

STATUS QUO

- Data practitioners split into producers and consumers
- Small group of producers
- Consumers unable to contribute, beholden to producers
- Data thrown over the wall, black box

Coronavirus COVID-19 | <https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6>

COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)

Total Confirmed
15,578,624

Confirmed Cases by Country/Region /Sovereignty

- 4,061,925 US
- 2,287,475 Brazil
- 1,288,108 India
- 799,499 Russia
- 408,052 South Africa
- 371,096 Peru
- 370,712 Mexico
- 338,759 Chile
- 299,499 United Kingdom
- 286,523 Iran
- 272,421 Spain
- 270,400 Pakistan
- 262,772 Saudi Arabia
- 245,590 Italy
- 226,373 Colombia
- 224,252 Turkey
- 218,658 Bangladesh
- 217,797 France

Admin0 Admin1 Admin2

Last Updated at (M/D/YYYY)
7/24/2020, 12:35:03 PM

Cumulative Confirmed Cases Active Cases Incidence Rate Case Fatality Ratio Testing

188
countries/regions

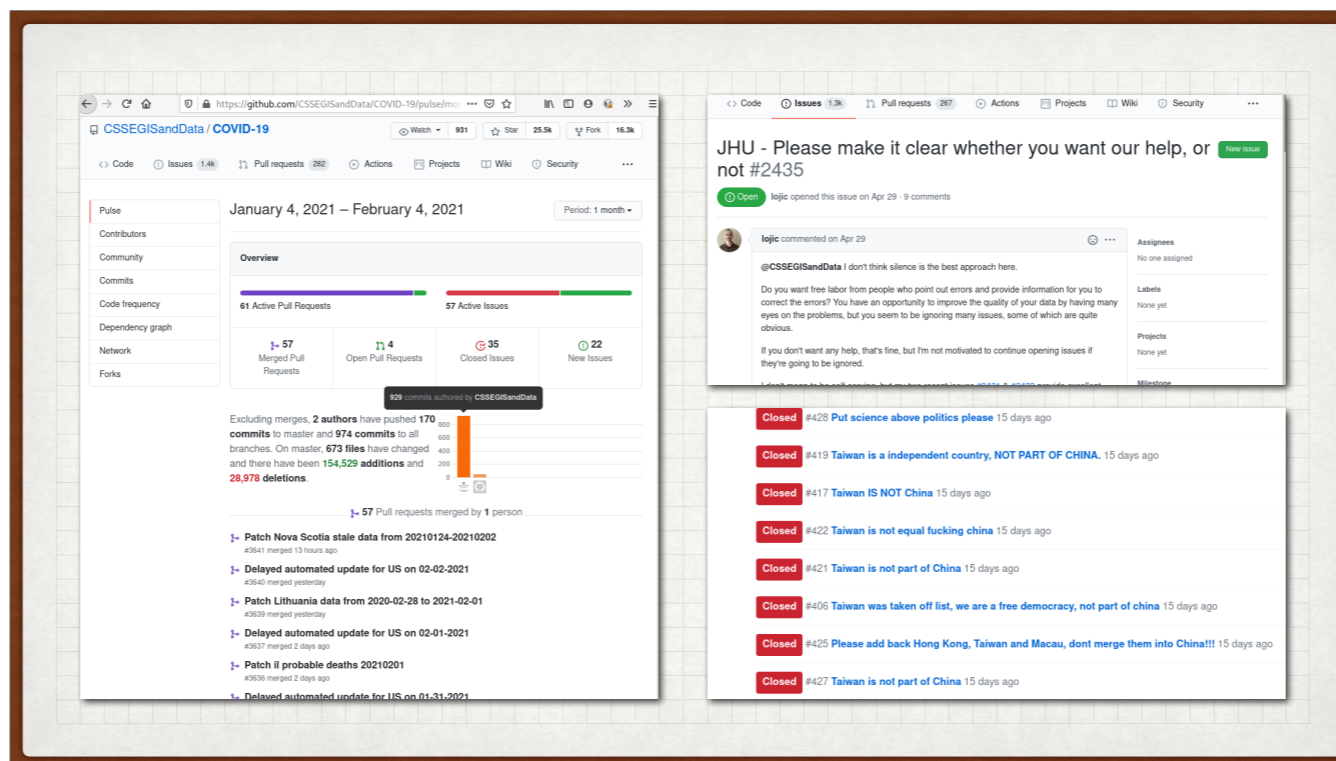
Lancet Inf Dis Article: [Here](#). Mobile Version: [Here](#).
Lead by JHU CSSE. Technical Support: Esri Living Atlas team and JHU and Stavros Niarchos Foundation. Resource support: Slack, Github and other JHU COVID-19 Research Efforts. [FAQ](#). Read more in this [blog](#). [Contact Us](#)

