

Designing for Impact in Civic Data Science

Andrew Therriault, PhD
Northeastern University
Harvard University

PyData Amsterdam - June 17, 2020

andrewtherriault.com | twitter.com/therriaultphd | github.com/therriault

GETTING STARTED

MY BACKGROUND

- Political scientist by training, data scientist by trade
- Led data science and analytics teams at the City of Boston, Democratic National Committee, and Facebook
- Currently working as an independent consultant and teaching at Harvard and Northeastern Universities
- Co-organizer of PyData Boston chapter

TOPICS WE'LL COVER

- The potential of *civic data science* and *open data* to improve communities and provide better services to those who need them
- Examples of civic data science projects built on open data
- The challenges of developing these projects, the reasons most of them fail, and how to overcome these problems

WHAT YOU'LL LEARN

- How to find, interpret, and work with open data
- How to design effective civic data science projects to solve real-world problems
- How to develop strong partnerships that maximize the impact of your work
- How to build your projects to sustainably deliver value over the long term

An Introduction to Civic Data Science & Open Data

Civic data science: the application of data science methods* to support the work of governments, nonprofits, advocacy groups, political campaigns, and other organizations which exist to deliver positive social impacts

** broadly defined—everything from fundamental data management, analysis, and reporting to predictive models, experimental tests, and complex data viz*

THE CASE FOR CIVIC DATA SCIENCE

CIVIC ORGS ARE VITAL

The tech industry talks about making the world better (and sometimes succeeds), but their ultimate motives are still financial. Civic orgs use money primarily as a tool to promote social goals, and their direct impacts on public well-being are orders of magnitude larger.

BIG POTENTIAL IMPACTS

Two core principles of applied data science: efficiency and effectiveness. With resources almost always scarce for civic organizations, this kind of optimization can be the difference between success and failure. And even if not, marginal improvements are still very meaningful.

EXCITING CHALLENGES

Beyond just being meaningful, civic data science projects are also some of the most exciting you can work on. Subjects are varied and unique, people are smart and highly motivated, and opportunities for creativity and innovation are unparalleled.

WHAT IS OPEN DATA?

From the Open Data Handbook (emphasis added):

Open data can be **freely used, re-used and redistributed by anyone** - subject only, at most, to the requirement to attribute and sharealike...


The data must be **available as a whole** and at no more than a reasonable reproduction cost... The data must also be available in **a convenient and modifiable form**... The data must be provided under terms that **permit re-use and redistribution** including the **intermixing with other datasets**... Everyone must be able to use, re-use and redistribute - there should be **no discrimination against fields of endeavour or against persons or groups**...


For our purposes today, we'll focus on government-provided open data.

