basware
Now it all just happens™

# SmartCoding

## Design and Implementation of ML project

# Contents

# Introduction

basware
Now it all just happens™

# Who am I?

- Olli-Pekka Valtonen
- Education: Software Engineer
- Various roles at Basware:
  - Software Developer
  - System Architect
  - Project / Program Manager
  - Team Lead
  - Technical Product Owner

# What is Basware?

- Provides leading world-class Networked Accounts Payable Saas service.
- Basware invented the Invoice Automation space
- Founded in Finland (came to Tampere through buying Analyste)
- Previously listed in HEX,
- Delisted, owned by AKKR
- ~1500 employees.

# Why SmartCoding?

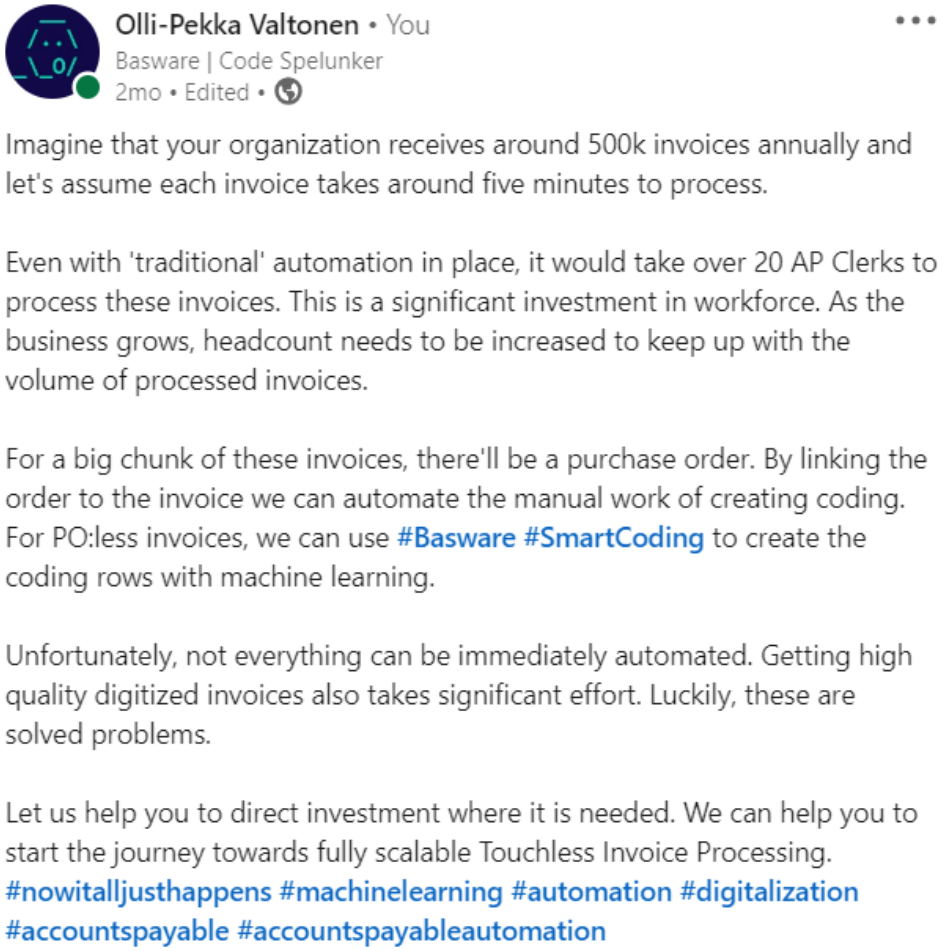
basware
Now it all just happens™

# Why SmartCoding?

## Invoices without purchase order are time consuming & painful

- Examples: Team Dinner, Services, Utilities, Coffee for the office
  - (These also could be with purchase order, Basware offers tools to automate most of these)
- PO:less invoicess need to be coded manually.
- This is work that nobody wants to do.

# Oppoturnity



Olli-Pekka Valtonen • You
Basware | Code Spelunker
2mo • Edited •

Imagine that your organization receives around 500k invoices annually and let's assume each invoice takes around five minutes to process.

Even with 'traditional' automation in place, it would take over 20 AP Clerks to process these invoices. This is a significant investment in workforce. As the business grows, headcount needs to be increased to keep up with the volume of processed invoices.

For a big chunk of these invoices, there'll be a purchase order. By linking the order to the invoice we can automate the manual work of creating coding. For PO:less invoices, we can use #Basware #SmartCoding to create the coding rows with machine learning.

