Machine Learning in Deutsche Telekom

AI

SHARED
SERVICES

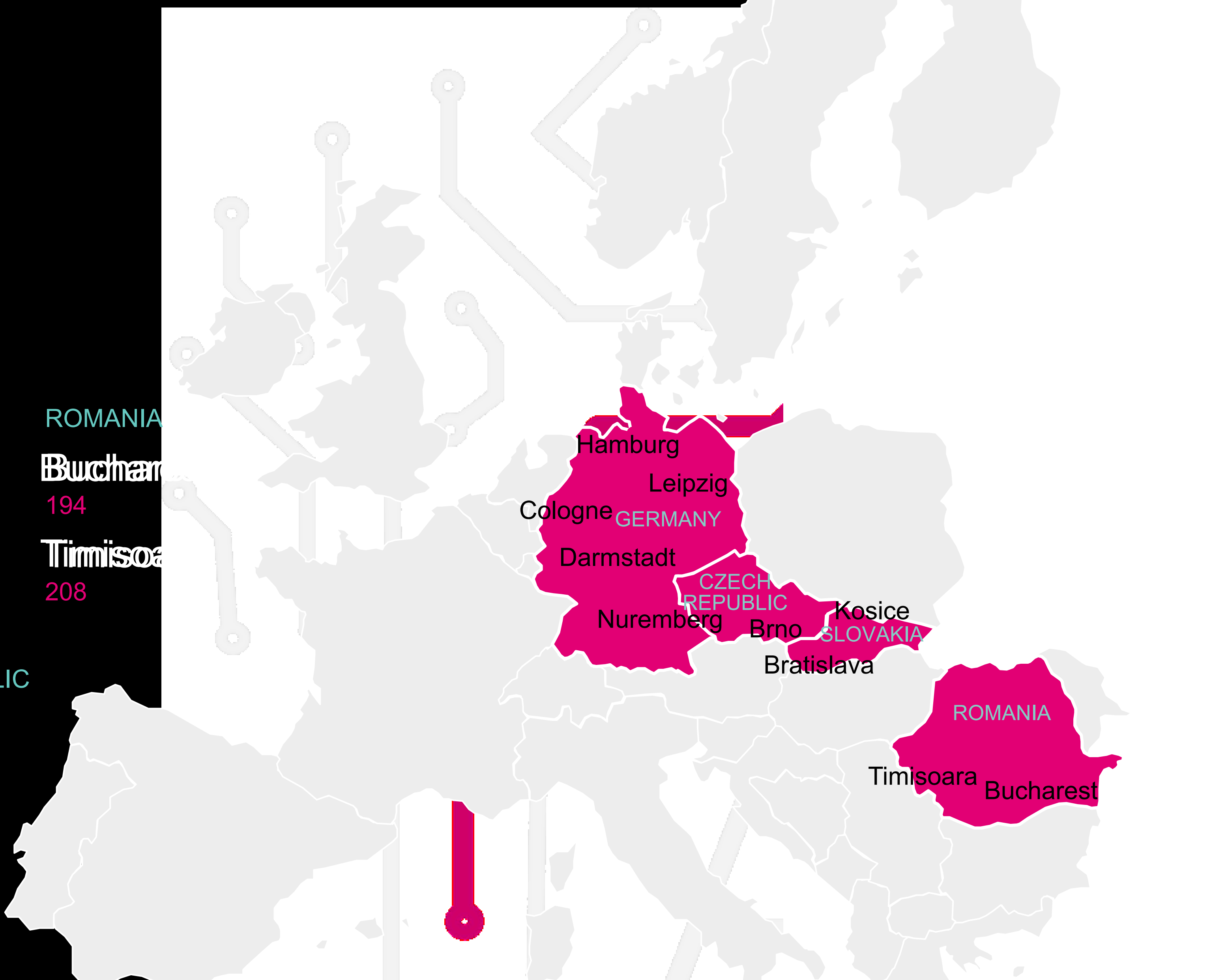
Who we are…

Deutsche Telekom Services Europe

Deutsche Telekom, T-Mobile CZ

DTSE in Europe

GERMANY	SLOVAKIA	ROMANIA
Cologne 1007	**Bratislava** 617	**Buchar** 194
Darmstadt 388	**Kosice** 237	**Timisoa** 208
Hamburg 146		
Leipzig 361	**CZECH REPUBLIC**	
Nuremberg 130	**Brno** 133	



[1] Financial Report 12.2019

AI Shared Services: your partner for the leading digital telco

We embrace AI and agile working to enable data-driven decision making

> 40
INTERNATIONAL TECH TALENTS

> 120
AI COMMUNITY MEMBERS

Our recipe for great AI projects

AGILE WORKING & RAPID DEPLOYMENT

EUROPEAN CROSS-FUNCTIONAL TEAMS

SMALL INVESTMENT & END-TO-END SERVICE

BEST TECHNOLOGY & CONTINOUS IMPROVEMENT

Follow us on LinkedIn

TEXT PROCESSING

PREDICTIVE ANALYTICS

ESG ACCELERATION

AND MUCH MORE…

Jakub Kondek, M.Sc.

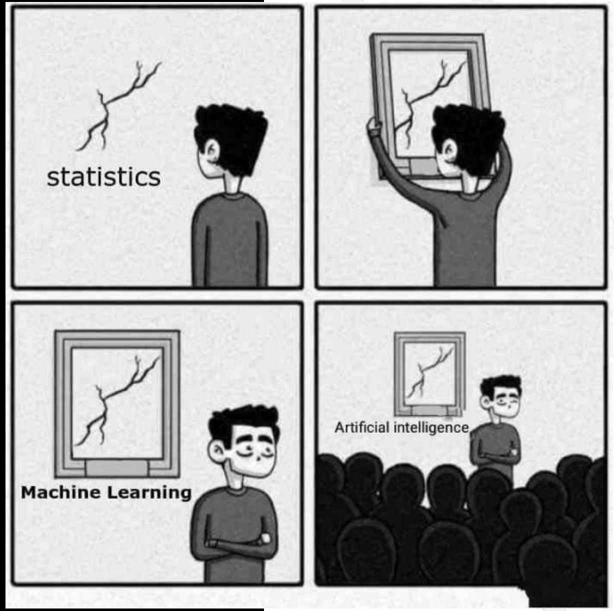
j.kondek@telekom.com

Senior Data Scientist

Turista

Blues enthusiast

stepan.vondracek@telekom.com



https://www.instag
s/

DAILY Challenges of Data Scientist

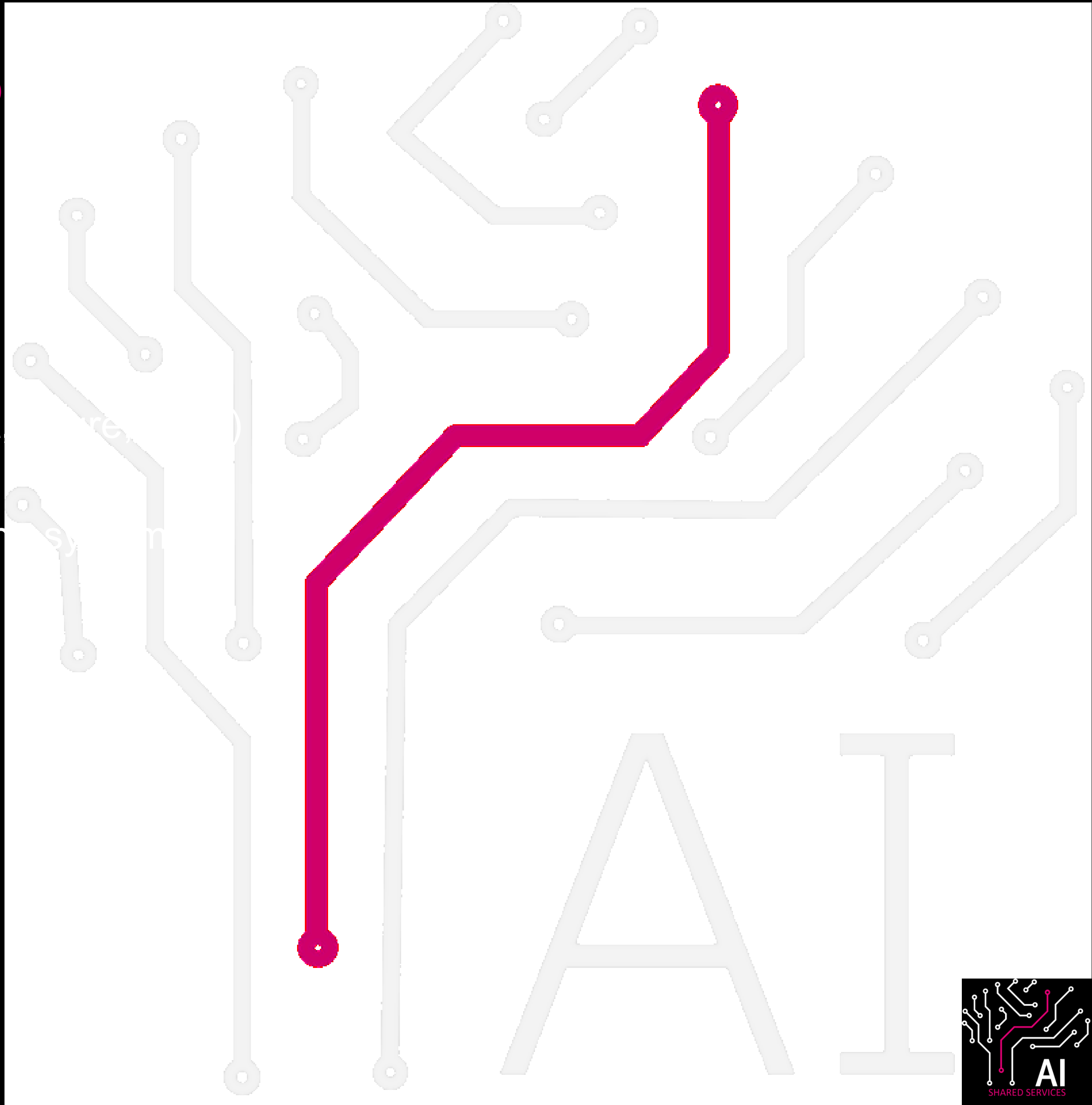
- A customers do not understand what ML does （can do）
- Communication of technological topics
- Unclear requirements *("we would like to impl*
  some AI...")
- Stated problems are much simpler
- System integration
  - Is data which was used for training available
    predicting
  - Do I have access/rights to data
- Can I send prediction somewhere reasonably

