Aktuárská věda vs. moderní data science - možné synergie a vzájemná inspirace

Petr Bednařík
petr.bednarik@datasentics.com

SAV 17.5.2019

Můj životní příběh a úhel pohledu

DataSentics

Matematické modelování

Finanční a pojistná matematika pro aktuárskou certifikaci

Deloitte.

1. Aktuárský konzultant
2. Specialista na datovou kvalitu a data management
3. Specialista na využití dat a IT v pojišťovnictví

9 let

Parciální diferenciální rovnice
Numerika Fraktály MATLAB
Trojúhelníky Áčka Úmrtnost
Stochastické metody
ALM Interní modely
Prophet Simulace
Solvency II
Master data DWH
Reporting ETL Cross-sell/up.sell
CRM Fraud CLV

DataSentics

Data science architekt a zakladatel

3 roky

Strojové učení AI
Digital Cloud
Big data technologie

Čím že se to teď zabývám

DataSentics

Accelerating the **digital transformation** of **enterprises** in Europe with the power of **machine learning**

How We Work

ML Center of Excellence in Prague & Bratislava serving Central and Western Europe

Data science
Machine learning
(incl. NLP, image, graphs…)
25+ specialists

Data engineering
Cloud
(Azure, AWS)
15+ specialists

We become your data innovation strike team within your digital transformation

ML use case ideas → Experiments/ prototypes → Automated ML solutions into production

Powered by

databricks

We are one of the first European partners

Microsoft

We are a Microsoft partner specialized on AI in Azure

aws

We are an AWS partner specialized on AI

Key Industry Focus

Financial Services

Retail/FMCG and ecommerce/digi

Key References

ERSTE, MONETA, Allianz, VIG, Asahi, MALL, IKEA, dentsu AEGIS network, pepsi, socialbakers, Red Bull

Main Solutions

Customer Engagement 360°
Combining digital and offline behavioral data and machine learning to personalize the customer experience

Fraud / AML / Risk
Using advanced methods such as NLP and graph analytics to identify anomalies, etc.

Contact center
Using NLP for automatic routing of emails/request, topic detection, etc.

Shelf Inspector
Computer-vision-based AI solution for automatically analyzing the quality of placement of my products in a store shelf

We will help you build your own best practice cloud data platform using our ADAP framework.
Getting the best "data science experience" for your teams unifying data scientists and data engineers.

Agile Data Analytics Platform (ADAP)

Best Open-source Big Data Engines: Apache Spark, python
+
Unified Analytics Environment: databricks
+
Infrastructure, Scalability and Automation: Azure or aws

- A Fully Open Data Platform for ML Use Cases in Production, Automation + CI/CD
- Best Practice Lambda Architecture Using Standard Cloud Components
- Deployed to Your Own Cloud Environment

0. Data science – co tím vlastně myslím

Data analytics hype-word ecosystem

CLOUD COMPUTING

MACHINE LEARNING

DATA VISUALISATION

BIG DATA

HADOOP

NEURAL NETWORKS

BUSINESS INTELLIGENCE

DATA SCIENCE

PREDICTIVE MODELLING

NATURAL LANGUANGE PROCESSING

DATA ANALYTICS

ARTIFICIAL INTELLIGENCE

DEEP LEARNING

DATA WAREHOUSING

DATA ENGINEERING

NOSQL

Data analytics hype-word ecosystem

DATA SCIENCE
DATA ANALYTICS

Improve something in your business by using a combination of
data + **data analytics method** (with more or less **mathematics)** + **data technology**

DATA ANALYTICS USE CASE

What business problem we want to tackle (improve success rates of follow-up calls, improve efficiency of marketing spend, …) + input data + methods used + outputs + how exactly will the outputs be used (ideally automatically – e.g. call centre agents will be calling the clients with the highest score first)

Mathematics light -> Business intelligence **Mathematics heavy -> Data science**

DATA ANALYTICS METHOD

BUSINESS INTELLIGENCE
DATA VISUALISATION
DASHBOARDING/REPORTING/KPIs

ARTIFICIAL INTELLIGENCE
MACHINE LEARNING
NATURAL LANGUANGE PROCESSING
PREDICTIVE MODELLING
NEURAL NETWORKS
DEEP LEARNING

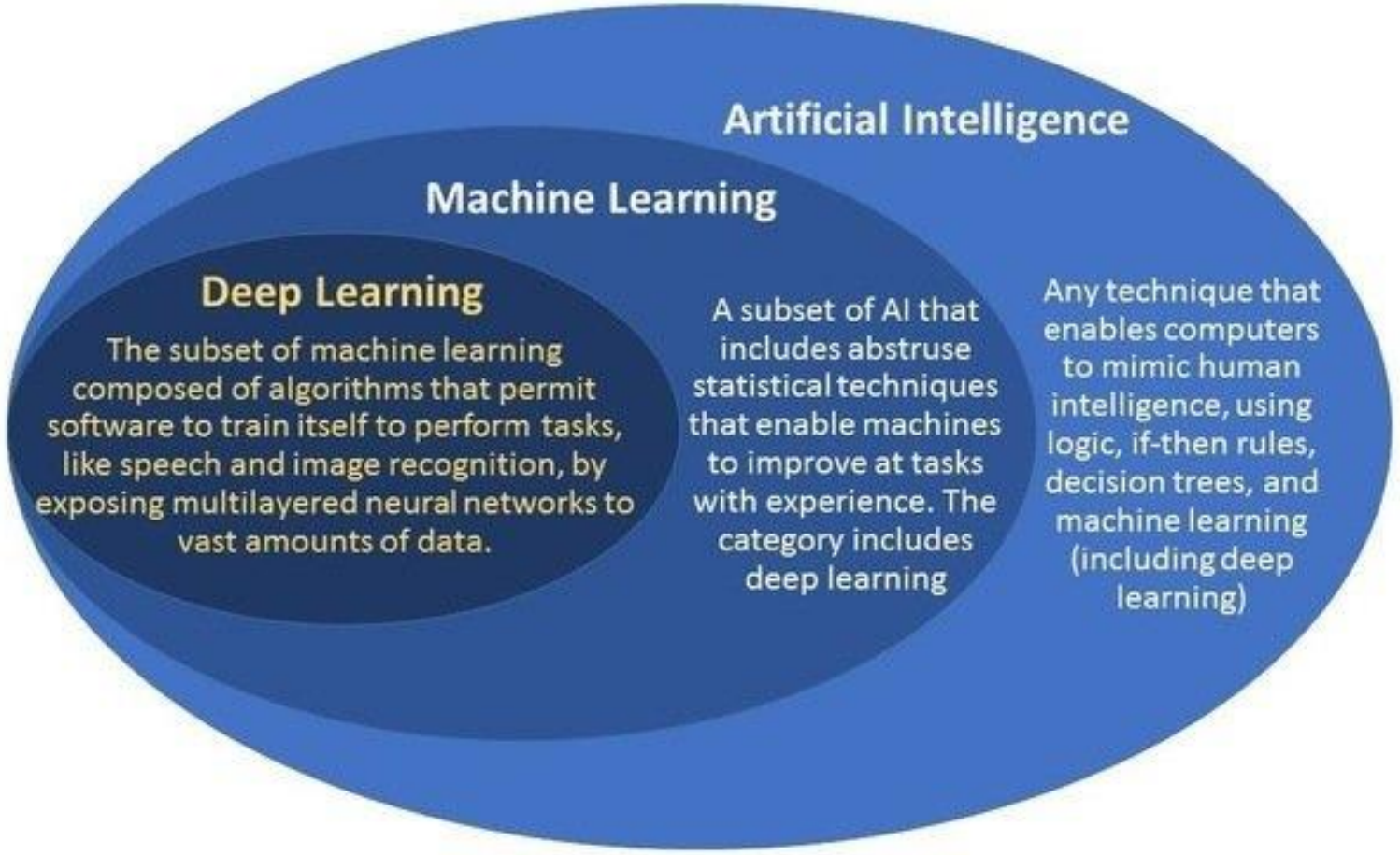
DATA TECHNOLOGY

DATA WAREHOUSING
BI TOOLS

DATA ENGINEERING
BIG DATA
HADOOP
NOSQL
CLOUD COMPUTING

DATA SCIENCE TOOLS

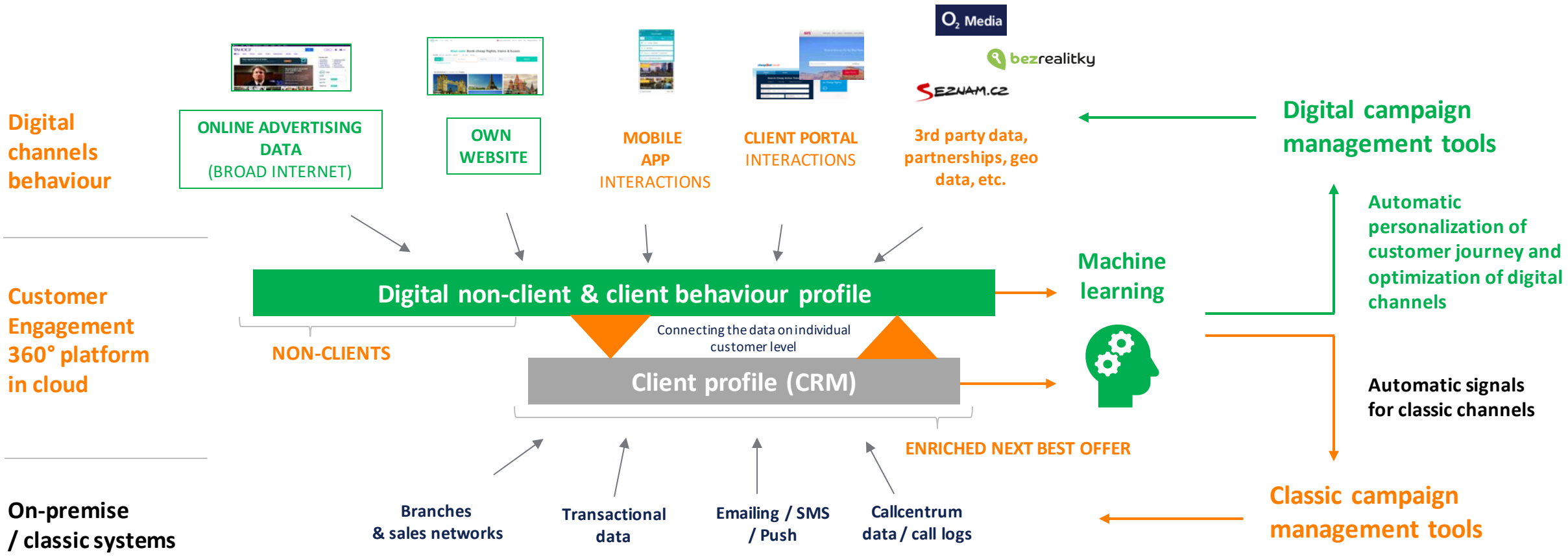
Data analytics hype-word ecosystem



Display ad purchasing optimization

DataSentics

Get more leads (serious interest in product) from RTB campaigns with the same budget (or the same amount of leads for less)



Display ad purchasing pipeline

DataSentics

1) Download data from ad system every day – currently working with 1,5 billion impressions

2) We see what person (anonymous but individual) visited, which sites / what articles he read, etc.:

CookieID	concated_urls
-9213710615765419756	https://www.lide.cz/ https://www.lide.cz/ https://www.lide.cz/
-9201554805977552101	https://www.novinky.cz/komentare/490653-komentar-milos-zeman-je-pry-arijec-jiri-pehe.html https://www.csfd.cz/film/297835-un-mari-de-trop/prehled/ https://www.csfd.cz/tvurce/2009-alain-delon/ https://www.csfd.cz/film/32506-zlocin/prehled/ https://www.csfd.cz/film/24978-stir/gale https://www.csfd.cz/film/32492-v-plnem-slunci/prehled/ r2b2.cz https://www.csfd.cz/tvurce/2008-jean-paul-belmondo/ https://www.csfd.cz/tvurce/2009-alain-delon/ r2b2.cz https://technet.idnes.cz/ https://www.csfd.cz/televize https://www.csfd.cz/film/32561-druhy-dech/prehled/ https://www.csfd.cz/film/159899-utajeny/prehled/ https://www.csfd.cz/film/32563-doterny-chlap/prehled/ https://www.csfd.cz/ r2b2.cz https://b.cpex.cz/01 r2b2.cz
-9181889006973565627	https://www.sport.cz/ostatni/ostatni/clanek/1067439-nevim-jak-to-petr-dela-rika-sablikova.html#hp-artcl
-9180351750282907576	https://www.sbazar.cz/ https://www.sbazar.cz/ https://www.sbazar.cz/ https://www.sbazar.cz/ https://www.sbazar.cz/

Showing the first 1000 rows.

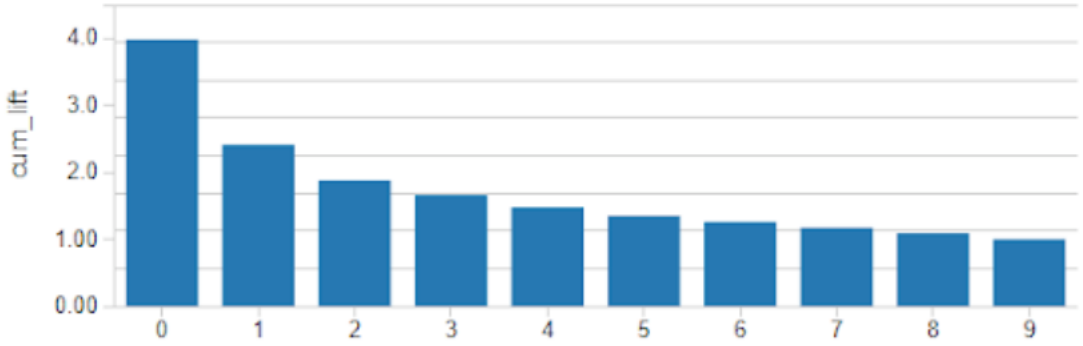
3) For each person we identify interests/topics based

CookieID	tokens
-8778597263115494681	["zpravy","aktualne","zahranici","donald","trump","interrupce","usa","spojene","staty","potrat","amerika"]
-8768341419222352654	["idnes"]
-8761763359002315682	["blesk","clanek","celebrity","serialy","filmy","kino","smutna","dohra","vymeny","manzelek","potrat","kvuli","stresu","stehovani","blesk","clanek","celebrity","serialy","filmy","kino","sm
-8759332933456071596	["penize","mesec","clanky","leasing","akontace","delka","splaceni"]
-8742275175582910903	["bezrealitky","nemovitosti","byty","domy","nabidka","pronajem","bytu","eliasova","praha","mail","centrum","source","centrumhp","content","default","term","position","expres","ivan
-8727196472789610958	["horoskopy","blesk","clanek","nove","horoskopy","astroclanky","astrologie","horoskop","lasky","pro","vodnare","beranem","nikdy","nebudete","nudit","kozorohy","vytacite"]
-8726898350420273305	["super","neni","pry","protekcni","dite","podivejte","jak","vypada","dcera","michala","suchanka","ktera","snazi","prosadit","divadle","novinky","kultura","zemrel","herec","jiri","pecha",
-8718405020472090768	["idnes","revue","spolecnost","petr","vancura","jan","cina","jsou","par","lidicky"]

Display ad purchasing pipeline

4) Predictive model correlating desired behaviour (serious interest in product on website) with broad internet behaviour and interests/topics – using NLP methods (TF-IDF, LDA, word2vec, supervised learning)

List of cookies IDs (audience) with high probability of interest in product
Top 10% of people based on model have a 4x serious interest in product



5) Predictive model to rate quality of ad space (visibility time, visibility quality, robots, …)

Whitelist of domain with coefficients describing quality

6) Campaigns automatically defined in display ad system every day = audience (who to target) + domain whitelist (where)

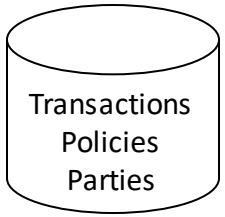
Actual A/B test results very close to theoretical model lift

Fraud/AML checker

DataSentics

Typical traditional fraud and AML solutions are based on expert rules and are difficult to tweak/maintain (react to changes in fraudster behavior, etc.) and do not scale well in case of increase complexity. Our solution enriches the traditional rule based approach by combining it with machine learning based algorithms (anomaly detection using clustering and graph methods + auto-weighting of rules) and thus giving a more reliable and accurate engine for detecting suspicious cases. All of this is powered by Spark making it easily scalable.

