

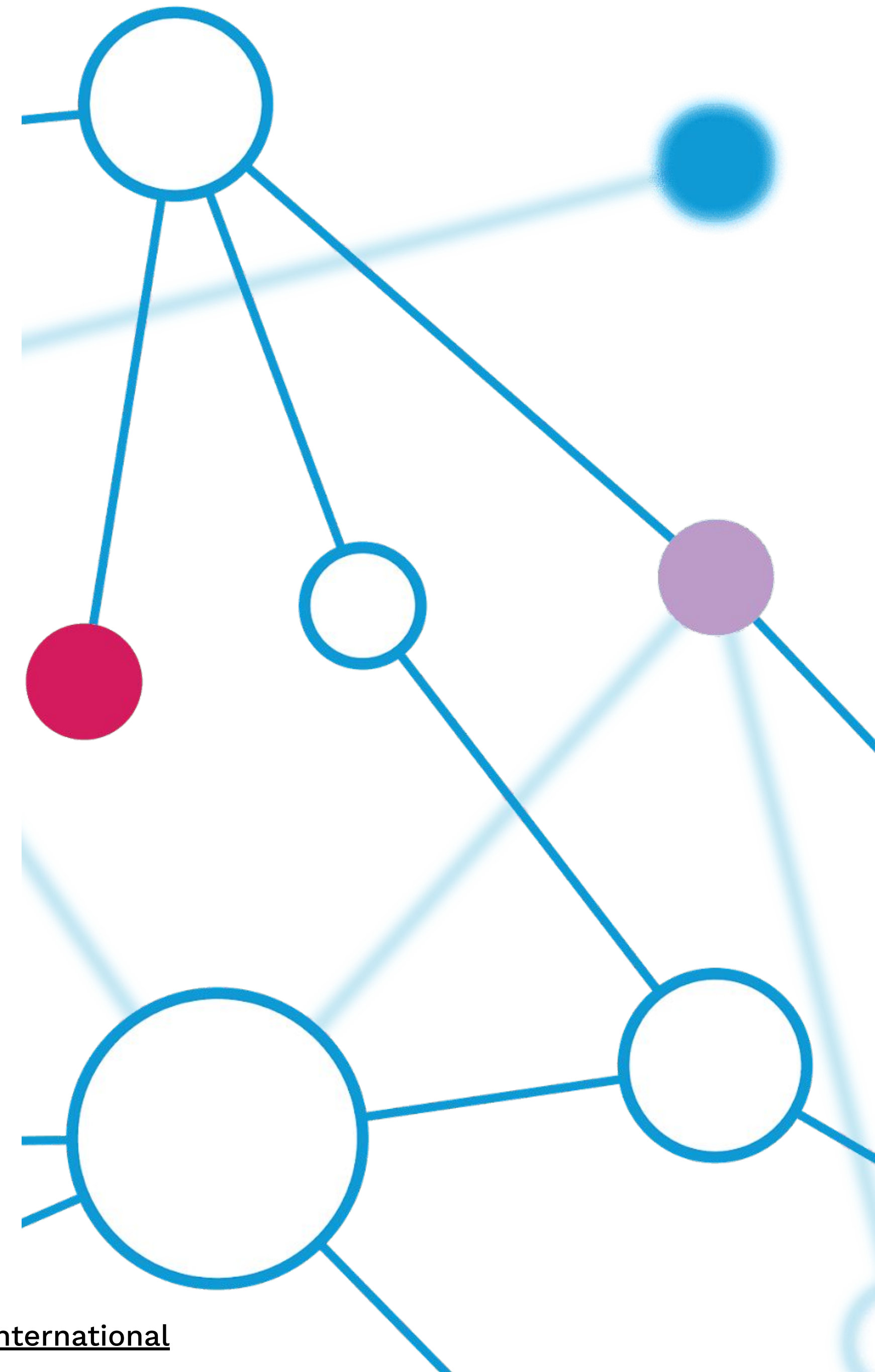


From Data Tsunami to Actionable Insights

SCALE 2024

Cali Dolfi, Red Hat
Dawn Foster, CHAOSS

 <https://chaoss.community/>
 <https://github.com/chaoss>
 @chaoss@fosstodon.org



Introductions



Cali Dolfi

Senior Data Scientist, Red Hat

www.linkedin.com/in/calidolfi



Dawn Foster

Data Science, CHAOSS

fastwonderblog.com/

hachyderm.io/@geekygirldawn

www.linkedin.com/in/dawnfoster

Data Tsunami

People can be overwhelmed
by a wall of metrics
and unsure how to start



Metrics Models

Using collections of metrics to focus on a particular topic

What is This?
Metrics related to outbound / upstream contributions to open source projects whether developed by your org or a 3rd party (e.g., development culture, collaboration, and DEI).

Metric Model: Starter Project Health

- Time to First Response
- Change Request Closure Ratio
- Bus Factor
- Release Frequency

CONTRIBUTION



This figure is intended to help organizations quickly and easily consider key open source community health related issues. The figure has two key areas of contribute and consume, and ways that you can improve and sustain these two areas.

MONITOR

What is This?
An ongoing process of monitoring metrics to see if improvements are effective

What is This?
An ongoing process using additional metrics to find more ways to improve

