

Data Science meets Software Data

Markus Harrer, Software Development Analyst

@feststelltaste

The Architecture Gathering, 17.10.2019



Über mich

**// Tools only
find, people
have to find
out!"**



Markus Harrer

Senior Consultant, Nürnberg

- Architektur-Reviews
- Software-Modernisierung und -Rightsizing
- Performance-Optimierung
- Datenanalysen in der Softwareentwicklung

INNOQ



cards42



Cards for Analyzing and
Reflecting on Doomed
Software



Foundation & IMPROVE

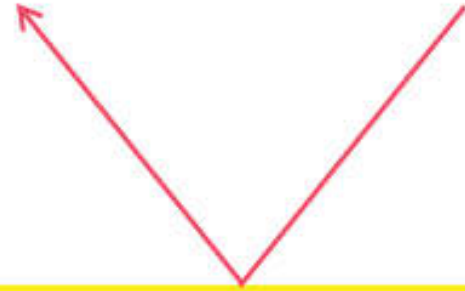


Datenanalysen in der Softwareentwicklung?

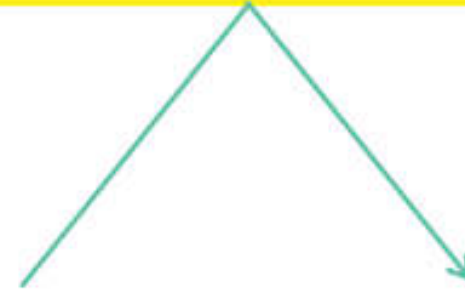
Das (Tr|D)auerthema

Management

Risikoappetit



Kommunikationsbarriere



Konsequenzen

Softwarearchitekt*innen

Das (Tr|D)auerthema

Management

Risikoappetit

Qualitäts-
anforder-
ungen ✓

Kommunikationsbarriere

Konsequenzen

Softwarearchitekt*innen



Das (Tr|D)auerthema



Lösung:

Data Science meets Software Data?

Data Science

Was ist Data Science?

"**Statistik** auf nem **Mac**."

Nach <https://twitter.com/cdixon/status/428914681911070720>

Meine Definition

Was bedeutet "data"?

"Without **data** you're just another person with an opinion."

W. Edwards Deming

=> Belastbare Erkenntnisse mittels **Fakten liefern**

Was bedeutet "science"?

"The aim of **science** is to seek the simplest explanations of complex facts."

Albert Einstein

=> Neue Erkenntnisse **verständlich herausarbeiten**

Was hat das mit Softwareentwicklung zu tun?

Was ist ein Data Scientist?

"Jemand, der mehr Ahnung von **Statistik** hat als ein **Softwareentwickler** und mehr Ahnung von **Softwareentwicklung** als ein **Statistiker**."

Nach zu <https://twitter.com/cdixon/status/428914681911070720>

=> Wir sind gar nicht so weit weg!

Software Data

Welche Arten von Softwaredaten?

Alles was aus der Entwicklung und dem Betrieb der Softwaresysteme so anfällt:

- Statische Daten
- Laufzeitdaten
- Chronologische Daten
- Daten aus der Software-Community

=> Eine sehr große Auswahl!

Data Science & Software Data:
Perfect match!

Software Analytics

Definition "Software Analytics"

"Software Analytics is analytics on software data for **managers** and **software engineers** with the aim of empowering software development individuals and teams to gain and share insight from their data to **make better decisions.**"

Tim Menzies and Thomas Zimmermann

Das (Tr|D)auerthema



~~Das (Tr|D)auerthema~~



Aber: Hier gibt es ein Problem!