1. Raw data



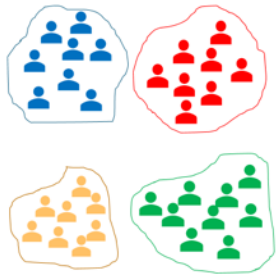
2. Input data preparation

3. Features (360 view) for each party

PARTY_SK	Large_Topup	Premium_Cash_AVG	Premium_Wire_AVG_3M
1000216066	210,7840684	0	393,8586
1003923840	0,366319333	12,72668067	0
100818586	2,542911333	57,443	0
1009082064	2,394485786	55,90654421	0

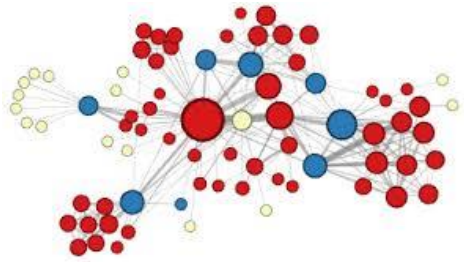
4. Combination of different predictive methods

Clusters



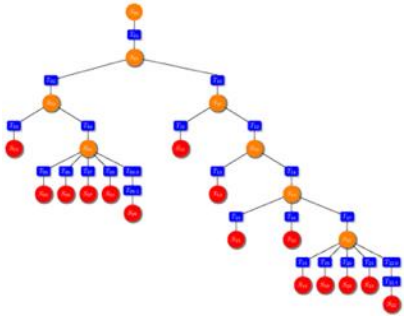
+

Node2Vec



+

Rule based system

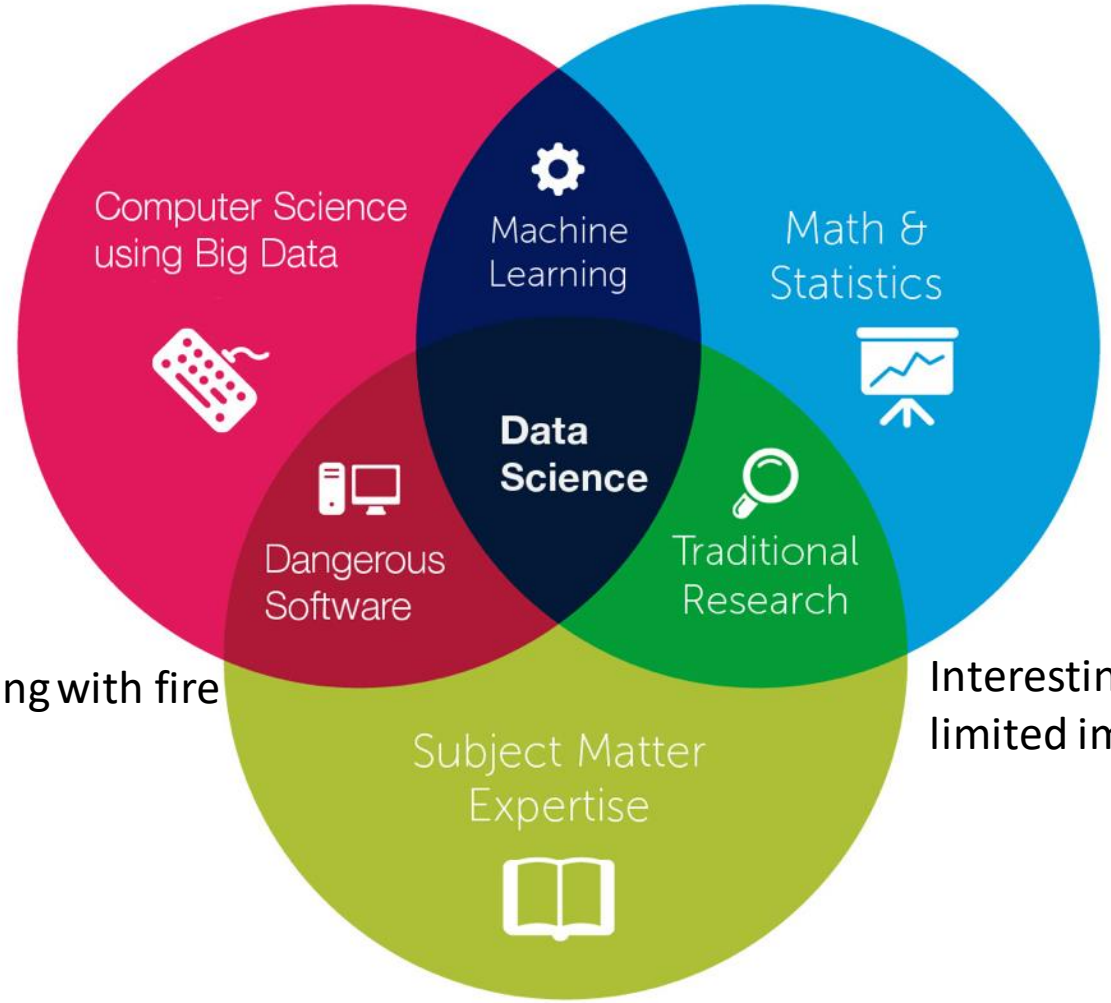


5. Parties sorted by risk score

PARTY_SK	RISK_SCORE
1015096836	0,96
1000216066	0,87
1034515283	0,79
1053521418	0,77
100818586	0,77
1010633942	0,72
1009082064	0,7
1009126212	0,68
1010359908	0,53

Data analytics hype-word ecosystem

Tech demos with limited real impact

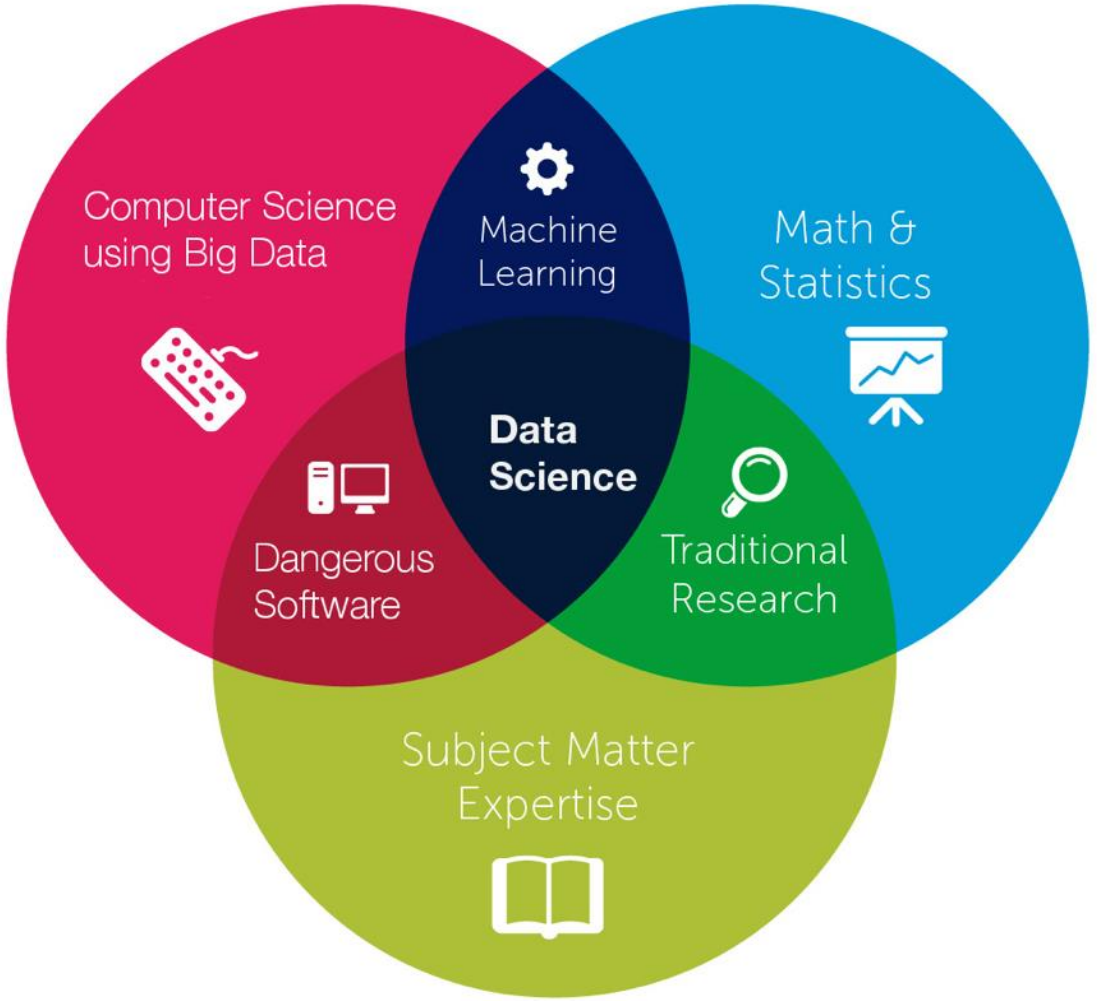


Developers playing with fire

Interesting slides and graphs but limited impact in real life

Aktuárská věda x data science?

- Profese s dlouhou historií – „původní data scientisti v pojišťovnách“
- Když vznikala, tak Computer Science a Big Data ještě neexistovaly
- Když vznikala, tak hlavní datové problémy v pojišťovnách byly kolem rezerv a pricingu – jiné věci pomocí dat moc nešly řešit (obchod, marketing, fraud, likvidace, …)
- Toto zaměření ale do velké míry přetrvalo
- Proto se profese data science a aktuárů začaly oddělovat / vzdalovat



Aktuárská věda x data science?

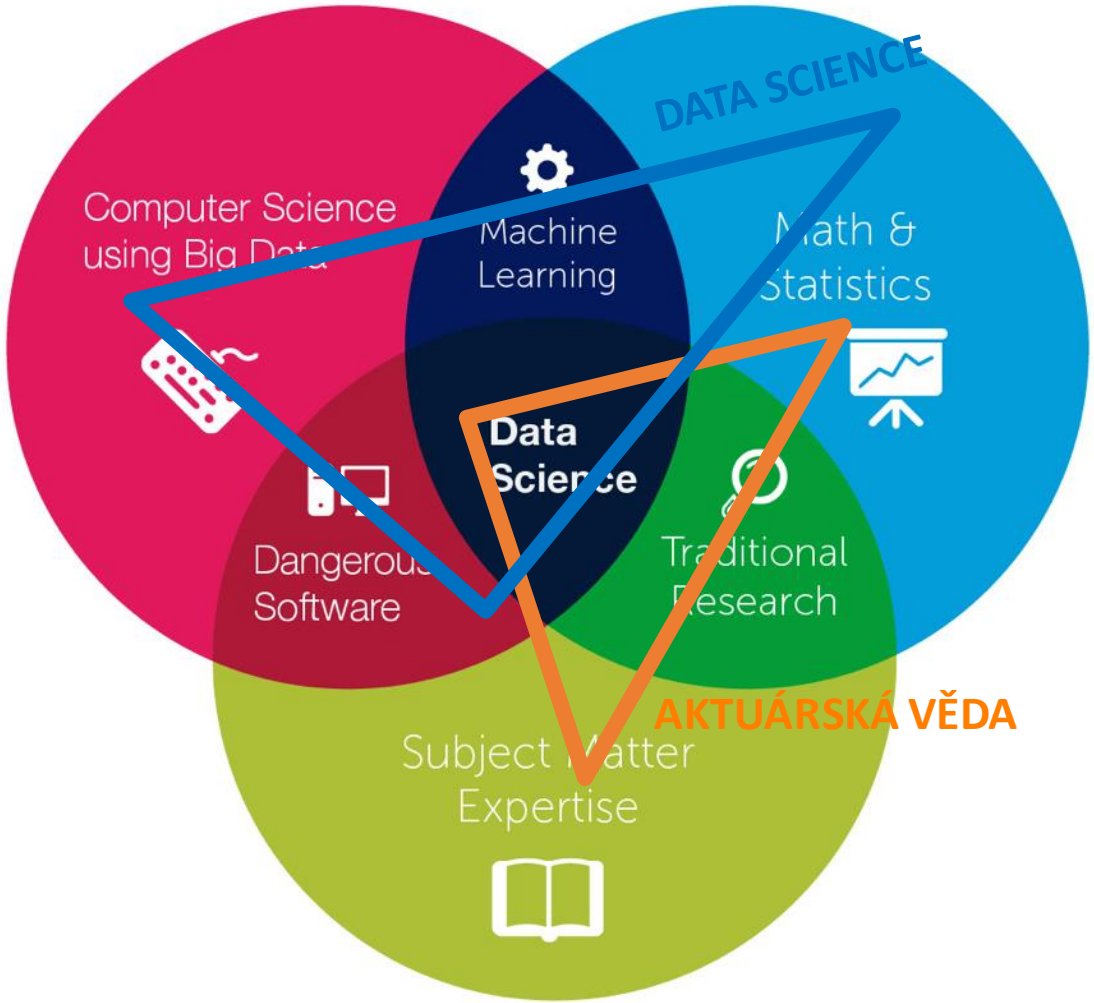
Věřím, že se aktuáři mohou v některých věcech inspirovat

DataSentics

Aktuárská věda zde často nedává velký důraz

Omezení plynoucí z typických technologii v aktuárské oblasti

Velká závislost a těžká spolupráce s IT – manuální procesy, nekvalitní data, vzájemné nepochopení a frustrace…



Je hodně moderních matematických metod, které jsou nyní dostupné díky technologiím (pokročilejší modely, neuronky, …) – aktuárská věda na ně nedává velký důraz

Nové metody z oblastí strojového učení a AI pronikají do aktuárské vědy velmi pomalu – tradiční metody jsou málo kdy vyzívané

V rámci pojišťoven je hodně data-driven procesů, na které aktuárská věda nedává velký důraz (obchod, marketing, likvidace, fraud, …)

Zajetí regulatorních problémů a fakt, že aktuárská práce a výstupy nemají vždy takový dopad na život firmy a užitek, jaký by mohly

Jak mi aktuárský background pomohl v data science

Pravděpodobností modelování dlouhodobého cash-flow

CLV model – Mana

- Performance of acq. cohorts in months
- **Consumption approach with no cancelation**

Akt. tydeni spotreba	Hodnota	Leden	Unor	Brezen	Duben	Kveten	Cerven	Cervenec	Srpen	Zari	Rijen	Listopad	Prosinec
do 1	0,5	0	36	62	83	99	113	125	136	145	153	160	167
do 2	1,5	0	25	50	75	100	124	147	168	188	206	223	238
do 4	3	0	43	88	126	156	177	191	199	204	205	204	202
do 10	7	0	194	277	305	307	297	283	268	253	240	229	218
nad 10	12	1000	703	523	411	338	289	255	229	211	196	184	175

Inputs:

- Relevant segmentation rule
- Relevant timeframe
- Product (SKU) transitions
- Margin
- …

	X12d	X35p	X24.96d	X70p	X108d	X140p
12d	27.7%	34.2%	25.6%	7.6%	2.3%	2.6%
35p	4.7%	70.6%	9.2%	7.4%	5.3%	2.8%
24-96d	2.9%	2.9%	54.9%	30.8%	0.1%	8.4%
70p	1.9%	3%	13.8%	67.1%	0.3%	13.9%
108d	2.8%	11.5%	2.4%	3.6%	72.6%	7.1%
140p	1.5%	1.8%	5.9%	11%	1.2%	78.6%

Jak mi aktuárský background pomohl v data science

Pravděpodobností modelování dlouhodobého cash-flow

Koncept předpokladů – kombinace historických dat a expertního úsudku, backtesting

Koncept best estimate / očekáváné hodnoty a distribuce výsledků / kapitálu – scénáře, jejich pravděpodobnosti, riziko

