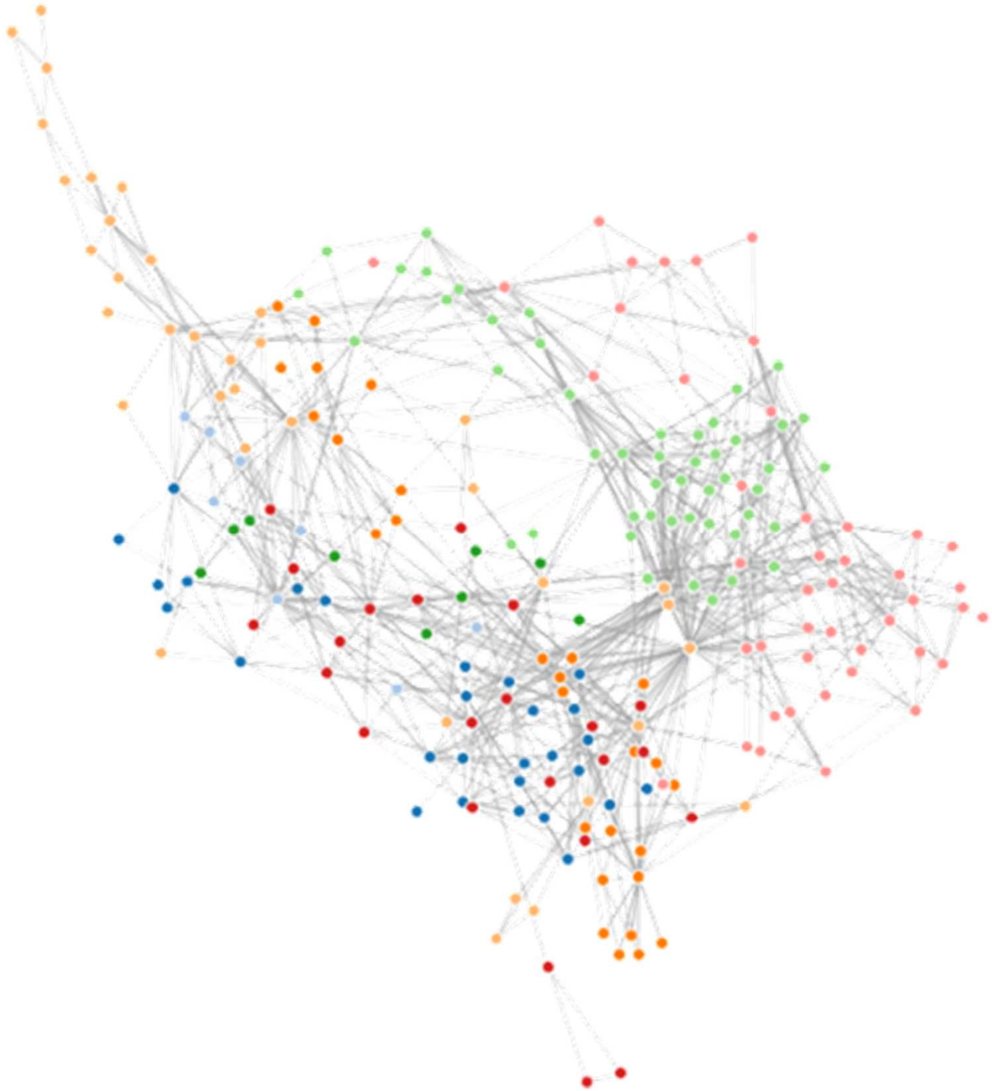


INNOQ TECHNOLOGY LUNCH, 21.10.2020

# Software-Systeme datengetrieben analysieren

**Markus Harrer**
Software Development Analyst

Twitter: @feststelltaste
Blog: feststelltaste.de

INNOQ

THE ORIGINAL HORROR SHOW

# LEGACY SYSTEMS

IMAGE: SETHJREID/PIXABAY

# Das Problem mit Legacy Systeme

Fachliche Features

Architektur

Erkannte Fehler

Strange Things

Nach https://pkruchten.files.wordpress.com/2013/12/kruchten-colours-yow-sydney.pdf

MANAGEMENT

# Kommunikationsbarriere

ENTWICKLUNG

# SHIPPING APP

LCARS 40274

02-654598

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2385 | 8578232 | 9 | 5789 | 3882 | 5893 | 9885 | 3489 | 3465 | 0846 | 9798 | 9629 | 29 |
| 2064 | 2064962 | 7 | 9776 | 626 | 1276 | 7612 | 126 | 97 | 6165 | 6626 | 876 | 74 |
| 34 | 279 | | 89 | 6589 | 6547 | 6587 | 3465 | 867 | 2347 | 5762 | 4588 | 05 |
| 4768 | 8967248 | 7 | 9798 | 8969 | 476 | 9047 | 8476 | 9749 | 0982 | 8969 | 0247 | 89 |
| 685 | 3478 | 8 | 867 | 346 | 34 | 48 | 49 | 8 | 89 | 897 | 38 | |
| 757 | 898990 | 8 | 200 | 285 | 923 | 9 | 387 | 238 | 578 | 875 | 87 | 9 |
| 484 | 947589 | 7 | 569 | 68 | 678 | 893 | 56 | 584 | 678 | 476 | 458 | 4 |

9886-234

0128-069

1014-819

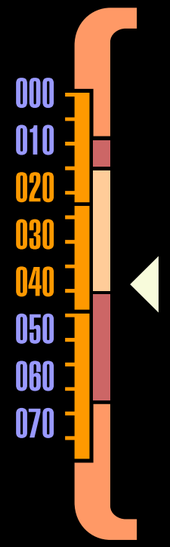
7232-838

03-975683

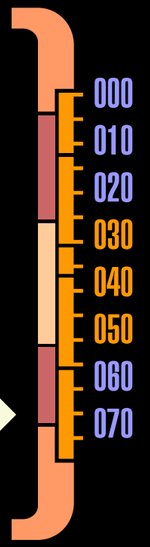
04-765466

05-224353

06-576565




DEVELOPER'S INSANITY




PROJECT'S BUDGET

# THE ULTIMATE QUALITY DASHBOARD

# THE EMPIRIC STRIKES BACK

Not a long time ago, from brains not far, far away....

