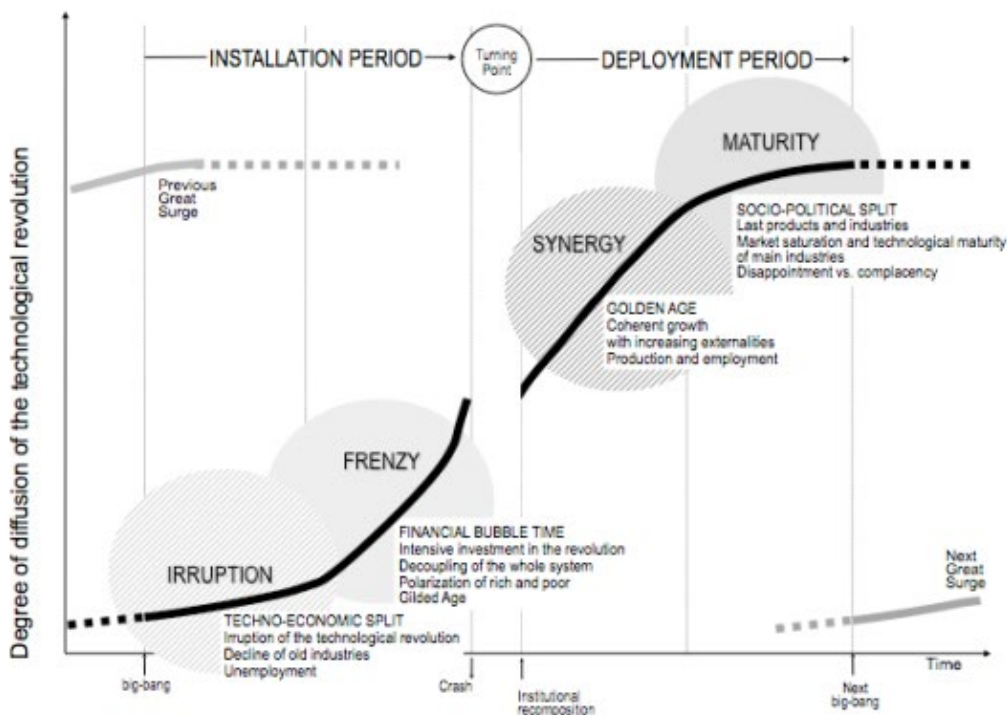# Data Engineering Day 2021

The Data Engineering Day is a yearly event held at BFH TI to foster interaction between academia and industry where we come together and try to tackle some of the big questions in the field of Machine Learning (ML). The 2021 edition focused on the identification of challenges, approaches and visions relating to production-grade ML applications.
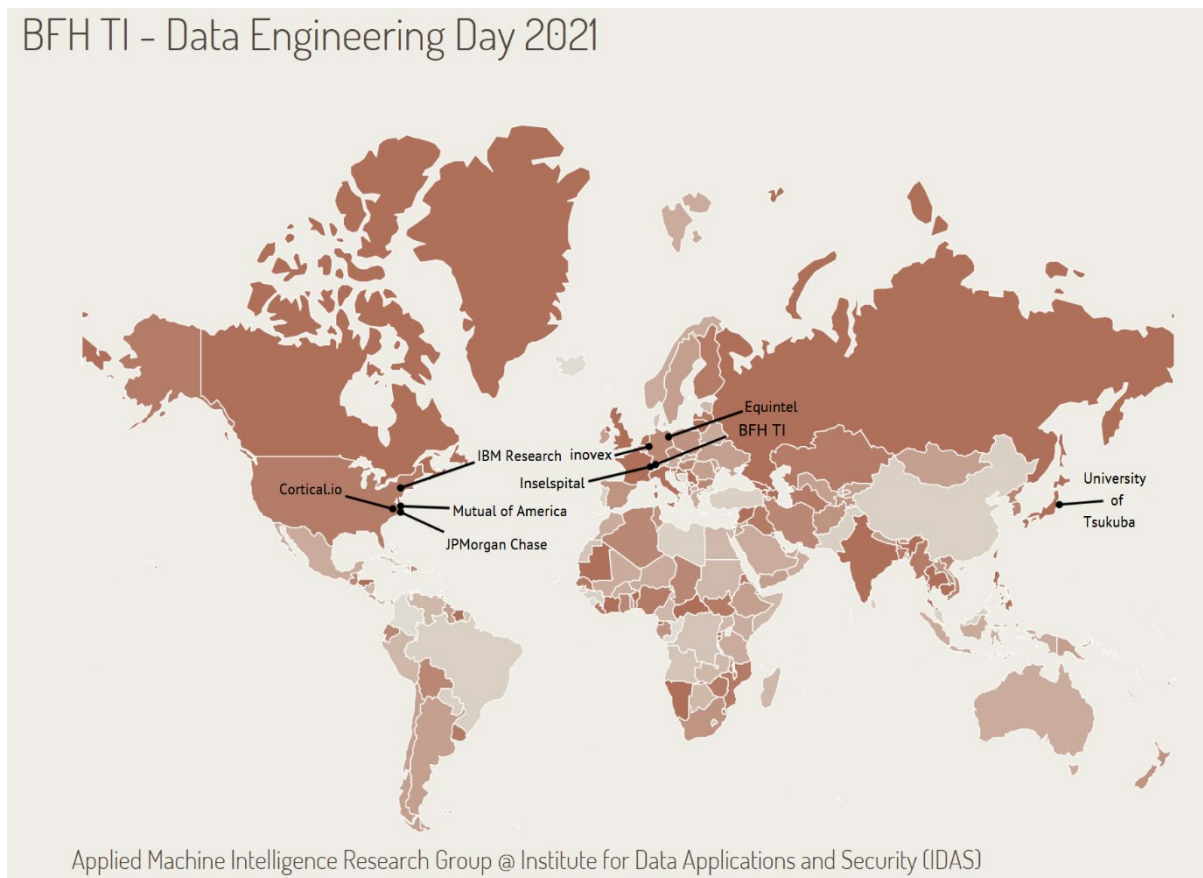
The perspective we take is based on the model of economist Carlota Perez. Carlota Perez studied all major technological revolutions since the industrial revolution and found that there are two phases to them. **The installation phase** when the technology undergoes rapid development, comes into the market and the infrastructure is built (rails for the railroads, assembly lines for the cars and network infrastructure for the internet), and **the deployment phase** when the technology is broadly established (e.g., the adoption of smartphones, ecommerce). In terms of the current ML and AI hype it can be argued that we are still in the installation phase and have not yet reached the state of wide adoption of ML powered products in the enterprise and consumer markets that would mark the deployment phase.



Figure 5.1    Recurring phases of each great surge in the core countries

We invited a set of leading figures from industry and academia from the U.S., Japan, Germany, and Switzerland to outline current challenges of applied ML, hear about the approaches they developed to overcome them, and to be inspired by their visions of future applications of ML.

# BFH TI – Data Engineering Day 2021

Equintel