Možná inspirace pro aktuáry

Moderní data science přístup a způsob práce

DataSentics

0. Data science: co tím vlastně myslím

1. Přístup: agilita a proces standardizace
Od explorace problému, přes prototyp k udržitelnému produkčnímu řešení

2. Tým: struktura data science týmu
Nastavení role vůči IT, jak dosáhnout větší nezávislosti a spolehlivosti

3. Technologie: jak mohou pomoct
Open source, big data, cloud, vizualizace, continuous integration, …

4. Data science metody: využití v aktuárských problémech
Prediktivní modely, clustering, analýza textu, …

Shift in data science happening…

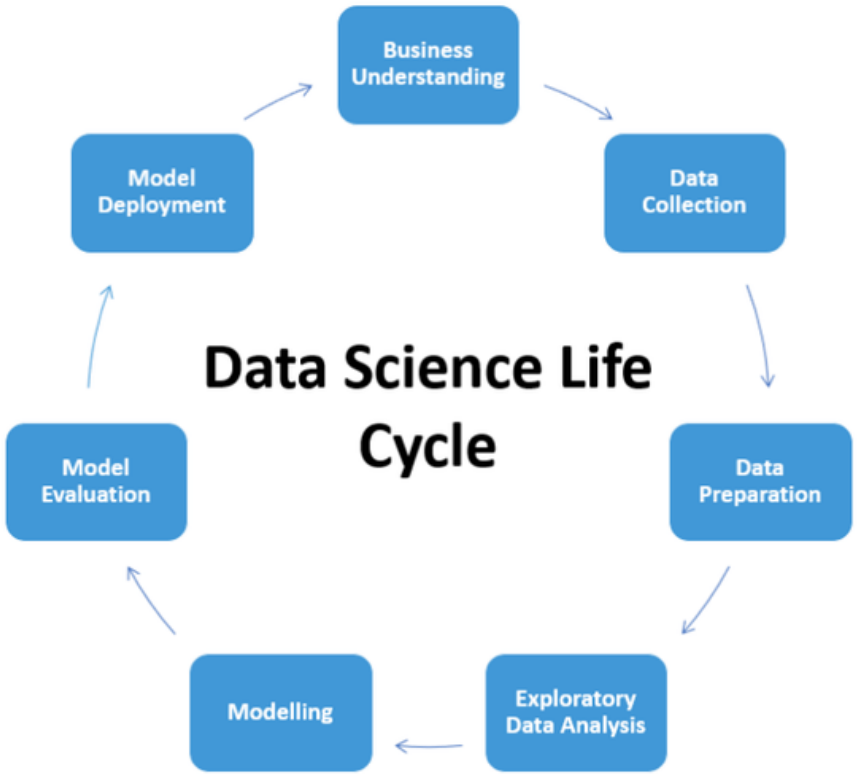
DataSentics

	Oldschool data analytics	Modern data science
Approach	Large projects with uncertain business value Black-box solutions Ends with a presentation and dashboards	**Agile experimentation**, fail fast, exploration Focus on owning the know-how, gradual improvement **Automated data solution in production**
Team	Siloes teams across business, IT, data mining Don't talk to each other, difficult coordination	Small cross-functional teams – data strike team concept
Data	100MBs, Manual extracts Mostly internal Monthly calculations	**GBs – TBs, Automated data pipelines** Internal and external/online **Real-time / daily**
Methods	Business intelligence Basic predictive analytics	Automated decision-making Advanced machine learning
Tools	Proprietary and closed Expensive Siloed	**Open-source based** Inexpensive Integrations first
Infrastructure	On-premises – in-flexible and expensive On your notebook – difficult collaboration Large servers – limited/expensive	**In the cloud – scaleable and cost-efficient** Notebook as a terminal **Distributed computing – scaleable/cheaper**

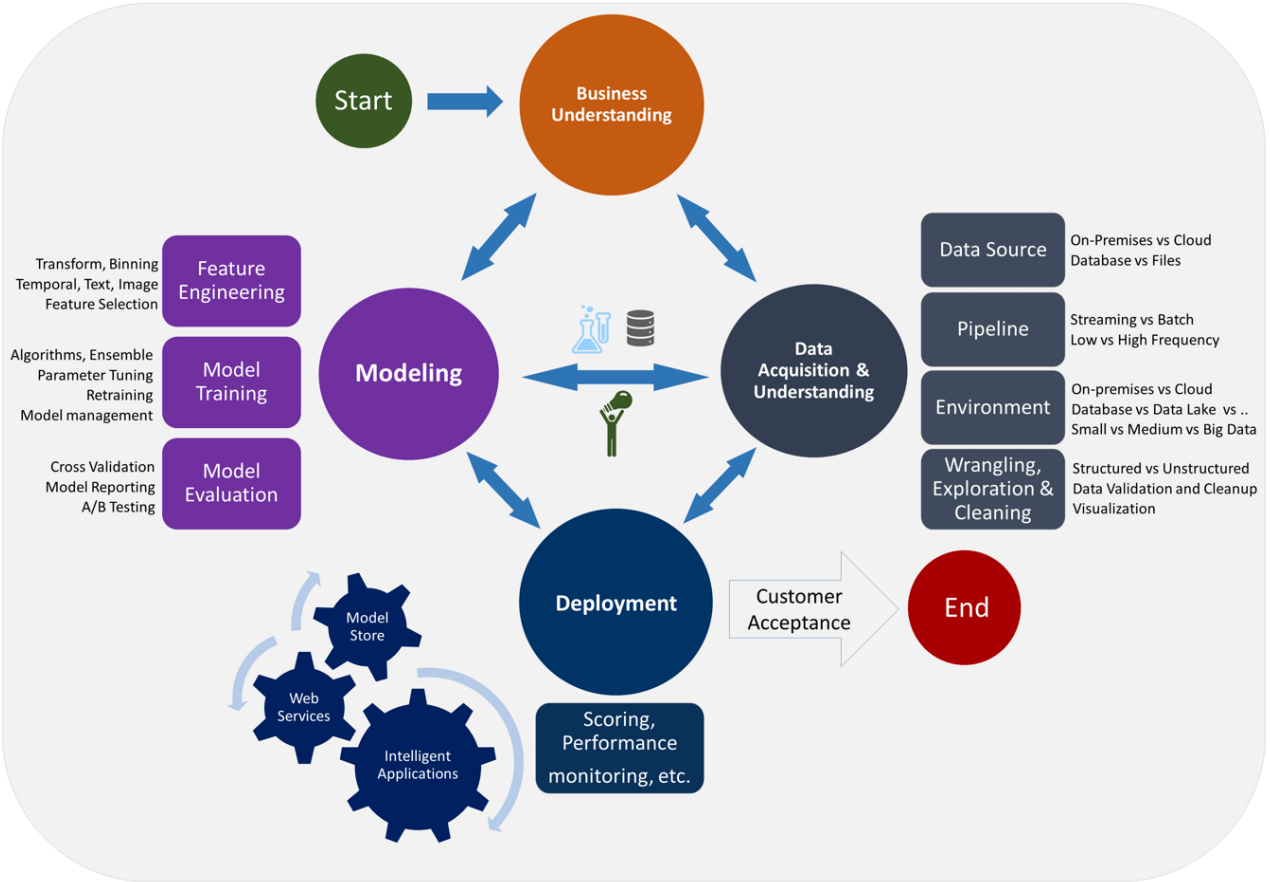
1. Přístup: agilita a proces standardizace

Data science process

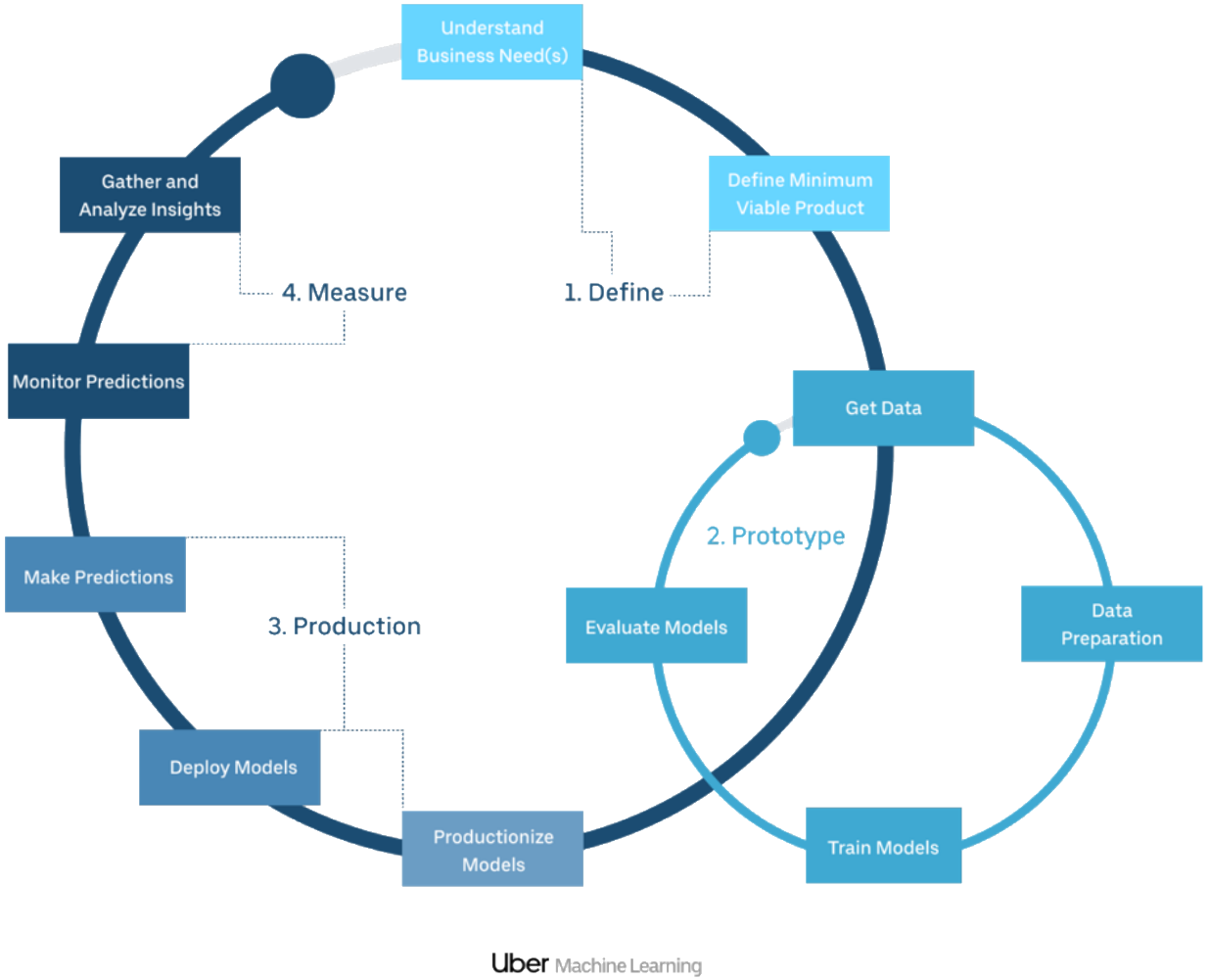
Textbook approach = linear



Reality = back and forth



Data science best practice process – agile/iterative



Example (1 month iterations):

Iteration 1: Pilot / PoC / Prototype / Experiment

Iteration 2: MVP (Minimum viable product) in light automation mode

Iteration 3: Key improvements to algorithm based on feedback from live runs

Iteration 4: Productionalization

Iteration 5: More advanced improvements

Data science best practice process – agile/iterative

DataSentics

Iteration 1 **Prototype**	Iteration 2 **MVP**	Iteration 3 **Version 2.0**	Iteration 4 **Robust**	Iteration 5 **Version 2.5**
Ad purchasing – test if url history predicts click using simple linear regression model	Simple automation – download data, score model predictions, send top 20% to ad systém	Clicks are not a good target -> use offline purchase data as new target for model	Standardize code, implement logging and model performance monitoring, …	Improve model performance by using model advanced model (xgboost)

Backlog:

- identify topics from urls
- download actual pages and analyze content
- detect sex and age from type of content
- …

Benefits:

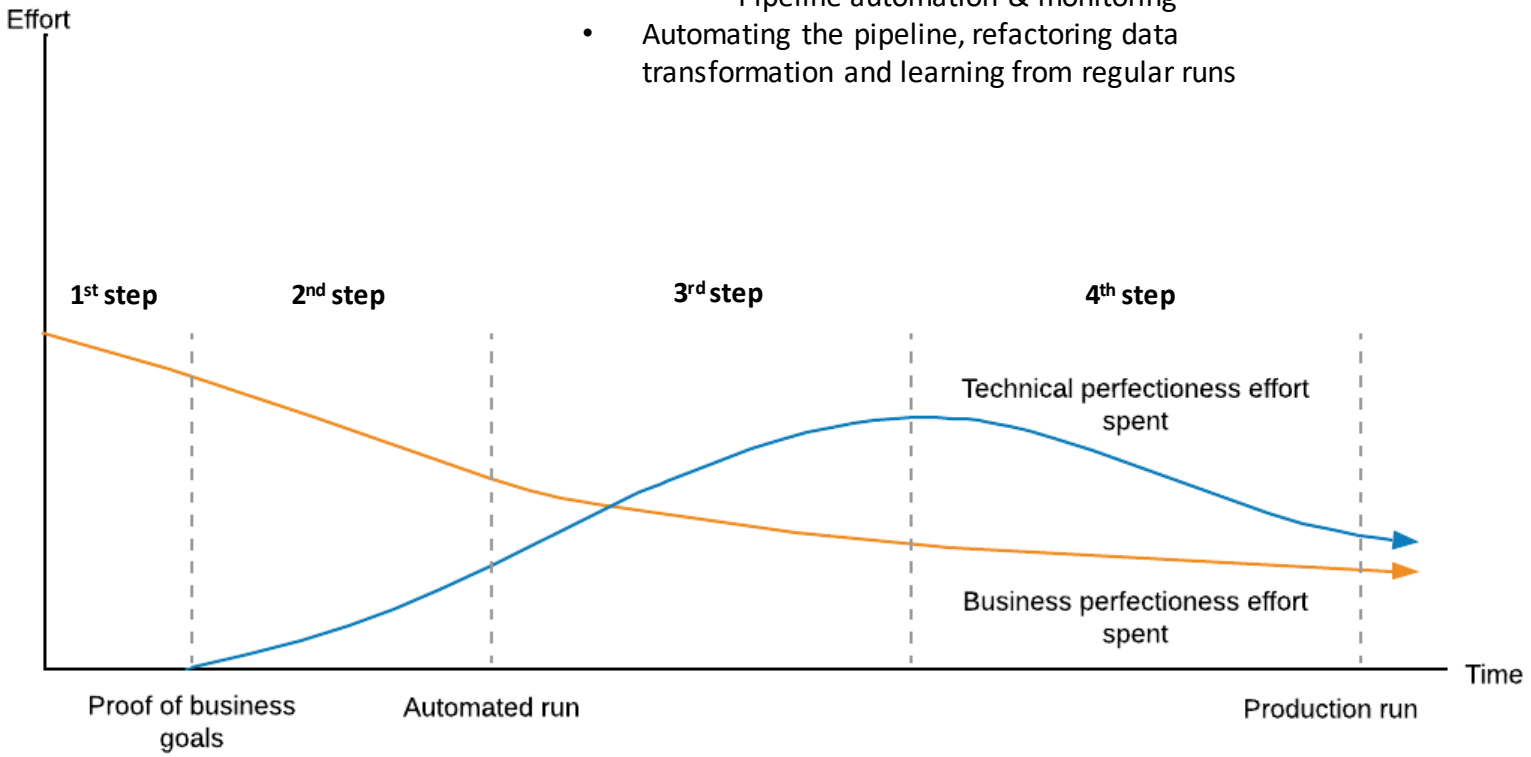
- At the start I only needed to invest into Iteration 1
- Already from point 2 I am getting business value
- I didn't know what the most important thing for iteration 3 to improve will be at the start
- I didn't know if the invest into Iteration 4 makes sense