COMMON OPEN DATA TOPICS

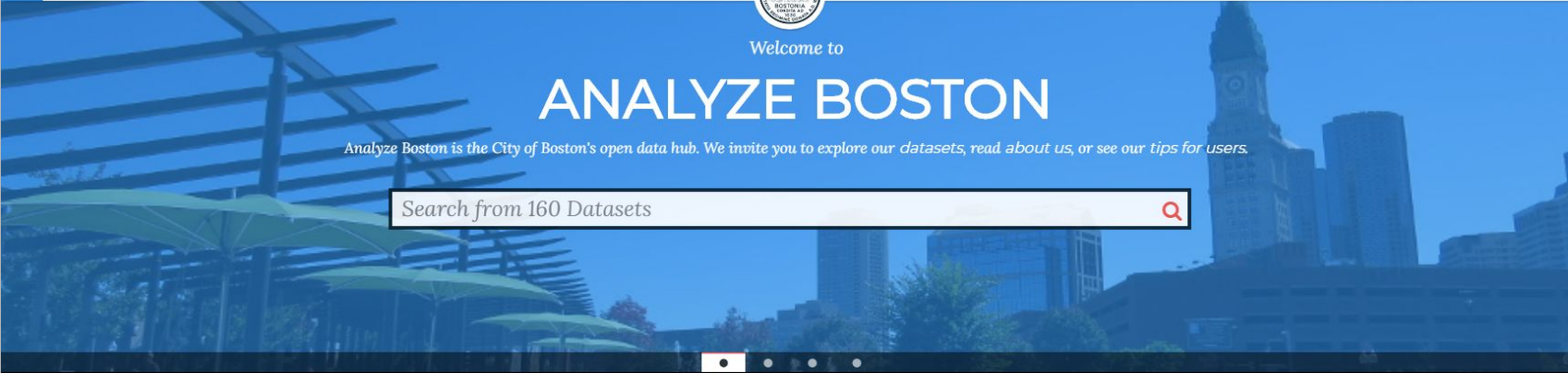
- Official research statistics
 - Population, demographics, economics, public health, weather, crime, environment
- Public services
 - Emergency services, public works, education, parking, nutrition, housing, parks
- Government operations
 - Budgets, salaries, expenditures, permits, licenses, facilities, energy usage, taxes
- Property information
 - Ownership, descriptions, assessments, transaction histories, zoning
- Geospatial data
 - Addresses, street segments, area boundaries, street features, traffic regulations

EXAMPLE: THE CITY OF BOSTON

**ANALYZE BOSTON**




DATASETS NEWS TIPS LOG IN SIGN UP CONTACT



Welcome to


ANALYZE BOSTON

Analyze Boston is the City of Boston's open data hub. We invite you to explore our *datasets*, read *about us*, or see our *tips for users*.


Search from 160 Datasets 

SHOWCASES


See what our users are doing with open data.




Our Progress Toward Carbon Neutrality



Beantown Solar




Climate Ready Boston Map Explorer




Imagine Boston 2030 Metrics Dashboard

EXAMPLE: THE CITY OF BOSTON

MENU

ANALYZE BOSTON



DATASETSNEWS TIPSLIGIN SIGN UPCONTACT

Home > Datasets

ORGANIZATIONS

Boston Maps (89)

Public Works Department (9)

Department of Innovation and Technology (6)

Environment Department (6)

Boston Transportation Department (5)

Boston Police Department (4)

Inspectional Services Department (4)

Office of Budget Management (4)

Archives and Record Management (3)

Boston Planning & Development Agency (3)

Show More Organizations

TOPICS

Geospatial (85)

City Services (17)

Environment (9)

Finance (6)


Permitting (6)

Public Safety (6)

Search datasets...

160 DATASETS FOUND


ORDER BY: Relevance



Central Library Electricity Usage

Electric power load at Boston Public Library's Central Branch (700 Boylston Street, in Copley Square) measured every five minutes.
Modified on June 15, 2020
2314 total views


CSV



Street Sweeping Schedules

This is a legacy dataset which contains detailed information on the timing and location of street sweeping service throughout the City. Daily street cleaning takes place April 1...
Modified on June 15, 2020
8946 total views


CSV



Public Works Active Work Zones

The Department of Public Works' Construction Inspection Unit (CIU) produces a daily report on all active utility and private sector construction projects taking place in the...
Modified on June 15, 2020
4876 total views

