Data Science in Software Engineering

Increasing Software Productivity

Dr. Eray Tüzün

Department of Computer Engineering

Bilkent University

@eraytuzun

@tuzuneray
http://linkedin.com/in/tuzuneray

@eraytuzun
http://twitter.com/eraytuzun

Software is Eating the World

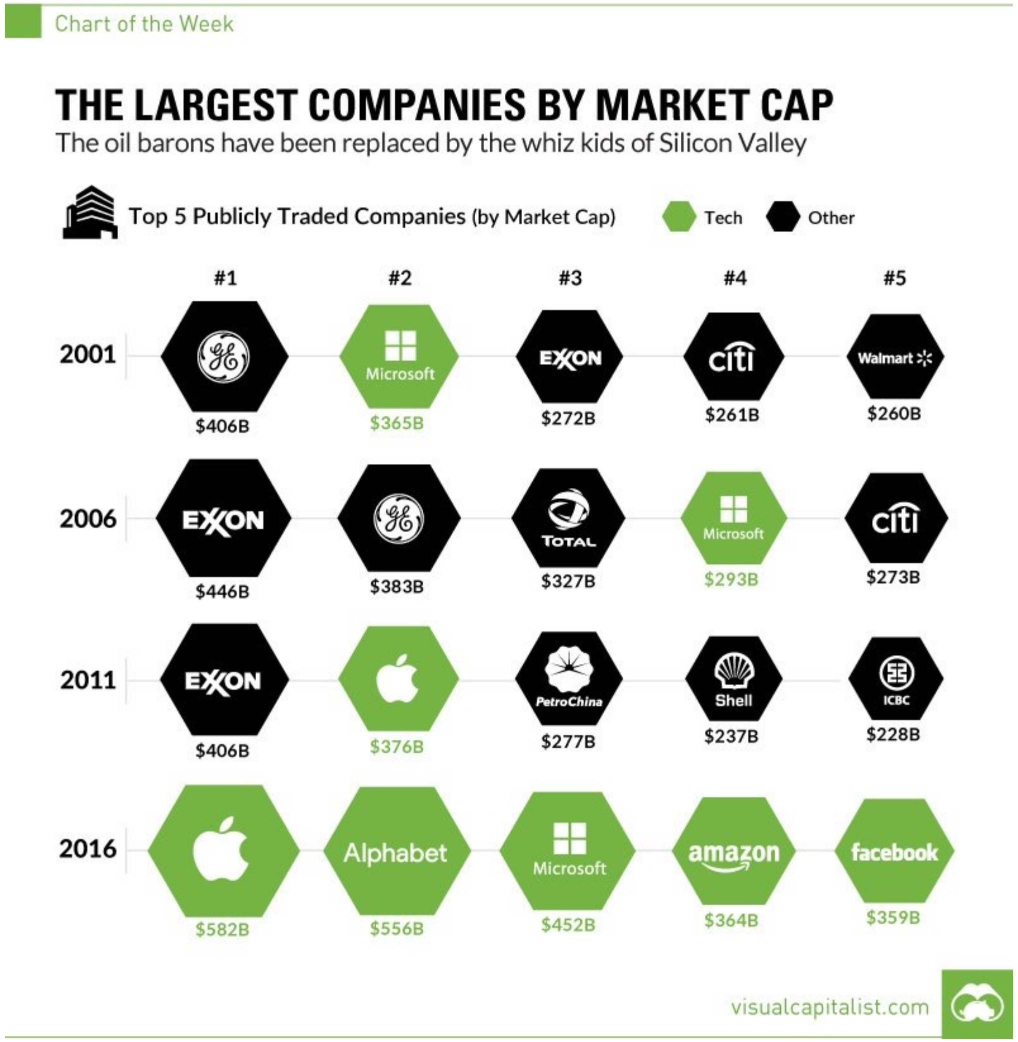


Chart of the Week

THE LARGEST COMPANIES BY MARKET CAP

The oil barons have been replaced by the whiz kids of Silicon Valley

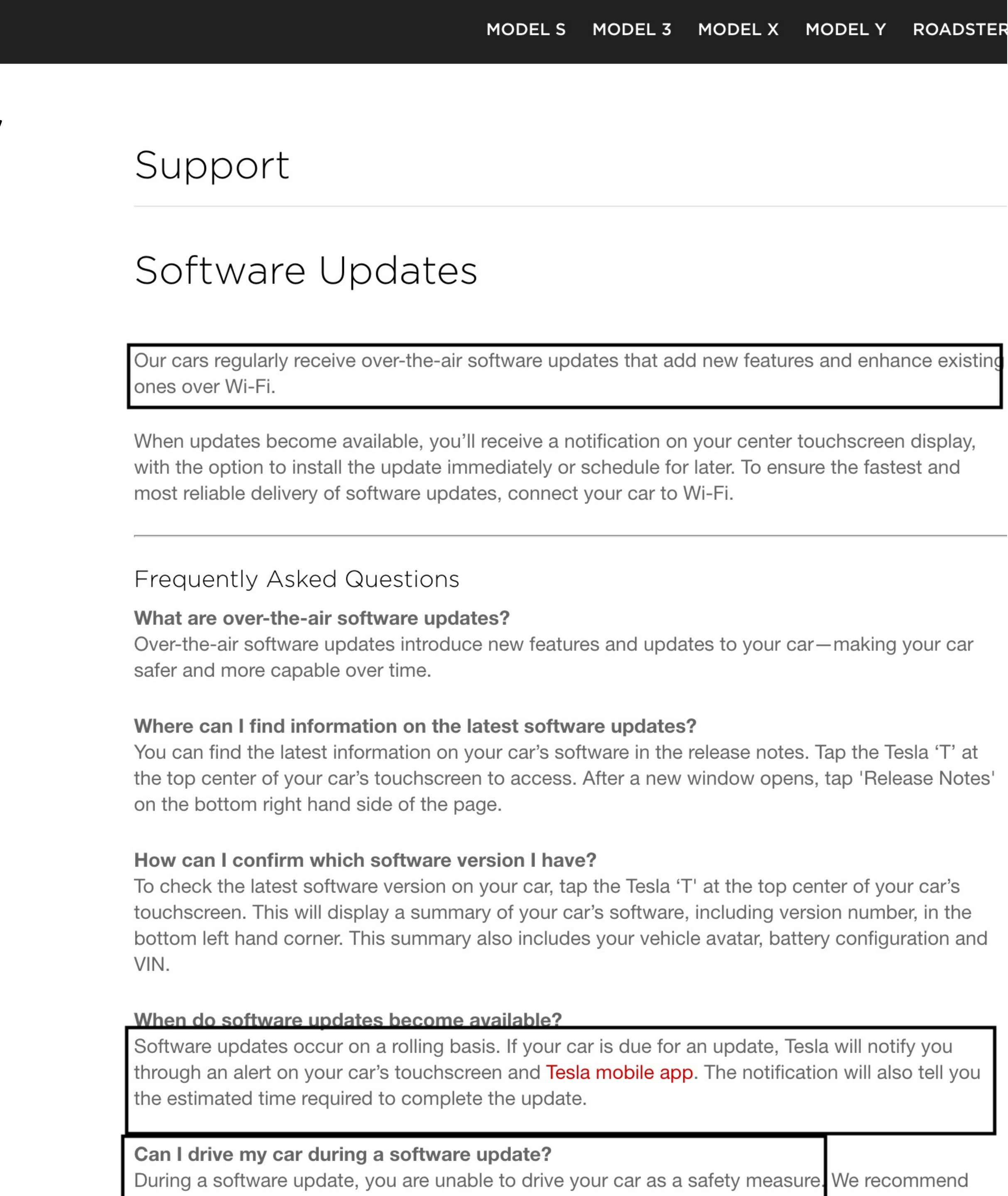
Top 5 Publicly Traded Companies (by Market Cap) 🟩 Tech ⬛ Other

	#1	#2	#3	#4	#5
2001	GE $406B	Microsoft $365B	EXXON $272B	citi $261B	Walmart $260B
2006	EXXON $446B	GE $383B	TOTAL $327B	Microsoft $293B	citi $273B
2011	EXXON $406B	$376B	PetroChina $277B	Shell $237B	ICBC $228B
2016	$582B	Alphabet $556B	Microsoft $452B	amazon $364B	facebook $359B

visualcapitalist.com

1 Andreessen, Marc. "Why Software Is Eating The World." The Wall Street Journal. https://www.wsj.com/articles/SB10001424053111903480904576512250915629460.

TESLA - A software company



Support

Software Updates

Our cars regularly receive over-the-air software updates that add new features and enhance existing ones over Wi-Fi.

When updates become available, you'll receive a notification on your center touchscreen display, with the option to install the update immediately or schedule for later. To ensure the fastest and most reliable delivery of software updates, connect your car to Wi-Fi.

Frequently Asked Questions

What are over-the-air software updates?
Over-the-air software updates introduce new features and updates to your car—making your car safer and more capable over time.

Where can I find information on the latest software updates?
You can find the latest information on your car's software in the release notes. Tap the Tesla 'T' at the top center of your car's touchscreen to access. After a new window opens, tap 'Release Notes' on the bottom right hand side of the page.

How can I confirm which software version I have?
To check the latest software version on your car, tap the Tesla 'T' at the top center of your car's touchscreen. This will display a summary of your car's software, including version number, in the bottom left hand corner. This summary also includes your vehicle avatar, battery configuration and VIN.

When do software updates become available?
Software updates occur on a rolling basis. If your car is due for an update, Tesla will notify you through an alert on your car's touchscreen and Tesla mobile app. The notification will also tell you the estimated time required to complete the update.

