Distributing power with open data.

catalyst.coop/work-with-us

Zane Selvans & Christina Gosnell
Catalyst Cooperative
csv,conf,v6 2021/05/05

hello@catalyst.coop
https://catalyst.coop
@CatalystCoop

Outline

- Intro
- **When is data useful in change making?**
- **Making data work for change.**
- **Getting data into the right hands.**
- Takeaways & Questions

Catalyst's Origin Story

- Xcel, Activism in CO.
- **Strategy**: Target expensive & unprofitable coal plants for early retirement.
- **Tactic**: Lots of hand-scraped PDFs of gov't documents.
- **Expansion**: One utility morphed into many.
- **New Tactic**: Tackle the bigger problem of data access.

Do you have a data problem?

More data != better outcomes

Often, it's political.

Ideological barriers (climate, police brutality).

Good (data) story is important but not sufficient.

"The data" can't make decisions for us. We use values.

Open data != better (not always, anyway)

https://www.flickr.com/photos/byzantiumbooks/

When is data useful?

Is this *actually* a data-informed decision making process?

Can data level the playing field? Information asymmetry.

Empower movements & bolster public support.

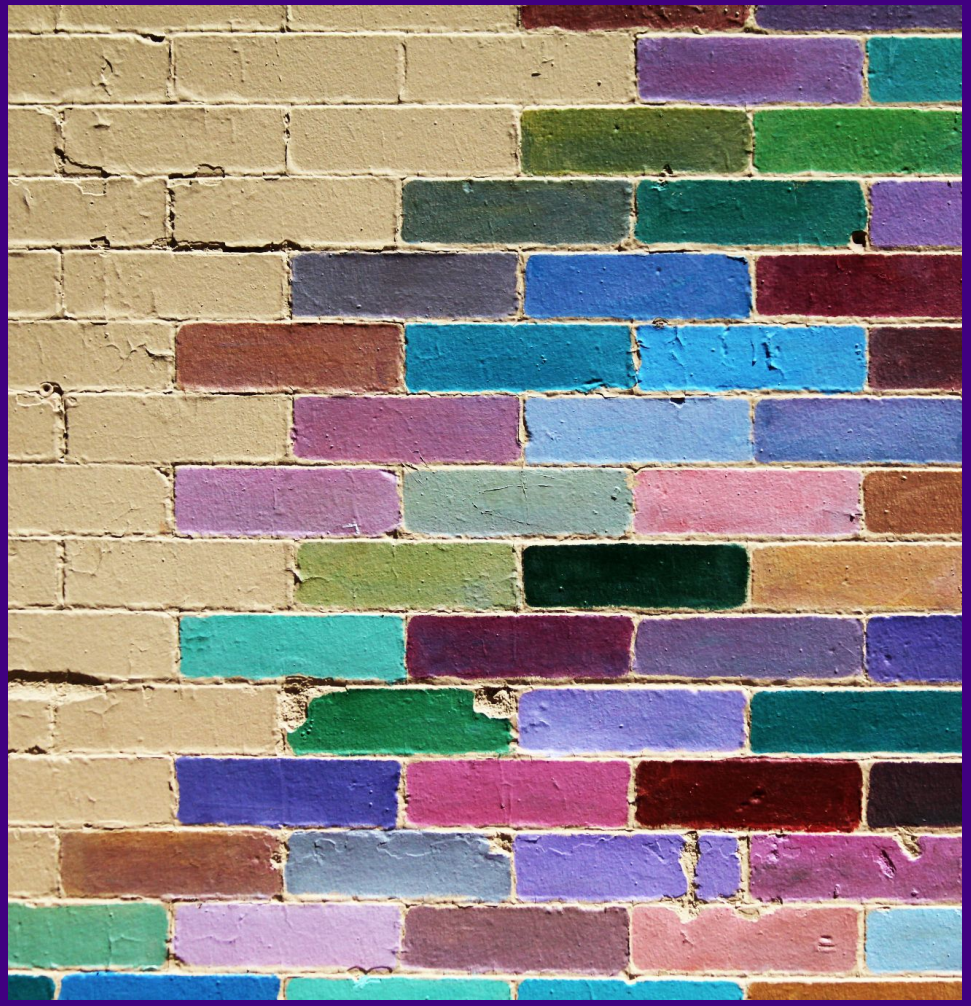
Is there public/available data?