IMPROVE

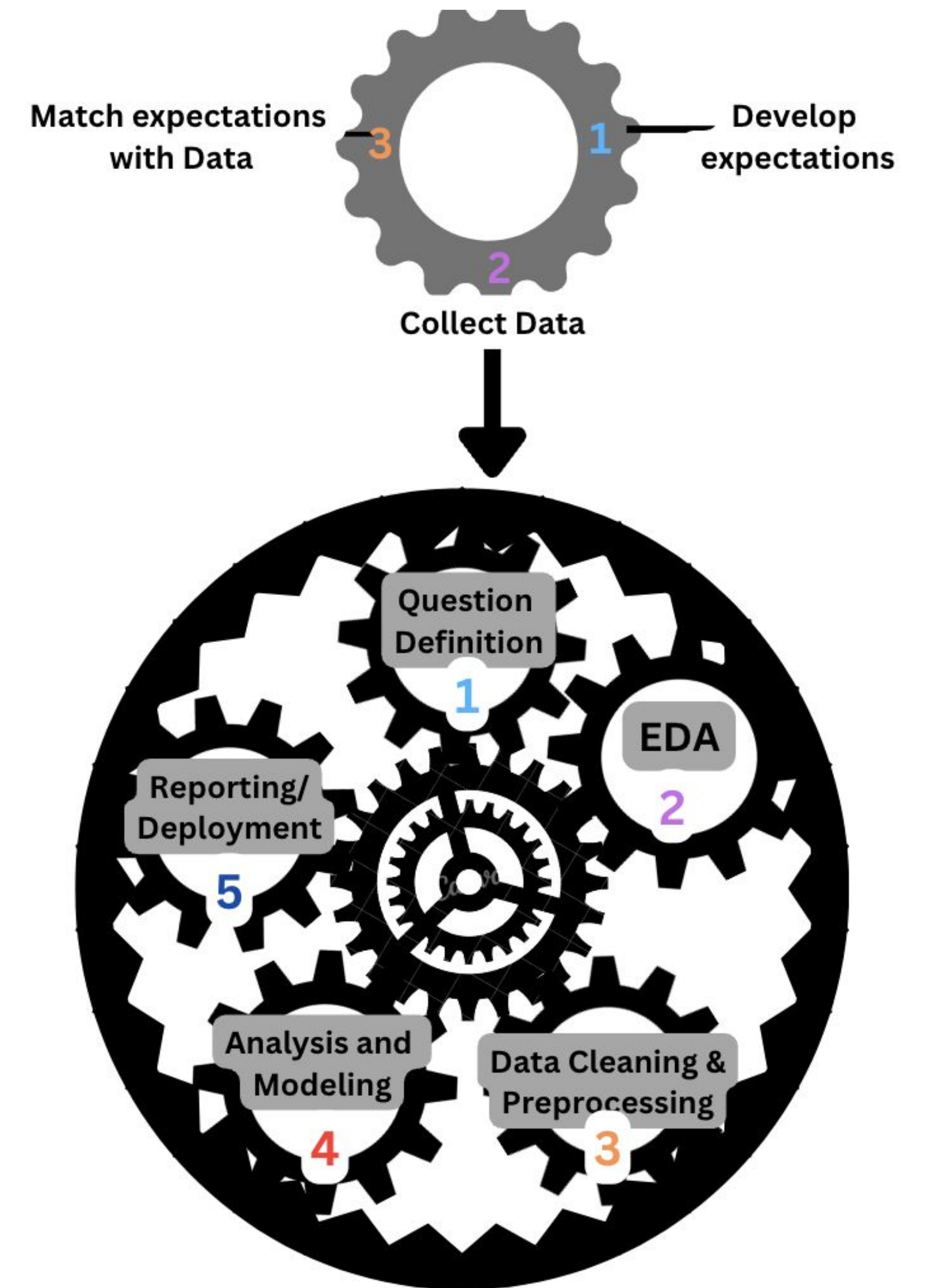
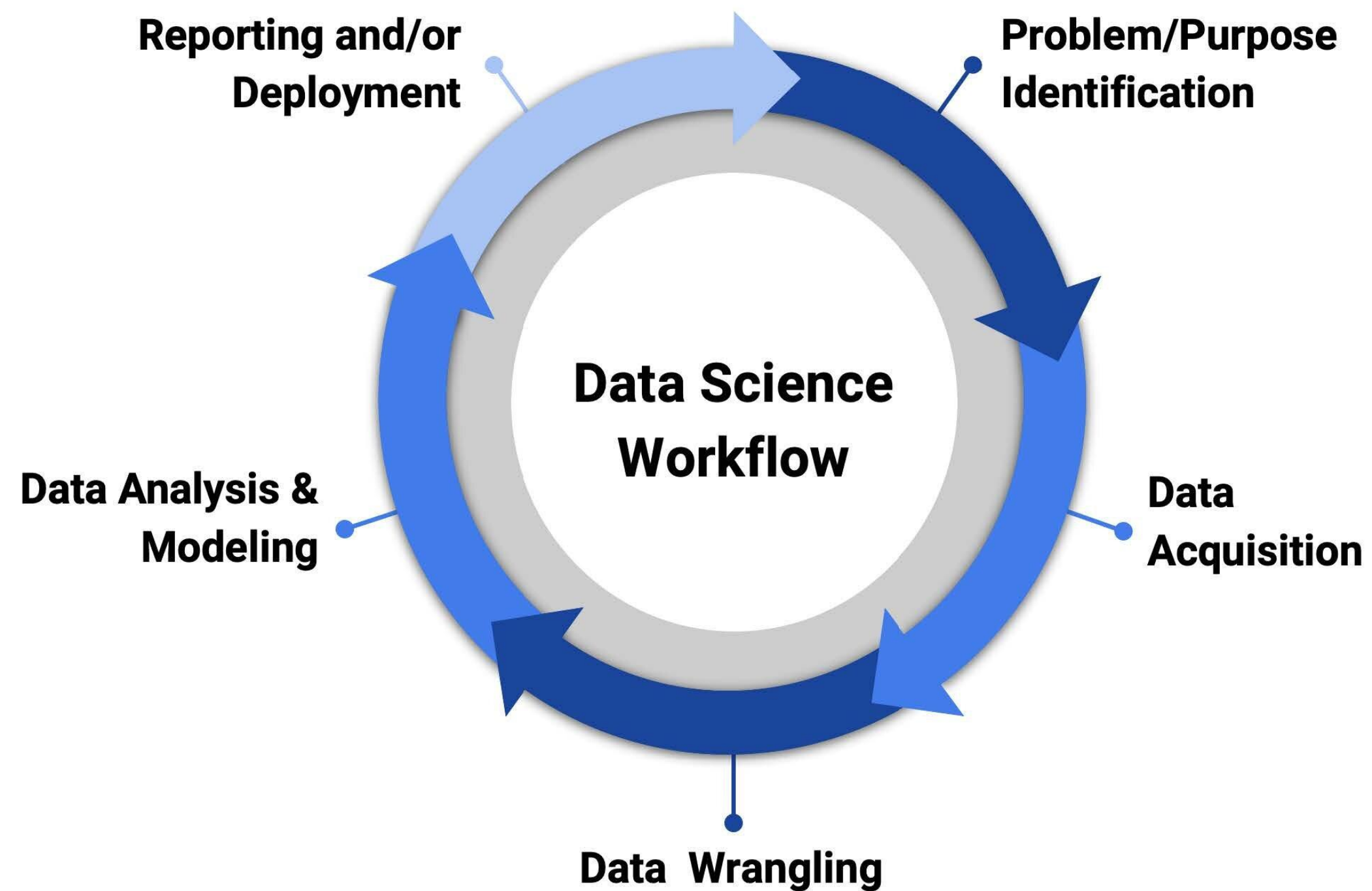
CONSUMPTION

What is This?
Metrics related to inbound / downstream consumption of open source software within an organization's products, services, and infrastructure (e.g., compliance, procurement, and viability).

Metric Model: OSS Project Viability Starter

- Bus Factor
- Elephant Factor
- Change Requests
- Change Request Closure Ratio
- Libyears
- OSI Approved Licenses

General Data Science Workflow

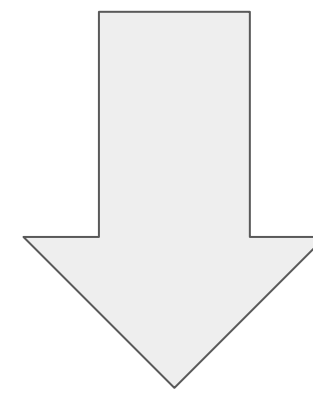


What do you want to learn about your community?



Converting a question to a metric

- ❑ Research on established metrics that could relate to the question
- ❑ Specific data points needed
- ❑ Visualization to represent the data
- ❑ Potential insights and actions



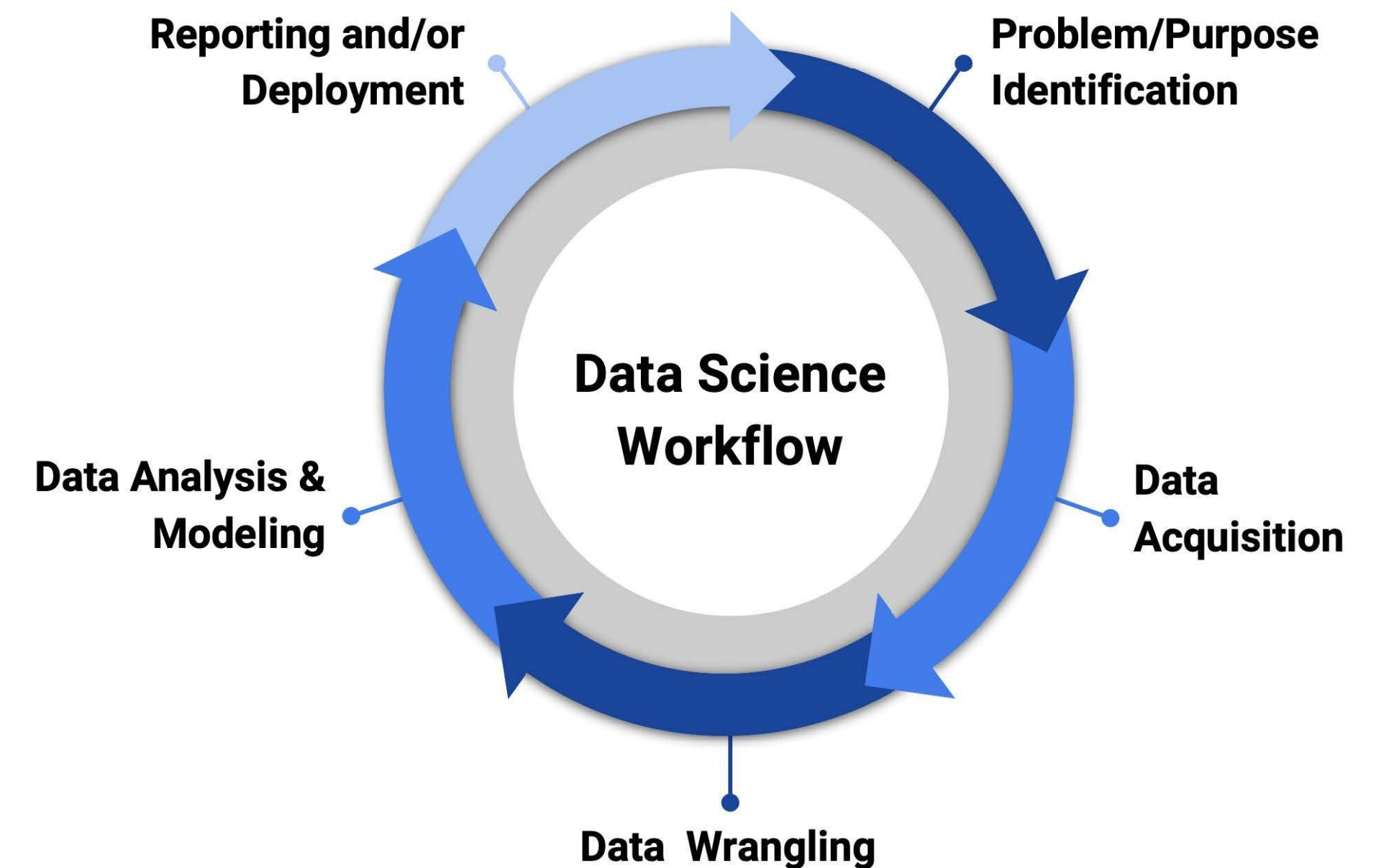
WIP Metric



Community Feedback

Let's talk data

- **Rarely:**
 - are you going to need a single data point or metric about a community
 - End with the same question, metrics, or visualization concept you started with
- **Always:**
 - Space leaves room for a concept to grow

