Large Scale Machine Learning at Zalando

Dr. Mikio L. Braun, Delivery Lead for Recommendation and Search

StrataHadoop 2016, London

mikio.braun@zalando.de
@mikiobraun
tech.zalando.com
One of Europes Biggest Fashion Retailers

★ 15 Countries
★ 3 warehouses
★ 19+ mn active customers
★ 2.9+ bn € revenue 2015
★ 135+ mn visitors per month
★ 9.900+ employees
★ IPO Oct 2014
★ ~7 bn € valuation (Oct 2015)

https://geschaeftsbericht.zalando.de/#de/report

Mikio Braun, Large Scale ML at Zalando, BBDC Fachtagung 2016
... Strong Technology Focus

★ 1000+ working in technology
★ 5 locations
  ○ Berlin
  ○ Dortmund
  ○ Hamburg
  ○ Dublin
  ○ Helsinki

tech.zalando.com

Mikio Braun, Large Scale ML at Zalando, BBDC Fachtagung 2016
How Zalando works

• Autonomous teams
• End-to-end responsibility
• Microservice architecture

Product Page ➔ Reco Plattform ➔ Reco Algorithms

Article Data

Mikio Braun, Large Scale ML at Zalando, BBDC Fachtagung 2016
Technology Stack

1. One AWS Account per Team
2. Deployment with Docker
3. Managed SSH Access
4. REST/OAuth 2.0
5. Changes fully auditable

STUPS

http://stups.io

amazon web services™

Mikio Braun, Large Scale ML at Zalando, BBDC Fachtagung 2016
Technology Stack

- Python, Jupyter/IPython notebooks locally
- Apache Spark
- Apache Flink
- Amazon Redshift
- But also “old school” Hadoop and a lot custom code…
Machine Learning at Zalando

- Recommendations  [https://tech.zalando.com/blog/feature-extraction-science-or-engineering/](https://tech.zalando.com/blog/feature-extraction-science-or-engineering/)
- Advertising
- Fraud protection [https://tech.zalando.com/blog/scalable-fraud-detection-fashion-platform/](https://tech.zalando.com/blog/scalable-fraud-detection-fashion-platform/)
- Search
- Zalando Research [https://tech.zalando.com/blog/zalando-launches-research-lab/](https://tech.zalando.com/blog/zalando-launches-research-lab/)

Mikio Braun, Large Scale ML at Zalando, BBDC Fachtagung 2016
Recommendations
Data Driven Recommendations

- Collaborative filtering
- Content based recommendation
- Personalised recommendations
- ...

Mikio Braun, Large Scale ML at Zalando, BBDC Fachtagung 2016
For Example, One-pass Ranking Models

Dataset: $\langle l, r, \mathbf{i} \rangle$
- $l_i = \text{relevance for user } i$
- $r_i = \text{rank } 1 \ldots n$
- $\mathbf{i}_i = \text{feature for item } i$

$$Q_w = X^T w$$

$$\min_w \left\{ \frac{1}{n} \sum_{j=1}^{m} \ell_{M_{(r(i), l(j))}}(X^{(j)}; \varphi_w) + \lambda_1 \|w\|_1 + \frac{1}{2} \lambda_2 \|w\|_2^2 \right\}$$

$$\ell_{M_{(r, l)}}(X; \varphi) = \sum_{r_i \leq r_j} \Delta_{M_{(r, l)}}(i, j) \mathcal{P}(\varphi(x_i), \varphi(x_j))$$

$$\Delta_{M_{(r, l)}}(i, j) = M(r, l) - M(r \setminus j, l)$$


Mikio Braun, Large Scale ML at Zalando, BBDC Fachtagung 2016
Hardcore Data Science to Production

- Usually one shot computation
- Sometimes done in Python
- Getting raw data hard initially

Mikio Braun, Large Scale ML at Zalando, BBDC Fachtagung 2016
Production System

- Realtime system
- Usually done in Java/JVM based
- Events and article data continually upgraded
Data Science vs. Production

- A/B Test ↔ offline evaluation
- Iterate on data science part
- Iterate on the whole system!

Mikio Braun, Large Scale ML at Zalando, BBDC Fachtagung 2016
Data Scientists and Developers

Exploration

Design

Coding

Data Science

Developer

very loose goals, abundance of data

evaluation, benchmarking

code to run for years in production

one off, throw away
DS&D: Coding

Very different approaches to coding...

← developers

data scientists →

Mikio Braun, Large Scale ML at Zalando, BBDC Fachtagung 2016
DS&D: Collaboration

• What is the most productive way?
• Ideally, interface on code, not just documentation
• Production logs often become data analysis input!
Organization

- Cross-functional teams
- Communication!
- Microservices, at Zalando: STUPS (Docker on AWS)
Summary

• Machine Learning plays an important role at Zalando for many different aspects
• Large Scale Learning for Recommendation, etc. …
• Organization of many teams with end-to-end responsibilities
• How to bring data science live