# SOFTWARE ANALYTICS

# SOFTWARE ANALYTICS

A definition of

*MENZIES* & ZIMMERMANN

Software Analytics

*is analytics* on software data for

*managers* and software engineers

*with the aim* of empowering

*software* development individuals

and teams

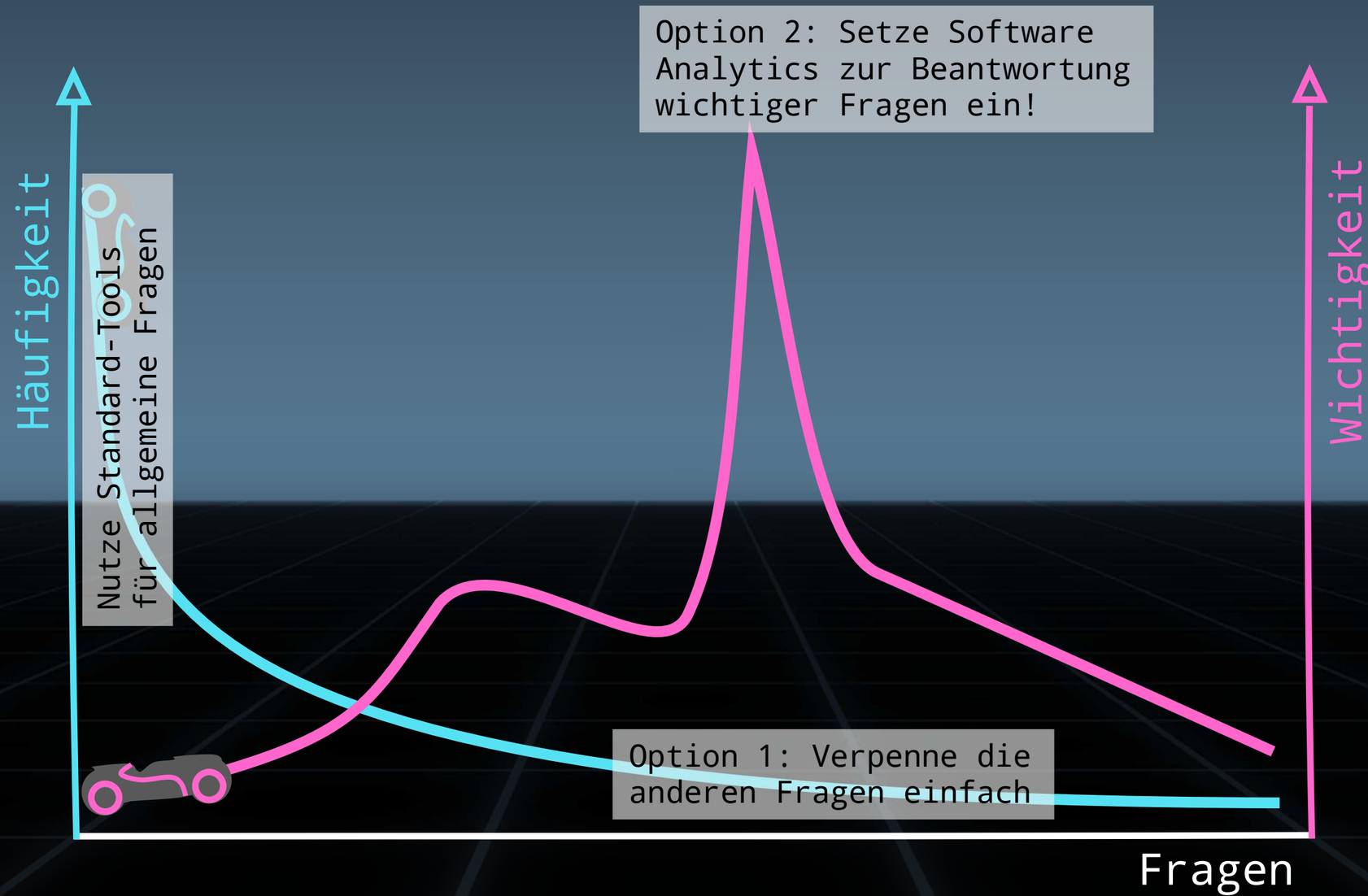
*to gain and* share insight from

*their data* to make better

*decisions.*


Statische Daten
Laufzeit-daten
DIE SOFTWAREDATEN VON
1 TEAM
VON ENTWICKLERN
Chronolo-gische Daten
Daten der Community

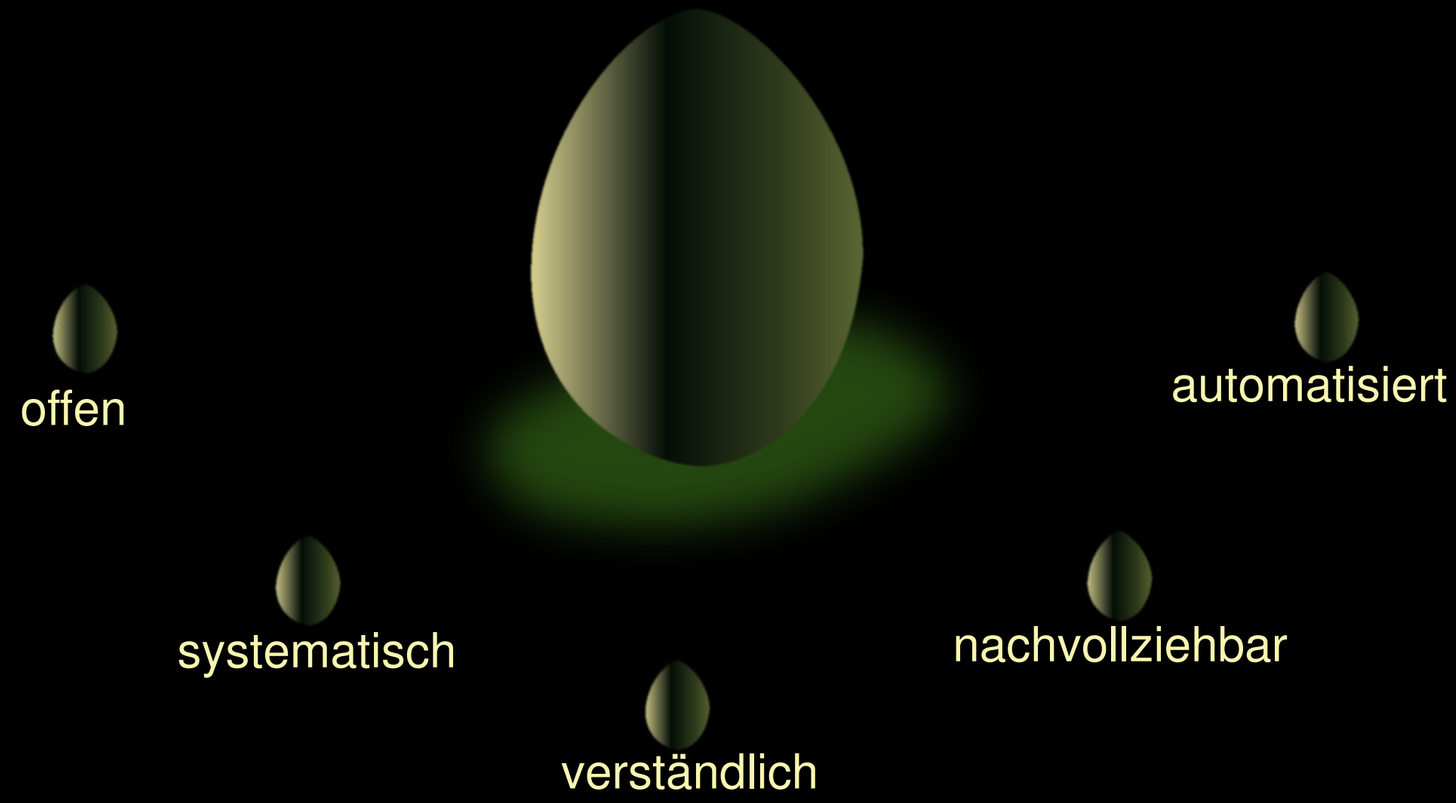
# Software Analytics: Keine 08/15 Fragen