Can I drive my car during a software update?
During a software update, you are unable to drive your car as a safety measure. We recommend

Software Engineering Goals

- Increase Productivity

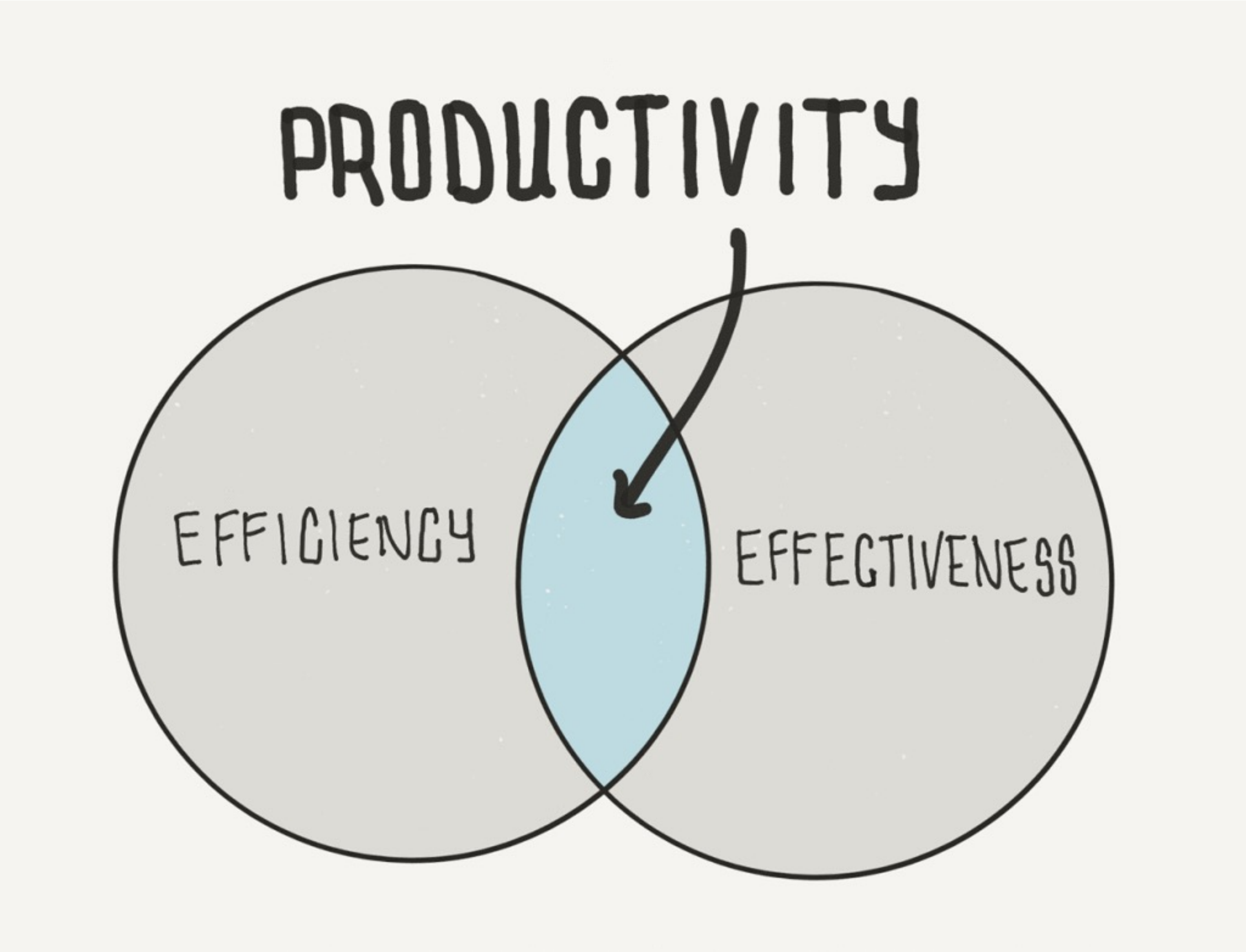
- Improve Quality

Why is software development productivity important?

Large-Scale Software Engineering

- Fixed release schedule.
- Fixed resources (time and money)
- Way too much work to finish.
- Develop more valuable products for lower costs ($)
- Decrease time-to-market for competition
- Hard to find (good) developers

How to achieve Productivity



Doing things right vs. Doing the Right Things

Engineering Complexity

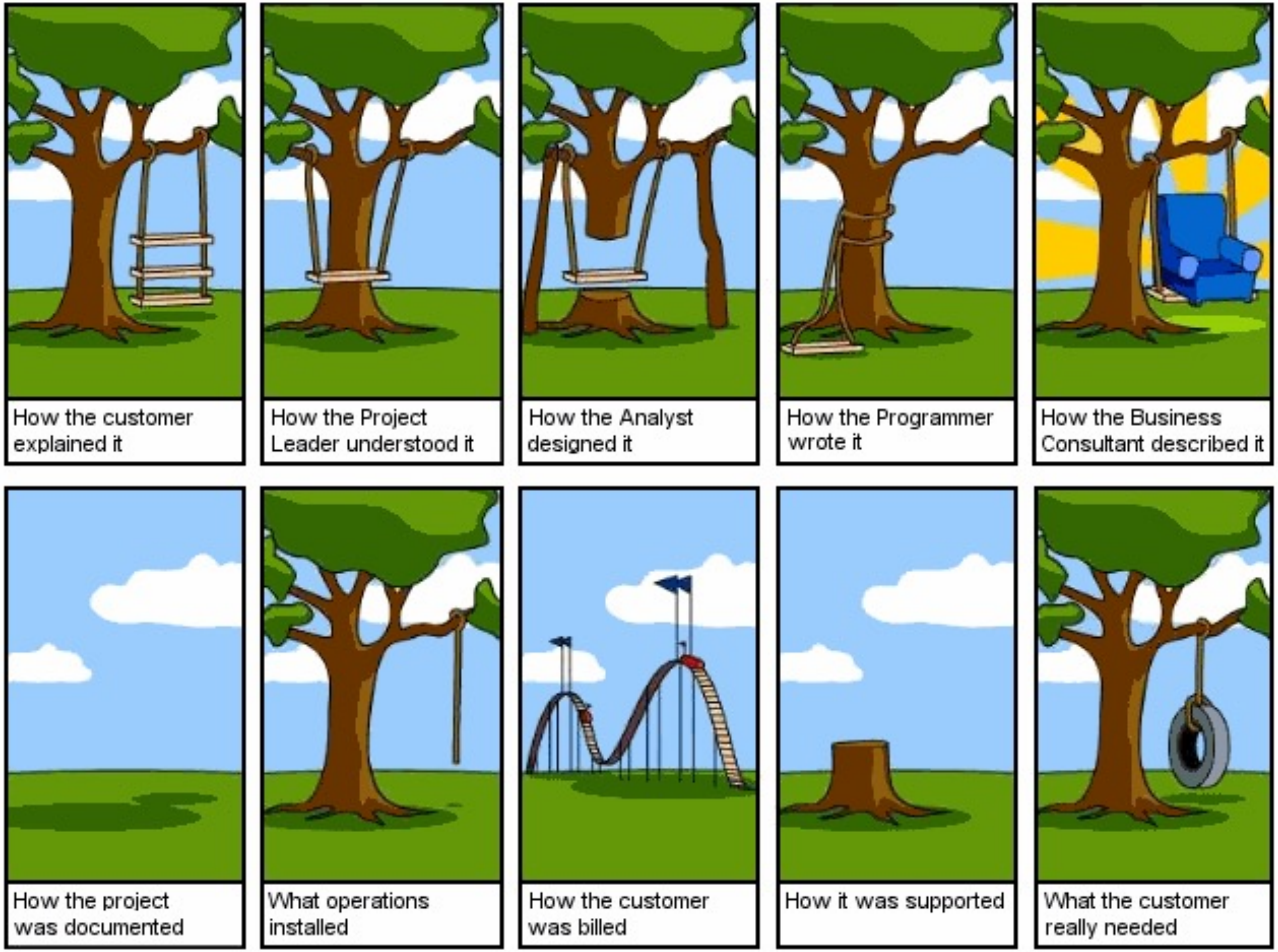
7.300.000 kg,
2.5 million rivets, all of them as old as the tower itself, hold the latticed structure together.

