

Einführung in Software Analytics

Markus Harrer

@feststelltaste

rheinJUG, Düsseldorf, 19. November 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



ACCREDITED
Trainer

Foundation & IMPROVE



LATEST NEWS



Datenanalysen in der Softwareentwicklung?

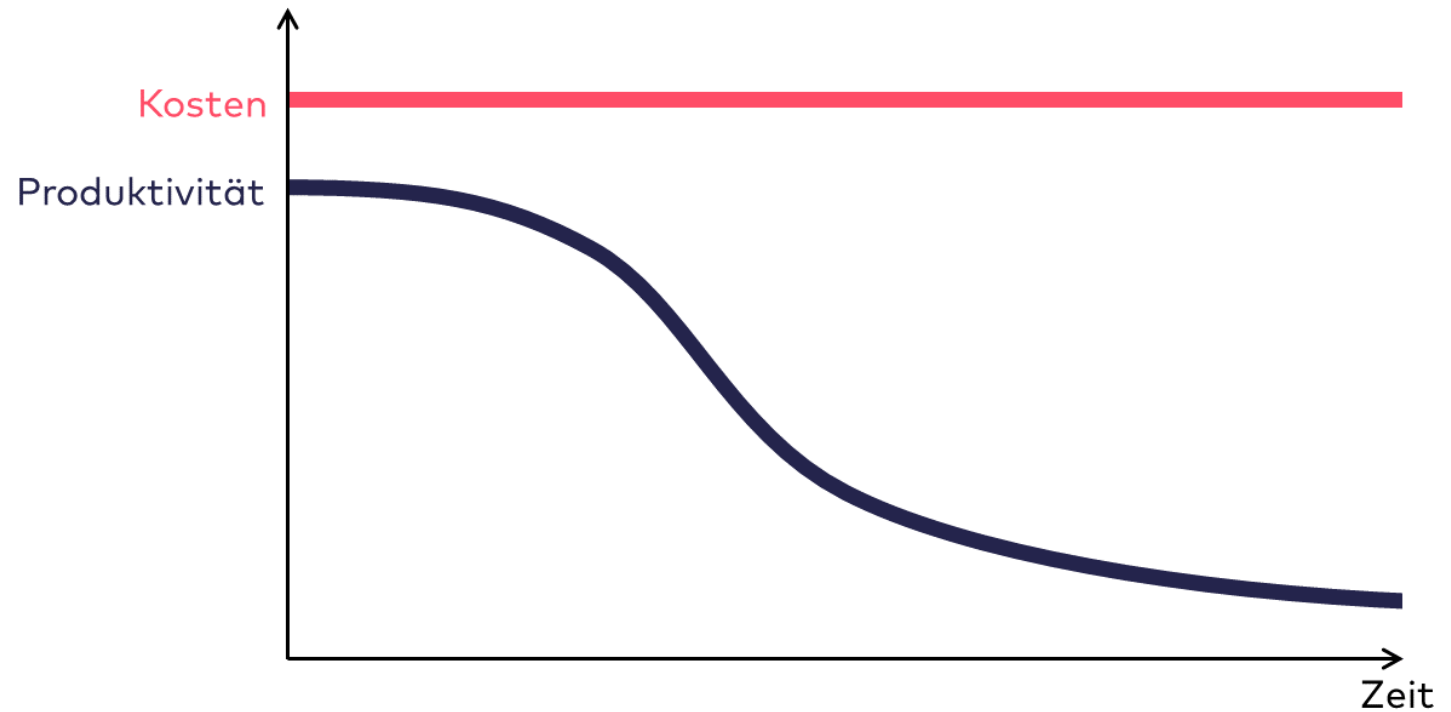
Warum?