Unfortunately, not everything can be immediately automated. Getting high quality digitized invoices also takes significant effort. Luckily, these are solved problems.

Let us help you to direct investment where it is needed. We can help you to start the journey towards fully scalable Touchless Invoice Processing.
#nowitalljusthappens #machinelearning #automation #digitalization #accountspayable #accountspayableautomation

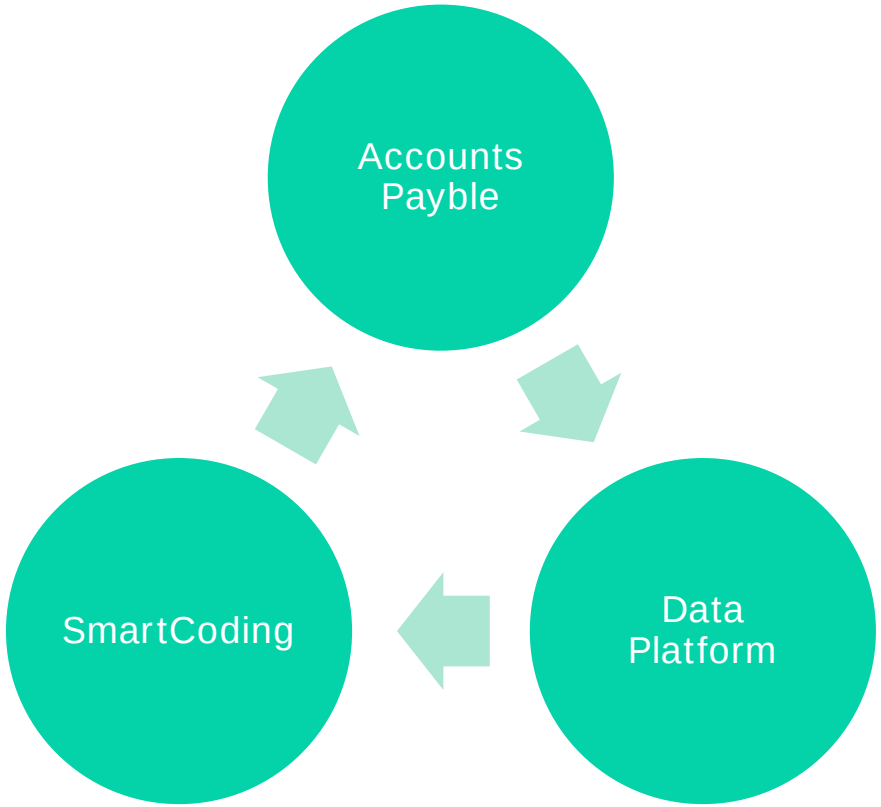
## Increased automation

- Fully automating processing of 50k invoices saves around 2 FTE.