CSV

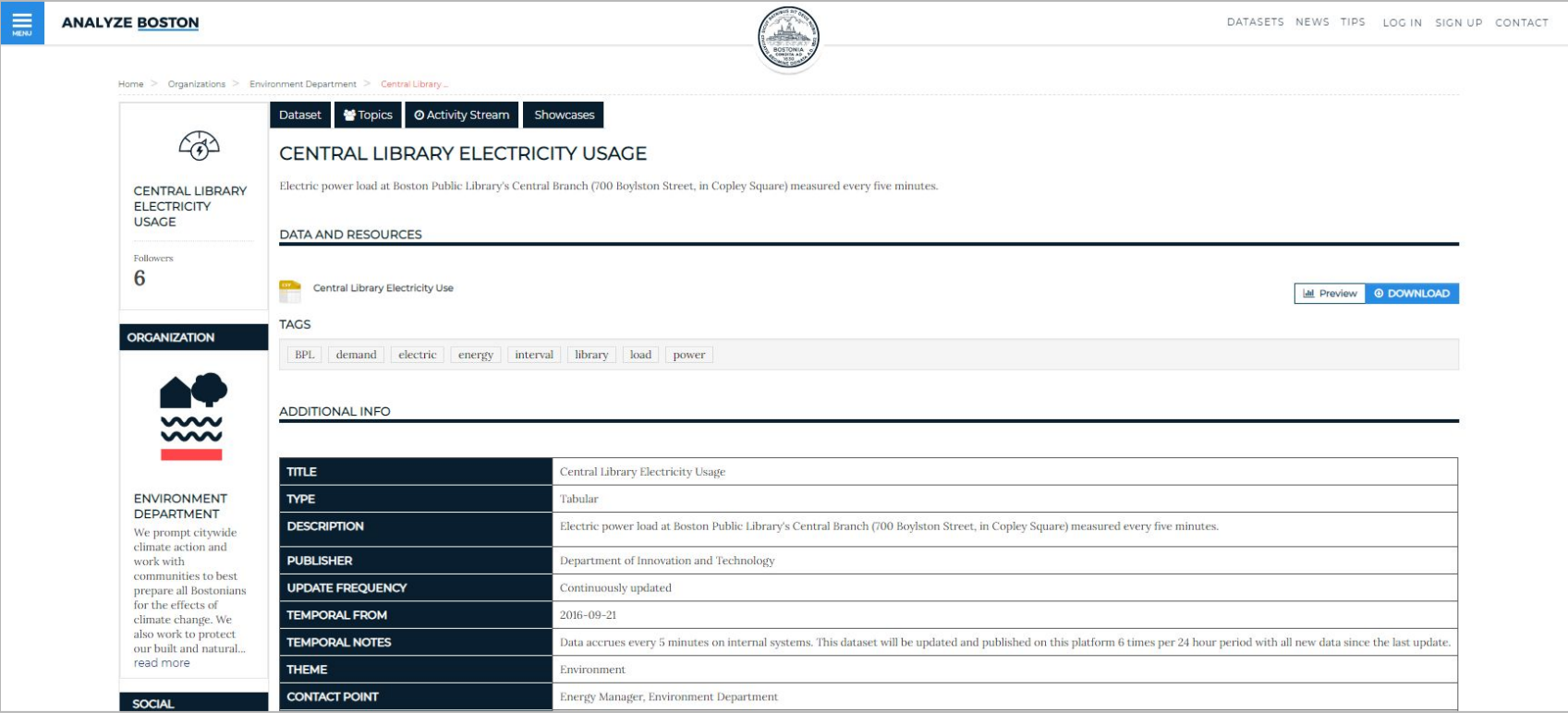


Building and Property Violations


Violations on Boston buildings or properties issued by inspectors from the Building and Structures Division of the Inspectional Services Department. Note: property_id is...
Modified on June 15, 2020
1189 total views

CSV


EXAMPLE: THE CITY OF BOSTON



EXAMPLE: THE CITY OF BOSTON

MENU

ANALYZE BOSTON



DATASETSNEWSTIPSLIGINSIGNUPCONTACT

Home > Organizations > Environment Department > Central Library ... > Central Library ...

CENTRAL LIBRARY ELECTRICITY USE

URL: <https://data.boston.gov/dataset/652762e9-2556-47cd-8e...>

FROM THE DATASET ABSTRACT

Electric power load at Boston Public Library's Central Branch (700 Boylston Street, in Copley Square) measured every five minutes.