# Beispiele von Datenanalysen zum Terminieren von Problemen

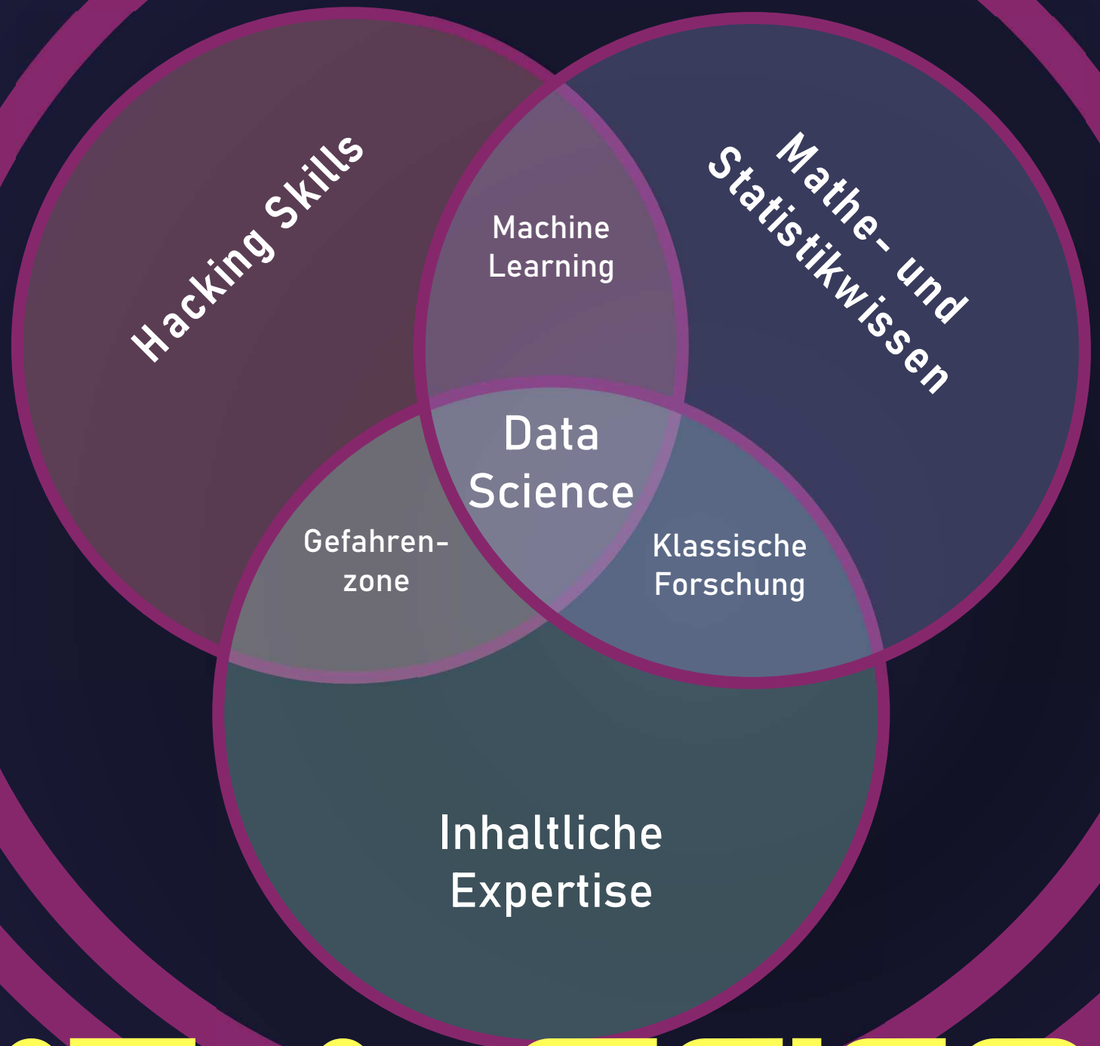
- Quantifizierung des Wissensverlusts bei Entwicklerfluktuation
- Verprobung von Modularisierungsvarianten (“virtuelles Refactoring”)
- Ermittlung von Performance-Hotspots über Call-Tree-Analyse
- Tracking von umfangreichen Code-Umbauarbeiten
- Analyse der Community-Aktivitäten um Open-Source-Software
- <weitere, ganz spezifische Analysen in ganz spezifischen Situationen>

# REPRODUCIBLE DATA ANALYSIS




= EIN WEG ZUR UMSETZUNG VON

**SOFTWARE ANALYTICS**




# DATA SCIENCE

Venn Diagramm von Drew Conway


TOOL
TIME

# Python 3

# PANDAS


... and matplotlib, numpy, scikit-learn, NLTK, Pygments, py2neo, requests, BeautifulSoup, Pygal ...