Global Deaths
634,744

- 144,552 deaths US
- 84,082 deaths Brazil
- 45,762 deaths United Kingdom
- 41,908 deaths Mexico
- 35,097 deaths Italy
- 30,601 deaths India
- 30,195 deaths

US State Level Deaths, Recovered

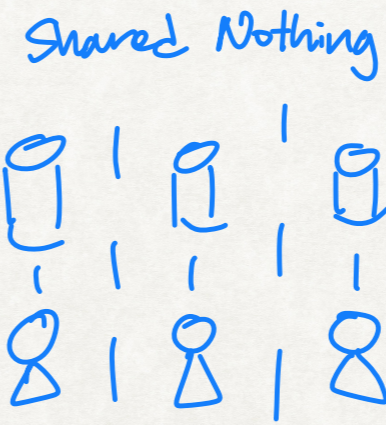
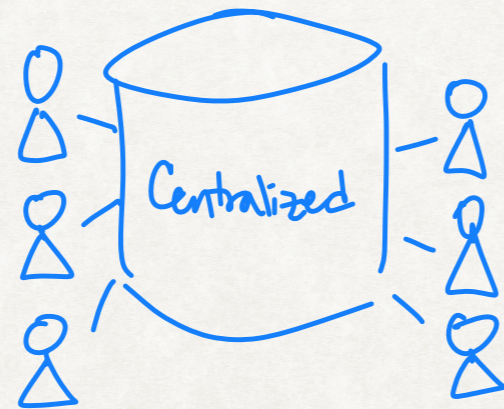
- 32,594 deaths, **72,466** recovered New York US
- 15,730 deaths, **31,925** recovered New Jersey US
- 8,484 deaths, **96,452** recovered Massachusetts US
- 8,201 deaths, **recovered** California US
- 7,560 deaths, **recovered** Illinois US
- 7,113 deaths, **78,268** recovered Pennsylvania US
- 6,395 deaths, **55,162** recovered



25k+ stars. 1.4k open issues. One committer

“Consumers beholden to producers”: JHU gets separate data updates from Taiwan and mainland China. A few months into the pandemic they started aggregating into a single value. Lots of upset comments.