# What was done?

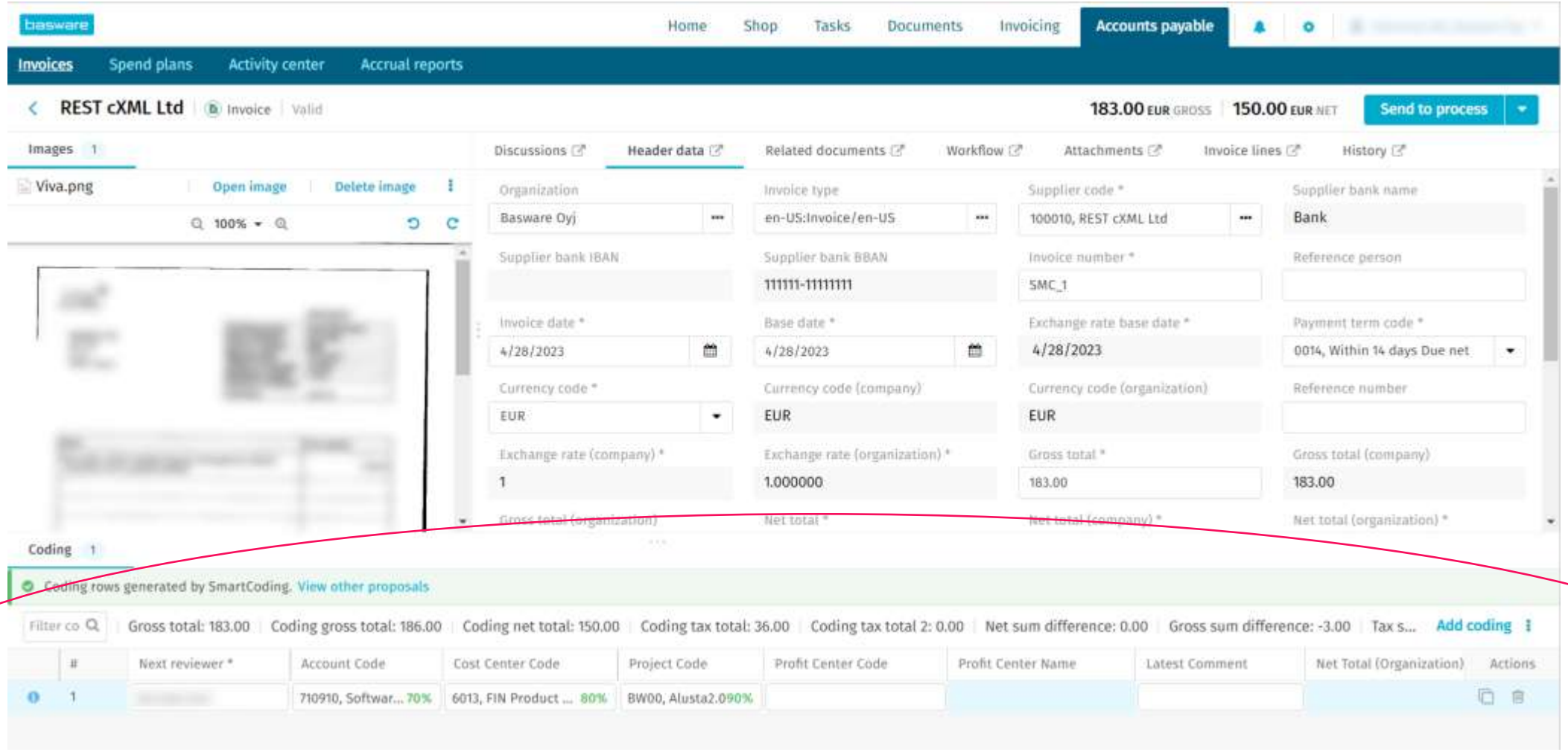
basware
Now it all just happens™

# SmartCoding




- Accounts Payable: Feed succesfully coded invoices into data platform.
- Clean, sanitize and dedup the data in Data Platform.
- Feed the data from Data Platform to SmartCoding microservice.
- In the microservice train machine learning model(s) that predict the best coding for given invoice.
- As PO:less invoice arrives to invoice process, ask the microservice to infer the best coding using ML and display it to the user.

