BOBKONF 2024, 15.03.2024

Software Analytics with Data Science on Software Data

Markus Harrer
Software Evolutionist @ INNOQ

Social: markusharrer.de

INNOQ

THE ORIGINAL HORROR SHOW

LEGACY SYSTEMS

IMAGE: SETHJREID/PIXABAY

The problem with the problems

Features

Architecture

Errors

Strange things

MANAGEMENT

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

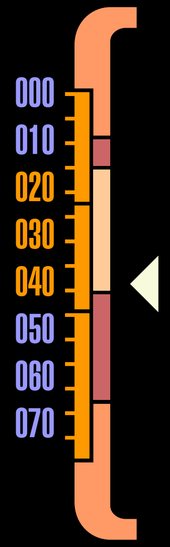
7232-838

03-975683

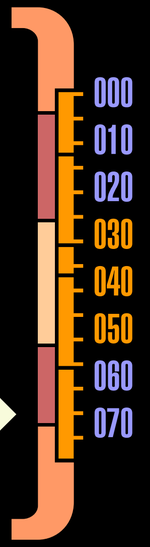
04-765466

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

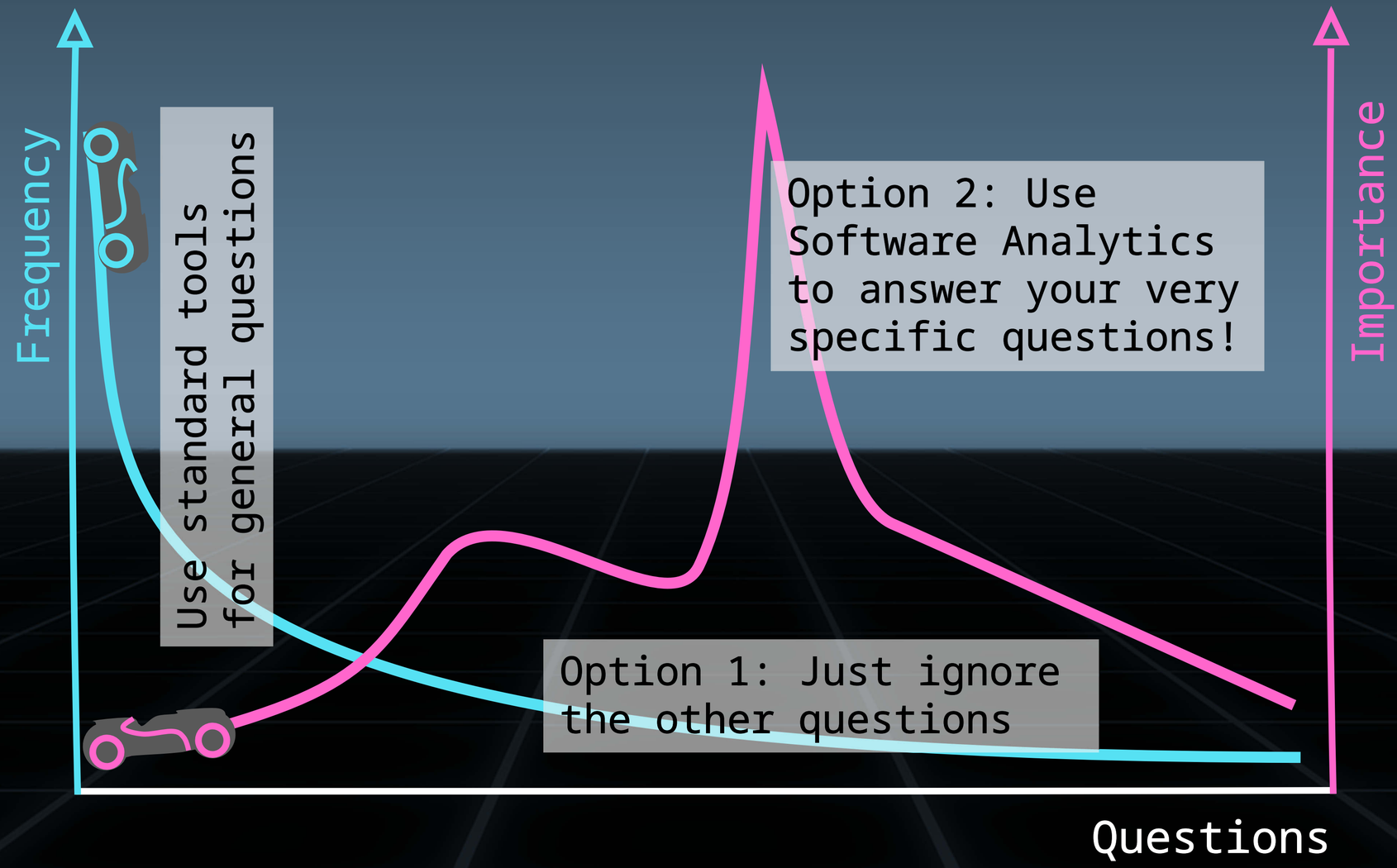
to gain and share insight from

their data to make better

decisions.

