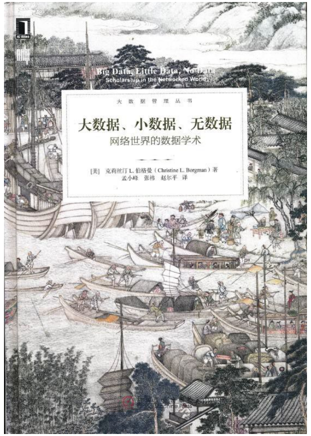
Capturing, Maintaining, and Exploiting Data for the Public Good

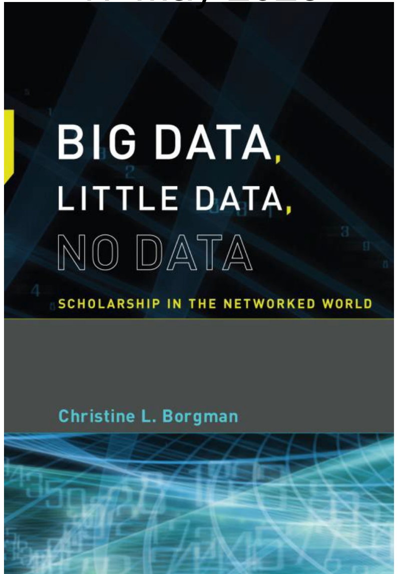
Christine L. Borgman

Distinguished Research Professor, Information Studies, UCLA

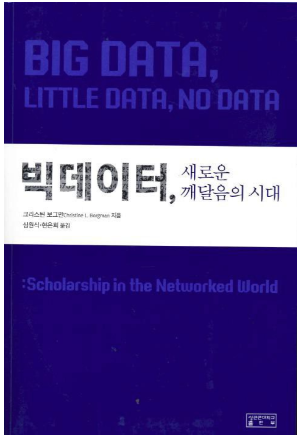
University of California Center, Sacramento, Public Lecture Series
17 May 2023

ビッグデータ・
リトルデータ・
ノーデータ
研究データと知識インフラ
クリスティン L. ボーグマン［著］
佐藤義則・小山憲司［訳］


Big Data, Little Data, No Data
Scholarship in the Networked World
大数据管理丛书
大数据、小数据、无数据
网络世界的数据学术
[美] 克莉丝汀 L. 伯格曼（Christine L. Borgman）著
孟小峰 张伟 赵尔平 译


BIG DATA,
LITTLE DATA,
NO DATA
SCHOLARSHIP IN THE NETWORKED WORLD
Christine L. Borgman

QU'EST-CE QUE
LE TRAVAIL SCIENTIFIQUE DES DONNÉES ?
Big data, little data, no data
CHRISTINE L. BORGMAN


BIG DATA,
LITTLE DATA, NO DATA
빅데이터, 새로운 깨달음의 시대
크리스틴 보그만Christine L. Borgman 지음
심원식·현은희 옮김
:Scholarship in the Networked World

Research data for the public good

Why share research data?

- Reuse
- Reproduce
- Transparent
- Educate
- Required
  - Funding agencies
  - Journals

How to share research data?

- Link datasets to publication
- Deposit in data archive
- Publish data documentation
  - Research protocols
  - Codebooks
  - Software
  - Algorithms
- Cite data and software
- Develop instructional materials

erc European Research Council
Supporting top researchers from anywhere in the world

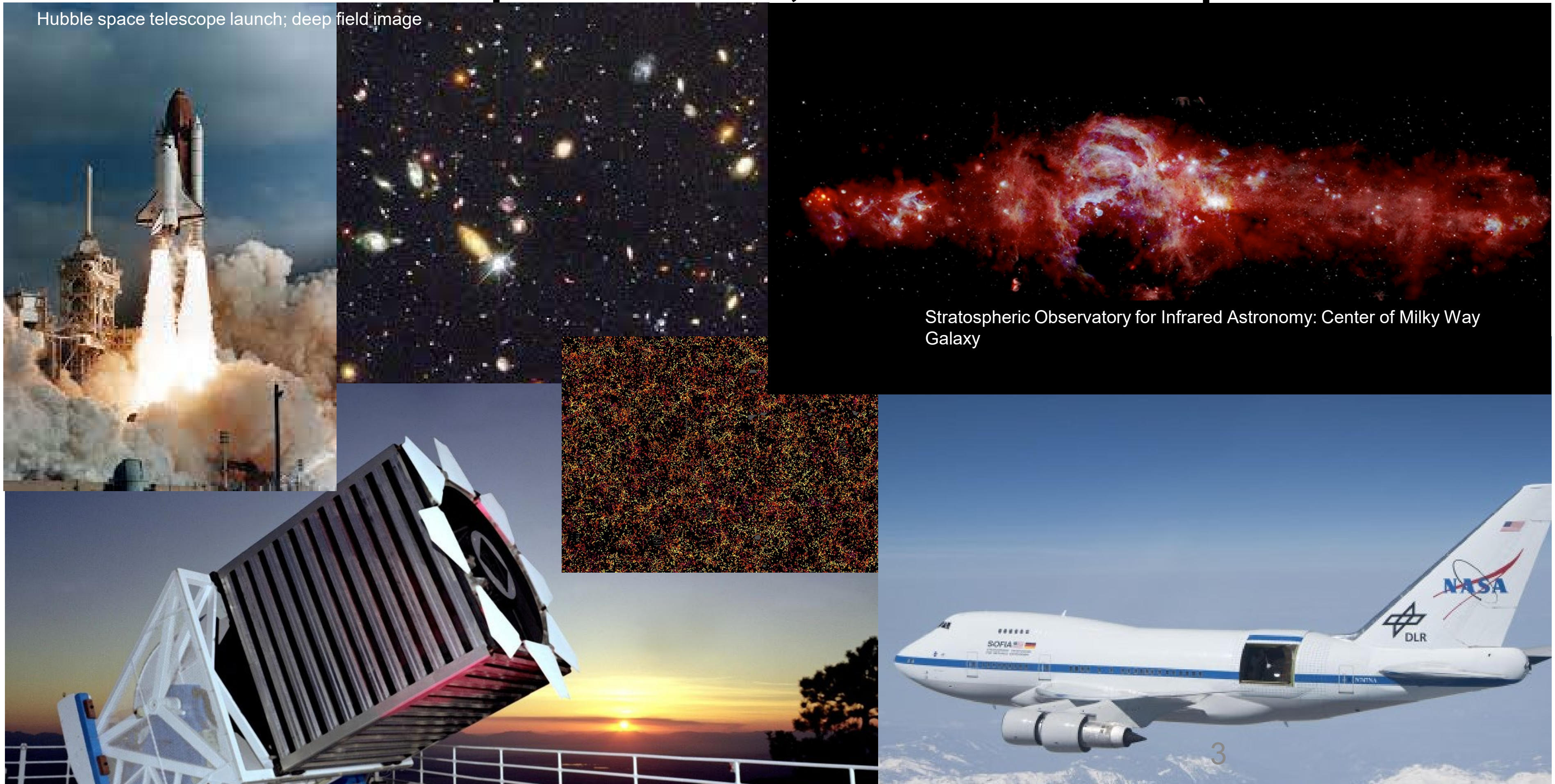
E·S·R·C
ECONOMIC & SOCIAL RESEARCH COUNCIL

National Science Foundation
WHERE DISCOVERIES BEGIN

University of California
CDL
California Digital Library

NIH CDE Repository

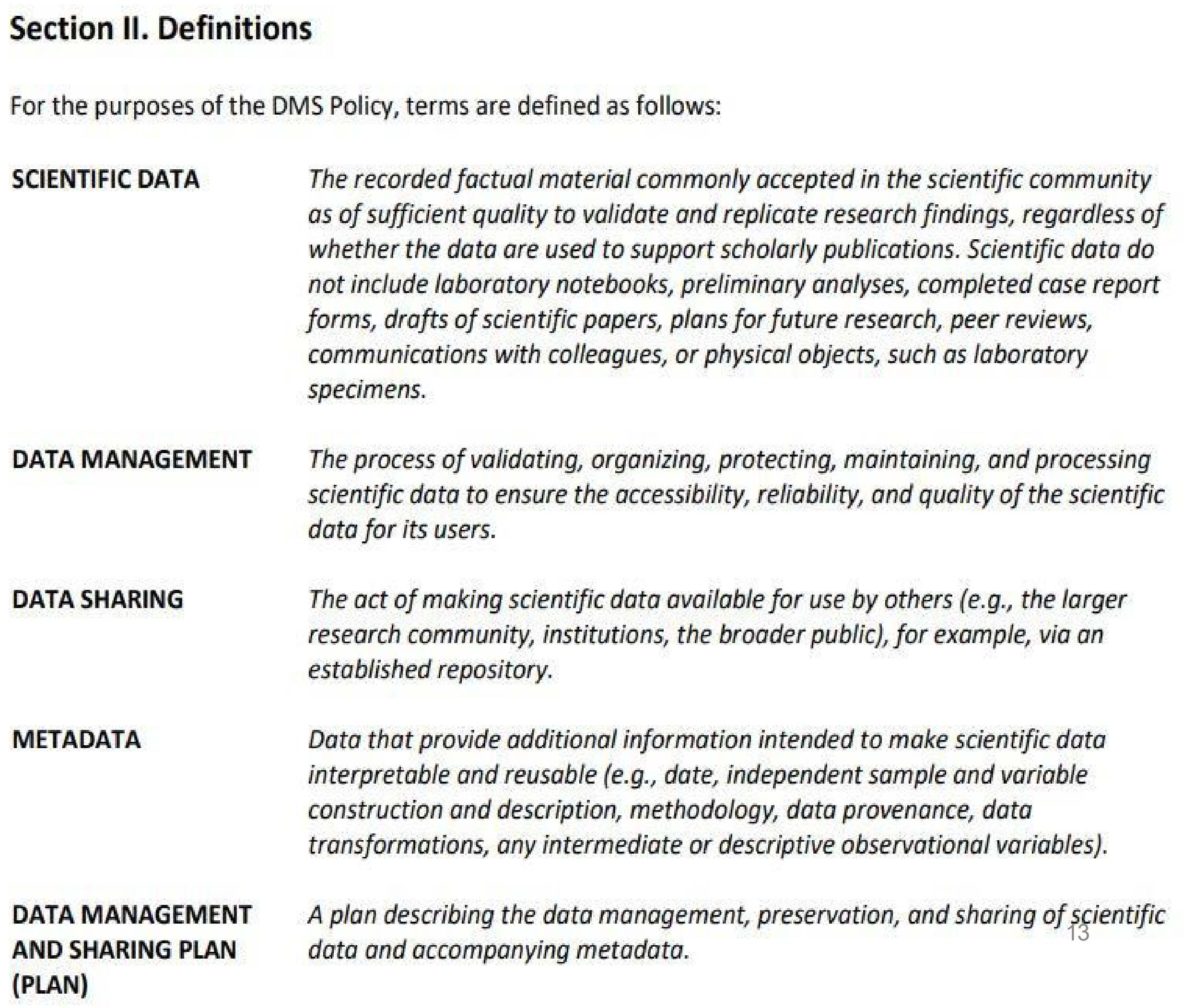
Decades to acquire data, decades to preserve



Hubble space telescope launch; deep field image

Stratospheric Observatory for Infrared Astronomy: Center of Milky Way Galaxy

National Institutes of Health Data Sharing Policy 2023



Section II. Definitions

For the purposes of the DMS Policy, terms are defined as follows:

SCIENTIFIC DATA *The recorded factual material commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications. Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens.*

DATA MANAGEMENT *The process of validating, organizing, protecting, maintaining, and processing scientific data to ensure the accessibility, reliability, and quality of the scientific data for its users.*

DATA SHARING *The act of making scientific data available for use by others (e.g., the larger research community, institutions, the broader public), for example, via an established repository.*

METADATA *Data that provide additional information intended to make scientific data interpretable and reusable (e.g., date, independent sample and variable construction and description, methodology, data provenance, data transformations, any intermediate or descriptive observational variables).*

DATA MANAGEMENT AND SHARING PLAN (PLAN) *A plan describing the data management, preservation, and sharing of scientific data and accompanying metadata.*

Knowledge infrastructures

Robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds (Edwards, 2010)

- Policy frameworks
- Scholarly practices
- Technical infrastructures
- Governance models

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. MIT Press.