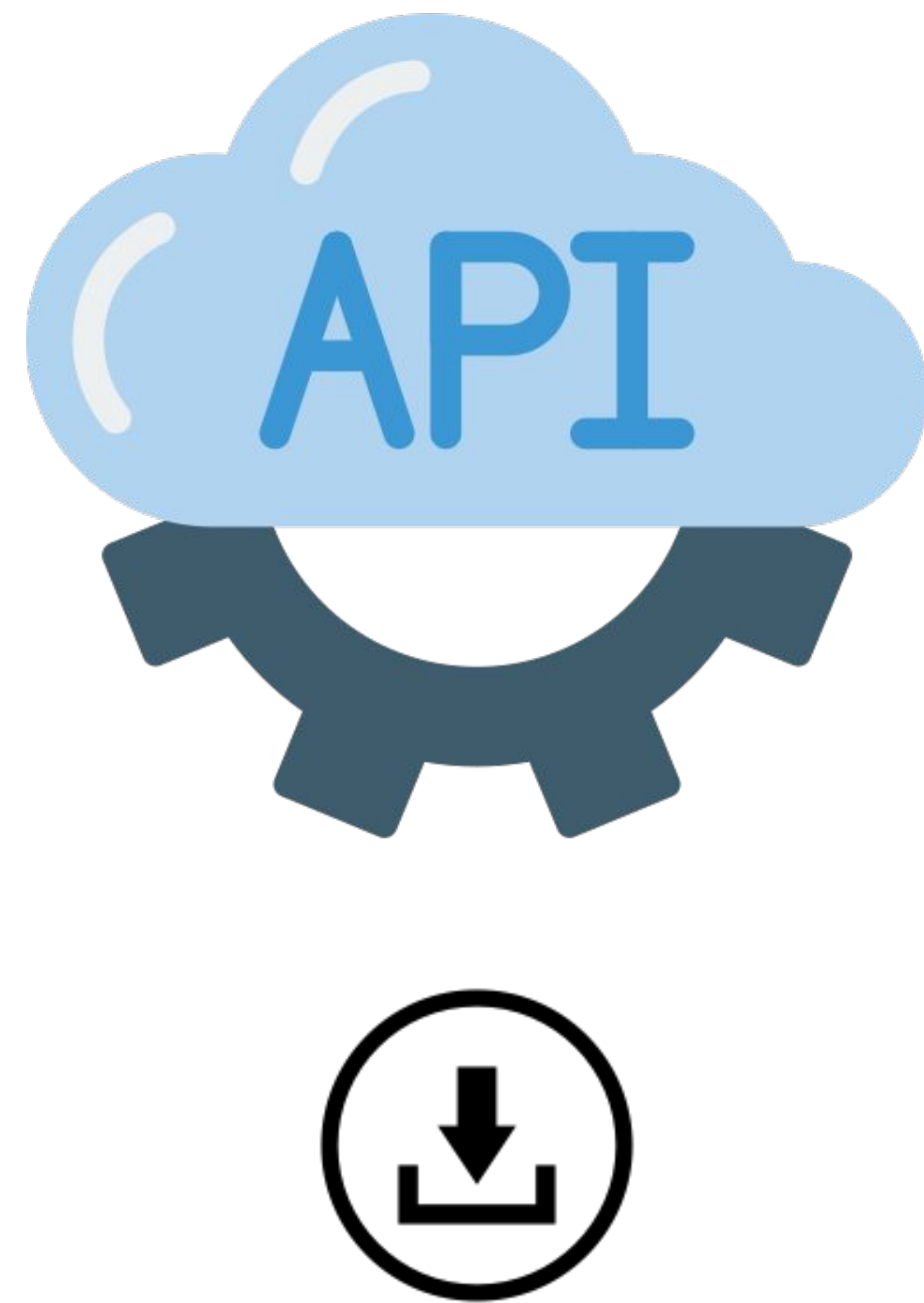


Common OSS Community Data Sources



Community OSS Data Collection

Retrieval Method



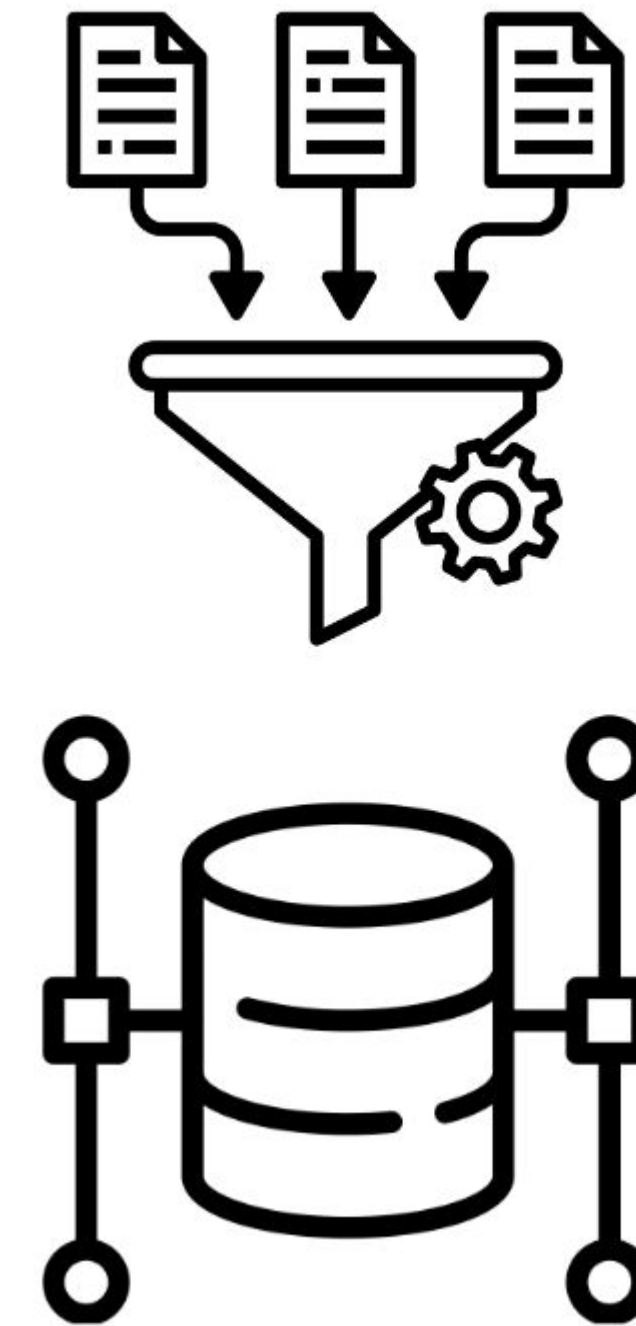
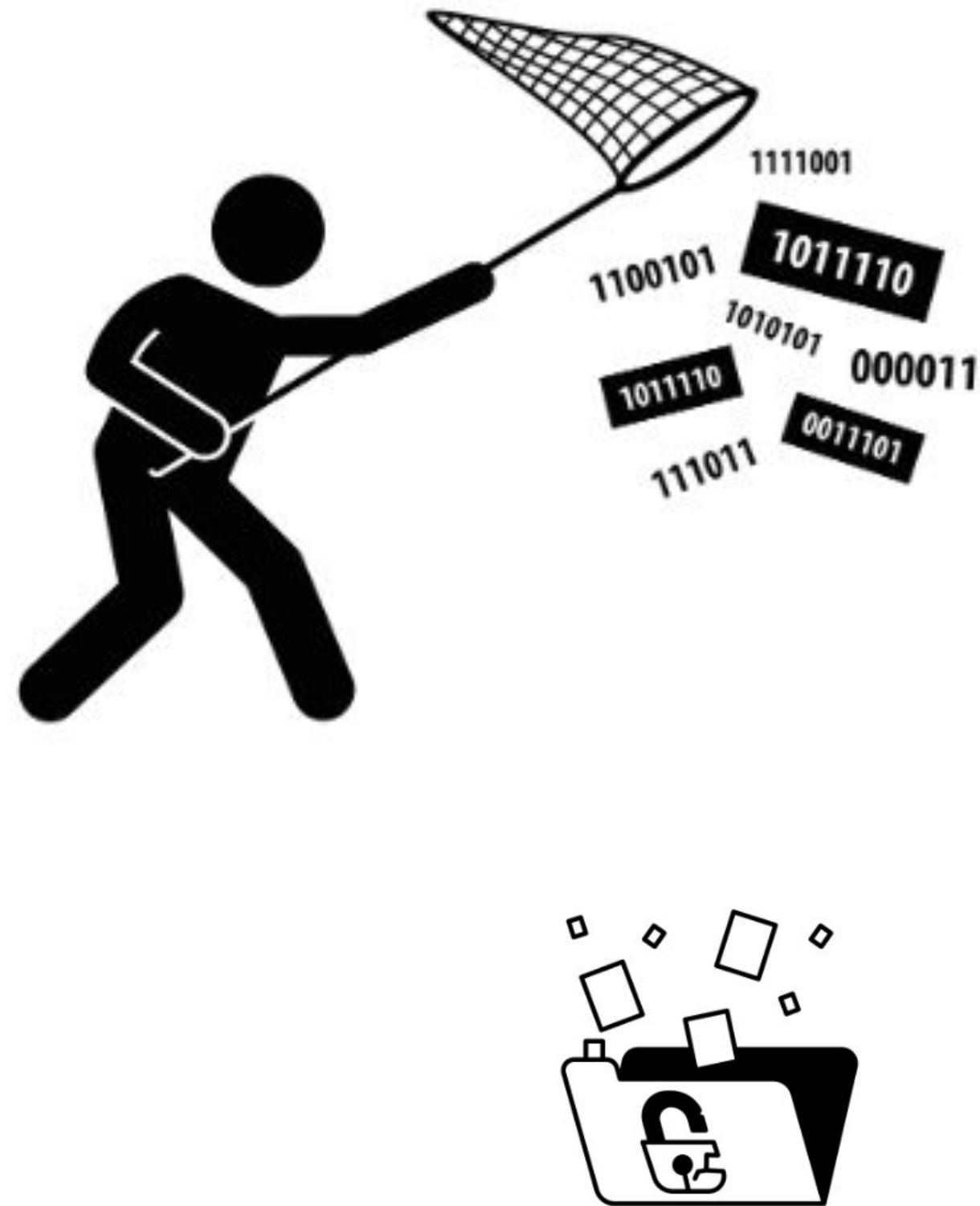
Output Format