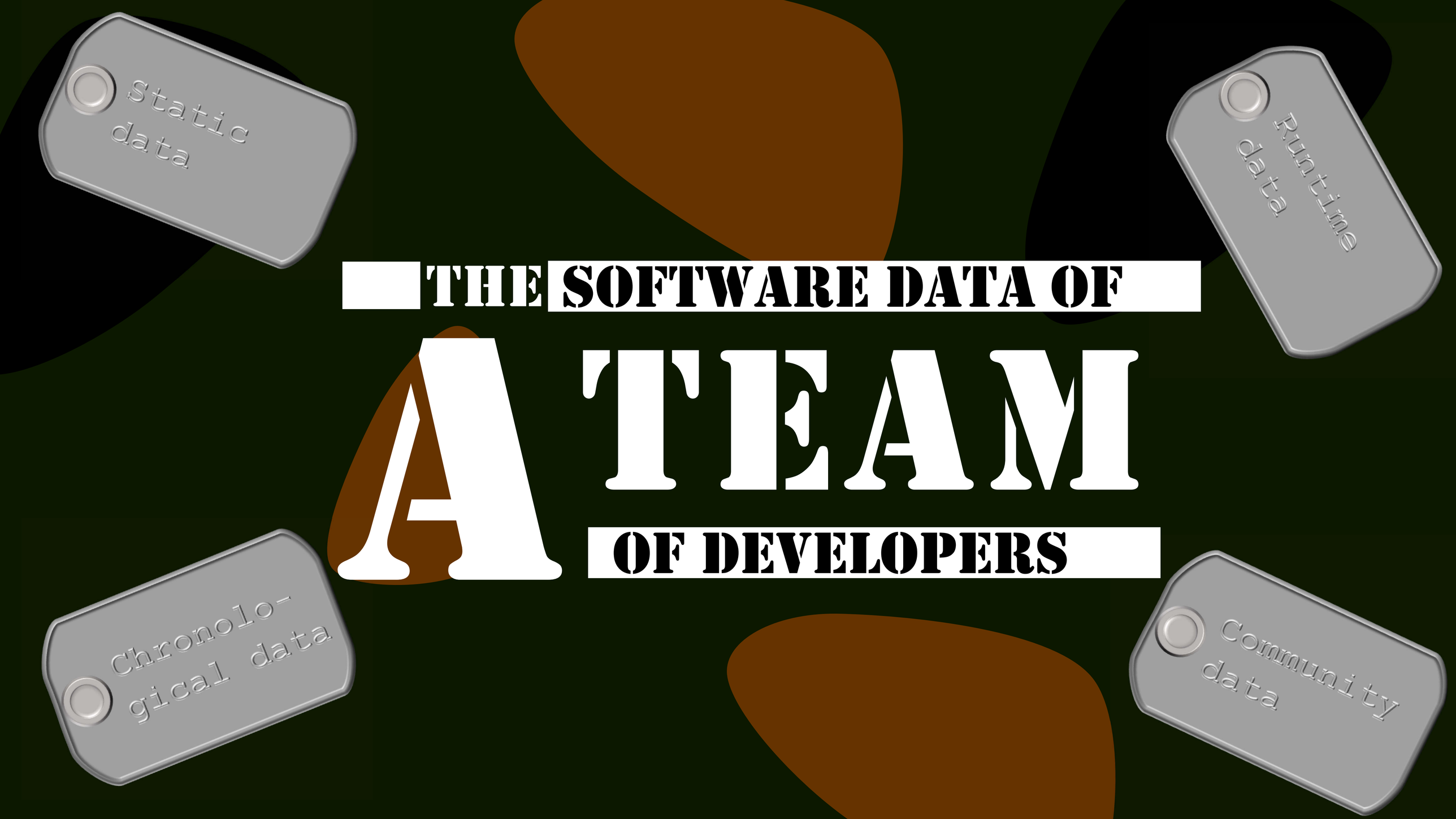

SOFTWARE
ANALYTICS
EPISODE II
A NEW HOPE

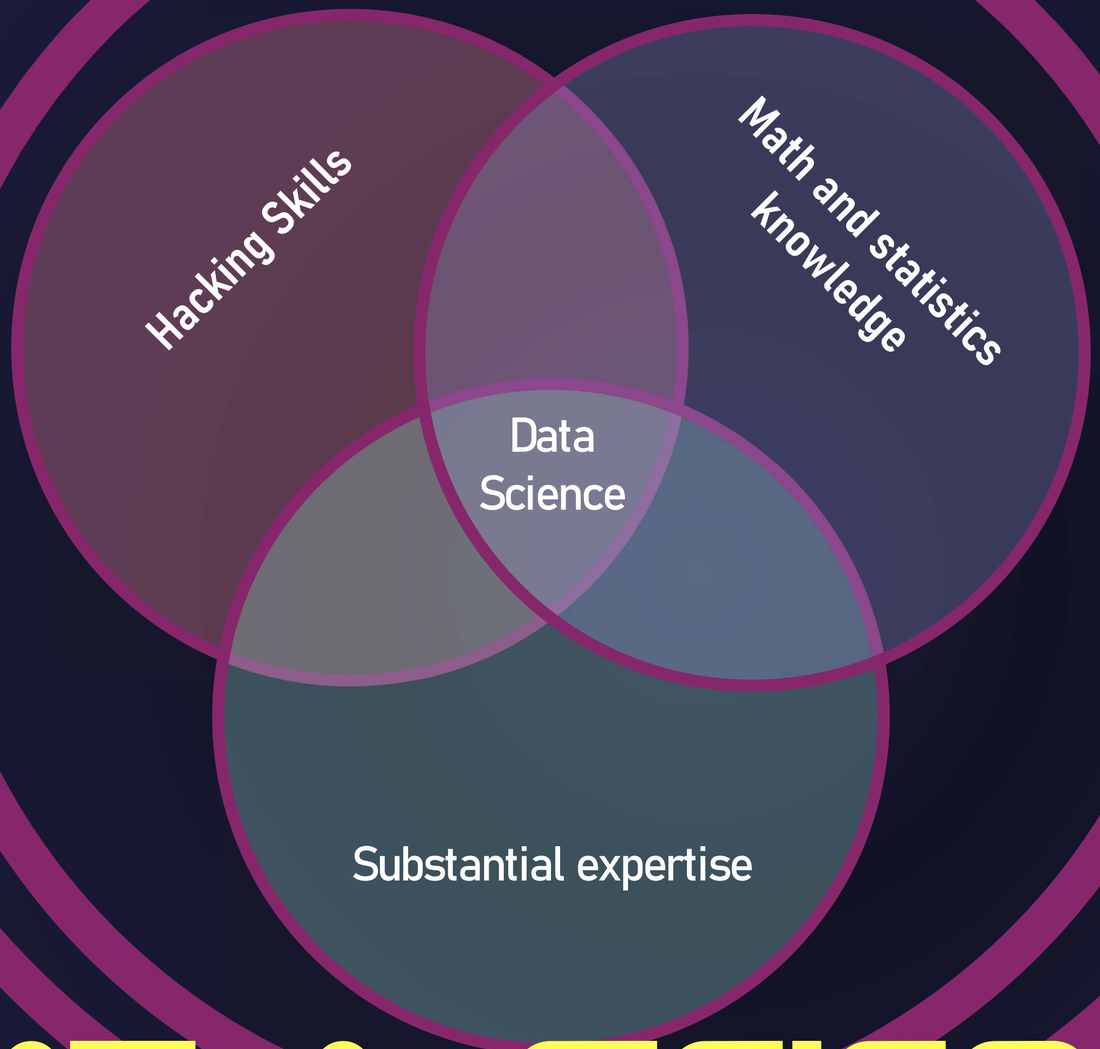
ANSWERING YOUR SPECIFIC QUESTIONS



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>

