

BOBKONF 2024, 15.03.2024

Software Analytics with Data Science on Software Data

Markus Harrer

Software Evolutionist @ INNOQ

Social: [markusharrer.de](https://www.markusharrer.de)

INNOQ

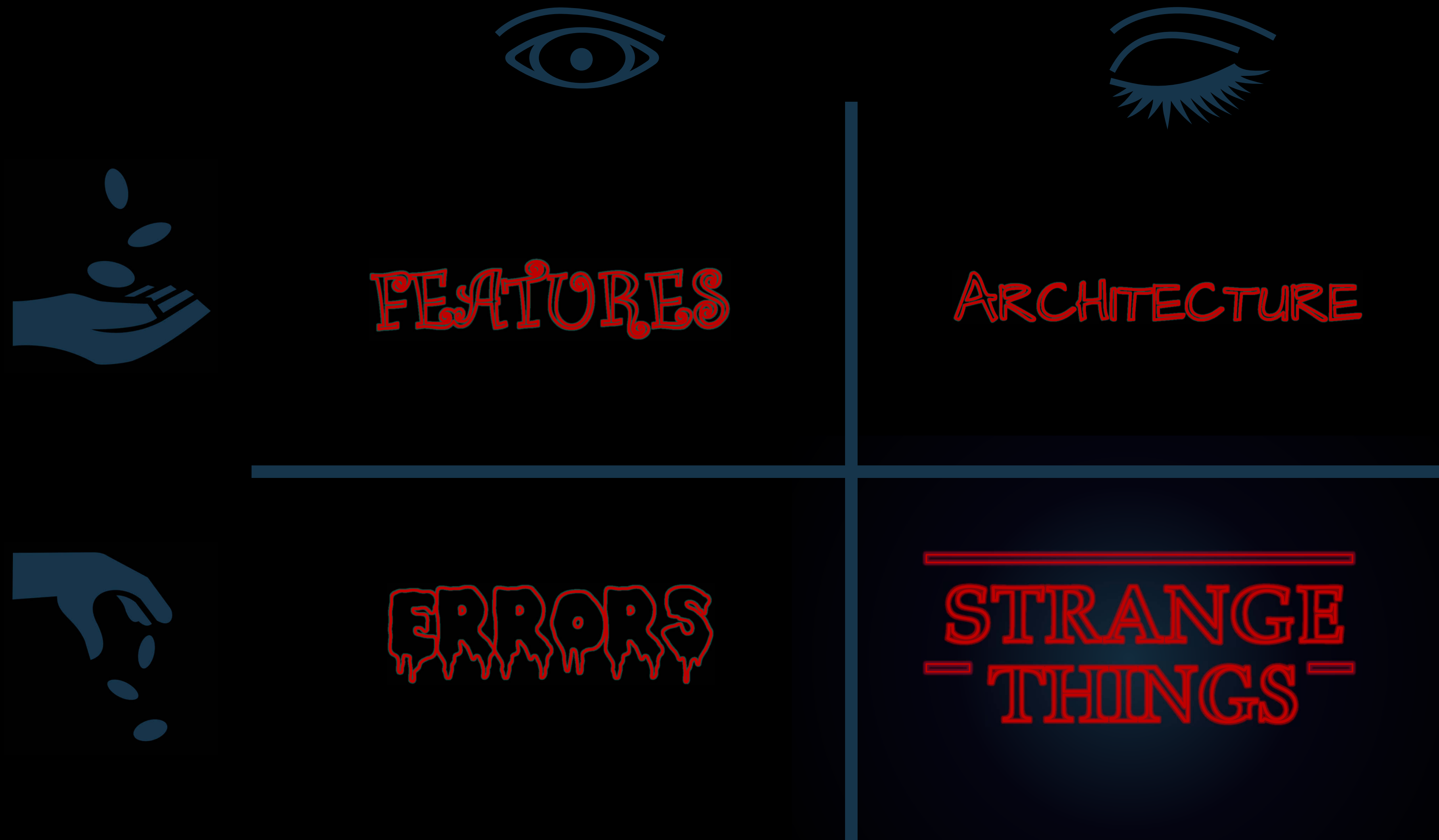


THE ORIGINAL HORROR SHOW



LEGACY SYSTEMS
LEGACY SYSTEMS

THE PROBLEM WITH THE PROBLEMS



MANAGEMENT

COMMUNICATION GAP

DEVELOPMENT

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

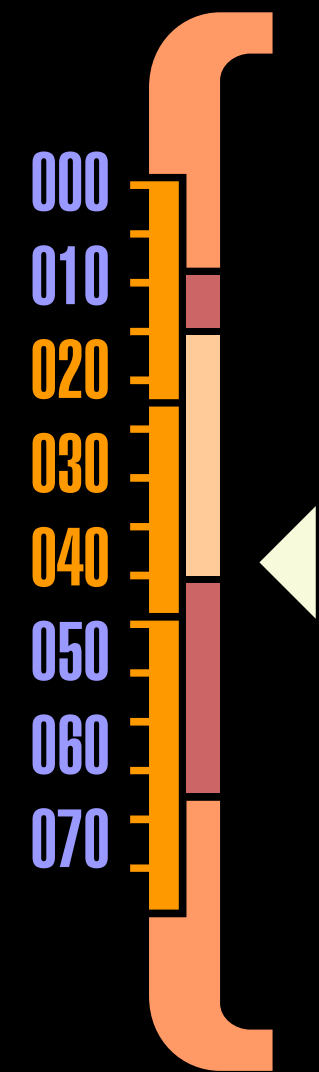
7232-838

03-975683

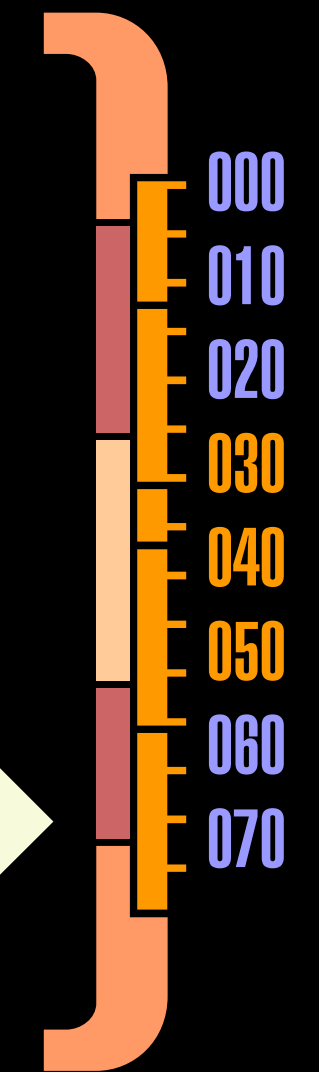
04-765466

05-224353

06-576565



DEVELOPER'S SANITY



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*

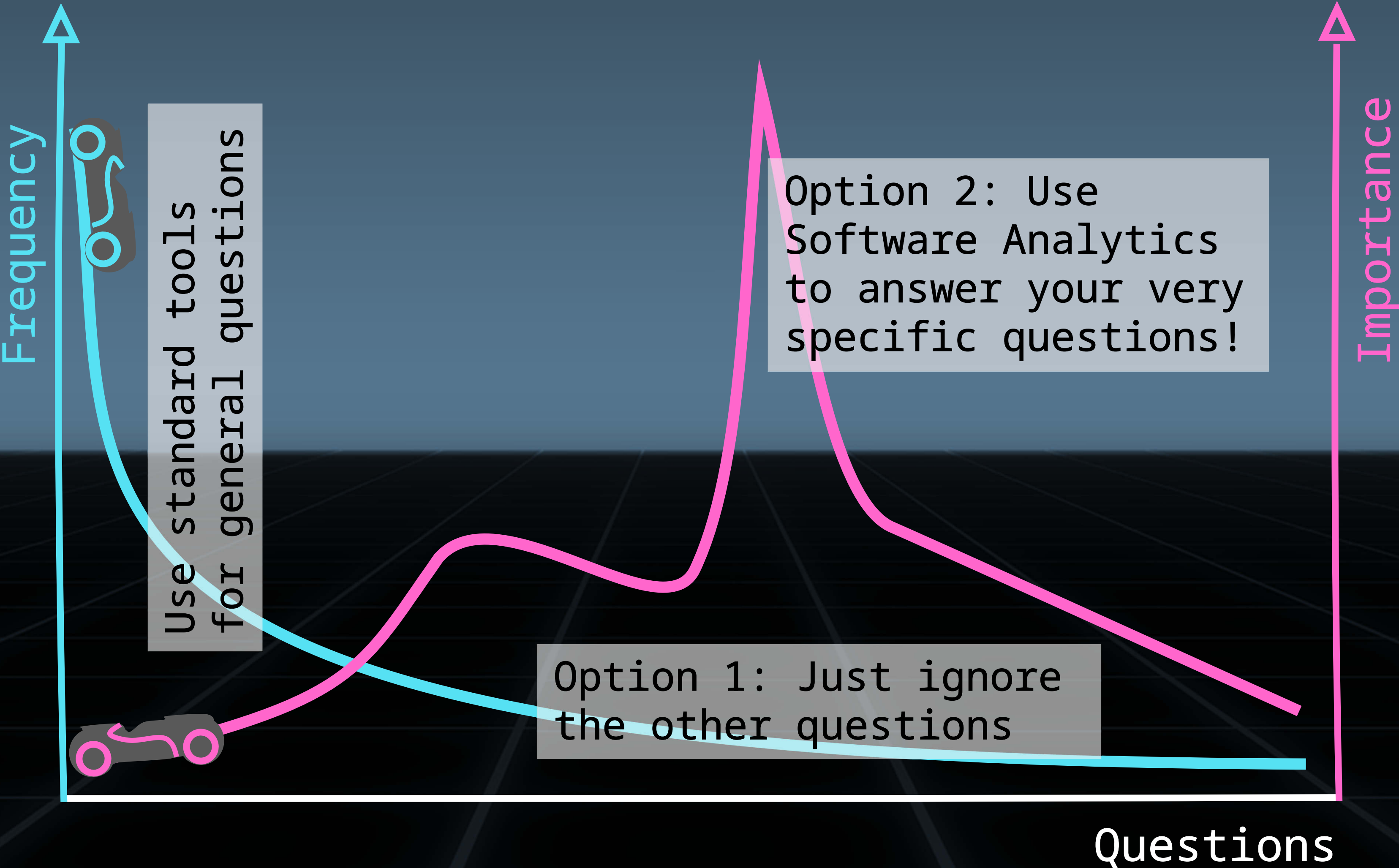
*to gain and share insight from
their data to make better
decisions.*

SOFTWARE
ANALYTICS

EPISODE II

A NEW HOPE

ANSWERING YOUR SPECIFIC QUESTIONS



TYPICAL ISSUES TO TERMINATE

- Spotting parts in source code no one knows of anymore
- Finding root causes of performance bottlenecks
- Identifying alternative modularization options
- Showing the progress of long-living restructurings
- Measuring the community activity around open source software
- <your very specific analysis in your very specific situation>