# Computational Notebook

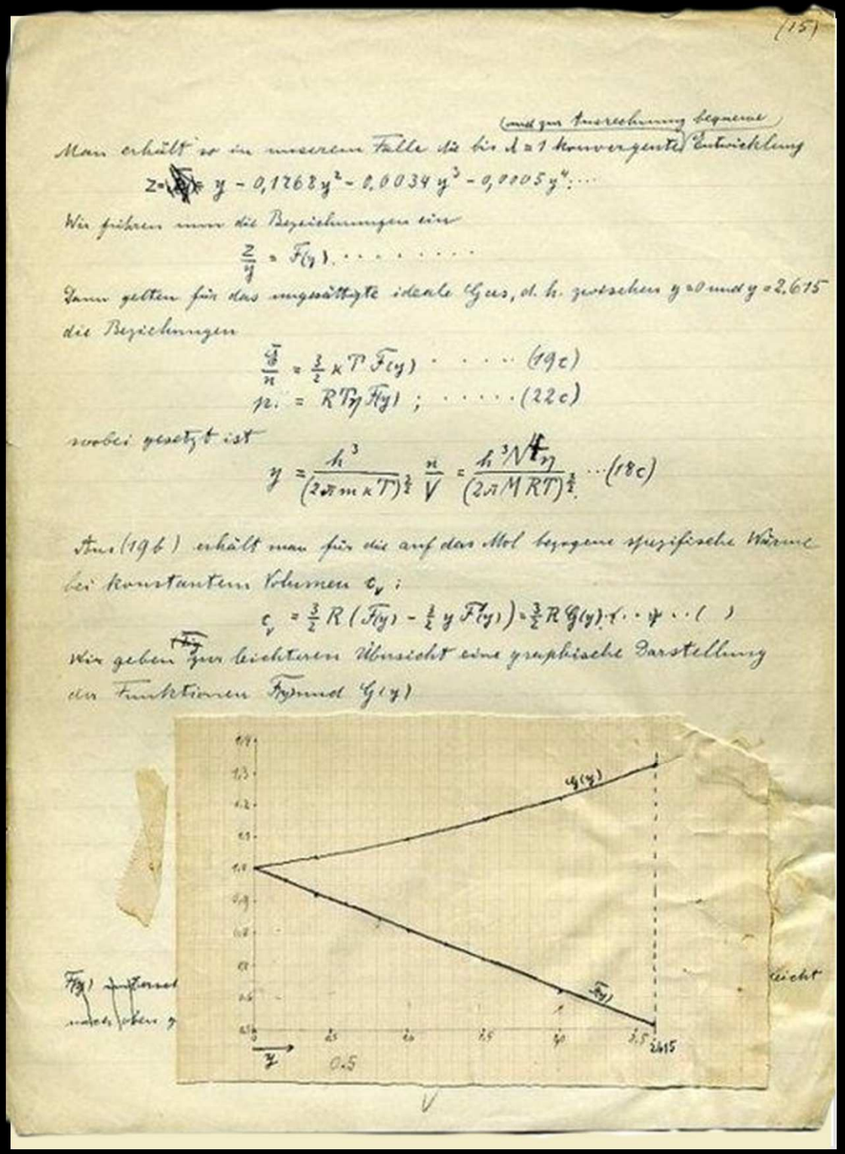
code and data in love

# Computational Notebook

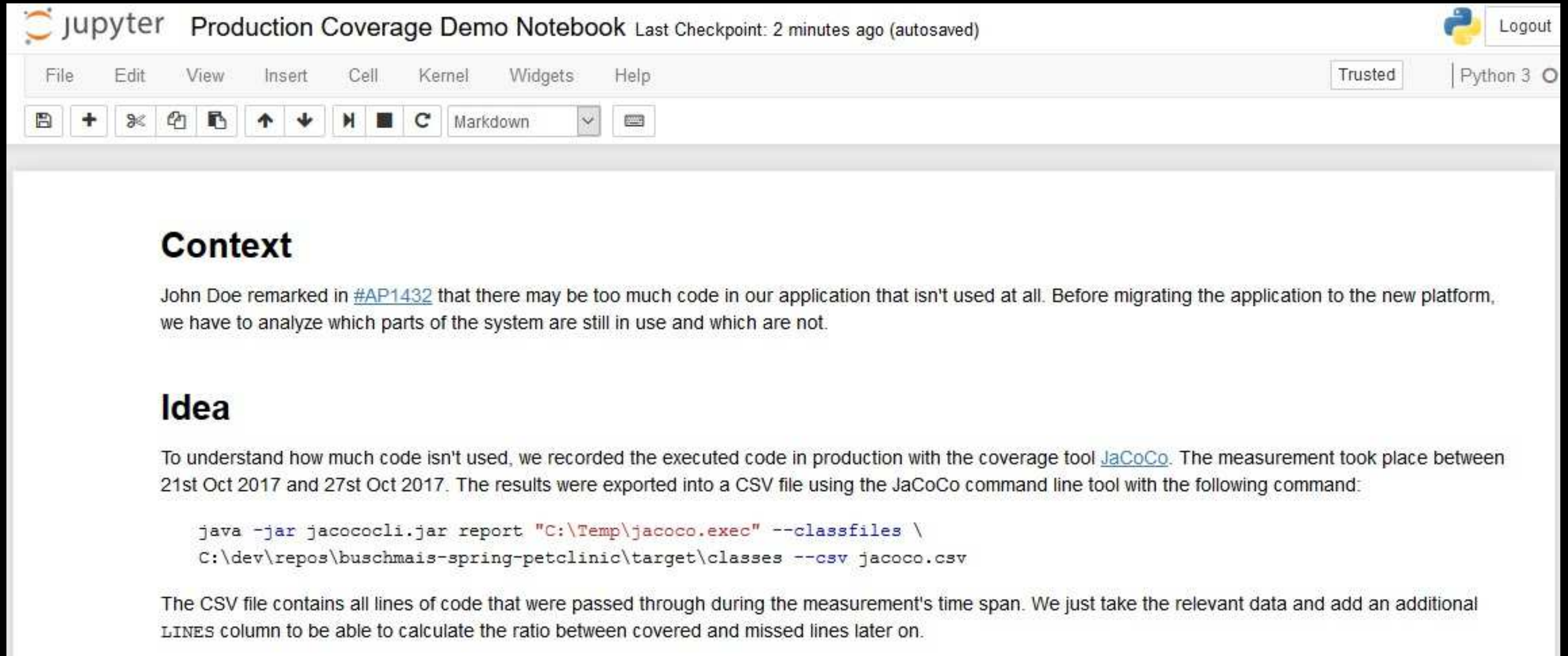
## Jupyter Notebook

- Kontext dokumentiert
- Ideen, Daten, Annahmen und Vereinfachungen aufgeführt
- Berechnungen verständlich dargelegt
- Zusammenfassungen erklärt

- Komplett automatisiert



# Jupyter Notebook



# Context

John Doe remarked in #AP1432 that there may be too much code in our application that isn't used at all. Before migrating the application to the new platform, we have to analyze which parts of the system are still in use and which are not.

# Idea

To understand how much code isn't used, we recorded the executed code in production with the coverage tool JaCoCo. The measurement took place between 21st Oct 2017 and 27st Oct 2017. The results were exported into a CSV file using the JaCoCo command line tool with the following command:

```
java -jar jacococli.jar report "C:\Temp\jacoco.exec" --classfiles \
C:\dev\repos\buschmais-spring-petclinic\target\classes --csv jacoco.csv
```

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

Out[1]:

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

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

Out[1]:

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

# Analysis

It was stated that whole packages wouldn't be needed anymore and that they could be safely removed. Therefore, we sum up the coverage data per class for each package and calculate the coverage ratio for each package.

In [2]:
```
grouped_by_packages = coverage.groupby("PACKAGE").sum()
grouped_by_packages['RATIO'] = grouped_by_packages.LINE_COVERED / grouped_by_packages.LINES
grouped_by_packages = grouped_by_packages.sort_values(by='RATIO')
grouped_by_packages
```

Out[2]:

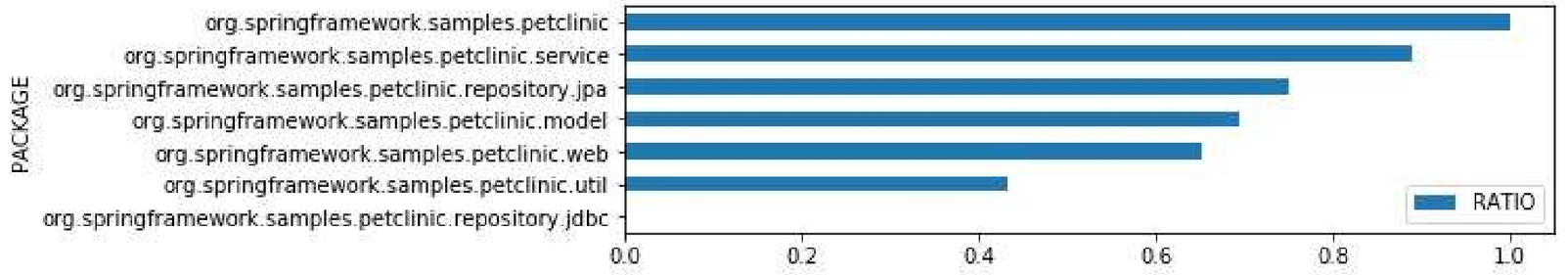
| PACKAGE | LINE_COVERED | LINE_MISSED | LINES | RATIO |
|---|---|---|---|---|
| org.springframework.samples.petclinic.repository.jdbc | 0 | 152 | 152 | 0.000000 |
| org.springframework.samples.petclinic.util | 13 | 17 | 30 | 0.433333 |
| org.springframework.samples.petclinic.web | 75 | 40 | 115 | 0.652174 |
| org.springframework.samples.petclinic.model | 75 | 33 | 108 | 0.694444 |
| org.springframework.samples.petclinic.repository.jpa | 21 | 7 | 28 | 0.750000 |
| org.springframework.samples.petclinic.service | 16 | 2 | 18 | 0.888889 |
| org.springframework.samples.petclinic | 24 | 0 | 24 | 1.000000 |

| | | | | |
|---|---|---|---|---|
| org.springframework.samples.petclinic.web | 75 | 40 | 115 | 0.652174 |
| org.springframework.samples.petclinic.model | 75 | 33 | 108 | 0.694444 |
| org.springframework.samples.petclinic.repository.jpa | 21 | 7 | 28 | 0.750000 |
| org.springframework.samples.petclinic.service | 16 | 2 | 18 | 0.888889 |
| org.springframework.samples.petclinic | 24 | 0 | 24 | 1.000000 |

We plot the data for the coverage ratio to get a brief overview of the result.