... ein typischer Projektverlauf

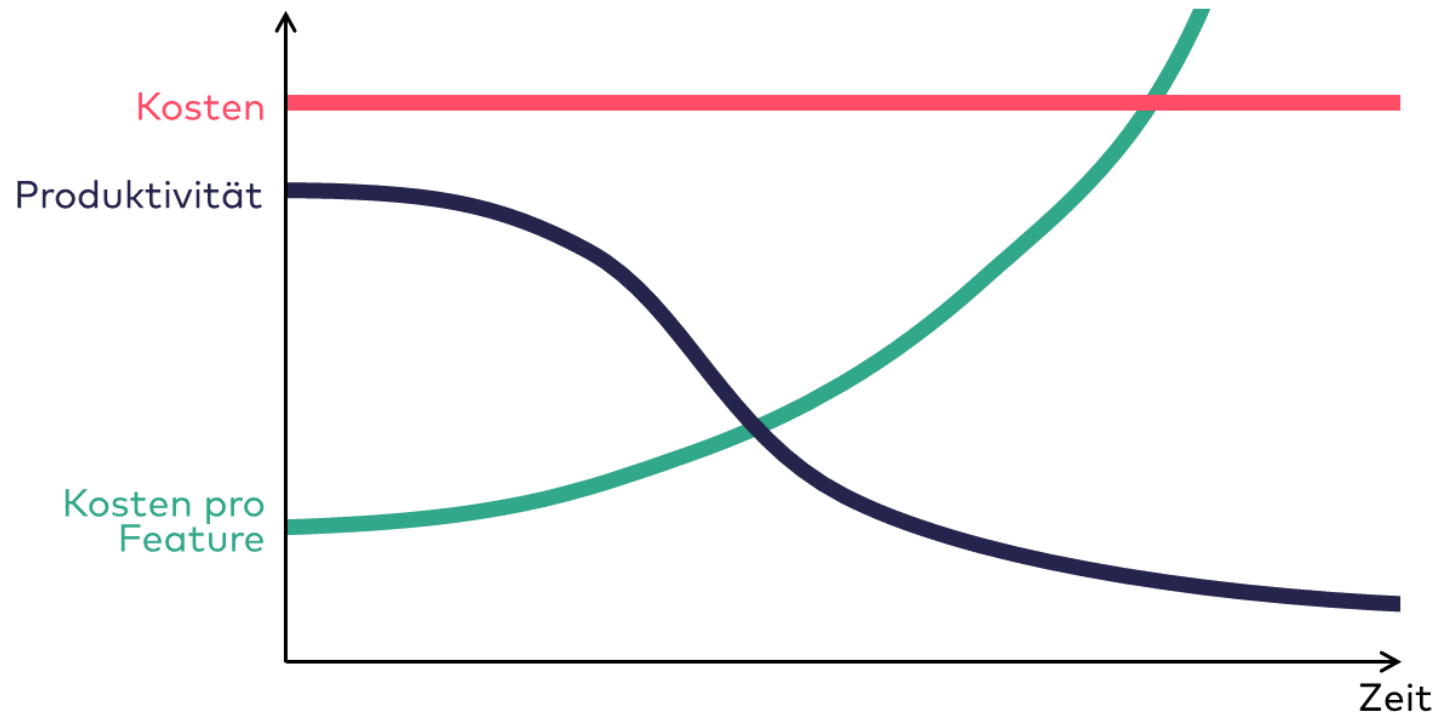
... ein typischer Projektverlauf



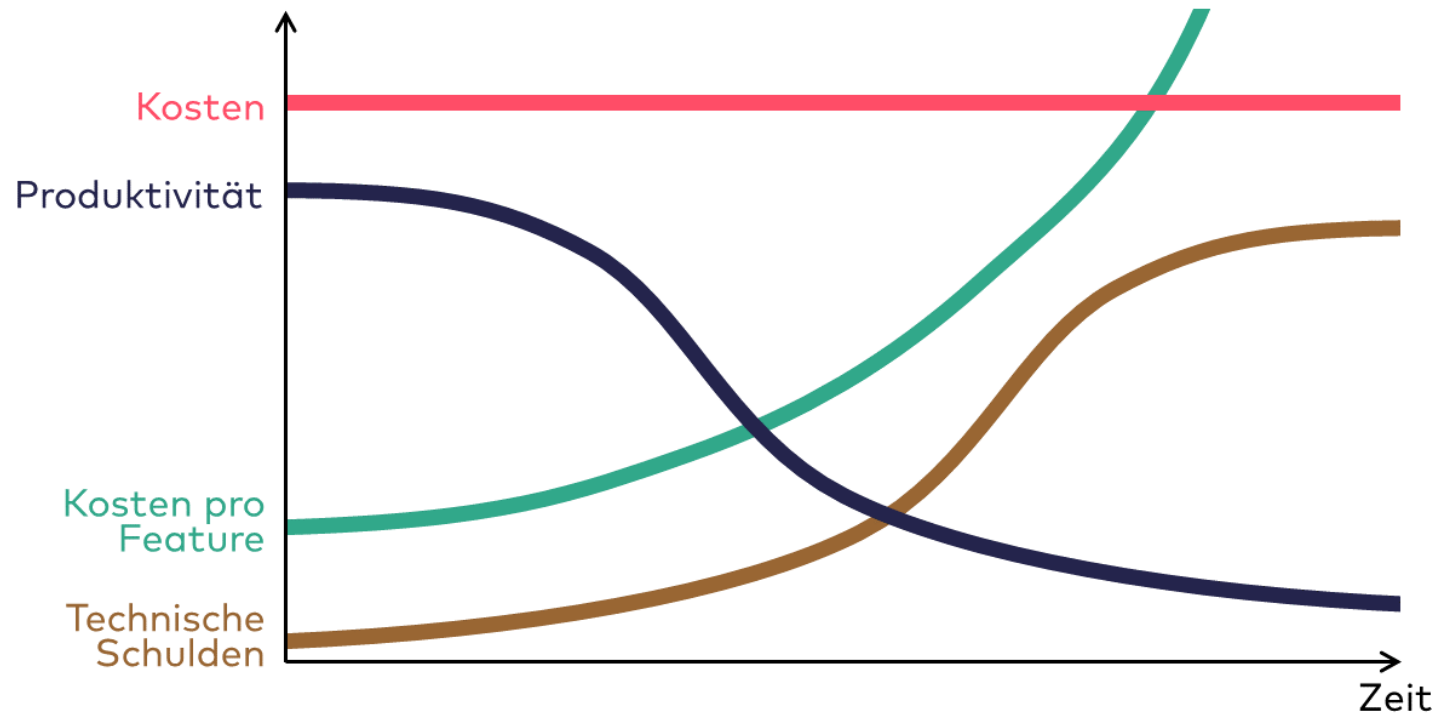
... ein typischer Projektverlauf