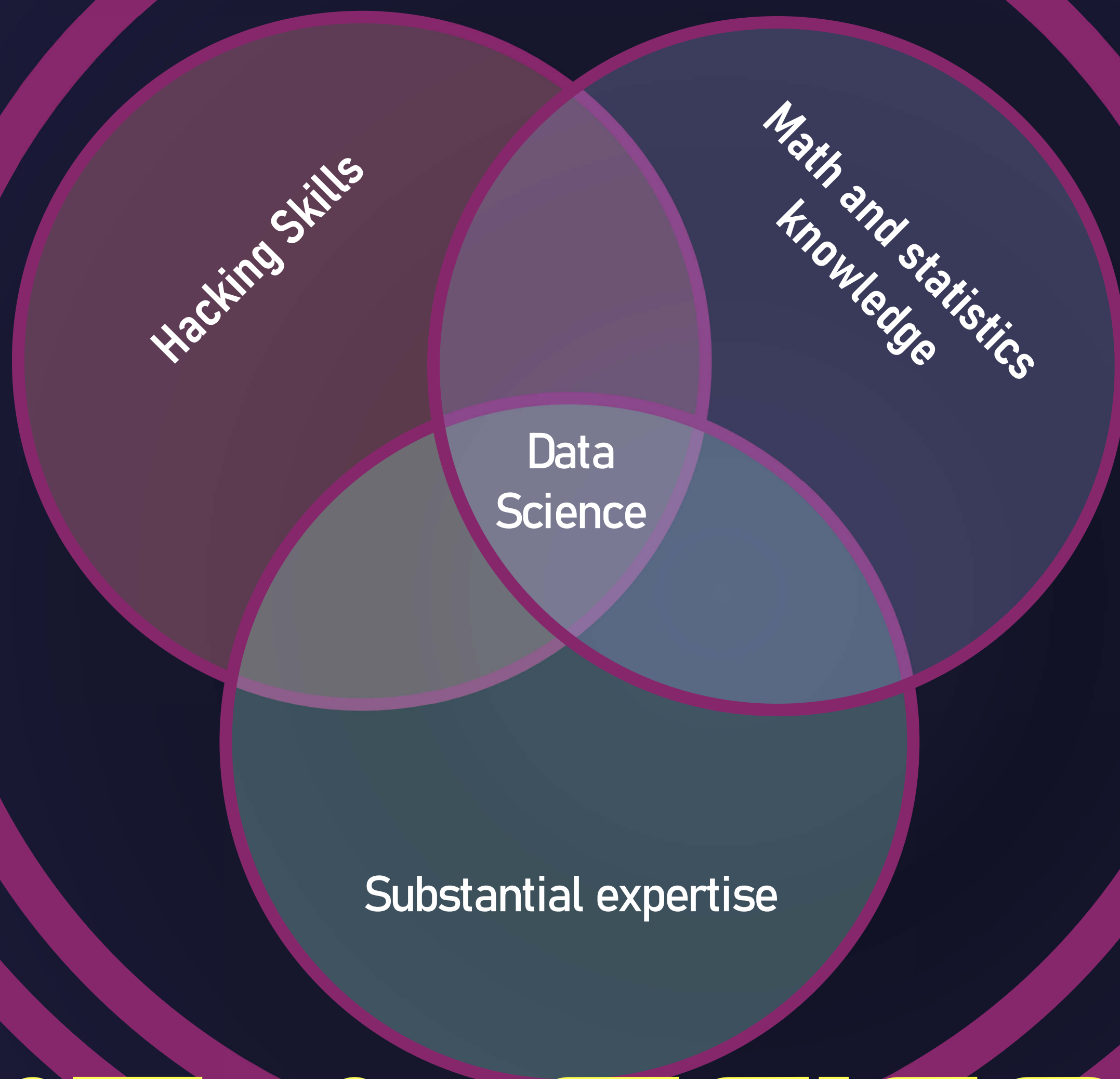


THE SOFTWARE DATA OF

A TEAM

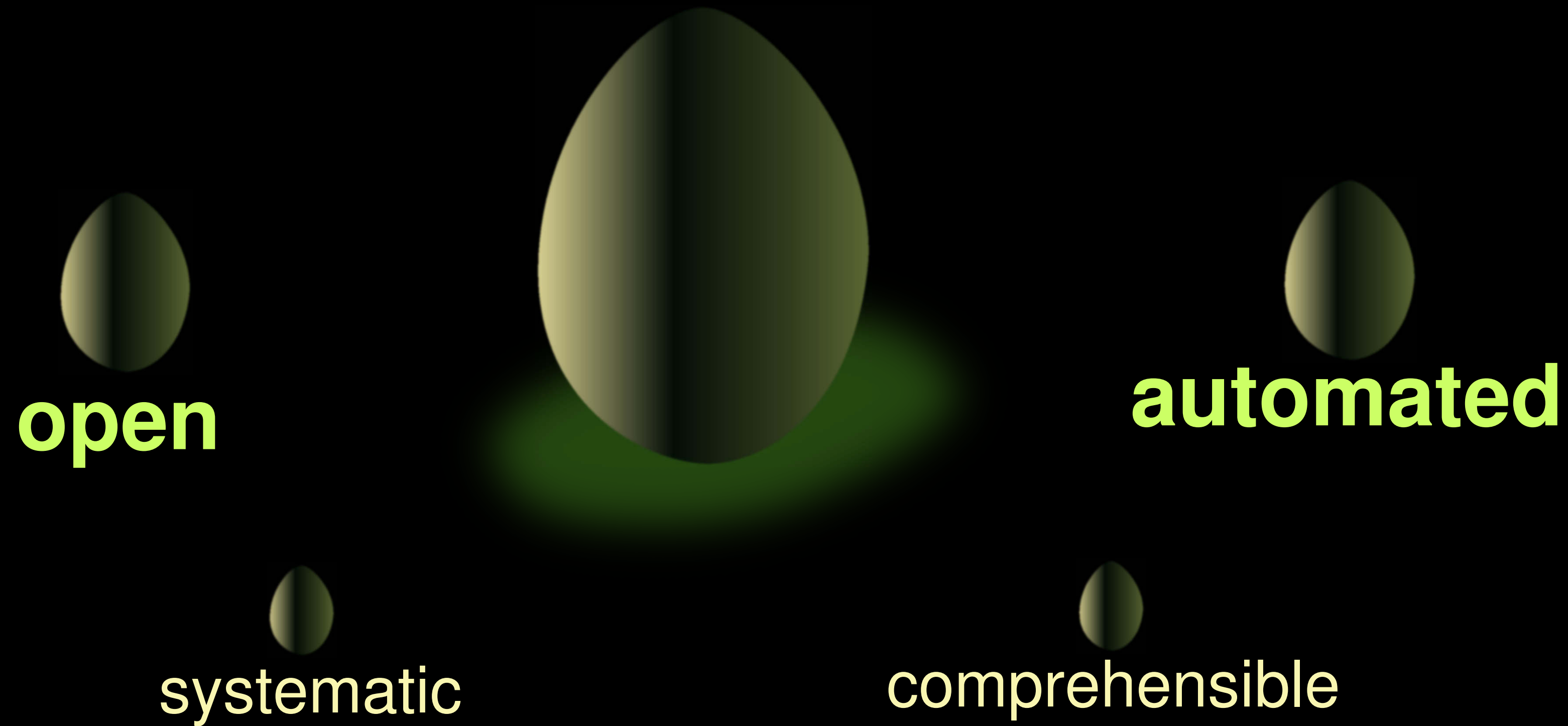
OF DEVELOPERS





DATA SCIENCE

REPRODUCIBLE DATA SCIENCE



= A WAY TO IMPLEMENT SOLID
SOFTWARE ANALYTICS

TOOOL
TIME
S

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

(15)

Man erhält so in unserem Falle die bis $d=1$ konvergente Entwicklung

$$z = y - 0,1768y^2 - 0,0034y^3 - 0,0005y^4, \dots$$

Wir führen nun die Bezeichnungen ein

$$\frac{z}{y} = F(y), \dots$$

Dann gelten für das ungesättigte ideale Gas, d. h. zwischen $y=0$ und $y=2,615$ die Beziehungen

$$\frac{p}{p_0} = \frac{1}{2} \times T \cdot F(y) \dots (19c)$$

$$p_0 = RT_0 F(y); \dots (22c)$$

wobei gesetzt ist

$$y = \frac{h^3}{(2\pi m kT)^{3/2}} \frac{n}{V} = \frac{h^3 N_A}{(2\pi M RT)^{3/2}} \dots (18c)$$

Aus (19b) erhält man für die auf das Mol bezogene spezifische Wärme bei konstantem Volumen c_v :