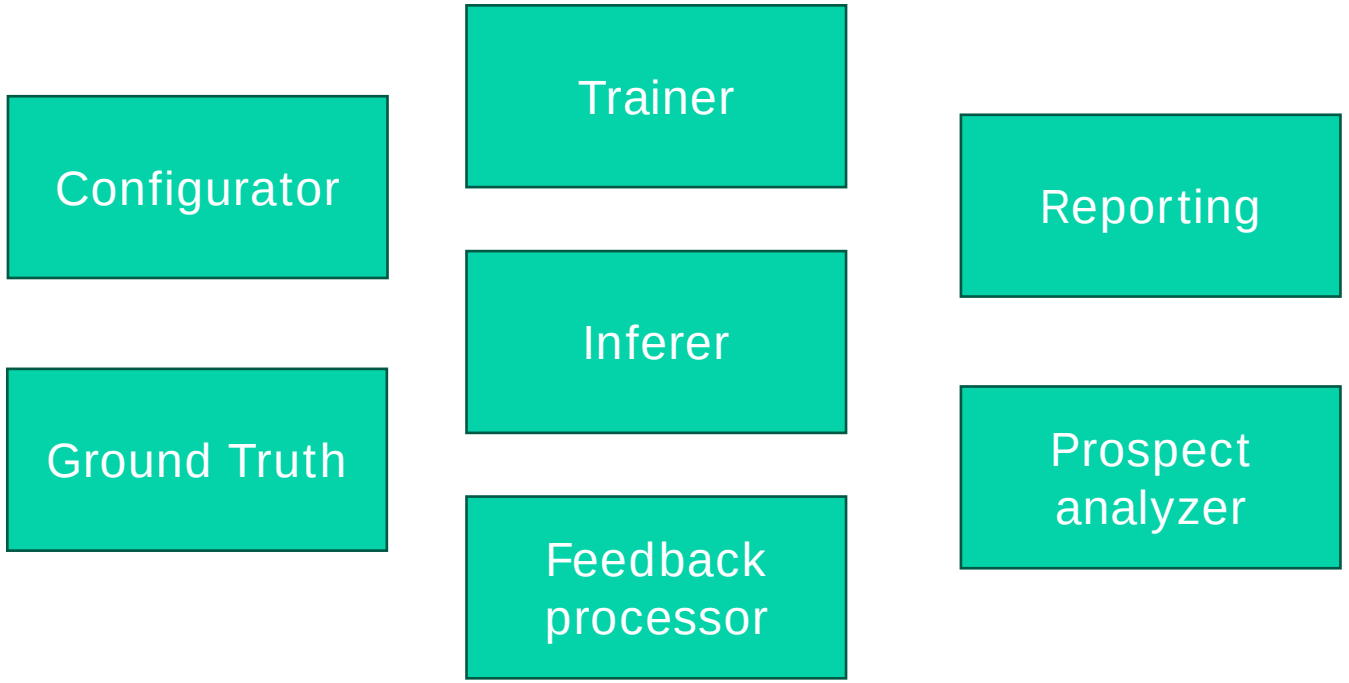
# Accounts Payable



# ML Microservice

## Responsibilities

- Inference
- Training
- Configuration
- Reporting
- MLOPS tool to analyze prospects
- Maintains the ground truth



# Data Platform

**We didn’t do this alone**

- Data Lake / Data Warehouse solution
- Responsible for ingesting and storing the raw data
- Responsible for computing and transforming the data to required format consumed by the ML microservice.

Built by another team – blackbox for the SmartCoding team

# The SmartCoding Project

Recruitment

Delivery to Customers

Research

Design

Implementation

Planning

Risks

Skill rampup

Operability

QA

Dependency to Data Platform

MLOPS

Internal Training

Operability

Business Iteration

# Step 1: Plans & POCs

## Plans

- Project plan
- Business plan
- Architecture plan
- E2E testing plan
- Delivery plan
- Syncing deliveries with Data Platform

## POCs & Research

- Training & Inference POC
- ML Model Research

Most plans were iterated along the way.

# Prereqs for success

Data

Well-formed data

Correct data

People

Data practitioners

Cloud native developers & QA.

Sales that can speak ML

Resources

Computation environment such as Azure or AWS

Operations resources

Time

Whew

# Data

basware
Now it all just happens™

# Let’s talk about data

**”.. bbbut data is the new oil…”**

- Basware has well-formed data from human created and operated systems.
- Basware has typically data that is valid and often correct as it is used for accounting purposes.

Is there machine learning to be done?

- First task for a data scientist is to figure out the nature of the problem (classification vs regression)
- Second task is to uncovern the patterns and relationships to determine whether to use statistical methods or machine learning.
  - Multiple models? A single model?
  - Which method?
  - Etc

**“Data has no value when** sitting idly in a database. This simply is not so. Just as inventory sitting on a shelf in a warehouse has discernible value, so do idle information assets. This is the difference between realized value and the true definition of an asset, which takes into consideration its probable future economic benefit. All information has a probable **future economic benefit.”**

Douglas B. Laney

# People

basware
Now it all just happens™

# Let’s talk about roles

## Data Scientist

Solves the data problem

Research &

Model Design

Has requirements for the data architecture

## Architect

Solves the software problem

Fitting inference into the system

Microservice architecture

Has opinion on the data architecture

## Dev & Test

Develops the service

Implementation of data pipelines

Test automation

Test tooling

E2E testing

## MLOPS

Configures the models

Monitors the system

Deploys the models

Retrains.

# Missing skills

## ML & MLOPS

- Data Scientist
- MLOPS engineer

## Architecture, Dev & QA

- Cloud native development
- Serverless microservice
- Serverless QA

**We didn't really find tasks for LinkedIn**
Influencers with MANIC ENERGY ☹

# Data Practitioner Rampup

## Data Scientist

- Critical for project success

## Rest of the organization

- Very difficult domain for multi-taskers
- Requires basic statistics / math skills
  - (High-school level)
- Both data and ML understanding needs to be developed in sales, marketing and consulting.