... ein typischer Projektverlauf



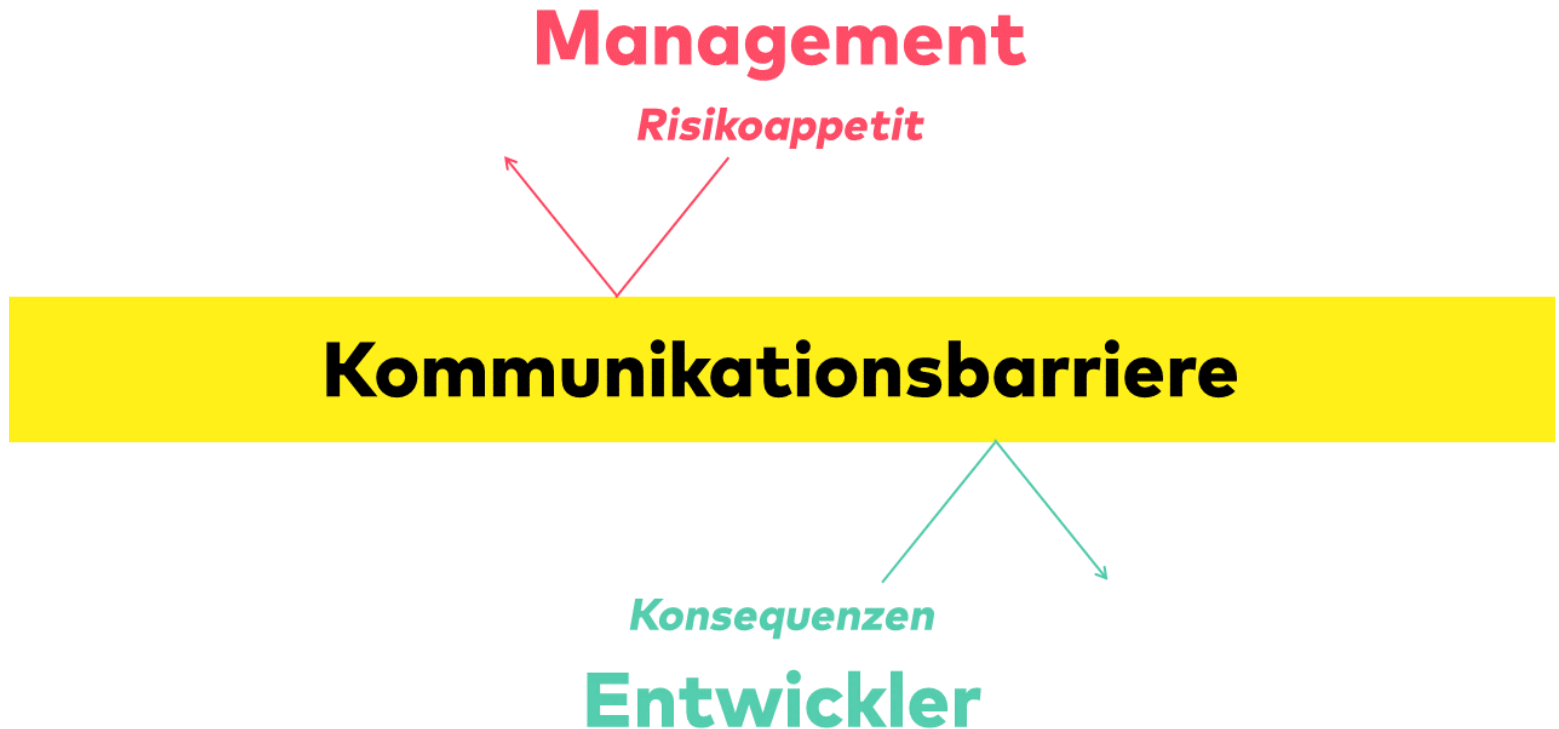
... ein typischer Projektverlauf



"Die Definition von Wahnsinn ist, immer wieder das Gleiche zu tun und andere Ergebnisse zu erwarten."