***

Possible challenge for statistics/econo
graduate -> mostly IT terminology

What We Use in AIS

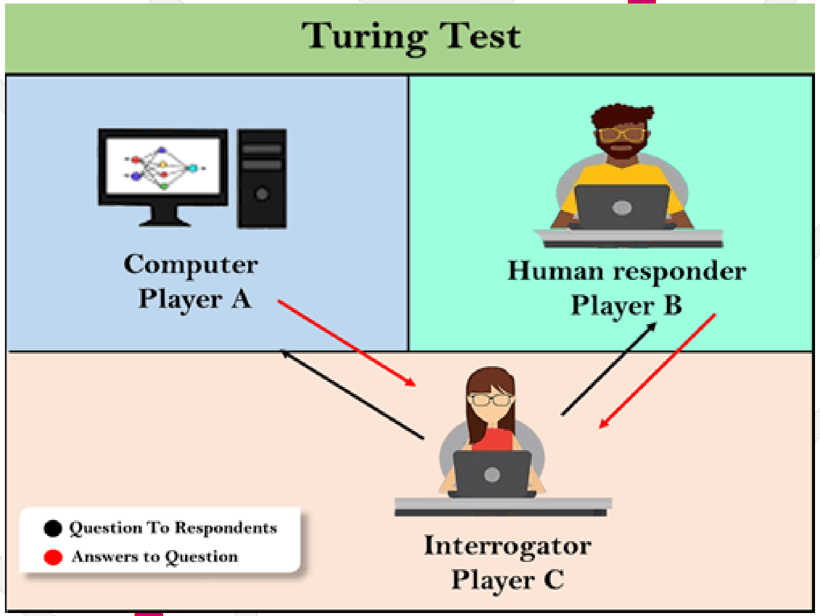
- Python，SQL
- Python IDEs （Pycharm，Visual Code）
- Dedicated ML server + Cloud (Openshift, …
- containerization (Docker, Kubernetes)
- REST API for communication between th…
- Webservice Frameworks (fastapi)



Part I. - Natural Language Processing

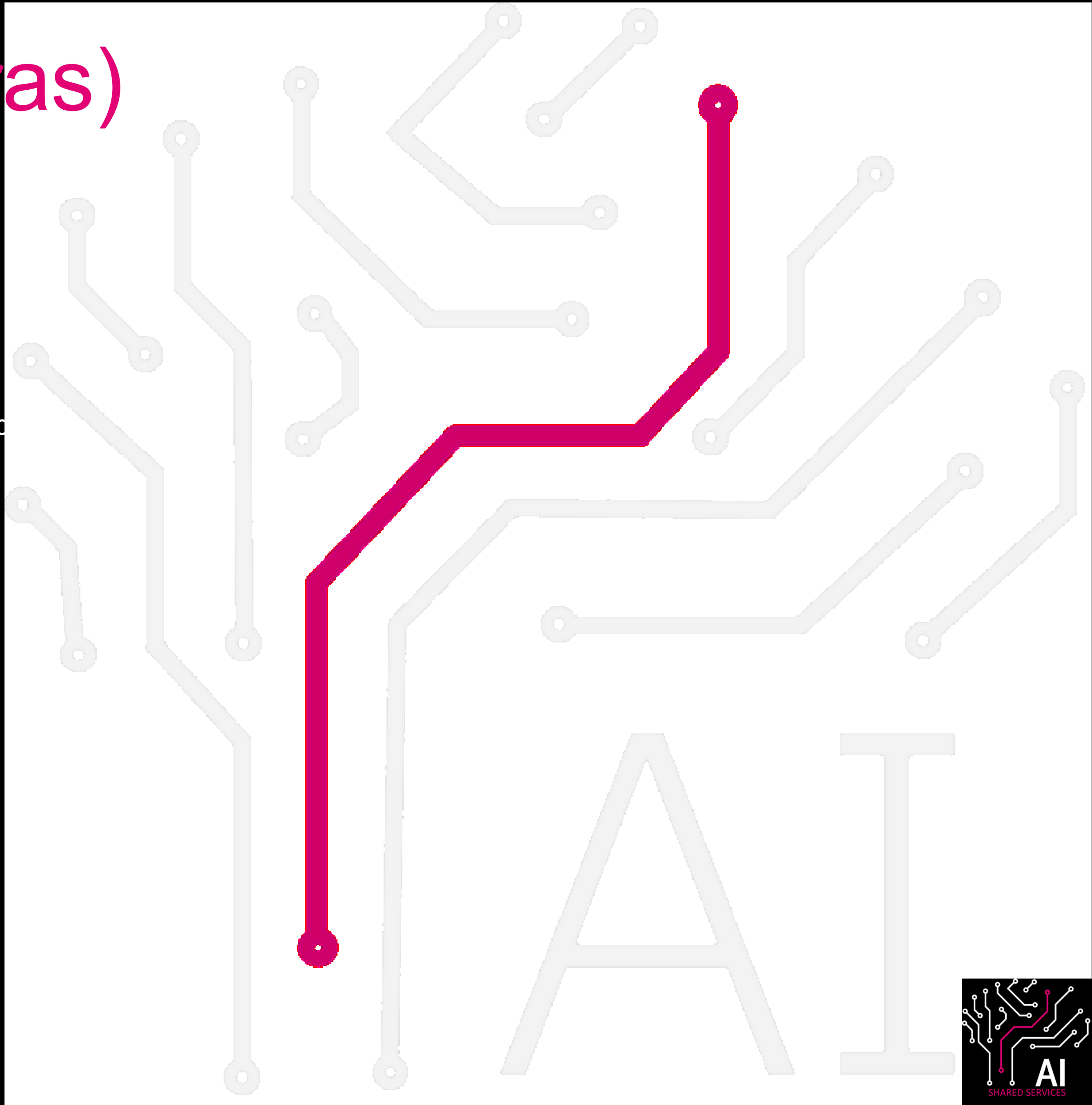
NATURAL Language Processing (NLP)

- A field of AI dealing with interaction between computers and humans using the natural language
- Started in 1950s (Turing test)
- Considered to be a difficult problem in computer science
- Research in NLP is still going on
- 3 main NLP subfields:
  - Speech Recognition
  - Natural Language Understanding
  - Natural Language Generation



NLP Model Types (eras)

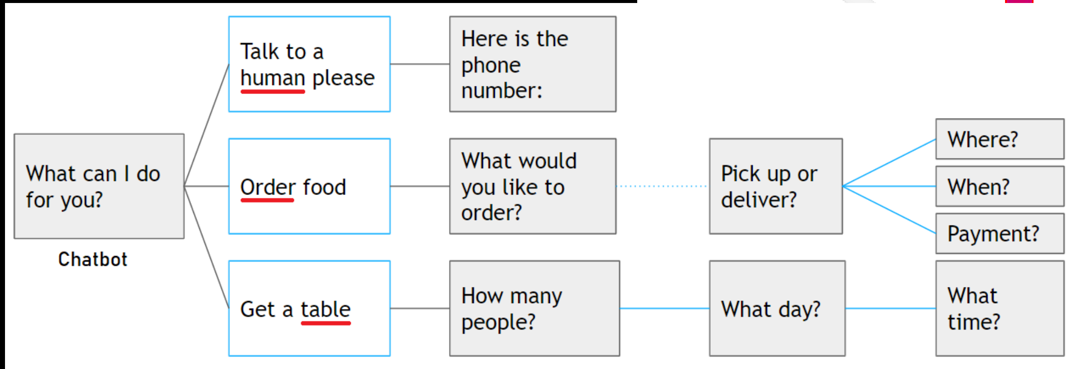
- Rule-based
- Statistical – the “traditional approach”
- Deep learning – the state of the art, ”modern” appro