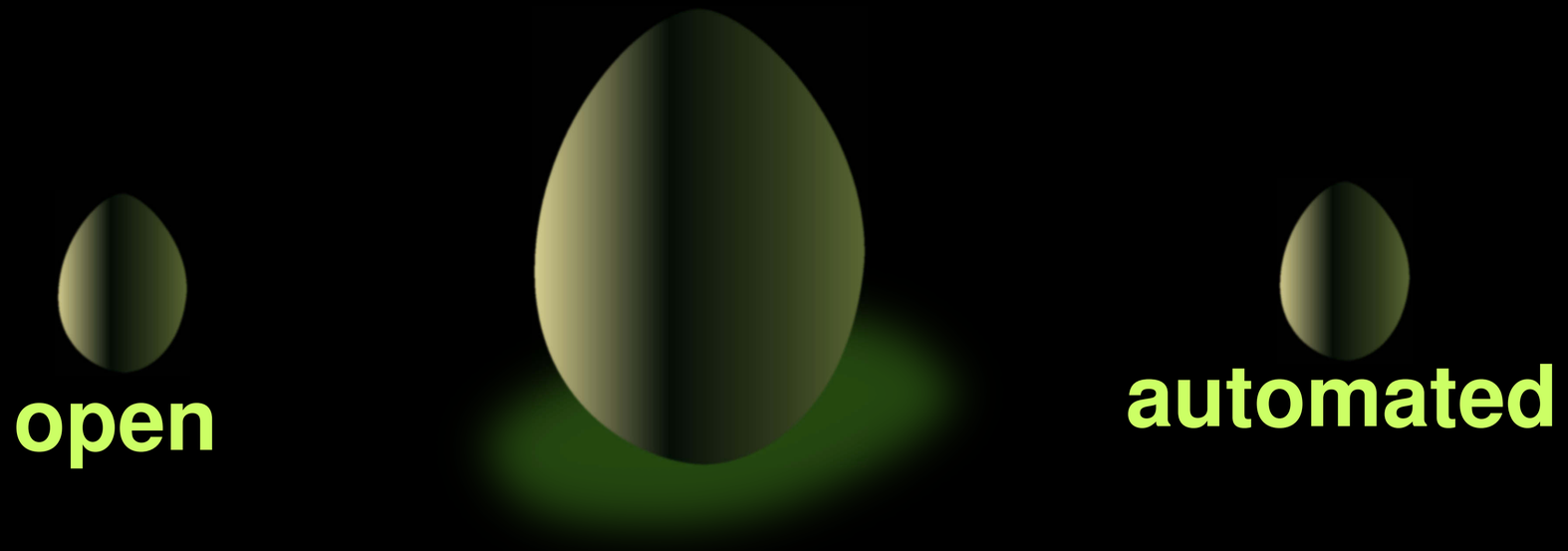

Static data
Runtime data
THE SOFTWARE DATA OF
A TEAM
OF DEVELOPERS
Chronological data
Community data



DATA SCIENCE

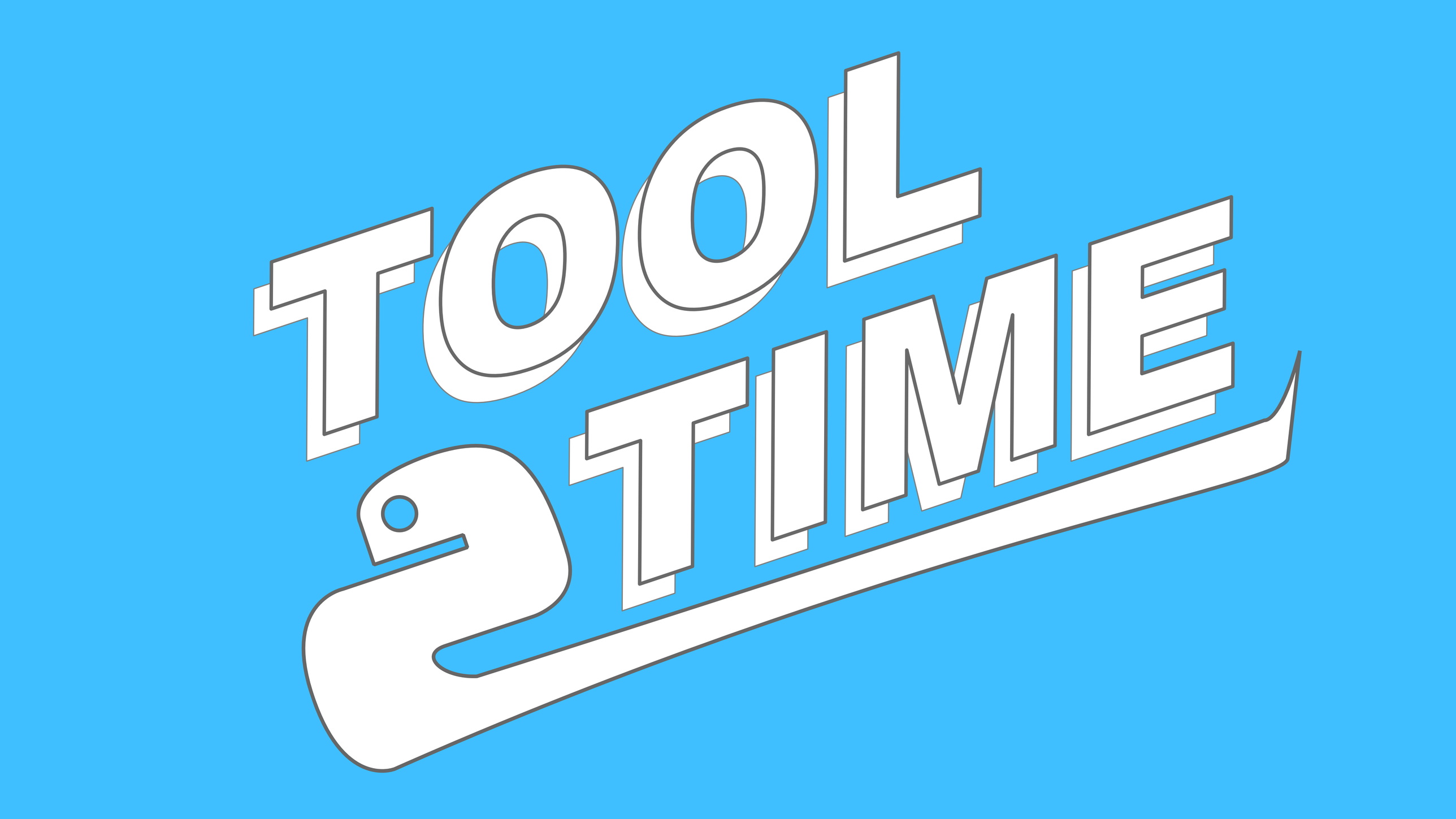
Venn diagram by Drew Conway (simplified)