– Albert Einstein

Das (Tr|D)auerthema



Das (Tr|D)auerthema



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

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

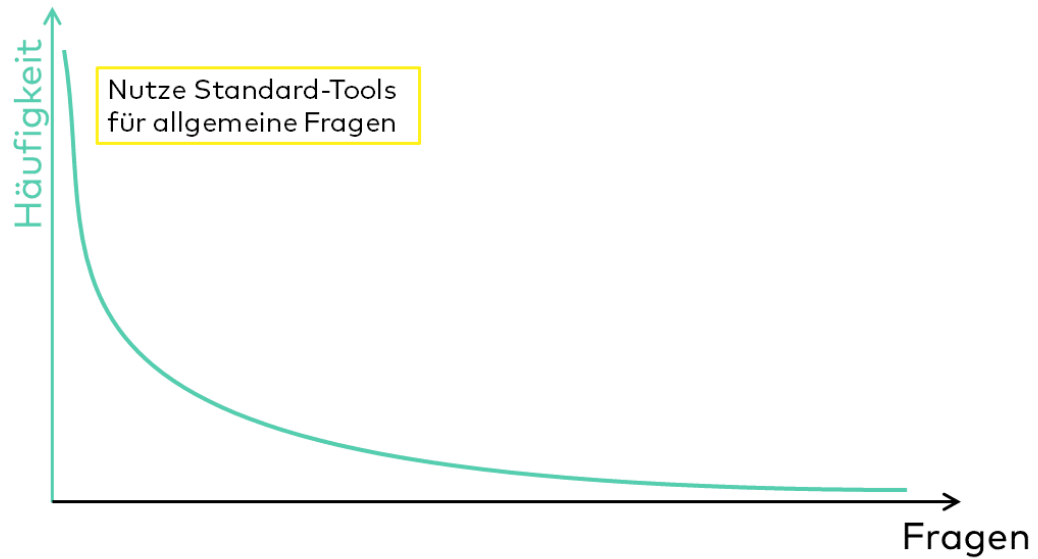
Sehr große Auswahl == sehr große Möglichkeiten?

"Software Analytics" heutzutage

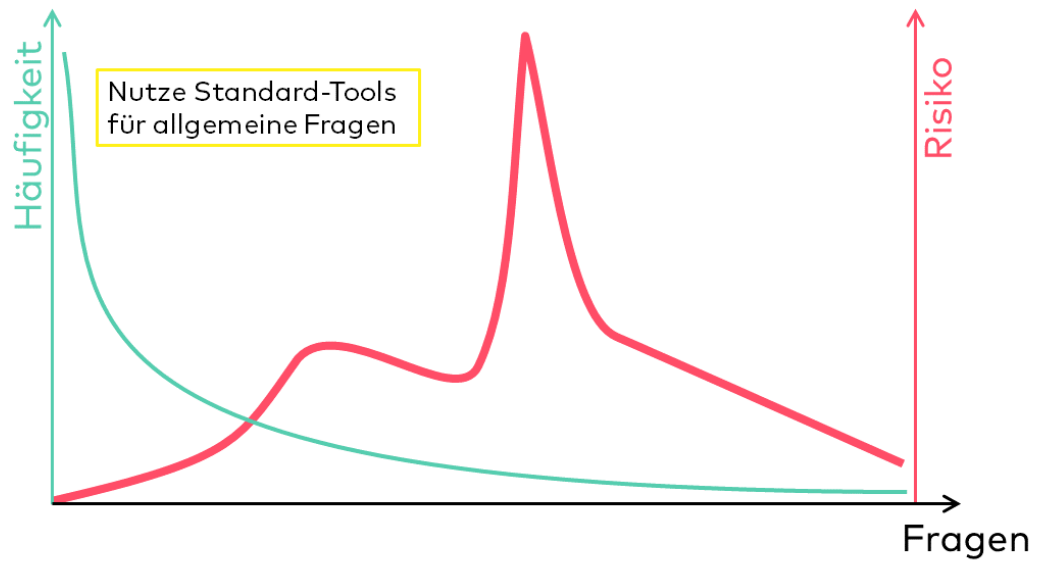


Markus traurig

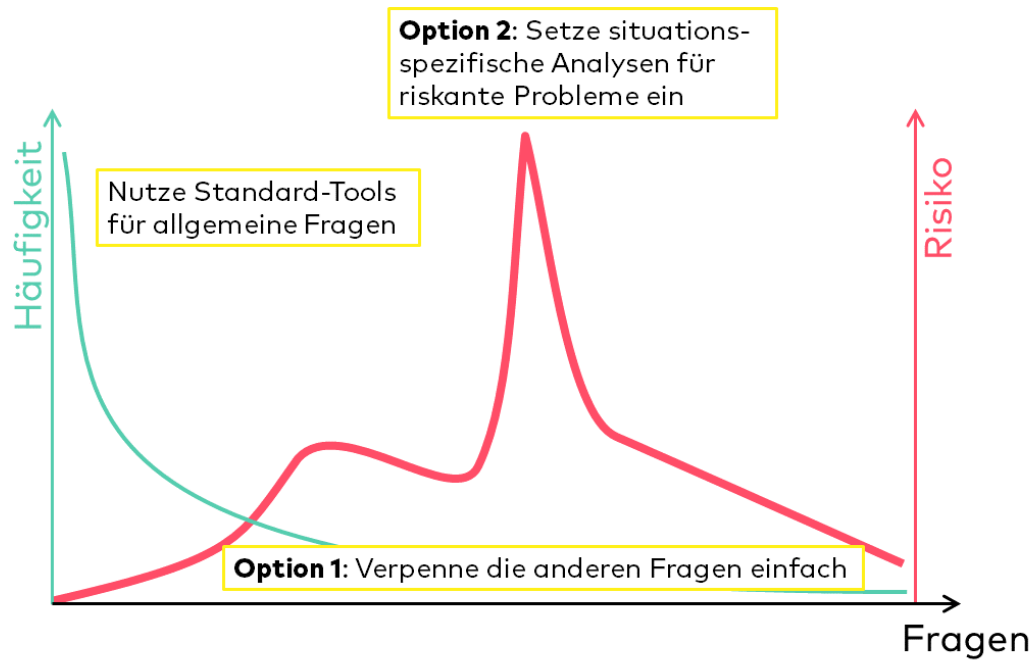
Unsere derzeitiges "Software Analytics"