Balance of business/algorithm vs. technical effort

In 1st step (DEFINE &PROTOTYPE)
- Meeting business goals
- Gathering data
- Basic environment setup

In 2nd step (LIGHT-AUTOMATION)
- Setting up automation tools and minimalistic governance based on our best practise (MVP)
  - Data lake & Databricks automation
  - Security
  - Pipeline automation & monitoring
- Automating the pipeline, refactoring data transformation and learning from regular runs

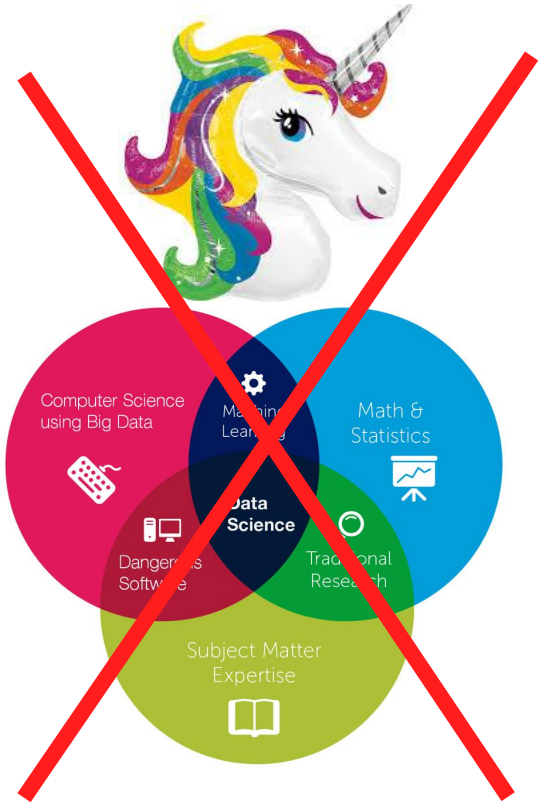
In 3rd step (IMPROVING BUSINESS OUTCOME)
- Automating the pipeline based on learnings from regular runs
- Customizing/Improving the governance & toolset based on actual priorities
- Creating a backlog for further development



In 4th step (PRODUCTION)
- Stabilizing business outcomes
- Agreeing on minimal production run requirements and meeting them
  - Security/privacy
  - Operation/Documentation
- Issuing/Improving standards for similar use cases
  - Including scaling up in terms of pipelines and business criticality
- Setting up basic operation process

2. Tým: struktura data science týmu

Rozbití mýtů – ideální data scientist



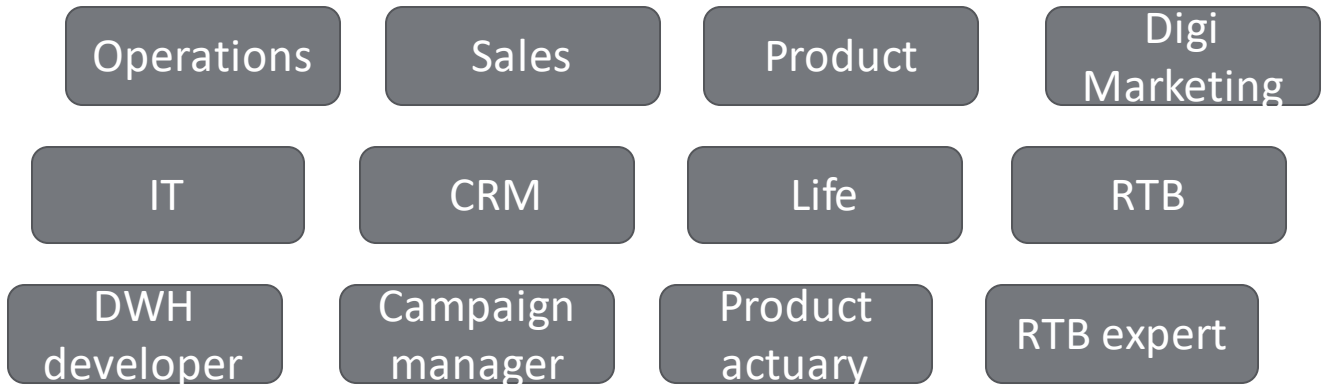
Ideální data scientist:

- Je geniální matematik
- Umí perfektně programovat a zná všechny nejnovější technologie
- Chápe do detailu byznysové fungování firmy
- Umí výborně komunikovat a prezentovat

NEEXISTUJE! (prakticky)

Rozbití mýtů – tradiční struktura

Příklad situace: *Digitální kampaň mířící na potenciální klienty podobní současným klientům s vysokou hodnotou*



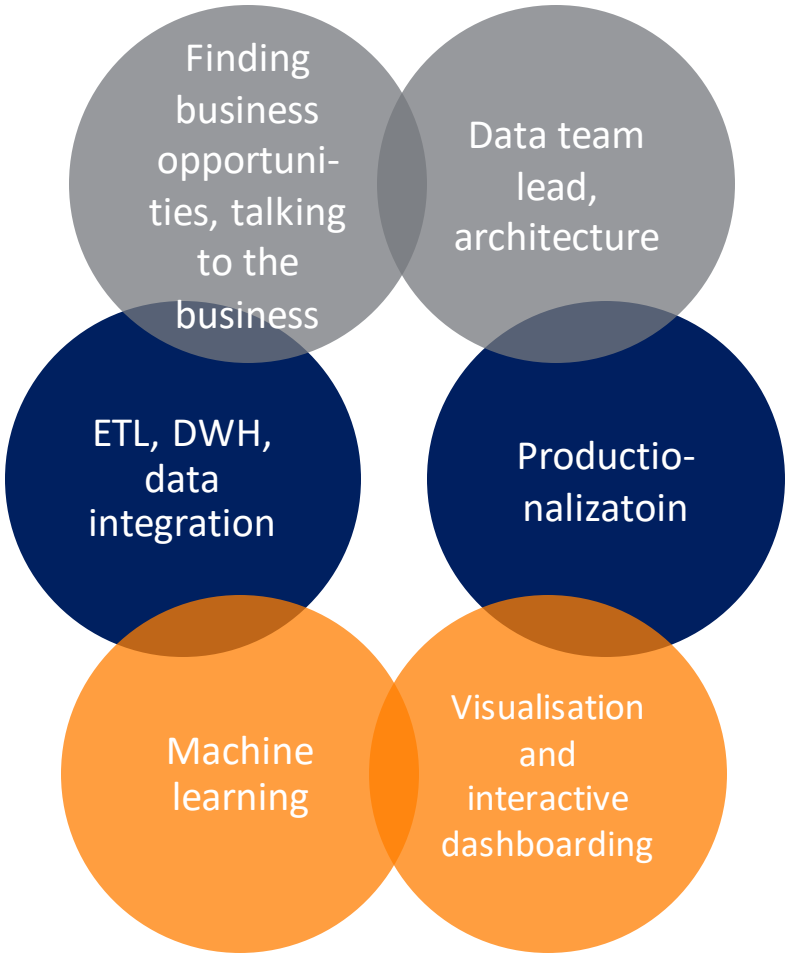
Ve firmě existujou experti, kteří by to byli schopní udělat

Málokdy se jim to podaří, protože každý podlehá jiným šéfům s jinými KPI a budgety - **sila**

Když po sobě něco chtějí musí si předem dát jasné zadání, obhájit byznys case, sehnat budget -> to se vyplatí jen u velkých projektů s jistým výsledkem

Takto ale nejde dělat agilní expertimenty s nejistým postupem a výsledkem

Co s tím? Data Strike Team koncept



Business architect / use case designer

Data Engineer

Data Scientist

Kombinovaný (cross-funkční) tým

- Nemusí být unicorn – lze pokrýt týmem
- Všechny klíčové kompetence v jednom týmu – sedí vedle sebe a konstantě kolaborují
- Takovíto tým může většinu témat posouvat vpřed nezávisle na součinnosti s jinými týmy (kapacity IT/DWH, …)

Klíčový je vztah k IT/DWH/data týmu

- Centrální IT tým pouze poskytuje platformu (sadu technologií a workflow) a definuje mantinely
- Implementaci samotného obsahu (konkrétní use case, projekt, model) už si ale zajišťuje data science tým sám, protože má silné technické kompetence
- Centrální tým jen pomáhá s komplexní produkcionalizací/standardizací a podporuje přepoužitelnost napříč týmy

3. Technologie: jak mohou pomoct

Na jakých technologických konceptech stavět agilní data science

DataSentics

Uživatelské rozhraní

Infrastruktura na pozadí

Přístup a podpůrné nástroje

1. Open-source datové platformy

python

Spark

2. Kolaborativní prostředí pro data science exploraci a prototypování

databricks jupyter

3. Vizualizace dat

Qlik Sense

+ableau

Power BI ...

4. Data lake přístup a paralelizace

hadoop Spark

5. Cloud a PaaS

Azure aws

6. Data engineering a automatizace

Airflow

7. Verzování a CI/CD

git GitHub

mlflow

1. Advantages of open-source data science tools

DataSentics

Traditional tools in actuarial space

Excel

Great for fast and simple things
Crazy for more complex and production things

SAS

More powerful analytics
Not so cheap, not open, limitations in algoritmhs, data integration, …

Specialized tools
Prophet
TW Radar
Remetrica

Closed / black box
Not that easy to integrate or scale

Open-source tools

- They are free
- Full-fledged programming languages – ideal for automation, data integration and more complex logic/algorithms
- The system is open – you can see exactly what's going on and modify it if needed
- Fast new development – the community is constantly developing new algorithms and methods for you to use – there is a library for everything
- It is more popular (easier to hire) – students at school use it, etc.
- It's code – much easier to version control, collaborate on, continuously integrate/deploy
- Cloud friendly, easier to scale
- You look much more cool!

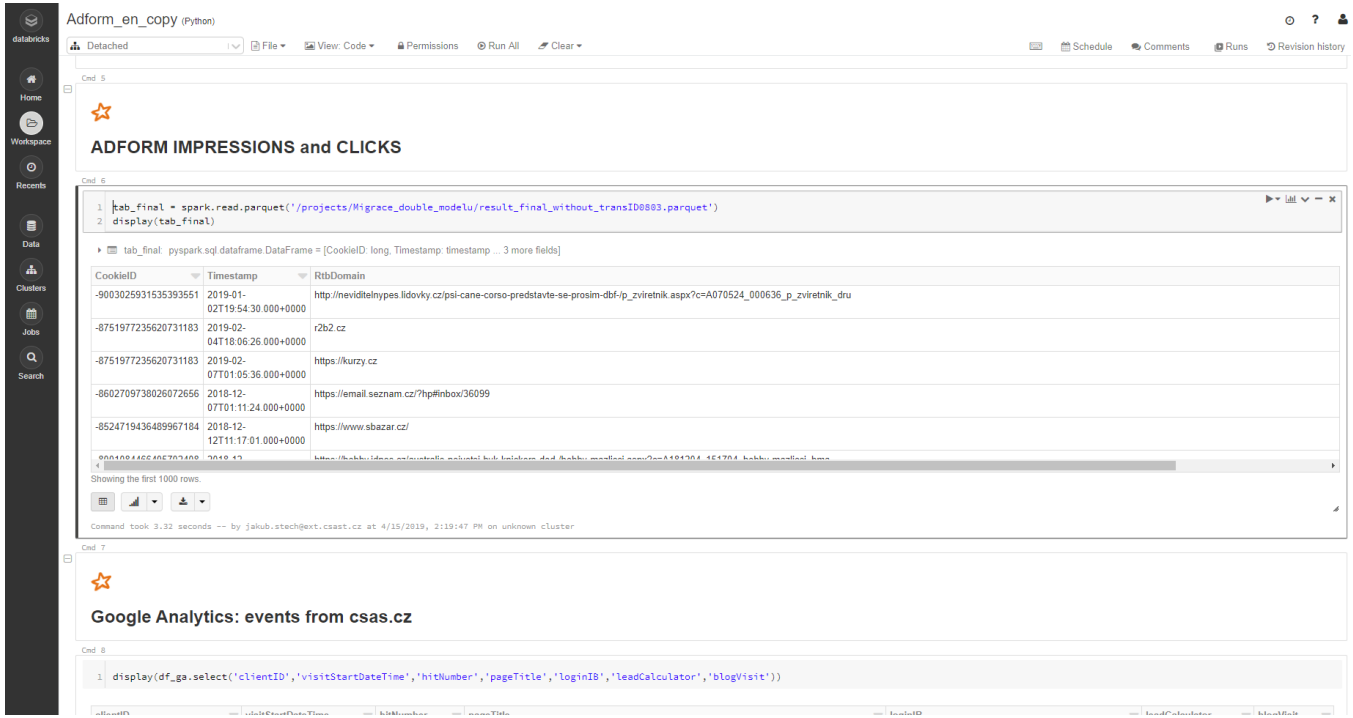
R python Spark

2. TYPICAL COLLABORATIVE DATA SCIENCE ENVIRONMENT – e.g. Databricks

No infrastructure maintenance (managing clusters)

GET STARTED IN SECONDS
Single click to launch your new R/Python/Spark environment

Automatically manage clusters in the background



TEAM COLLABORATION
Projects, common data, work on the same notebook in real-time while tracking changes with detailed revision history

INTERACTIVE EXPLORATION
Explore data using interactive notebooks with support for multiple programming languages including R, Python, Scala, and SQL

Deploy Production Jobs & Workflows
Execute jobs for production pipelines on a specific schedule, …

3. Using modern visualization tools

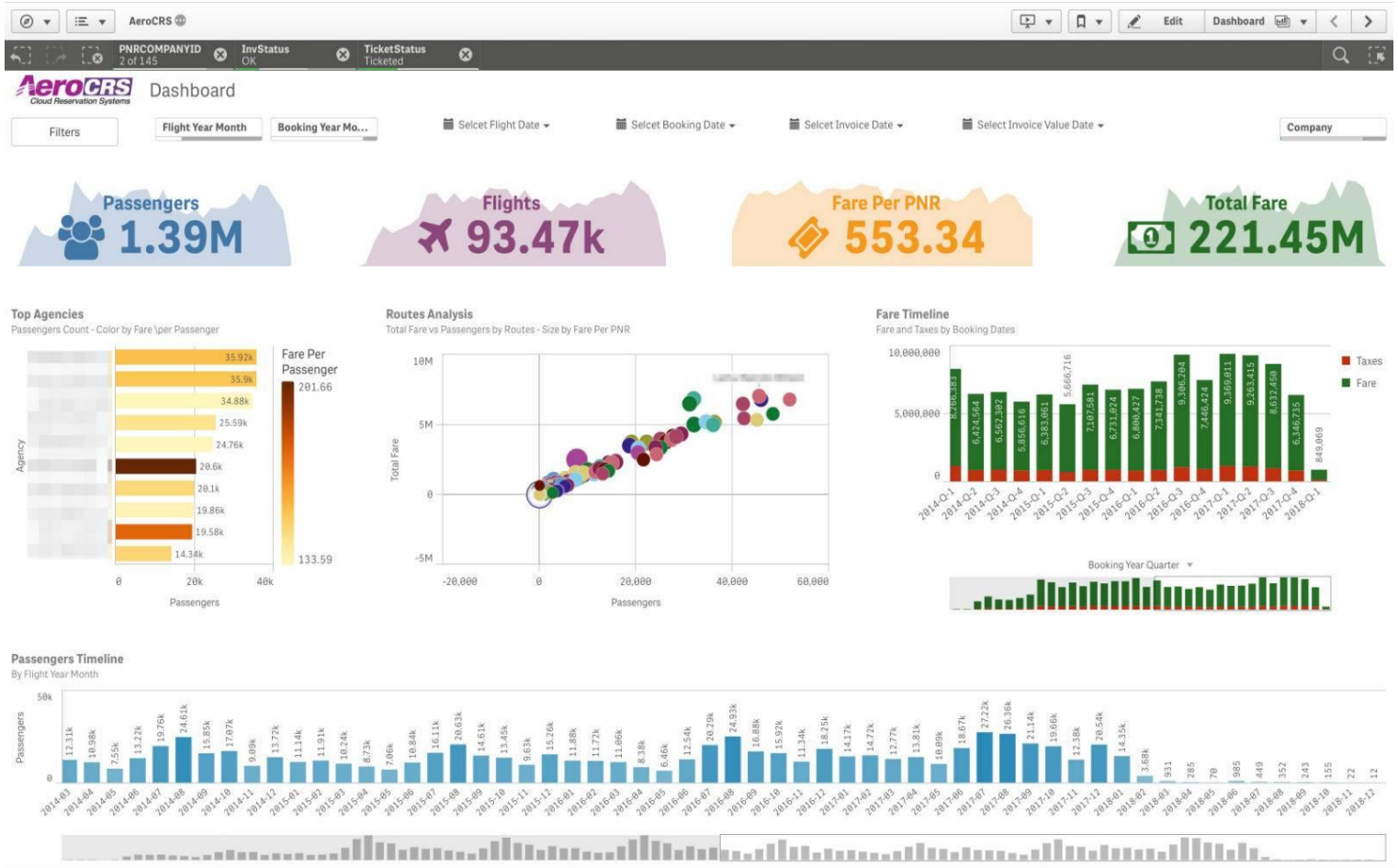
- **Traditional tools do not offer any dynamic and interactive visualization platform:**

Interactive:
If I want to explore just some data / filter them (time window, just one product, just some age etc.), I just click a button and everything is recalculated and redrawn.