Rule-Based Modeling

- A hand-crafted system of grammar rules based on lin
  - regular expressions, context-free grammars
  - often requires a skilled expert – a linguist
- Useful when we don't have enough data
- Very good interpretability
- Poor generalization and maintenance

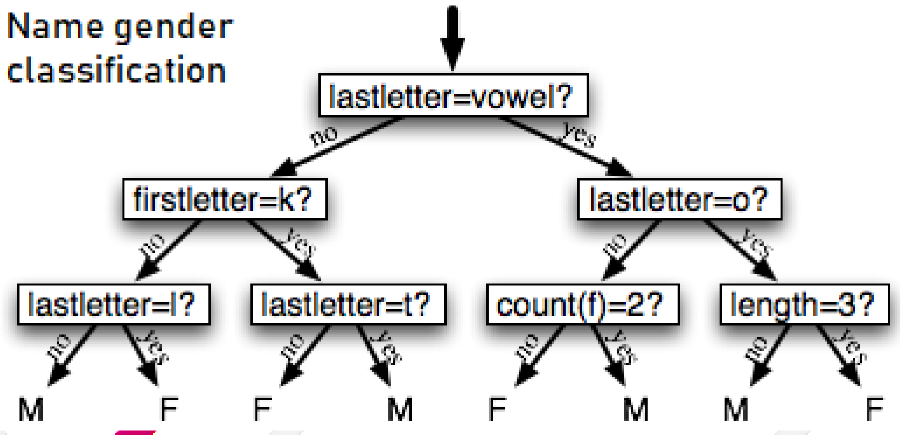
- E.g. rule-based chatbots



Statistical Modeling

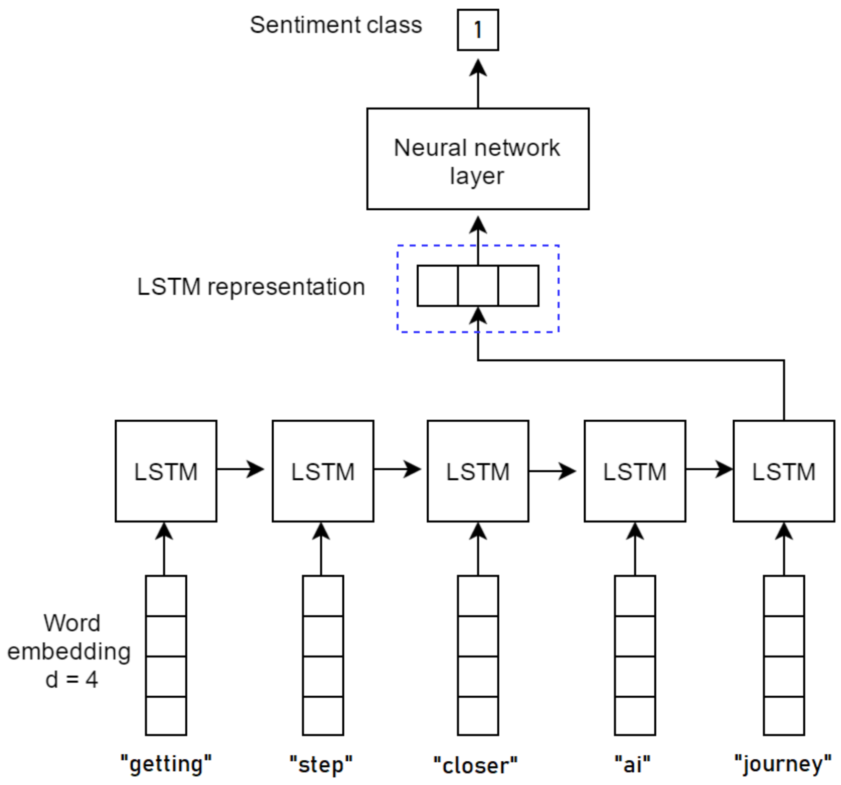
- Traditional ML models – training and testing data
- Requires moderate amount of data with annotations
- Heavy on feature engineering
  - word frequency, number of characters, edit distance
    capitalized, plural etc.
- Linear classifiers, Decision trees etc.

- Language Model
  - a probability distribution over sequences of words
  - can be also used for Language Generation
- N-gram Language Model
  - N-gram frequencies pre-counted on training corpus
  - P(“closer”|“getting a step”) > P(“coffee”|“getting



Deep Learning

- Feature engineering is generally skipped
  - raw data as an input (word embeddings)
  - network learns important features itself
- Large training corpus
- Good generalization
  - transfer learning – reusing models trained on different tasks
- Poor interpretability
- Sequence models
  - Recurrent Neural Networks (RNN, LSTM, GRU)
  - Temporal Convolutional Networks (TCN)
- Tasks
  - Classification, Regression, Sequence-to-sequence



Selected ”methods” of working with text

BAG-of-words (BOW)

- A simplifying representation, disregards word order
- Text is represented as a bag (multiset) of its words
  - multiplicity – number of occurrences of each word


```
from sklearn.feature_extraction.text import CountVectorizer
```

```
vectorizer = CountVectorizer(tokenizer=nltk.word_tokenize, vocabulary=dictionary)
```

```
bow = vectorizer.transform([sentence])
```

```
list(zip(dictionary.keys(), bow.toarray()[0])) #word counts
```

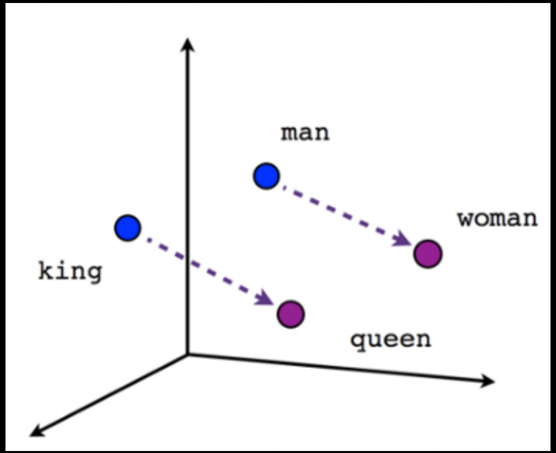
```
[('.', 2),
 ('a', 1),
 ('ai', 2),
 ('am', 1),
 ('closer', 1),
 ('getting', 1),
 ('i', 2),
 ('journey', 1),
 ('love', 1),
 ('my', 1),
 ('on', 1),
 ('step', 1)]
```

```
bow.toarray() #bag of words representation
```

```
array([[2, 1, 2, 1, 1, 1, 2, 1, 1, 1, 1, 1]])
```

WORD EMBEDDINGS

- Mapping of words into vectors of real numbers
- Words are closer to each other, if they occur in similar context
- King − Man + Woman = Queen
- Most popular embedding models are Word2Vec (Tomas Mikolov), FastText, or BERT




```python
import fasttext

model = fasttext.load_model("./fastText/cc.en.300.bin")

cosine_distance(model.get_word_vector("journey"), model.get_word_vector("voyage"))
```

```
0.42419618368148804
```

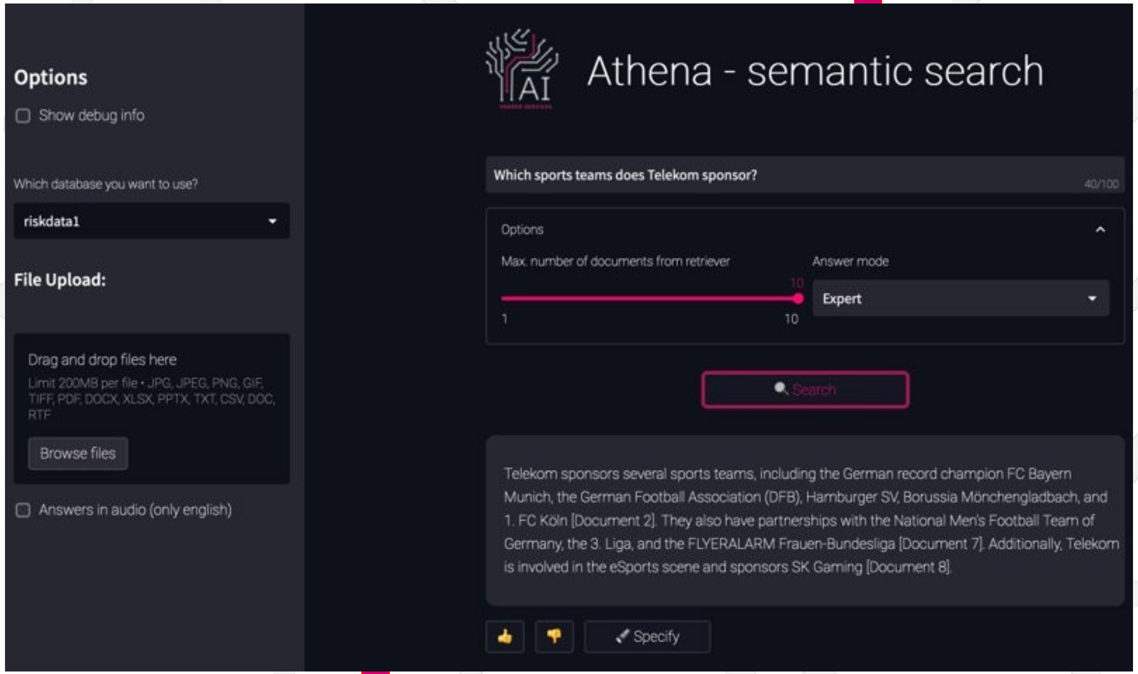
```python
cosine_distance(model.get_word_vector("journey"), model.get_word_vector("coffee"))
```

```
0.9299457967281342
```