REPRODUCIBLE DATA SCIENCE



= A WAY TO IMPLEMENT SOLID

SOFTWARE ANALYTICS


TOOL
TIME

Python 3

PANDAS

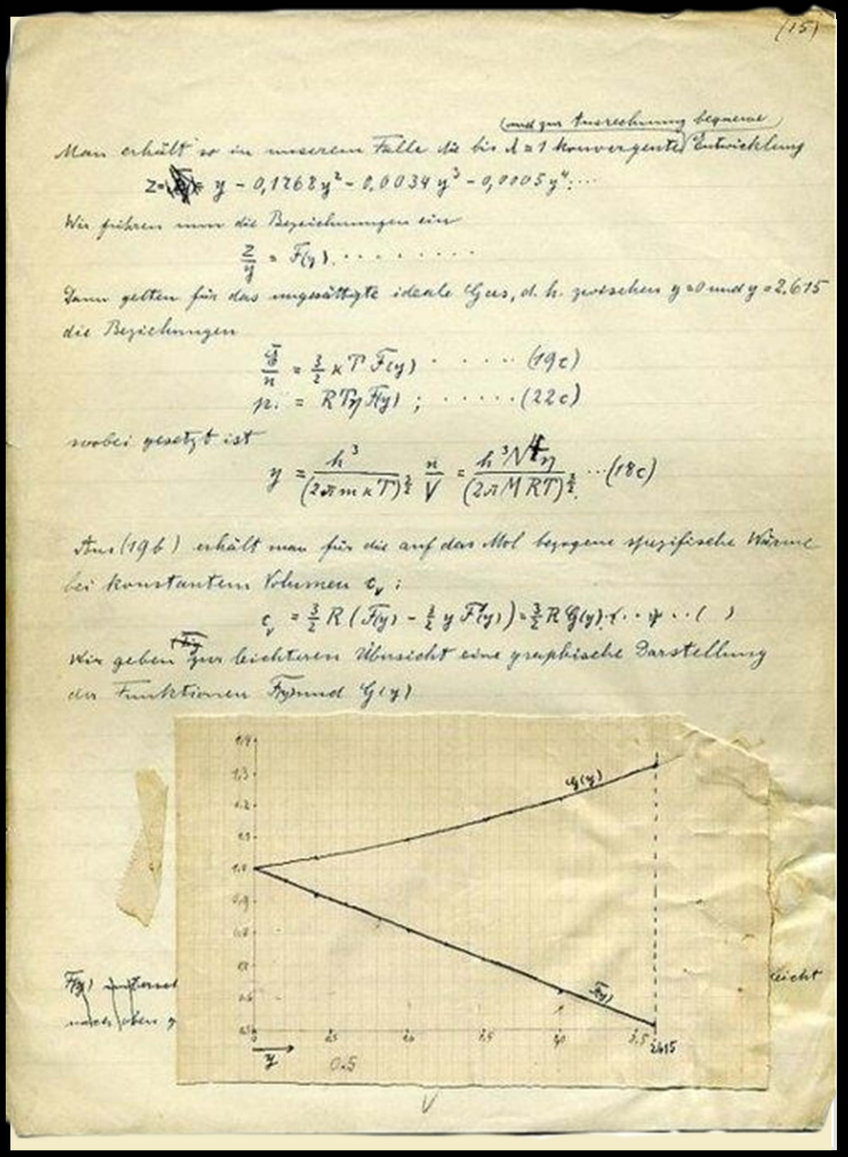
... and matplotlib, numpy, scikit-learn, NLTK, Pygments, py2neo, requests, BeautifulSoup, Pygal ...