In [3]:
```
%matplotlib inline
grouped_by_packages[['RATIO']].plot(kind="barh", figsize=(8,2))
```

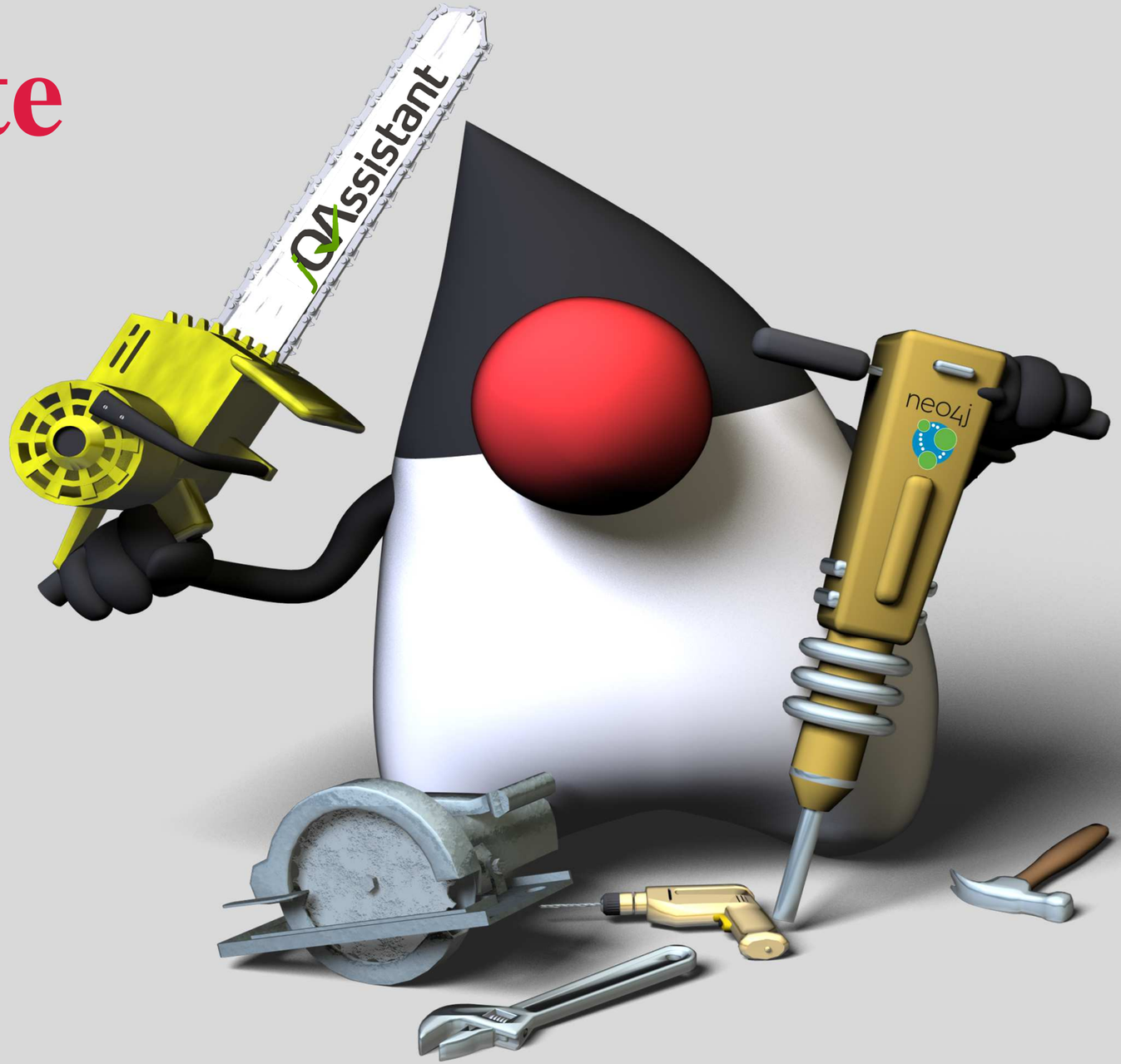
Out[3]: `<matplotlib.axes._subplots.AxesSubplot at 0x1874cdde9e8>`




# Conclusion

The JDBC package `org.springframework.samples.petclinic.repository.jdbc` isn't used at all and can be left out safely when migrating to the new platform.

# Graph-basierte Analyse




https://commons.wikimedia.org/wiki/File:Netbeans-Duke.png

# jQAssistant & Neo4j

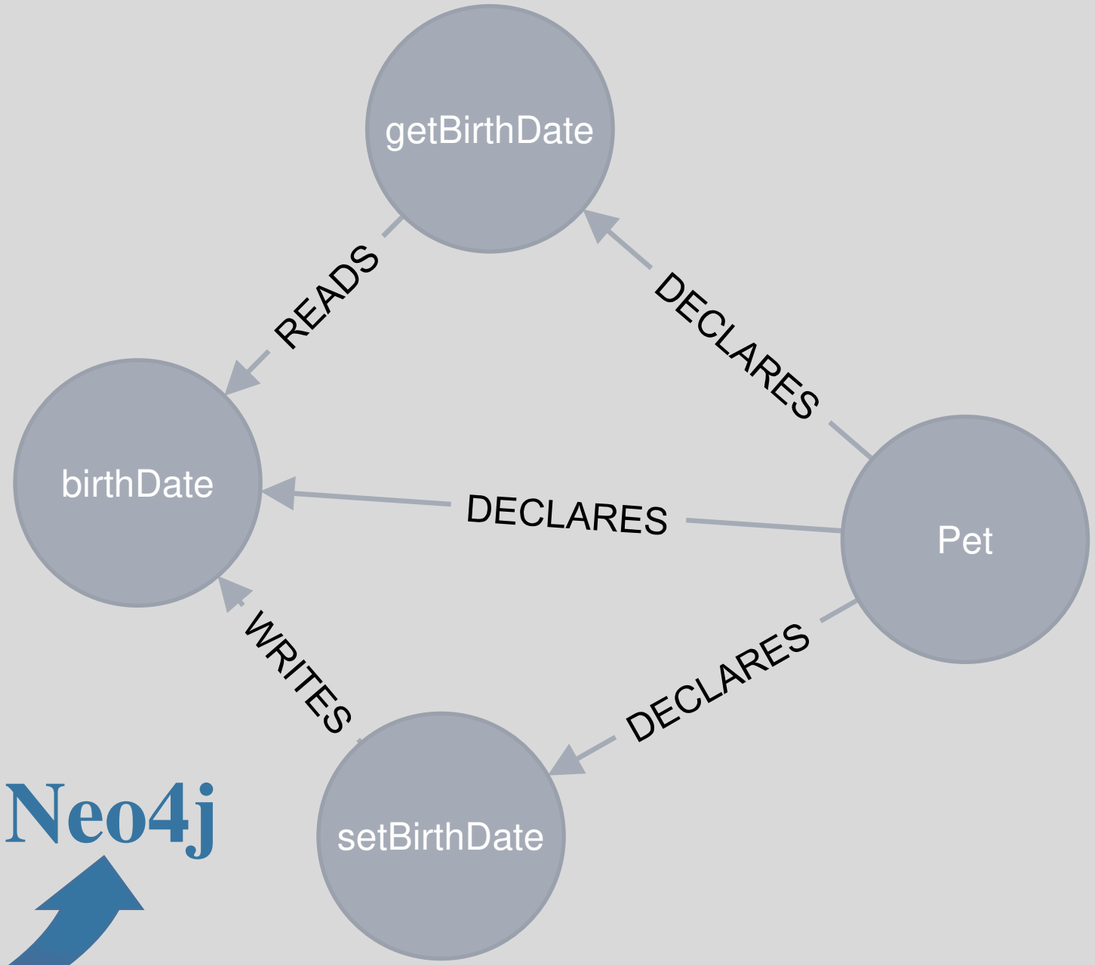
```java
public class Pet {

 private LocalDate birthDate;

 public LocalDate getBirthDate(){
     return this.birthDate;
 }

 public void setBirthDate(LocalDate birthDate){
      this.birthDate = birthDate;
 }
```