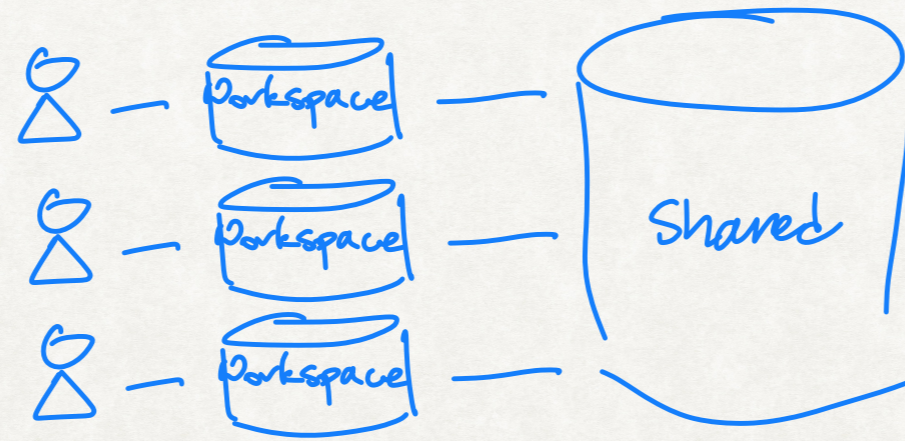
DATA ARCHITECTURES



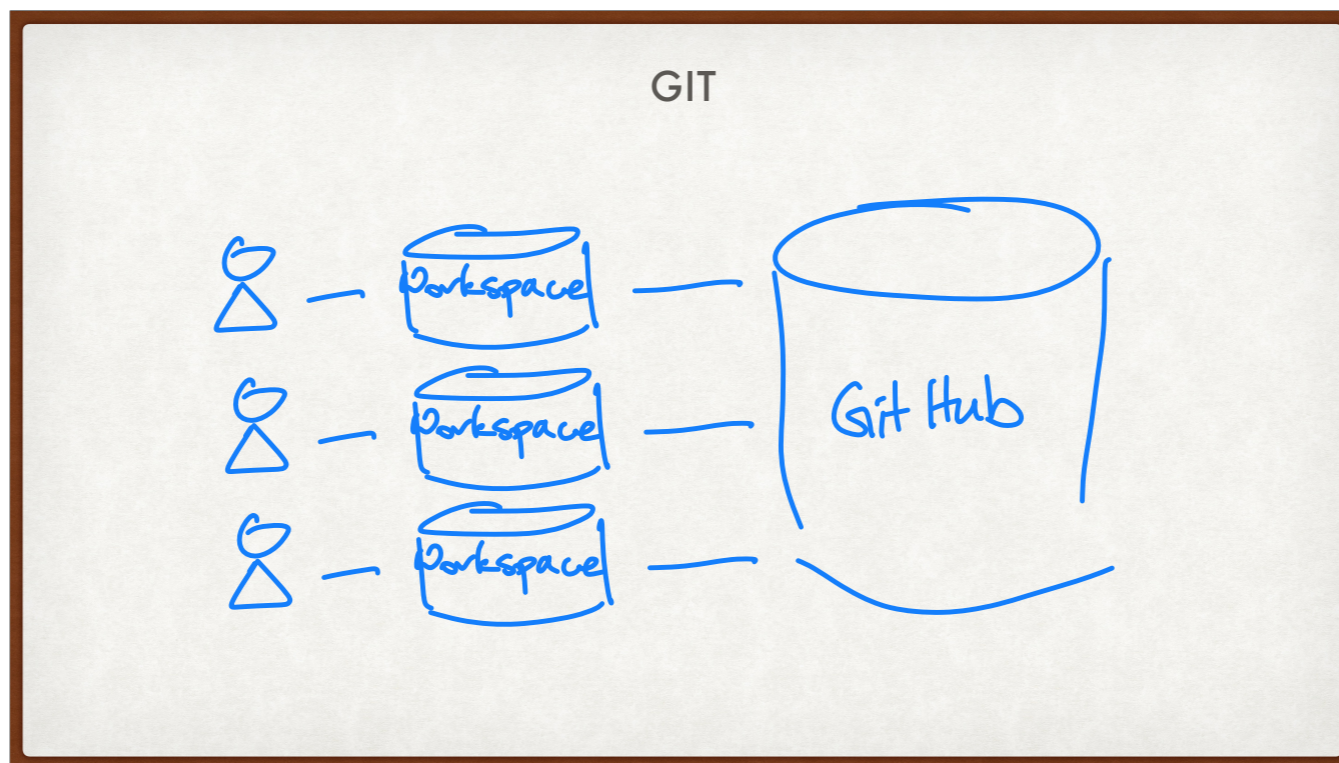
Despite being on GitHub, JHU Covid is example of centralized data architecture. Data warehouse is a common centralized architecture. Risk averse. Hard to coordinate changes. Ask for permission bureaucracy. Waterfall development.

Shared nothing: each practitioner has their own environment, no sharing with other users. Workable in some professional service-type organizations. Autonomy to make changes anywhere. Can't break other users (or be broken). Collaboration is expensive.