$$c_v = \frac{3}{2} R (F(y) - \frac{1}{2} y F'(y)) = \frac{3}{2} R G(y) \dots ()$$

Wir geben $F(y)$ zur leichteren Übersicht eine graphische Darstellung der Funktionen $F(y)$ und $G(y)$.

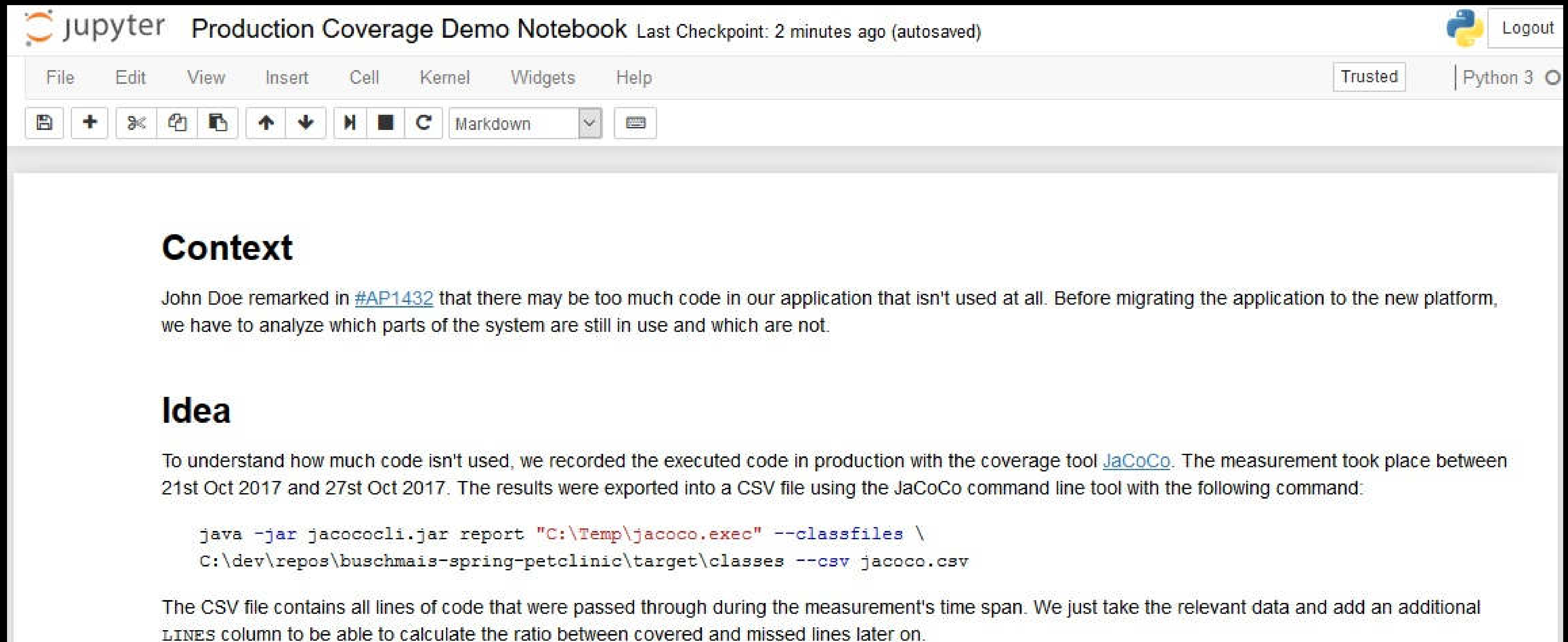
V

nicht

F(y) = ...
nicht

Literate Programming

with Jupyter Notebook



The screenshot shows a Jupyter Notebook interface. At the top, the title bar reads "jupyter Production Coverage Demo Notebook Last Checkpoint: 2 minutes ago (autosaved)". On the right side of the title bar, there is a Python logo and a "Logout" button. Below the title bar is a menu bar with options: File, Edit, View, Insert, Cell, Kernel, Widgets, and Help. To the right of the menu bar, there is a "Trusted" status indicator and a "Python 3" kernel selector. Below the menu bar is a toolbar with various icons for file operations (save, new, open, close), navigation (up, down, home, stop, refresh), and a dropdown menu currently set to "Markdown".