Engineering Complexity

5-10 million lines of code

The failure of any line of code could lead to total system failure

Space of inputs even to the smallest app on the phone > 10^{80}



How the customer explained it

How the Project Leader understood it

How the Analyst designed it

How the Programmer wrote it

How the Business Consultant described it

How the project was documented

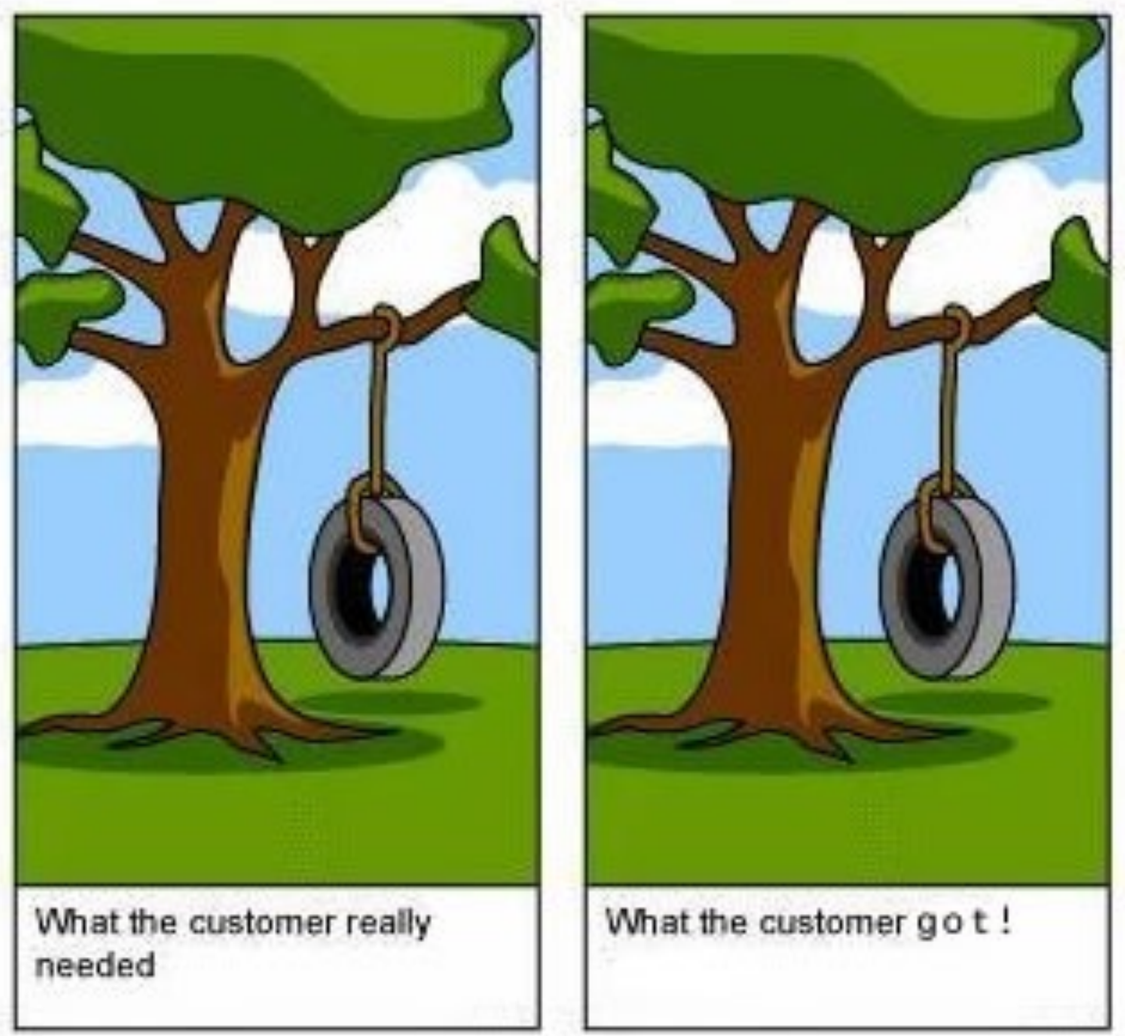
What operations installed

How the customer was billed

How it was supported

What the customer really needed

Software development is complex



Software Decision Making

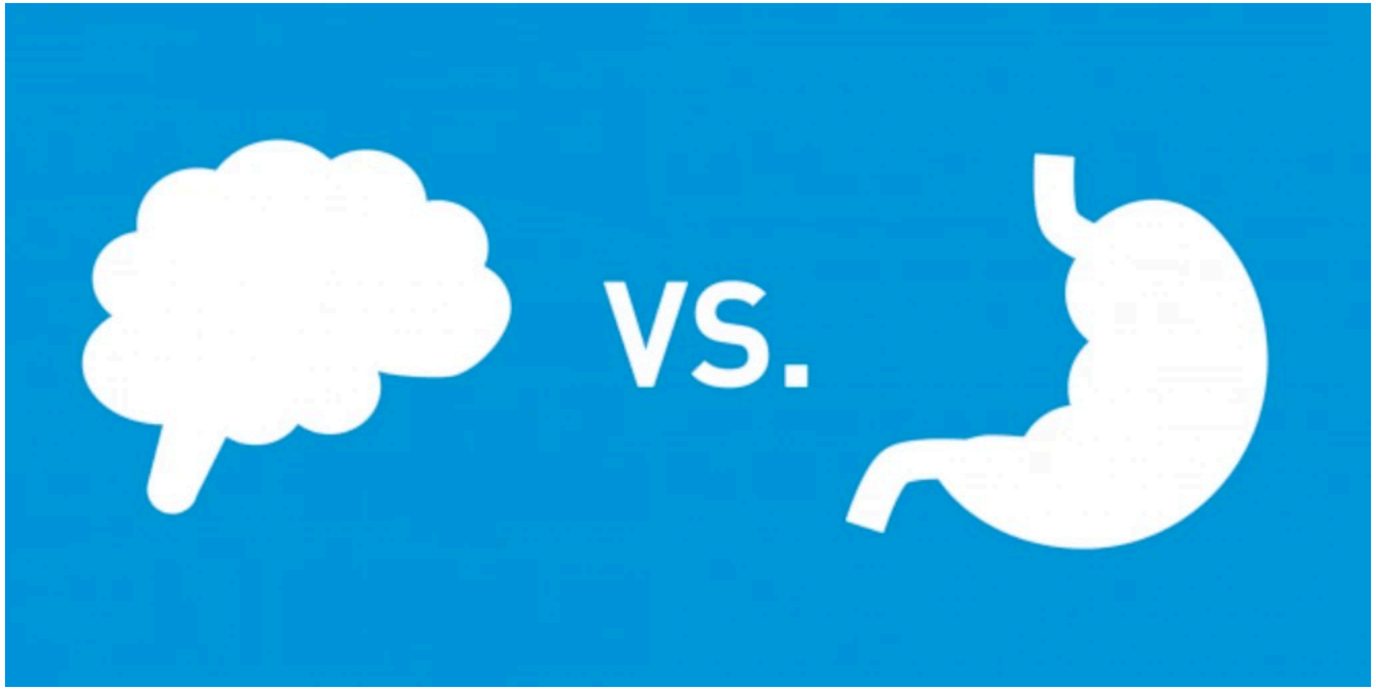
Software Practitioners rely on their **prior experiences** to plan software project projects, fix bugs, prioritize testing, etc.

What are key “Decisions” we are after?

- Guess the location of undetected bugs
- Effort Estimation
- Who should review my code?
- Finding duplicated bugs
- Finding potential reopen bugs
- Who is the Most Valuable Developer in my team?
- Which features are mostly used in my application?
- Who in the team is most likely to leave in the next 2-4 weakest?

- When will this project be ready to ship?
- Which components of our application most need to be tested?
- Who should fix this bug?
- What parts of my API do people find hardest to use?
- Will My Patch Be Accepted?
- Finding code clones
- Why my bugs are reopened?
- ….

Decision Making



Data Driven vs Gut Feeling

Data Science

The field of study that uses various **methods** to extract useful insights and knowledge from the **data** to make data-driven **decisions**.

Methods can include/require; domain expertise, programming skills (i.e. scripting to process data), statistical modelling (i.e. machine learning algorithms), visualization techniques.

What is Data Science?

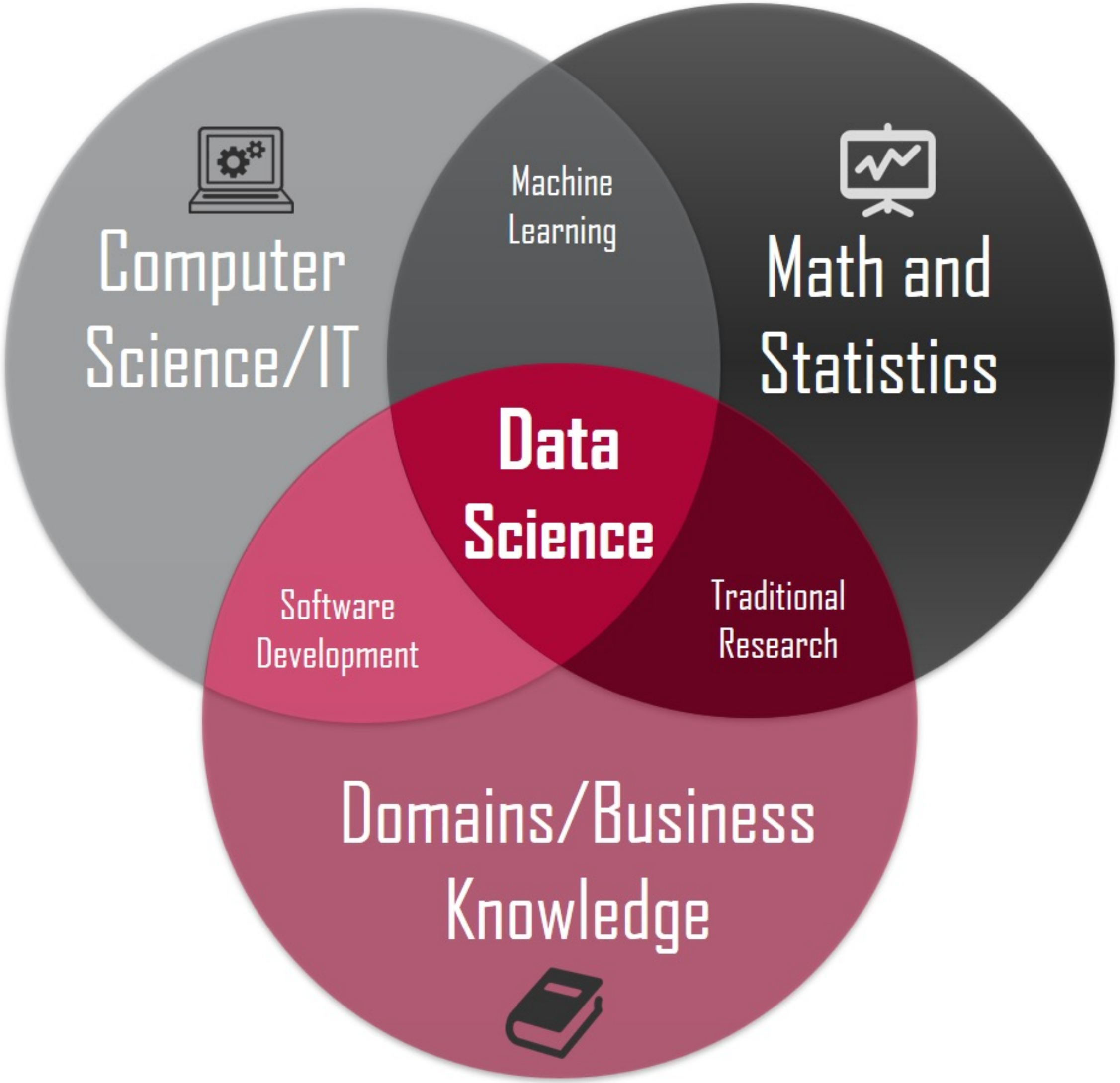
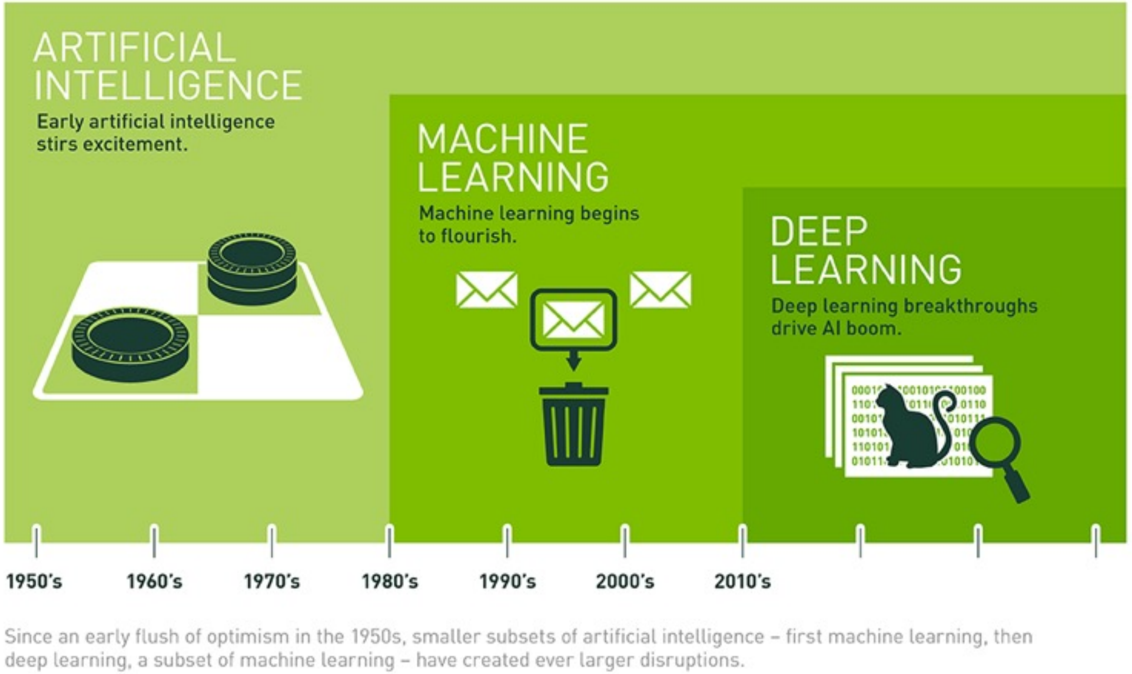


