

# 10 Pragmatic Lessons for Building Data Collaboration

Patrick McGarry – 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.





Q Search



dash/Power Grid A PRIVATE dated Jun 15, 2017 · Add a license

Contributors Discussion Activity Settings Overview

### **Power Grid Data**

Edit

+ Add to project

### Overview

The next installment in the AP-APME infrastructure initiative examines the state of the nation's electric power infrastructure: its vulnerability to cyberattacks from foreign lands, including exclusive details about one such intrusion against a major power producer; problems involved in paying for upgrades to confront extreme weather events; general reliability issues and challenges related to alternative forms of energy.

The package includes a number of data sets moving in advance to help AP members tailor the material to their audiences. The data sets include a state-by-state breakdown of the number of extreme weather-related events that have caused outages, the average duration and frequency of outages for U.S. utilities and an index of electricity providers nationwide. An index of enforcement actions taken against utilities, broken down by region, also is provided.

i	<b>Data dictionary</b> Aggregated metadata from this dataset's 6 tabular files and 6 other files							View data dictionary			
6	Drag and drop or connect to a data source. You can add up to 1020.53 MB to this dataset.							Add data			
12 files								Filter <del>-</del>	Sort -		Ξ
	cleaning.ip script Pyt	<b>ynb</b> hon notebook used to cl	ean the	data. Edit						<u>+</u>	:
				View notebo	ok						
$\triangle$	This file has	150 warnings. View al	ľ								
XLS	eiaservicet raw data	erritories2014.xls An index of electricity pr	oviders	nationwide & the counties v	vhere the	ey Edit			II Explore	<u>+</u>	:
🛗 da	ata_year 🗸 🗸	# utility_number	~	I utility_name	~	T state	~	T county	× 1		-
1 2014			24	City of Abbouillo - (SC)		00		466-122			

Edit

_ Download - Lau
Details
QUERIES (15)
Top 10 Causes of Outag
Top 10 Causes of Outag ③ @dash · 9 months ago
Violations by Region ③ @dash • 9 months ago
Total Probable Cyber At ⑤ @dash · 9 months ago
Outages per Year by Sta ③ @dash • 9 months ago
<b>FAGS</b> (4)

power grid electricity

CONTRIBUTORS (16)



😵 data.world

## **Data Practices Manifesto / CPEDS**

### https://datapractices.org

### Manifesto for Data Practices

- First draft baked
- Community-driven
- 3. Sign Today!
- **Origination:** <u>ODSLS</u> 4.
- Focus: Data work 5.

### **Community Principles on Ethical Data Sharing**

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



💮 data.world

**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!**





🛞 data.world



## **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?