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.

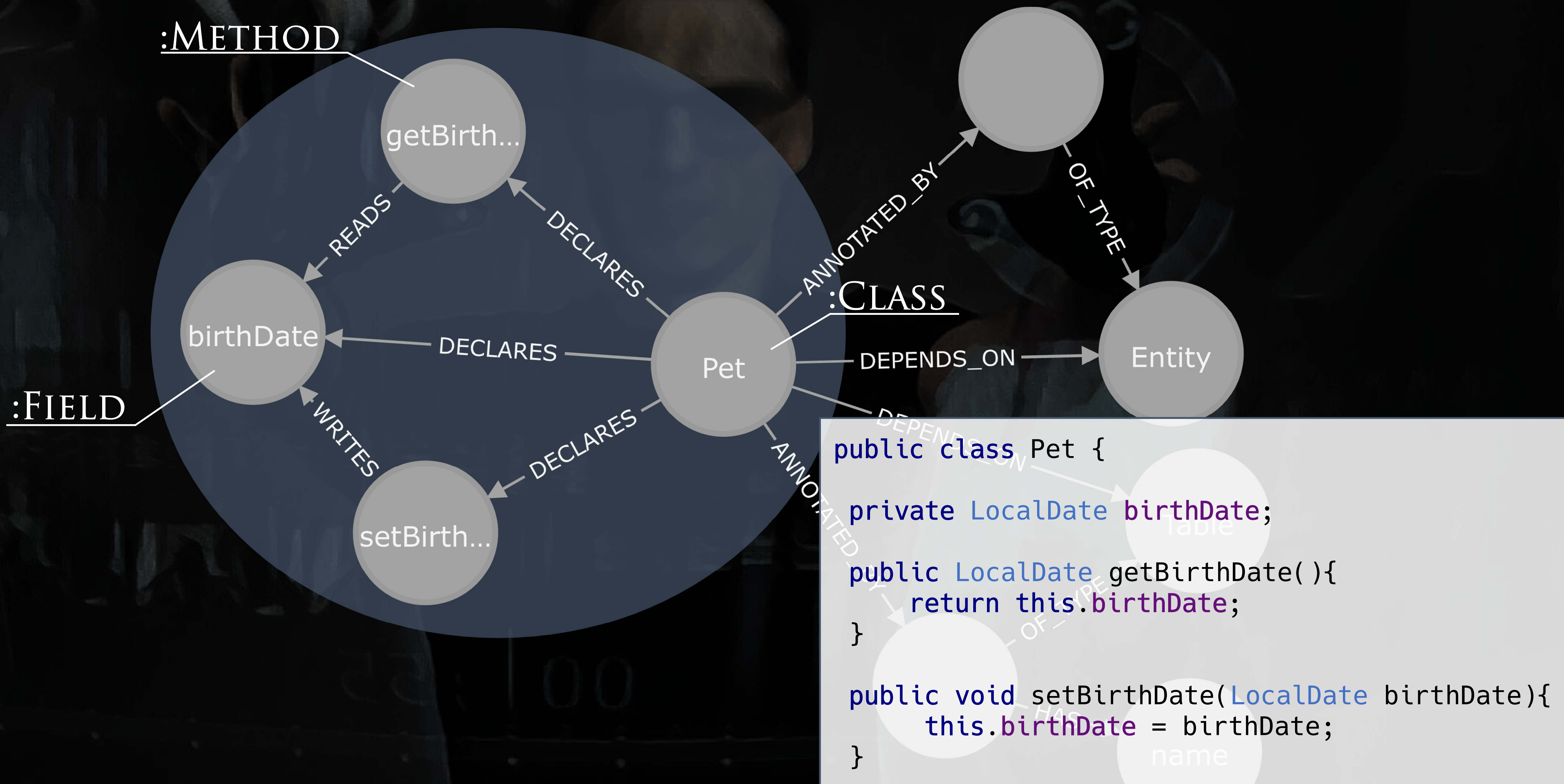


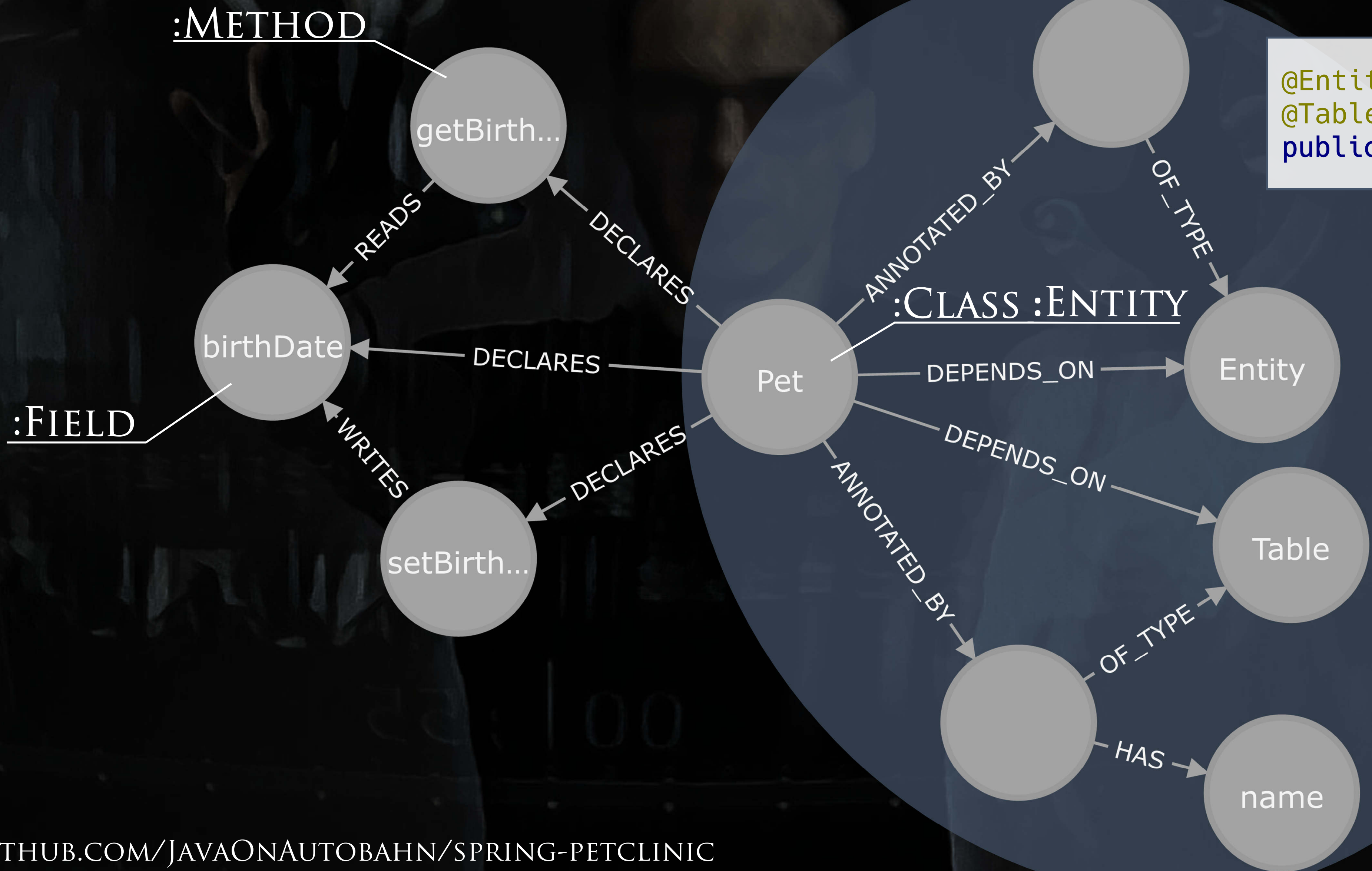
GRAPH DATA SCIENCE



JQASSISTANT

NEO4J





```