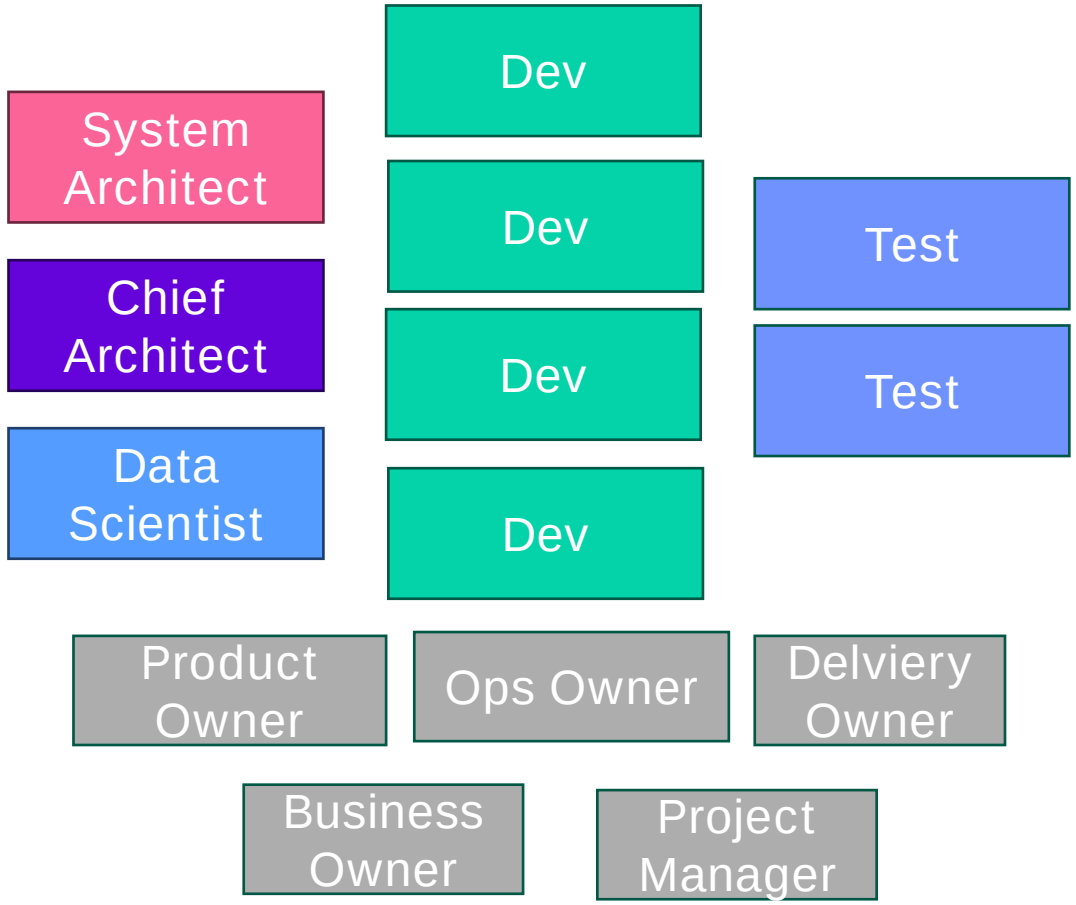
## MLOPS

- Training existing personnel
- Basic Math  / Data Skills
- Ops skills
- Curious mindset

# Team Composition

## Normal software development tasks

- Architect roles are vital to tie together the 'data problem', the 'ML problem' and the 'software problem'.
- Wide variety of tasks: data pipelines,writing ML code, writing tests, writing user interfaces, cost estimations, security and writing code.



It takes around three months of 100% full-time-equivalent on the job learning to bring competent Senior Developer familiar with AWS to get fluent with Serverless Microservice Architecture on AWS

# Resources

basware
Now it all just happens™

# No cloud computing? No ML (for us)

## Mature production stack

- Access to cloud computing resources
- AWS SageMaker for the ML
- Python based microservice stack to run everything.
  - Python seems to be the language of ML
- Something like Splunk will make your life much easier.
  - You want to be able to trace the ’an action’ thoughout the system.

# Operability

## Right skill for the correct task.

- Using R&D Data Scientist for MLOPS tasks is overkill.
  - Results in unhappy people.
- Overlap with production tasks
  - Monitoring, configuration, problem solving
- Navigating data practitioners can be hard: who is the right person?
  - Production Data Scientist, Data Engineer, ML Engineer...

## Growing new muscles is very hard...

- and should not be underestimated.

# Delivery

Think very hard before you make anything configurable...

- Especially if you intend to take it into use for 1000+ customers.
- Could it still be automated?

Whether you like it or not, somebody needs to support the delivery...