BFH TI
IBM Research  inovex
Inselspital
Cortical.io
Mutual of America
JPMorgan Chase
University of Tsukuba

Applied Machine Intelligence Research Group @ Institute for Data Applications and Security (IDAS)

### Challenges
The speakers presented a wide range of challenges towards ML adoption across different domains and engineering aspects.

### Adaption to Domain-Specific Requirements
A recuring challenge listed by the speakers is the need for ML to adapt to the requirements of specific domains. Oana Diaconu, Executive Director Data Science and Analytics who joined us from Wallstreet, highlighted that ML models must meet the model governance requirements defined by the regulations of the financial industry. The implementation of these regulations was identified as a prerequisite for the success of ML in the financial industry. Pablo Gonzalvez Garcia, Director of Engineering at cortical.io U.S in New York, struck a similar tone by emphasizing that ML-powered applications need to address data and application security from the ground up to be enterprise-ready. Stavros Zervoudakis, an Adjunct Professor at NYU and senior Fin-tech executive, demonstrated the challenges posed by unbalanced datasets based on the application scenario of credit card fraud. Florian Wilhelm, Head of Data Science at inovex in Cologne, outlined the hurdles that need to be taken for Data Science to make it into production. He explained how the transition to productions poses challenges on the team organizational and software development level. ML adoption it seems is often delayed by difficulties stemming from organizational and domain specific aspects.

