data.world

# 10 Pragmatic Lessons for Building Data Collaboration

Patrick McGarry — data.world

data.world

# ./whoami

- Slashdot/Sourceforge
- Alcatel-Lucent
- Perforce
- Inktank (Ceph)
- Red Hat
- data.world

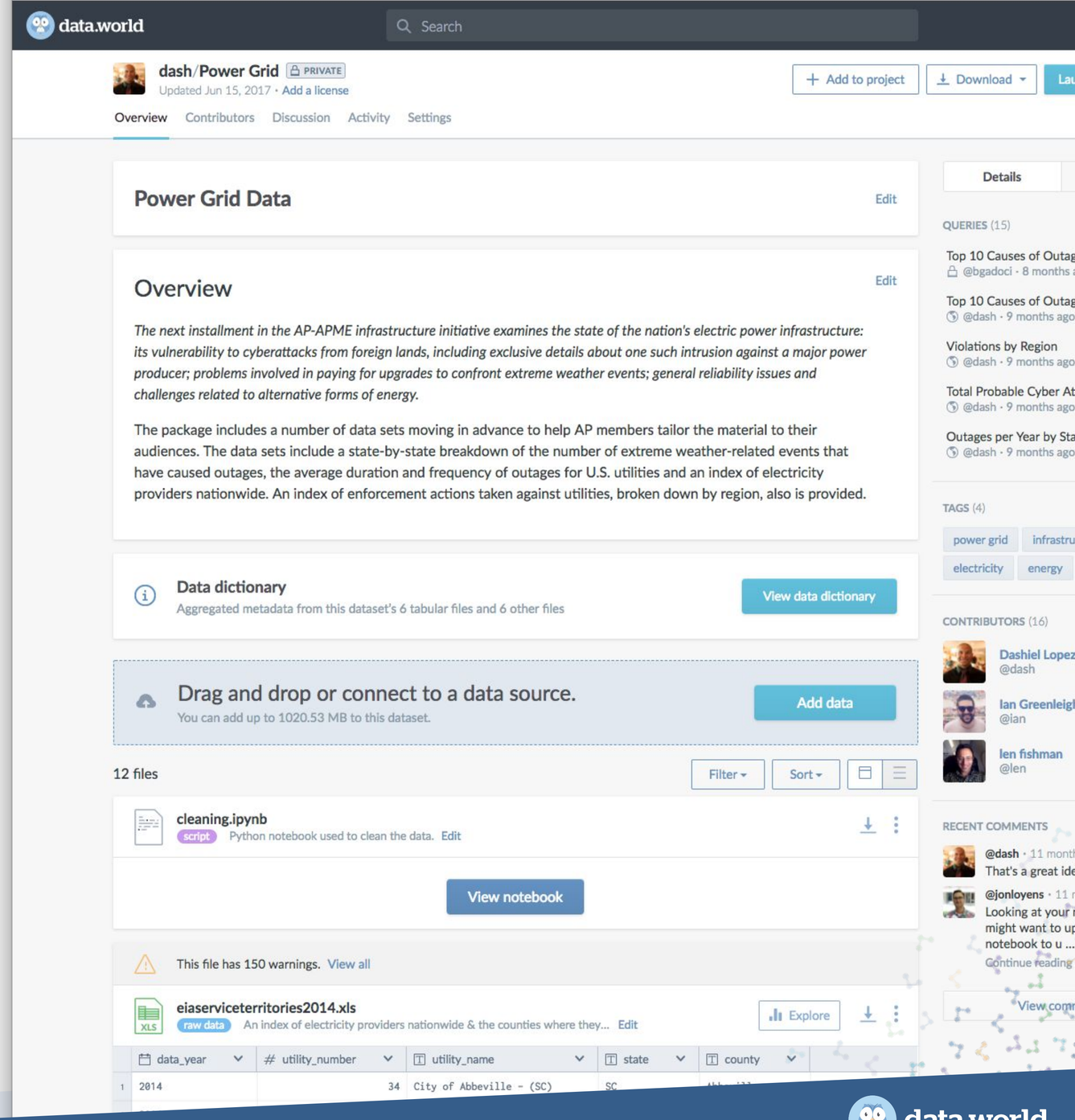
# data.world

We're building the most **meaningful**, **collaborative**, and **abundant** data resource in the world by dismantling the barriers between data and people.