Research data infrastructure: Stakeholders

- Research funding agencies
- Individual scientists and scholars
- Academic institutions
  - Academic leadership
  - Research Computing
  - University libraries
  - Schools and departments
- Students and teachers
- General public

Photo by Mihai Surdu on Unsplash

Borgman, C. L., & Bourne, P. E. (2022). Why It Takes a Village to Manage and Share Data. *Harvard Data Science Review*, *4*(3).
Borgman, C. L., & Brand, A. (2022). Data blind: Universities lag in capturing and exploiting data. *Science*, *378*(6626), 1278–1281.

Research data interdependencies

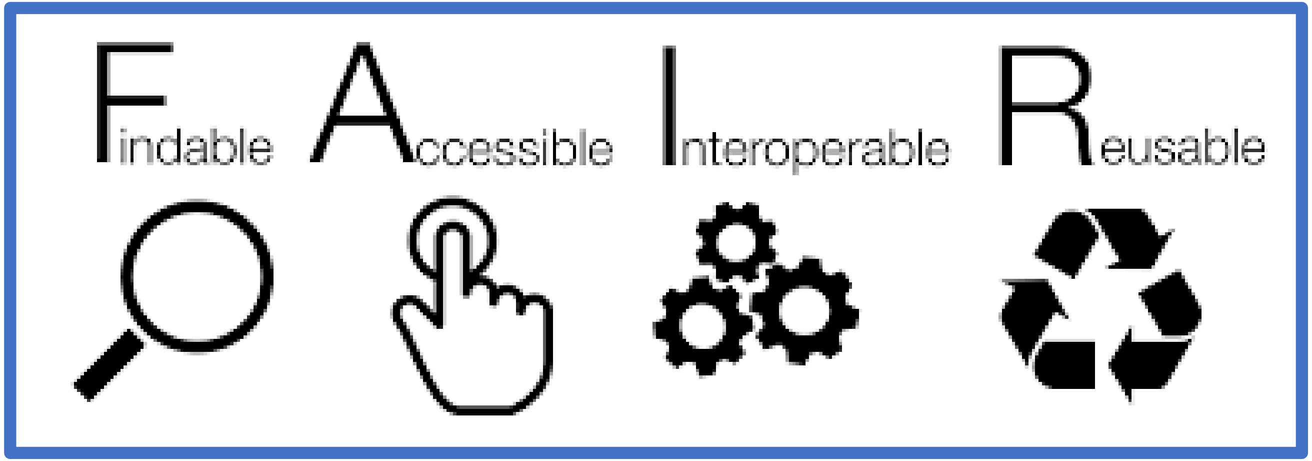
Thinking globally, acting locally

- What data to share
- Data, context, and credit
- Data access and discovery
- Data assets as research methods
- Intellectual property in data
- Domain expertise in data
- Misuses of data

Borgman, C. L., & Bourne, P. E. (2022). Why It Takes a Village to Manage and Share Data. *Harvard Data Science Review*, *4*(3).
Borgman, C. L., & Brand, A. (2022). Data blind: Universities lag in capturing and exploiting data. *Science*, *378*(6626), 1278–1281.

Royce Hall, UCLA

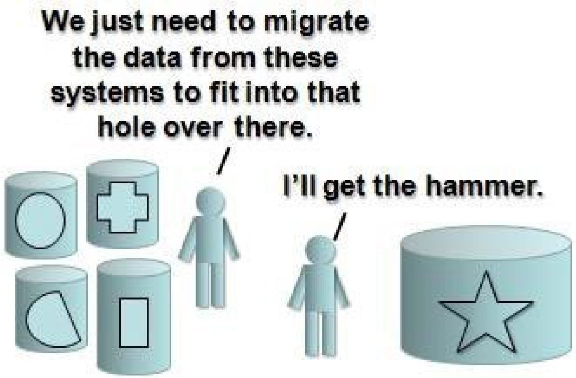
Data Sharing and Stewardship: The Ideal



Wilkinson, et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, http://dx.doi.org/10.1038/sdata.2016.18

Data Stewardship: The Reality

http://www.information-age.com/cloud-computing-pharmaceutical-industry-123462676/


We just need to migrate the data from these systems to fit into that hole over there.
I'll get the hammer.

http://www.datamartist.com/data-migration-part-1-introduction-to-the-data-migration-delema

Mount Wilson Solar Observatory, 2017

Getty Research Institute

http://gsa.rice.edu/

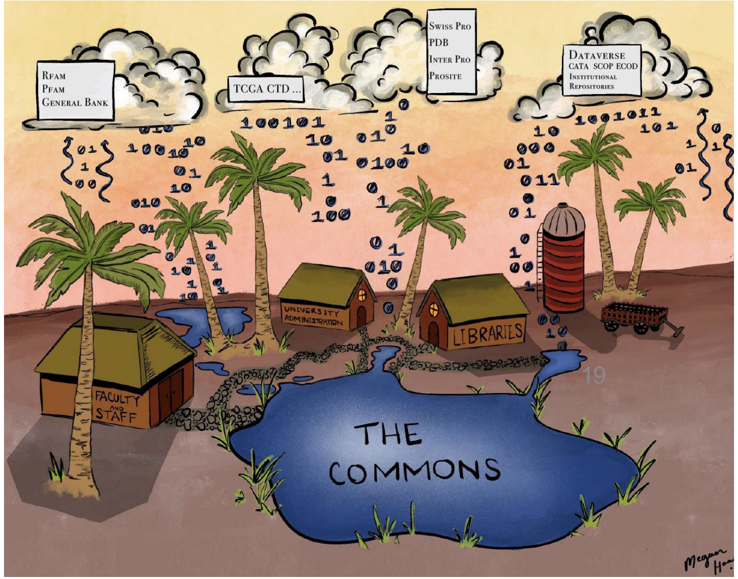
Graduate students

https://techcrunch.com/2019/07/24/researchers-spotlight-the-lie-of-anonymous-data/

General public

Governance: Building the Village

- Data sharing is a ‘collective action problem’
- Holistic approaches to sharing infrastructure
  - Distribute responsibility among stakeholders
  - Invest in data management expertise
  - Reframe goals in collective terms
- Fund the commons
  - Public support for data repositories
  - International exchange of best practices
- Invest in sustainable strategies



Borgman, C. L., & Bourne, P. E. (2022). Why it takes a village to manage and share data. *Harvard Data Science Review.* Illustration by Megan Haas

Research data for the public good

- Capture
  - Common data formats
  - Metadata, documentation
  - Software
- Maintain
  - Knowledge infrastructures
  - Data archives
  - Stewardship investments
- Exploit
  - FAIR principles
  - Technical and domain expertise
- Educate
  - Create, access, utilize
  - Assess, trust, interpret

California

Accelerating Data Excellence Across the State

UC Center Sacramento | 17 May 23

Joy Bonaguro, Statewide Chief Data Officer

Statewide Data Goals

Equipping ourselves to navigate the data landscape

Build the data roads

streamline data access

Craft the rules of the road

improve data management and governance

Boost the travelers

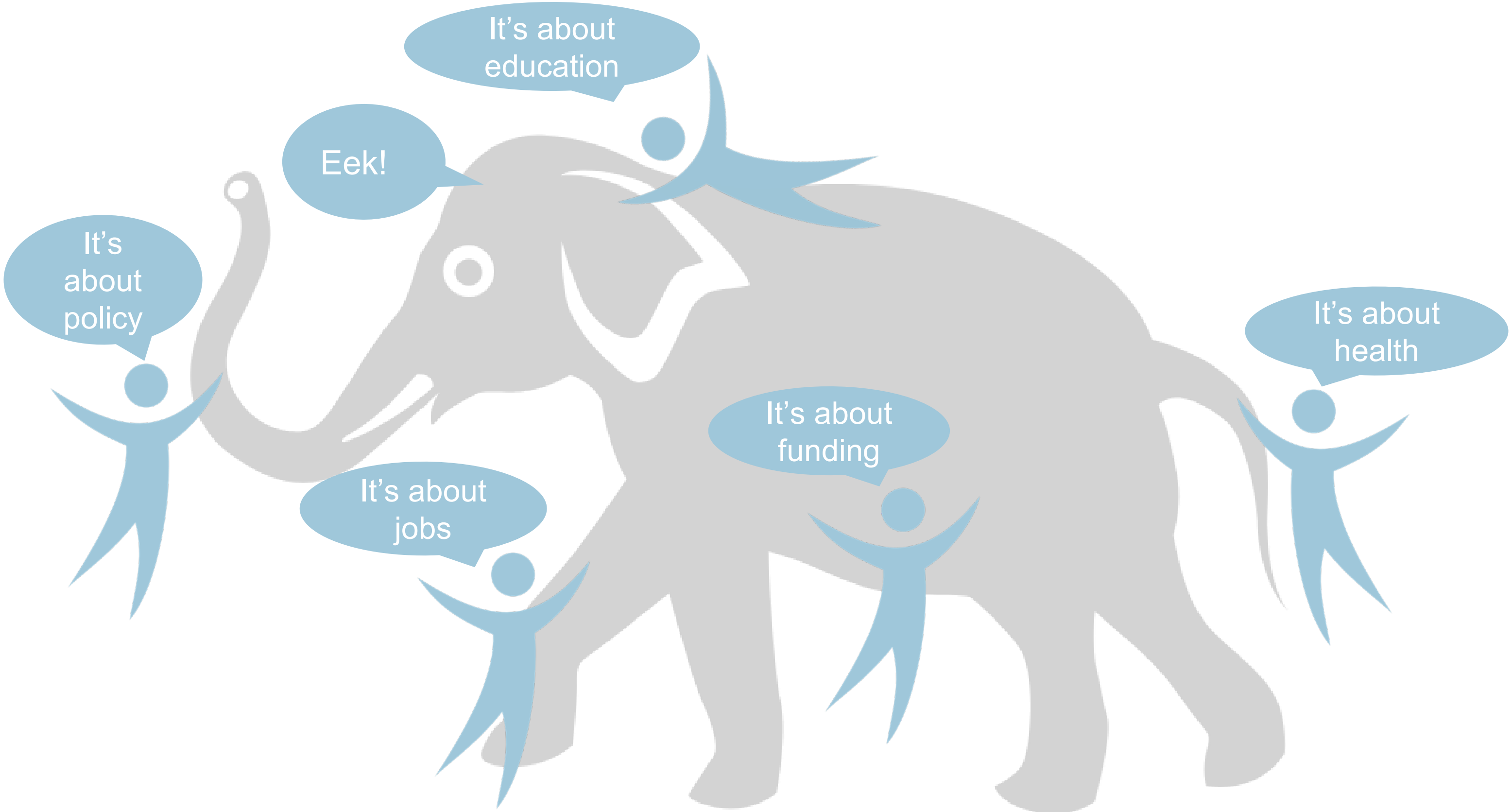
spur data use and ability

Why we need to intentionally build our data roads

Our data is housed in silos

Why data access matters:

If you only see part of the problem, you may solve for the wrong thing



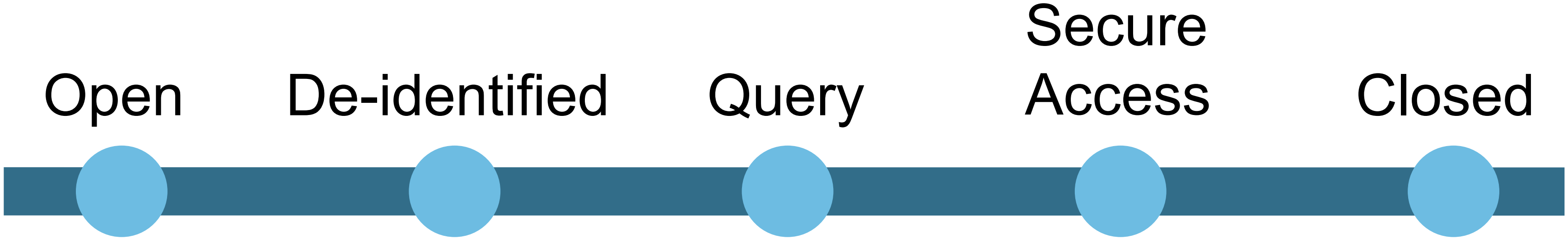
Task: Obtain Data

Hey! You got this data?
Maybe...
Can I have it?
Hello ????
Here it is...
Thanks! Some ?s...
Question death spiral

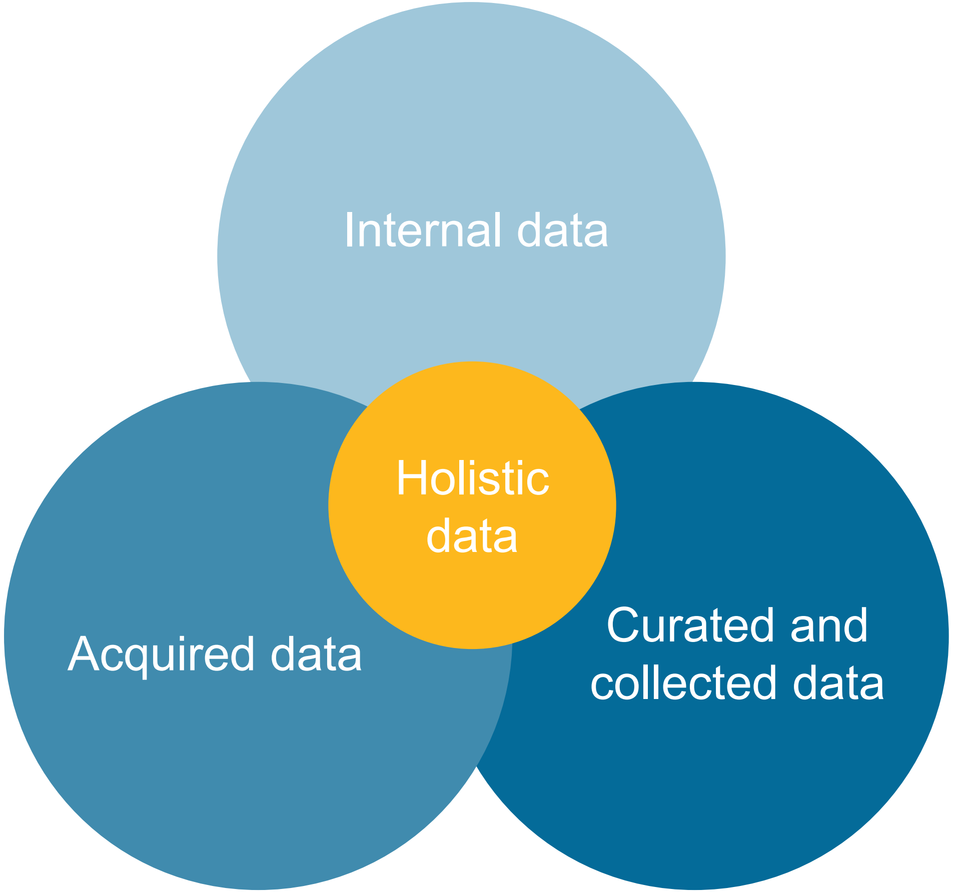
Task: Update Data

See previous

We must architect a spectrum of access



Across the full spectrum of data sources



Strategic Goals:

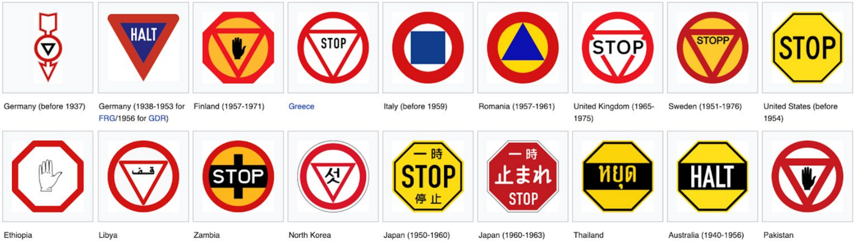
Equipping ourselves to navigate the data landscape

Build the data roads
streamline data access

Craft the rules of the road
improve data management and governance

When would you want to travel?

A: Before 1968



Germany (before 1937)

Germany (1938-1953 for FRG/1956 for GDR)

Finland (1957-1971)

Greece

Italy (before 1959)

Romania (1957-1961)

United Kingdom (1965-1975)

Sweden (1951-1976)

United States (before 1954)

Ethiopia

Libya

Zambia

North Korea

Japan (1950-1960)

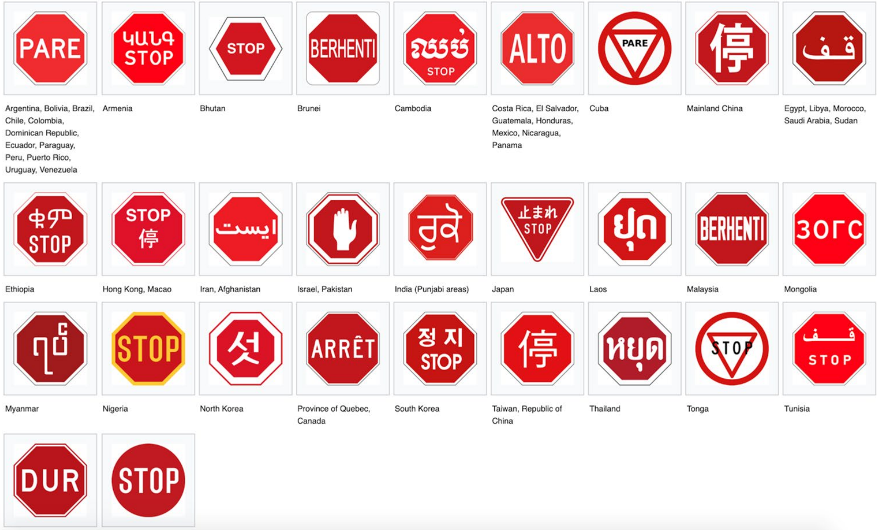
Japan (1960-1963)

Thailand

Australia (1940-1956)

Pakistan

B: After 1968



Argentina, Bolivia, Brazil, Chile, Colombia, Dominican Republic, Ecuador, Paraguay, Peru, Puerto Rico, Uruguay, Venezuela

Armenia

Bhutan

Brunei

Cambodia

Costa Rica, El Salvador, Guatemala, Honduras, Mexico, Nicaragua, Panama

Cuba

Mainland China

Egypt, Libya, Morocco, Saudi Arabia, Sudan

Ethiopia

Hong Kong, Macao

Iran, Afghanistan

Israel, Pakistan

India (Punjabi areas)

Japan

Laos

Malaysia

Mongolia

Myanmar

Nigeria

North Korea

Province of Quebec, Canada

South Korea

Taiwan, Republic of China

Thailand

Tonga

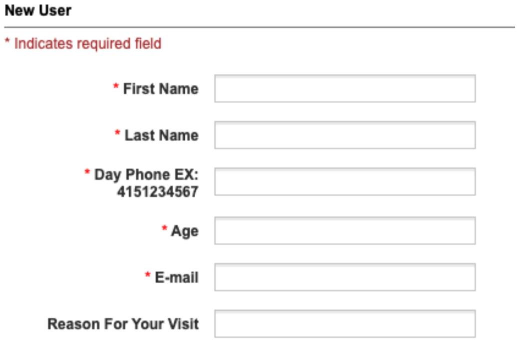
Tunisia

“At least we are consistently inconsistent.”

http://www.matusonconsulting.com/blog/2015/05/inconsistent-consistency/

4 vaccine signups; 4 race/ethnicity options

a) Not included

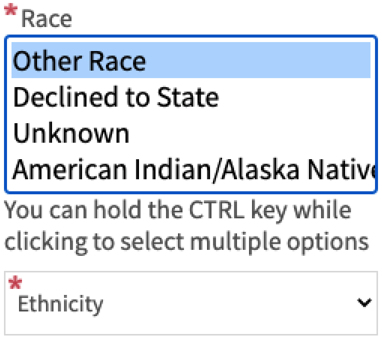


New User

* Indicates required field

* First Name	
* Last Name	
* Day Phone EX: 4151234567	
* Age	
* E-mail	
Reason For Your Visit	

c) Scroll forever



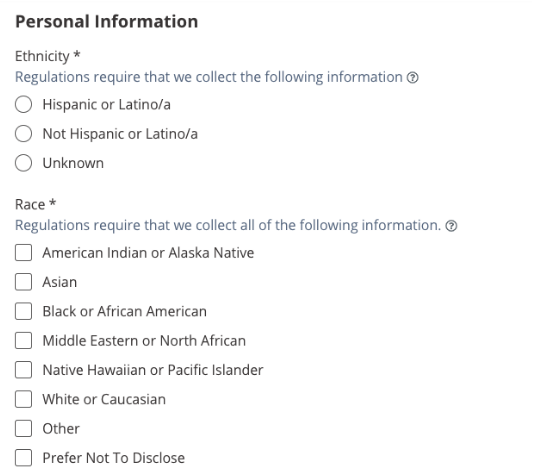
*Race

Other Race
Declined to State
Unknown
American Indian/Alaska Native

You can hold the CTRL key while clicking to select multiple options

* Ethnicity

d) At a glance



Personal Information

Ethnicity *

Regulations require that we collect the following information

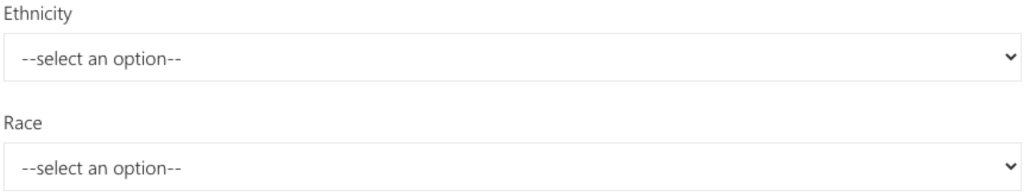
○ Hispanic or Latino/a
○ Not Hispanic or Latino/a
○ Unknown

Race *

Regulations require that we collect all of the following information.