Unsere derzeitiges "Software Analytics"



Unsere derzeitiges "Software Analytics"

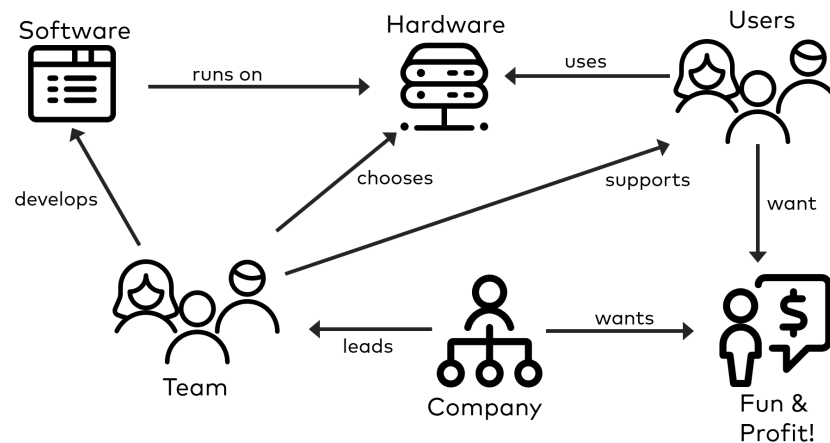


"Software Analytics" heutzutage

Leider Option 1: Wir verpennen Risiken

Es fehlt Option 2: Situationsspezifische Analysen

"Es kommt drauf an!" aka Kontext



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

Situationsspezifische Analysen mit Mitteln aus dem

Data Science

**= Eine leichtgewichtige Umsetzung von Software
Analytics**

Data Science

Was ist Data Science?

"Statistics on a Mac."

<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."

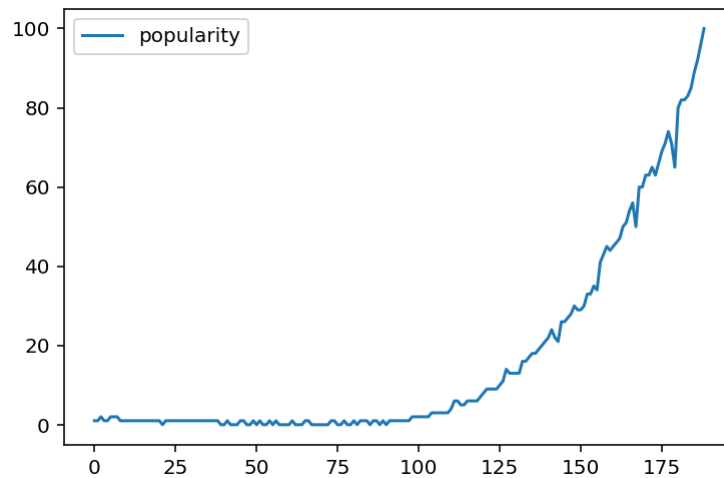
Albert Einstein

=> Neue Erkenntnisse **verständlich** herausarbeiten

Warum Data Science?

Data Science liegt immer noch im Trend!

```
In [3]: pd.read_csv("../dataset/google_trends_datascience.csv").plot();
```



"100" == max. Beliebtheit!

Große (Online-)Community

- Kostenlose Online-Kurse, -Videos und -Tutorials (z. B. DataCamp mit über 4,7 Mio. Mitgliedern)
- Direkte Hilfestellungen (z. B. Stack Overflow oder Blog-Artikel)
- Lernen mit anderen durch Online-Herausforderungen (z. B. Kaggle)

Wiederverwendbarkeit

- Analyse von Geschäftsdaten oder Softwaredaten? Fast das Gleiche!
- Vorgehen für Analysen? Bewährt!
- Best Practices für Analysen? Frei Haus!
- Visualisierungen: Spart immens Zeit!

**Wie weit weg sind
SoftwareentwicklerInnen
von Data Science?**

Was ist ein Data Scientist?