{JSON}



In one time use cases or EDA, this format can be sufficient, but that data tsunami grows quick....

Setting up for a general use case - Relational Databases



Standardized data structure - preprocessing out of the way and get to reuse work across different visualizations

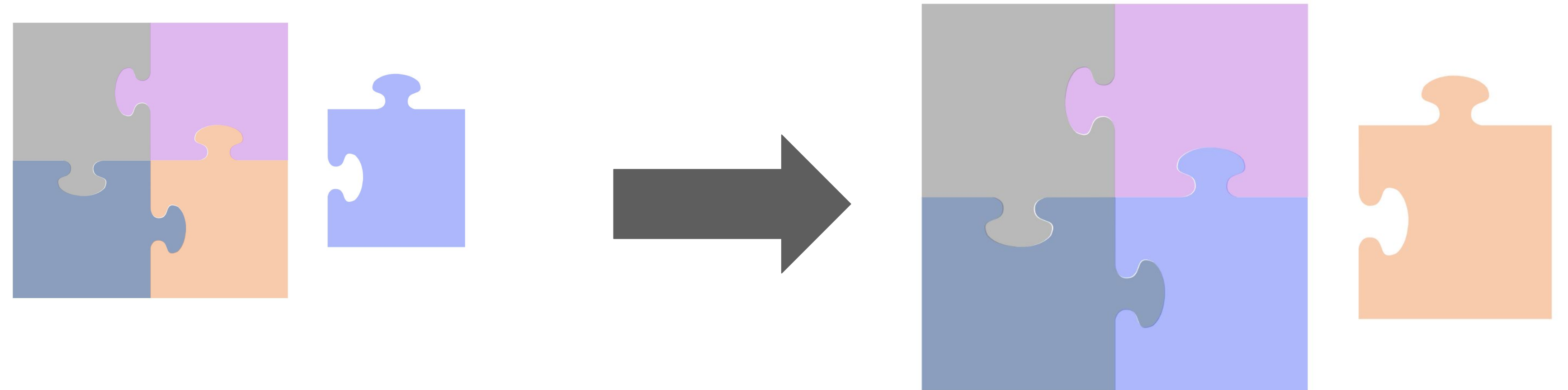
Structural similarity:

Time to first response:

Issues vs PRs

Staleness:

Issues vs PRs



Data Similarity:

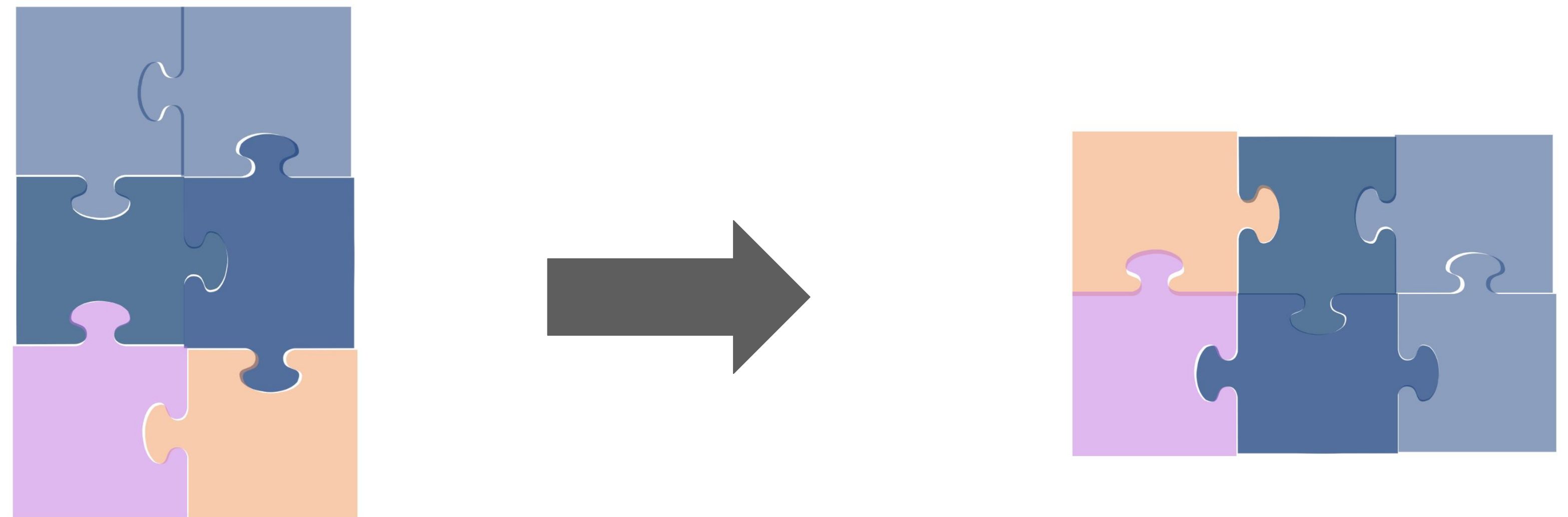
PR Review Assignments per contributor vs status counts

Fly By Contributors by month

vs Repeat Contributors by month

vs Repeat Contributors by month

month



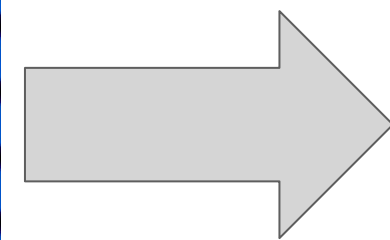


How does this look in action?