LLMs @AI Shared Services – Product Athena

Athena Overview

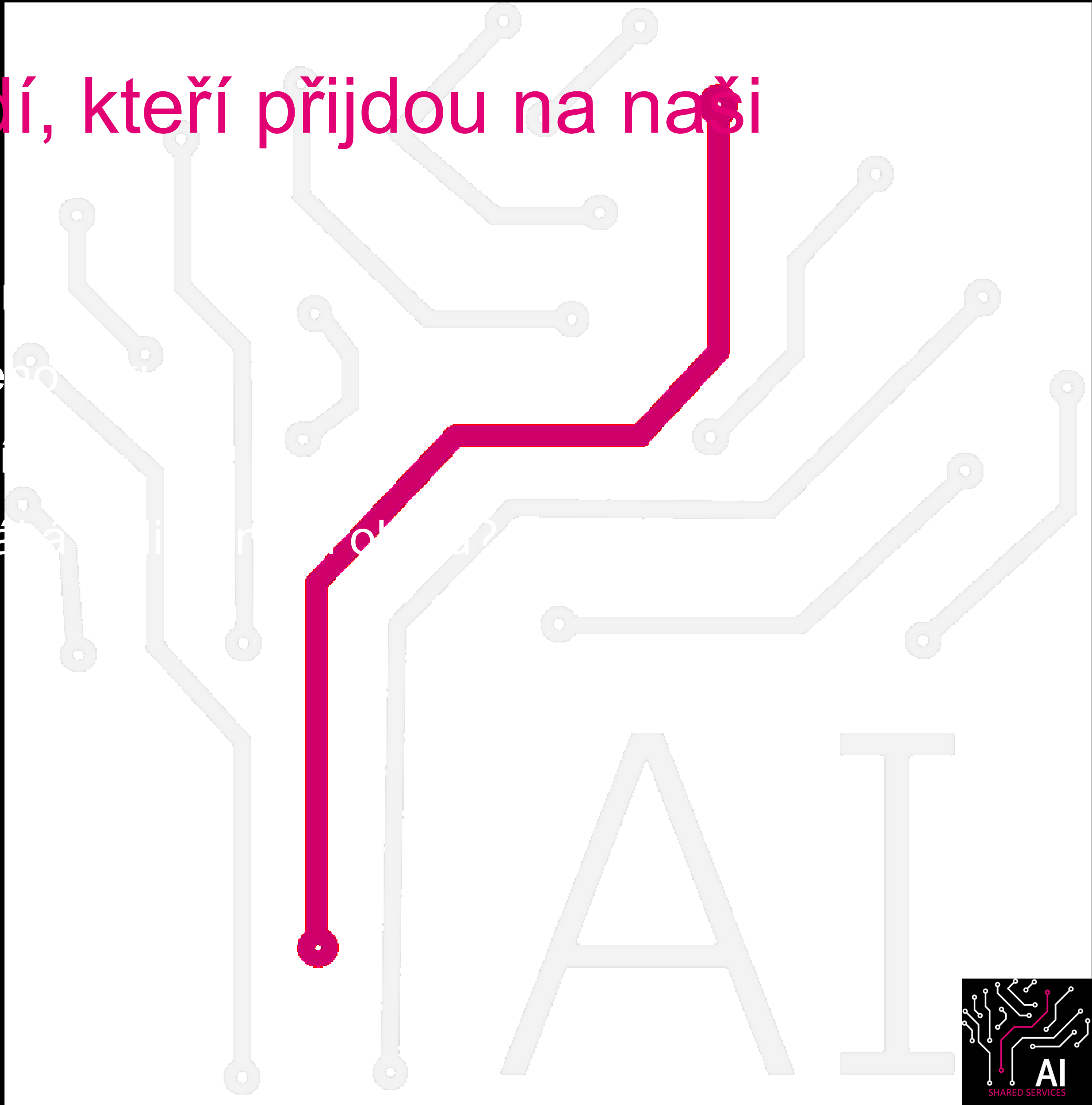
- Semantic search over various data types
- Enables fast orientation in vast amount of data
- Features:
  - Language independent
  - Summarization
  - Similarity matching
  - Text generation
  - Connection to web data



Part II. - Time Series Forecasting in DTSE

Jak odhadnete počet lidí, kteří přijdou na naši přednášku?

- Počet lidí na semináři minulý týde
- Počet lidí ve druhém týdnu loňské
- Zohlednění počtu studentů v roční
- Geniální anotace, která určitě přilá



Predikce časových řad

1. PRAVIDELNĚ uspořádané hodnoty
2. Predpoklad, že (některé) aspekty z
   pokračovat v budoucnosti a ovlivní
3. Možnost zahrnout další vysvětlující

Date	Position	Value
2022-01-01 00:00:00	Revenues	104
2022-02-01 00:00:00	Revenues	101
2022-03-01 00:00:00	Revenues	103
2022-04-01 00:00:00	Revenues	108
2022-05-01 00:00:00	Revenues	112
2022-06-01 00:00:00	Revenues	115
2022-07-01 00:00:00	Revenues	129
2022-08-01 00:00:00	Revenues	134
2022-09-01 00:00:00	Revenues	127
2022-10-01 00:00:00	Revenues	116
2022-11-01 00:00:00	Revenues	111
2022-12-01 00:00:00	Revenues	?
2023-01-01 00:00:00	Revenues	?
2023-02-01 00:00:00	Revenues	?

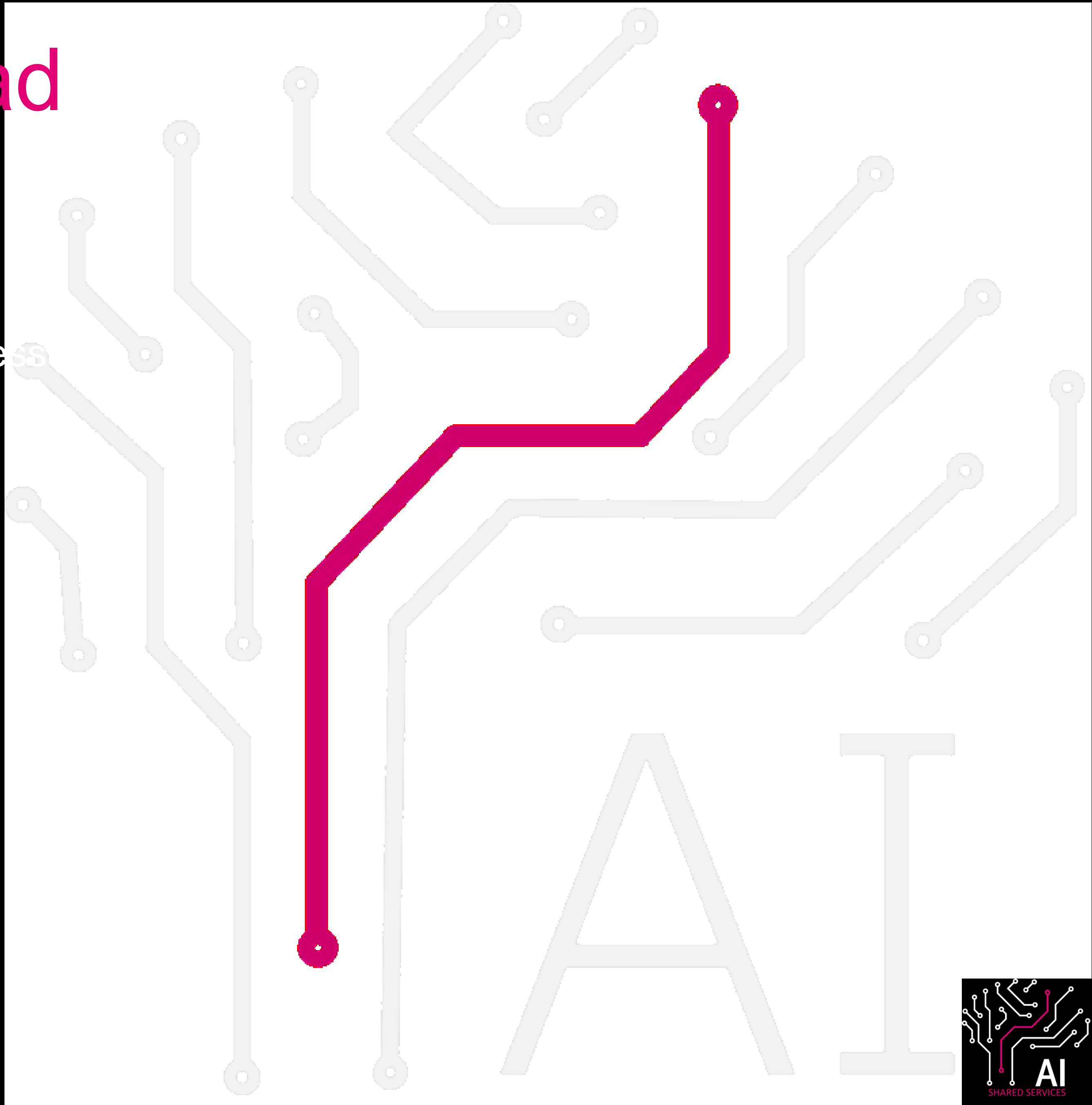
- V Deutsche Telekom hlavně měsíčn

Proč?