□ American Indian or Alaska Native
□ Asian
□ Black or African American
□ Middle Eastern or North African
□ Native Hawaiian or Pacific Islander
□ White or Caucasian
□ Other
□ Prefer Not To Disclose

b) Not required



Ethnicity

--select an option--

Race

--select an option--

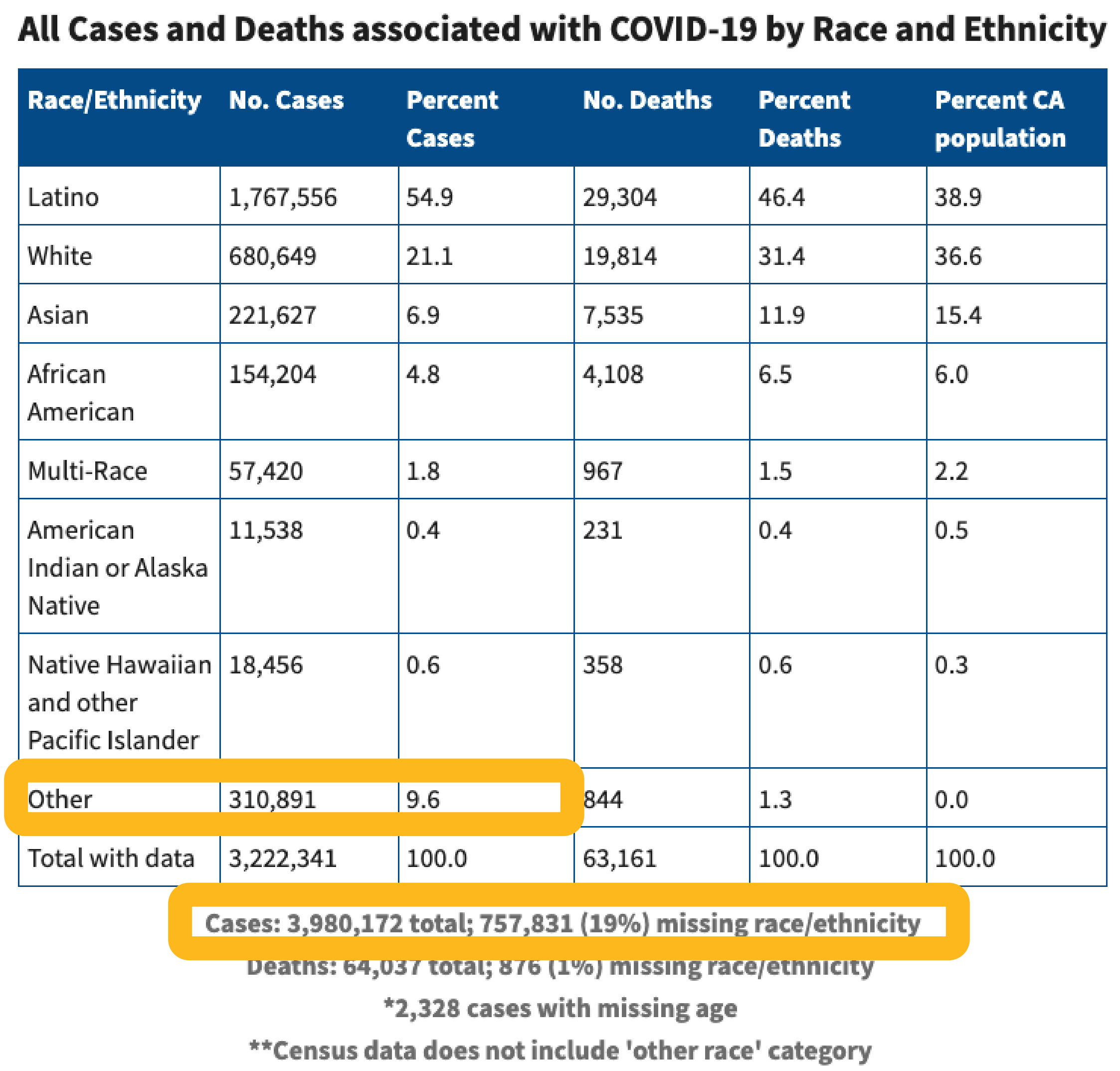
Silent audience poll: Who do you think has the best data?

a) Not included

b) Not required

c) Scroll forever

d) At a glance



All Cases and Deaths associated with COVID-19 by Race and Ethnicity

Race/Ethnicity	No. Cases	Percent Cases	No. Deaths	Percent Deaths	Percent CA population
Latino	1,767,556	54.9	29,304	46.4	38.9
White	680,649	21.1	19,814	31.4	36.6
Asian	221,627	6.9	7,535	11.9	15.4
African American	154,204	4.8	4,108	6.5	6.0
Multi-Race	57,420	1.8	967	1.5	2.2
American Indian or Alaska Native	11,538	0.4	231	0.4	0.5
Native Hawaiian and other Pacific Islander	18,456	0.6	358	0.6	0.3
Other	310,891	9.6	844	1.3	0.0
Total with data	3,222,341	100.0	63,161	100.0	100.0

Cases: 3,980,172 total; 757,831 (19%) missing race/ethnicity

Deaths: 64,037 total; 876 (1%) missing race/ethnicity

*2,328 cases with missing age

**Census data does not include 'other race' category

Strategic Goals:

Equipping ourselves to navigate the data landscape

Build the data roads
streamline data access

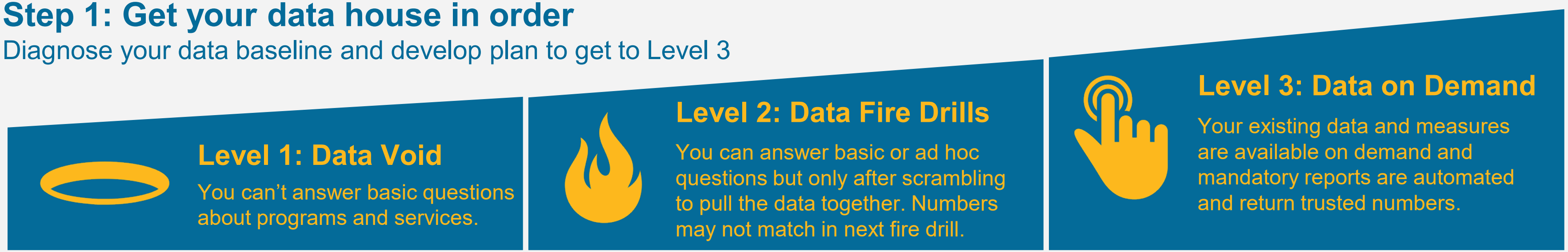
Craft the rules of the road
improve data management and governance

Boost the travelers
spur data use and ability

2 step plan to build data maturity

Step 1: Get your data house in order

Diagnose your data baseline and develop plan to get to Level 3



Level 1: Data Void

You can't answer basic questions about programs and services.

Level 2: Data Fire Drills

You can answer basic or ad hoc questions but only after scrambling to pull the data together. Numbers may not match in next fire drill.

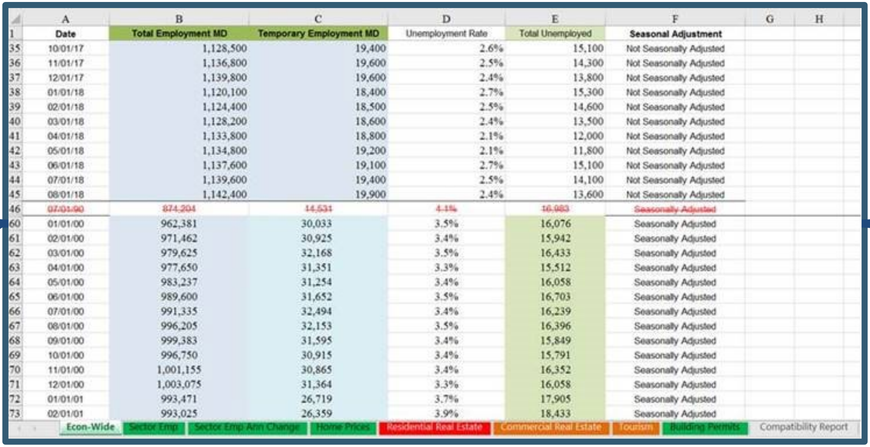
Level 3: Data on Demand

Your existing data and measures are available on demand and mandatory reports are automated and return trusted numbers.