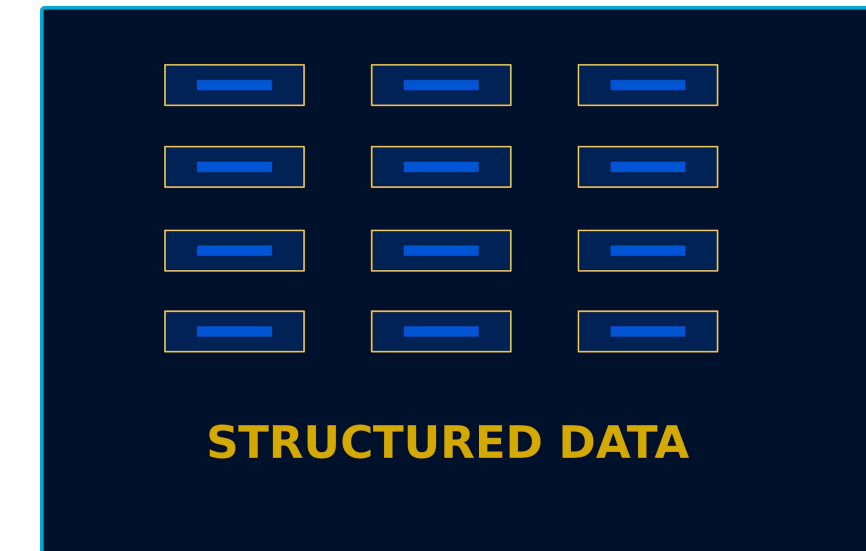
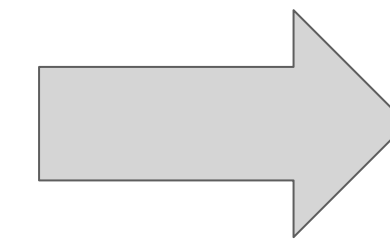
Augur: A path to Data Science through a relational DB



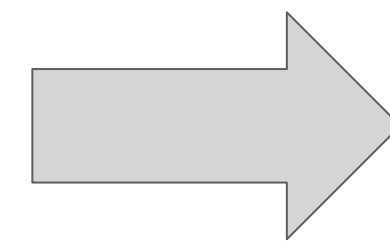
Mountains of Data



6 Years of Data
Carpentry

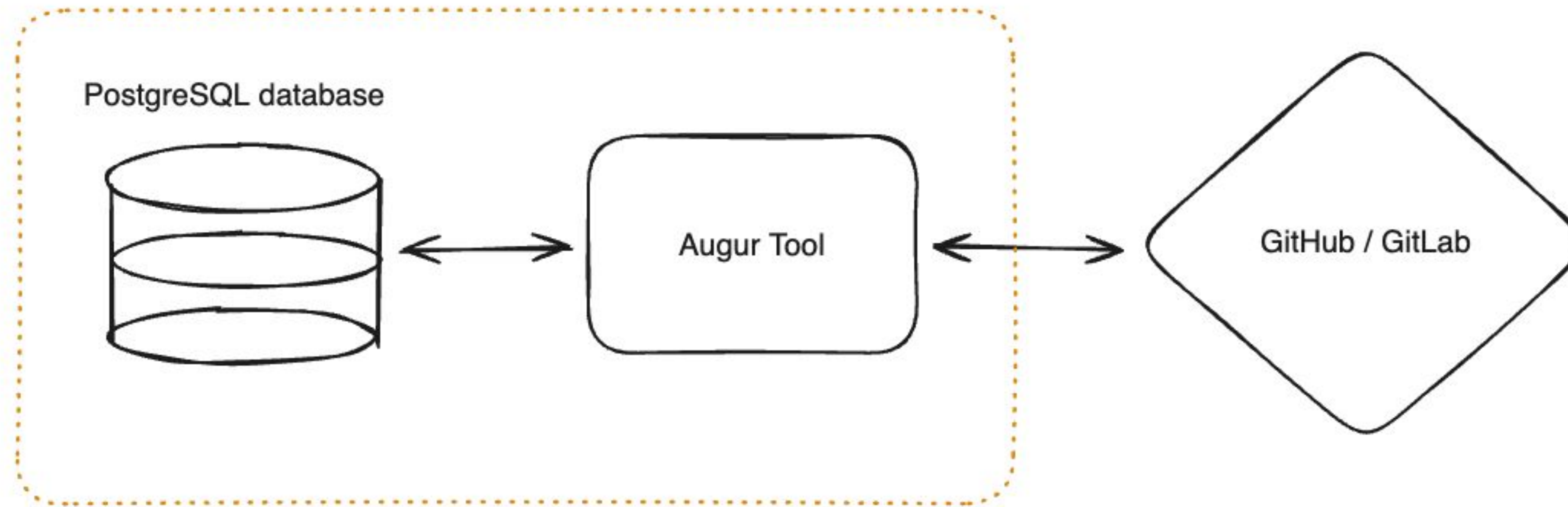


Structured
Data

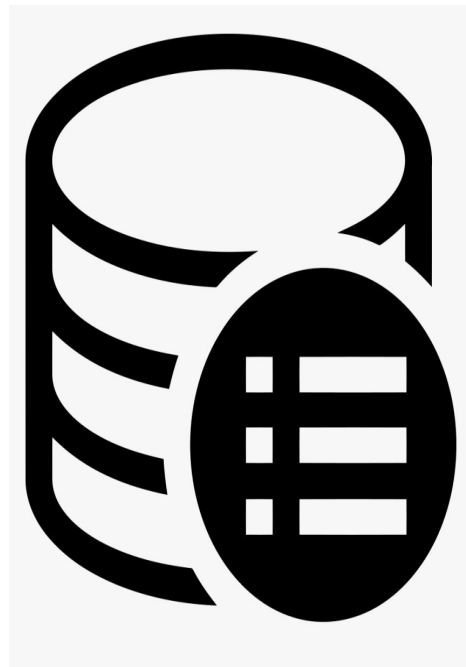


Validated
Data

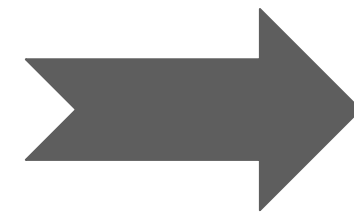
Augur High-level Architecture



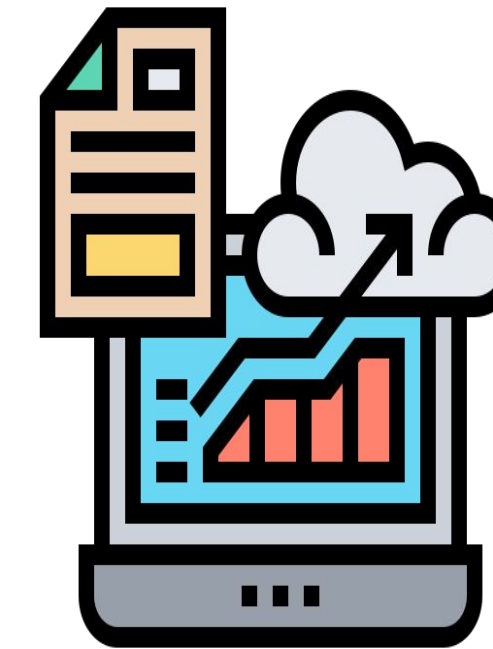
Augur Database



Relational database
with organized Git
Platform data and an
enforced data
validation