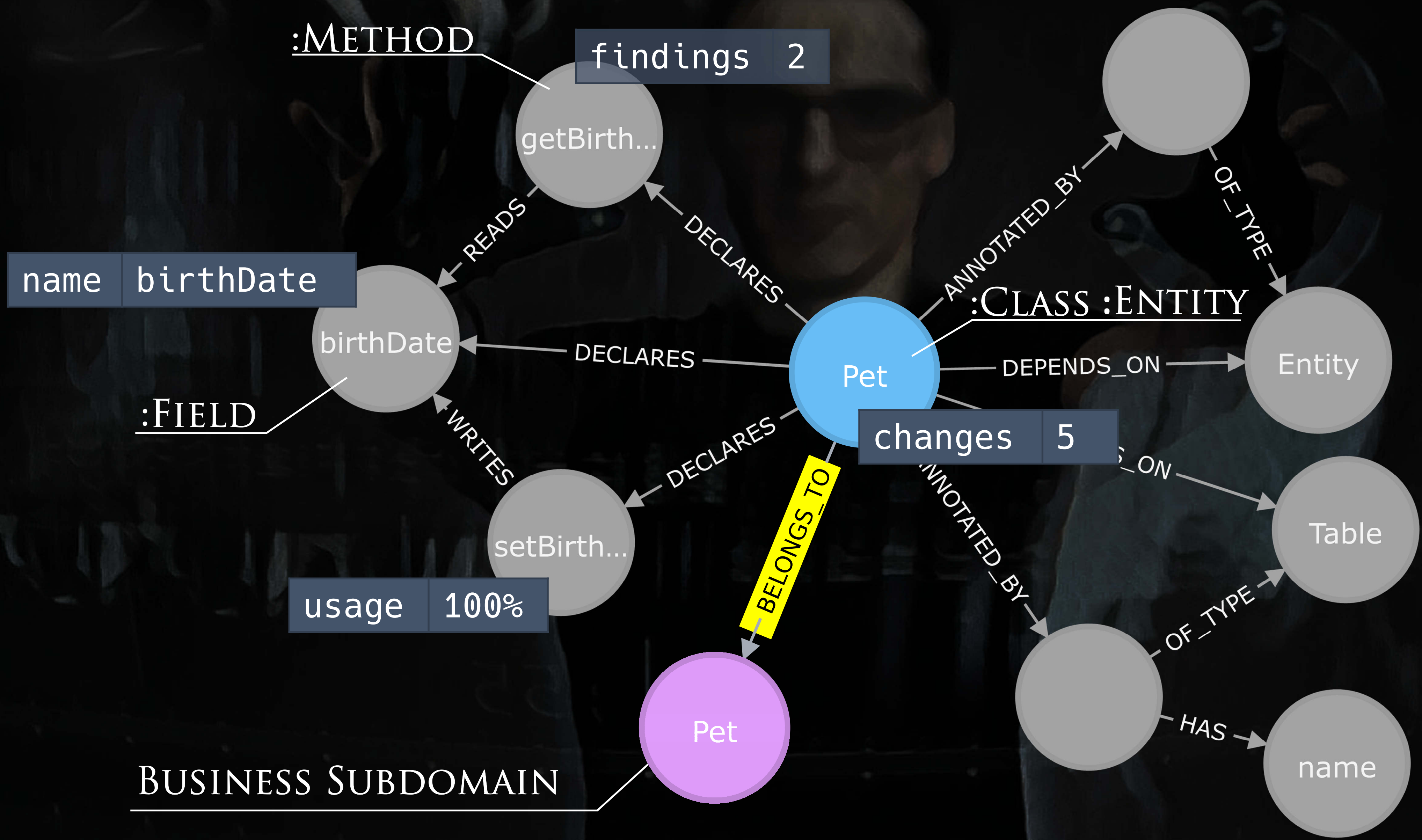
@Entity
@Table(name = "pets")
public class Pet {

```

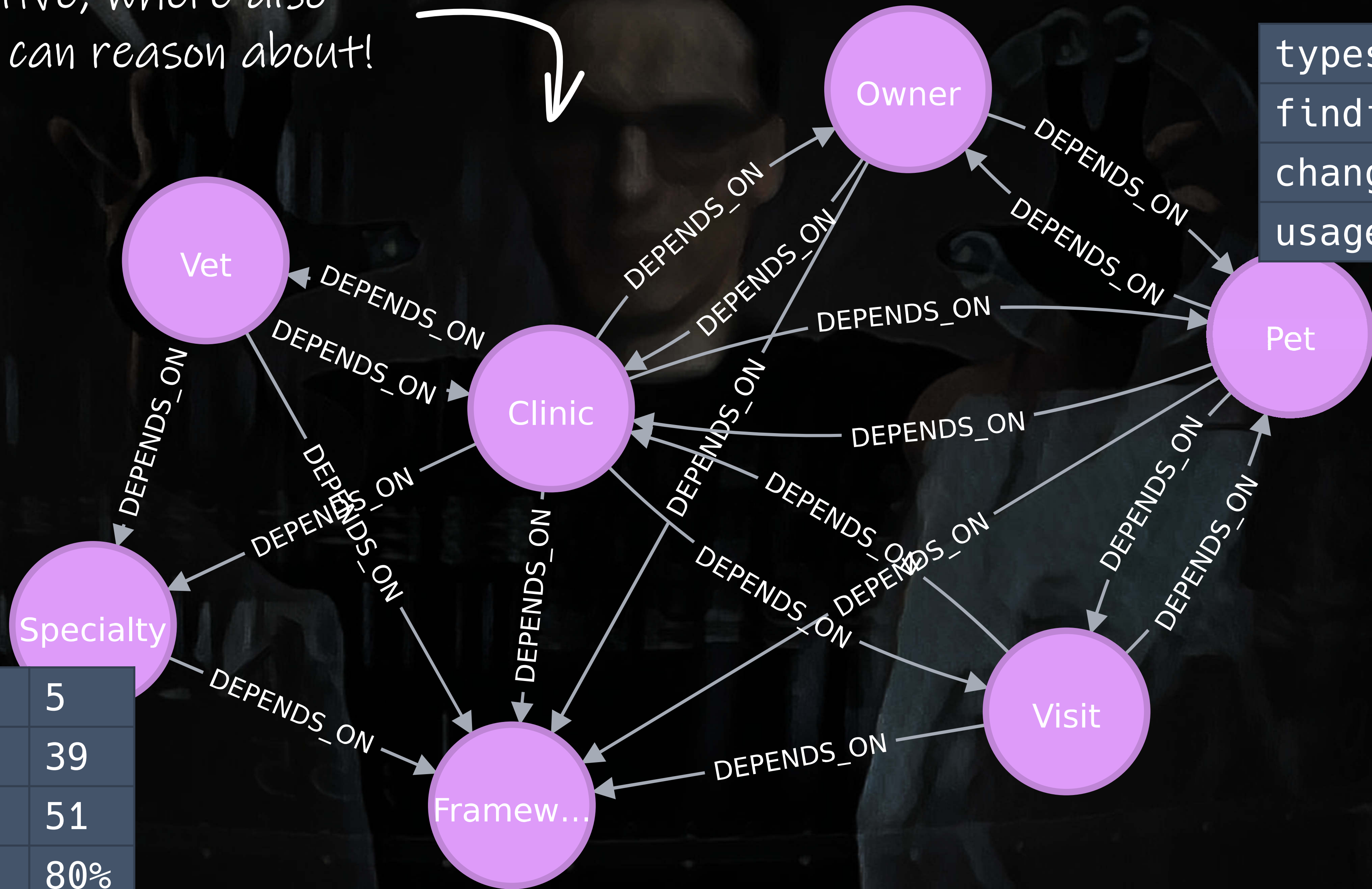
:METHOD

:FIELD

:CLASS :ENTITY



A perspective, where also managers can reason about!