Dynamic (some of them):
Visualiziton of the data is live.

Easy to use:
Predefined charts, nice user interface, no need to write any code.

Can handle quite big data, can run in cloud.



4. Data lake approach

Classical way – ETL + DWH:

- Load all data into a database – first step of standardization
- Normalize the data - dimensional data models, etc. – high quality data but slow process and not friendly to changes
- A central team prepares data for business users / limited sandbox capability
- Access to raw data only for selected few
- Enterprise data warehouse – need for always-on expensive infrastructure
- SQL as main language

...

Flat-file approach - more friendly for data science

Ingest new data much faster

Have the full granularity of data -> do your own preparation

Data lake centric approach:

- Just store the data as it is in files (csv, txt, json, …) – storage is cheap
- Perform basic standardization (schema, etc.) using for example Parquet or Avro file formats
- Send only selected data in DWH/database
- Give access to the data to all
- Separation of compute and storage
- Distributed computing -> scalability but also complexity
- Wider use of languages – R, Python, Spark, Java, …

Storage ... Compute ...

Real-time streaming

- Special set of tools for real-time
- Ingestion and processing

...

Orchestration

- Scheduling
- Monitoring
- Right order

...

Downloading external/online data

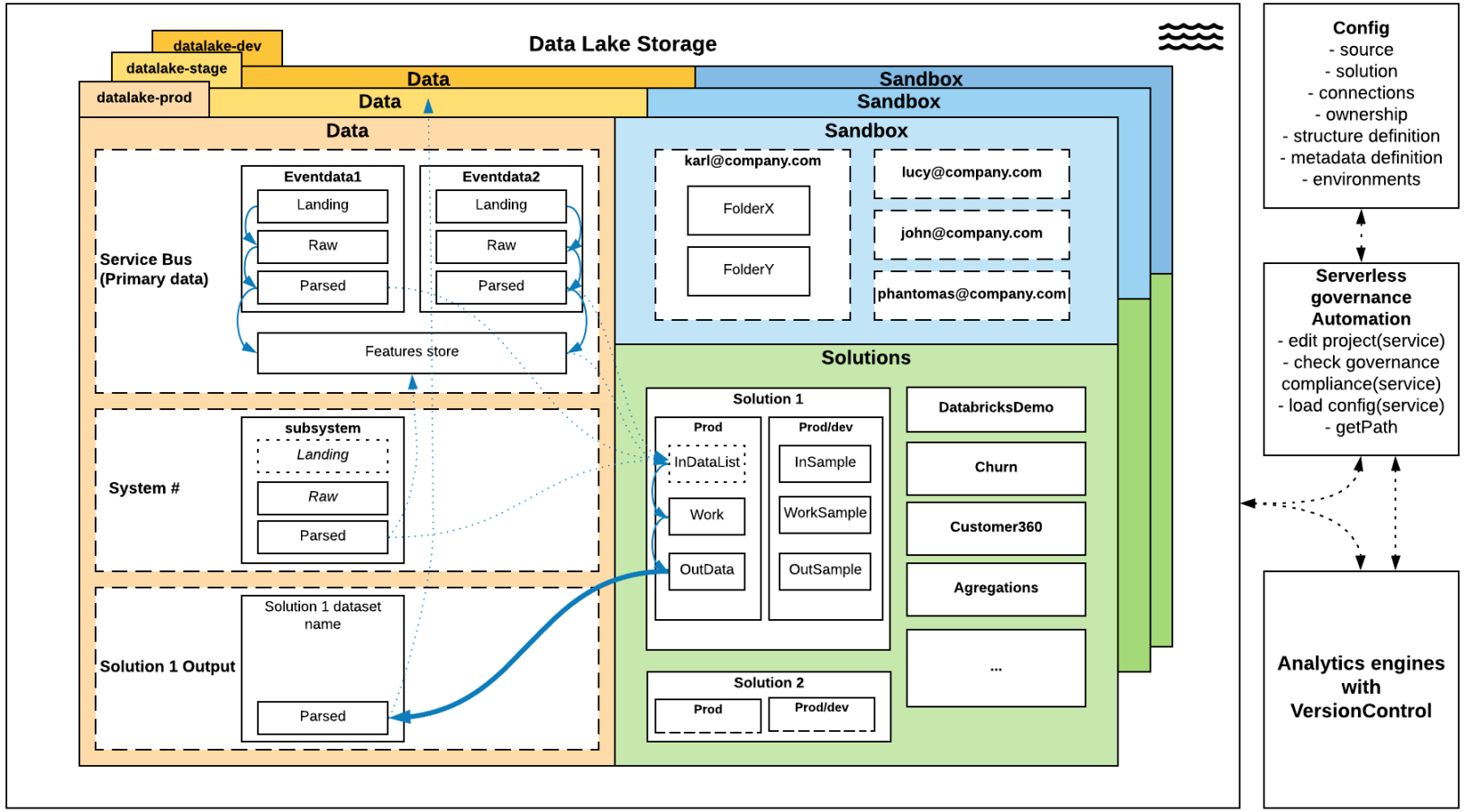
- Calling APIs
- Downloading files
- …

AGILE DATA ANALYTICS PLATFORM: DATALAKE STRUCTURE

DataSentics

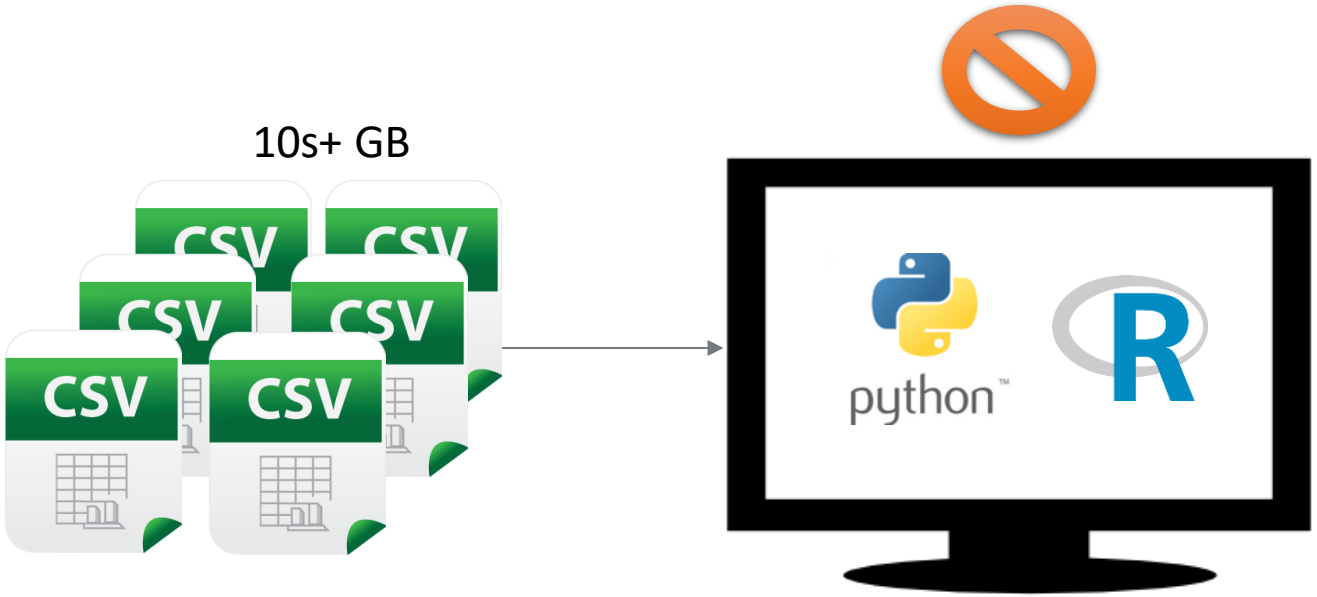
Data Lake structure

- Automatically generated and checked data structure
- Automated CI/CD
- Config library
- Versioning