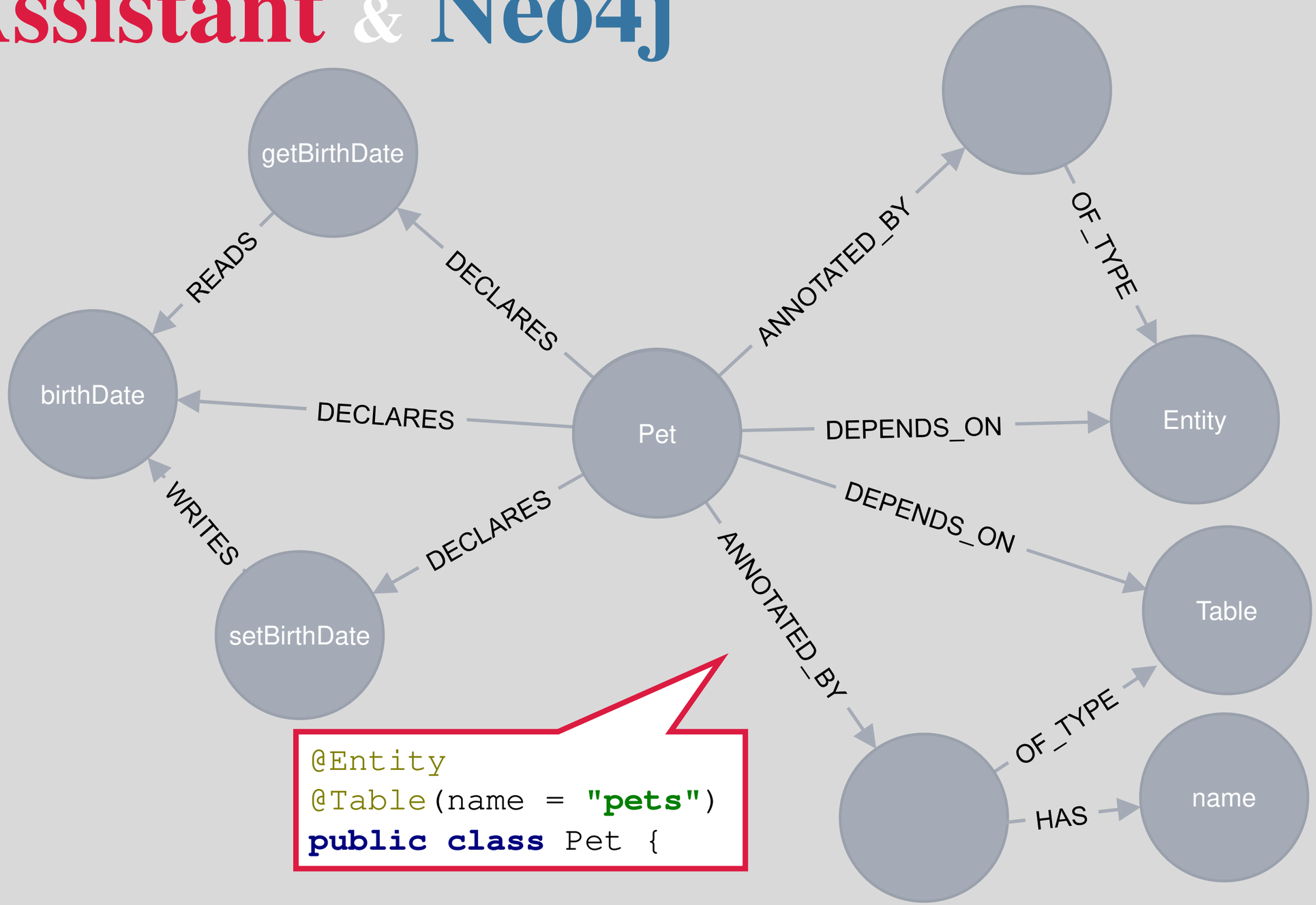
Java Code




Neo4j

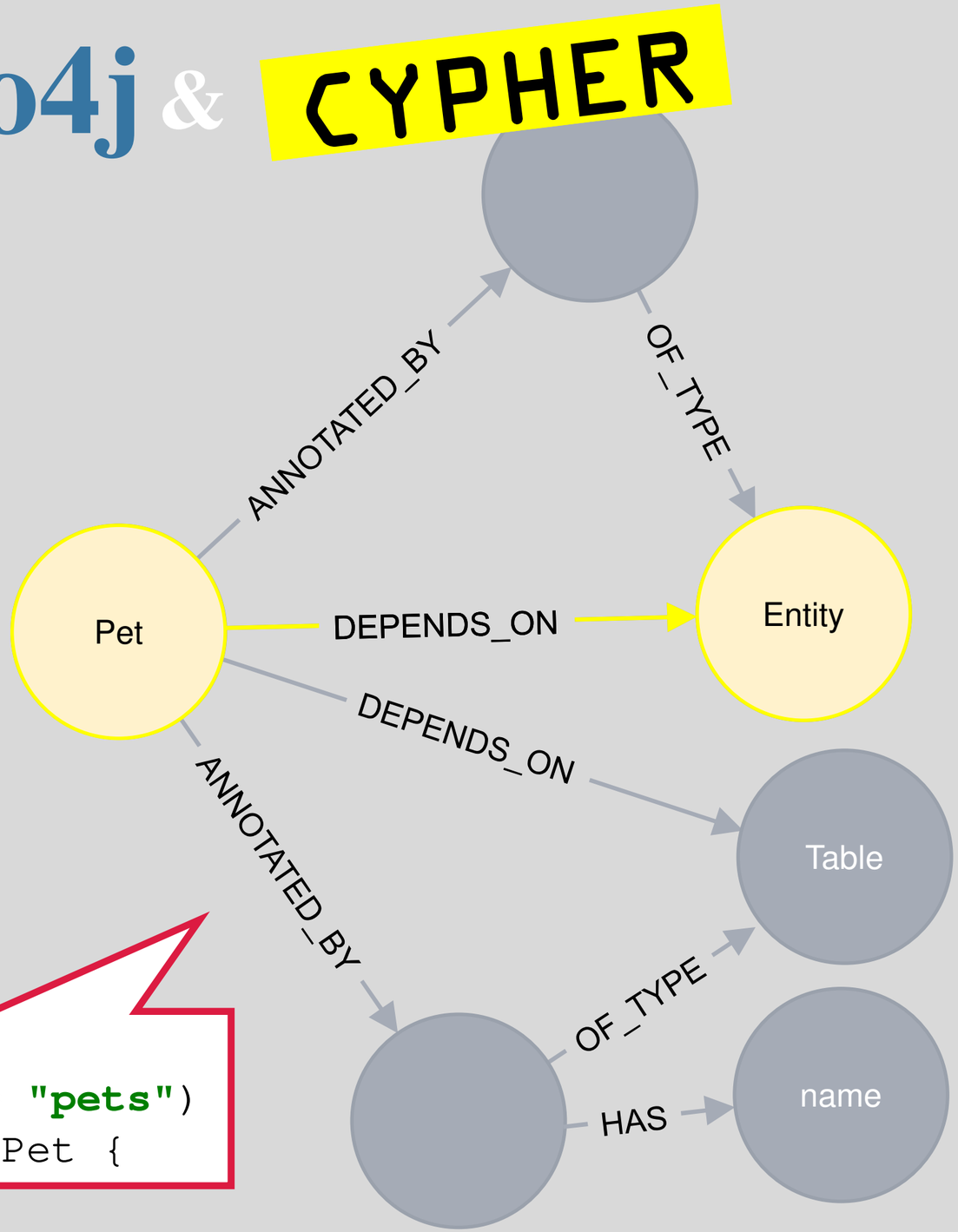
jQAssistant

# jQAssistant & Neo4j



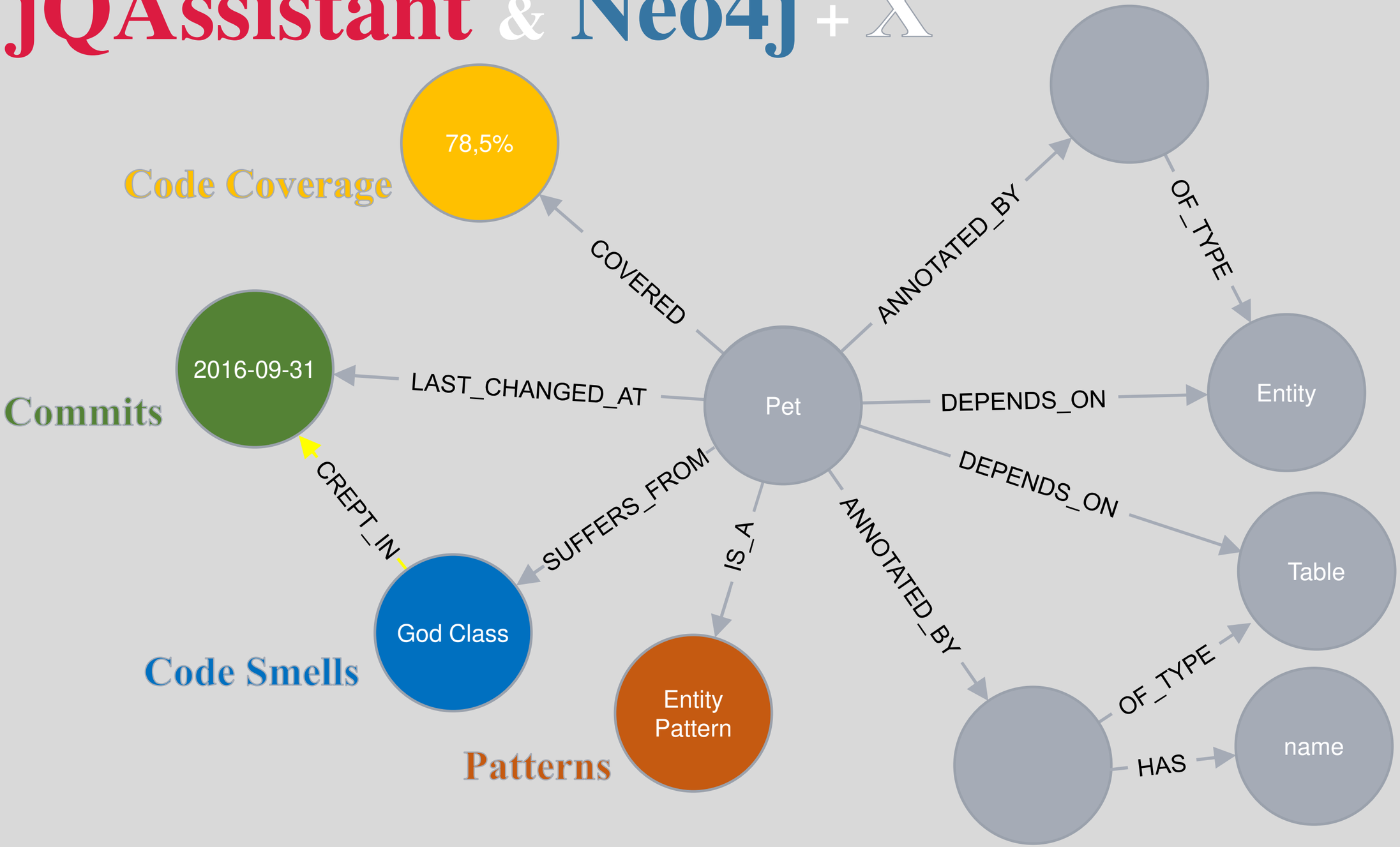
# jQAssistant & Neo4j & CYPHER

```
MATCH
    (p:Class)-[:DEPENDS_ON]->(e:Type)
WHERE
    p.name = "Pet" AND
    e.name = "Entity"
RETURN p, e
```




```
@Entity
@Table(name = "pets")
public class Pet {
```

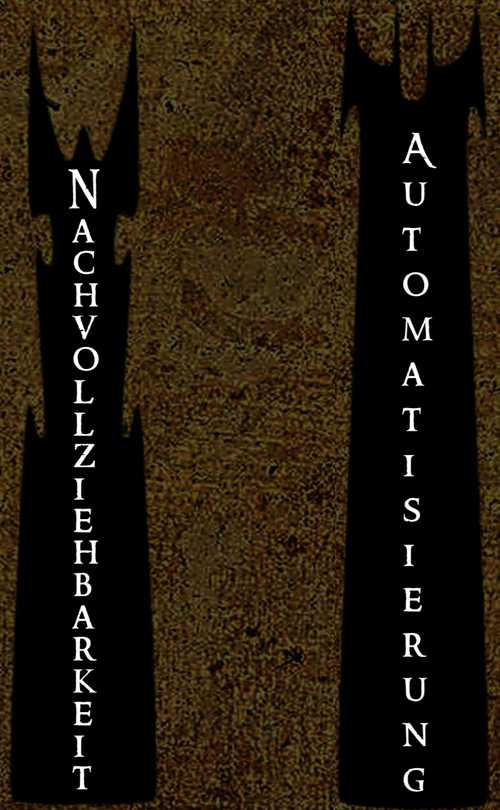
# jQAssistant & Neo4j + X



# Werde wieder Herr der Dinge

## mit datengetriebenen Softwareanalysen


### Die Gefahren




### Die Zwei Tipps


Anzahl der gelösten Probleme

### Die Rückkehr der Vernunft

# QUESTIONS

## ASK ´ EM ALL

# Infos zu den Demos

## Jupyter Notebook, Python, pandas, matplotlib

Repo

https://github.com/feststelltaste/software-analytics/tree/master/demos/20201021_INNOQ_Technology_Lunch

Interaktive Online-Version

https://mybinder.org/v2/gh/feststelltaste/software-analytics/master?filepath=demos%2F20201021_INNOQ_Technology_Lunch%2FRefactoring.ipynb