- Zlepšená schopnost rozhodování
- Zlepšená schopnost plánování
- Konkurenční výhoda
- Řízení rizika

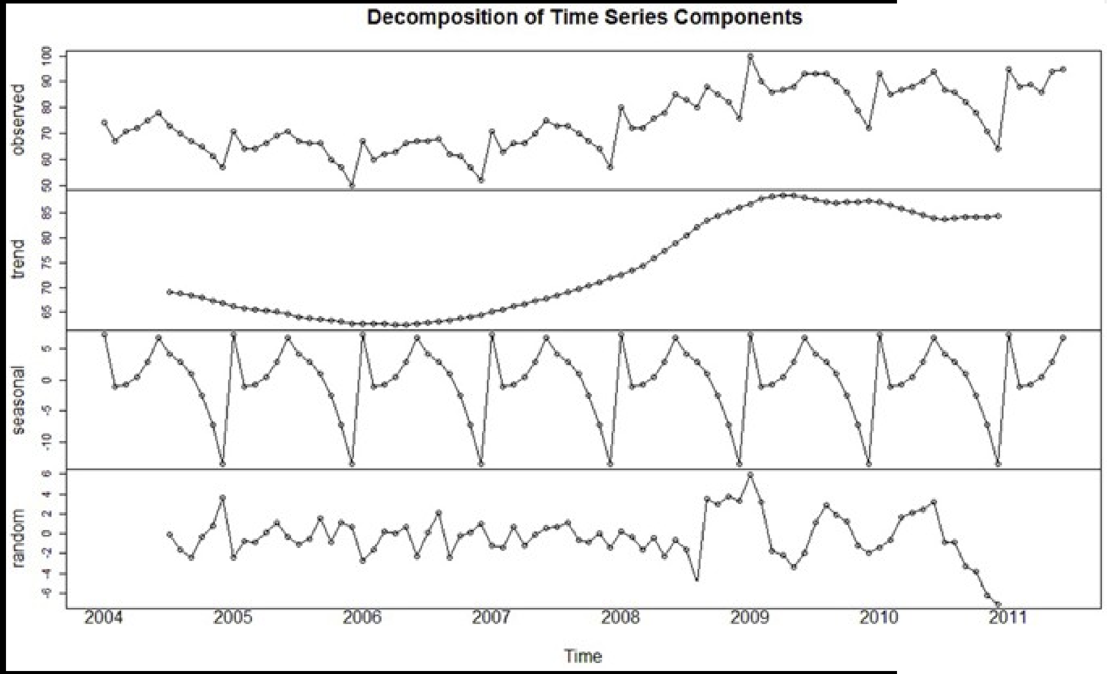
Predikce časových řad

- Konvenční postup:
    1. Definice problému
    2. Sběr dát a získávání insights od business
    3. EDA – vizualizace, popisné statistiky
    4. Preprocessing
    5. Výběr modelů a jejich hyperparametrů
    6. Cross-validace
    7. Výběr finálního modelu (ensembling?)
    8. Predikce
    9. Postprocessing
    10. Delivery

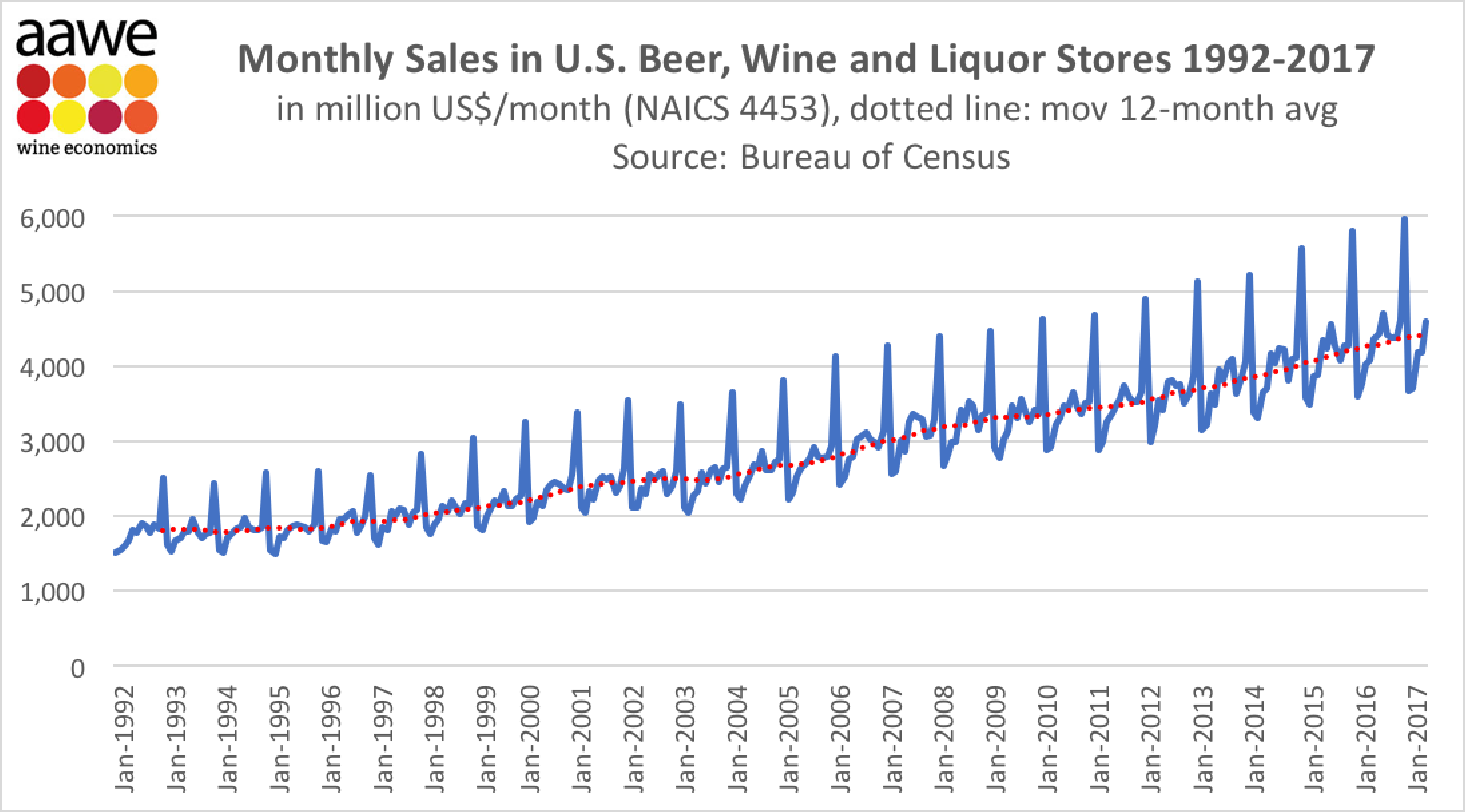


Komponenty časových řad

Komponenty časových radov

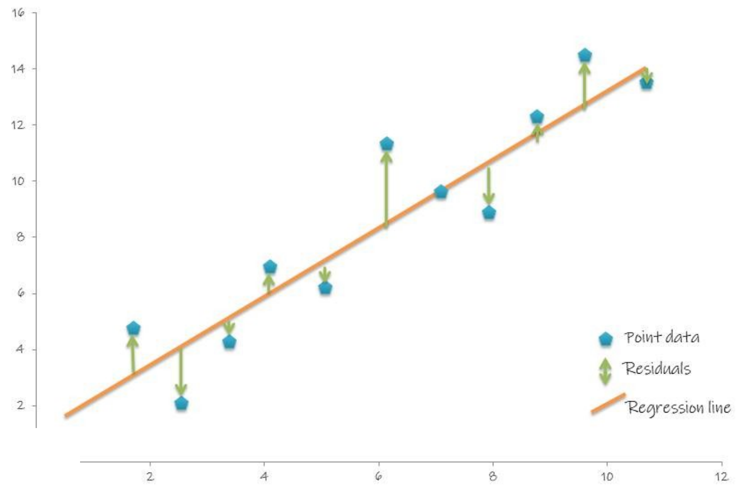


Trend
Seasonality
Residuals


aawe
wine economics
Monthly Sales in U.S. Beer, Wine and Liquor Stores 1992-2017
in million US$/month (NAICS 4453), dotted line: mov 12-month avg
Source: Bureau of Census
6,000
5,000
4,000
3,000
2,000
1,000
0
Jan-1992
Jan-1993
Jan-1994
Jan-1995
Jan-1996
Jan-1997
Jan-1998
Jan-1999
Jan-2000
Jan-2001
Jan-2002
Jan-2003
Jan-2004
Jan-2005
Jan-2006
Jan-2007
Jan-2008
Jan-2009
Jan-2010
Jan-2011
Jan-2012
Jan-2013
Jan-2014
Jan-2015
Jan-2016
Jan-2017