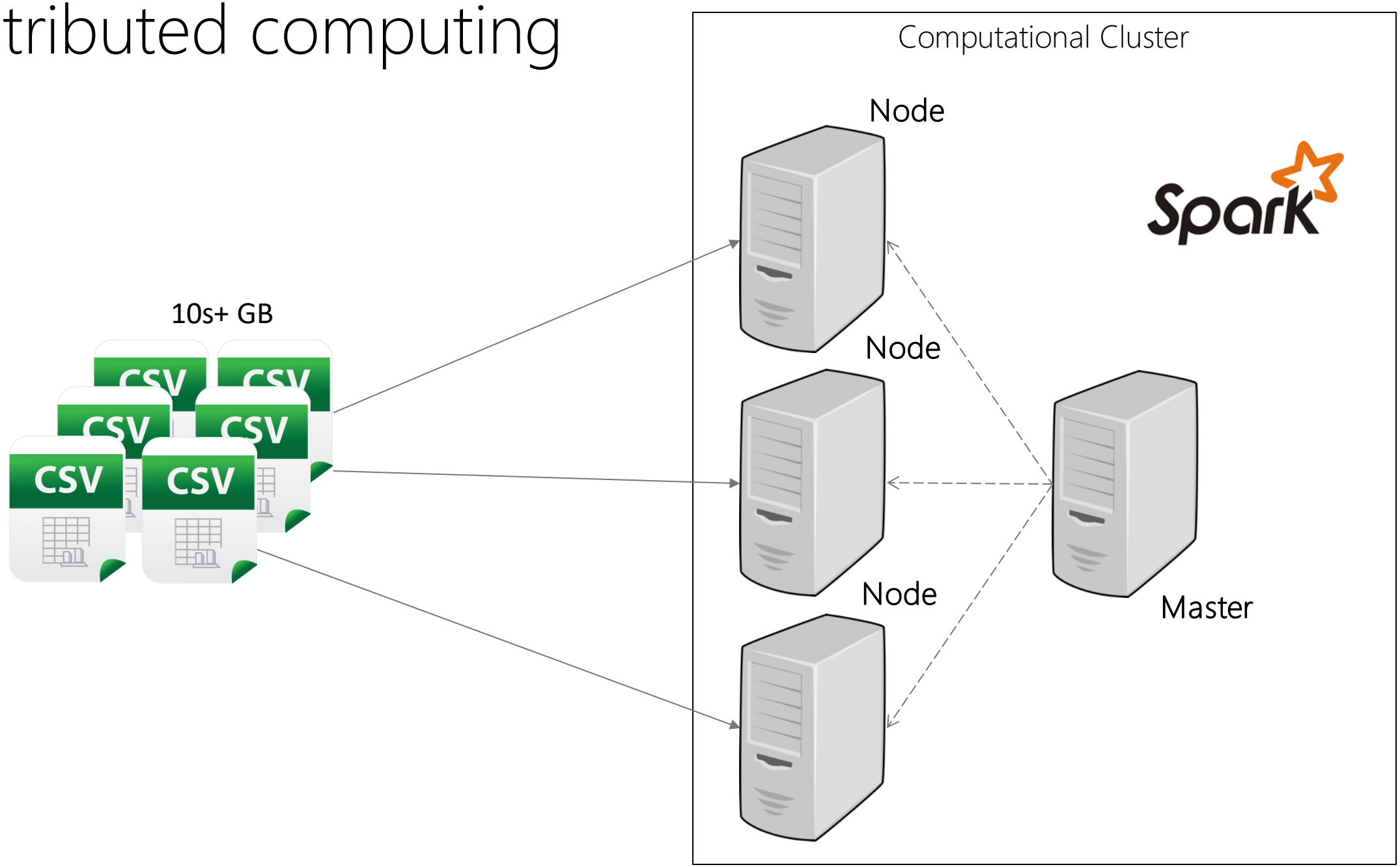


Non-Distributed computing

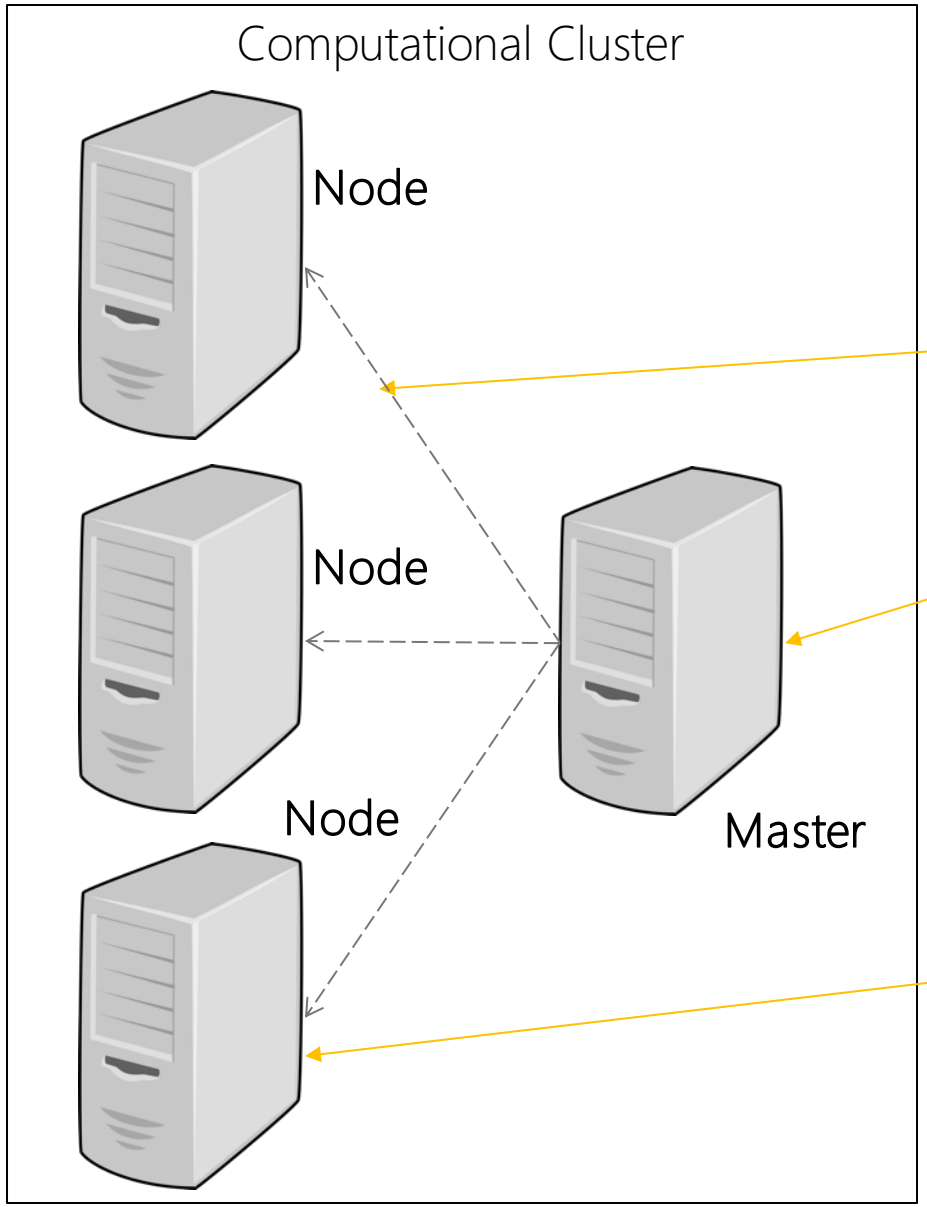
Non-Distributed computing



Distributed computing

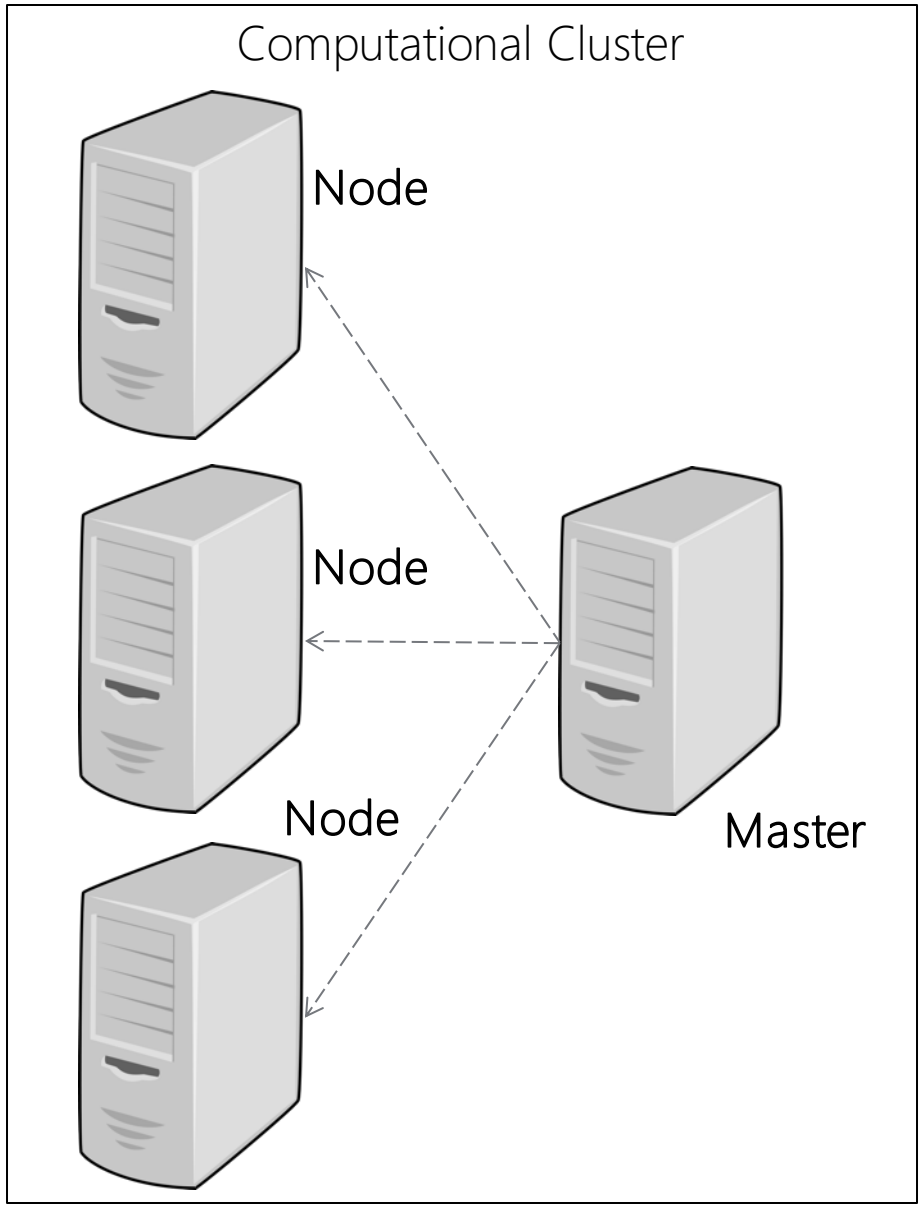


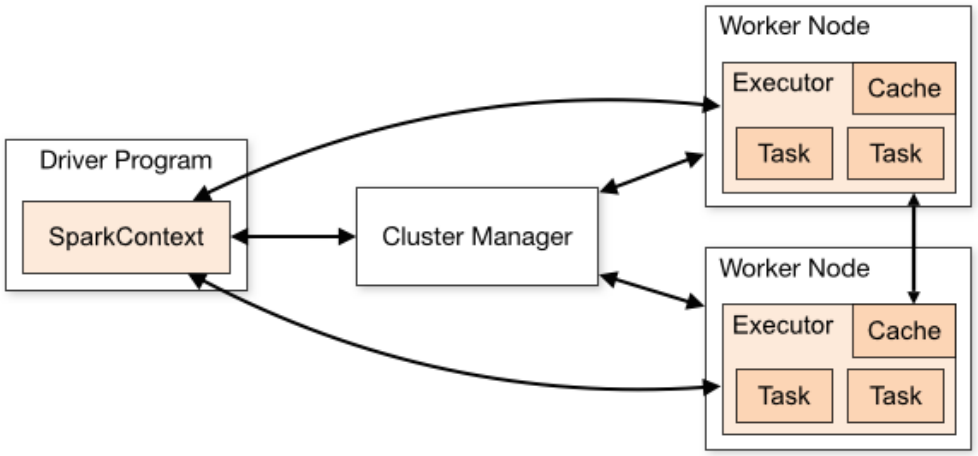
Distributed computing – not an easy task




```
import requests
import csv

ip_addresses = {
‘node1’: 127.0.0.5,
‘node2’: 127.0.0.6,
‘node3’: 127.0.0.7,
‘node4’: 127.0.0.8
…}

# connect to the worker
try:
    r=requests.post(‘..’)
except:
    …

# read csv
with file(‘csv1’) open as f:
…

# calculate sum
len(csv)
…
```

Distributed computing –> use framework






```
df = spark.read.csv('test.csv)
df.count()
```

Open-source framework wars

Data Engineering | *Data Science*

Data Manipulation, ETL

Machine Learning

Streaming

Graph Operations

Open-source framework wars

Data Engineering | *Data Science*

Data Manipulation, ETL

Machine Learning

Streaming

Graph Operations

Unified & unifying Spark

Data Engineer

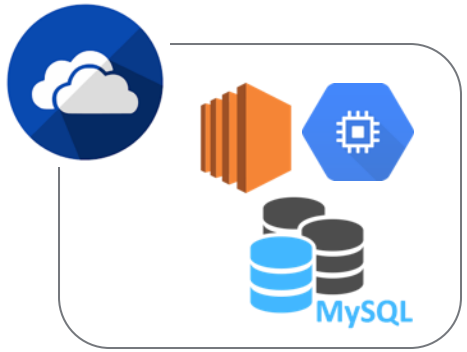
- ETL
- Data Integration
- Deployment of models
- Scala, Java, Python, SQL

Spark

Data Scientist

- Data Analyses
- ML models
- Notebooks
- Python, R

5. Where to Build This – On-Premise vs Cloud

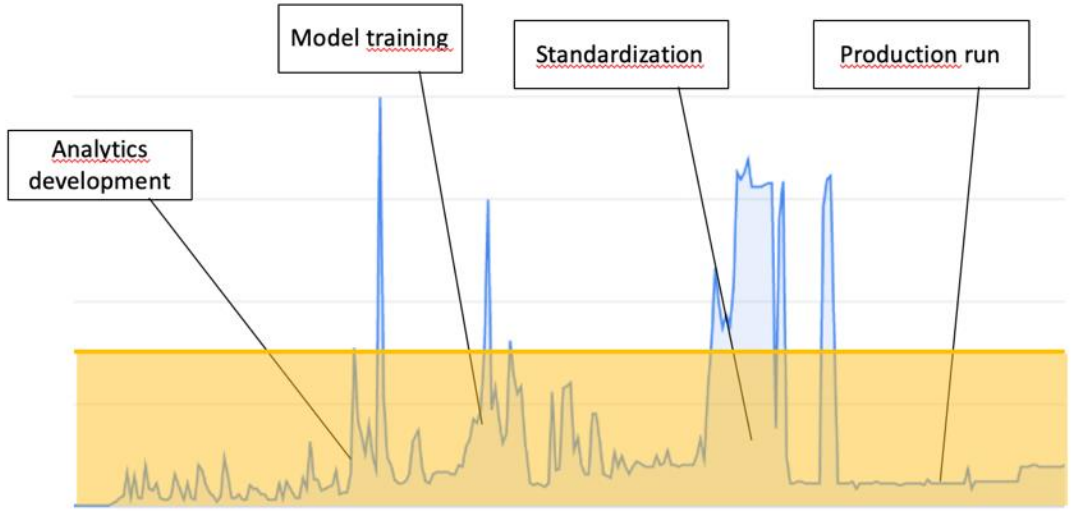


Feature	On-Premise	Cloud VM	Cloud Serverless
Control over underlying hardware	Yes	No	No
Control over underlying software	Yes	Yes	No
Software (typically)	Proprietary / open-source	Open-source / proprietary	Cloud services
Maintenence requirements	High	Medium	**Almost none**
Scaling	Difficult (+costly)	Medium	**Seamless**
Cost (depends)	Usually high	Medium	**Low**

Why the Public Cloud Serverless is cost efficient

DataSentics

- Data science / data analytics workloads are typically **very irregular and spread in time**
  - When I am training a deep learning network I need a lot of performance
  - But I only do this for example once a month/week
  - Similarly with recalculation of the DWH/datamarts/data cleansing/unification/etc.
  - A lot more reports are generated towards the end of the month / end of the year
  - A lot more performance needed during the day than during the night, etc.
- **On-prem dilemma** of whether to buy a lot of machines to cover peaks but then have them standing idle or to sacrifice peak performance to lower idle time
- In cloud serverless there is no such dilemma – you just pay for the actual needed performance at the given time, adapts automatically at a seconds notice

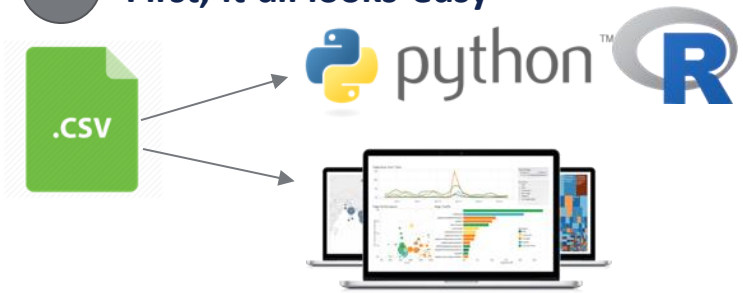


Public cloud – pay for actual computing you used (blue)

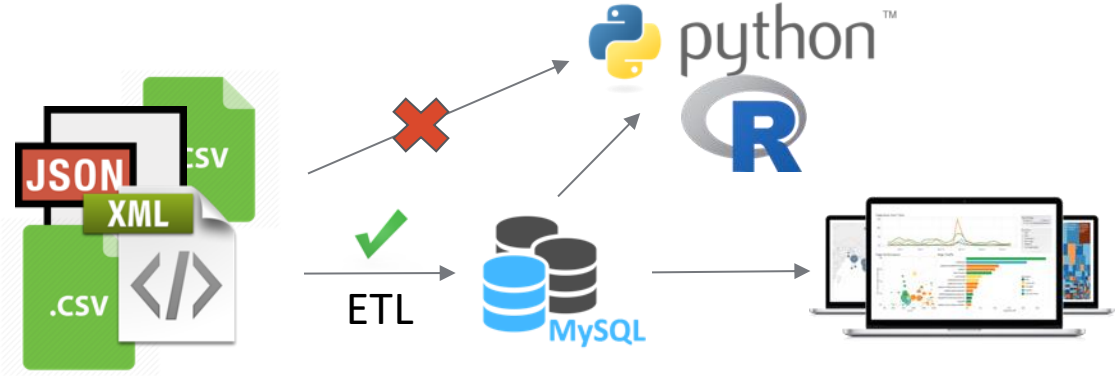
Private cloud – the majority of the time you have more than you need and in the peaks (typically the most important times) you have less power than you need. Applies to both hardware and software (typically node based licencing).

6. Data engineering – why?

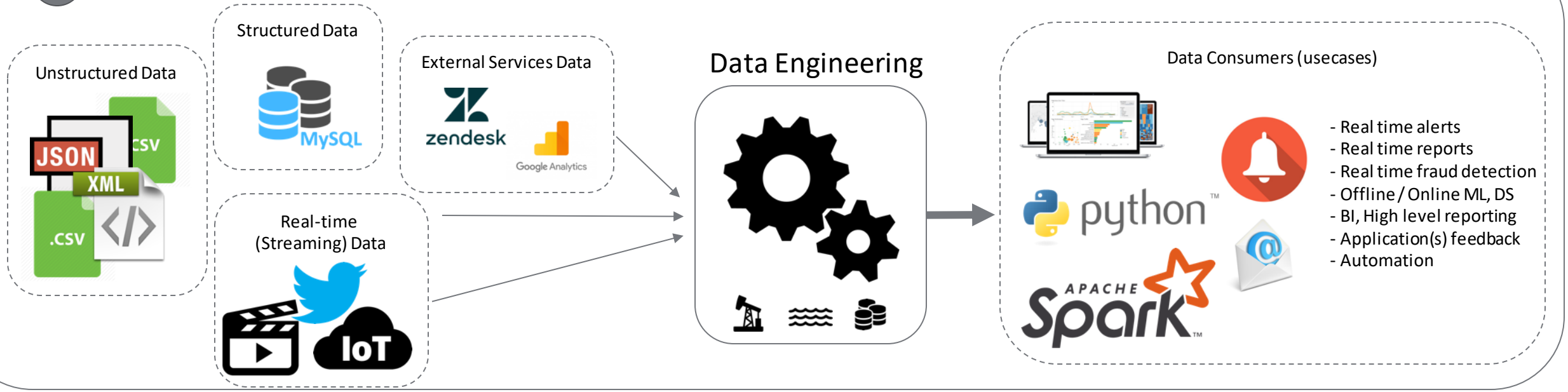
1 **First, it all looks easy**



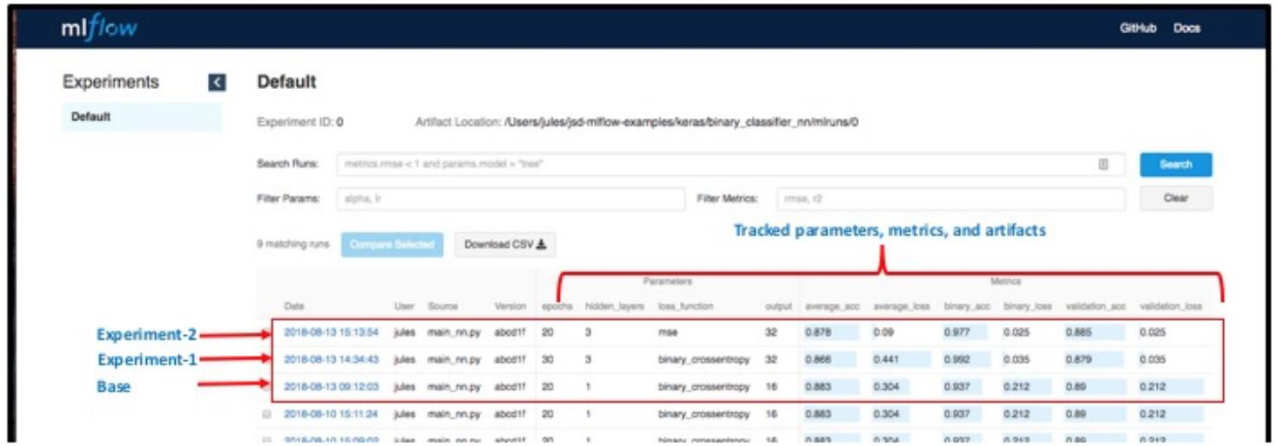
2 **But then.. more data from various sources..**