Thomas Zimmermann in "One size does not fit all":

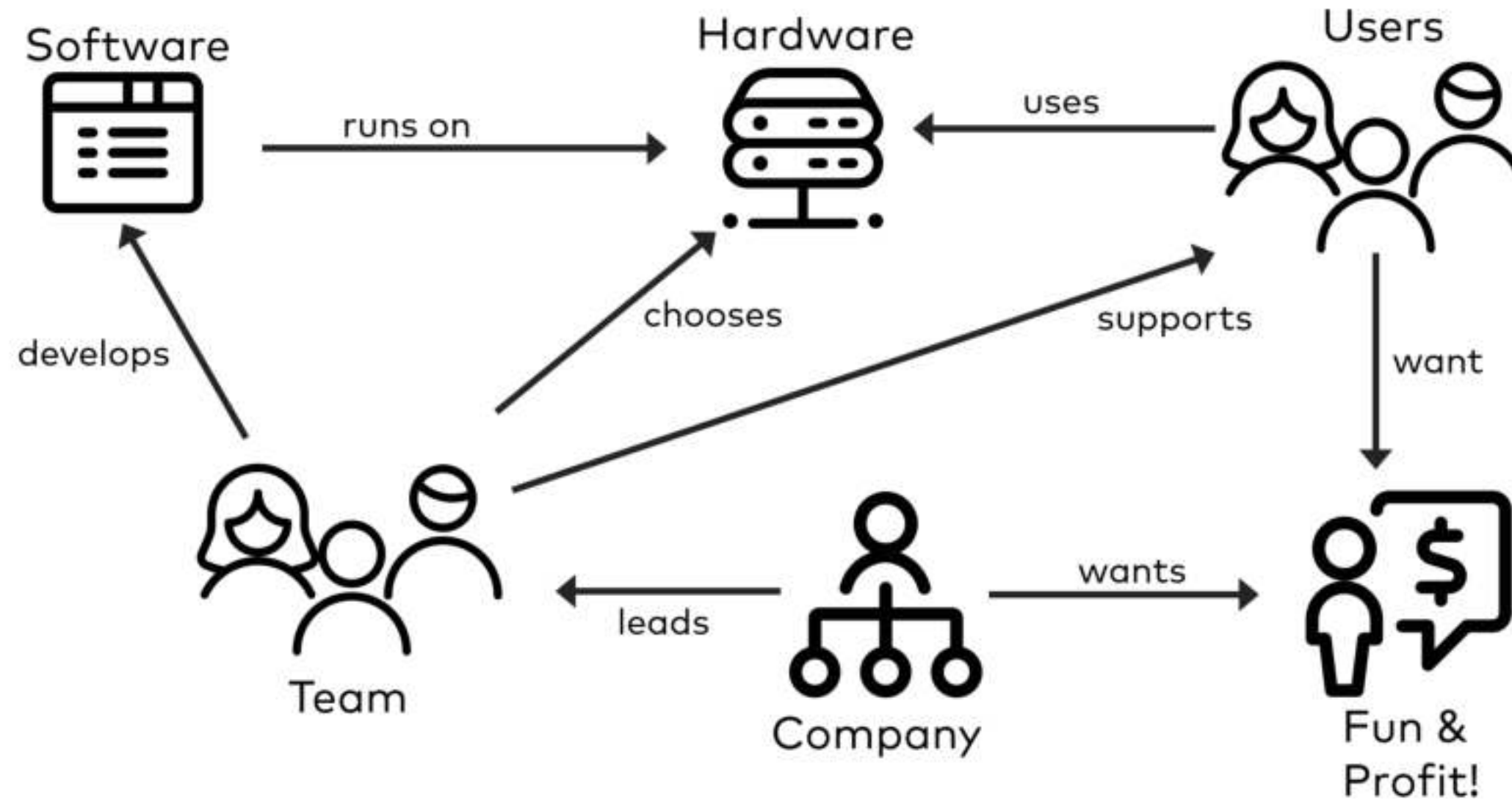
"Even if you find models that work for most, they will not work for everyone.

There is much **academic research** into **general models**. In contrast, **industrial practitioners** are often fine with **models that just work for their data** if the model provides some insight or allows them to work more efficiently."