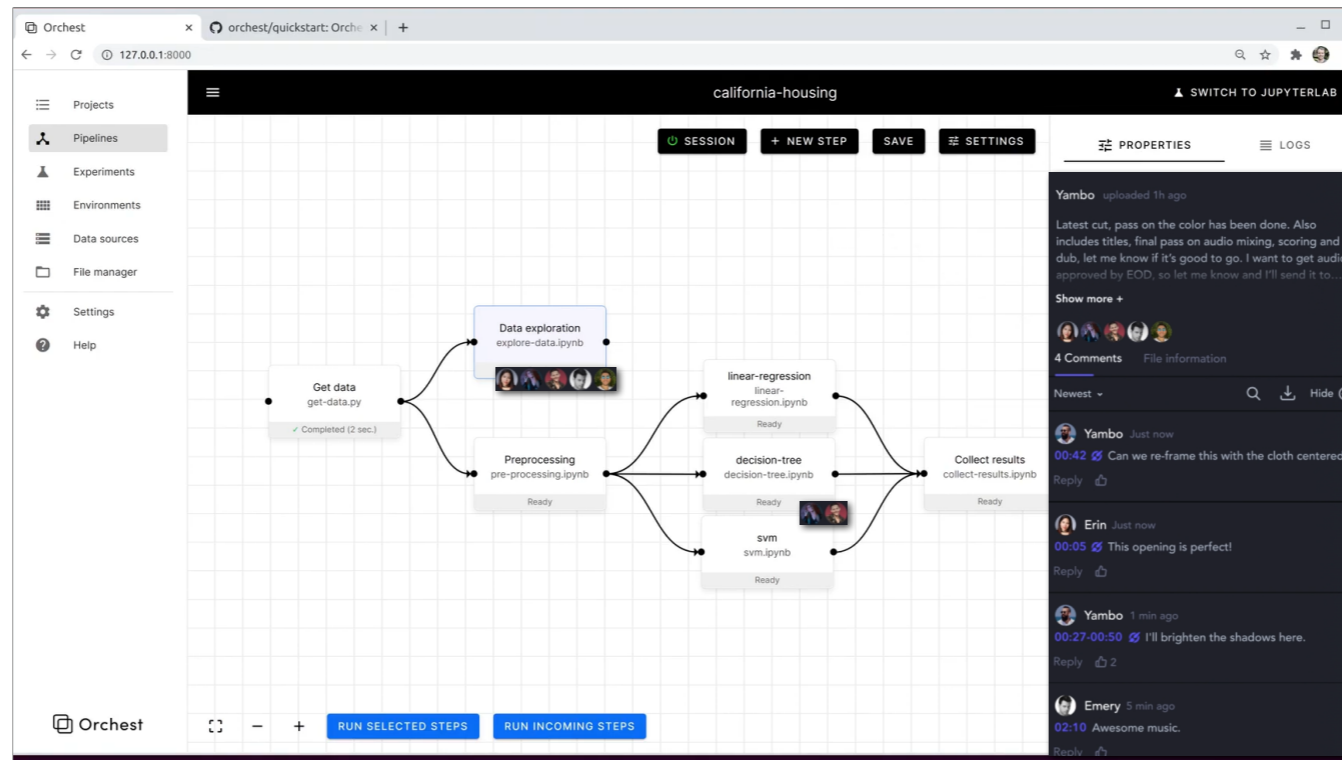
DISTRIBUTED



Best of both worlds? Each practitioner gets their own workspace, but can publish and share with others.



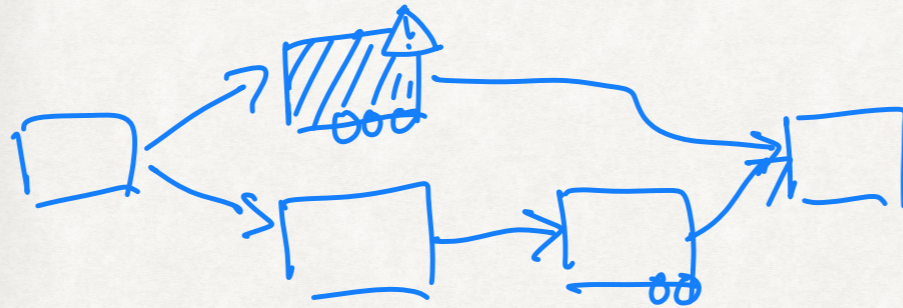
Exactly the model of Git!