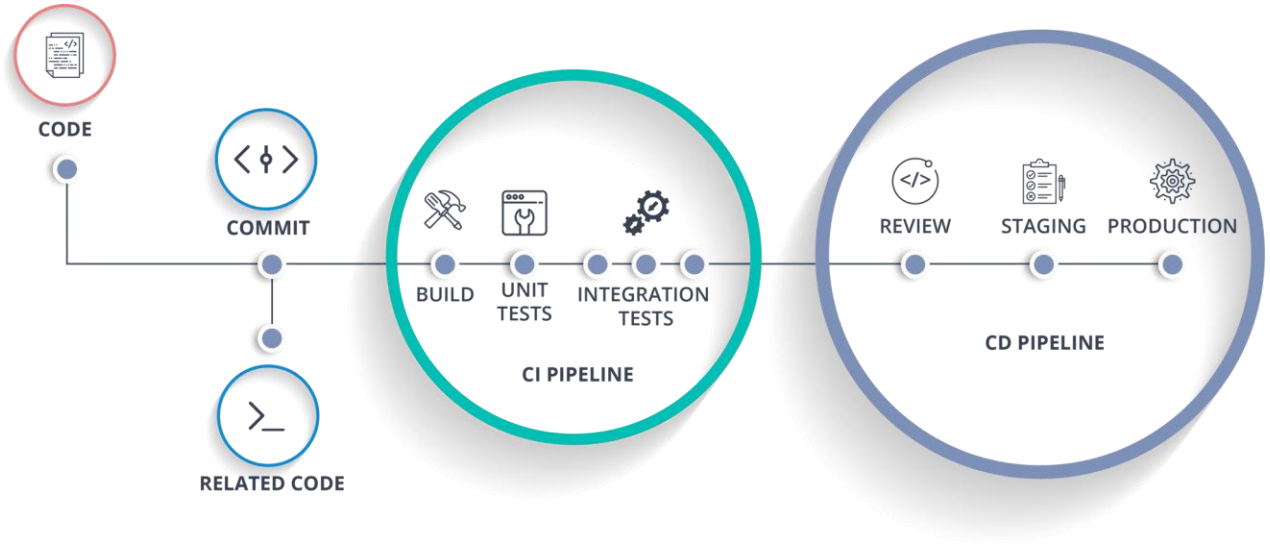


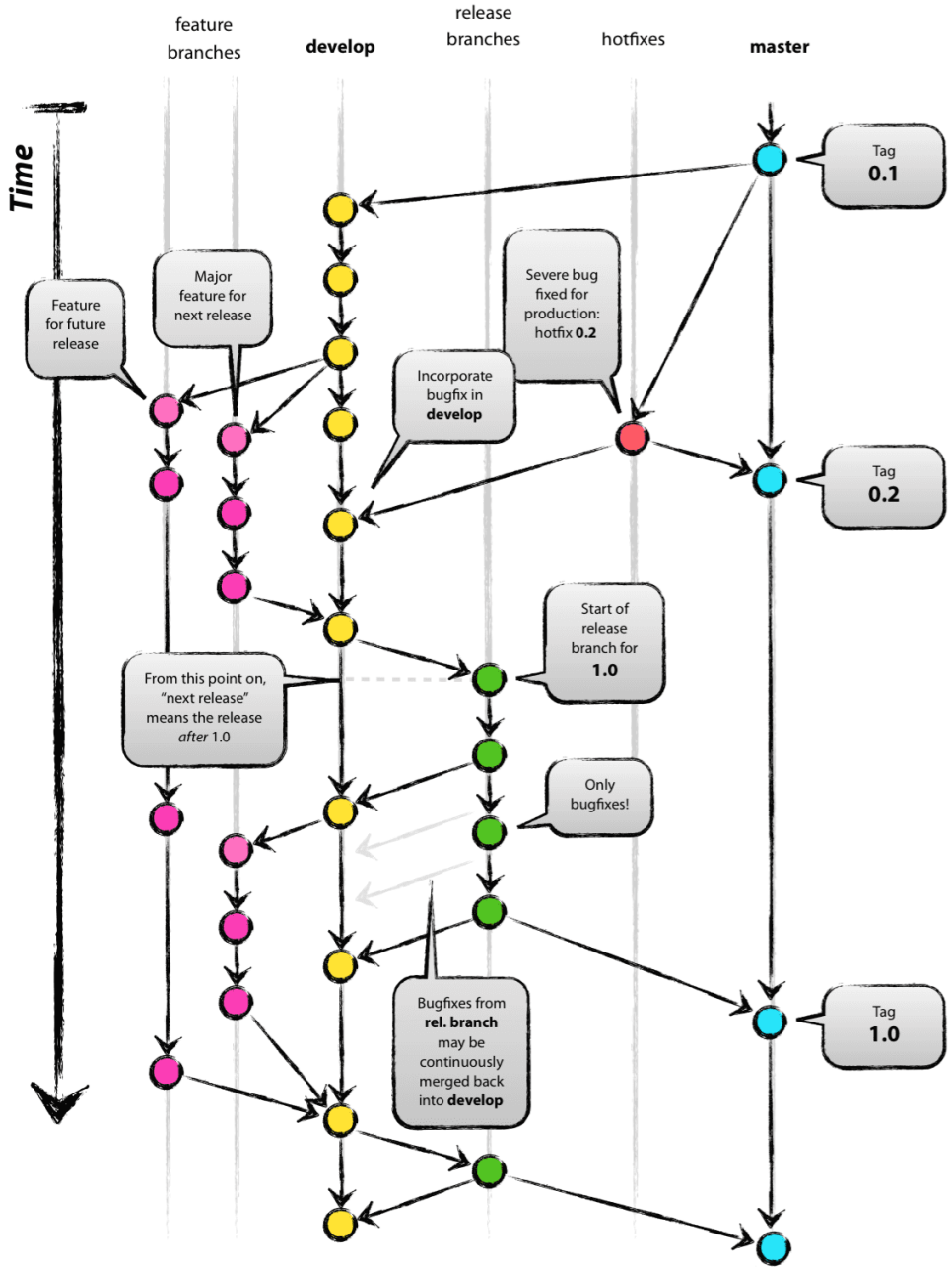
3 **And then...**



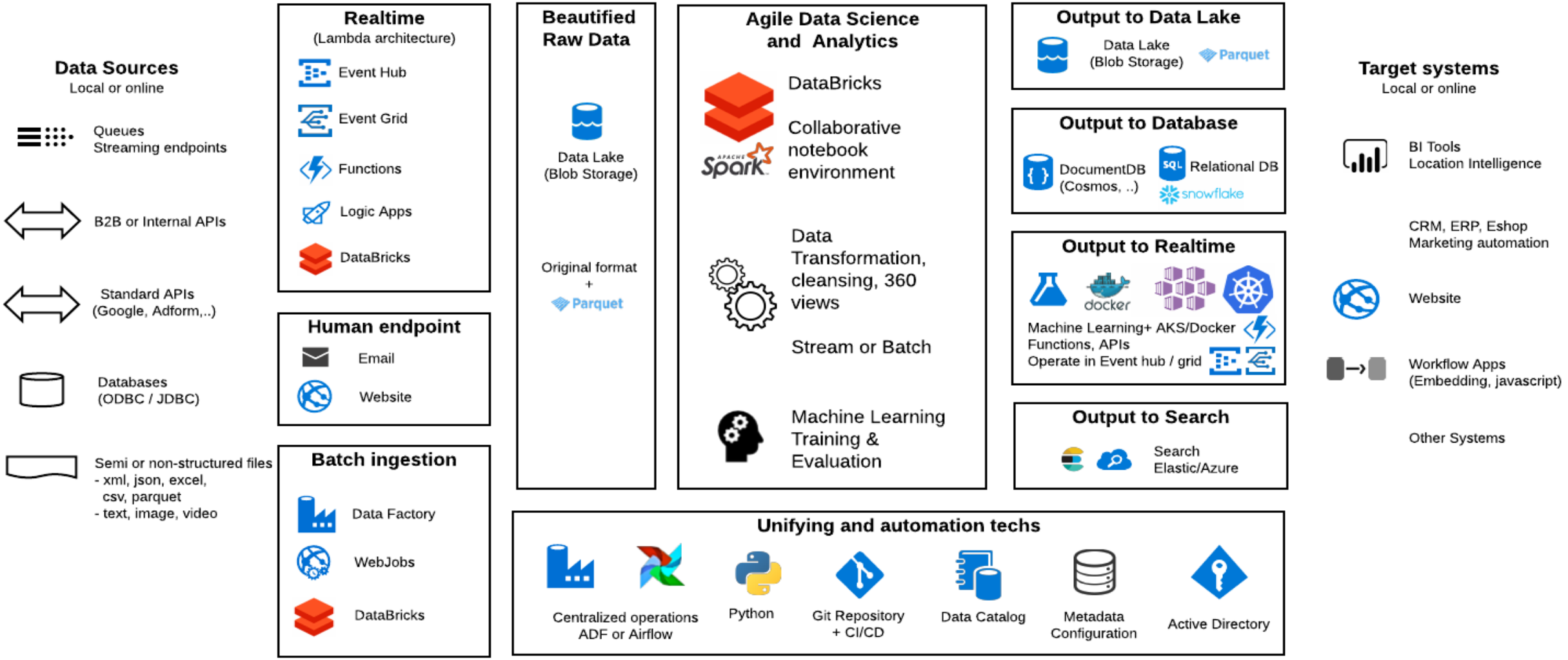
7. Versioning and CI/CD







Microsoft Azure Agile Cloud Analytics by DataSentics



Similarly also on AWS

4. Data science metody: využití v aktuárských problémech

Příležitosti

O technologiích / přístupu

- Python/R jako silnější a udržitelnější řešení než Excel
- Příprava dat vlastním a agilnějším způsobem, ale pořád s dobrou štábní kulturou a úrovní automatizace a robustnosti
- Paralelizace a efektivnější masivní výpočty – stochastické ALM modely, apod.
- Použití vizualizace a dashboardů pro monitoring a vyhodnocování výsledků modelů, analýza změny, apod.
- Automatizovanější a udržitelnější procesy, rychlejší schopnost experimentů a změn

Použití pokročilých data science metod

- Příprava personalizovaných předpokladů pro cash-flow modely – úmrtnost, stornovost, náklady, škodovost, …
- Využití cash-flow a CLV modelů pro řízení marketingu / hledání optimální nabídky
- Pricing pomocí pokročilejších metod (random forest, xgboost, …)
- Využití online dat (kotace, prohlížení webu, …) a externích dat (scrapping, apod.) pro získaní informací o poptávce a nabídce v závislosti na vlastnostech klienta/předmětu pojištění -> obohacení rizikového pricingu o tržní pohled / elasticitu, apod.

Back down to earth (Excel)

- **Use case: <u>Assume that you have dataset and you have to prepare it for pricing</u>**
  - You have following data about cell phones:
    The goal is to standardize Brand names and calculate Sum of Brutto Premium for each brand

ID	Brand	Price	Sold	Insured	Brutto	Commission
1	Smasung	12876	01.02.2010	01.02.2010	149	52
2	Samsung	24652	14.08.2010	14.08.2010	259	91
3	huwei	33027	01.06.2010	01.06.2010	259	91
4	Samsung	23449	19.08.2010	19.08.2010	259	91
5	bb	3923	19.11.2010	19.11.2010	99	35
6	apple	20999	09.09.2010	09.09.2010	259	91
7	samsung	3483	12.09.2010	12.09.2010	99	35
8	HTC	20954	26.11.2010	26.11.2010	259	91
9	sams	28704	20.09.2010	20.09.2010	259	91
10	smasung	7151	20.09.2010	20.09.2010	129	45
11	BlackBerry	23790	17.01.2010	17.01.2010	259	91
12	HTC	23183	06.09.2010	06.09.2010	259	91
13	iphone	24543	28.12.2010	28.12.2010	259	91
14	Huawei	21631	30.01.2010	30.01.2010	259	91
15	huawei	30146	29.10.2010	29.10.2010	259	91
16	Smasung	27286	20.04.2010	20.04.2010	259	91
17	Motorola	19081	27.07.2010	27.07.2010	159	56
18	huavei	24294	23.05.2010	23.05.2010	259	91
19	sms	10147	19.11.2010	19.11.2010	149	52
20	Nokia	12821	22.08.2010	22.08.2010	149	52
21	smasung	13426	11.04.2010	11.04.2010	149	52

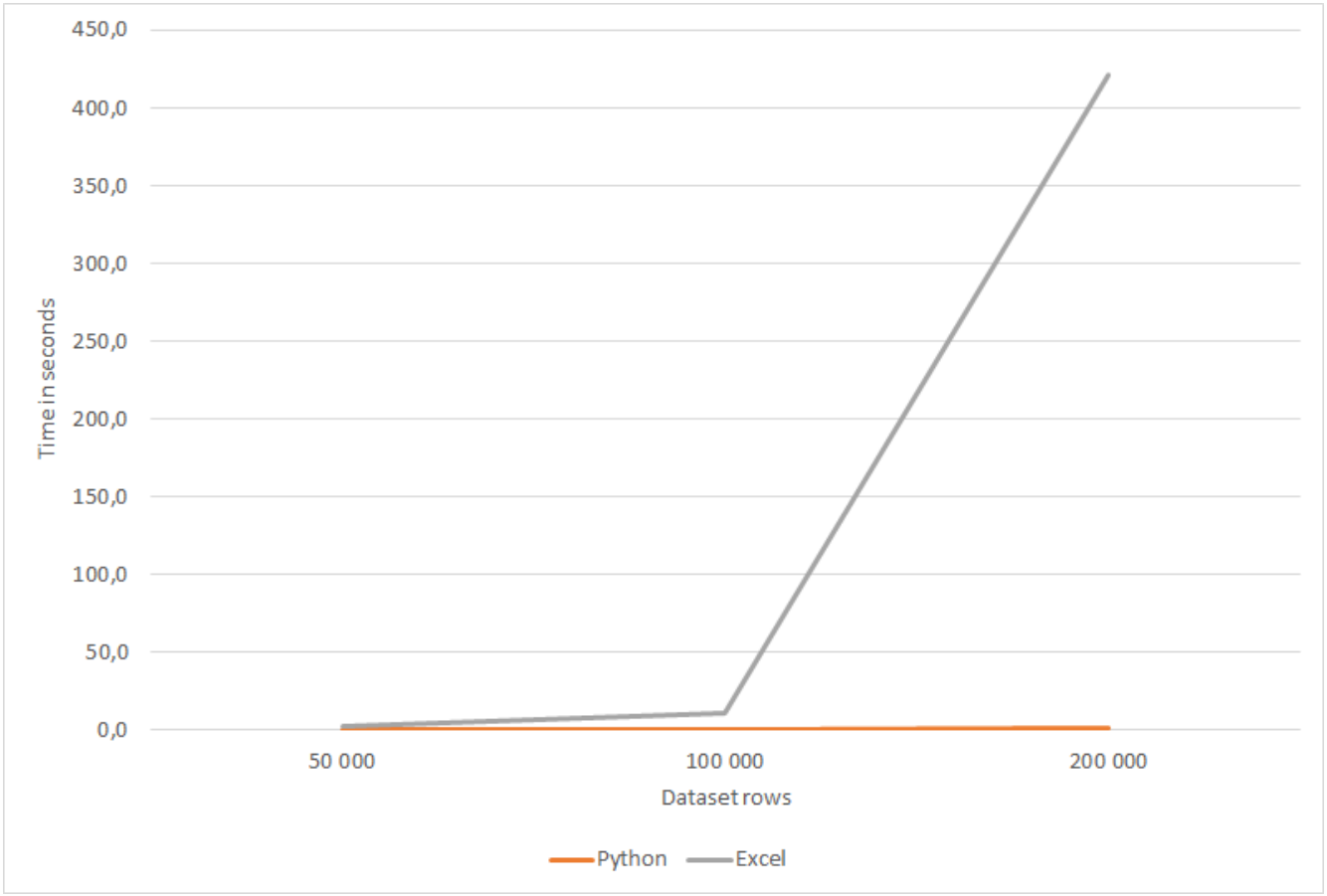
Excel vs Python

- Assume that you have 50.000, 100.000 and 200.000 rows of such dataset
- To standardize the brand names you need to use something like Levenshtein distance
  - In Excel -> there is no package / library to do this. You need to use multiple text functions or write VBA script (60 rows of code)
  - In Python -> there is library to do the trick for you (10 lines)