Do others have the same data problem?

When is it worth investing in common infrastructure?

Is there a large enough appetite to support a common project?

Will the reduced duplication of effort outweigh the extra work required to create a common project?



Making Data Work for Change

Free as in Puppy: Toil is a Paywall

- **Toil:** manual, repetitive, automatable, non-scalable work that provides no enduring value.

- Unless it's analysis ready, open data is usually "free-as-in-puppy".

- The need to wrangle is often an avoidable barrier to data access.

- If toil is unavoidable, centralize it. Watch for new tasks that can be brought upstream.

Data: Pristine or Usable?

- Providing “analysis ready” data requires making hard choices on behalf of users.
- Poorly curated data requires interpretation.
- Cleaning, re-shaping, outlier detection: all judgement calls.

- Archive the process as well as the products!

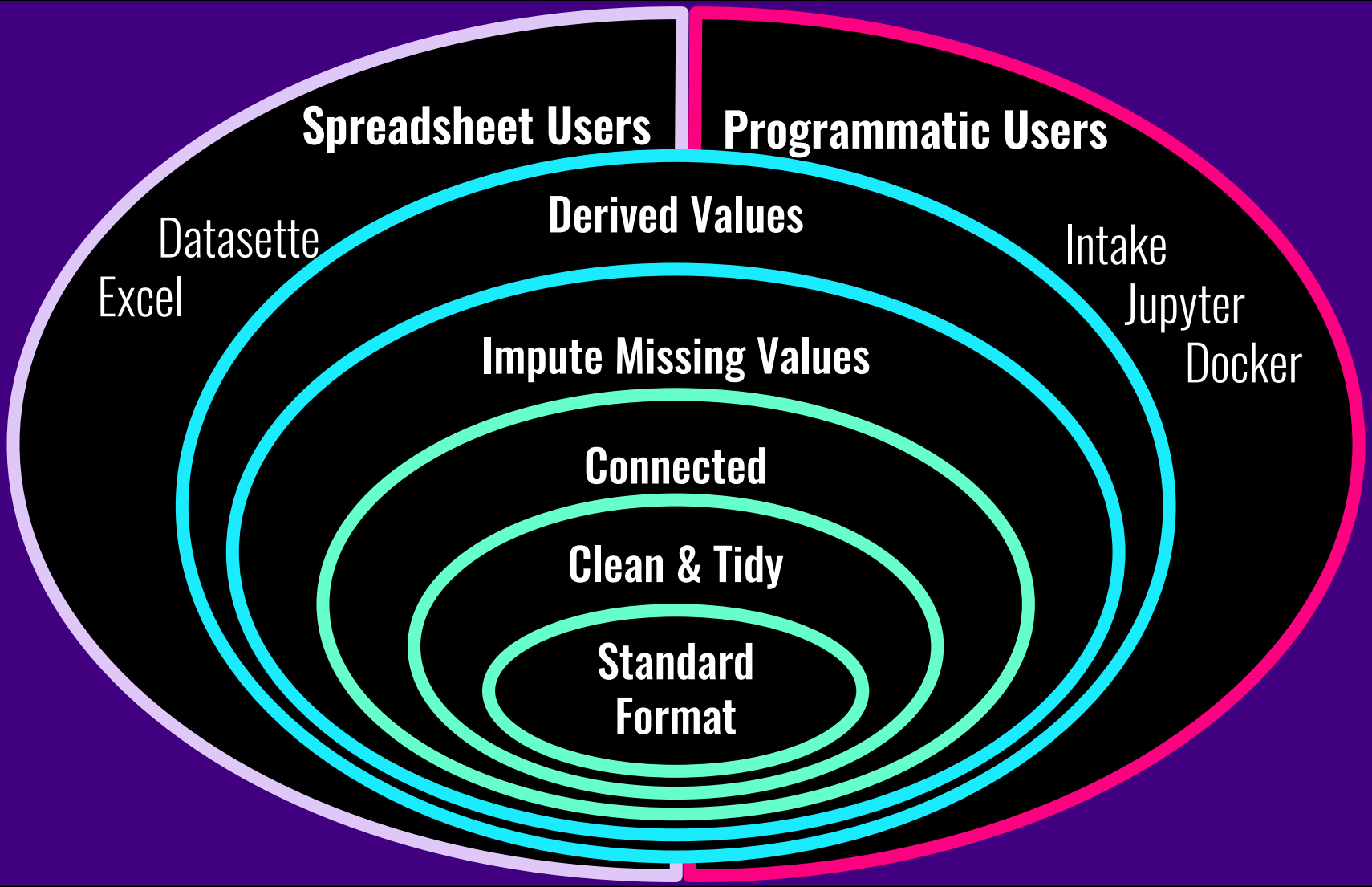
The Means of (Data) Production

- As a small team, treating data as a process, we automate a lot.
- Our tools have diverged from those of our users.

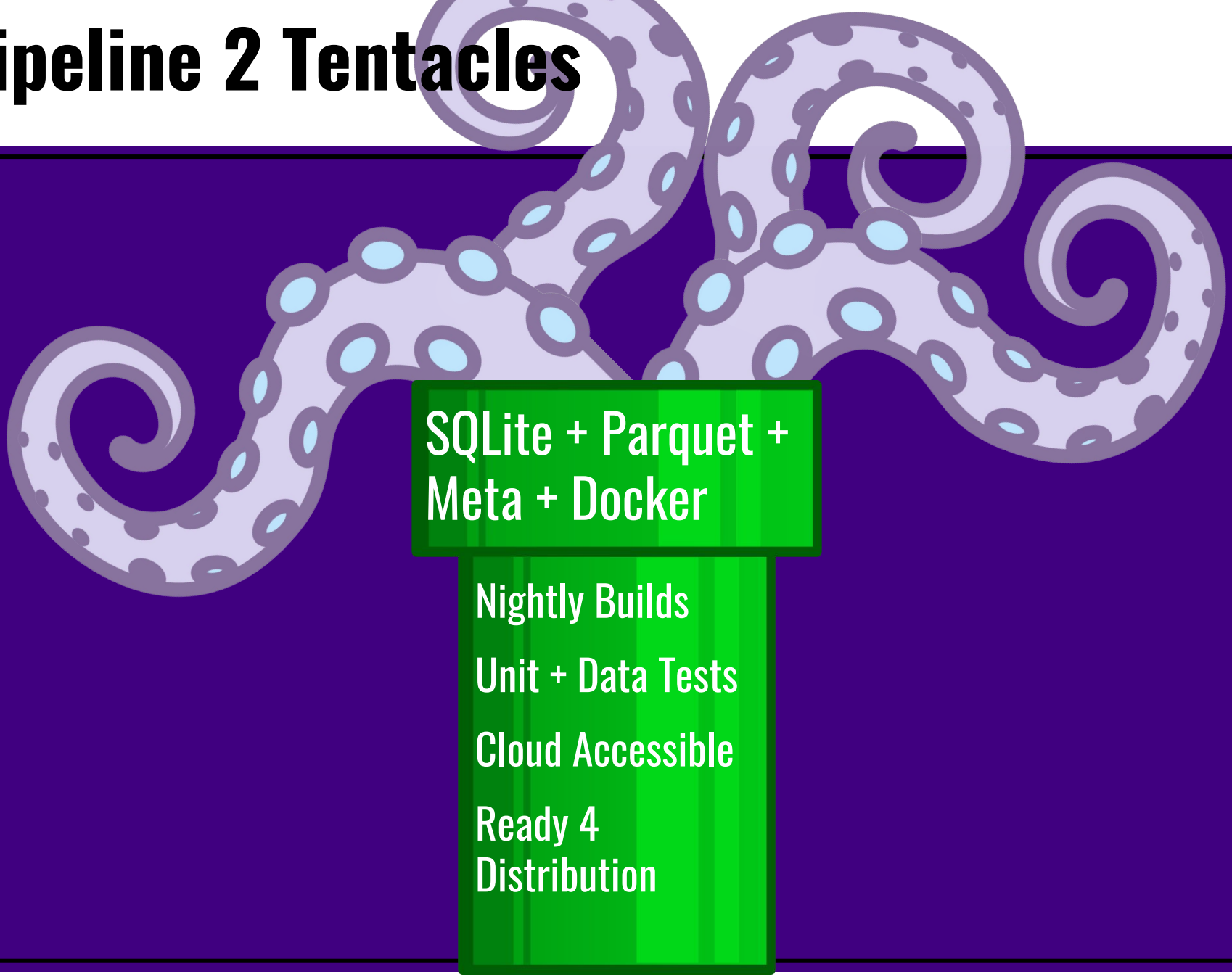
- Have to take care that we bridge this gap. Don't just replace toil with technical barriers.