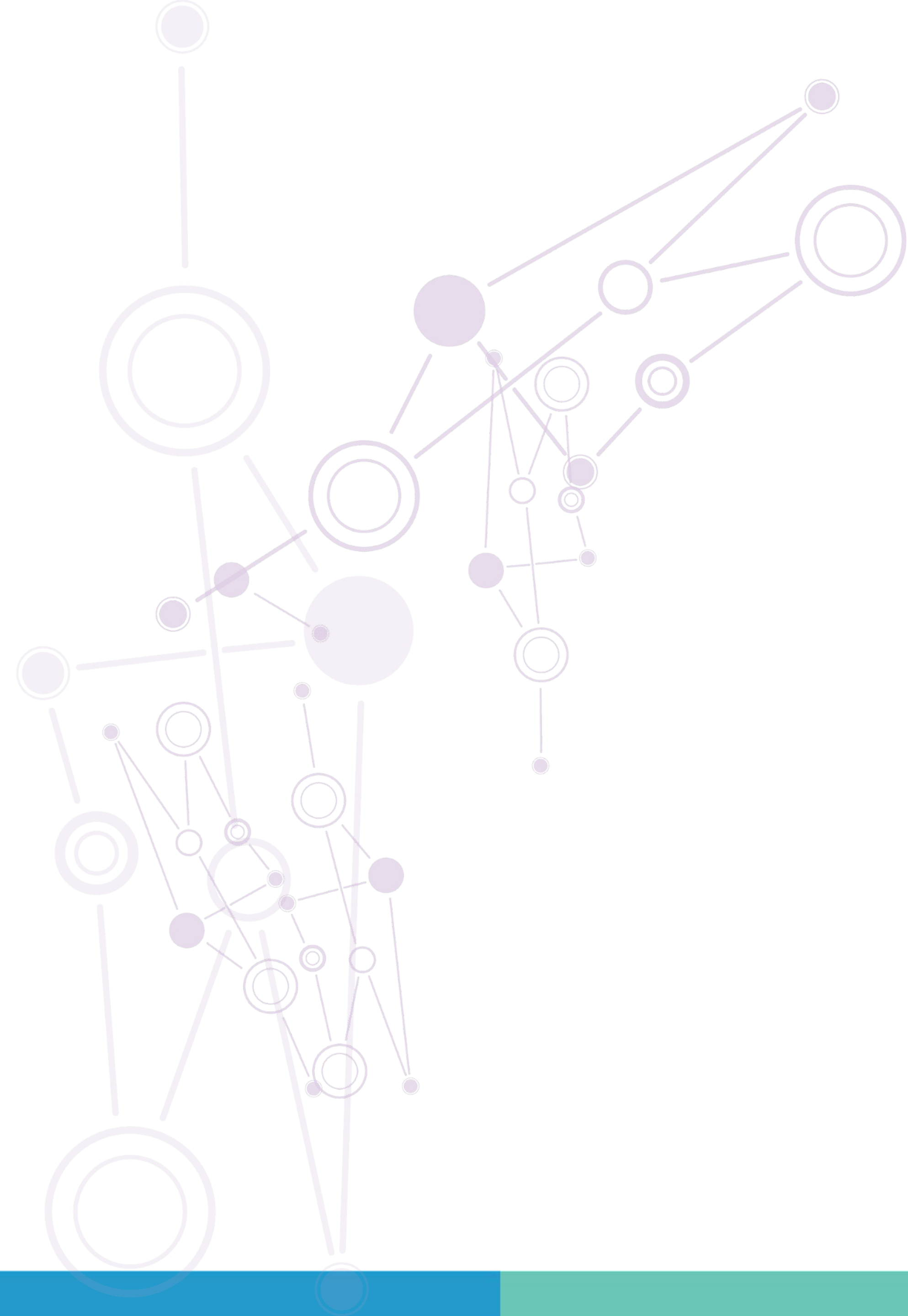



8Knot Dashboard



Data Science Tools:
Dash-Plotly dashboard
with the structure to
visualize any analysis of
the Augur data