"A data scientist is someone who is better at **statistics** than any **software engineer** and better at **software engineering** than any **statistician**."

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

Nicht so weit weg wie gedacht!

Einige individuellen Analysen aus der Praxis

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

**Wie Software Analytics mit Data Science
beginnen?**

Bewährte Ansätze nutzen

Roger Pengs "Stages of Data Analysis"

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

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

Gedanken zur Analyse strukturieren

Software Analytics Canvas Projekt: _____			
 Fragestellung <i>Was ist es, was wir über die Software / Prozesse / Nutzung / Organisation / o. ä. wissen wollen?</i>	 Datenquellen <i>Welche Daten können uns unsere Frage möglicherweise beantworten? Welche Informationen brauchen wir?</i>	 Heuristiken <i>Welche Annahmen wollen wir treffen, die uns die Beantwortung unserer Frage vereinfachen?</i>	 Validierung <i>Welche Ergebnisse erwarten wir aus unserer Analyse, wie werden diese überprüft und verständlich dargestellt?</i>
 Implementierung <i>Wie können wir die Analyse Schritt für Schritt und nachvollziehbar umsetzen?</i>		 Ergebnisse <i>Was sind die wichtigsten Erkenntnisse aus unserer durchgeführten Analyse?</i>	 Nächste Schritte <i>Welche Folgeaktionen können wir aus den Erkenntnissen ableiten? Wen müssen wir als nächsten ansprechen?</i>

Wie nachvollziehbar umsetzen?

Verwende Literate Statistical Programming

(Intent + Code + Data + Results)

* Logical Step

+ Automation

= Literate Statistical Programming

Vehikel: Computational notebooks

Beispiel "Computational Notebook"

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

Nutze Data Science Standardwerkzeuge

z. B. einen der populärsten Stacks

- Jupyter Notebook
- Python 3
- pandas
- matplotlib

Andere Technologien

Jupyter Notebook arbeitet auch mit anderen Technologieplattformen zusammen, z. B. mit

- bash
- jQAssistant Scanner / Neo4j Graphdatenbank
- JVM-Sprachen via beakerx / Tablesaw

=> Also auch was für Javaianer dabei!

Hands-On

Programmierbeispiel

Fallbeispiel

IntelliJ IDEA

- IDE für Java-Entwickler
- Fast komplett in Java geschrieben
- Großes und lang aktives Projekt

I. Fragestellung (1/3)

- Schreibe die Frage explizit auf
- Erkläre die Analyseidee verständlich

I. Fragestellung (2/3)

Frage

- Welche Quellcodedateien sind besonders komplex und änderten sich in letzter Zeit häufig?

I. Fragestellung (3/3)

Umsetzungsideen

- Werkzeuge: Jupyter, Python, pandas, matplotlib
- Heuristiken:
 - "komplex": viele Quellcodezeilen
 - "ändert ... häufig": hohe Anzahl Commits
 - "in letzter Zeit": letzte 90 Tage

Meta-Ziel: Grundmechaniken kennenlernen.

II. Explorative Datenanalyse

- Finde und lade mögliche Softwaredaten
- Berechne und filtere die Rohdaten

Wir laden einen Datenexport aus einem Git-Repository.

```
In [4]: log = pd.read_csv("../dataset/git_log_intellij.csv.gz")
log.head()
```

Out[4]:

	additions	deletions	filename	sha	timestamp	author
0	4	0	java/java-impl/src/com/intellij/codeInsight/hi...	be6247932aa9	2019-07-01 14:40:24	Roman.Ivanov
1	4	6	java/java-impl/src/com/intellij/codeInsight/hi...	ee2032b77eca	2019-07-01 11:36:14	Roman.Ivanov
2	0	3	java/java-impl/src/META-INF/JavaPlugin.xml	fbeb4d639dc1	2019-06-26 11:14:34	Roman.Ivanov
3	6	3	java/java-impl/src/com/intellij/codeInsight/hi...	fbeb4d639dc1	2019-06-26 11:14:34	Roman.Ivanov
4	11	6	java/java-impl/src/com/intellij/codeInsight/hi...	a3d5a9b855fe	2019-06-14 10:32:15	Roman.Ivanov

Wir sehen uns Basisinfos über den Datensatz an.

In [5]: `log.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1128819 entries, 0 to 1128818
Data columns (total 6 columns):
additions      1128819 non-null object
deletions      1128819 non-null object
filename       1128819 non-null object
sha            1128819 non-null object
timestamp      1128819 non-null object
author         1128819 non-null object
dtypes: object(6)
memory usage: 51.7+ MB
```

1 DataFrame (~ programmierbares Excel-Arbeitsblatt), **6 Series** (= Spalten),
1128819 entries (= Reihen)

Wir wandeln die Zeitstempel von Texte in Objekte um.