Image source: Rob Tibshirani, Stanford Stats 101


ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's

Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Data Science makes use of AI, ML, DL

https://blogs.nvidia.com/blog/2016/07/29/whatsdifference-artificial-intelligence-machine-learningdeep-learning-ai/

Data Science in Software Engineering

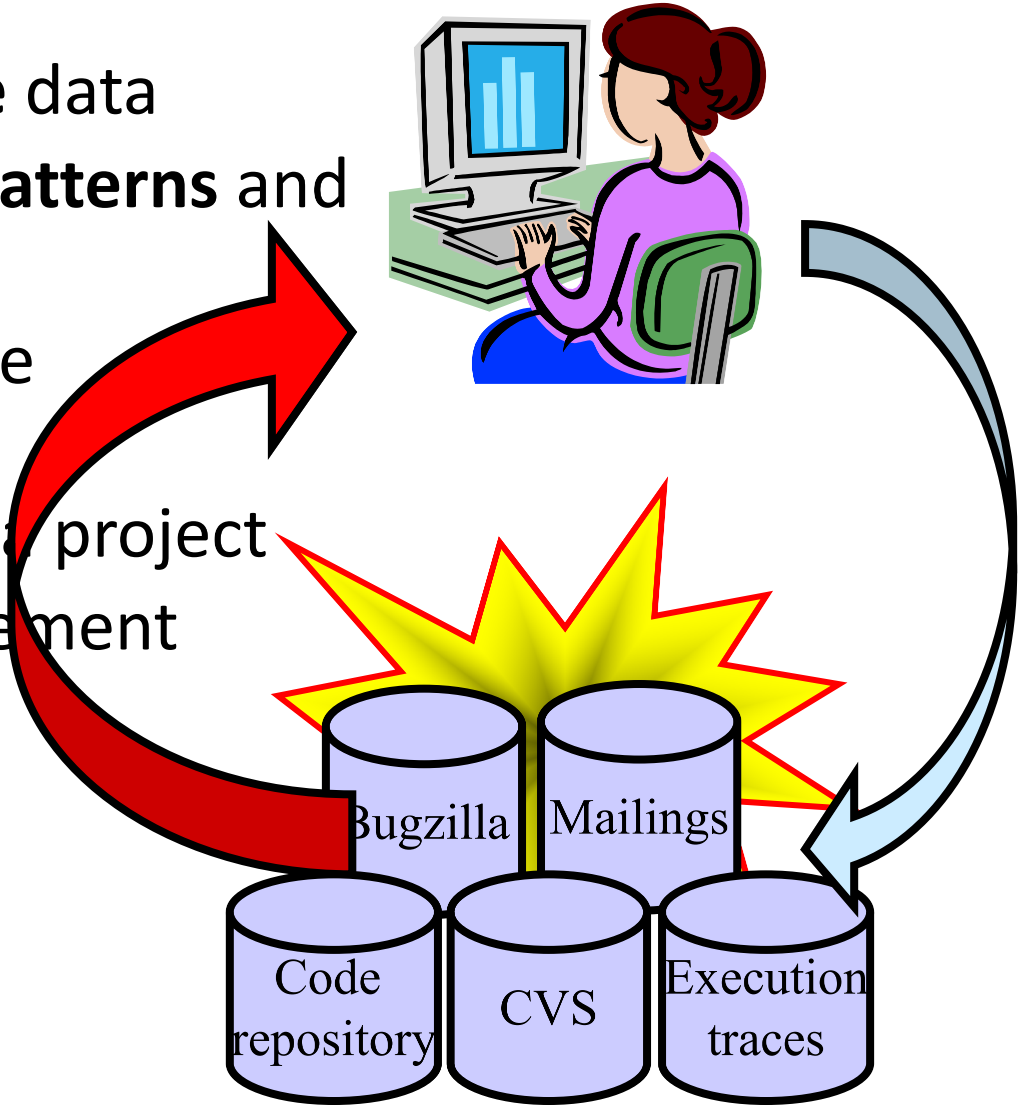
Transform static record-keeping SE data to **active** data

Make SE data actionable by uncovering hidden **patterns** and **trends**

Gain empirically-based understanding of software development

Predict, plan, and understand various aspects of a project

Support future development and project management activities



Definitions

“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.”

Menzies T, Zimmermann T. Software analytics: so what? IEEE Software 2013;30(4):31-7.

“Software analytics is to enable software practitioners to perform data exploration and analysis in order to obtain insightful and actionable information for data-driven tasks around software and services (and software practitioners typically include software developers, testers, usability engineers, and managers, etc.).”

Zhang D, Dang Y, Lou J-G, Han S, Zhang H, Xie T. Software analytics as a learning case inpractice: approaches and experiences. MALETS 2011.

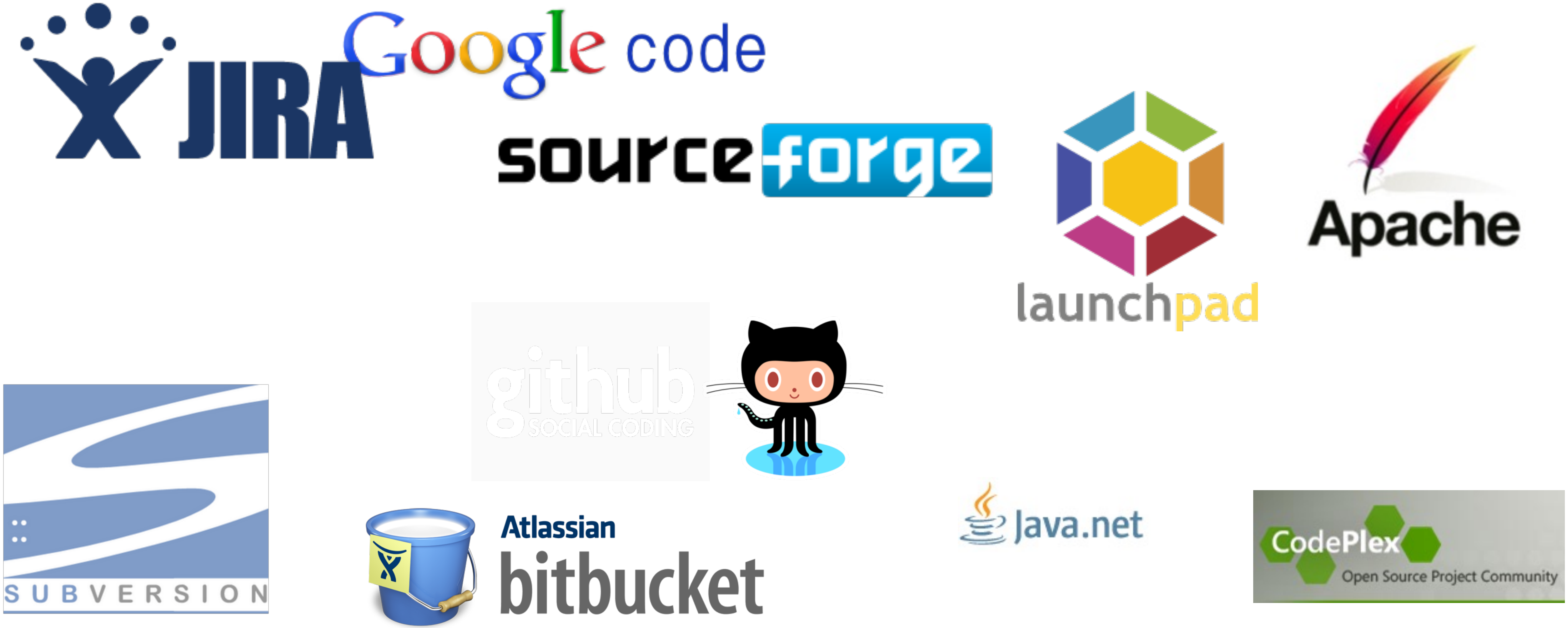
Terminology

- Data Science in Software Engineering
- Software (Data) Analytics
- Software Intelligence
- Software Development Analytics
- Mining Software Repositories

What do we mean by “data” in Software Engineering?

- Large amount of artefacts are generated in the software development process
- Increased amount of data available in software archives through large open-source projects

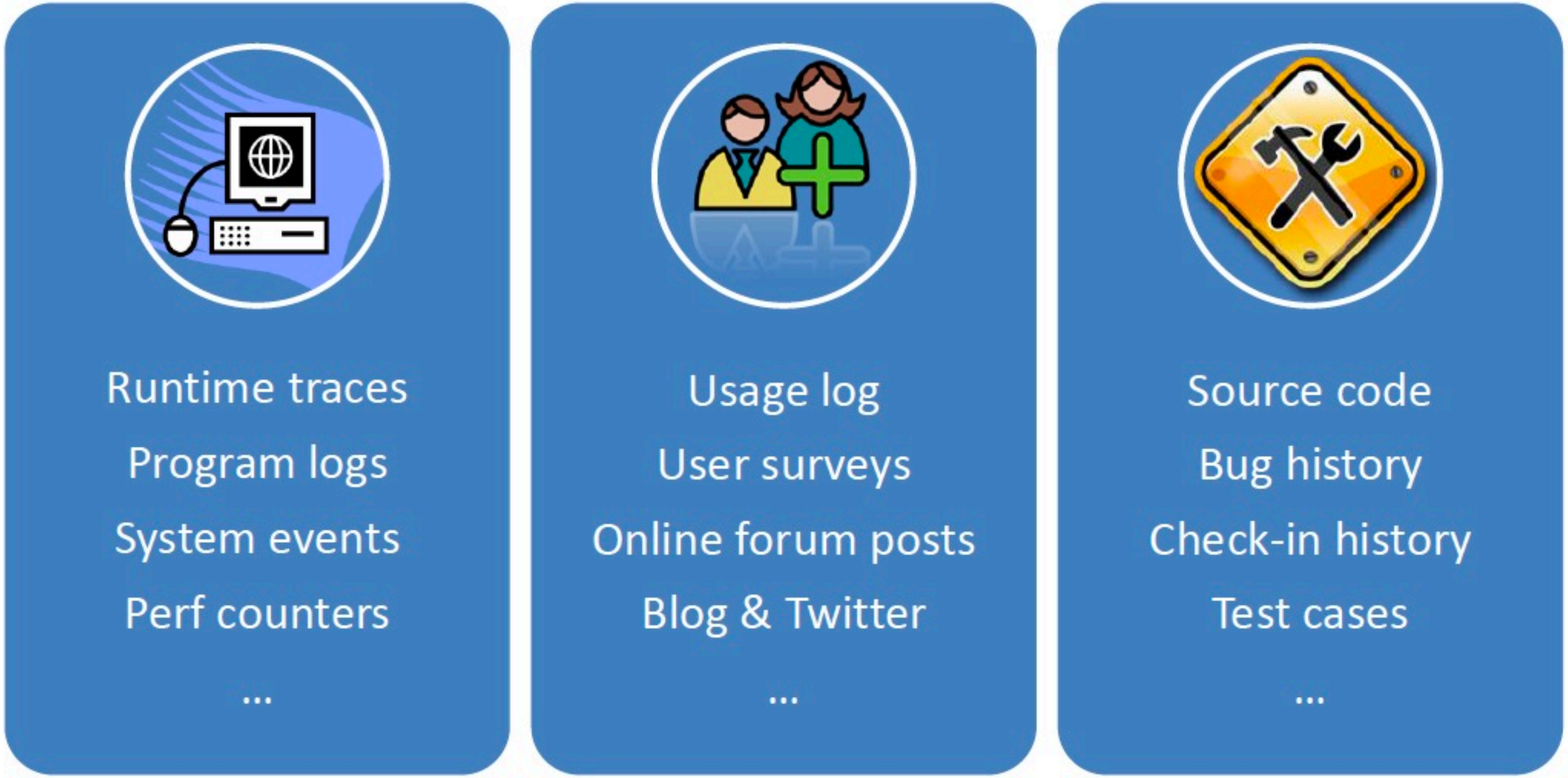
Where is the Data?