8Knot/Augur Demo

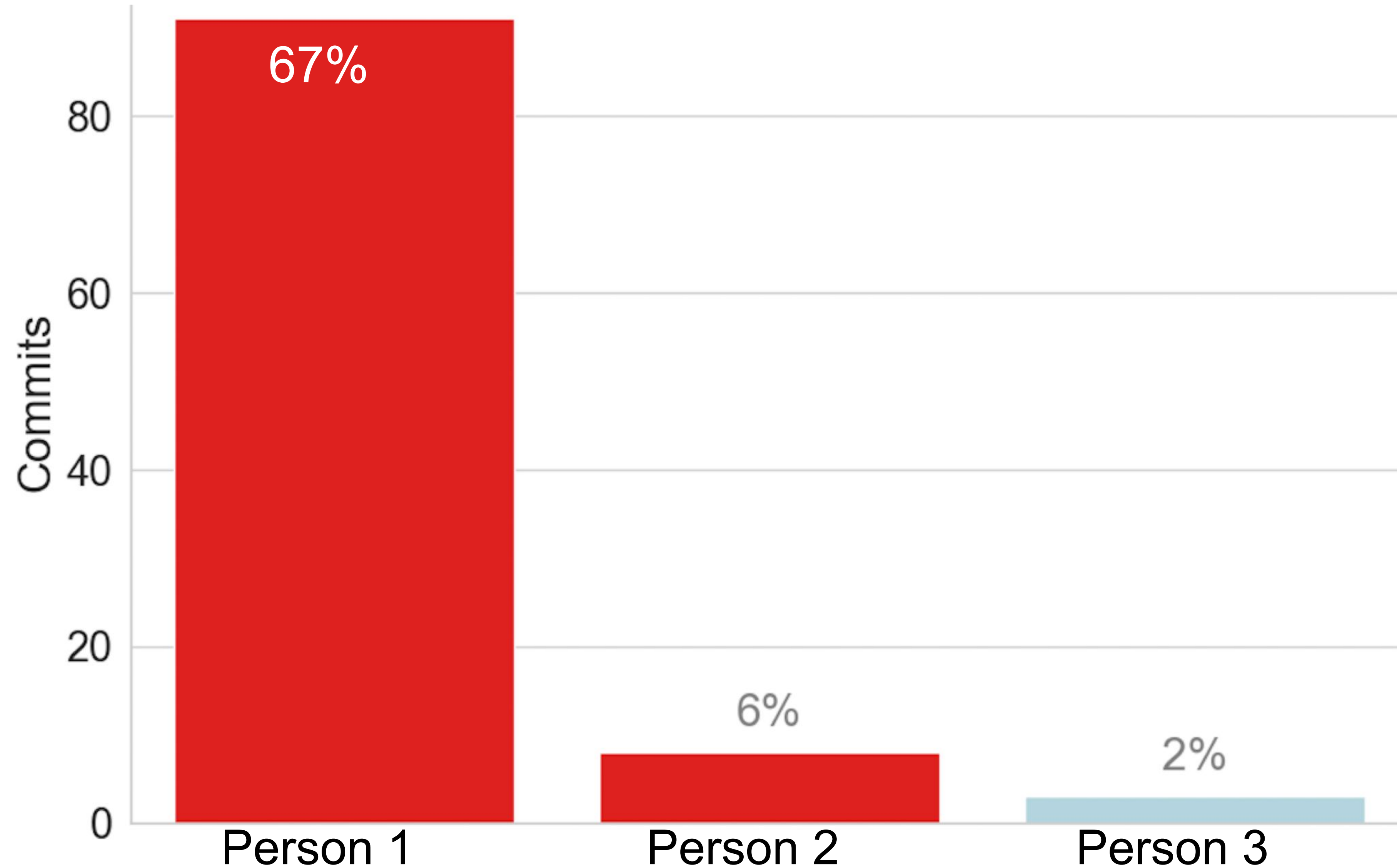




Interpretation: Improving your project and community

Bus Factor for Contributor Sustainability

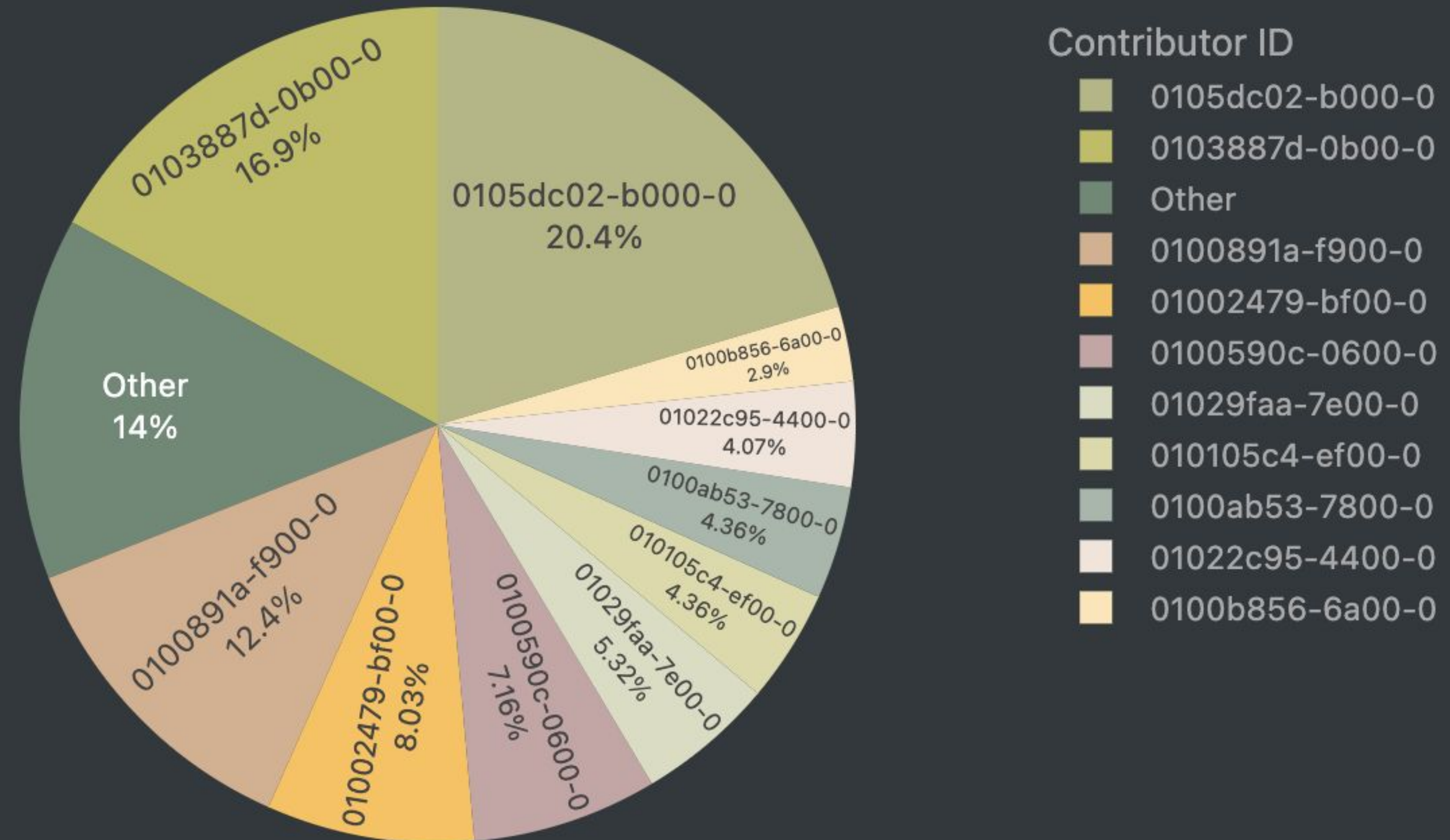
How big of an issue is it?



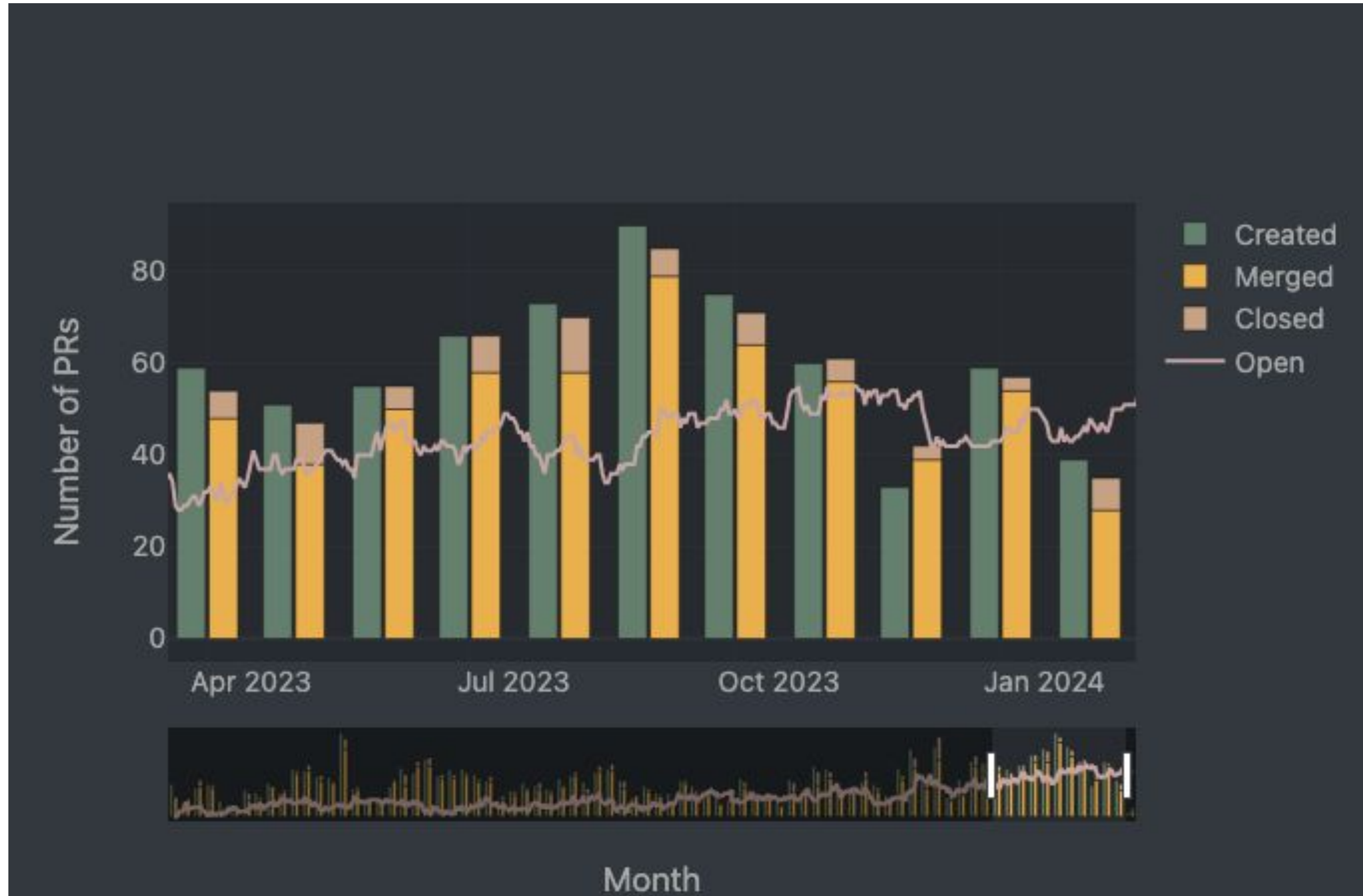
Bus Factor for Contributor Sustainability

Who might be ready to move into a leadership position?

