)
 - ▶ NIPS (NeurIPS) 2019, Vancouver
 - ▶ NIPS (NeurIPS) 2018, Montreal
 - ▶ ODSC 2019, London
 - ▶ ICML 2018, Stockholm
 - ▶ ICLR 2017, Toloun
 - ▶ Simula 2017, Oslo
 - ▶ ...
- ▶ We **write articles**, e.g. on Medium
 - ▶ *How to improve the performance of a machine learning model with post processing employing **Levenshtein distance***
 - ▶ ***Word embeddings**, what are they really?*

Projects made

Past 5 years: +100 applied AI/ML projects delivered*

Energy: Predictive Maintenance:

Failure diagnostics from +200 sensors to minimize field worker involvement



Healthcare: Image diagnostics for identification of **malignant tumors** in the colon



Inmeta has found a sweet spot for taking applied machine learning into the business world

*See appendix for many examples

Inmeta is a prominent ML consultancy in Europe



- ▶ 2019 Global AI/ML* Partner of the Year
No. 1 among 3,000 competing partners worldwide
- ▶ International AI trainer of Microsoft partners
Inmeta ML-group held workshops in: Paris, London, Munich, Copenhagen, Amsterdam, Lisbon, Singapore, Sidney, New Zealand



- ▶ First European ML Competency Partner of AWS
Currently one of two partners

*ML = Machine learning

Work process: From problem definition to deployment

1. Opportunity assessment, concept proofing (~ 3 weeks)
 - ▶ **Articulate problem**, business value, and desired outcome
 - ▶ Conceptualize solution
 - ▶ Determine data source & quality
 - ▶ **Iterate & validate** solution concept
2. Exploratory data analysis (6-8 weeks)
 - ▶ **Collect raw data**
 - ▶ Data visualization and interpretation
 - ▶ Identify features and label sources, feature engineering
 - ▶ Select sampling strategy
 - ▶ Data- **cleaning**, **processing** & assessment (statistically sane)
3. Model construction & training (3-4 weeks)
 - ▶ Choose methodology and algorithms
 - ▶ **Define metric** to align business objectives with model performance
 - ▶ Train, validate, and evaluate model
4. Model optimization (3-4 weeks)
 - ▶ Optimize model performance and size iteratively, based on evaluation metrics
5. Deployment & operation (TBD)

The biggest dairy product cooperative in Norway (15k farmers)

- ▶ **Predict milk production:** How much milk will farmers deliver to Tine, each week, the coming 2 years. Basis for **stock planning**
 - ▶ **Model lactation curve** of cow after calving
(varies from farm to farm)
 - ▶ Predict number of cows on each farm, 2 years into future
 - ▶ Model health/sickness of cows on each farm
- ▶ \$10s million in savings for the industry as a whole



(Tech used, at different stages in the project: RNN, Catboost, XGBoost, CNN)

*Worlds largest classification society, 13k ships & offshore units / y,
12.5k employees in 100 countries*

- ▶ **Automatic classification and routing** of ~70k/y advanced technical e-mail (to 400 experts, 650 different categories) saving USD \$200k annually trained on 250k historical records
- ▶ **Smart Survey booking:** Scheduling vessel inspection globally for certificate renewal is highly complex.
\$10 mn in admin cost savings (annual ROI 875%)



+30 additional models in production & pipeline over past 4y by 2-4 by data scientists

- ▶ Thousands of technical docs submitted to engineers **each day**.
 - ▶ Re-use previous work: previous similar design rejected/approved?
 - ▶ Similarity **search of 15M technical drawings** < 1 second.
 - ▶ New documents added every hour
- ▶ Solution: **TechDoc search**
 - ▶ Version 1: **Autoencoder**, compress 65k pixels → 512
 - ▶ Version 2: **SimCLR contrastive learning*** (rotation invariant)