How does the data look like?

- https://github.com/apache/hadoop/pulls
- https://issues.apache.org/jira/projects/HADOOP/issues/HADOOP-16856?filter=resolvedrecently

Types of SE Data



Runtime traces

Program logs

System events

Perf counters

…

Usage log

User surveys

Online forum posts

Blog & Twitter

…

Source code

Bug history

Check-in history

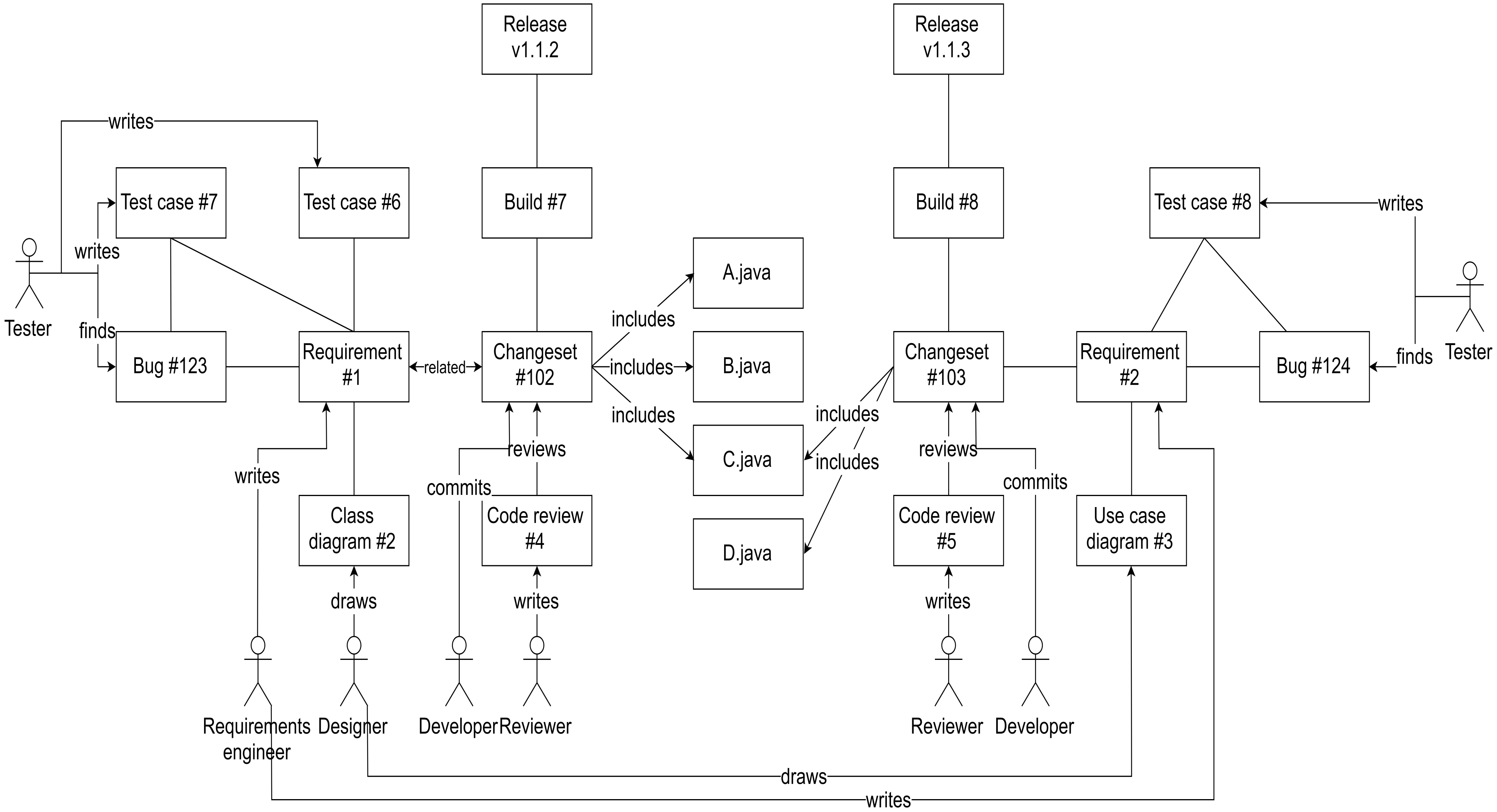
Test cases

…

How big is the data?

- Size
  - Linux >10M LOC
- Developer
  - >1K in Windows
- Code comments
  - 2 Million in Eclipse, 1 million in Mozilla
- Commit logs
  - 222K for Linux, 31K for PostgreSQL
- Bug reports
  - 641K in Mozilla, 18K in Linux, 7K in Apache

Software Data is Connected



Who uses the data?