```
In [8]: log['timestamp'] = pd.to_datetime(log['timestamp'])  
log.head()
```

Out[8]:

	additions	deletions	filename	sha	timestamp	author
0	4	0	java/java-impl/src/com/intellij/codeInsight/hi...	be6247932aa9	2019-07-01 14:40:24	Roman.Ivanov
1	4	6	java/java-impl/src/com/intellij/codeInsight/hi...	ee2032b77eca	2019-07-01 11:36:14	Roman.Ivanov
2	0	3	java/java-impl/src/META-INF/JavaPlugin.xml	fbeb4d639dc1	2019-06-26 11:14:34	Roman.Ivanov
3	6	3	java/java-impl/src/com/intellij/codeInsight/hi...	fbeb4d639dc1	2019-06-26 11:14:34	Roman.Ivanov
4	11	6	java/java-impl/src/com/intellij/codeInsight/hi...	a3d5a9b855fe	2019-06-14 10:32:15	Roman.Ivanov

Wir sehen uns nur die jüngsten Änderungen an.

```
In [13]: recent = log[log['timestamp'] > log['timestamp'].max() - pd.Timedelta("90 days")]
recent.head()
```

Out[13]:

	additions	deletions	filename	sha	timestamp	author
0	4	0	java/java-impl/src/com/intellij/codeInsight/hi...	be6247932aa9	2019-07-01 14:40:24	Roman.Ivanov
1	4	6	java/java-impl/src/com/intellij/codeInsight/hi...	ee2032b77eca	2019-07-01 11:36:14	Roman.Ivanov
2	0	3	java/java-impl/src/META-INF/JavaPlugin.xml	fbeb4d639dc1	2019-06-26 11:14:34	Roman.Ivanov
3	6	3	java/java-impl/src/com/intellij/codeInsight/hi...	fbeb4d639dc1	2019-06-26 11:14:34	Roman.Ivanov
4	11	6	java/java-impl/src/com/intellij/codeInsight/hi...	a3d5a9b855fe	2019-06-14 10:32:15	Roman.Ivanov

Wir wollen nur Java-Code verwenden.

```
In [16]: java = recent[recent['filename'].str.endswith(".java")]
         java.head()
```

Out[16]:

		additions	deletions	filename	sha	timestamp	author
27	1	10		platform/smRunner/src/com/intellij/execution/t...	f4ed78c8f574	2019-06-28 18:28:41	Ilya.Kazakevich
29	28	0		platform/smRunner/testSrc/com/intellij/executi...	f4ed78c8f574	2019-06-28 18:28:41	Ilya.Kazakevich
30	6	3		plugins/InspectionGadgets/InspectionGadgetsAna...	a724467ad1a5	2019-07-01 19:47:38	Roman Shevchenko
31	2	2		plugins/InspectionGadgets/test/com/siyeh/igfix...	a724467ad1a5	2019-07-01 19:47:38	Roman Shevchenko
32	2	2		plugins/InspectionGadgets/test/com/siyeh/igfix...	a724467ad1a5	2019-07-01 19:47:38	Roman Shevchenko

III. Formale Modellierung

- Schaffe neue Sichten
- Verschneide weitere Daten

Wir zählen die Anzahl der Änderungen je Datei.

```
In [21]: changes = java.groupby('filename')[['sha']].count()  
changes.head()
```

Out[21]:

	sha
filename	
RegExpSupport/gen/org/intellij/lang/regexp/_RegExLexer.java	1
RegExpSupport/src/org/intellij/lang/regexp/RegExpCapability.java	1
RegExpSupport/src/org/intellij/lang/regexp/RegExpFileType.java	5
RegExpSupport/src/org/intellij/lang/regexp/RegExpLanguageHost.java	16
RegExpSupport/src/org/intellij/lang/regexp/RegExpLanguageHosts.java	15

Wir holen Infos über die Code-Zeilen hinzu...

```
In [24]: cloc = pd.read_csv("../dataset/cloc_intellij.csv.gz", index_col=1)
cloc.head()
```

Out[24]:

	language	blank	comment	code
filename				
java/java-tests/testData/psi/resolve/ThinletBig.java	Java	299	1140	20125
java/java-tests/testData/psi/parser-full/declarationParsing/class/LongClass.java	Java	10121	10164	10166
python/gen/com/jetbrains/python/console/protocol/PythonConsoleBackendService.java	Java	1971	591	10086
jps/jps-builders/src/org/jetbrains/jps/api/CmdlineRemoteProto.java	Java	502	3066	8605
plugins/java-decompiler/engine/testData/obfuscated/aj.java	Java	551	1	8043

...und verschneiden diese mit den vorhandenen Daten.

```
In [37]: hotspots = changes.join(cloc)[['sha', 'code']].dropna()  
hotspots.head()
```

Out[37]:

	sha	code
filename		
RegExpSupport/gen/org/intellij/lang/regexp/_RegExLexer.java	1	1190.0
RegExpSupport/src/org/intellij/lang/regexp/RegExpCapability.java	1	34.0
RegExpSupport/src/org/intellij/lang/regexp/RegExpFileType.java	5	40.0
RegExpSupport/src/org/intellij/lang/regexp/RegExpLanguageHost.java	16	92.0
RegExpSupport/src/org/intellij/lang/regexp/RegExpLanguageHosts.java	15	168.0

VI. Interpretation

- Erarbeite das Kernergebnis der Analyse heraus
- Mache die zentrale Botschaft / neuen Erkenntnisse deutlich

Wir zeigen nur die TOP 10 Hotspots im Code an.