Our platform helps data people solve problems faster by creating new ways to discover, prep, and collaborate.

data.world

Certified B Corporation



# Data Practices Manifesto / CPEDS

https://datapractices.org

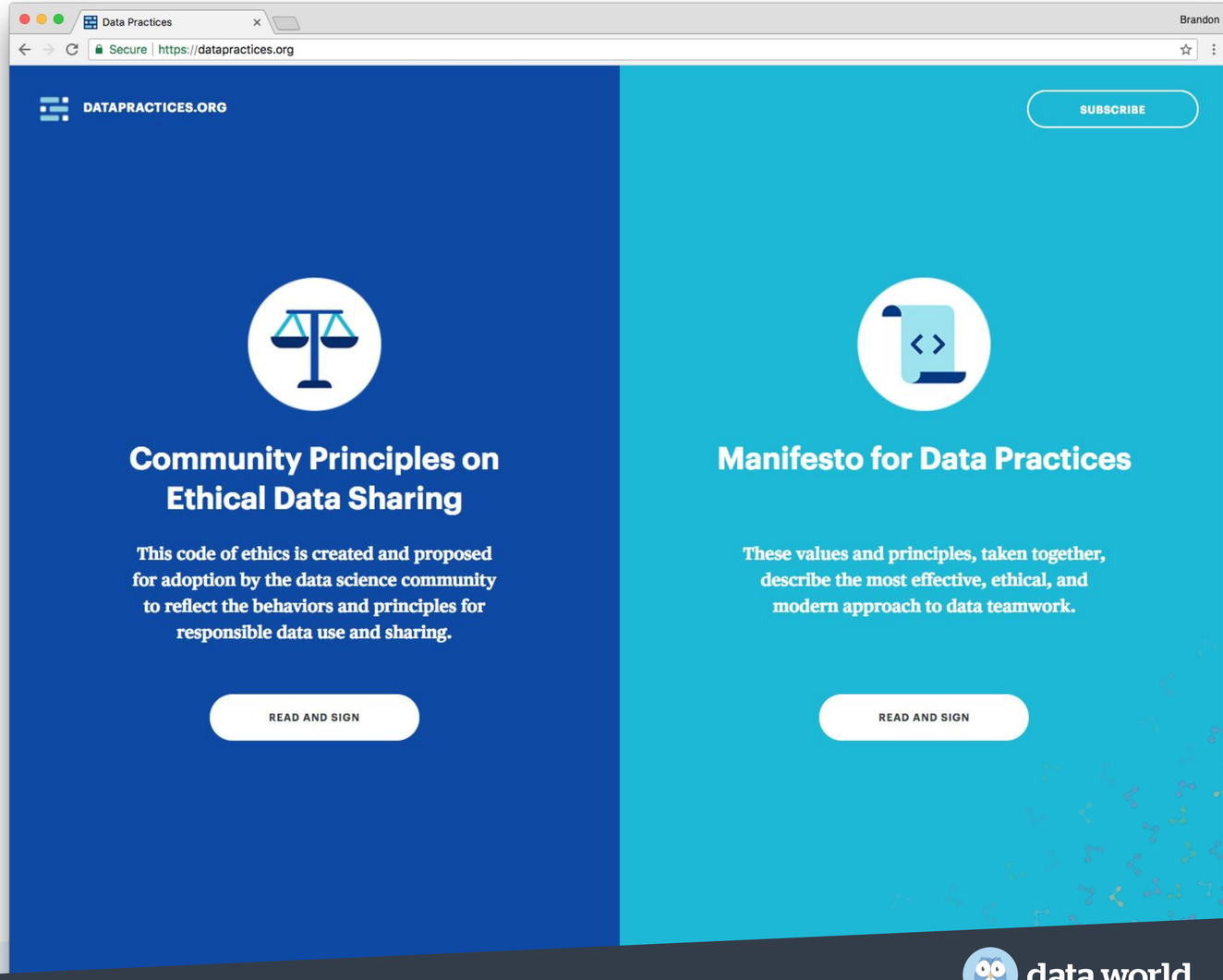
Coming soon: Exercises & workshops

## Manifesto for Data Practices

1. First draft baked
2. Community-driven
3. Sign Today!
4. **Origination:** ODSLS
5. **Focus:** Data work

## Community Principles on Ethical Data Sharing

1. Community Principles on Ethical Data Sharing
2. Work in progress
3. Community-driven
4. Community draft coming soon!
5. **Origination**: Bloomberg, D4D, Brighthive
6. **Focus**: Data Ethics