Enough architecture hand waving, what could this look like?

Not my product. Mock made by mashing up two products: a non-collaborative data pipeline tool, and a collaborative video editing tool. What if stakeholder interactions were front-and-center while working on our data pipelines.

Complicated data flow



DISCUSSION

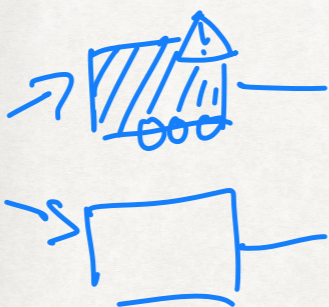
○ Help me!

○ m

○ m

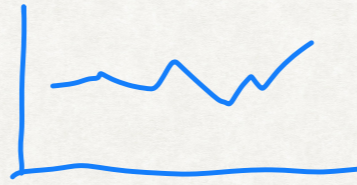
Same slide, sketched

duplicate



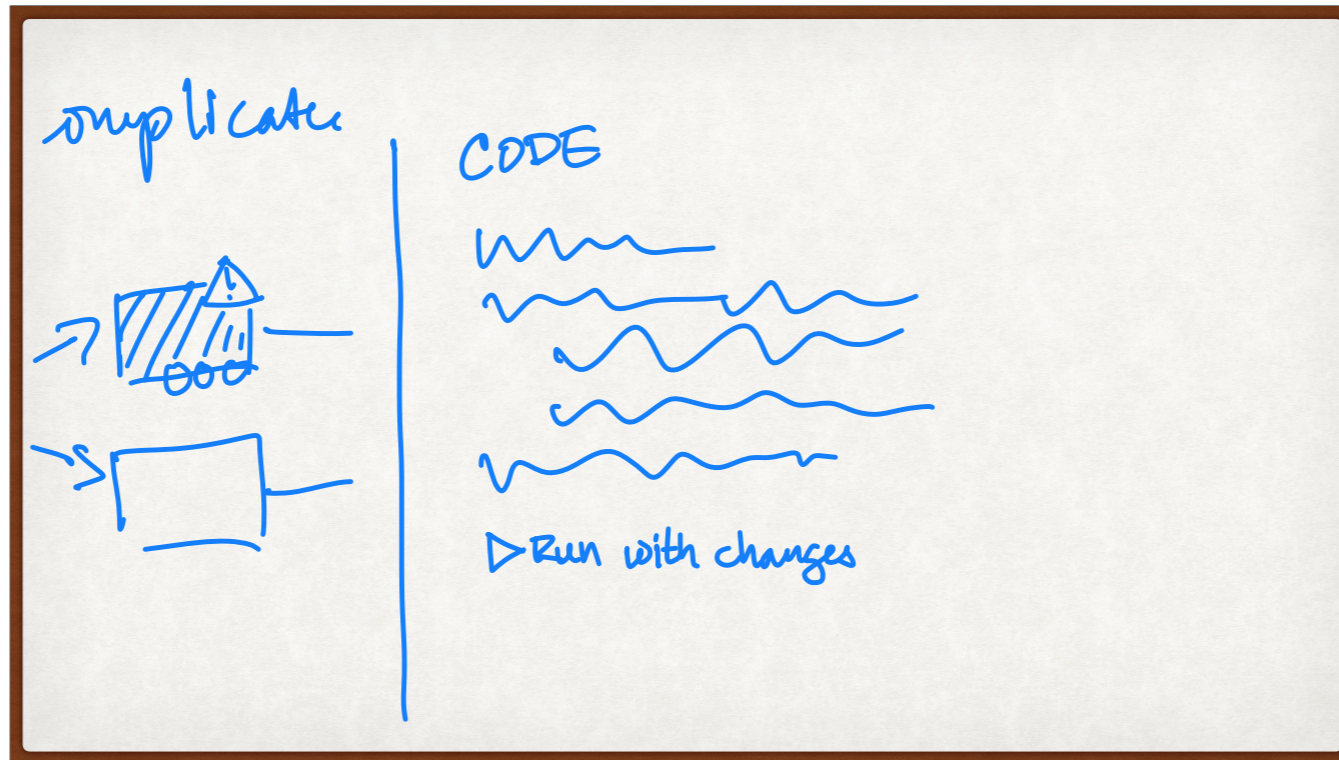
DATA REVIEW

- 2 data alerts
- ⚠ check duplicate names
- ⚠ make respondents > 70%

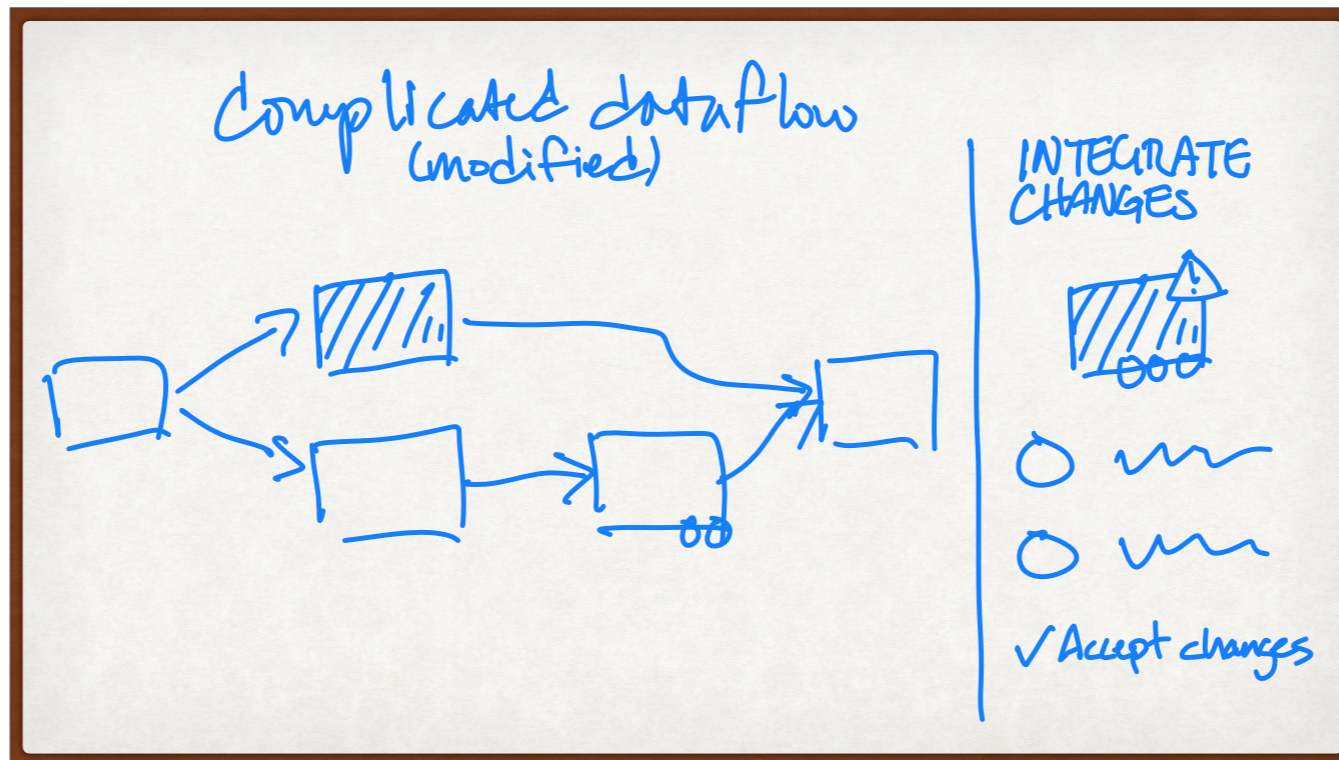


○ Can we check our input data sources?