Target audience – software practitioners

Program Manager

Developer

Management personnel

Designer

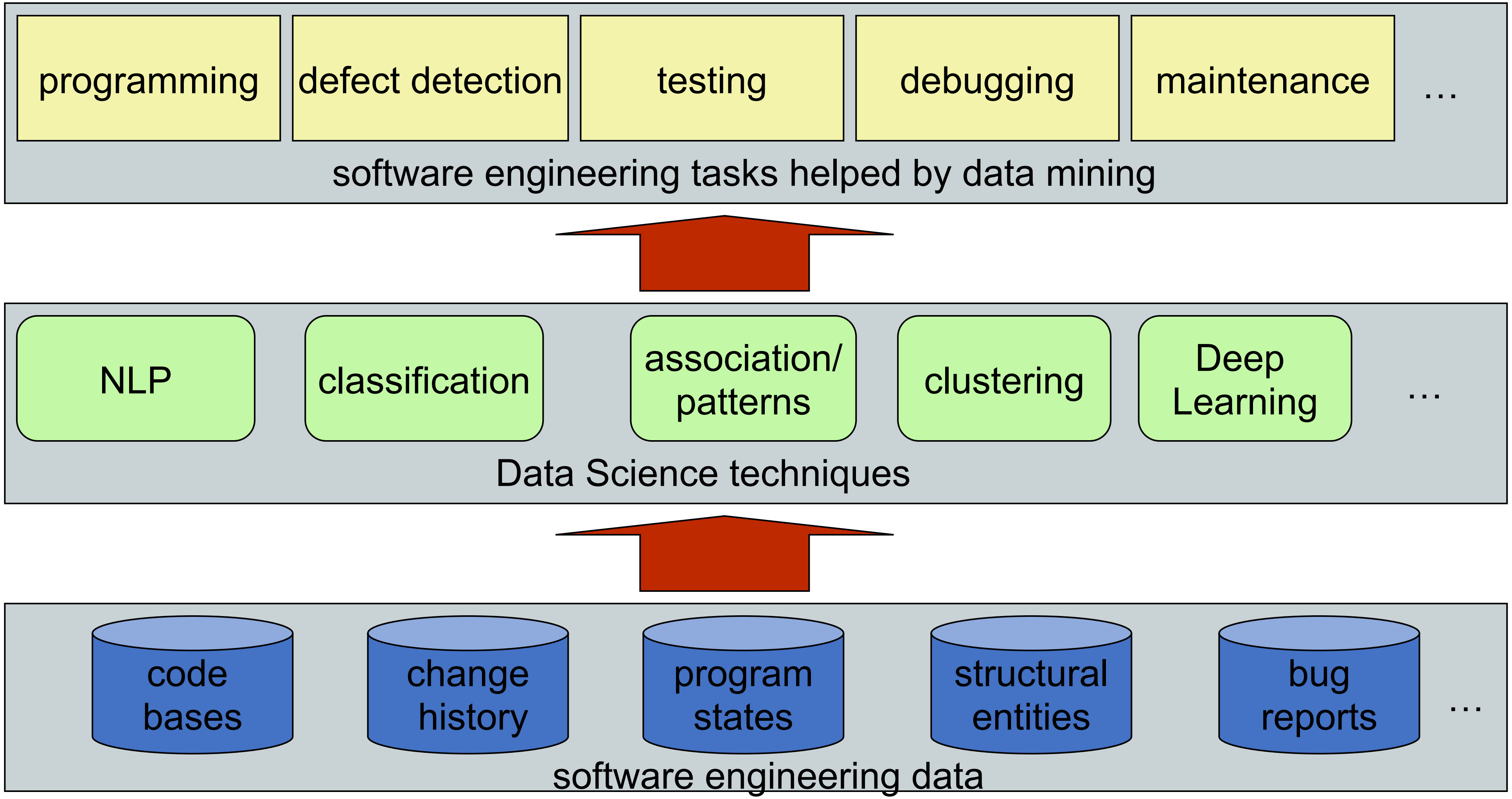
Tester

Support engineer

Operation engineer

Usability engineer

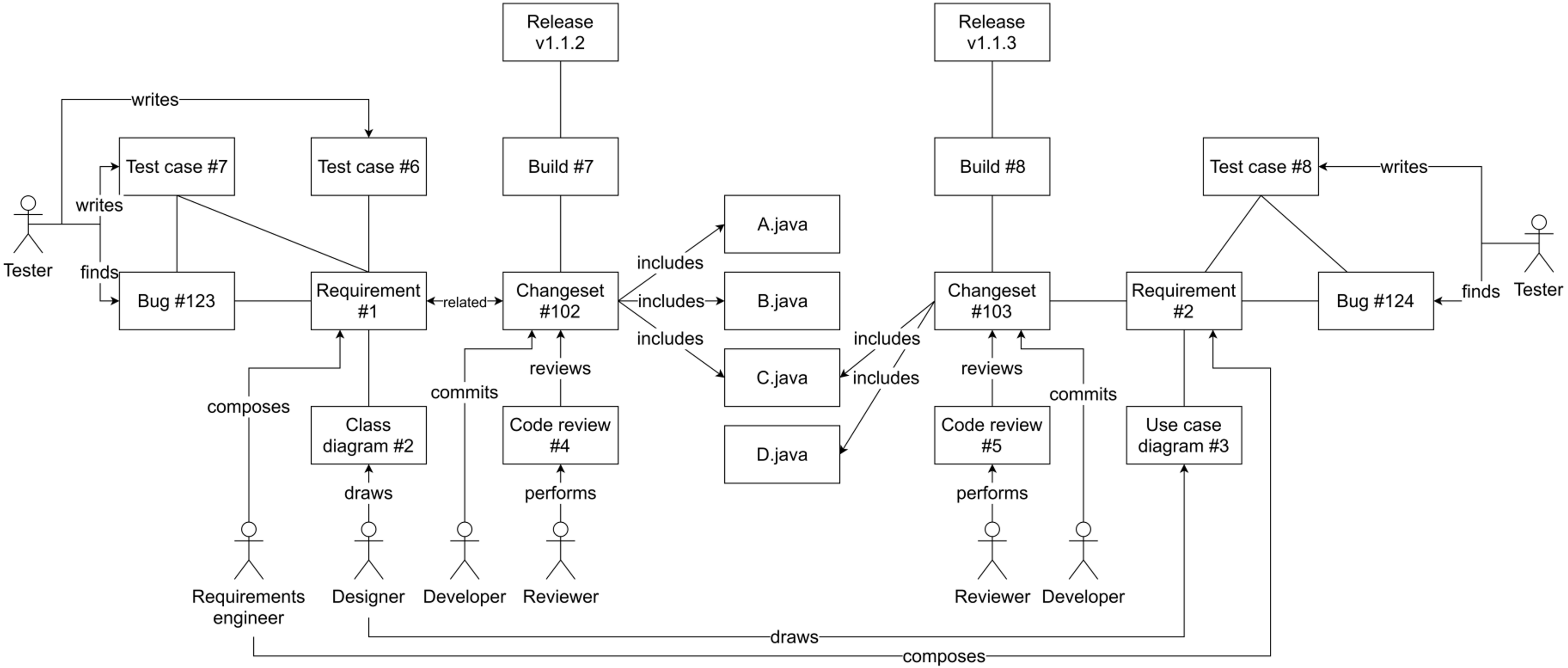
Overview of Data Science in SE



Questions

- How can we make data useful to a wide audience, not just to developers but to anyone involved in software?
- What can we learn from the vast amount of unexplored data?
- How can we learn from incomplete or biased data?
- How can we better tie usage analytics to development analytics?
- When and what lessons can we take from one project and apply to another?
- How can we establish smart data science as a discipline in software engineering practice and research as well as education?

Software Artifacts Traceability Graph



How does RSTrace work?

Create an artifact traceability graph using the available repositories

For an artifact that requires reviewer recommendation

- Run “the algorithm” that assigns scores to developer-artifact pairs
- Sort developers by scores in descending order
- Recommend top-k developers

The Algorithm

We define a new metric called **know-about** that measures the knowledge of a developer about an artifact (e.g. source code file).

Know-about score of a developer for an artifact is calculated as follows:

$$KA_{Developer-Artifact} = \sum_i \frac{1}{Length(Path_i(Developer - Artifact))}$$

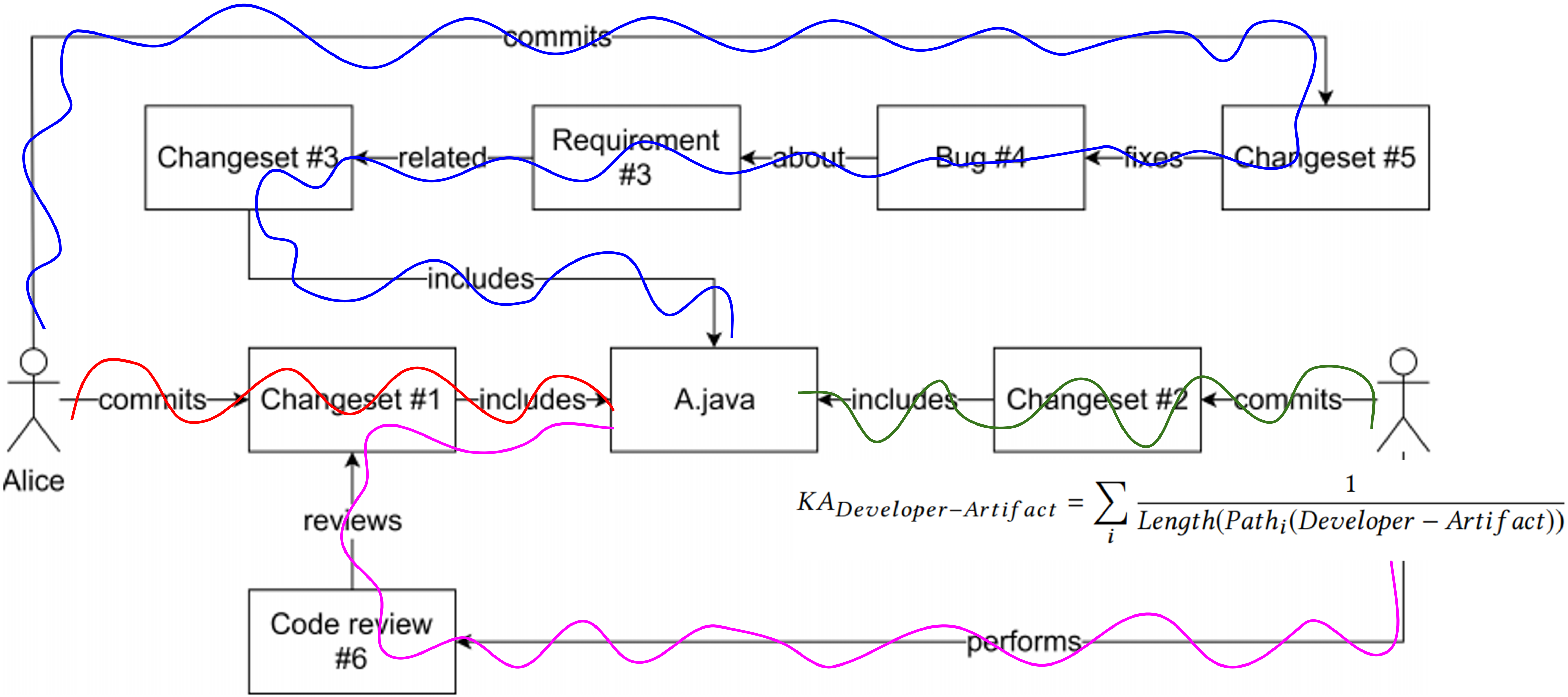
Know-about score of a developer for a changeset is calculated as follows:

$$KA_{Developer-Changeset} = \sum_i KA_{Developer-Artifact(i)}$$

Let's see an example

$$where\ Artifact(i) \in Changeset$$

Calculate know-about score of Alice and Bob for A.java



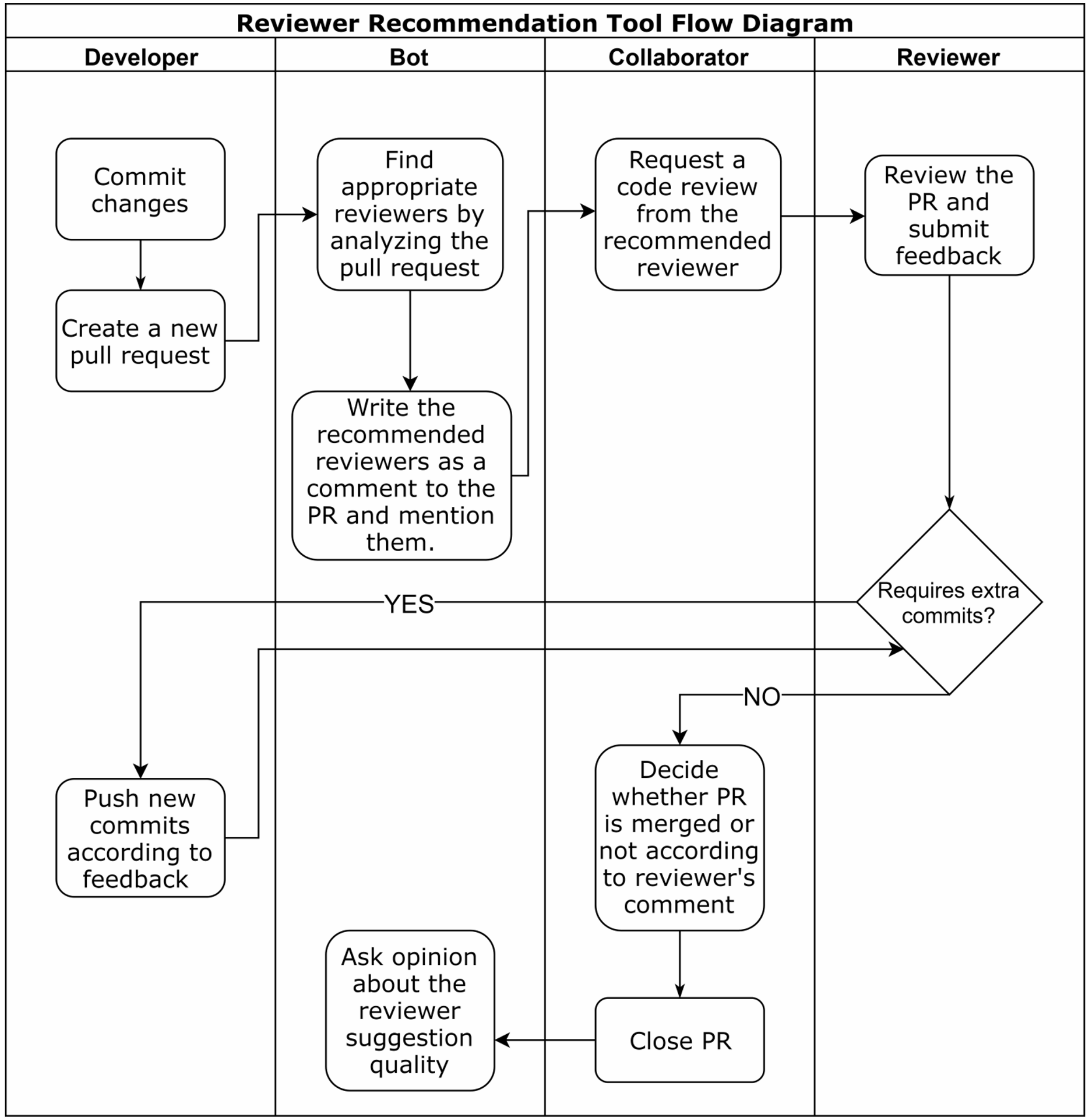
$$KA_{Developer-Artifact} = \sum_{i} \frac{1}{Length(Path_i(Developer - Artifact))}$$