```
In [39]: top10 = hotspots.sort_values(by="sha", ascending=False).head(10)
top10
```

Out[39]:

	sha	code
filename		
platform/structuralsearch/source/com/intellij/structuralsearch/plugin/ui/StructuralSearchDialog.java	83	1097.0
platform/platform-impl/src/com/intellij/idea/IdeaApplication.java	71	366.0
platform/core-impl/src/com/intellij/ide/plugins/PluginManagerCore.java	67	1425.0
platform/platform-impl/src/com/intellij/openapi/project/impl/ProjectManagerImpl.java	65	810.0
platform/lang-impl/src/com/intellij/build/BuildTreeConsoleView.java	63	936.0
java/java-analysis-impl/src/com/intellij/codeInspection/dataFlow/TrackingRunner.java	63	1300.0
platform/platform-impl/src/com/intellij/idea/StartupUtil.java	62	551.0
platform/platform-impl/src/com/intellij/ide/plugins/PluginManagerConfigurableNewLayout.java	60	1284.0
platform/platform-impl/src/com/intellij/ide/ui/laf/darcula/ui/DarculaComboBoxUI.java	59	606.0
platform/platform-impl/src/com/intellij/openapi/project/impl/ProjectImpl.java	57	299.0

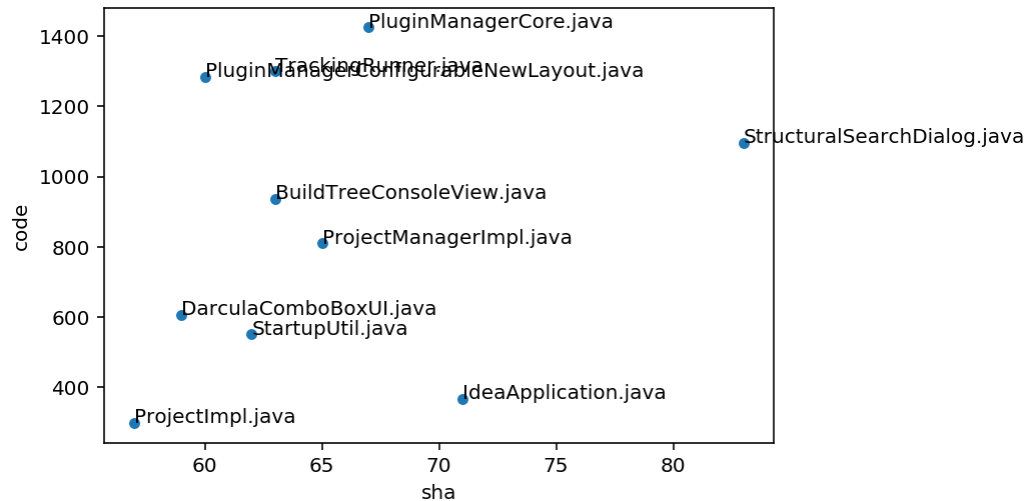
V. Kommunikation

- Transformiere die Erkenntnisse in eine verständliche Visualisierung
- Kommuniziere die nächsten Schritte nach der Analyse

Wir erzeugen ein XY-Diagramm aus der TOP 10 Liste.

```
In [45]: # %load ugly_plotting_code.py
ax = top10.plot.scatter('sha', 'code');

for k, v in top10.iterrows():
    ax.annotate(k.split("/")[-1], v)
```



Ende der Demo

Weitere Analysebeispiele

- Analyse der CPU-Auslastung mit Hilfe von `vmstat`
- Abhängigkeitsanalyse mit `jdeps` und `D3`
- Analyse der Änderungskohäsion mit `Git`
- Analyse der Community-Aktivitäten mittels Daten von `StackOverflow`
- Prüfung der Architektur-Governance-Vorgaben mit `jQAssistant`

Buchempfehlungen

- Adam Tornhill: Software Design X-Ray
- Janelle Arty Starr: Idea Flow
- Tim Menzies, Laurie Williams, Thomas Zimmermann: Perspectives on Data Science for Software Engineering
- Jeff Leek: The Elements of Data Analytic Style
- Wes McKinney: Python For Data Analysis

Meine Empfehlungen zum Einstieg

Meine TOP 5's

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

Kurse, Videos, Blogs, Bücher und mehr (im Aufbau).

Mini-Tutorial

<https://github.com/feststelltaste/software-analytics-workshop>

Zusammenfassung

1. Software Analytics mit Data-Science-Werkzeugen ist 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!

Danke! Fragen?

Markus Harrer

innoQ Deutschland GmbH

markus.harrer@innoq.com

@feststelltaste