Computational Notebook

code and data in love

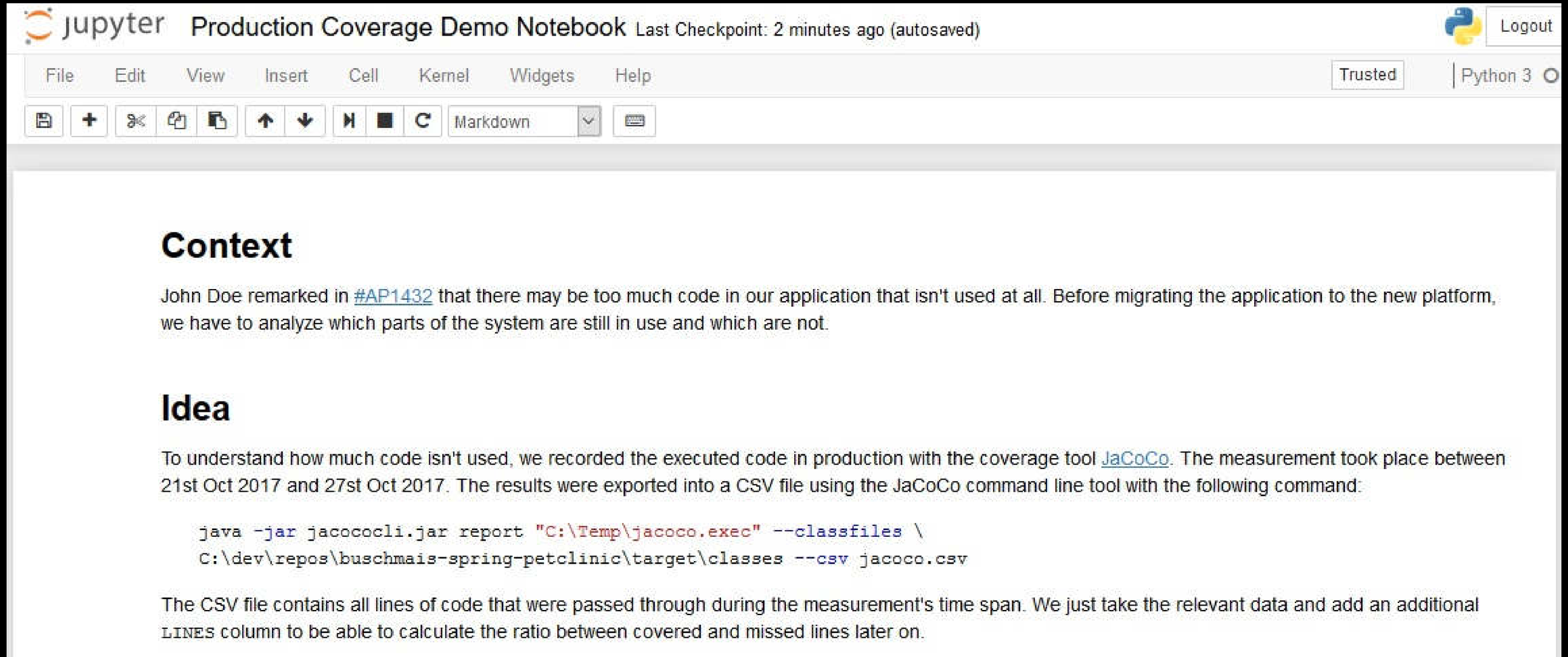
Computational Notebook

Jupyter Notebook



Literate Programming

with Jupyter Notebook

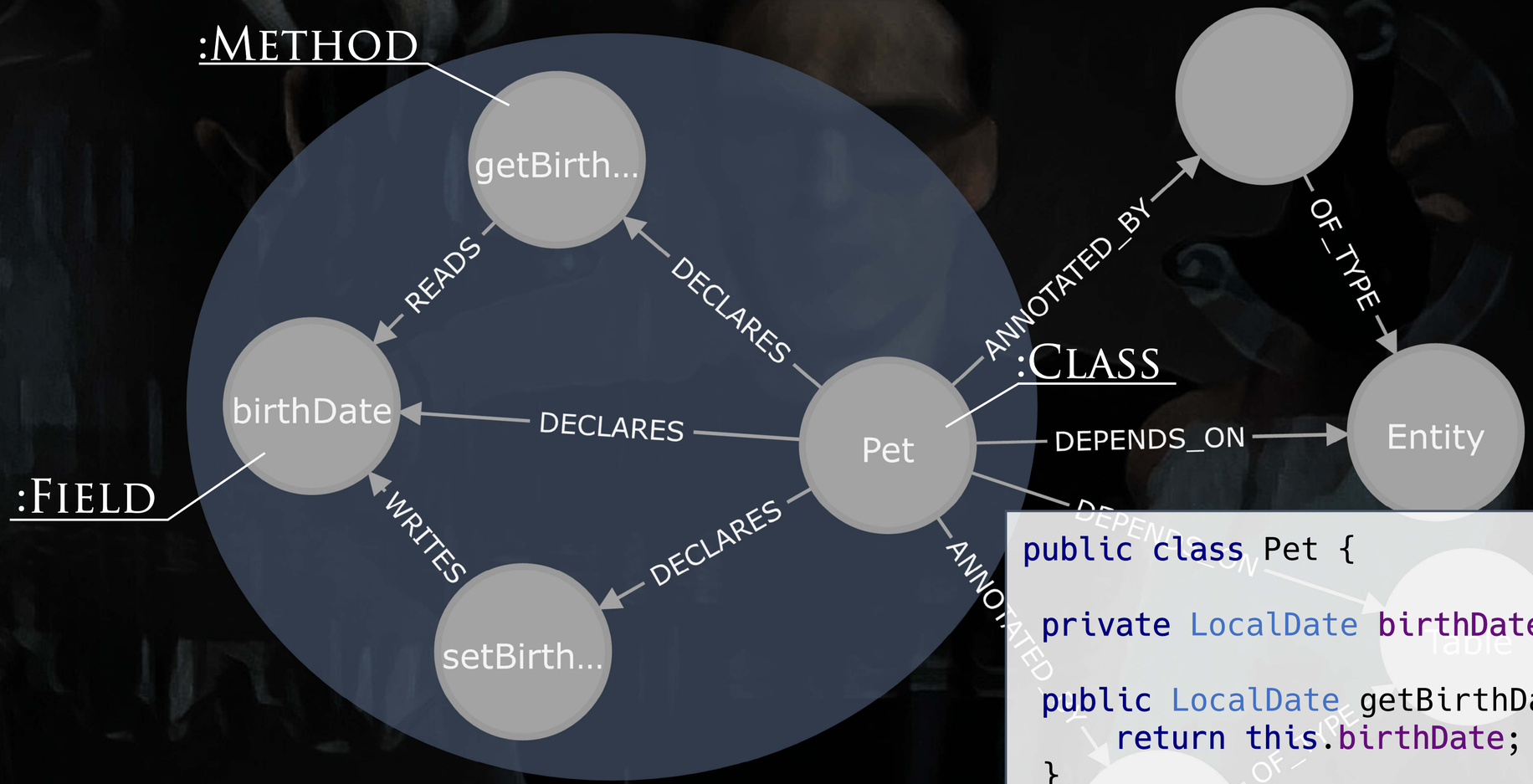


Attribution: Tobias ToMar Maier, https://commons.wikimedia.org/wiki/File:VHS_tape_with_time_scale.jpg