Source: Central Library Electricity Usage

Data Table

Add Filter

Show 10 entries

Showing 1 to 10 of 343,151 entries

_id	datetime_utc_measured	total_demand_kw
1	2020-06-15 18:00:00	1209.6
2	2020-06-15 17:55:00	1188
3	2020-06-15 17:50:00	1173.6
4	2020-06-15 17:45:00	1180.8
5	2020-06-15 17:40:00	1180.8
6	2020-06-15 17:35:00	1188
7	2020-06-15 17:30:00	1159.2
8	2020-06-15 17:25:00	1173.6
9	2020-06-15 17:20:00	1188
10	2020-06-15 17:15:00	1195.2

Showing 1 to 10 of 343,151 entries

Hide/Unhide ColumnsCopyDownloadPrint

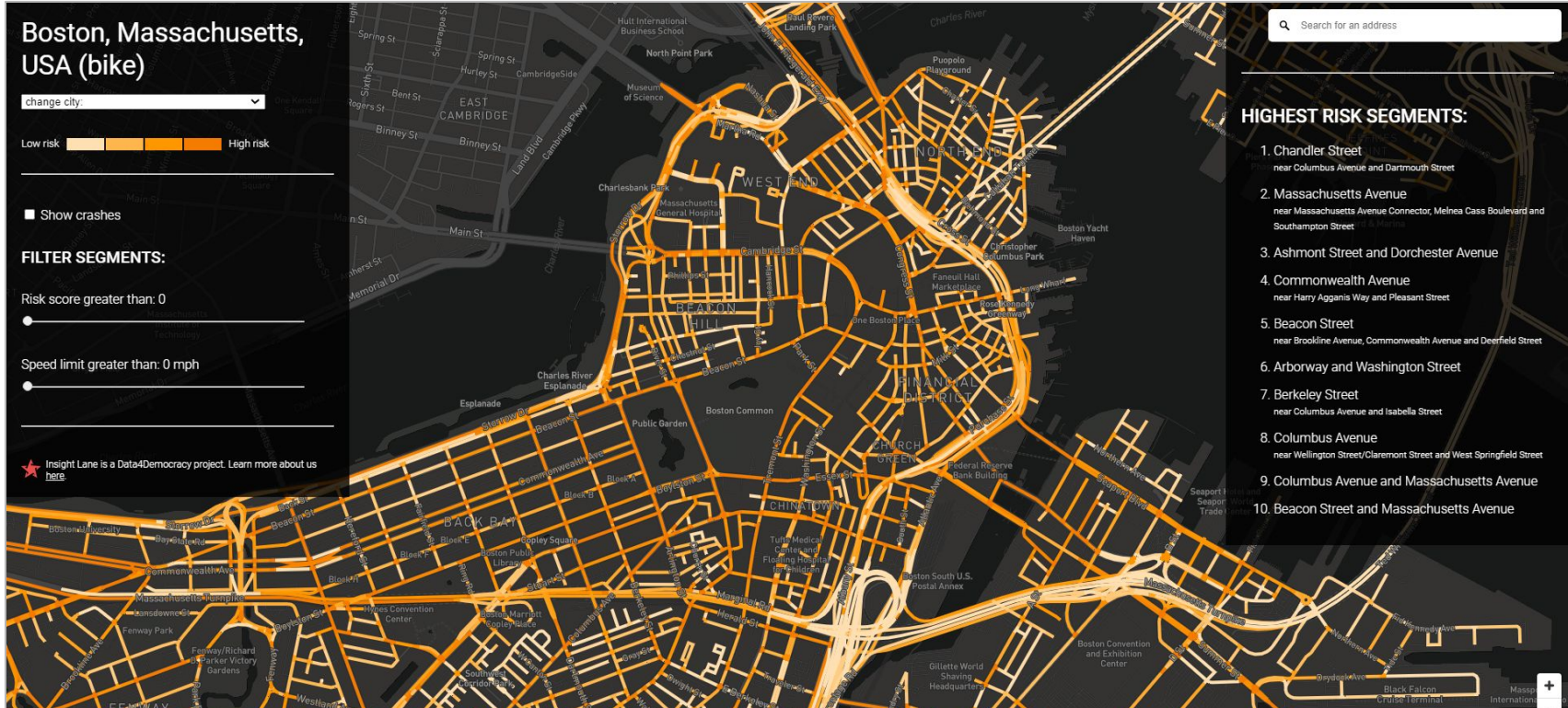
Search:

Embed

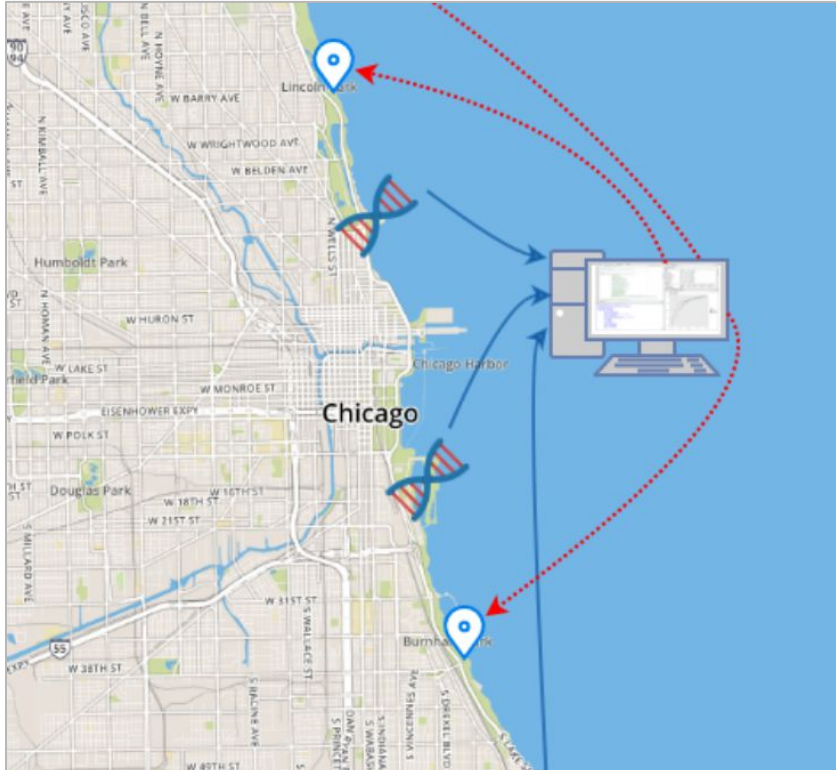
Previous12345...34316Next

A Few Example Projects

INSIGHT LANE - CRASH RISK MODELING FOR CITIES



CHICAGO CLEAN WATER - E. COLI PREDICTION



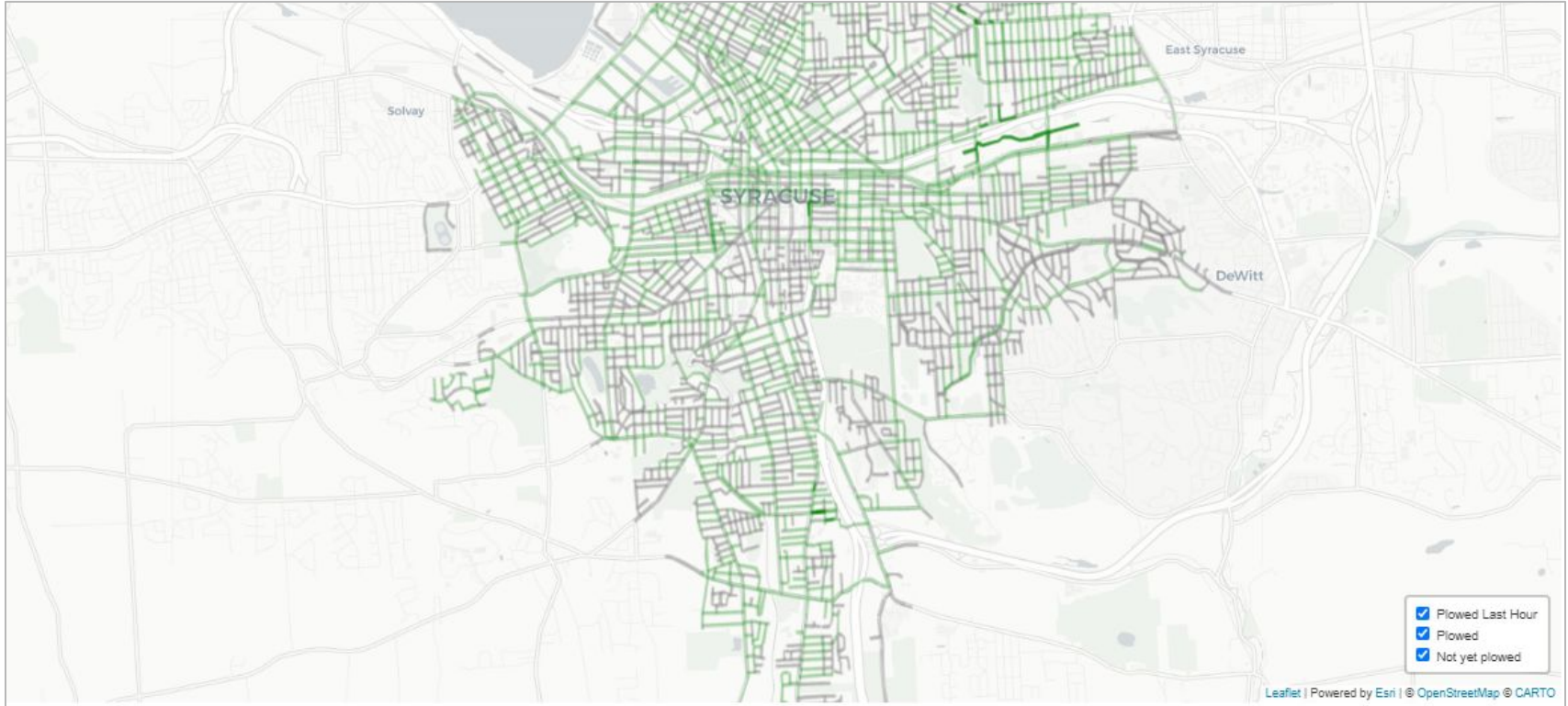
INNOVATIVE SCIENCE

Prior efforts to forecast beach water quality in Chicago and elsewhere have adopted similar approaches. Researchers would collect meteorological data near a swimming site, and then predict contamination levels in the water. However, this methodology can be unreliable and often does not identify days with high *E. coli* levels.

This project developed a new approach that takes advantage of rapid DNA testing and the lessons learned from data exploration. That approach begins by acknowledging that just five beaches contribute to about 56% of poor water quality beach days. These beaches, which are some of the hardest to predict, should be routinely rapid tested due to their volatility. Water quality patterns at the remaining beaches can then be separated into clusters. In the new approach, one beach from each of these clusters would be rapid tested to get an immediate result. The remaining beaches would be predicted by the model.

A key feature of the new approach is its cost effectiveness. The increased cost of rapid testing would be offset because only half the beaches would be tested. Yet it performs better and provides more accurate notifications to the public.