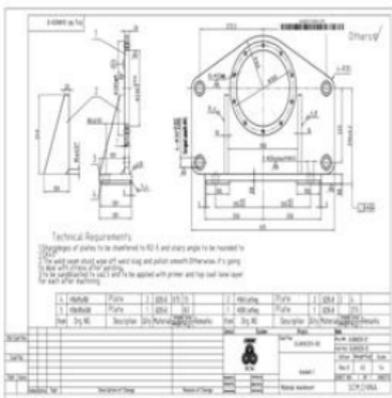
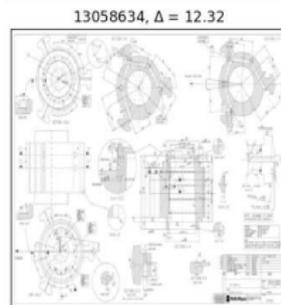
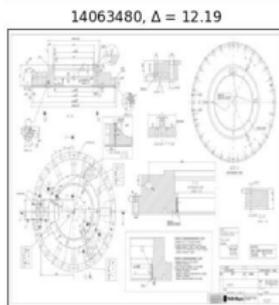
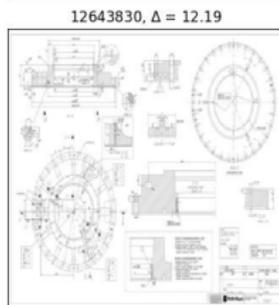
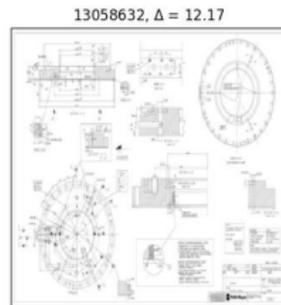
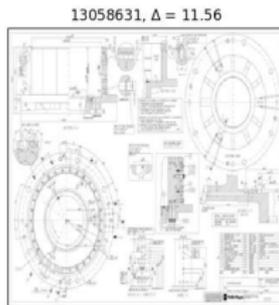
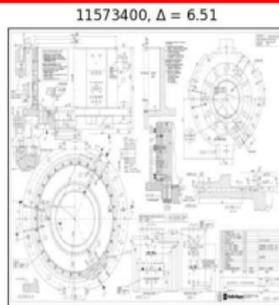


Figure: Technical drawing (low resolution for confidentiality)



Top left drawing ($\Delta = 0.0$) is input, rest are most similar in embedding space of v1. auto encoder

My projects

(Most customers are confidential)

- ▶ **Recommender system** for large Scandinavian retailer
- ▶ **Homomorphic encryption** collaboration with **Microsoft Research in Redmond**
- ▶ **Cyber-grooming detection (NLP)** from chat messages from popular (400M users) online game for children
(Article in [https://www.aftenposten.no/...](https://www.aftenposten.no/))
- ▶ Modelling of (highly detailed) **football statistics** for start-up to improve football team decisions
- ▶ Statistical analysis of debt holders at large **Norwegian debt collector** provider
- ▶ Helsedirektoratet (Ministry of Health) **cost analysis**
- ▶ **Customer segmentation** and prediction models at Telia Norge

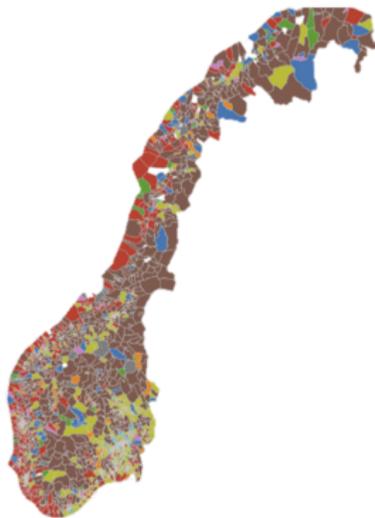
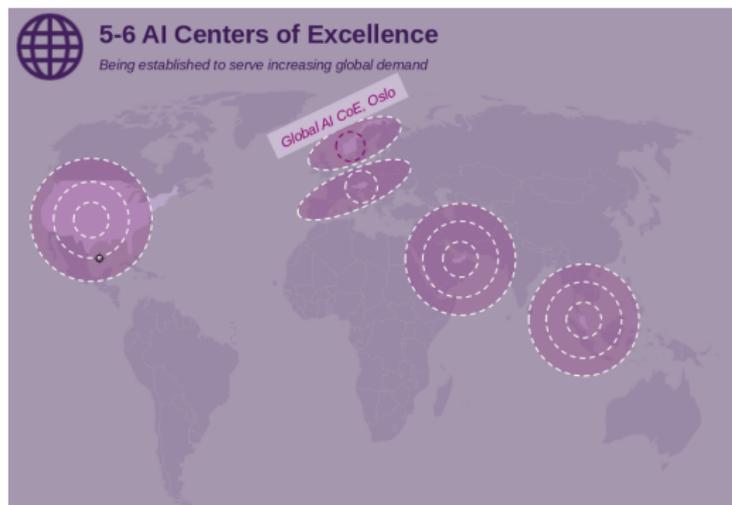


Figure: Analysis of **retail buying patterns** in Norway (from embedding space of recommender model).