types	16
findings	17
changes	15
usage	70%

types	5
findings	39
changes	51
usage	80%



Attribution: Tobias ToMar Maier, https://commons.wikimedia.org/wiki/File:VHS_tape_with_time_scale.jpg

BECOME THE
LORD OF THE THINGS
BY ANALYZING SOFTWARE IN A DATA-DRIVEN WAY



**THE FELLOWSHIP
OF THE BLING**



THE TWO TIPS



NUMBER OF
SOLVED PROBLEMS

**THE RETURN
OF REASON**

QUESTION



ASK 'EM ALL

Appendix

Demos

Jupyter notebook, python, pandas, matplotlib

Repo

https://github.com/feststelltaste/software-analytics/tree/master/demos/20240315_BOBKonf_2024

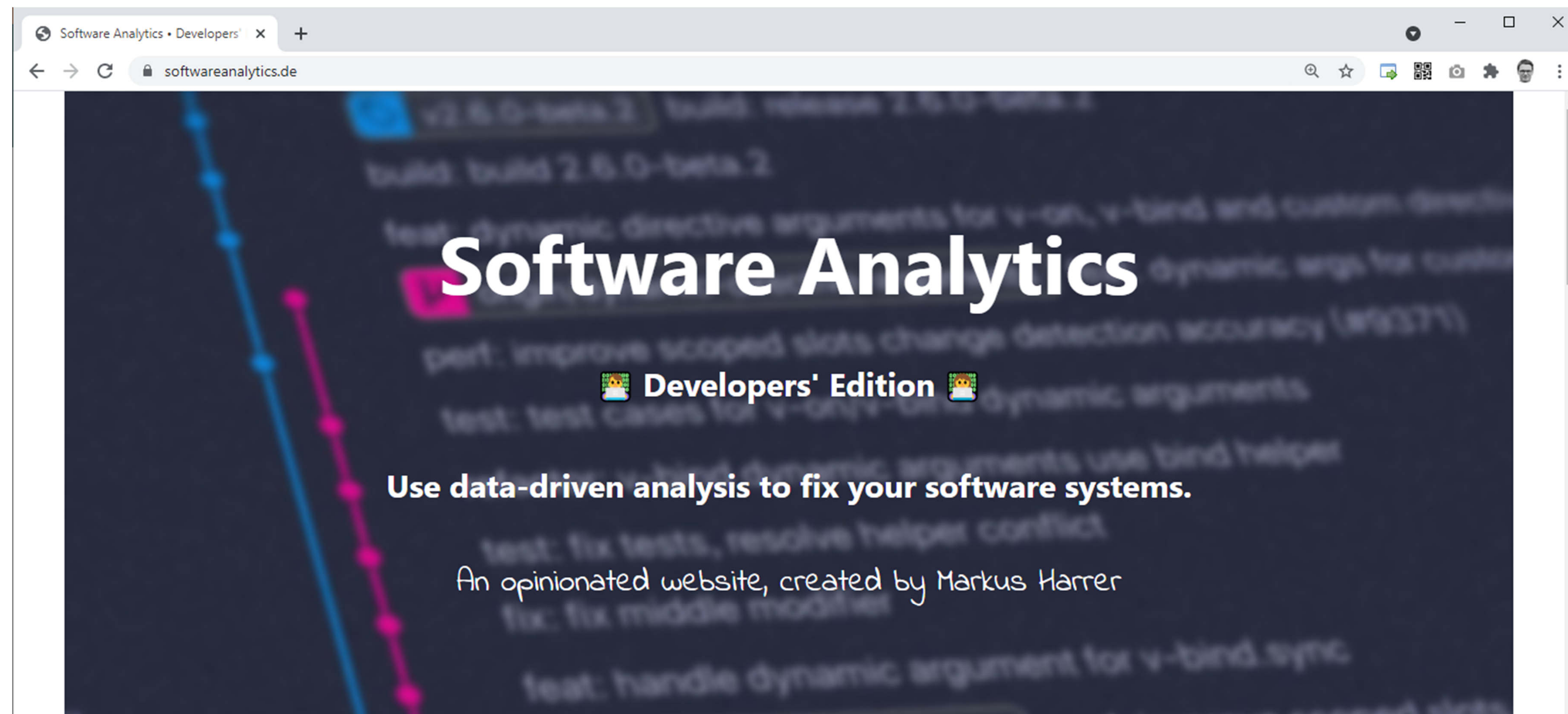
jQAssistant & Neo4j

Repo Spring PetClinic

<https://github.com/JavaOnAutobahn/spring-petclinic>

More on Software Analytics

softwareanalytics.de

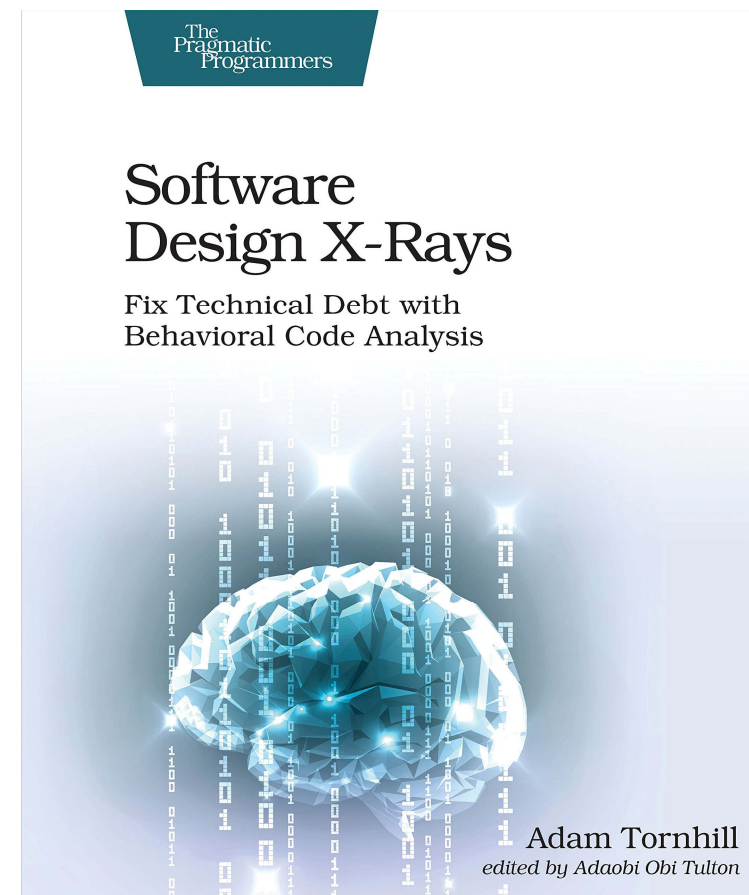


What is "Software Analytics?"