Beispiele für Analysen aus der Praxis

- Vorhandenen Modularisierungsschnitt analysieren
- Performance-Probleme in verteilten Systemen identifizieren
- Potenzielle Wissensverluste ermitteln
- Eingesetzte Open-Source-Projekte bewerten
- ...

"Es kommt drauf an!" aka Kontext



Individuelle Systeme => individuelle Probleme => individuelle Analysen => individuelle Erkenntnisse!

~~Das (Tr|D)auerthema~~



~~Das (Tr|D)auerthema~~



Data Science mit Softwaredaten

**Eine leichtgewichtige Umsetzung von Software
Analytics**

Best Practices

Bewährtes Vorgehen

Nach Roger Pengs "Stages of Data Analysis"

- I. Fragestellung
- II. Explorative Datenbeschaffung
- III. Formale Modellierung
- IV. Interpretation
- V. Kommunikation

=> von der Frage über die Daten zur Erkenntnis!

Literate Statistical Programming

(Intent + Code + Data + Results)

* Logical Step

+ Automation

= Literate Statistical Programming

Implementierung: **Computational notebooks**

Computational Notebooks

The CSV file contains all lines of code that were passed through during the measurement's time span. We just take the relevant data and add an additional `LINES` column to be able to calculate the ratio between covered and missed lines later on.

In [1]:

```
1 import pandas as pd
2 coverage = pd.read_csv("../input/spring-petclinic/jacoco.csv")
3 coverage = coverage[['PACKAGE', 'CLASS', 'LINE_COVERED', 'LINE_MISSED']]
4 coverage['LINES'] = coverage.LINE_COVERED + coverage.LINE_MISSED
5 coverage.head(1)
```

Out[1]:

	PACKAGE	CLASS	LINE_COVERED	LINE_MISSED	LINES
0	org.springframework.samples.petclinic	PetclinicInitializer	24	0	24

Flexible Standardwerkzeuge (1/2)

Klassischer Data-Science-Werkzeugkasten

- **Jupyter** (mit RISE)
- Python 3
- pandas
- matplotlib

Flexible Standardwerkzeuge (1/2)

Jupyter funktioniert und integriert sich auch mit

- jQAssistant / Cypher / Neo4j
- JVM-Sprachen über beakerx / Tablesaw
- bash
- ...

Praktischer Teil

Weitere Analysebeispiele

- Analyse der CPU-Auslastung mit Daten von vmstat
- Abhängigkeitsanalyse mit Daten von jdeps und Visualisierung mit D3
- Identifizierung von Modularisierungsoptionen basierend auf reinen Codeänderungen in Git
- *Call-Graph-basierte Analyse von Performance-Hotspots mit jqAssistant / Neo4j*
- *Automatisierte Prüfung der (Architektur-)Governance-Vorgaben mit jqAssistant*

Zusammenfassung

1. **Softwareanalysen** mit Data-Science-Werkzeugen sind möglich
2. Wer mehr will bekommt auch **mehr!**
3. Es gibt unglaublich **viele Quellen** für Daten in der Softwareentwicklung

=> von der **Frage** über die **Daten** zur **Erkenntnis!**

Einstieg in das Thema

- Adam Tornhill: Software Design X-Ray
- Wes McKinney: Python For Data Analysis
- Jeff Leek: The Elements of Data Analytic Style
- Tim Menzies, Laurie Williams, Thomas Zimmermann: Perspectives on Data Science for Software Engineering
- Mini-Tutorial unter <https://github.com/feststelltaste/software-analytics-workshop>

Schnelleinstieg

Meine TOP 5's*

<https://www.feststelltaste.de/category/top5/>

Kurse, Videos, Blogs, Bücher und mehr...

*einige Seiten befinden sich noch in der Entwicklung

Vielen Dank! Fragen?

Markus Harrer

innoQ Deutschland GmbH

markus.harrer@innoq.com

@feststelltaste