Graph Data Science

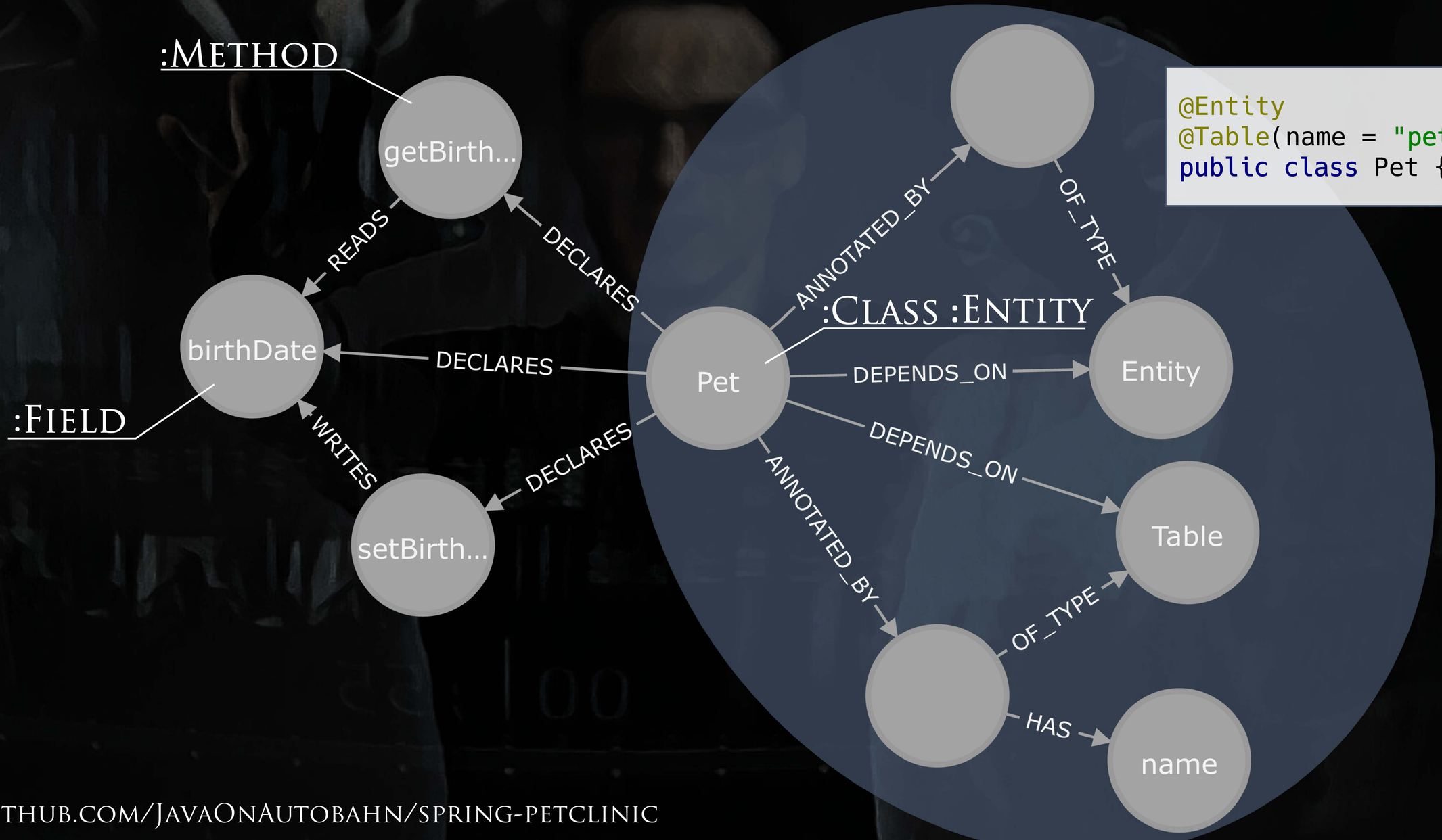
jQAssistant

Neo4j

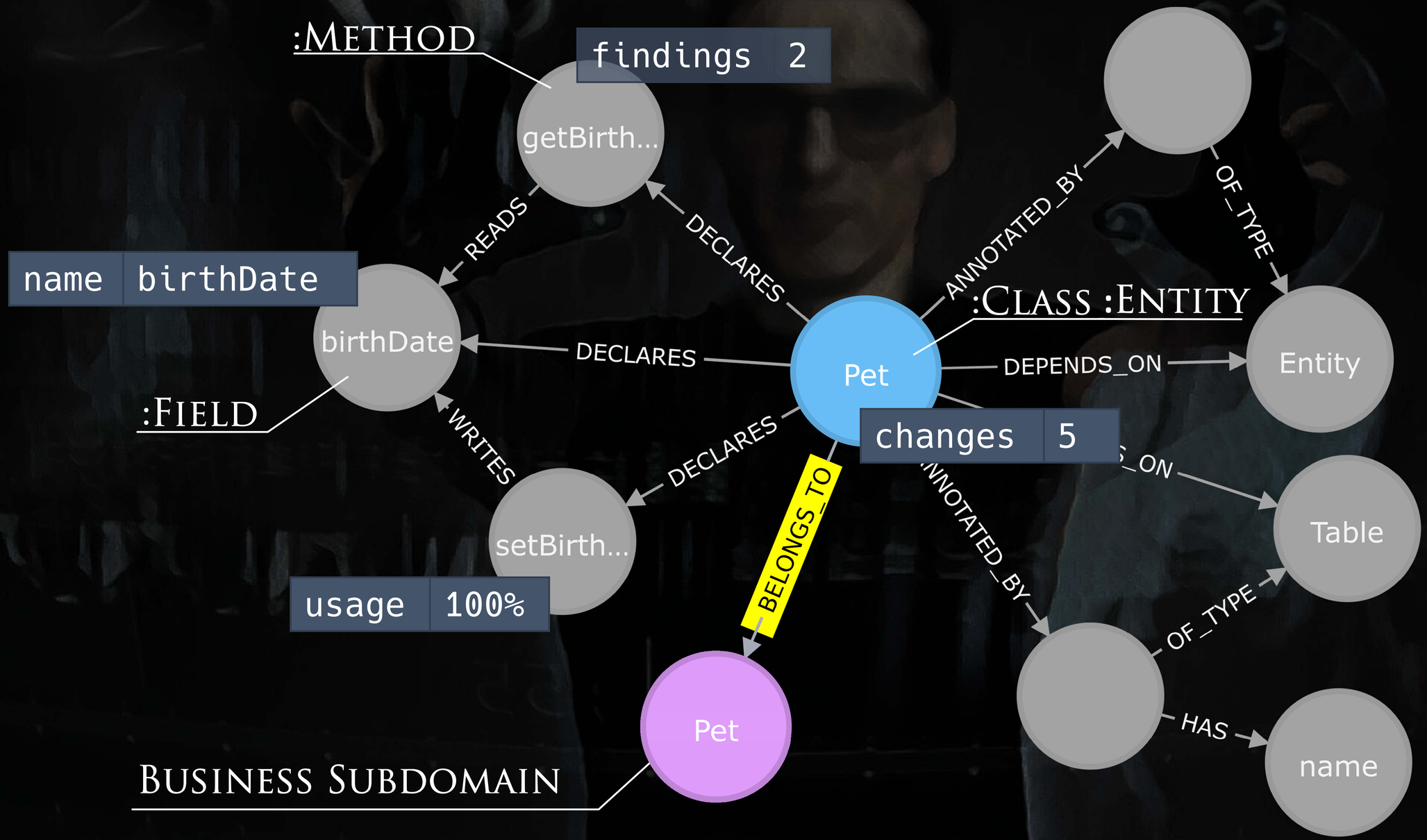



```
public class Pet {

 private LocalDate birthDate;

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

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

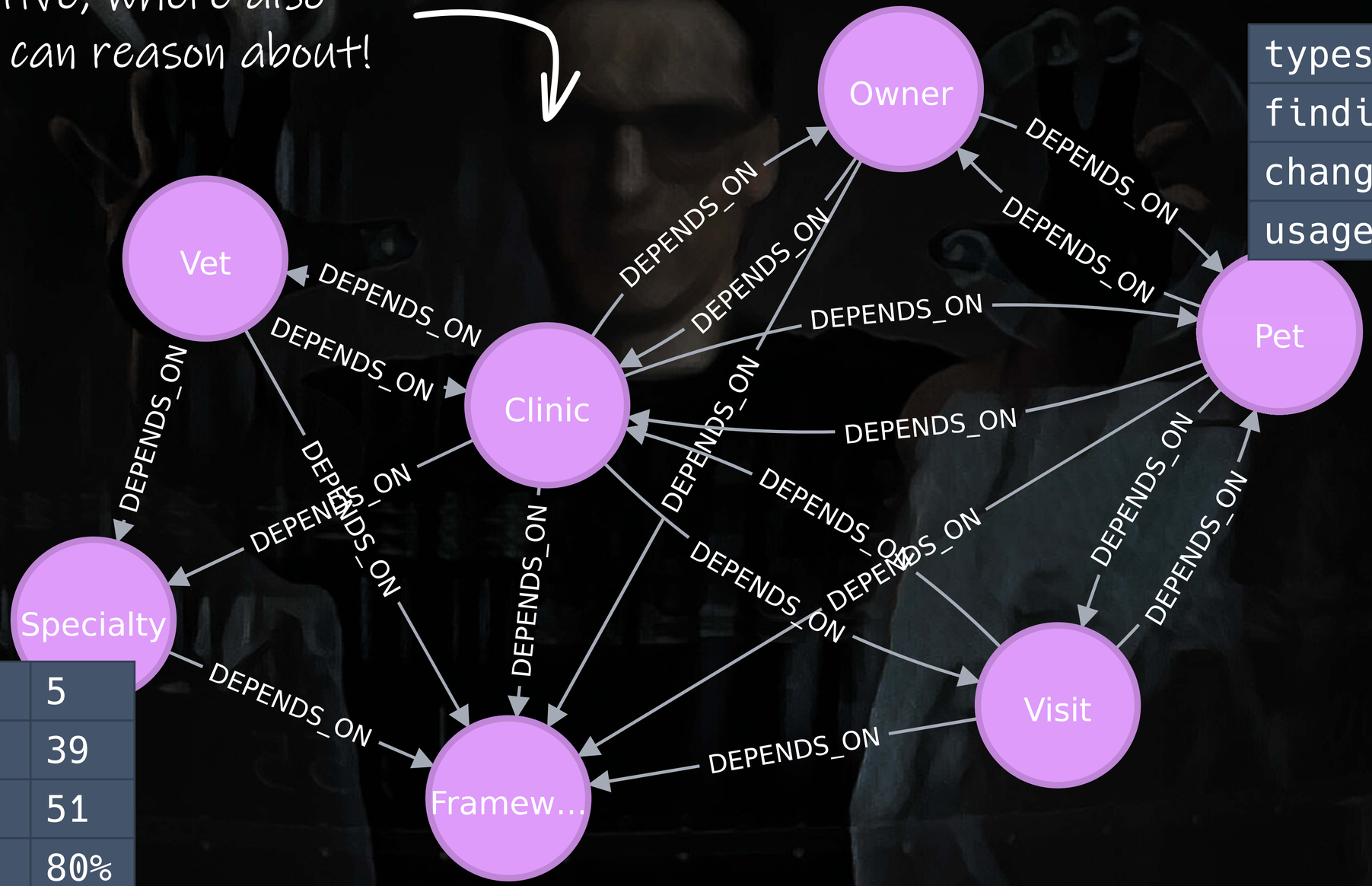
https://github.com/JavaOnAutobahn/spring-petclinic



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


:METHOD
findings 2
getBirth...
READS
DECLARES
name birthDate
birthDate
:FIELD
DECLARES
Pet
ANNOTATED_BY
OF_TYPE
:CLASS :ENTITY
Entity
DEPENDS_ON
changes 5
WRITES
DECLARES
setBirth...
usage 100%
BELONGS_TO
ANNOTATED_BY
Table
OF_TYPE