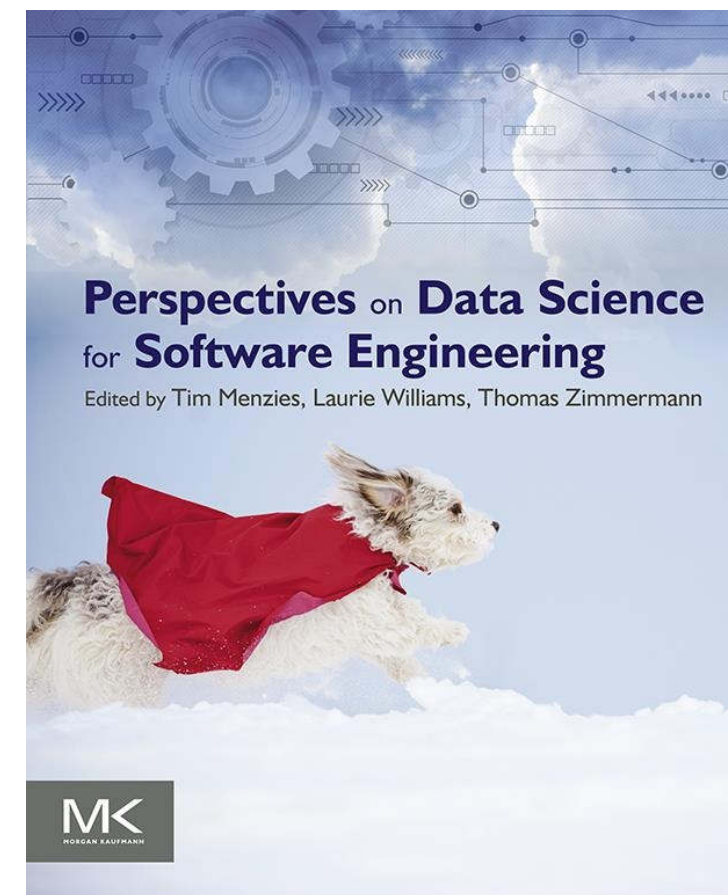
This is the best definition of Software Analytics I know so far:

Software analytics is analytics on software data for managers and software engineers with the aim of empowering software

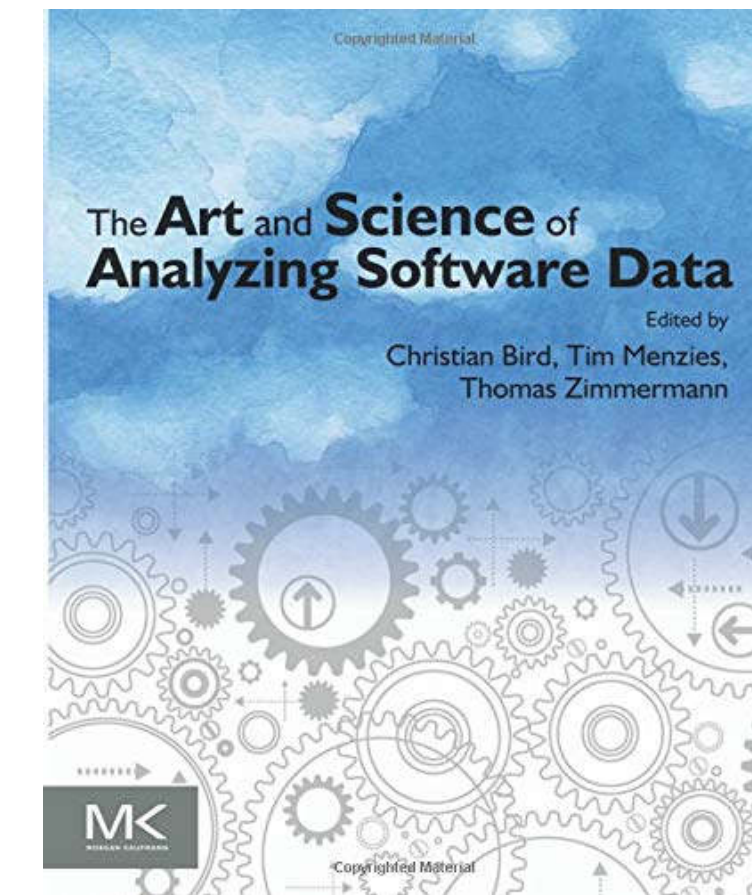
More on Software Analytics



Adam Tornhill:
Software X-Ray

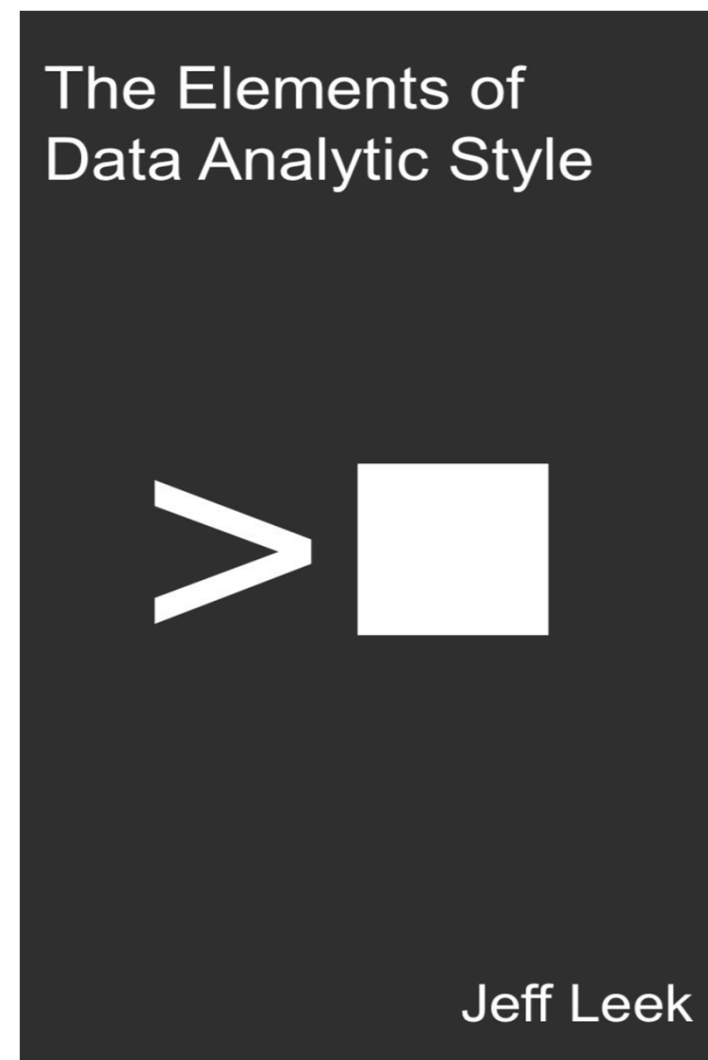


Tim Menzies, Laurie Williams,
Thomas Zimmermann:
*Perspectives on Data Science for
Software Engineering*



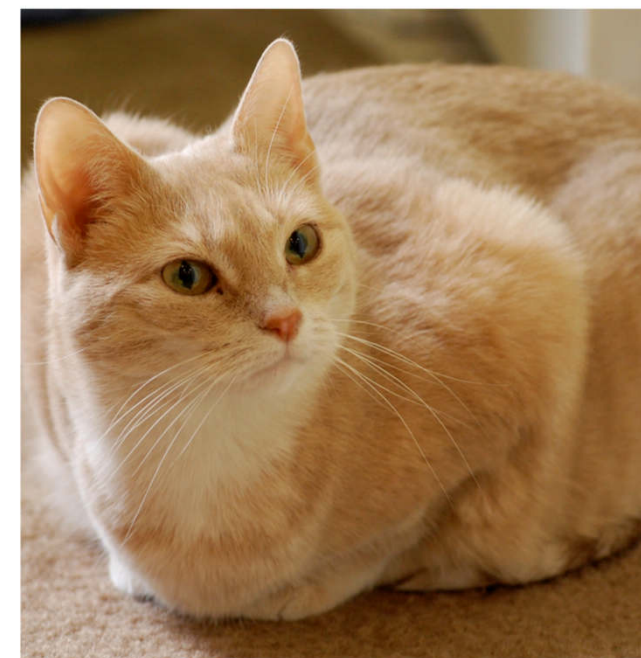
Christian Bird, Tim Menzies,
Thomas Zimmermann:
*The Art and Science of Analyzing
Software Data*