Future: expansion

- ▶ Inmeta is owned by software company **Crayon AS**, 1300 employed world wide, 35 countries
- ▶ Now **building global AI consultancy practice**, (already in Vienna & US)



Please send me your CV if interested to join our team:

karl.fogelmark@inmeta.no

Provides high resolution product images for commercial adverts

- ▶ **Automatic image background removal**
- ▶ Developed to eliminate third party spend of \$2-\$4 per image on ~50k images annually ~\$15k
- ▶ Current solution **reduce labour cost** by 50% based current output quality



Taxonomy of AI/ML Team

Data Scientists — Core of the ML-team

- ▶ Develop models, statistics, numerical analysis
- ▶ Background: mostly physics PhD, some with master degree
- ▶ Typically we use Python, with libraries e.g. numpy, scipy, keras, pytorch, tensorflow, nltk, gensym, etc.

ML/Data Engineers — Support Data Scientists

- ▶ Responsible for Data accessibility and preparation (Data bases, cloud)
- ▶ Data architecture, data management

AI Advisers — "Suits"

- ▶ Deal with the customer, manage sales
- ▶ Identify AI/ML opportunities at customer

Projects overview

Industry	Computer vision	NLP	Time series / Structured data	Example
Maritime	x	x	x	Prediction of time estimate for maritime vessel classification
Health	x	x	x	Image diagnostics for identification of colon cancer
Retail	x		x	Personalized recommendation of skin products based on face scan
Farming			x	Bottom-up prediction of national milk production
Comms & media	x	x		Segmentation and churn prediction across customer portfolio
Public gov.			x	Risk based survey prediction for non-compliance across population
Energy	x	x	x	Prediction of technical failures in windmill farms
Safety	x		x	Predictive maintenance for wastewater purification
Insurance			x	Disability insurance pricing model
Finance		x	x	Customer profiling; product recommender savings products
Process Industry			x	Computer vision for failure detection in solar cell production
Transportation			x	Optimization of load utilization and route planning

Example projects (1/3)



Estimation of ventricular wall thickness based on location of heart valve from ultrasound images



Identification of colon cancer using computer vision in colonoscopy



Classification of vessel sensor name and tag into correct vessel information structure (SensorTag)



Adjustment of inspection scope and frequency based on risk assessments across a multitude of information sources (IHS)



Development of "ML Factory" to operationalize and manage ML models in production



Prediction of milk yield per individual cow to improve farming operations (feeding, disease etc.)



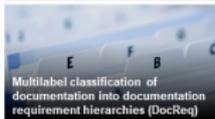
Evaluation of potential to do calculations on encrypted datasets through homomorphic encryption



Generation of anonymous test data based on data input containing sensitive personal data (POC)



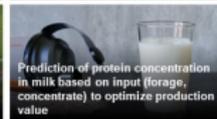
Classification of inspection findings e.g. non-conformities to requirements and regulations into correct vessel information structure



Multilabel classification of documentation into documentation requirement hierarchies (DocReq)



Prediction of short-term (1-21 days) milk production volumes at individual farm level



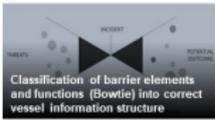
Prediction of protein concentration in milk based on input (forage, concentrate) to optimize production value



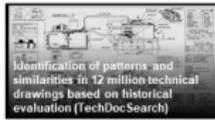
Analysis of activity based cost allocation for financial modeling within the public health system



Automation of e-mail routing to the right case handler (DATE - Direct Access to Technical Experts)



Classification of barrier elements and functions (Bowtie) into correct vessel information structure



Identification of patterns and similarities in 12 million technical drawings based on historical evaluation (TechDocSearch)



Prediction of long-term (24 months) milk production volumes at national level



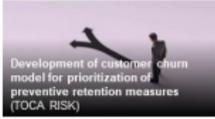
Optimization of slaughter order for cows considering meat quality in relation to weight



Categorization of different prostate cancer types from analysis of standardized treatment process



Development of several models for similarity-based search system used by case handlers (DATE - "Similar Case Identification")



Development of customer churn model for prioritization of preventive retention measures (TOCA RISK)



Prediction of time estimate for maritime vessel classification (Smart Survey Booking)



Prediction of fat concentration in milk as result of input (forage, concentrate) to optimize production value