## jQAssistant & Neo4j

Repo Spring PetClinic

https://github.com/javaonautobahn/spring-petclinic

Repo DesignSmells

https://github.com/feststelltaste/designsmells


Notebook hiermit ausführen

# Einstieg in Software Analytics

**Mein Blog**

https://feststelltaste.de

**TOP 5 Software Analytics**

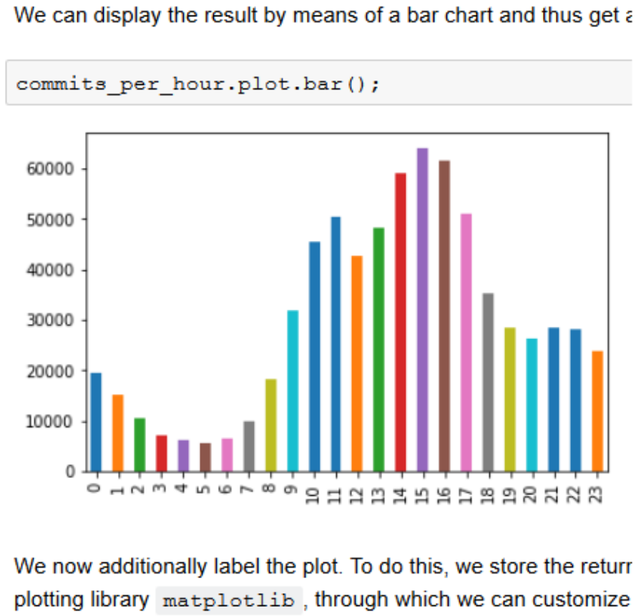
https://www.feststelltaste.de/top-5-software-analytics/

**Mein Software Analytics Repository**

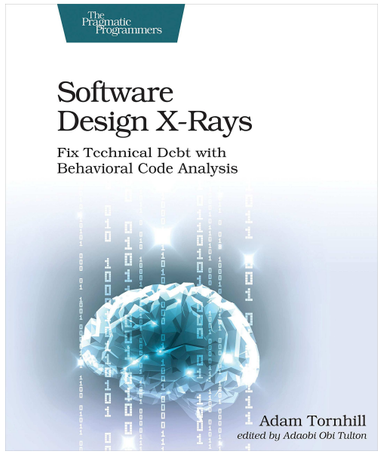
https://github.com/feststelltaste/software-analytics

**Mini-Tutorial und mehr zu Software Analytics**

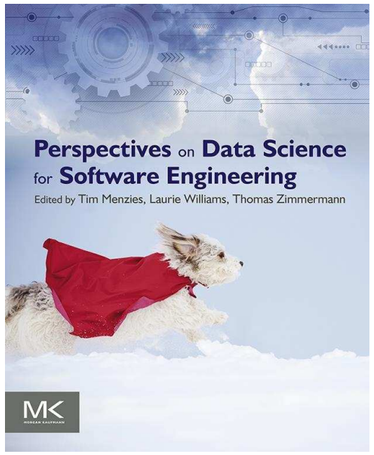
https://github.com/feststelltaste/software-analytics-workshop



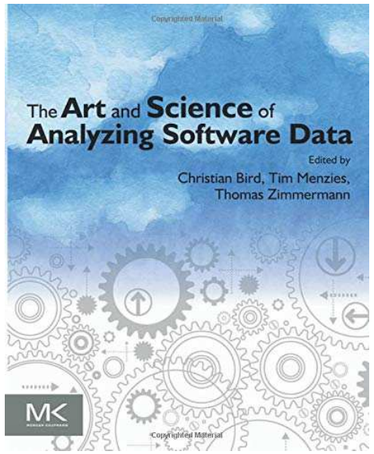
# Mehr Infos zu Software Analytics




Adam Tornhill:
*Software X-Ray*



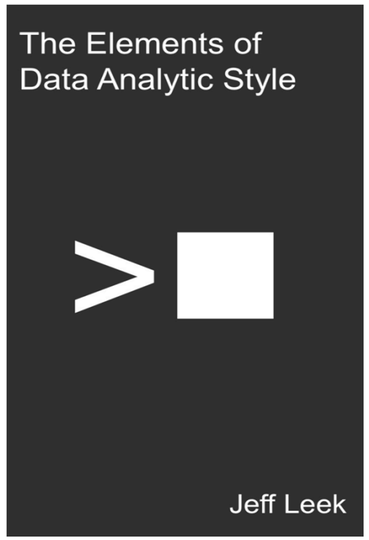

Tim Menzies, Laurie Williams, Thomas Zimmermann:
*Perspectives on Data Science for Software Engineering*




Christian Bird, Tim Menzies, Thomas Zimmermann:
*The Art and Science of Analyzing Software Data*

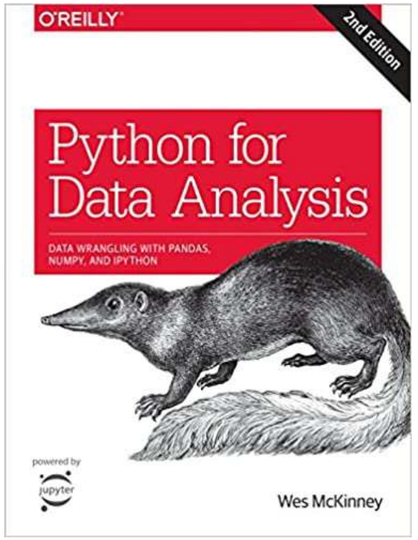
Noch mehr: https://github.com/feststelltaste/awesome-software-analytics

# Infos zu Data Science




Jeff Leek:
*The Elements of Data Analytic Style*


Roger D. Peng:
*Report Writing for Data Science in R*




Wes McKinney:
*Python for Data Analysis*

**Meine TOP5s**

- https://www.feststelltaste.de/top5-jupyter/
- https://www.feststelltaste.de/top5-pandas/
- https://www.feststelltaste.de/top5-python/

# Weitere Details zu jQAssistant/Neo4j

https://easychair.org/publications/preprint/893N

## Towards an Open Source Stack to Create a Unified Data Source for Software Analysis and Visualization

Richard Müller*, Dirk Mahler†, Michael Hunger‡, Jens Nerche§ and Markus Harrer¶
*Leipzig University, Germany
Email: rmueller@wifa.uni-leipzig.de
†buschmais GbR, Dresden, Germany
Email: dirk.mahler@buschmais.com
‡Developer Relations, Neo4j Inc., Malmö, Sweden
Email: michael.hunger@neo4j.com
§Application Development, Kontext E GmbH, Dresden, Germany
Email: j.nerche@kontext-e.de
¶Software Development Analyst, Freelancer, Roth, Germany
Email: contact@markusharrer.de

***Abstract*—The beginning of every software analysis and visualization process is data acquisition. However, there are various sources of data about a software system. The methods used**

Creating, storing, and querying the data captured by such graphs is very challenging. Diehl et al. summarize the most

# Software Analytics Trainings


Öffentliche Termine ab Januar 2021 remote, firmenintern nach Abstimmung

https://www.innoq.com/de/trainings/software-analytics

# Vielen Dank!




www.innoq.com

Markus Harrer
markus.harrer@innoq.com

@feststelltaste

**innoQ Deutschland GmbH**

Krischerstr. 100
40789 Monheim am Rhein
Germany
+49 2173 3366-0

Ohlauer Str. 43
10999 Berlin
Germany

Ludwigstr. 180E
63067 Offenbach
Germany

Kreuzstr. 16
80331 München
Germany

**innoQ Schweiz GmbH**

Gewerbestr. 11
CH-6330 Cham
Switzerland
+41 41 743 01 11

Albulastr. 55
8048 Zürich
Switzerland