More on Data Science



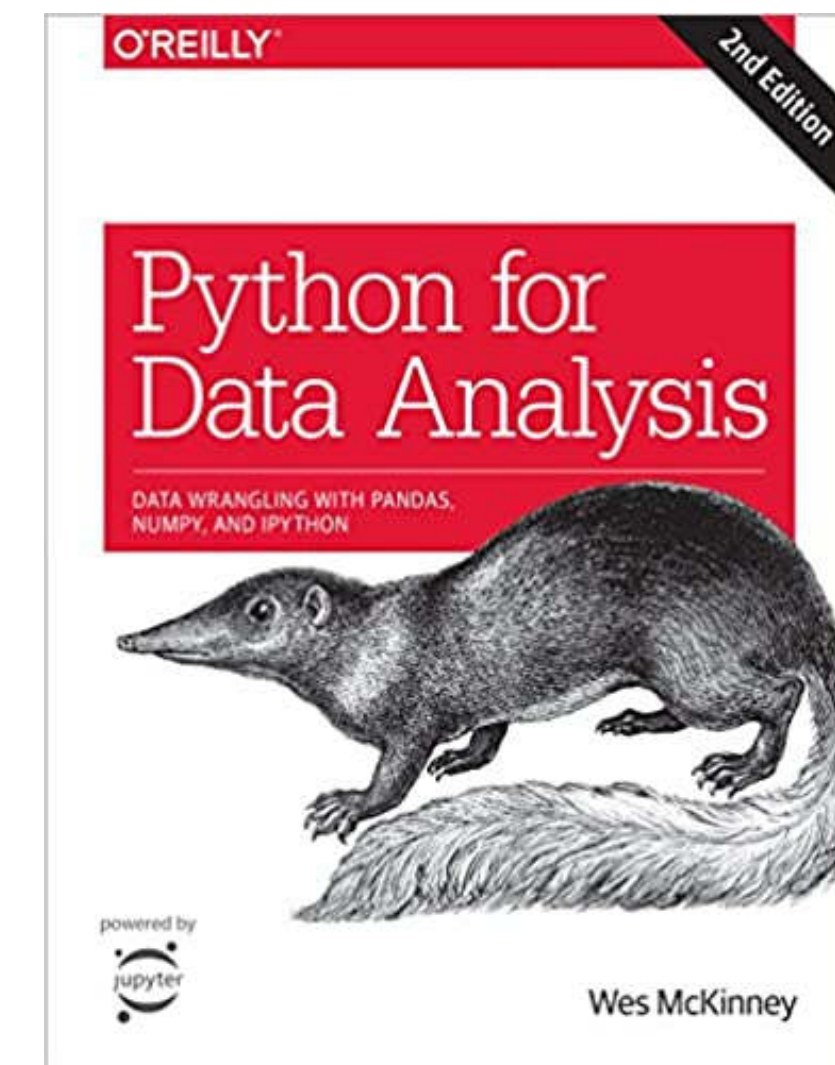
Jeff Leek:
*The Elements of Data
Analytic Style*

Report Writing for
Data Science in R



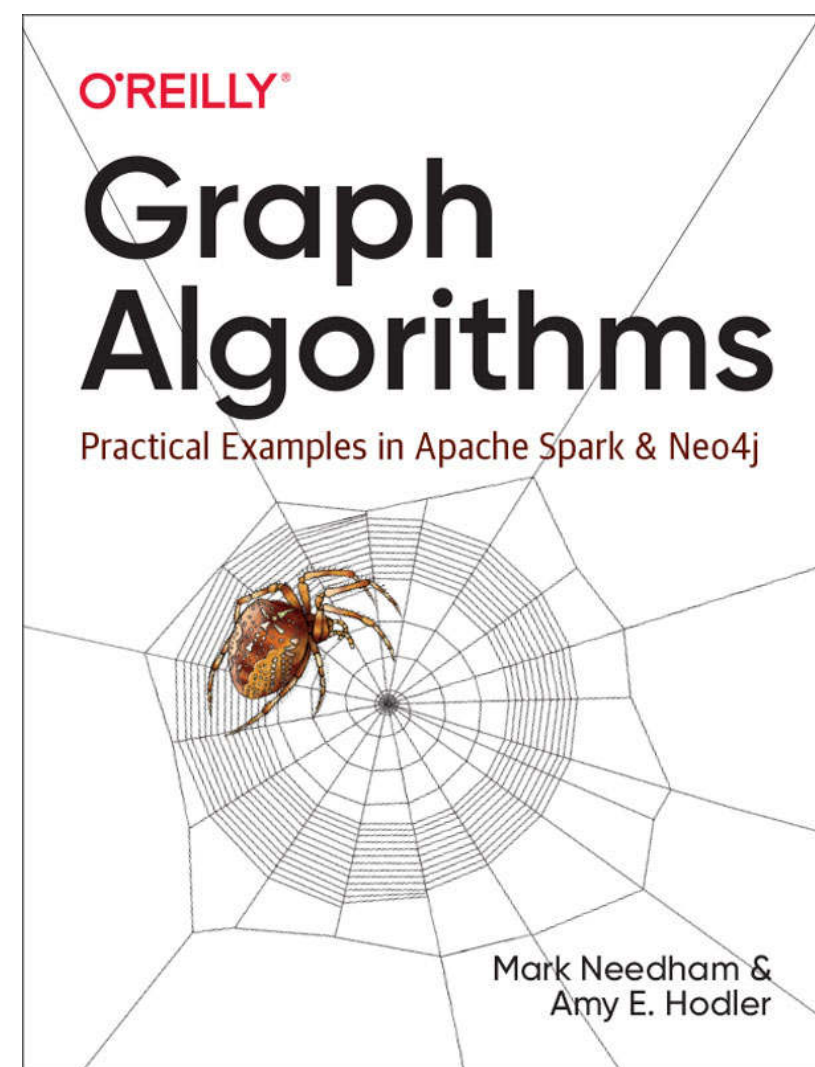
Roger D. Peng

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



Wes McKinney:
Python for Data Analysis

More on Graph Analytics



*Mark Needham & Amy Hodler:
Graph Algorithms*

This is a screenshot of the Neo4j website's product page for 'Neo4j Graph Data Science'. The page has a clean, modern design with a light yellow background. At the top, there is a navigation bar with the Neo4j logo and several menu items: 'Products', 'Use Cases', 'Developers & Data Scientists', 'Pricing', and 'Learn'. On the right side of the navigation bar, there are links for 'Aura Login', 'Partners', 'Company', and 'Support', along with a search icon. Below the navigation bar, there are two buttons: 'Contact Us' and 'Get Started Free'. The main content area starts with the heading 'WHAT IS IT?' followed by the title 'Neo4j Graph Data Science'. Below the title is a paragraph of text explaining that Graph Data Science is an analytics and machine learning solution that analyzes relationships in data. At the bottom of this section are two buttons: 'Read 5 Graph Data Science Basics' and 'Learn What's New'. To the right of the text is a large, colorful isometric illustration featuring various data science and business icons, such as a bar chart, a pie chart, a network graph, a clock, and a magnifying glass, all connected by lines to a central data cylinder.

<https://neo4j.com/product/graph-data-science/>

Paper about 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 important questions in this respect [2].



Thank you very much!

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 Munich
Germany

innoQ Schweiz GmbH

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

Albulastr. 55
8048 Zurich
Switzerland