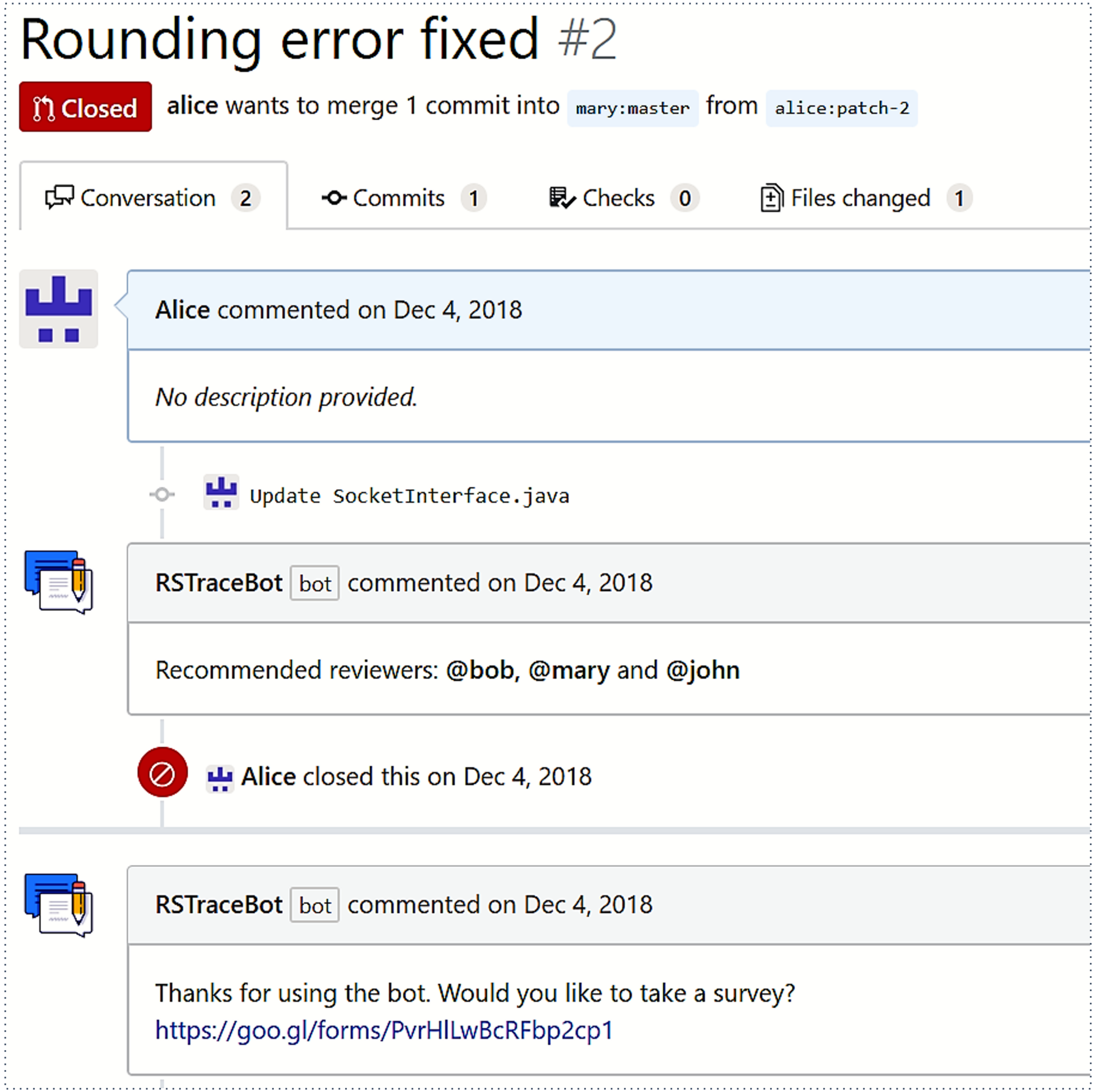
Alice: $1/2 + 1/5 = 0.70$

Bob: $1/2 + 1/3 = 0.83$

Bob is a better reviewer for *A.java* file because his know-about score is higher than Alice's

GitHub Bot


Reviewer Recommendation Tool Flow Diagram
Developer
Bot
Collaborator
Reviewer
Commit changes
Create a new pull request
Find appropriate reviewers by analyzing the pull request
Write the recommended reviewers as a comment to the PR and mention them.
Request a code review from the recommended reviewer
Review the PR and submit feedback
Requires extra commits?
YES
NO
Push new commits according to feedback
Decide whether PR is merged or not according to reviewer's comment
Close PR
Ask opinion about the reviewer suggestion quality



Rounding error fixed #2

Closed alice wants to merge 1 commit into `mary:master` from `alice:patch-2`

Alice commented on Dec 4, 2018

No description provided.

`Update SocketInterface.java`

RSTraceBot bot commented on Dec 4, 2018

Recommended reviewers: **@bob**, **@mary** and **@john**

Alice closed this on Dec 4, 2018

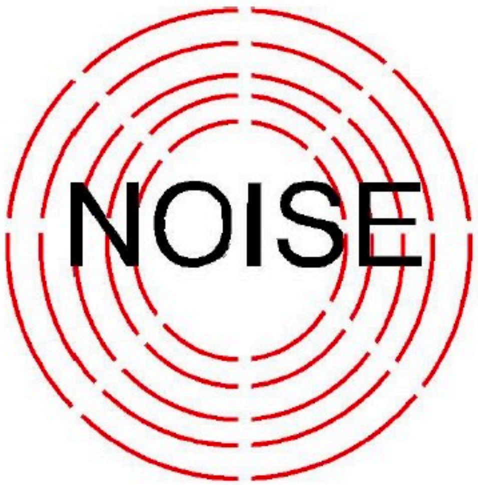
RSTraceBot bot commented on Dec 4, 2018

Thanks for using the bot. Would you like to take a survey?
https://goo.gl/forms/PvrHlLwBcRFbp2cp1

Understand the Application Domain

Domain Knowledge + Close(r) Inspection

- Make sure you manually examine the repositories. Do not fully automate the process!