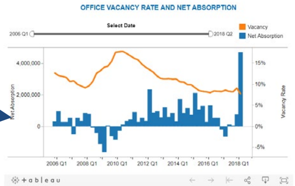
Analytics Accelerator

Behind the scenes of many data teams across the state

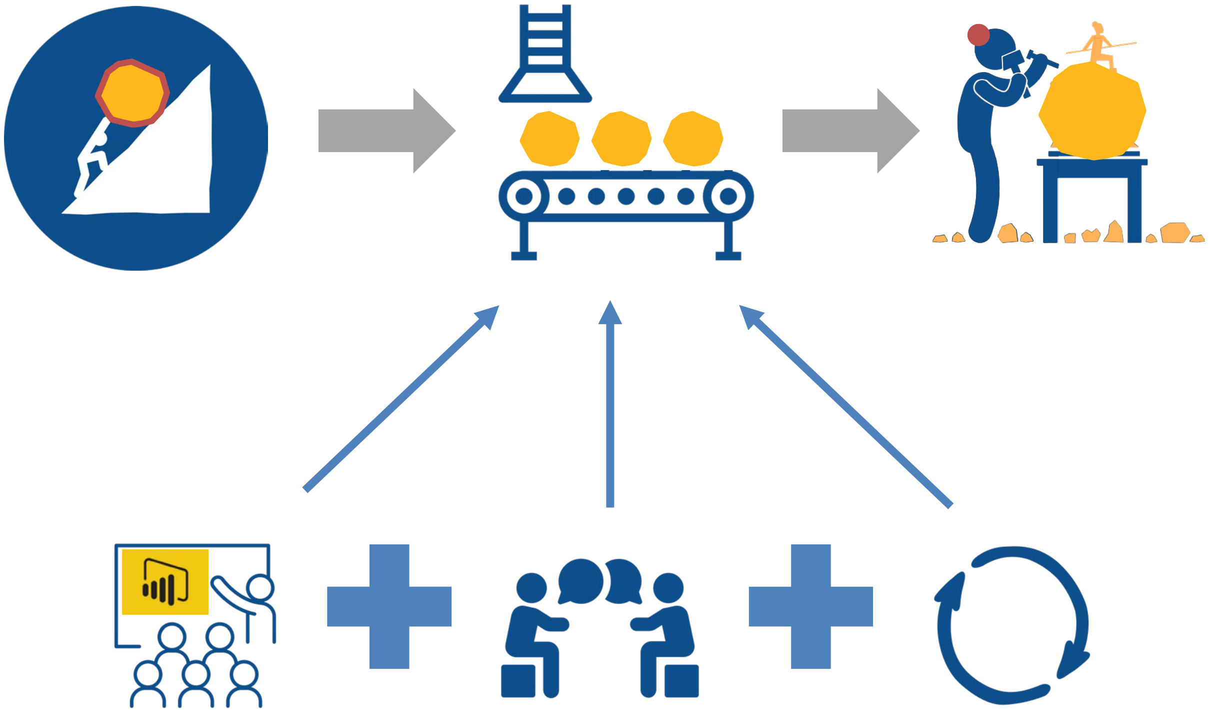
Courtesy of Arizona Historical Society



	A	B	C	D	E	F	G	H
1	Date	Total Employment MD	Temporary Employment MD	Unemployment Rate	Total Unemployed	Seasonal Adjustment		
35	10/01/17	1,128,500	19,400	2.6%	15,100	Not Seasonally Adjusted		
36	11/01/17	1,136,800	19,600	2.5%	14,300	Not Seasonally Adjusted		
37	12/01/17	1,139,800	19,600	2.4%	13,800	Not Seasonally Adjusted		
38	01/01/18	1,120,100	18,400	2.7%	15,300	Not Seasonally Adjusted		
39	02/01/18	1,124,400	18,500	2.5%	14,600	Not Seasonally Adjusted		
40	03/01/18	1,128,200	18,600	2.4%	13,500	Not Seasonally Adjusted		
41	04/01/18	1,133,800	18,800	2.1%	12,000	Not Seasonally Adjusted		
42	05/01/18	1,134,800	19,200	2.1%	11,800	Not Seasonally Adjusted		
43	06/01/18	1,137,600	19,100	2.7%	15,100	Not Seasonally Adjusted		
44	07/01/18	1,139,600	19,400	2.5%	14,100	Not Seasonally Adjusted		
45	08/01/18	1,142,400	19,900	2.4%	13,600	Not Seasonally Adjusted		
46	~~07/01/90~~	~~874,204~~	~~14,531~~	~~4.1%~~	~~16,983~~	~~Seasonally Adjusted~~		
60	01/01/00	962,381	30,033	3.5%	16,076	Seasonally Adjusted		
61	02/01/00	971,462	30,925	3.4%	15,942	Seasonally Adjusted		
62	03/01/00	979,625	32,168	3.5%	16,433	Seasonally Adjusted		
63	04/01/00	977,650	31,351	3.3%	15,512	Seasonally Adjusted		
64	05/01/00	983,237	31,254	3.4%	16,058	Seasonally Adjusted		
65	06/01/00	989,600	31,652	3.5%	16,703	Seasonally Adjusted		
66	07/01/00	991,335	32,494	3.4%	16,239	Seasonally Adjusted		
67	08/01/00	996,205	32,153	3.5%	16,396	Seasonally Adjusted		
68	09/01/00	999,383	31,595	3.4%	15,849	Seasonally Adjusted		
69	10/01/00	996,750	30,915	3.4%	15,791	Seasonally Adjusted		
70	11/01/00	1,001,155	30,865	3.4%	16,352	Seasonally Adjusted		
71	12/01/00	1,003,075	31,364	3.3%	16,058	Seasonally Adjusted		
72	01/01/01	993,471	26,719	3.7%	17,905	Seasonally Adjusted		
73	02/01/01	993,025	26,359	3.9%	18,433	Seasonally Adjusted		



What is the Analytics Accelerator?



Facilitated Trainings

Wrap around support and 1:1 consulting

Champion development & enduring training materials

Expected outcomes

- Automate manual reporting tasks
- Extra staff capacity to tackle impactful projects
- Faster more robust analytics via training in data concepts (Data Viz, Data Modeling, Data Security/Governance)

Modern Data Stack Accelerator

An example

Modern Data Stack Accelerator

We're having trouble tracking our supply of personal protective equipment across two separate systems…can your team help?

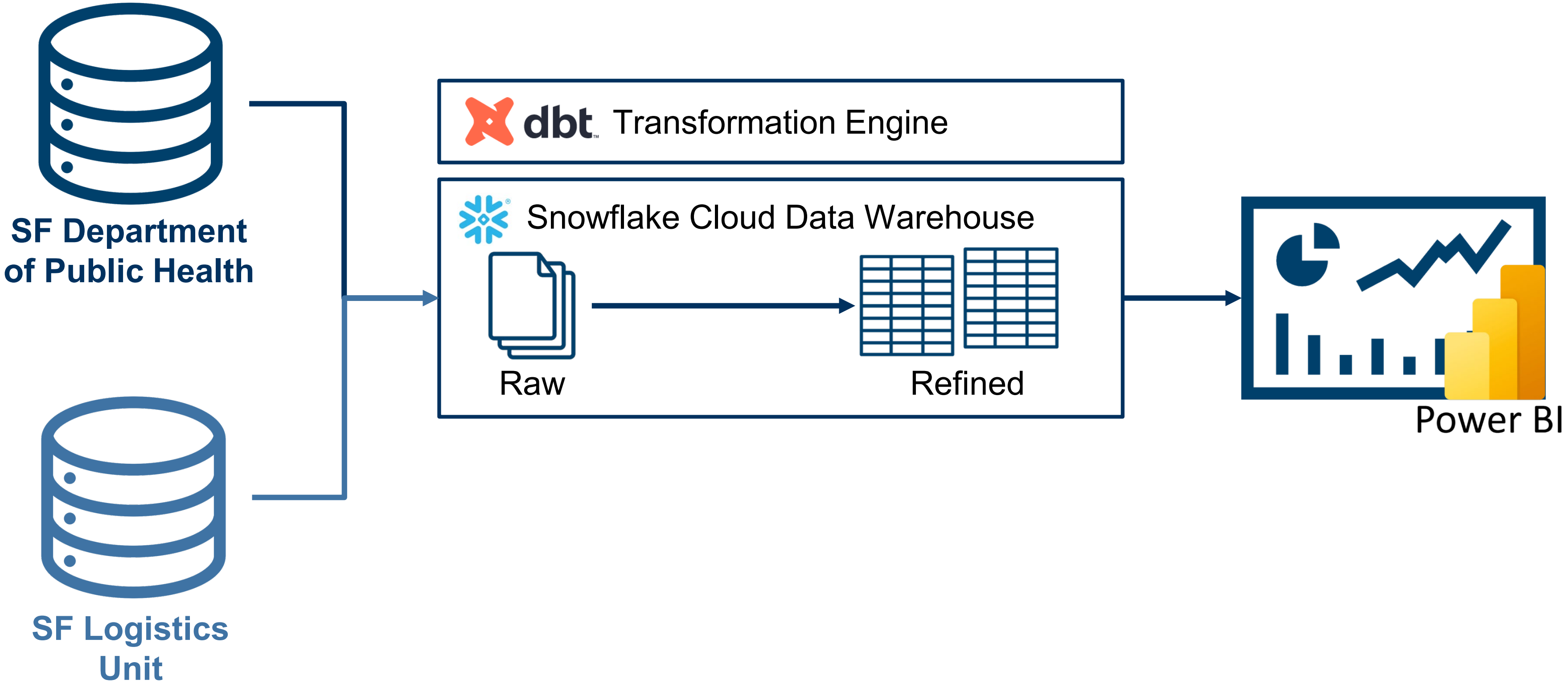
SF Department of Public Health

SF Logistics Unit

Exam Gloves	1 box
Nitrile Gloves	100 pairs
Medical Gloves	200 gloves

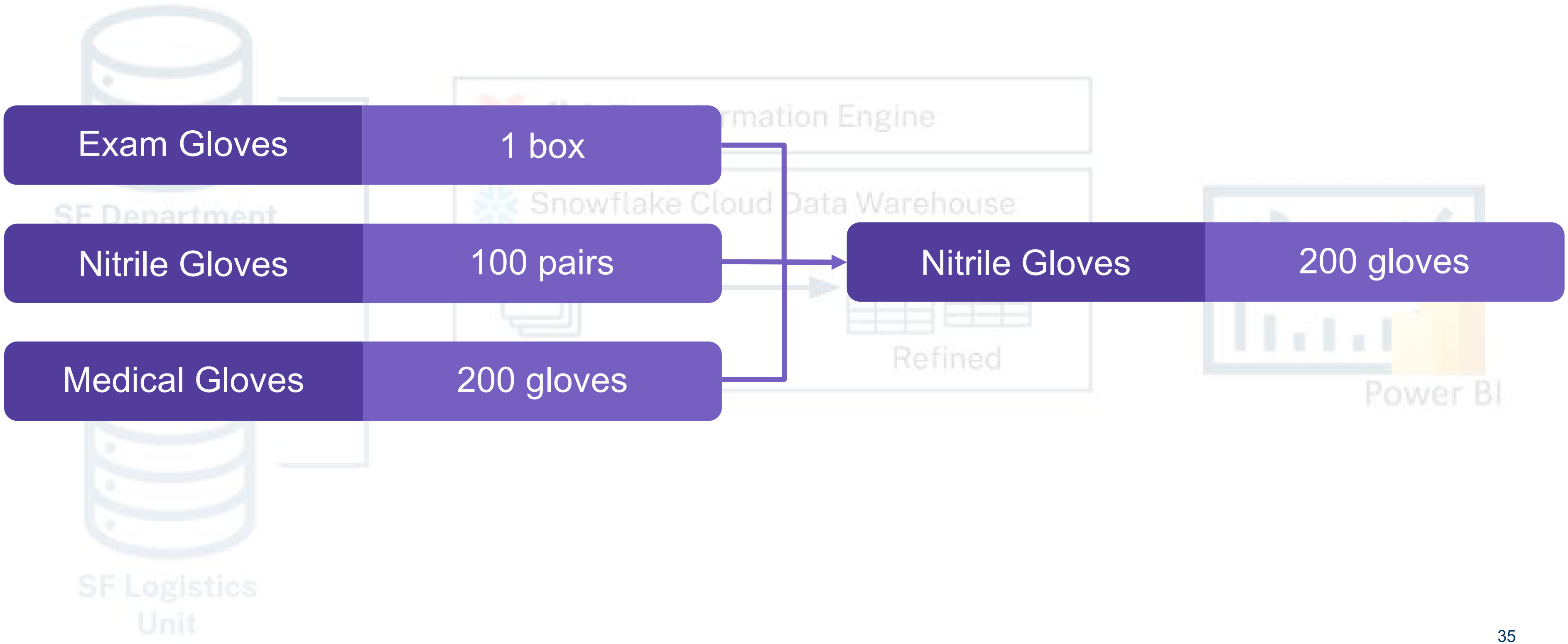
An example

Modern Data Stack Accelerator


SF Department of Public Health
SF Logistics Unit
dbt Transformation Engine
Snowflake Cloud Data Warehouse
Raw
Refined
Power BI

An example

Modern Data Stack Accelerator



Exam Gloves	1 box
Nitrile Gloves	100 pairs
Medical Gloves	200 gloves

→

Nitrile Gloves	200 gloves

What problems can be solved?

Modern Data Stack Accelerator

Difficulty combining multiple datasets efficiently and automatically for analysis and reporting

Difficulty automating manual data quality efforts for high visibility analyses or reporting

Difficulty moving, querying, and analyzing large datasets

Difficulty assessing or trying new data tools to see if they work for you before making an investment

2 step plan to build data maturity

Step 1: Get your data house in order

Diagnose your data baseline and develop plan to get to Level 3

Level 1: Data Void

You can't answer basic questions about programs and services.