- To calculate Sum of Brutto per Brand
  - In Excel -> Sumifs functions will do that
  - In Python -> GroupBy and Sum function

- The graph demonstrates efficiency of both approaches:

*(for python it is time including loading data)

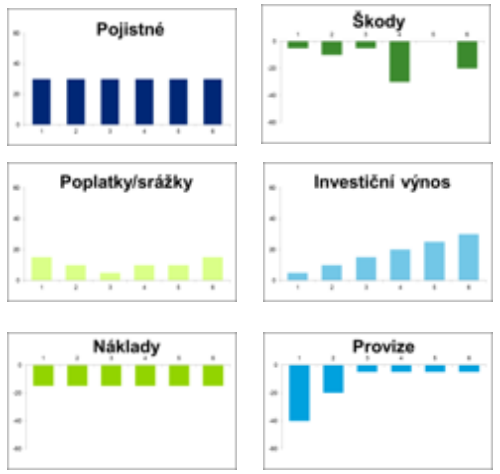


Procesor:	Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz 2.00 GHz
Nainstalovaná paměť	8,00 GB (použitelné: 7,73 GB)

CLV model – life insurance company example

- Each month events can happen: death, claim, policy cancellation, up-sell of new risk, top-up into investment fund
- Inputs: contract details of given customer, age, sex, acquisition distribution channel
- Lapse rate: depends on length of history with insurer and distribution channel, Death: age and sex, Claims: age and sex, number of claims in past, Up-sell: age, sex, region, distribution channel, contract configuration, family

	2017		Further years 2018	2019	2020
Premium	90	Premium in given year	80	70	60
Commissions	-2		-2	-2	-2
Claims	-45		-55	-20	-20
Charges	8		8	8	8
Expenses	-4		-4	-3	-3
Change of fund	-20		-20	-30	-30
Investment income	5		5	7	7
…		Profit in given year			
Profit	**32**		**12**	**30**	**20**

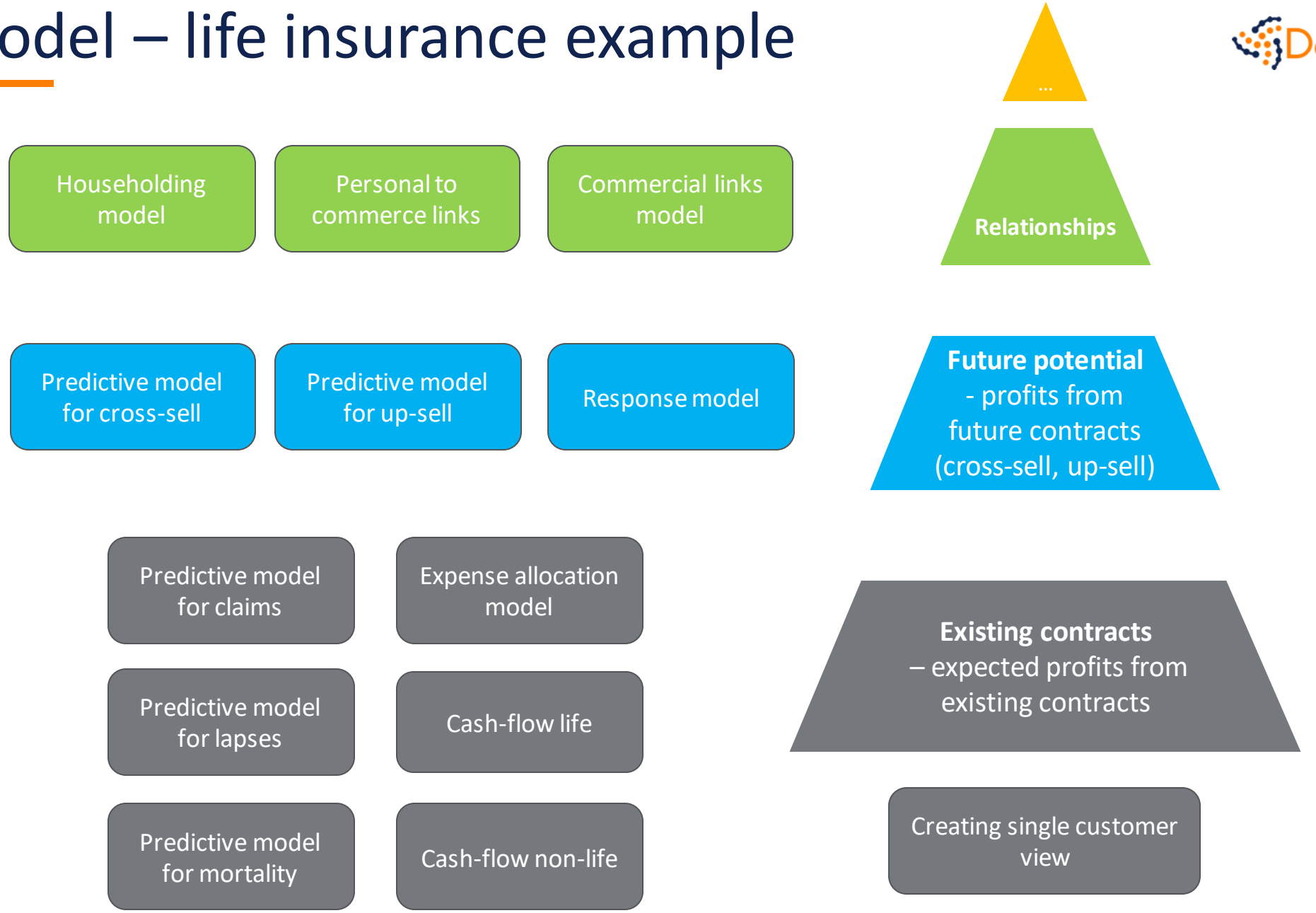


Discounting

CLV = 230

Discounting = taking into account time value of money (interest rates)

CLV model – life insurance example

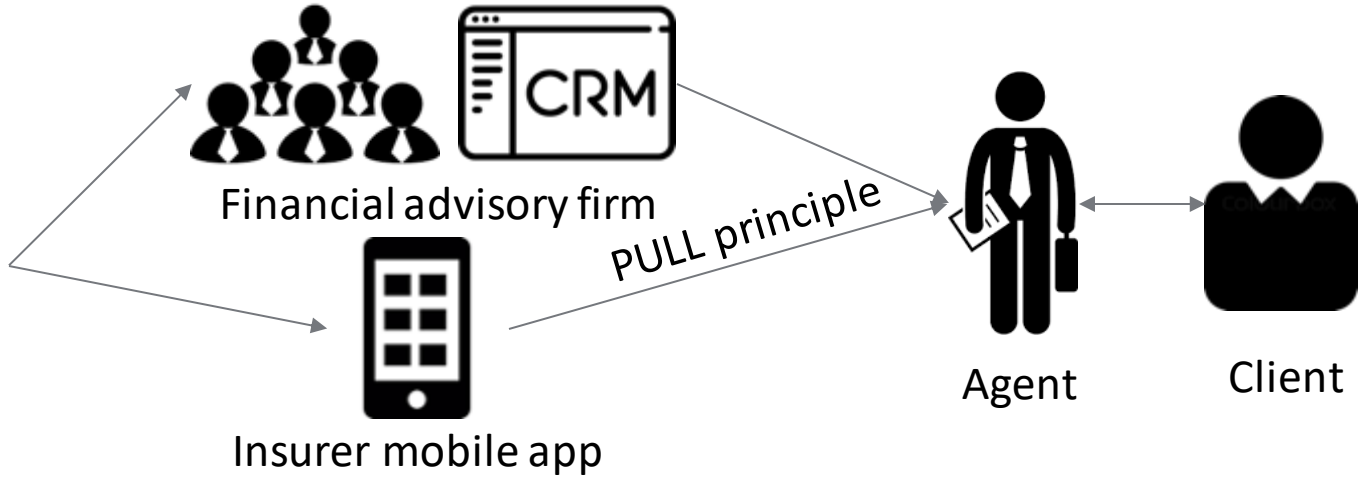


Subscribe/up/cross-sell – life insurance example

- **Use case: <u>Sales signals (opportunities and threats) for physical sales agent network</u>**
  - Life insurance is a **complex product** for solving complex needs (family, health, financial security)
  - **Importance of human touch** - up to this day the role of physical agents is still pivotal to explain to clients how things work and how to setup a contract
  - **Centrally organized push campaigns** based on analytics have **never really worked** in this area – resistance from network
  - **Pull principle is more successful** – provide agents with signals (based on analytics) with which they can work according to their own judgement – sales signal model:
    - Inputs: customer 360 view (contract history, contract events, interaction history, household information, demo and regional info, sales partner info, etc.)
    - Method: predictive model for each opportunity/threat (xgBoost) + clustering to create stories/life situations
    - Output: most relevant stories/life situations for given client

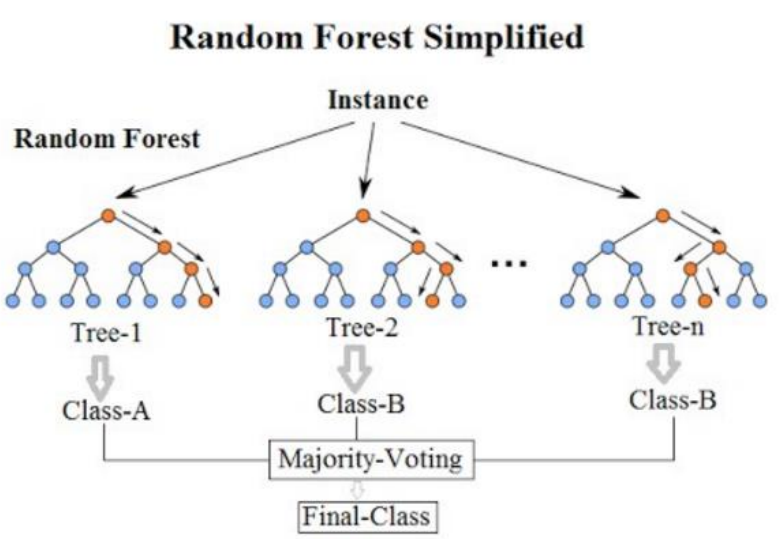
Insurer

- Cross-sell opportunity for serious illness insurance – life situation 1 (founding family, etc.)
- Cross-sell opportunity for disability insurance – life situation 2 (suboptimal insurance policy setup)
- Policy maturity
- Claim
- Changing address
- Marriage
- Risk of lapse
- Risk of non-payment / non-payment issue
- …



Retention process in car insurance

- **Use case: How big discount to offer in order to prevent client from cancelling contracts.**
- Car insurance is usually a fixed-term contract
- Before the insurance end a renewal contract is offered (usually with higher price – the car gets older) and the client can cancel it. If he fulfils certain conditions, he is being offered some discount (it can get up to 30 % in some cases).
- Every year 25% of contracts are cancelled (but - includes also car selling).
- Huge opportunity
- Typically a single decision tree with few criteria is fitted (nr. of claims during some time period, insured amount, type of product, …)
- Potential to use random forests – repeatedly fitted trees.
- Then modus returned as the output.
- Potential to use A/B testing – validation of the models outcomes.
  - Quick results given the amount of contracts.



Automatization of key models

- **Use case: Most of the insurance models are refitted manually on a monthly basis – it takes a lot of time and lot of effort just rerunning existing code.**

- Using platforms such as Databricks allows key models to be run daily/weekly - cloud allows orchestration (running automatically at certain time).

- Because of the bigger computing power, "old" and "new" model can be run parallel, the results as well as the data can be stored.

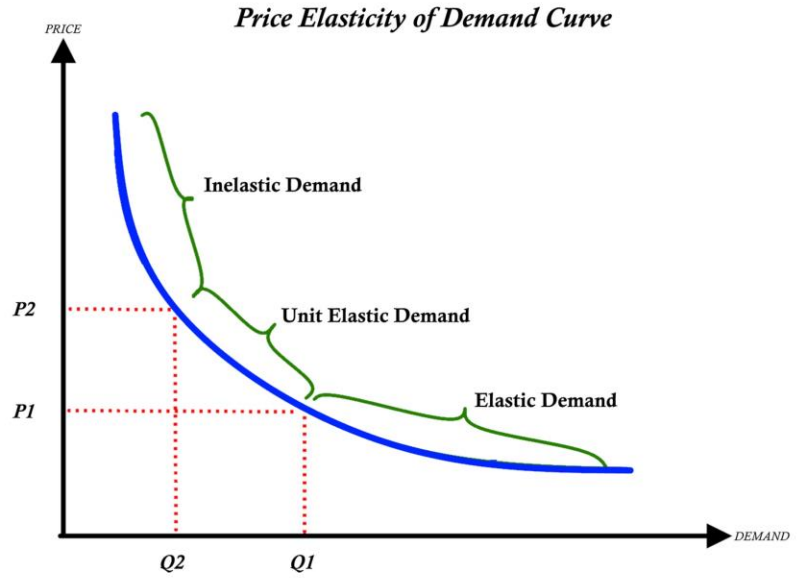
- It is necessary to check the stability of the models.

- Advantages:
  - Human needed only when something goes wrong.
  - Automatic control on the input data – if they do not differ too much
    - Typically reveals changes in DWH etc.
  - Automatic control on the coefficients changes
    - Early warning on the structural changes for example in clients behavior.
  - Automatic control on the model performance metrics (fit or prediction)

Usage of online data – personalized offers, discounts

- **Use case: Online data is the ultimate way of understanding the demand for the insurance product and its relation to the price.**

- What information can we gain from online data:
  - What insurance the client is and is not interested in
    - If we knew he is interested mainly in disability insurance and less in death insurance, we can offer a discount for the death insurance.
  - We can have a model determining clients probability of conversion as a function of the insurance price – we can find optimal discount maximizing profit.
    - This is closely related to the law of demand
    - Price elasticity of demand
    - Instead of Quantity demanded by a population, we can work equivalently with the conversion probability (when we fix population size).
    - A parametric model is already described and explored – we can just fit the parameters.



Interest - Motor insurance example

- Online motor insurance sales -> **leads** (filled data and received quote but did not buy on web – i.e. similar to abandoned cart) -> **call-centre is performing follow-up calls**
- **Propensity-to-buy model**:
  - Inputs: driver (age, sex, region), car (brand, age, value, …), quoted insurance configuration, quote interaction behaviour (modified default packages, etc.), channel
  - Method: predictive model (regression + random forest)
  - Output: probability that the given customer will buy motor insurance after follow-up call
- CLV model:
  - Some types of contracts/drivers/cars are more profitable for the company than other

- **Use case 1**: **Call-centre prioritization** – use propensity-to-buy * CLV as a measure of priority for follow-up calls = call first/most tries/biggest discount/etc. to customers with highest combination of propensity-to-buy and CLV
  - additional problem: probability of a fake number input – neural network returning probability
- **Use case 2**: **Marketing targeting** - Use PtB*CLV as measure of priority for purchasing online advertisement – invest more marketing budget into segments with higher PtB*CLV

Master pricing model – relation to the cancellation

- **Use case: Pricing is usually done using only claim model. But the CLV of a client does not depend only on claims – what about cancellation?**

- For example in life insurance, there are plenty of expenses for the insurance company on the beginning of the contract (administration costs, agent provisions). By unit-link life insurance, there are cancellation fees in the first years to cover them (client has some capital the insurer can take). By term-insurance, there is no such option. The initial costs are somehow transmitted to the tariffs, but even though clients who cancel quickly never pay for the expenses they brought to the insurer.

- As term-insurance is becoming more and more popular, clients features to the odds of cancellation is becoming an important factor that should be taken into account when offering the price for the insurance.

DataSentics

Contact

Petr Bednarik, founder

Washingtonova 17/1599
Prague 1, 110 00
Czech Republic

+420 733 161 533

petr.bednarik@datasentics.com

www.datasentics.com