SYRACUSE SNOWPLOW TRACKER



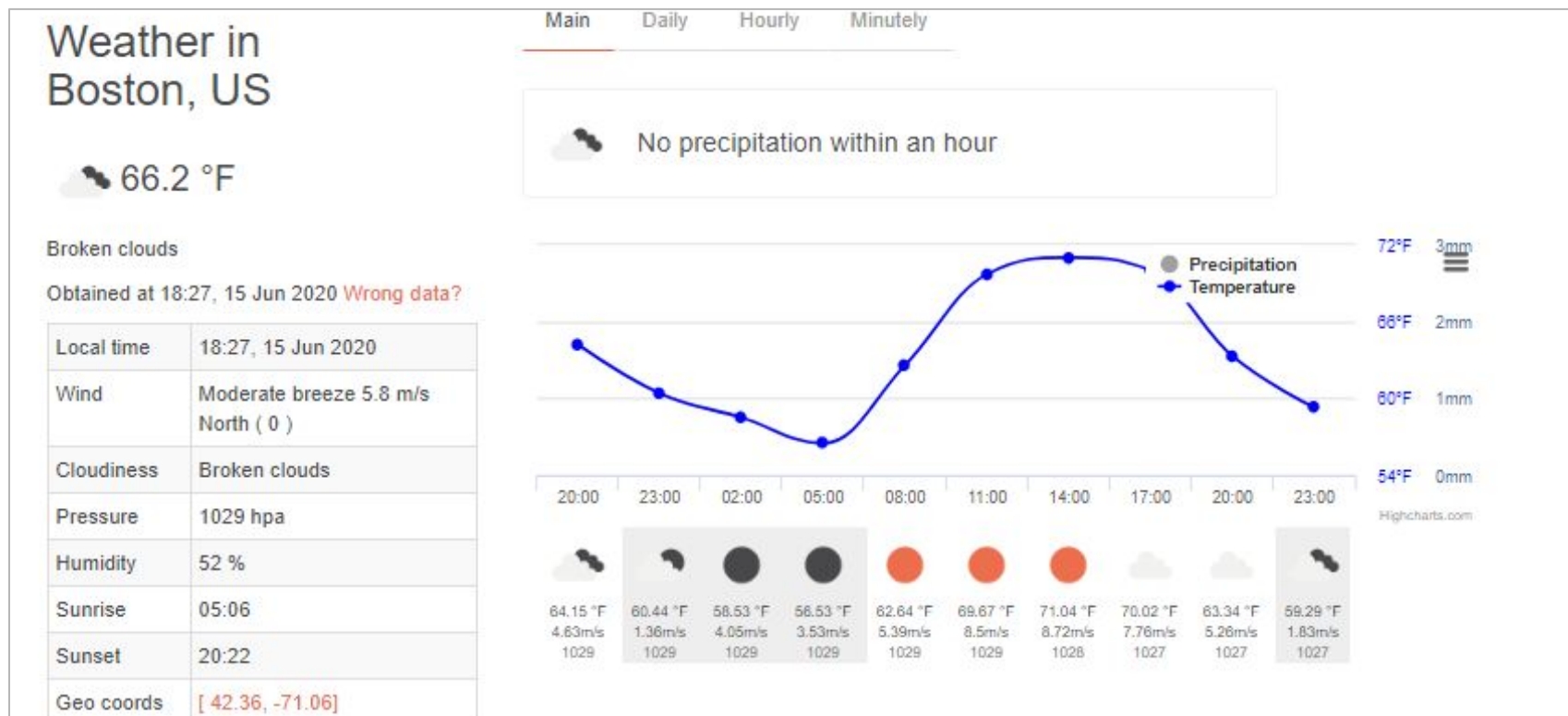
<http://www.innovatesyracuse.com/blog/snowplowmapdevelopment>

SAN DIEGO FOR EVERY CHILD: CHILD POVERTY MAP



<https://childpovertymap.sandiegoforeverychild.org/>

OPEN WEATHER FORECASTS AND CONDITIONS



Challenges and Solutions

WHY DO MOST PROJECTS FAIL TO DELIVER?

Common problems that keep civic data science projects from succeeding:

1. Poor understanding of the data, subject matter, stakeholders, or audience
2. No clear “theory of change” or specific use case to design around
3. Failure to connect with potential end users (e.g., government agencies)
4. Over-reliance on single points of failure with limited availability
5. Lack of quality control and refinement of end products
6. Insufficient plans for deployment and ongoing support/maintenance

Addressing these problems is what differentiates a successful civic data science project from the thousands of hackathon submissions and class projects built on open data every year.