Chyba predikce

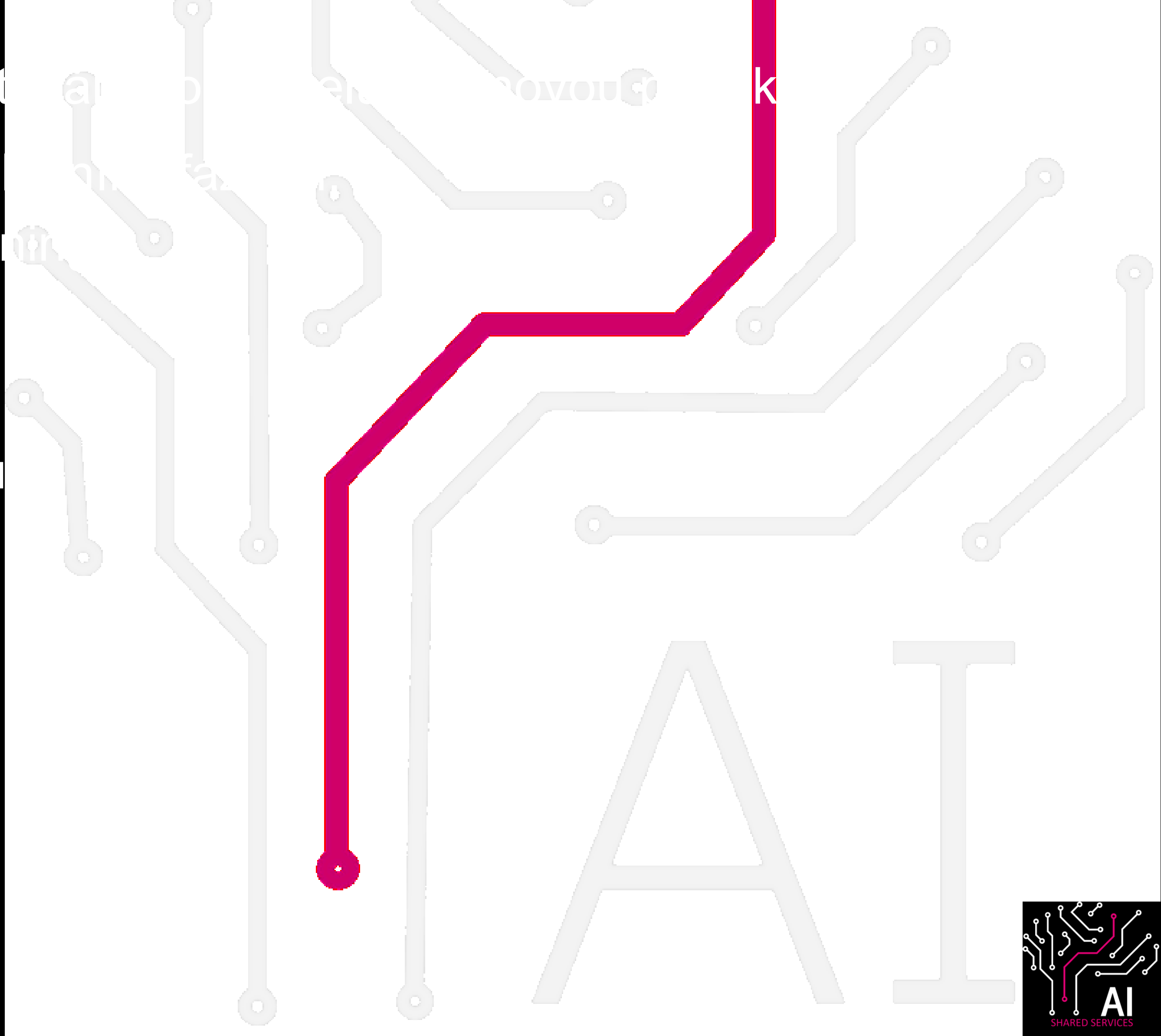
$$e_t = y_t - \hat{y}_t$$



- Common error metrics:
  - MAE (Mean Absolute Error) $MAE = mean(|e_t|)$
  - MSE (Mean Squared Error) $MSE = mean(e_t^2)$
  - RMSE (Root Mean Squared Error) $RMSE = \sqrt{mean(e_t^2)}$
  - MAPE (Mean Percentage Error) $MAPE = mean(|100 * e_t / y_t|)$

Auto ML řešení - Delphi Forecasting

- Každý měsíc probíhá nový výběr opt k
- Delphi algoritmus prechádza třemi h
  - Model training (‘hyperparameter tu
  - Model selection / evaluation
  - Prediction
- Separátně pro každou časovou řadu