Data Science Pipeline

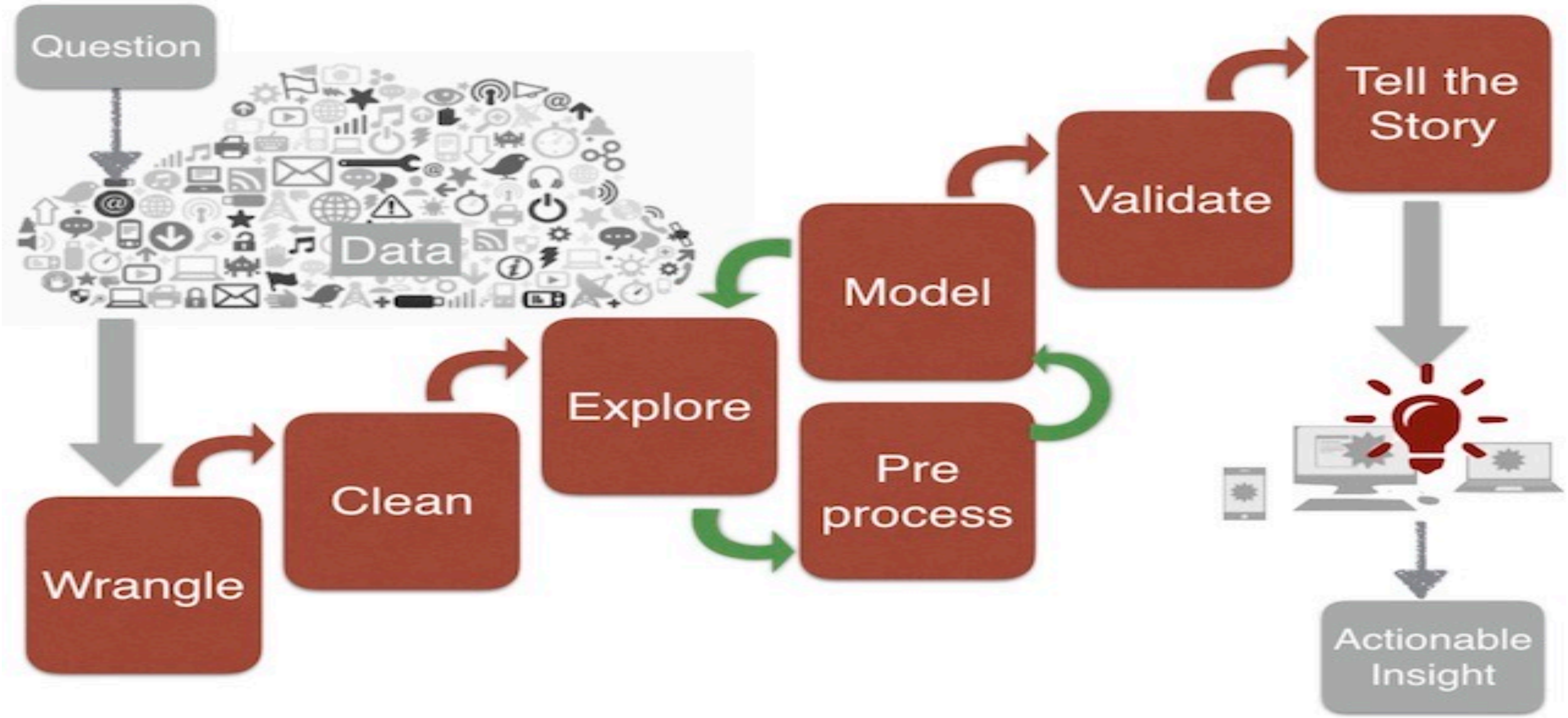
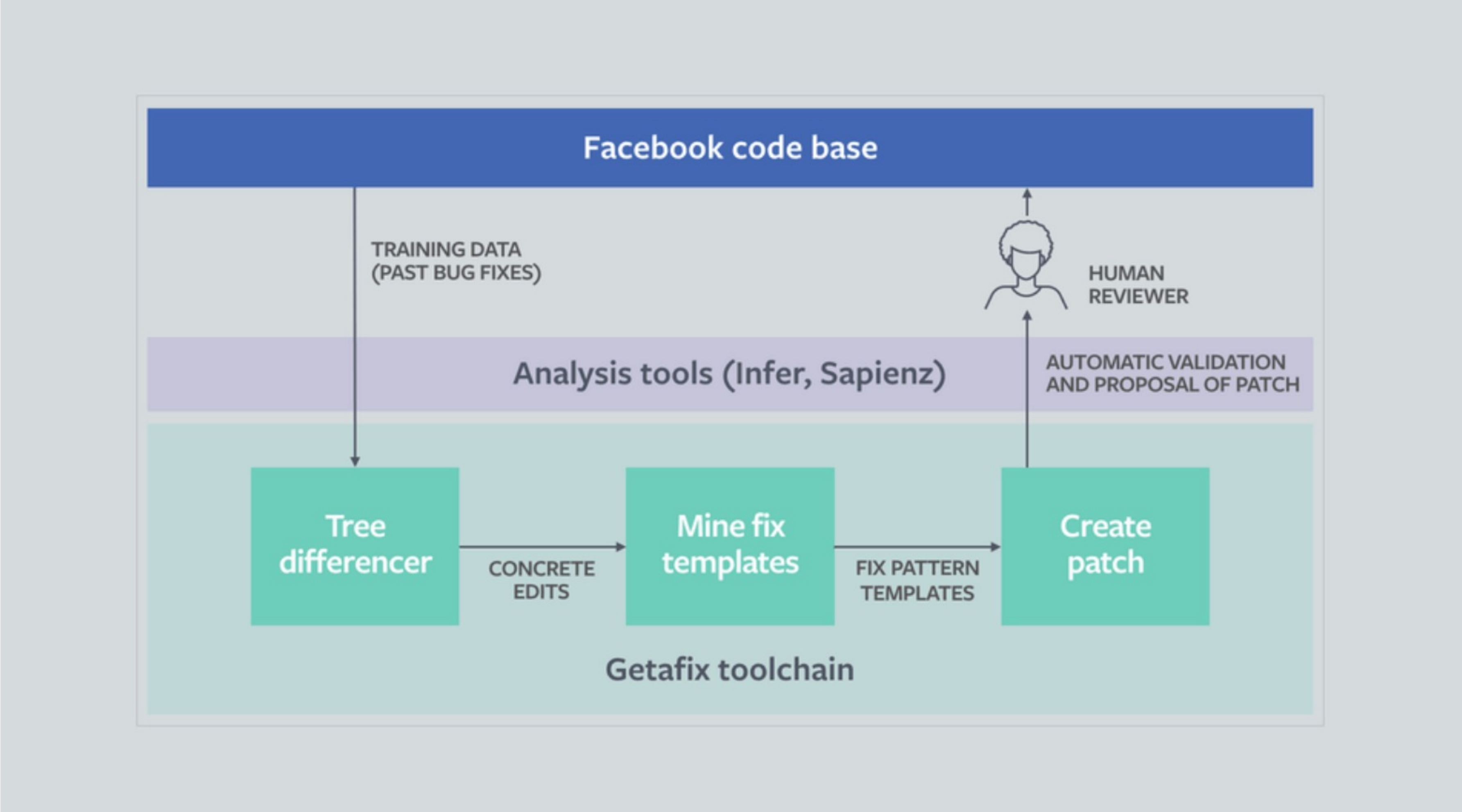


Image credit Wolfram Research

Facebook - How Facebook tools learn to fix bugs automatically

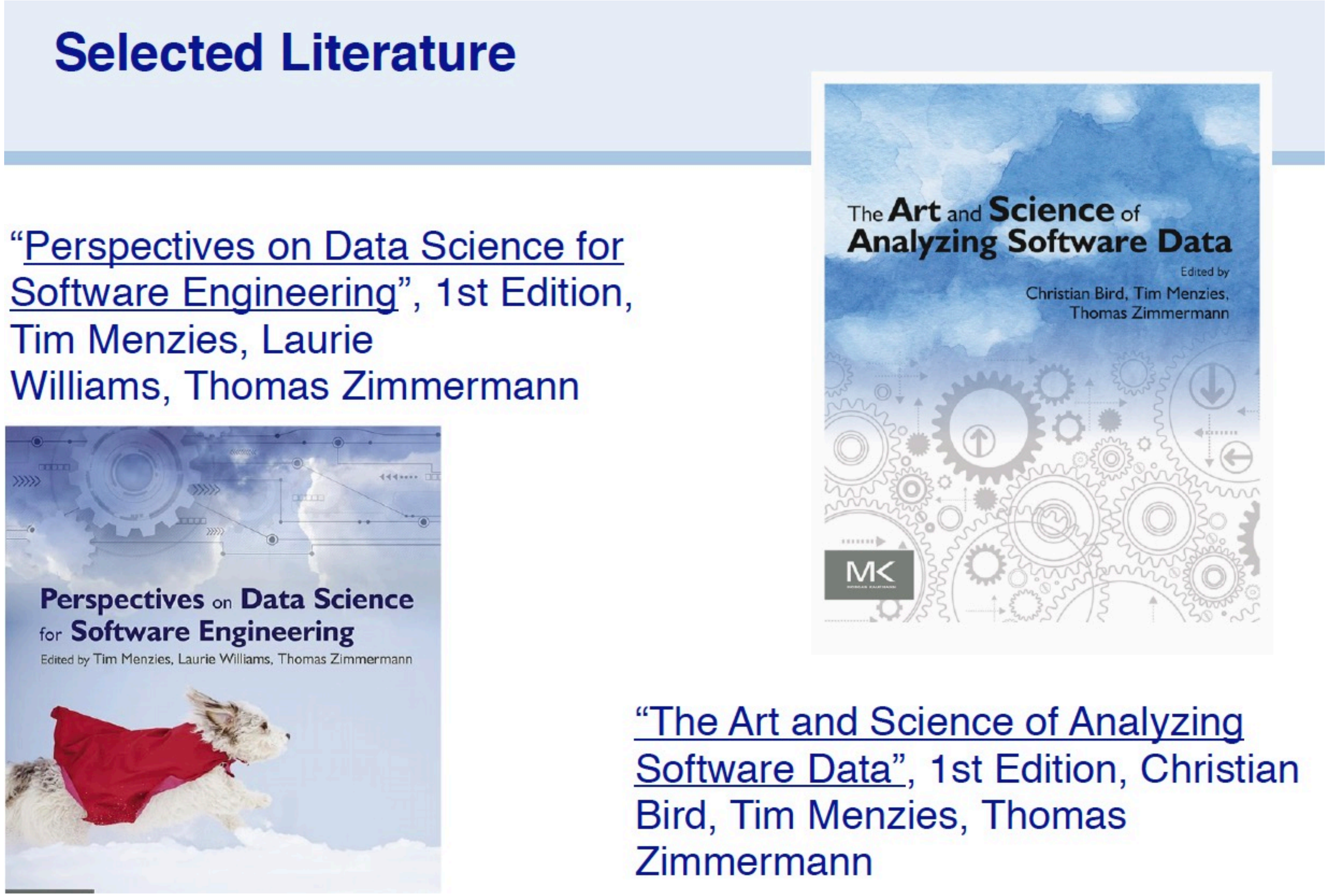


https://engineering.fb.com/developer-tools/getafix-how-facebook-tools-learn-to-fix-bugs-automatically/

GitHub CoPilot

- https://copilot.github.com

References



Selected Literature

"Perspectives on Data Science for Software Engineering", 1st Edition, Tim Menzies, Laurie Williams, Thomas Zimmermann

"The Art and Science of Analyzing Software Data", 1st Edition, Christian Bird, Tim Menzies, Thomas Zimmermann

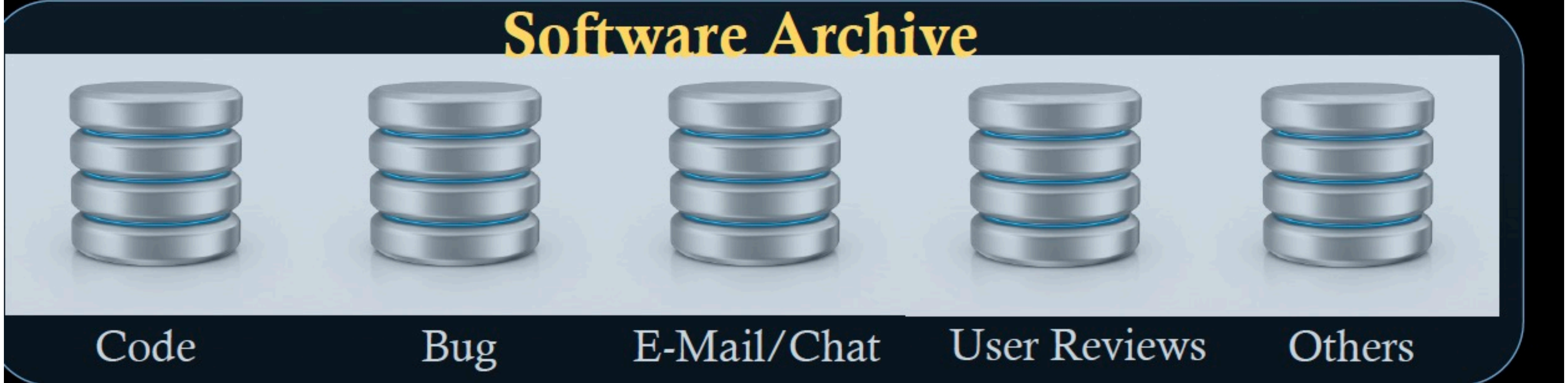
The Art and Science of Analyzing Software Data
Edited by
Christian Bird, Tim Menzies, Thomas Zimmermann
MK
Perspectives on Data Science for Software Engineering
Edited by Tim Menzies, Laurie Williams, Thomas Zimmermann

Data Science in Software Engineering

Analyze software data

Make informed data-driven decisions

Supporting decision making using facts instead of fortune tellers!



Software Archive

Code

Bug

E-Mail/Chat

User Reviews

Others

Slide Credits

- Tao Xie
- Ahmed Hassan