Level 2: Data Fire Drills

You can answer basic or ad hoc questions but only after scrambling to pull the data together. Numbers may not match in next fire drill.

Level 3: Data on Demand

Your existing data and measures are available on demand and mandatory reports are automated and return trusted numbers.

Step 2: Use data in decision-making

Define your business need and select the right data approach

Performance Management

Suite of tools for selecting, developing, and managing with metrics for existing programs, services, or contracts

Evaluation & Experiments

Suite of tools for assessing the impact of a program or service or testing a new program or service

Advanced Analytics

Suite of tools for exploring business questions, developing new insights, or developing new decision tools

Ongoing Exploratory Data Analysis and Data Development

Suite of activities to explore existing data to inform new efforts and to identify the need and plan for new datasets. This feeds all the other activities.

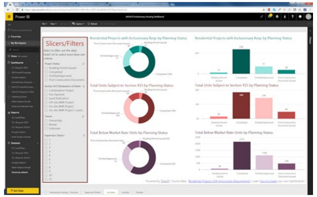
Data Science Accelerator

The Truffle Pig Problem:

Identifying good data science projects is the single greatest barrier to adoption

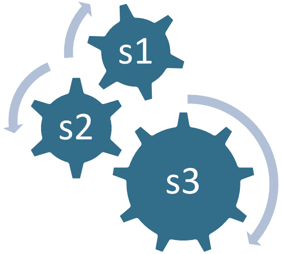
You know you have a truffle pig problem if…

Can I have a dashboard?


Power BI
Slicers/Filters
Residential Projects with Inclusionary Reqs by Planning Status
Total Units Subject to Section 415 by Planning Status
Total Below Market Rate Units by Planning Status

Can you build a warehouse?

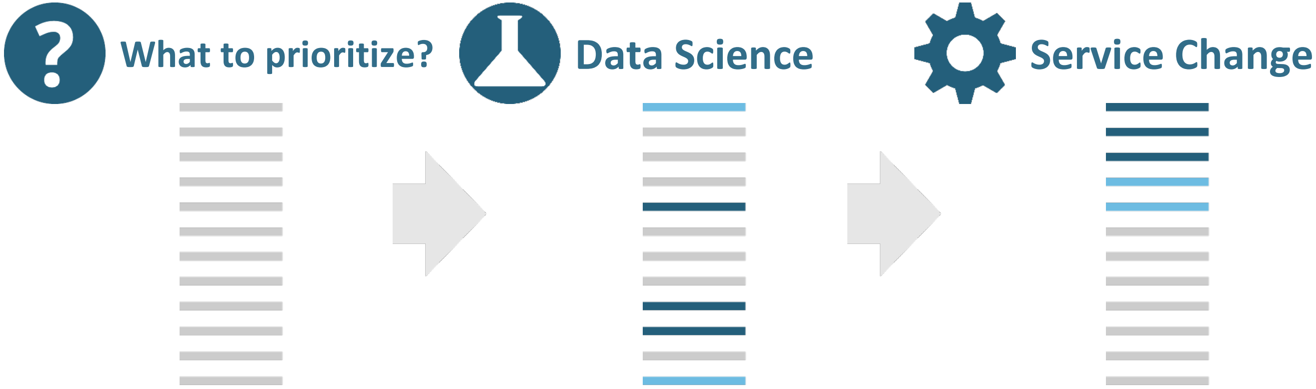
Can you automate this process?



Solution: The project typology

- Find the needle in the haystack
- Prioritize your backlog
- Flag “stuff” early
- AB test something
- Optimize your resources
- Some combination
- Something else…

How it works: “Prioritize your backlog”



What to prioritize?

Data Science

Service Change

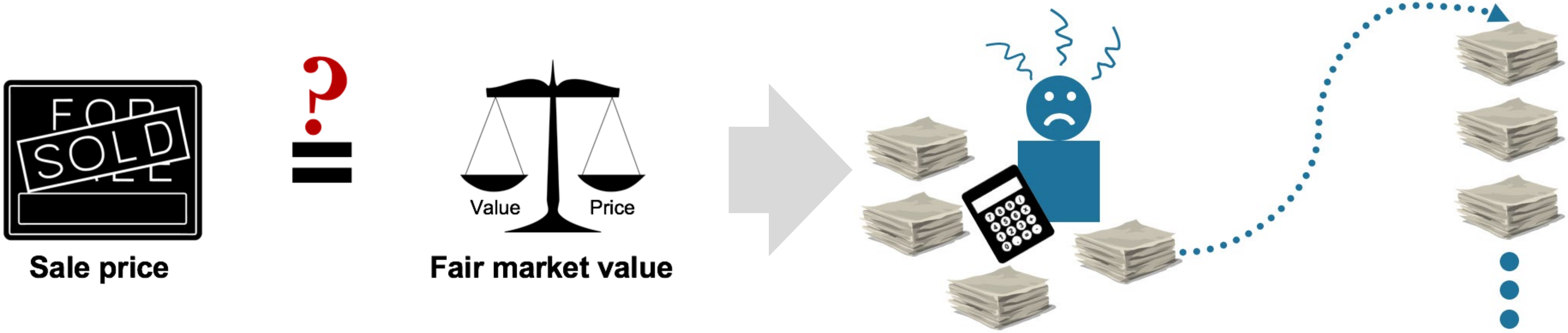
Service Issue:
Backlog is tackled via first in, first out (FIFO)

Data Science Process:
Create a model to categorize and group past and current cases

Service Change:
Prioritize cases based on categories in order of risk, need or opportunity

Result: Department addresses high priority cases first

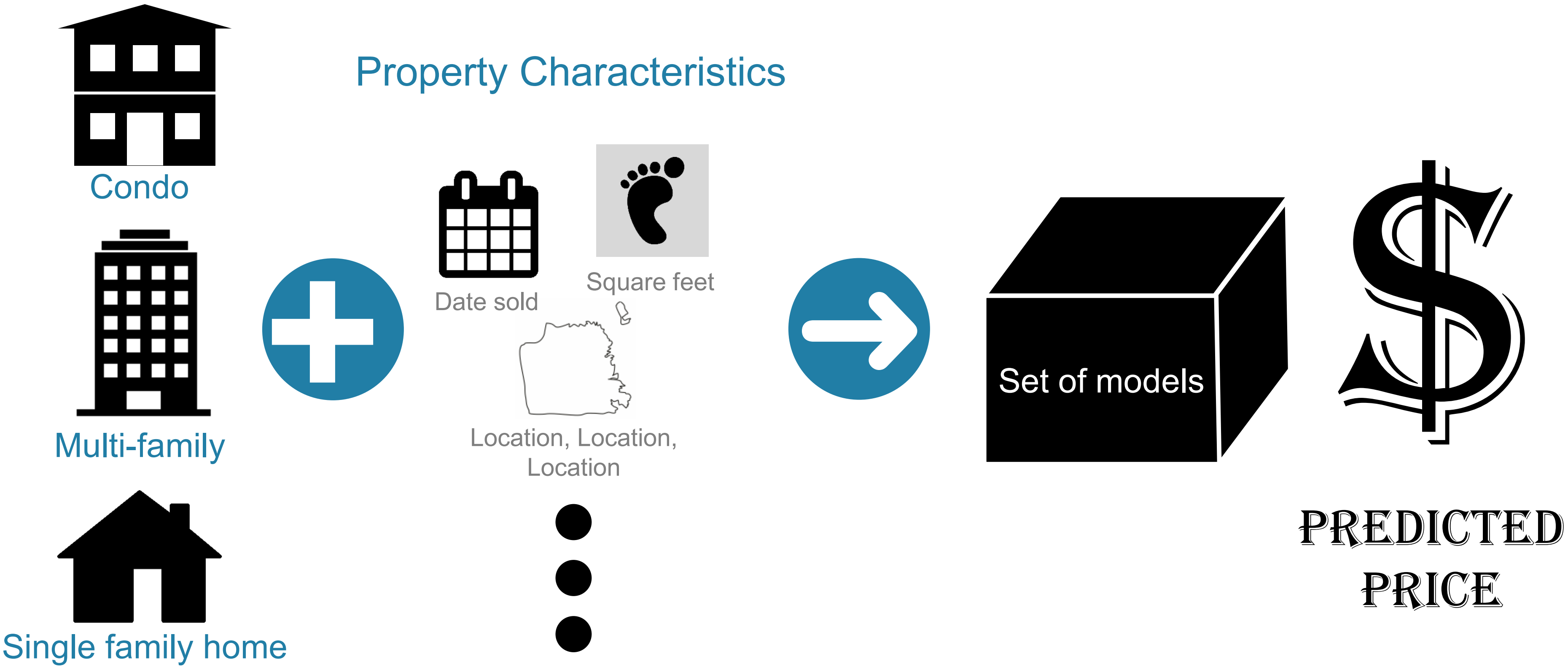
Service Issue: How to better process a giant backlog of property sales?