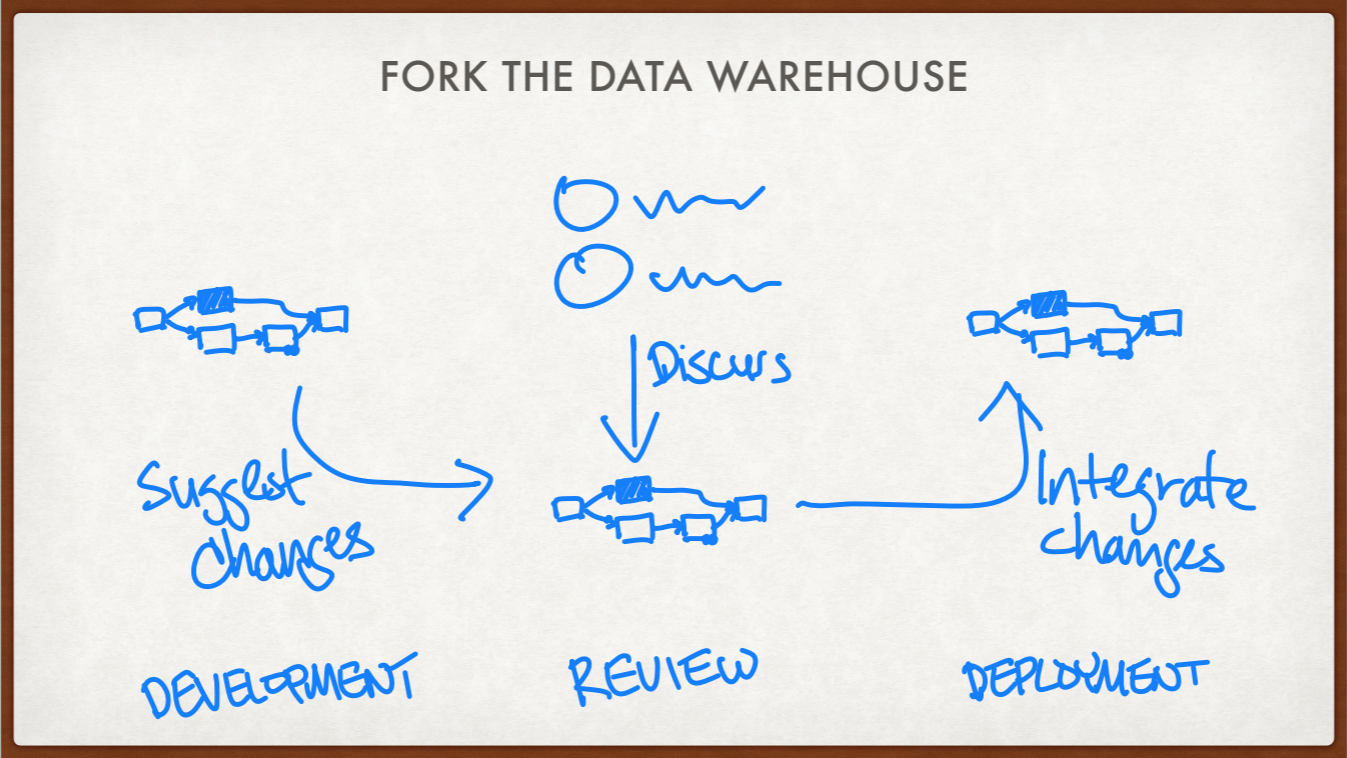
Ask for others to look at your changes. Like GitHub pull request. Surface automated test results, anomaly detection, trends, discussion



Discussion -> suggestions. Data needs to be explored. What if reviewers could try their own ideas out?