- Which is particularly annoying when you can’t rely on on person due to timezone differences.

Training an organization to do something for the very first time...

- is very frustrating.

Machine learning is not a deterministic system.

Biggest challenge is to tell an end user that the system is predicting a particular case with 65% confidence and that it might be wrong.

Yet looking at the statistics, the system is predicting with 85% accuracy and performing well.

# Data Scientists require access to prod data

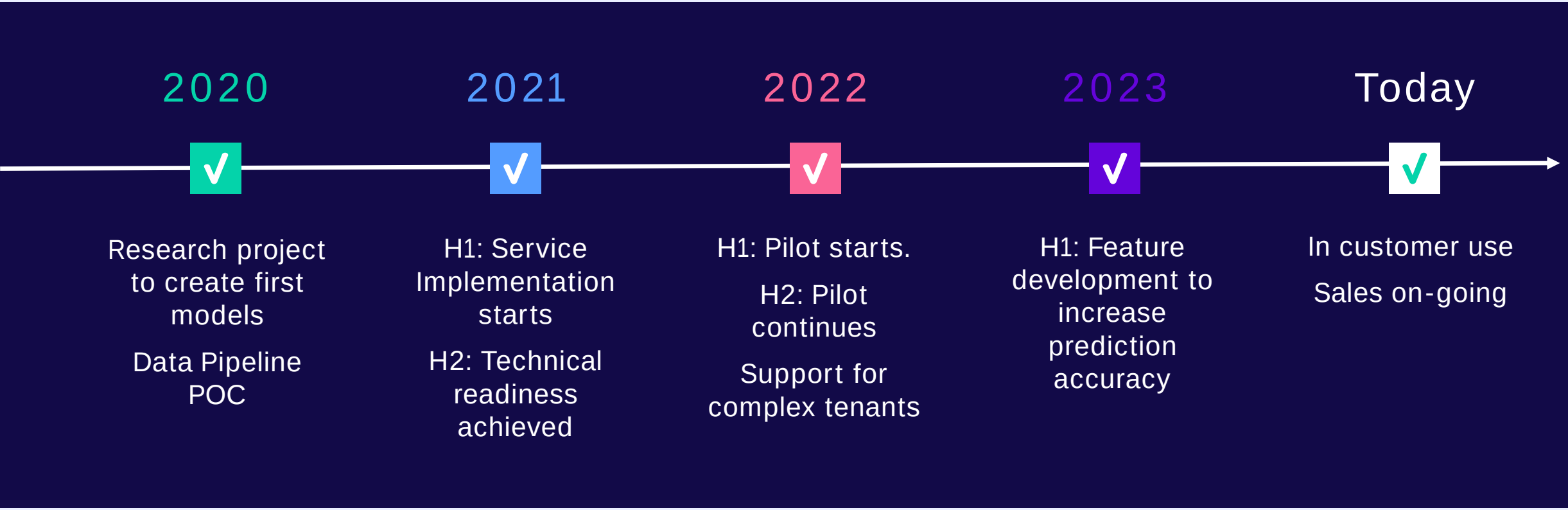
In some orgs R&D **doesn't have production** access...

- This changes with the introduction of data scientist. Production data is important resouce for any ML project.
- Access should be lightweight.
- Building suitable test data is difficult for complex systems
  - Map, pertubate, mask but retain statistical distribution of prod data and bring it to lesser environments.

# Time

basware
Now it all just happens™

# Let’s talk about time



# Conclusion

basware
Now it all just happens™

# Key learnings 1/2

## Planning & Pocs

- Time spent on building POCs & training was pretty small.

## Still mostly a software problem (for project manager)

- 85% software problem, max 15% ML problem

## Machine learning is not deterministic

- People's math skills suck and conceptualizing confidence vs accuracy is difficult.

## Building & operating a distributed system is expensive

- Requires relatively rare talent and and advanced skills through the organization.

## Training competent people is easy, but...

- Finding the right talent is hard.

# Key learnings 2/2

## Leave room for business iteration

- We redid many things as we learned: introduced multiple models per tenant, redid billing

## Product Management is important

- Users are impacted by the non-determinism.

## Working with R&D people is easy

- So easy it distorts your reality. With non R&D people not so easy.

## Do not underestimate developing new skills and talent

- Skill rampup and recruitment take time.

**You don’t really need big data solution in the beginning…**

- but it is useful later on.

# Thank you

Olli-Pekka Valtonen


basware
Now it all just happens™