Sale price

Fair market value

Data Science



Property Characteristics

Condo

Multi-family

Single family home

Date sold

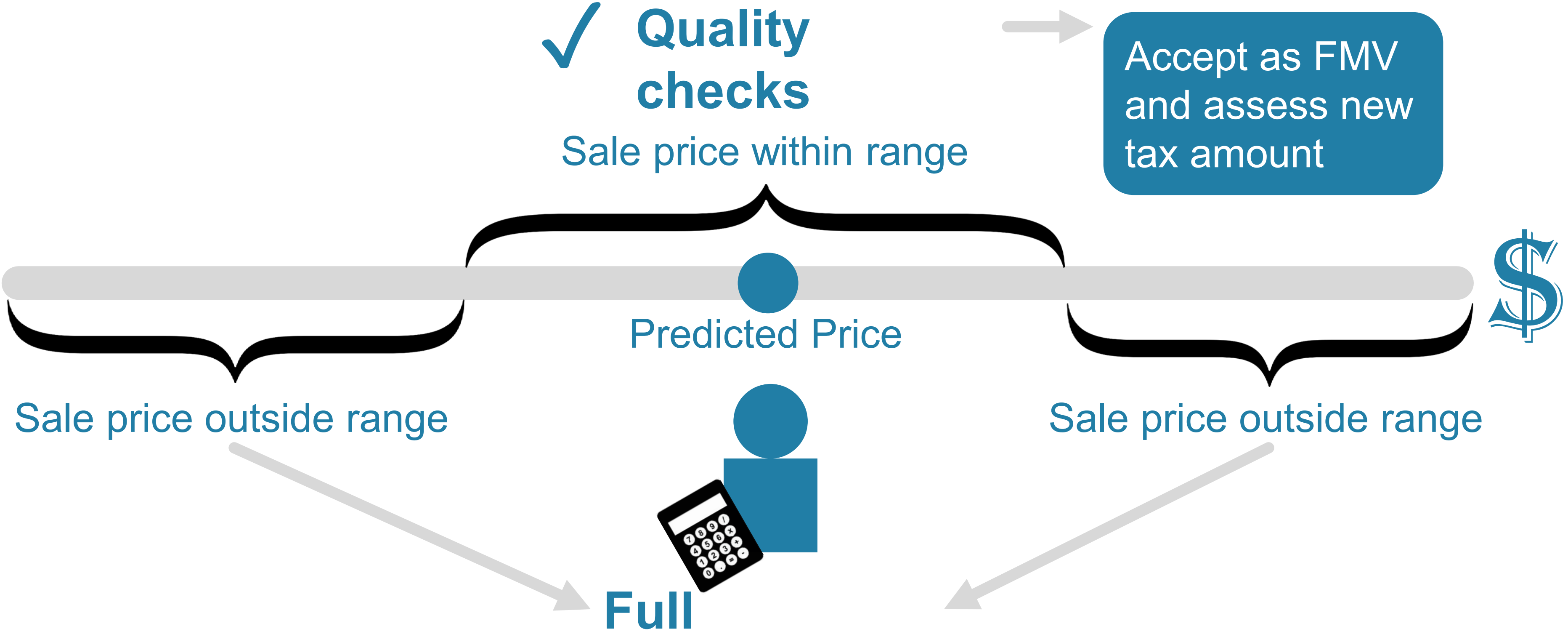
Square feet

Location, Location, Location

Set of models

PREDICTED PRICE

Service Change



✓ Quality checks

Sale price within range

Accept as FMV and assess new tax amount

Predicted Price

Sale price outside range

Sale price outside range

Full

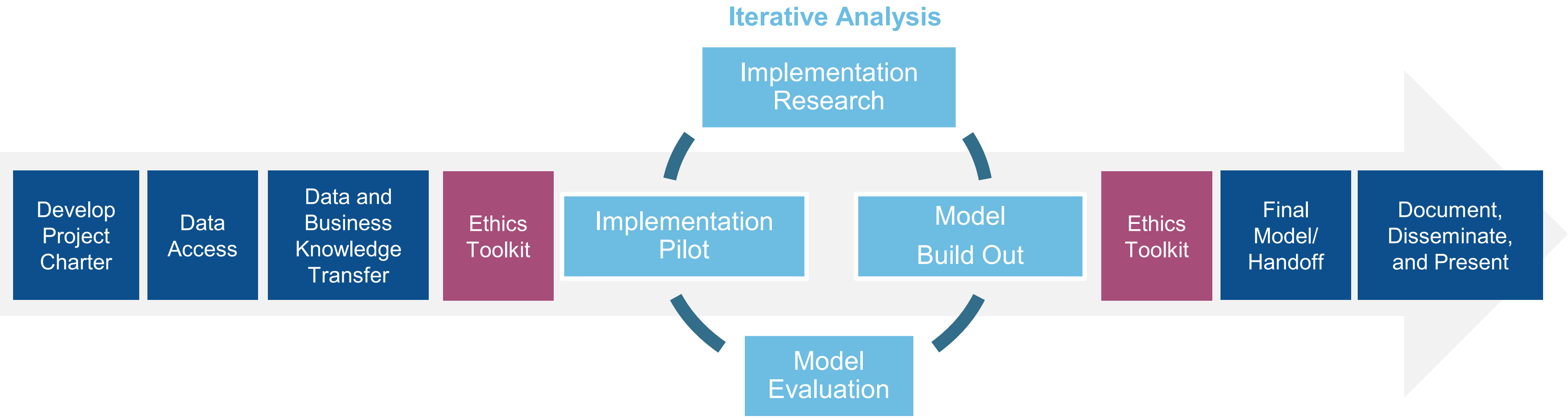
Results

1st model run reduced backlog 10%:

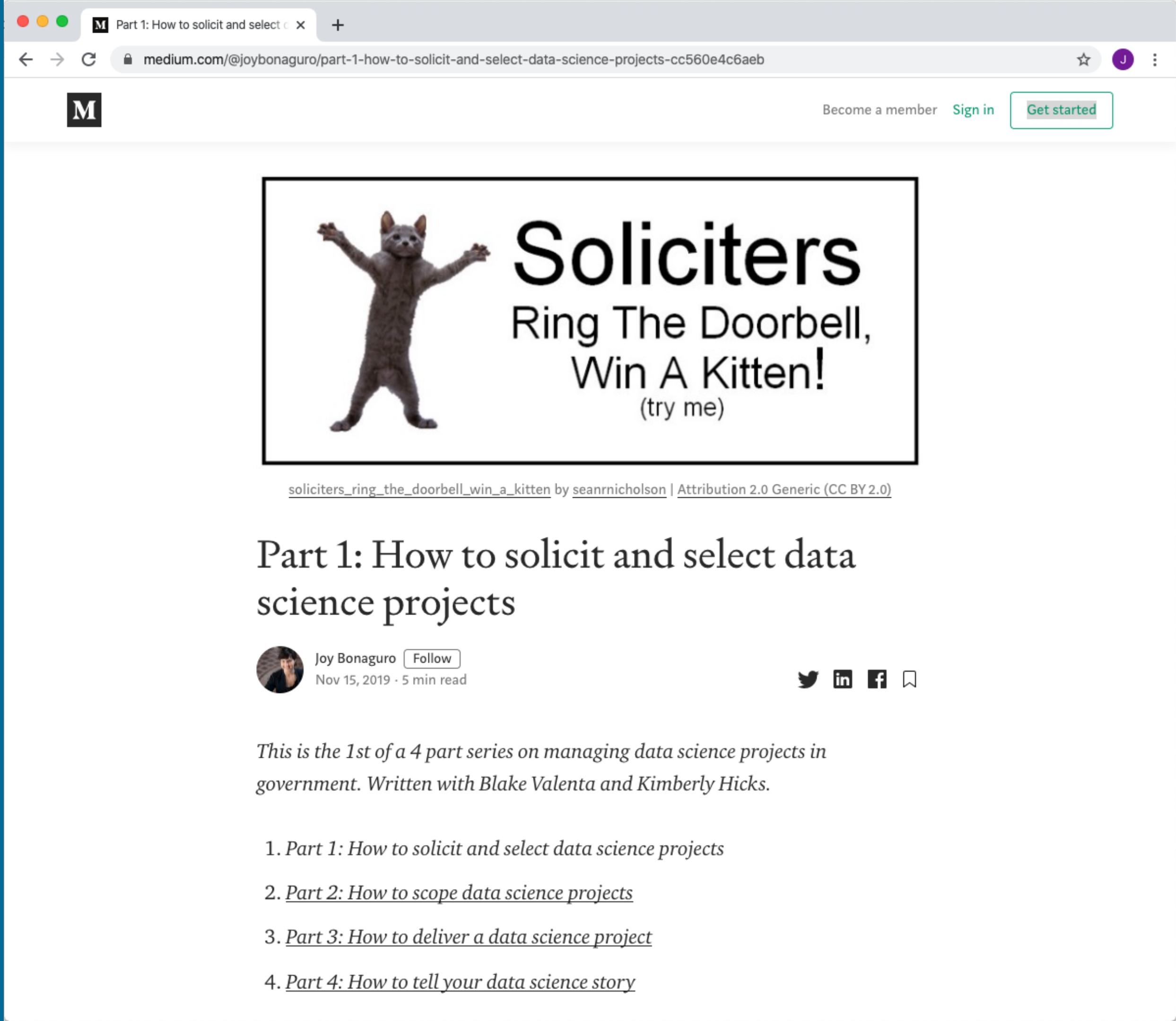
$239 M in roll value → **$2.8 M** in tax revenue

Ethics Toolkit

Ensuring responsible use of algorithms



- Business centered NOT data or technology centered
- Developed business / program muscle to opportunity spot


Part 1: How to solicit and select d
medium.com/@joybonaguro/part-1-how-to-solicit-and-select-data-science-projects-cc560e4c6aeb
M
Become a member
Sign in
Get started
Soliciters
Ring The Doorbell,
Win A Kitten!
(try me)
Follow

solicitors_ring_the_doorbell_win_a_kitten by seanrnicholson | Attribution 2.0 Generic (CC BY 2.0)

Part 1: How to solicit and select data science projects

Joy Bonaguro
Nov 15, 2019 · 5 min read

This is the 1st of a 4 part series on managing data science projects in government. Written with Blake Valenta and Kimberly Hicks.

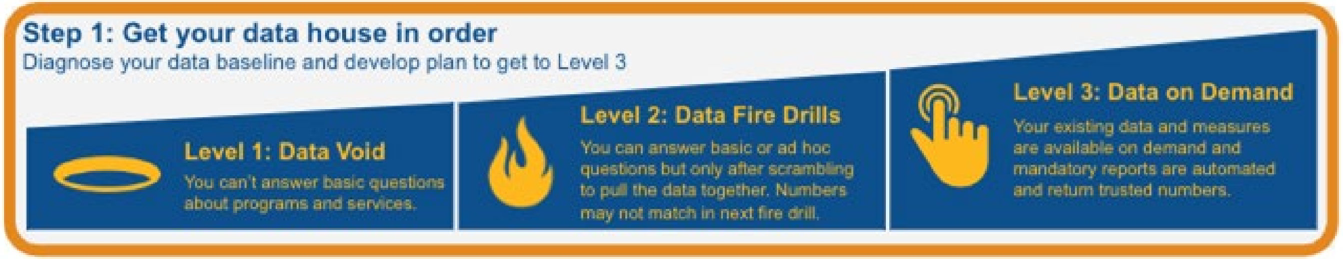
1. *Part 1: How to solicit and select data science projects*
2. *Part 2: How to scope data science projects*
3. *Part 3: How to deliver a data science project*
4. *Part 4: How to tell your data science story*

Data Services

Empowering departments at every step

MDSA

AA



Step 1: Get your data house in order

Diagnose your data baseline and develop plan to get to Level 3

Level 1: Data Void

You can't answer basic questions about programs and services.

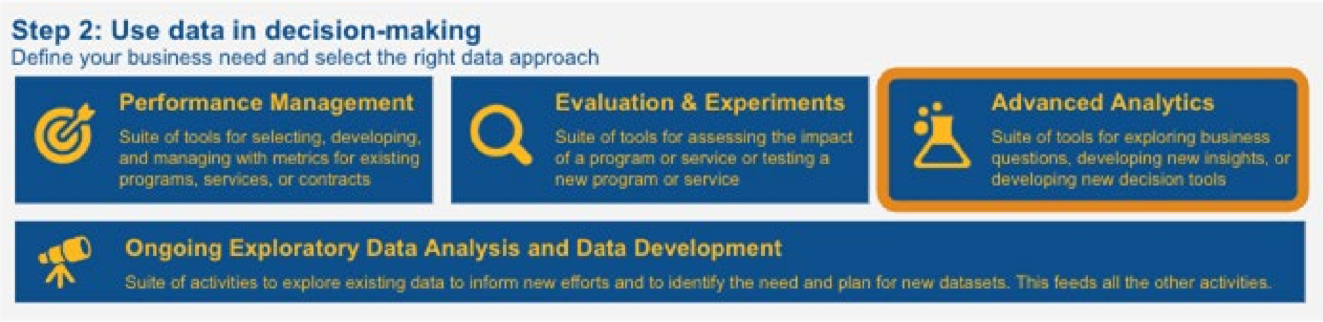
Level 2: Data Fire Drills

You can answer basic or ad hoc questions but only after scrambling to pull the data together. Numbers may not match in next fire drill.

Level 3: Data on Demand

Your existing data and measures are available on demand and mandatory reports are automated and return trusted numbers.