# 10 Pragmatic Lessons

# 1. Maximize Inclusion


## SME or bust

Need a subject matter expert involved to better understand data, context, and reach

## Diversity of:

- Inputs (broader data for better perspective)
- Collaborators (different outlooks will provide a more complete picture)
- Outputs (sharing data is hard, viz/numbers/procedures is important)

# 2. Foster Experimentation

## Continuous iterative testing and analysis is key!

## Close the feedback loop

- Analyst / data scientist, data engineer, SME, decision-maker
- Concurrent work when possible

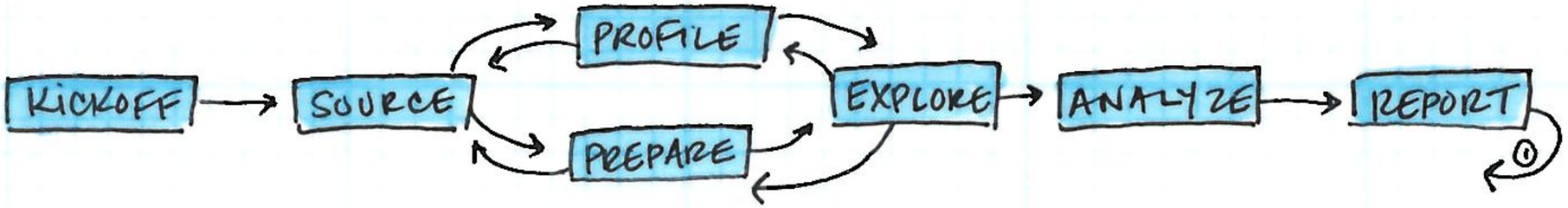
## Insights lead to future work/refinement

# 3. Start With a Question or Hypothesis

Clarity of thinking

Accountability

Focus and direction



# 4. Use Canonical Data Sources

Don’t email spreadsheets!