HAS
name
Pet
BUSINESS SUBDOMAIN

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

QUESTIONS

ASK ' EM ALL

@feststelltaste

Appendix

Demos

Jupyter notebook, python, pandas, matplotlib

Repo

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

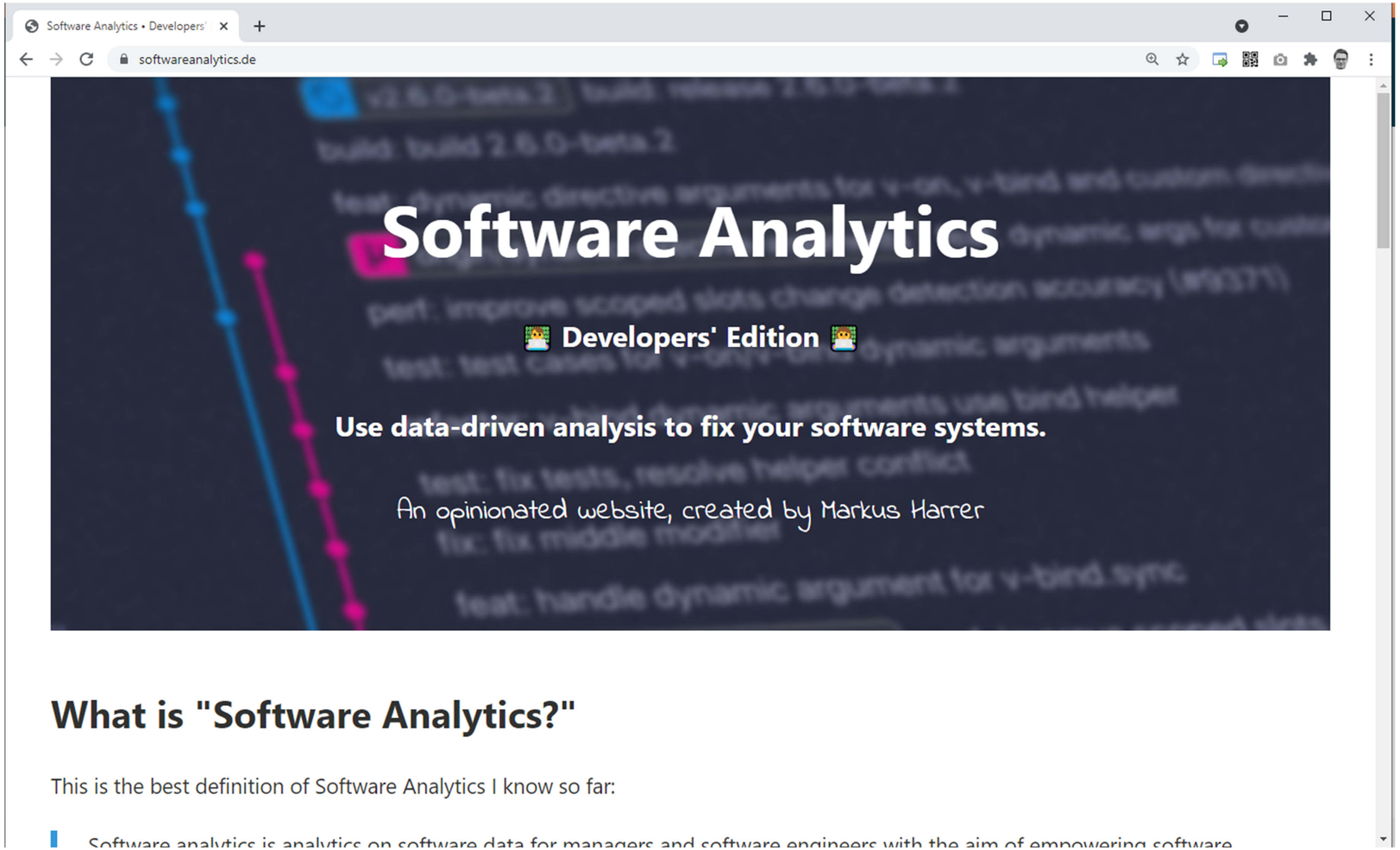
jQAssistant & Neo4j

Repo Spring PetClinic

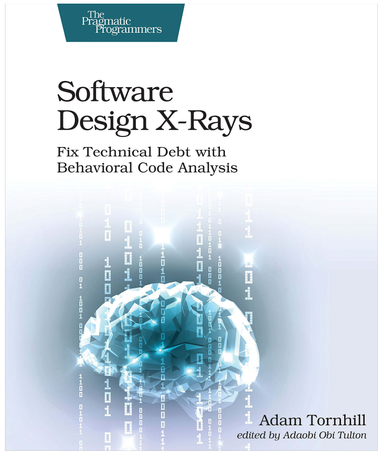
https://github.com/JavaOnAutobahn/spring-petclinic

More on Software Analytics

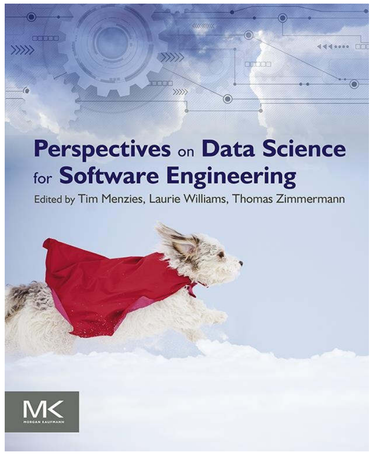
softwareanalytics.de



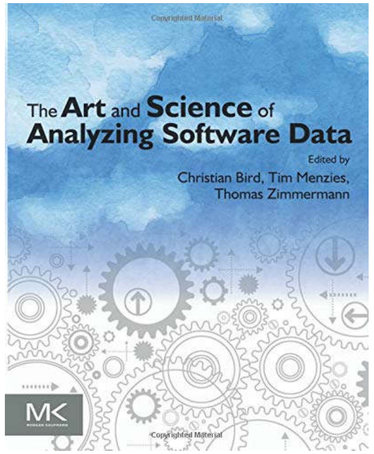
More on Software Analytics



Adam Tornhill:
Software X-Ray

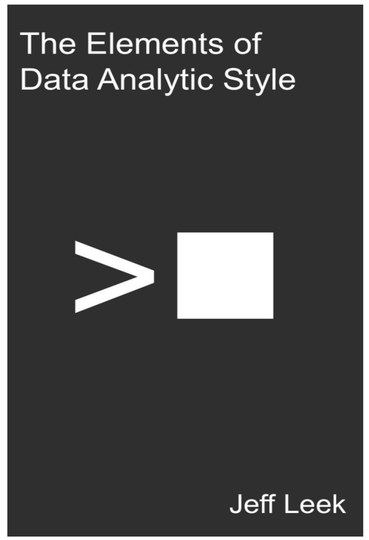