DSA



Step 2: Use data in decision-making

Define your business need and select the right data approach

Performance Management

Suite of tools for selecting, developing, and managing with metrics for existing programs, services, or contracts

Evaluation & Experiments

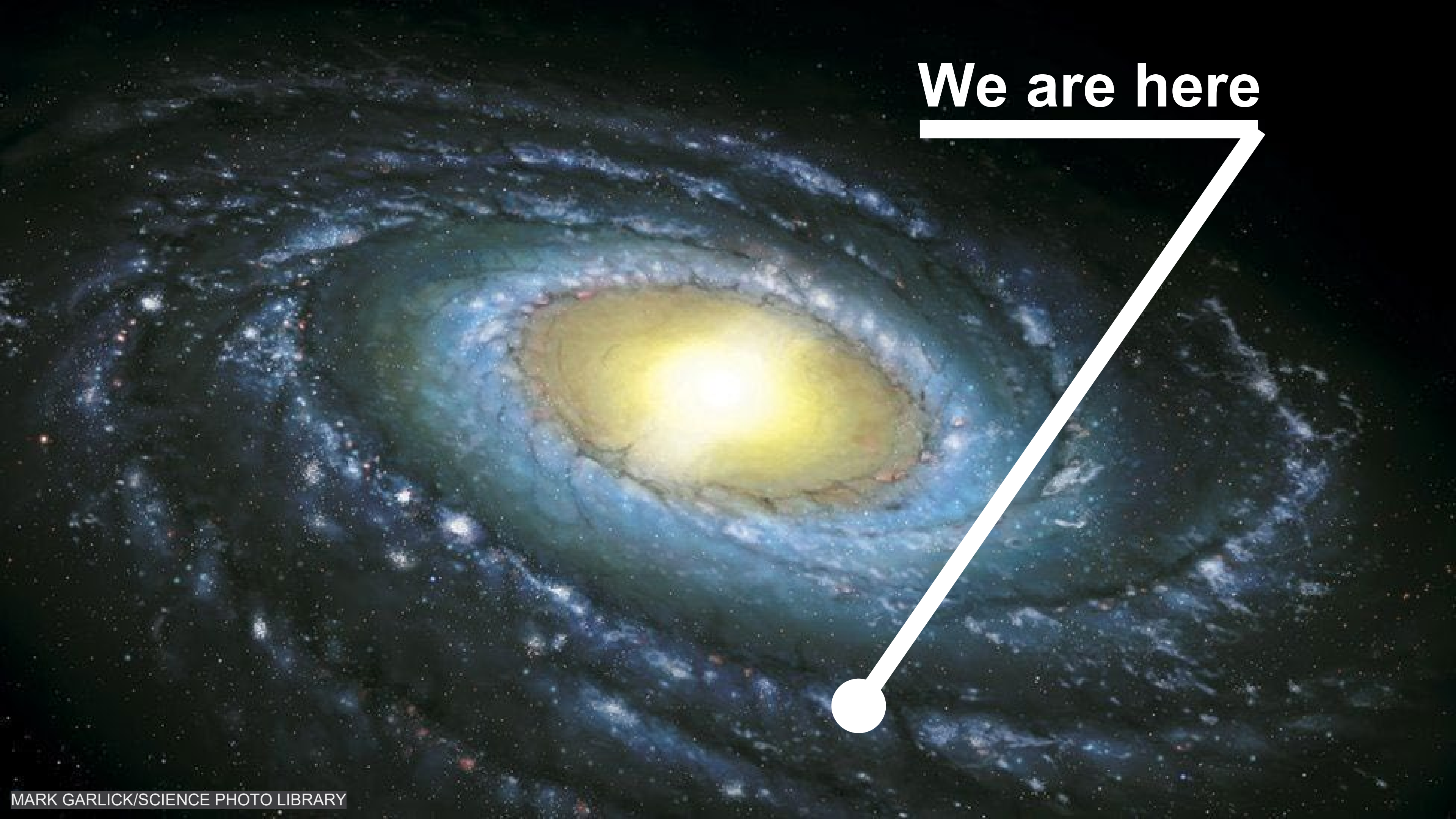
Suite of tools for assessing the impact of a program or service or testing a new program or service

Advanced Analytics

Suite of tools for exploring business questions, developing new insights, or developing new decision tools

Ongoing Exploratory Data Analysis and Data Development

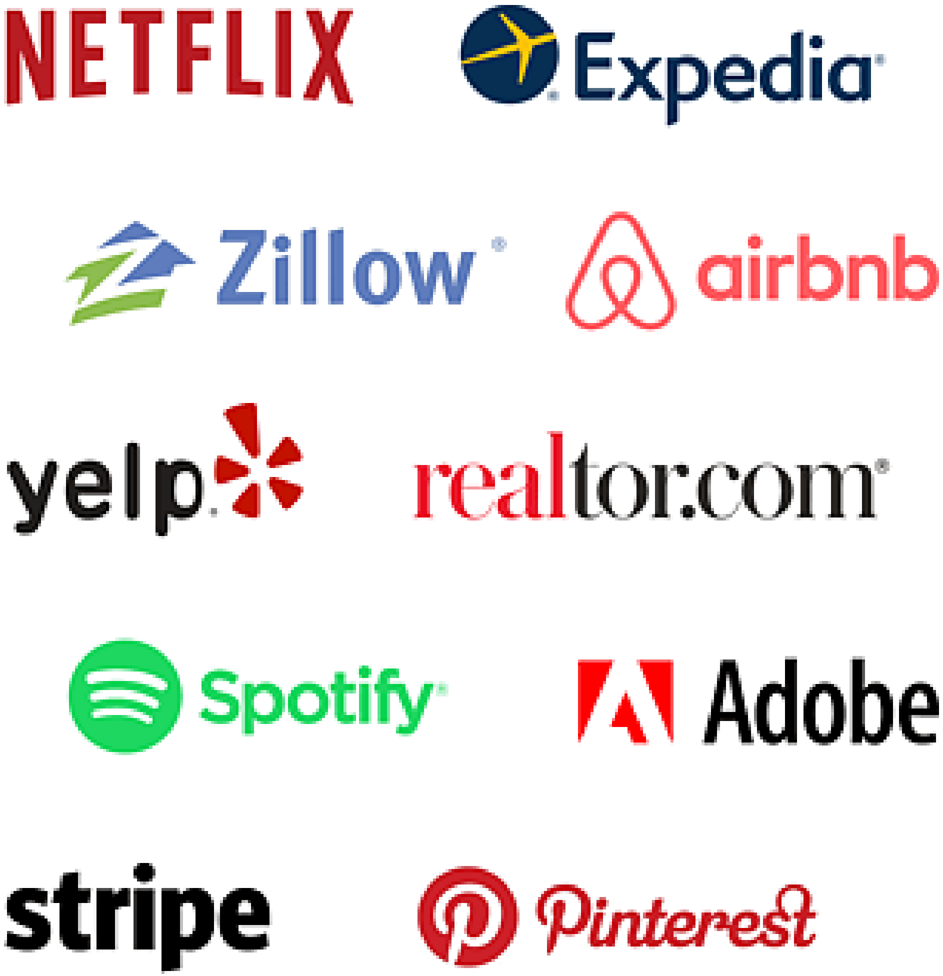
Suite of activities to explore existing data to inform new efforts and to identify the need and plan for new datasets. This feeds all the other activities.



We are here

MARK GARLICK/SCIENCE PHOTO LIBRARY

What do these companies have in common?



We rebuild the same stuff over and over again...with variable quality

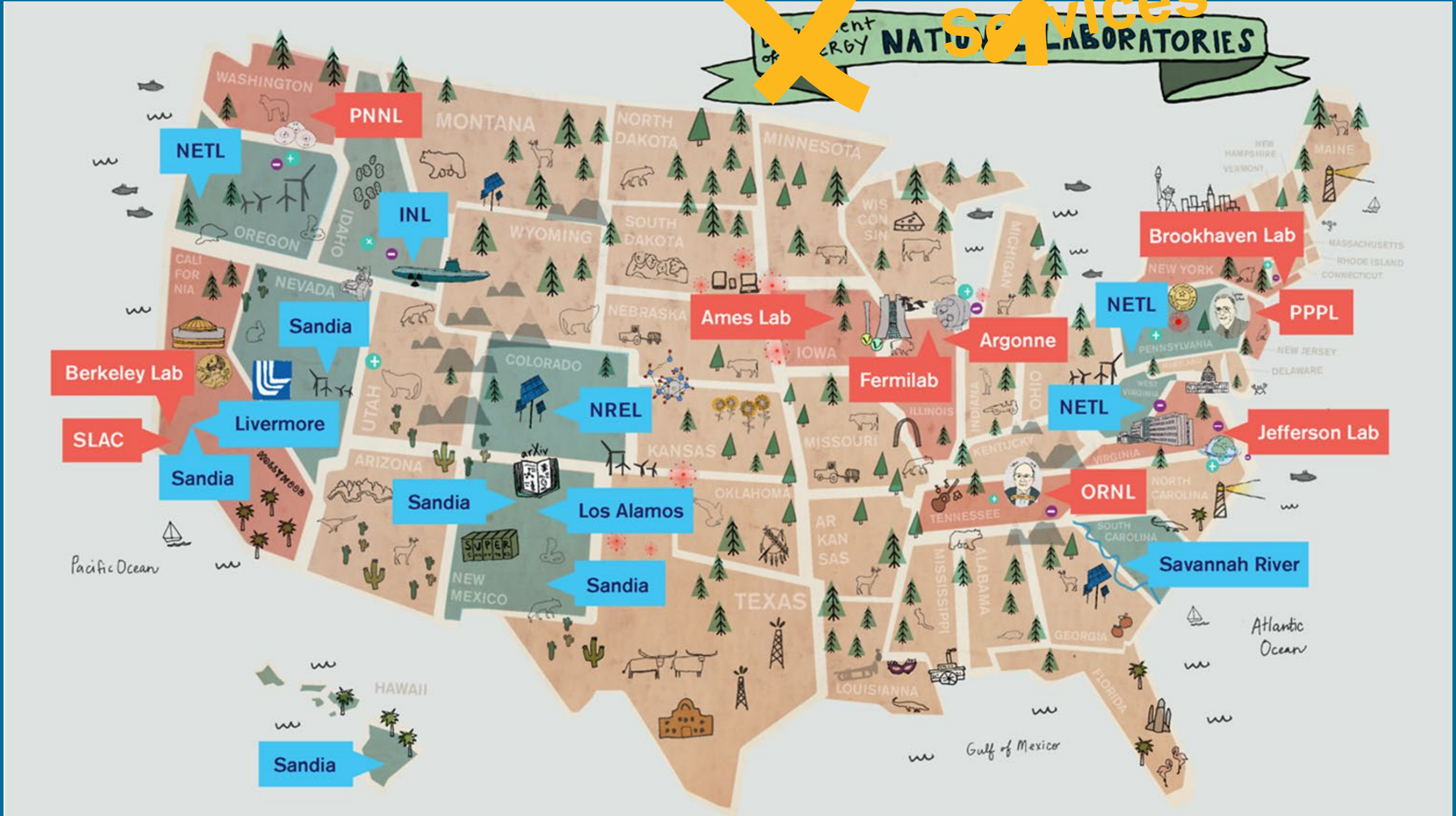
Why do we keep repeating the same “undifferentiated heavy lifting”?

By domain

By geography

By level of gov’t

Gov't Services



Thank you!

Questions / feedback / feelings / reactions / thoughts Welcome!

www.innovation.ca.gov

@joybonaguro | medium.com/@joybonaguro