Centralize work and make it portable

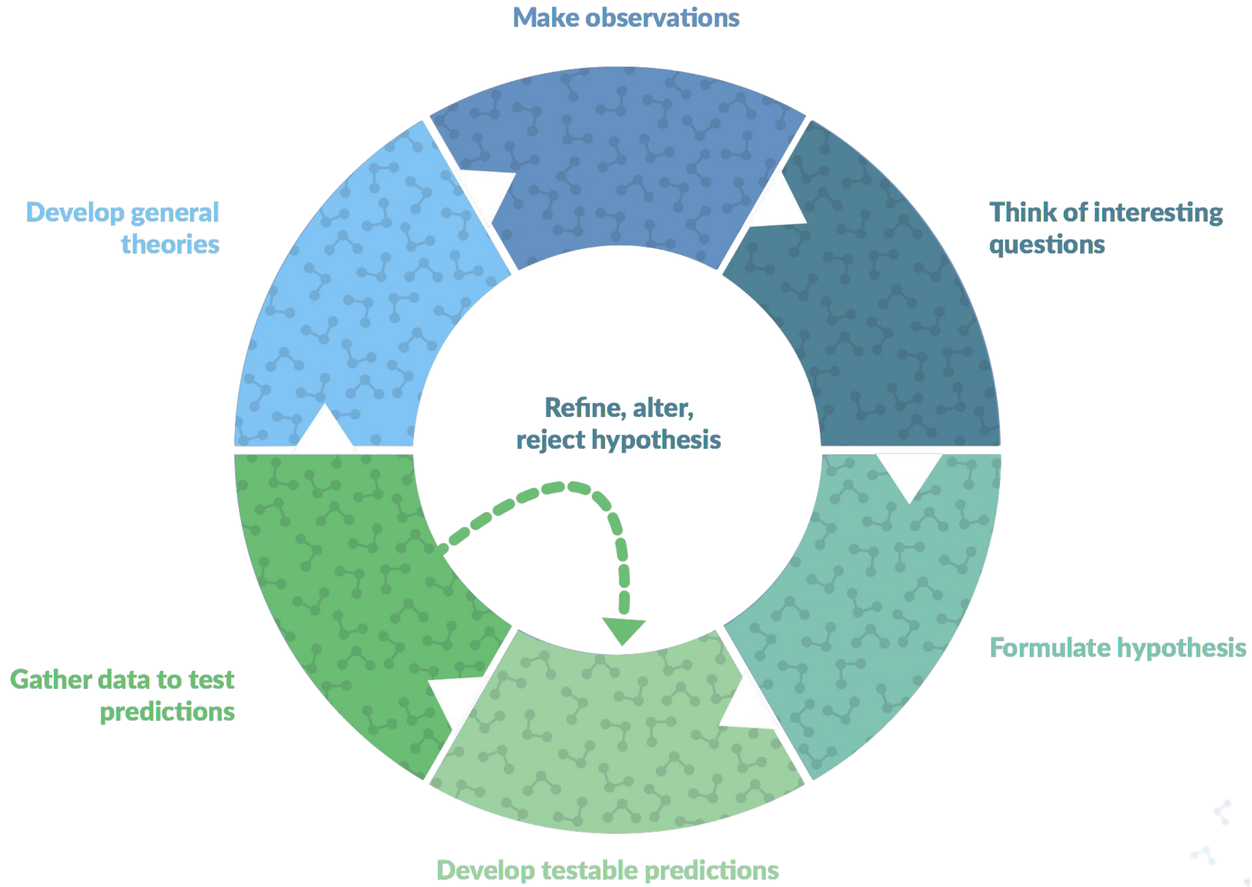
Employ the power of linked data!

# 5. Practice Good Science

## Even if you “aren’t a scientist”!

## Scientific Method

- Observation, measurement, experimentation
- Formulation, testing, and modification of hypothesis



# 6. Document Work and Processes

## Do what (good) software engineers have done for ages

## Data science methodologies are especially important to capture

## Open Approach

- Data (raw and resultant)
- Provenance
- Algorithms / code (where appropriate

# 7. Continually Build Context

## What is this?

- Conversations
- Tribal wisdom
- Q&A
- Data dictionary

## Minimize duplicative effort

## Many tools to help with this

# 8. Find and Understand Prior Work

*“Science is standing on the shoulders of ideas you believe were arrived at with sound method and challenging ideas you believe were reached by unsound methods”*

Provenance is important

Pass on to all work

# 9. Encourage Tool Agnosticism

Different personas use different tools

Different problems require different tools

Better results when multiple modes of examination occur

Portable data and open source is important for reproducibility

# 10. Consider and Measure Impact

Important to prioritize projects with well-defined goals

Design projects to achieve measurable, substantive outcomes

Introspection on “what to measure” and “what constitutes success” is powerful

# Questions?

data.world