### Integrating the End-User in the Solution
The interaction with end-users and ML based functionality is seen as a second major challenging aspect. Benjamin Ellenberger and Violeta Vogel, the Head of Data Science at Insel Gruppe in Bern identified that the requirements of ML based solutions in a hospital context do not stop at high accuracy but also require interpretability and explainability. Hideo Joho, Associate Professor at

Tsukuba University in Japan, provided us with an in-depth look at the challenges stemming from specific user behaviors in search applications. Janna Lipenkova, AI entrepreneur and founder of equintel, demonstrated the challenge to establish the users' trust into ML-powered news curation. The challenges posed by the necessity to gain trust from the users' side were shown by Werner Geyer, Senior Manager AI interaction at IBM Research in Boston Massachusetts. He showed that this applies also to developers and ML practitioners themselves, and that this prevails even in situations where objective measures are available to underline the quality of ML based suggestions. The interaction of human users with ML powered applications exhibits a variety of challenges; many of those essentially rooted in distrust from the users' side.

## Approaches to ML in Production

The speakers provided us with insights towards their own and industry-wide methodologies and approaches that were developed to mitigate and overcome the noted challenges to get ML into production.

### Achieving ML Success through Enabling User Interaction

The interaction of the end-users with the ML-specific aspects of a solution was seen as a key driver of the success of these applications.

Winning user trust is based on two core features of the application functionality developed by Janna Lipenkova and her team. To achieve this, the demonstrated AI Trends & News Monitor provides relevance scores representing the model's decisions and allows the user to extensively drill down into all actual news sources. Werner Geyer at IBM research shared how latest research results shape the development of current and future products. He outlined that providing information that lets the user comprehend the behavior of the underlying ML and AI algorithms, is key to the adoption of these tools.

The involvement of subject matter experts (SMEs) in the ML solution as a key success factor was a sentiment also shared by Insel Gruppe and cortical.io. Pablo Gonzalvez Garcia demonstrated this based on the ability of cortical.io'ss systems to be interpretable from the ground up based on sparse distributed representations, and the tool support for training and evaluating models via an end-user focused user interface. He also showed that the involvement of the SMEs in the curation of ML models benefits the flexibility of ML-powered applications and is key to the company's success in the enterprise market.

### Composition of ML-Specialized Teams

How organizational aspects can contribute to the success of ML applications in production was both outlined by our speakers from Insel Gruppe and inovex. Distributing responsibility for the ML solutions throughout the full development team, and ideally including the subject matter experts is seen as a key factor of solution success. The approach is similar to the "2-pizza-rule" practiced at Amazon.com that dictates that teams should be kept small and hold responsibility for all aspects of a product or feature. Florian Wilhelm and Pablo Gonzalvez both underlined the importance of this approach for production-grade ML development. Avoiding splitting the responsibility between engineering and ML aspects and involving the team throughout the full software development lifecycle is seen as a key ingredient to application success.

## Vision

### Intelligence Amplification

Werner Geyer of IBM research stated that in the near-term production grade success is seen at a level of human – AI cooperation where the human remains in charge of things. He showed the potential of this vision based on demonstrating how ML can automate aspects of software development and the training of ML solutions themselves. The potency of the combination of human and machine intelligence was shown based on the latest shared scientific results. The solutions developed by equintel for information curation and by cortical.io in the enterprise space added further evidence

regarding the enormous potential of such combinations of human wit and ML-powered solutions (leading up to 80% efficiency gains as outlined by Pablo Gonzalvez).

<span style="color:orange">Novel Applications</span>
A very concrete and compelling vision how advances in ML and other technical areas can be combined was shown by Professor Joho from Japan. He outlined how the availability of augmented reality devices and ML models capable of real-time text detection (OCR) could be applied to personalize the automated delivering information in specific contexts such as reading a book. Insel Gruppe demonstrated the potential of developing new types of solutions targeting specific aspects of the university hospital workflow if SMEs and ML engineering come together.

While the spotlight is often on the algorithmic and purely technical aspects of Machine Learning, the 2021 Data Engineering Day has shown us that the success of ML applications requires us to think beyond those aspects. Failure to integrate the end-users and to adapt to domain-specific requirements will prevent us from achieving success. It has also emphasized the need to be visionary and to actively seek to identify new use cases and applications that fit well with the capabilities, but also with the limitations of Machine Learning.