Používané technologie

OPENSHIFT

PostgreSQL

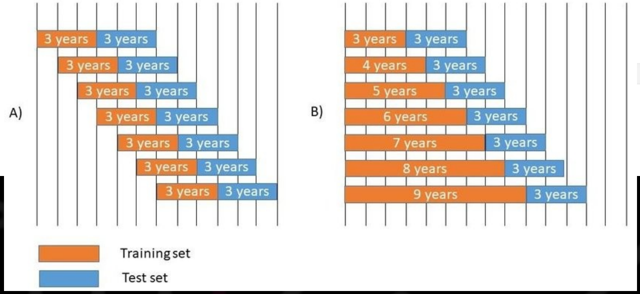
dagster

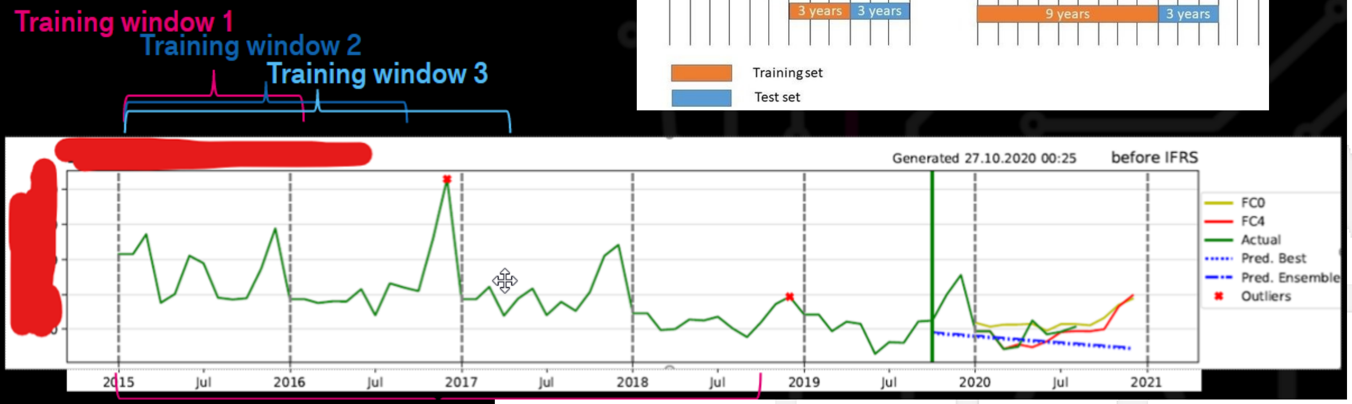
scikit learn

python

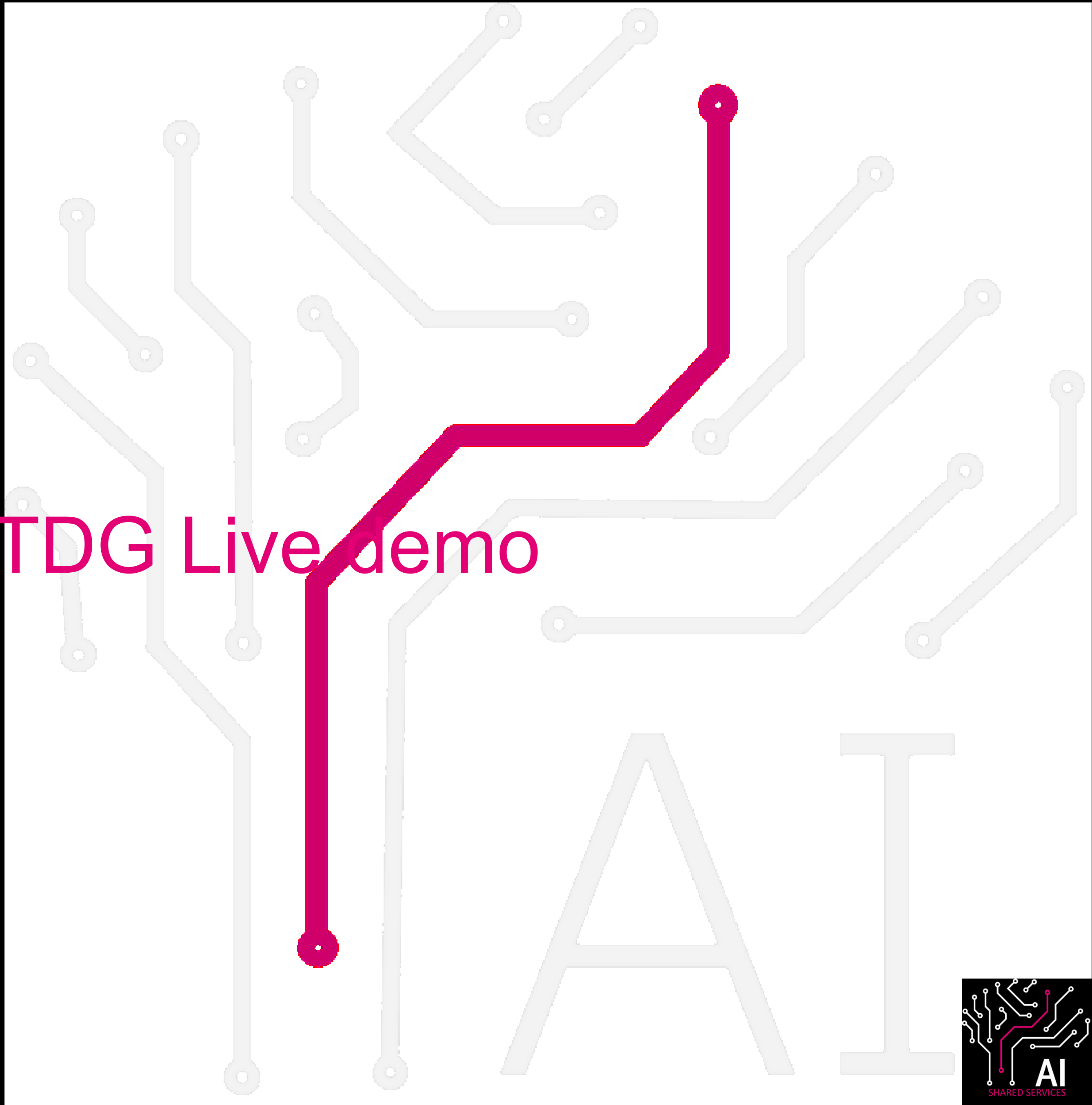
Tréninování Modelu

- Každý model má hyperparametry, které s … brát vysvělující proměnné?)
- K rychlejšímu výběru parametrů využívá … hyperparamery, které se nezdají, že by m …
- Model se fituje na základě zvolené komb … a následně … error metriky





DelPy TDG Live demo



Delphi Studio tour (1/3): allows to visualize and compare historical, forecasted, and adjusted data

Delphi Studio tour (2/3): adjust forecast to arrive at ML-assisted expert projection

Type	Aug 2021
Original Prediction	0.7
Adjustment in %	0.0
Adjustment abs.	0.2
Adjusted Prediction (und…	0.9
Elimination adj. ACT	
Adjusted Prediction (rep…	0.9

DelphiStudio

Value Driver

in [currency]

Select an entity: [anonym]

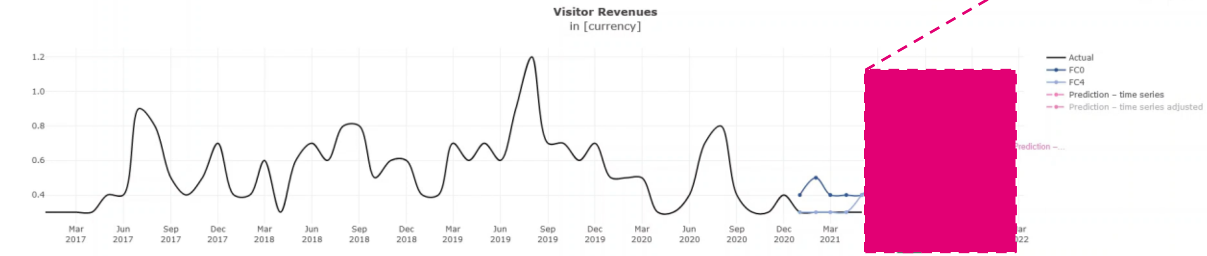
Select a position: ▶ 2.1.2 Visitor Revenues

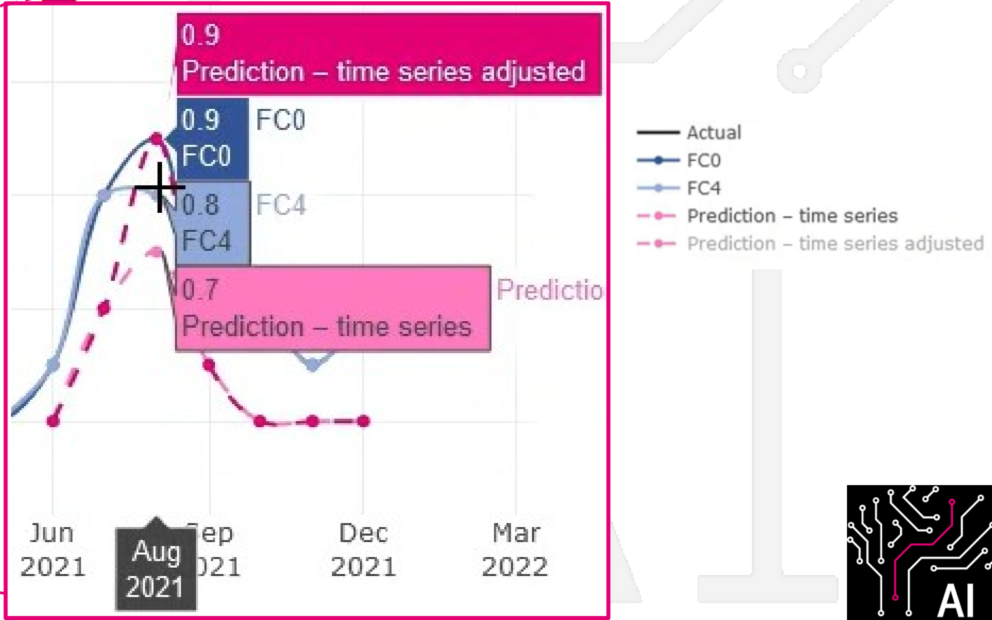
Visitor Revenues

Type	PY (full year)	FC4 (full year)	Prediction (full year)	Jan 2021	Feb 2021	Mar 2021	Apr 2021	May 2021	Jun 2…	[illegible]	…1	Sep 2021	Oct 2021	Nov 2021	Dec 2021
Original Prediction	5.4	5.8	4.7	0.3	0.3	0.3	0.3	0.3	[illegible]	[illegible]	0.7	0.5	0.4	0.4	0.4
Adjustment in %			0.0						[illegible]	[illegible]	0.0	0.0	0.0	0.0	0.0
Adjustment abs.			0.0						[illegible]	[illegible]	0.0	0.0	0.0	0.0	0.0
Adjusted Prediction (und…	5.4	5.8	4.7	0.3	0.3	0.3	0.3	0.3	[illegible]	[illegible]	0.7	0.5	0.4	0.4	0.4
Elimination adj. ACT															
Adjusted Prediction (rep…	5.4	5.8	4.7	0.3	0.3	0.3	0.3	0.3	[illegible]	[illegible]	0.7	0.5	0.4	0.4	0.4