Creation of production environment for ML solutions



Detection of fall accidents to notify health personnel when a patient needs help



Estimation of hour requirement to fulfill 5-year vessel classification contracts (PSACP - Periodic Service Agreement Cost Prediction)



Estimation of noise levels in vessel compartments (crew cabins etc.) based on historic noise measurements on similar vessels



Exploration of VERA, a Maritime Class chatbot-similar concept



Prediction of growth curves for cows to optimize farming operations e.g. feed adjustment and right timing for slaughter



Development of data lake

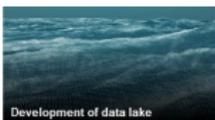
Example projects (2/3)



Identification of root causes for transmission loss of data in AMS meters (smart meters)



Detection of unsharp edges in product images used in retail



Development of data lake



Prediction of customer churn based on product profile, behavior, demographics, triggers etc.



Assignment of value potential from AI across product portfolio (hereunder autotagging metadata and optimizing search)



Prediction of environmental non-compliance for probability-based-precision-targeting of auditing efforts



Prediction of wind turbine failures from 200+ sensors and alarms to minimize field worker involvement



Detection of gloss and blur in product images used in retail



Generation of competitor insights e.g. variations in price based on geography.



Anonymization of data



Management of data science project including deployment of ML models in "stormskin"



Prediction of price changes in currency fluctuations



Automation of background removal in product images used in retail



Development of recommender solution for upsales based on basket items (3rd party)



Estimation of optimal daily day prices to optimize financial performance



Clustering of customer data for more granular segmentation



Recognition of vehicle number plates from images and data augmentation



Identification of patterns in large data set for development of new business models



Establishment of Joint Venture to commercialize image processing model



ML Pilot - recommender of goods



Optimization of grocery baskets based on patterns across offering and purchasing behavior



Analysis of voice of customer based on sentiment related to e.g. campaigns, events etc.



Prediction of wastewater purity based on mud slip rate on components used to dig out mud from treatment plants



Development of recommender solution to improve online conversion rate



Automation of dust/scratch removal in product images used in retail



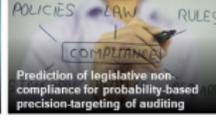
Recognition of patterns driving loyalty amongst customers



Identification of skin type and condition (using computer vision) to recommend individualized skin care products



Prediction of click backs in SMS-campaigns based on recipient and timing



Prediction of legislative non-compliance for probability-based-precision-targeting of auditing



Improvement of newsletter relevance by rank ordering products based on individual customer preferences (segments)

Example projects (3/3)



Evaluation of impact from drilling mud emissions on coral health



Structuring of large information data base and development of advanced search engine e.g. knowledge graph



Production setting of ML model (developed by third party)



Establishment of multidisciplinary, world leading imaging research consortium for development of the next generation CT imaging



Automation of customer service interactions by implementing customized chatbot solutions with integrated AI/ML applications



Ranking of urban green space across 40 cities globally using satellite images and machine learning



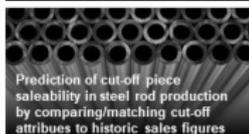
Prediction of attainable oil and gas reserves in existing wells



Development of advisory platform to connect core systems and channels (mostly architecture)



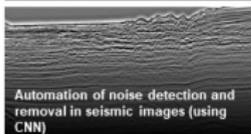
Optimization of production process for high quality mono crystalline silicon used for Photovoltaic solar panels



Prediction of cut-off piece saleability in steel rod production by comparing/matching cut-off attributes to historic sales figures



Development of data lake



Automation of noise detection and removal in seismic images (using CNN)



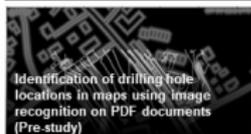
Development of recommendation model for upselling savings products (in development)



ML-training for MSFT partners with global scope (multi-continental roadshow)



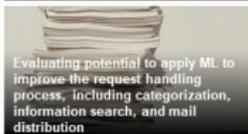
Data analysis and platform development



Identification of drilling hole locations in maps using image recognition on PDF documents (Pre-study)



Interpretation and matching of massive unstructured invoicing data to optimize collection of songwriter royalties



Evaluating potential to apply ML to improve the request handling process, including categorization, information search, and mail distribution



Identification of debtor call patterns to improve collection effectiveness and efficiency



Development of Machine Learning platform



Development of new data platform



AI-Monitoring of applications in DevOps program