Lottery Factor: Top 10 Contributors by Commit



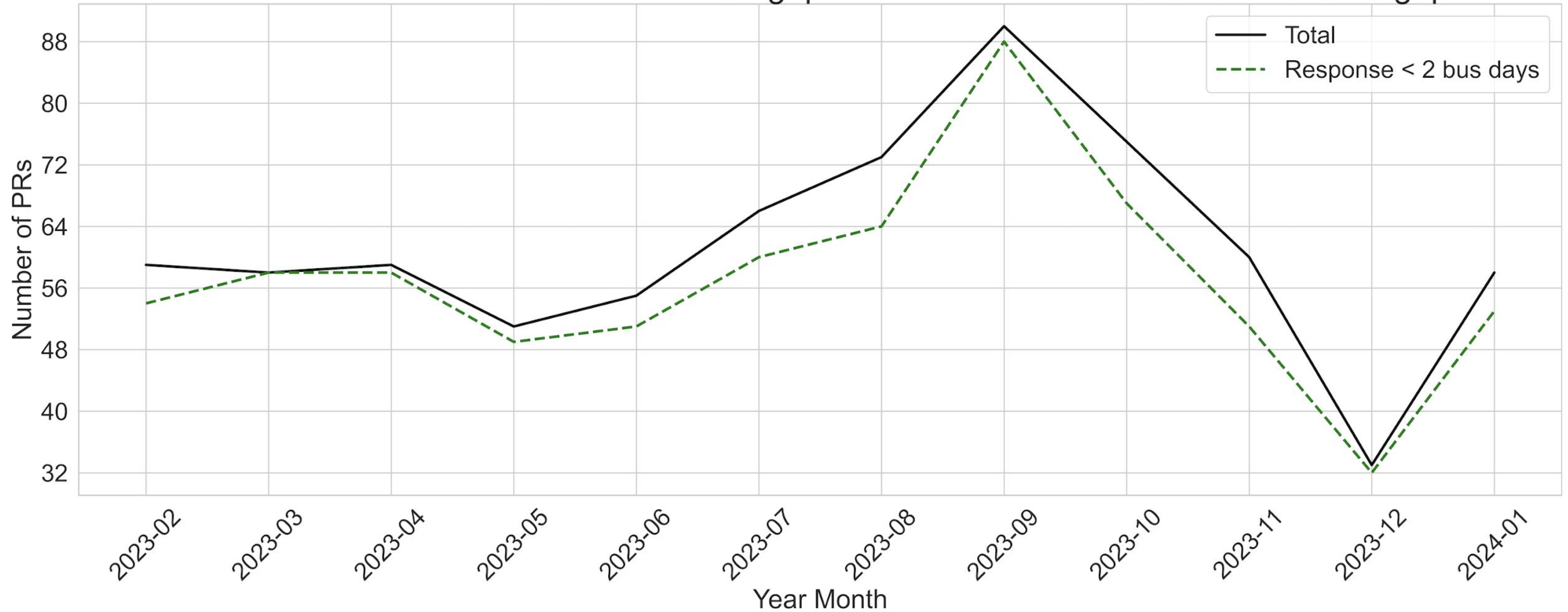
Responsiveness: Closure Ratio



Responsiveness

Time to First Response

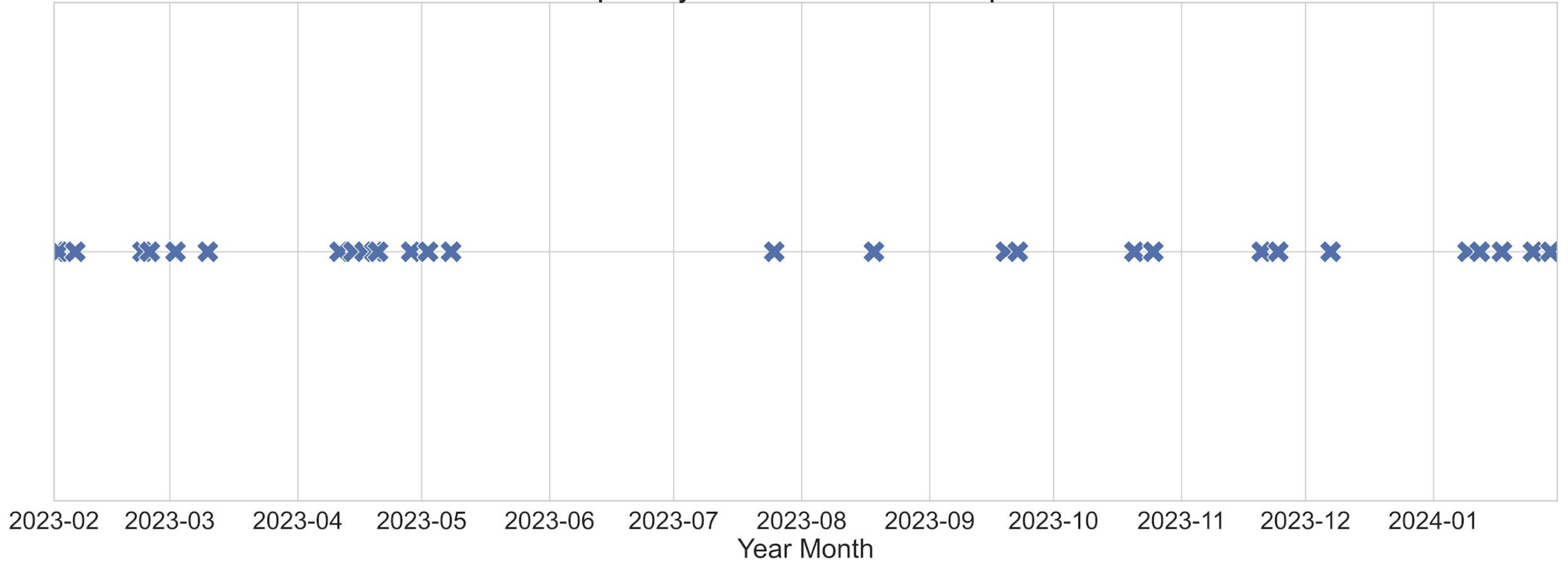
Trend: Positive - the 2023-11 - 2024-01 gap is smaller than the 2023-08 - 2023-10 gap.



Interpretation: Healthy projects will have little or no gap. A large or increasing gap requires attention.

Releases

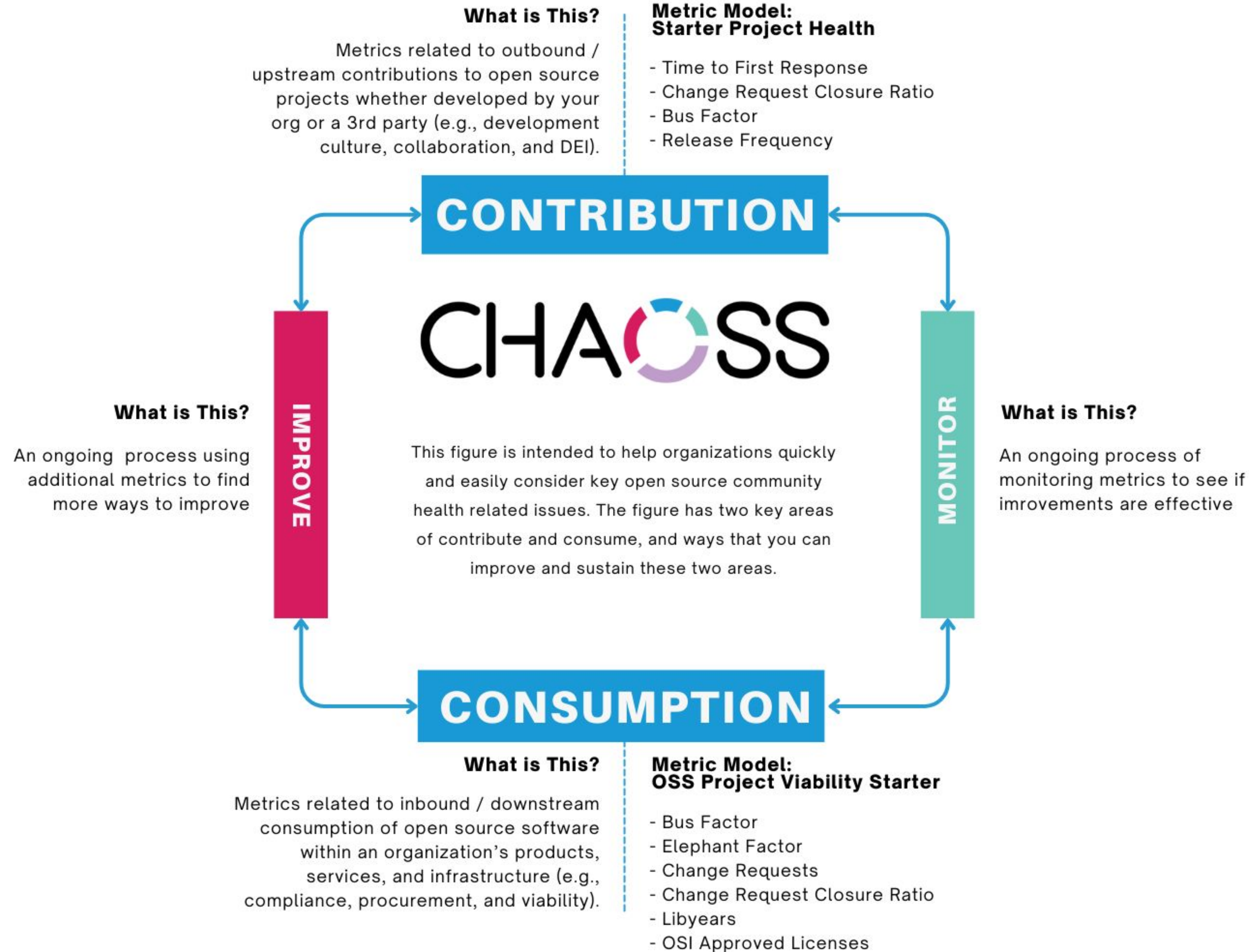
Release Frequency: 13 releases in the past 6 months.



Interpretation: Healthy projects will have frequent releases with security updates, bug fixes, and features.

Ongoing Cycle

Measure, improve,
monitor,
and repeat



Resources

<https://chaoss.community>

<https://metrix.chaoss.io>

<https://github.com/oss-aspen/8Knot>

<https://cacm.acm.org/practice/beyond-the-repository>

Final Thoughts


Use a data science workflow to convert the tsunami of data into actionable insights to improve your project.



Photo by [Isaac Smith](#) on [Unsplash](#)

THANK YOU!

Any Questions?

 <https://chaoss.community/>
 <https://github.com/chaoss>
 @chaoss@fosstodon.org