Tim Menzies, Laurie Williams, Thomas Zimmermann:
Perspectives on Data Science for Software Engineering



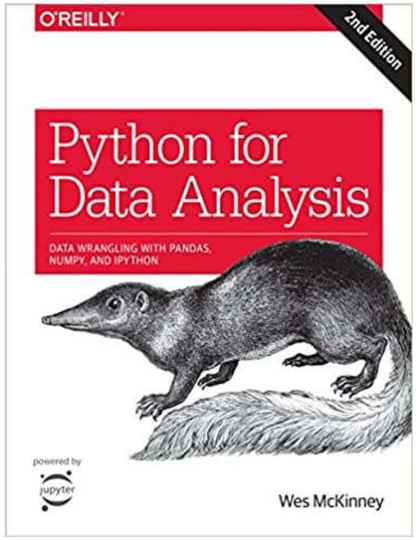
Christian Bird, Tim Menzies, Thomas Zimmermann:
The Art and Science of Analyzing Software Data

More on Data Science



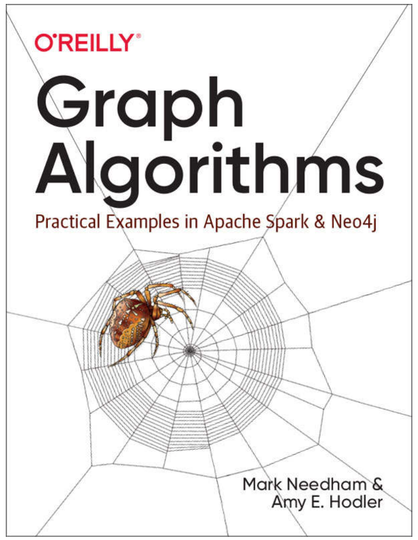
Jeff Leek:
The Elements of Data Analytic Style

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

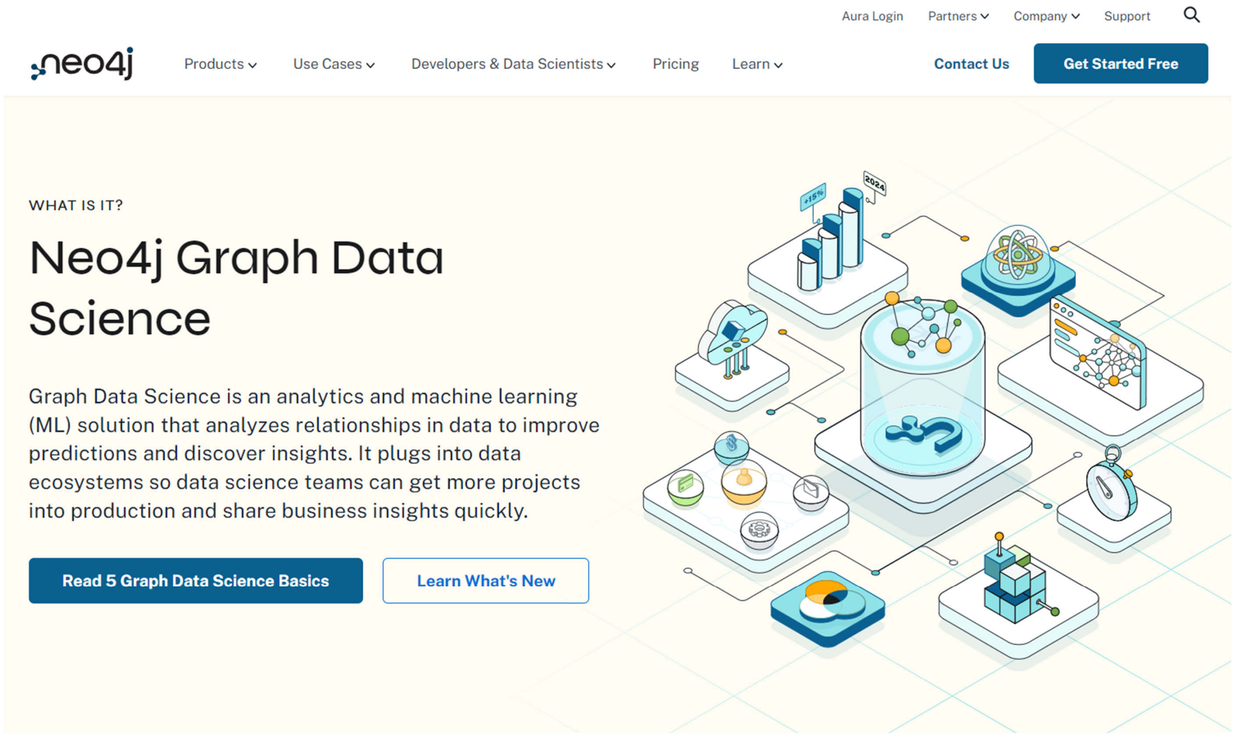


Wes McKinney:
Python for Data Analysis

More on Graph Analytics



Mark Needham & Amy Hodler: Graph Algorithms



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

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