THE IMPORTANCE OF PROJECT DESIGN

- Data science projects combine traditional research design with engineering, product development, and project management
- The design phase of a project requires you to gather information and make decisions about what you want to accomplish
- What a project design offers:
 - Clarifies the purpose, scope, and deliverables of your project
 - Identifies the key stakeholders, partners, and end users
 - Defines the requirements for success and highlights potential risks
 - Provides an opportunity to get concrete feedback before development starts
 - Gives the project team a shared vision and roadmap to work from
- Good design is key to *all* kinds of data science projects, but the particular challenges of civic projects make it even more critical for success

10 KEY QUESTIONS TO ASK WHEN DESIGNING

1. What is the overall goal you're working toward?
2. Who are your stakeholders, partners, and end users?
3. What do your users want to know, and how will that help them succeed?
4. What inputs do you need, and how will you get them?
5. How will you transform this data, and what exactly does that tell you?
6. What kind of infrastructure, outside support, resources, time, or other requirements are necessary to complete the project?
7. What potential risks could make your project fail to deliver an accurate or useful result in a timely manner?

10 KEY QUESTIONS TO ASK WHEN DESIGNING

8. How will you check the accuracy and validity of your results, and are there any external reference points you can compare them to?
9. What's the best way to provide your results to end users, and what kind of support (training, documentation, etc.) needs to come with it?
10. What happens to your project after it's released?

BUILDING EFFECTIVE PARTNERSHIPS

- Civic data science projects will almost always fail if they do not include partners working in the subject area involved (i.e., not data scientists!)
- These partners help throughout the process:
 - Understanding the subject matter, identifying research questions, and defining use cases
 - Finding and interpreting relevant data
 - Putting end results into practice through changes to policies or programs
 - Potentially taking long-term ownership or finding resources for support
- Recognize the incentives and limitations of potential partners, and be willing to consider multiple alternatives (especially early)
- Be respectful of their time and expertise, and be open to receiving feedback and changing direction

DESIGNING FOR LONG-TERM SUSTAINABILITY

- Unless your end product is a one-time analysis, your project is *not* done when you release a “finished” version
- Just some of the long-term needs you’ll want to think about:
 - Documentation and training, both for users and future maintainers
 - Ongoing infrastructure resources (servers, storage, etc.)
 - Changes to data sources, dependencies, or integrated applications
 - Future monitoring, quality control, refinements, and break-fixes
 - Eventual retirement or replacement
- A poorly-supported project can be costly to your end users, *and might be a worse outcome than if you hadn’t given them anything at all*
- If at all possible, consider handing off project to a person / group that gets paid for maintaining it, rather than relying on volunteers

Final Thoughts

A FEW LAST SUGGESTIONS

- Always remember that data scientists add value by providing information to those who can make use of it - *we really don't do anything on our own!*
- If you're passionate about a specific topic or organization, spend some time learning how it all works before diving head-first into data
- You don't need to have all the answers before getting started, but you should at least know what you don't know
- Be open to changing direction or walking away from a project if the original plan isn't working out
- If things do go well, think about how your work can be scaled or repurposed for use in other locations or context

WHERE TO GO FROM HERE

There are plenty of ways to get started:

- Start exploring available open data resources on your own
- Connect with the people running your local open data program
- Get involved in community groups and talk to data producers
- Find potential collaborators in your local PyData chapter or similar groups
- Participate in hackathons and other open data events

No matter how you start, just remember: if you want to have a truly meaningful impact, you need to design your project for long-term success, then execute accordingly.

Thank you!

*I'll be available for questions afterward on Discord,
or reach out on Twitter @therriaultphd. Slides are
available at <https://github.com/therriault/slides>.*