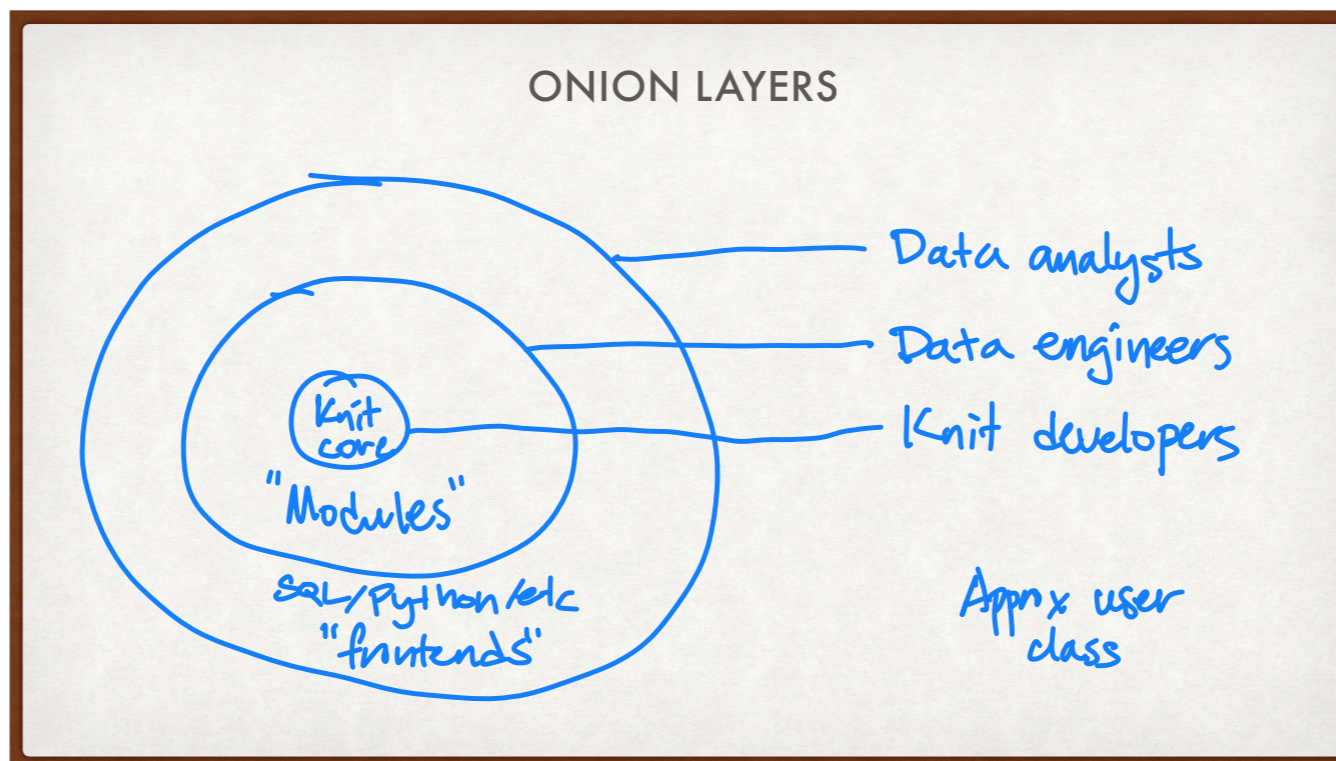
Integrate changes AFTER everyone has reviewed and seen the results



Conceptually like having unlimited copies of the data warehouse

DATA VS METADATA

- Metadata is data about data: where, who, when, from what, etc
- Metadata is fully managed and immutable
- Data can be mutable
- Data is external



Knit's functionality built up in layers.

Very compact domain-agnostic core on inside. Domain specialists (data analysts) on outside. Middle layer of "modules" that bridge between them

MODULES

PROVIDE EXTENSIBILITY

- Backing stores: S3, Snowflake, Postgres
- Data transformations: SQL, pandas, scikit-learn
- Frontends: pipeline code, GUI, notebook
- Runtime dependencies: virtualenv, Docker

Kind of like PyPI or NPM