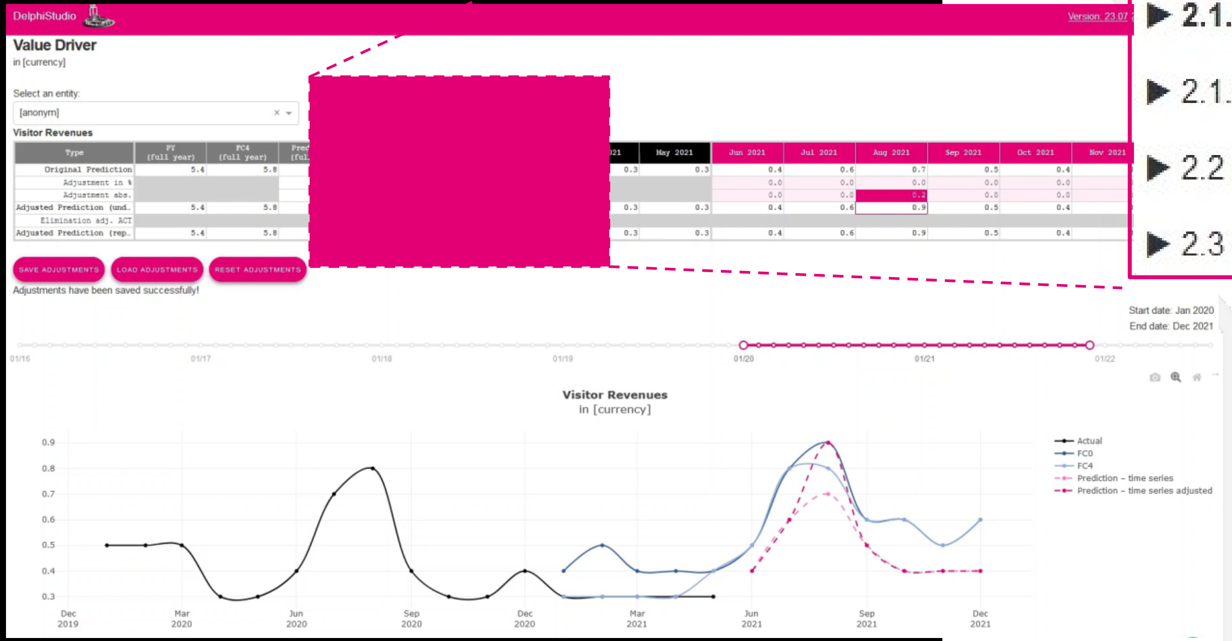
SAVE ADJUSTMENTS LOAD ADJUSTMENTS RESET ADJUSTMENTS

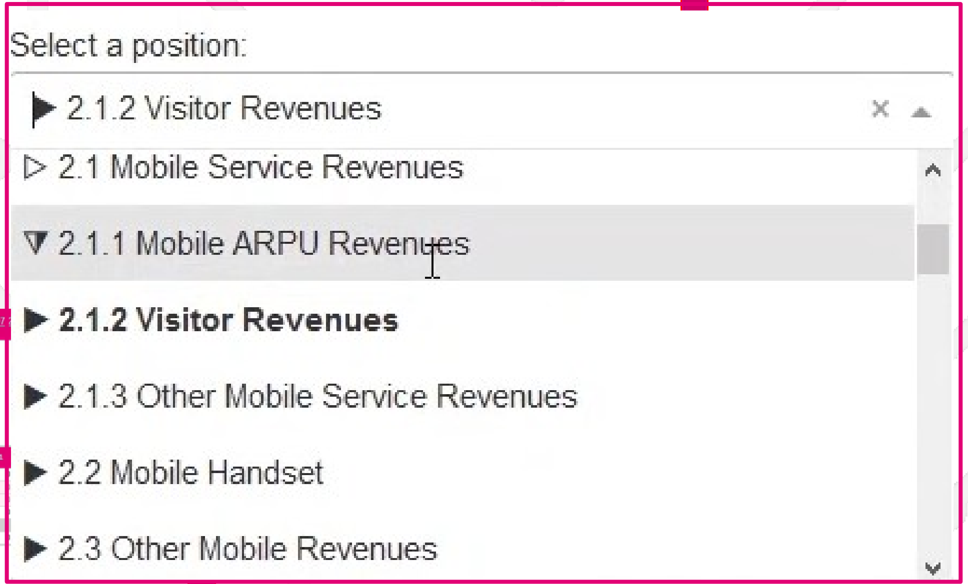
Start date: Jan 2017
End date: Dec 2021





Delphi Studio tour (3/3): switch between various P&L positions





ČÁST III. - Názory. Chcete někdo názory?

Specializace je vždy relativní

Naučte se psát kód koncepčně a tvořit úhledné projekty

- Čistý kód
- Vhodné pojmenovávání proměnných
- Modularita
- Pouze relevantní části
- Snadněji se tak buduje portfolio projektů
- Role engineeringu/architektury ML platform postup
  AutoML frameworks a množství dotupných předko

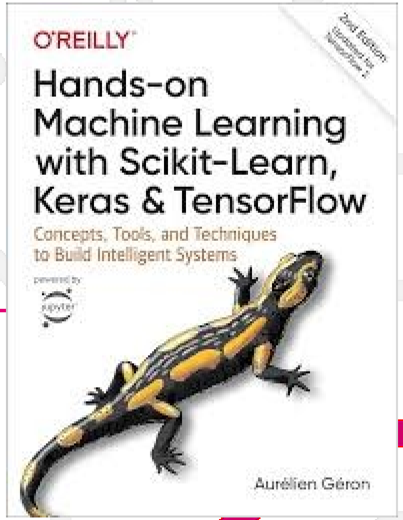
Neztraťte se v záplavě "cool data science"

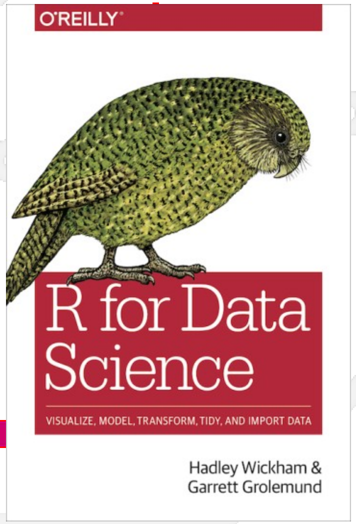
- Najděte balanc mezi užitečnými všeobecnými zna
- Don't skip the basics
- Nejtěžší je získat svou první práci v oboru
- Neztraťe se v záplavě kurzů a materiálů, opravdu

Zdroje ke studiu

- Science as Amateur Software Development
- Value in Data Science Beyond Models in Production | RStudio
- Forecasting – Principles and Practice
- Hands on ML
- Andrew Ng Courses
- R for DS + tidyverse tutoriály

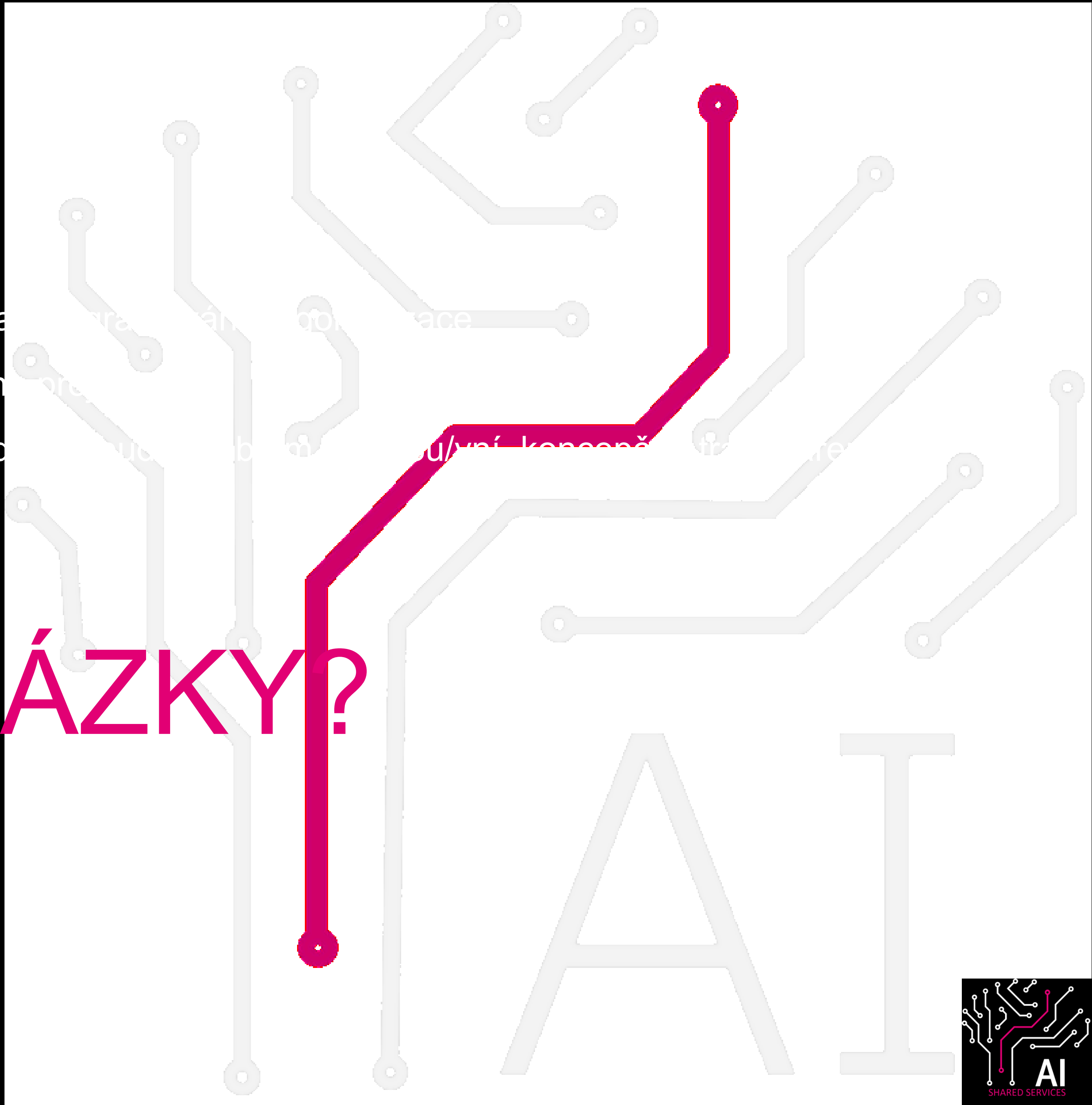
- LinkedIn





Závěr

- Data jsou a budou relevantní obor
- Nezapomínejte na základy - matematika, statistika
- Buďte trpěliví (při studiu, hledání práce, při práci n
- Soft skills jsou v IT někdy trochu opomíjená doved



OTÁZKY?