Getting Data to the People

Circles of Data ~~Hell~~ Access



Pipeline 2 Tentacles



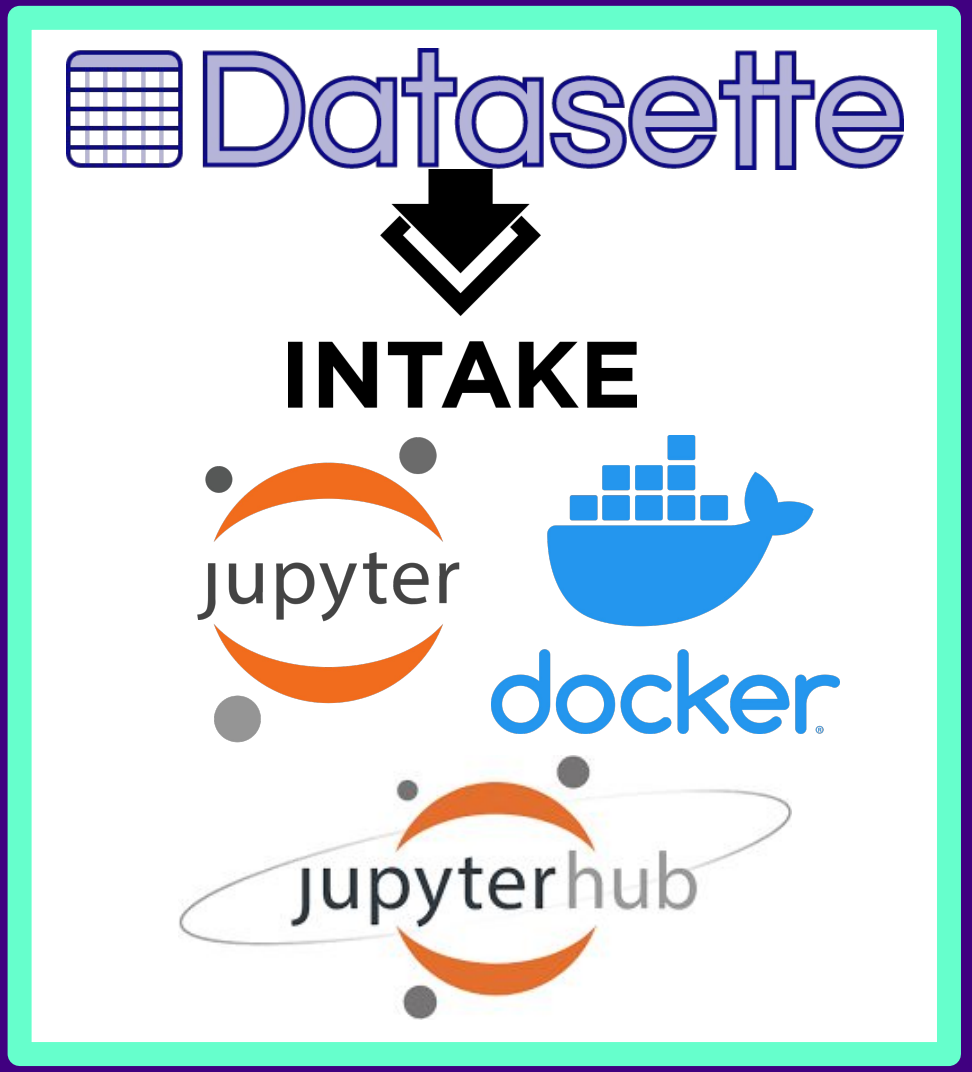
Open Access Tools

Datasette:
For web and spreadsheet users.

Intake Data Catalogs:
Cloud data you can install with conda.

Jupyter Notebooks + Docker:
For savvy Python users and archives.

JupyterHub + 2i2c:
Hosted no-setup notebooks, at a price.



Medium Data Tooling Gap

Big data tools can make it easy to distribute medium data!

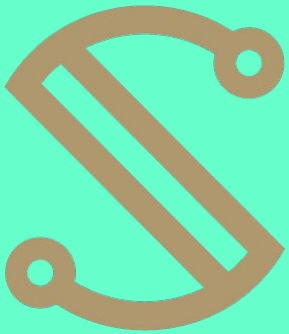
Storage / compute for 10-100GB of data is cheap!

But we need better off-the-shelf options, or the initial setup effort / costs are too big.

Acknowledgements

Open Knowledge Foundation

FRICTIONLESS DATA



ALFRED P. SLOAN FOUNDATION

ENERGY INNOVATION
POLICY & TECHNOLOGY LLC

GridLAB

Takeaways & Questions

- Develop a data-centric theory of change first.
- More Data != Better Outcomes.
- Toil is a Paywall: data isn't liberated until it's easy to use.
- Open isn't **always** better for making change.
- Data is a process as much as a product.
- Tools for data production != tools for data access & use.
- We need better medium-data access tools.

We're hiring!

https://catalyst.coop/work-with-us

References & Resources

Catalyst's Public Utility Data Liberation Project (PUDL):

- GitHub: https://github.com/catalyst-cooperative/pudl
- Example Notebooks: https://github.com/catalyst-cooperative/pudl-examples
- Datasette (browsable database): https://data.catalyst.coop
- Zenodo Archives: https://zenodo.org/communities/catalyst-cooperative/

Books:

- Short Circuiting Policy by Leah Stokes
- Democratizing our Data by Julia Lane
- Data Cleaning by Ihab Ilyas & Xu Chu

Tool: Datasette https://datasette.io

- Overview: https://simonwillison.net/2021/Feb/7/video/
- For Open Data: https://towardsdatascience.com/making-open-data-more-accessible-with-datasette-480a1de5e919

Paper: Open Data for Electricity Modeling (BMWi, 2018)

- https://www.bmwi.de/Redaktion/EN/Publikationen/Studien/open-Data-for-electricity-modeling.html

Paper: Cloud Native Repositories for Big Scientific Data (Abernathey et al. 2021)

- Full Text: https://ieeexplore.ieee.org/document/9354557
- JupyterCon Talk: https://youtu.be/lg7-qi4dEZ8

Talk: Science as Amateur Software Development (Richard McElreath, 2020)

- https://youtu.be/zwRdO9_GGhY