



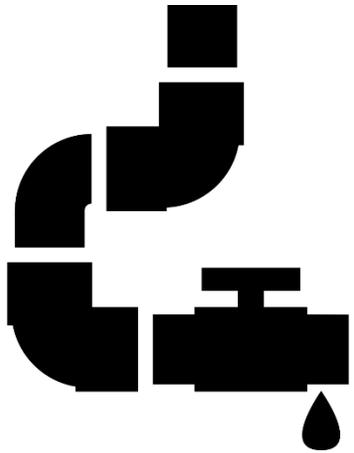
# Introducing Data Science

in Legacy Organizations

Joy Bonaguro | Statewide Chief Data Officer |  
State of CA



Objective: Demonstrate power of data science and eventually embed in culture and fabric



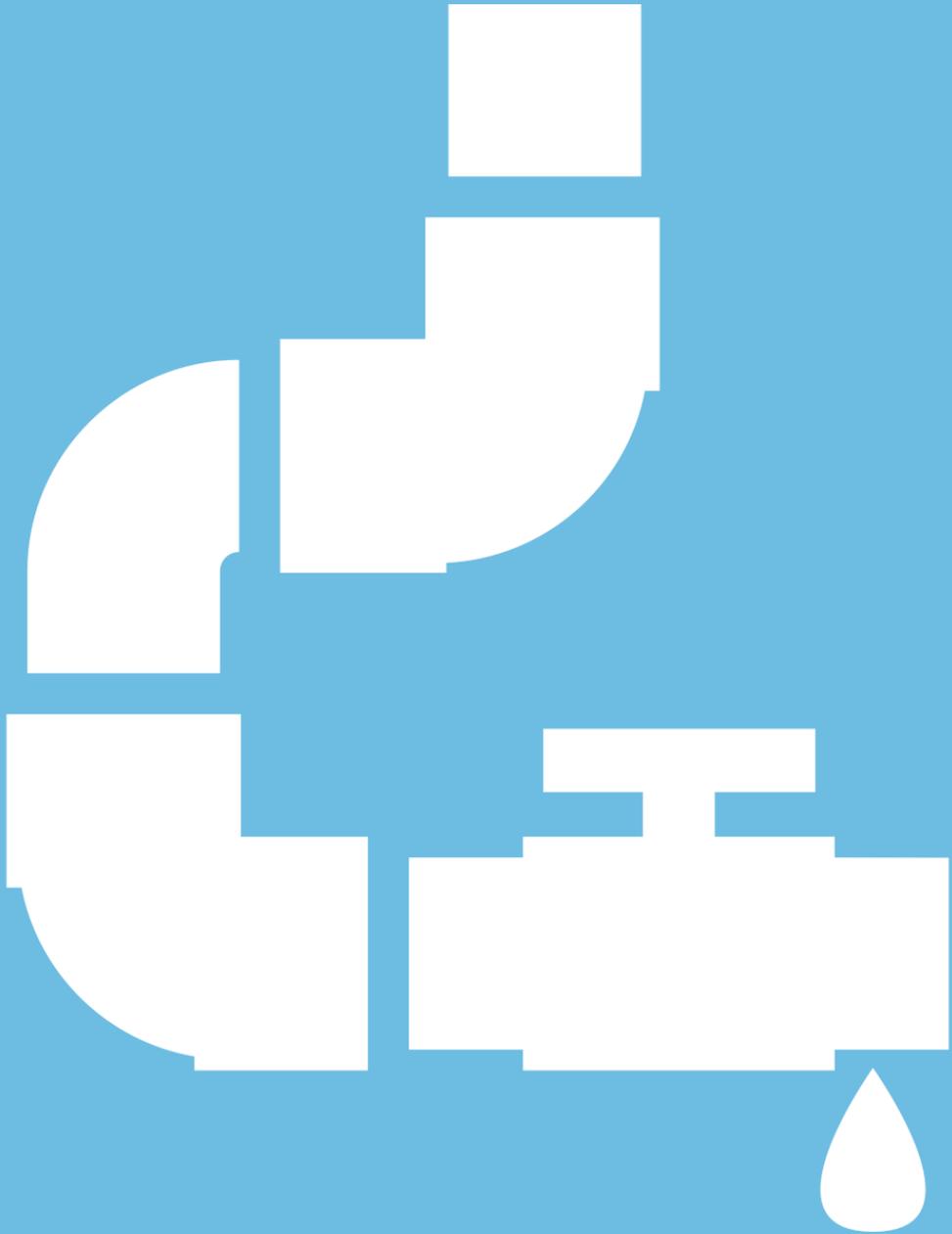
Develop



Deliver



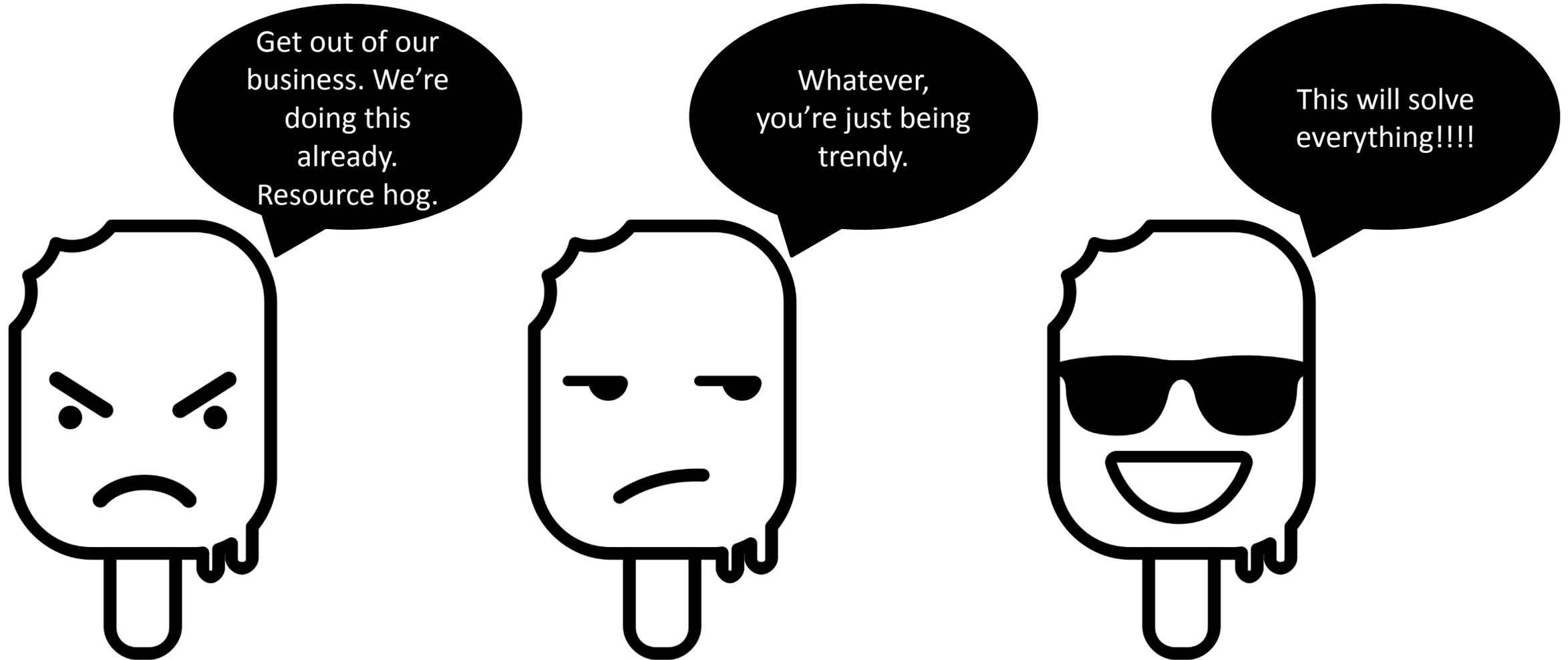
Celebrate



# Developing

**your data science pipeline**

# Introducing data science can evoke a range of reactions



# What is data science?



## Data Science

Applying advanced statistical tools to existing data to generate new insights



## Service Change

Converting new data insights into (often small) changes to business processes



## Smarter Work

More efficient and effective use of staff and resources

# What's in the data science toolkit?

## Statistical Methods

Sentiment analysis

Multilevel modeling

Survival analysis

AB testing

Propensity score matching

## Tools

Time series analysis

Missing data imputations

Pattern recognition

Machine learning

Logistic, multinomial and multiple linear regression techniques

## User Experience Research

Data mining

Classification and clustering

Principal component and factor analysis

Forecasting

Network analysis

# What's in the data science toolkit?

Statistical Methods

Tools

User Experience Research

## Languages

Python  
R  
SQL  
Javascript  
NodeJS

## Libraries

SciPy  
Pandas  
Scikit-learn  
GPText  
OpenNLP  
Mahout  
+many others

## Data Engineering

Profiling  
ETL  
Job notices  
APIs  
Optimized data  
pipelines  
Optimized data  
storage/access

## Visualization

D3.js  
Gephi  
R  
Leaflet  
PowerBI  
ggplot2  
shiny

# What's in the data science toolkit?

Statistical Methods

Tools

User Experience Research

Iterative  
Prototyping

Photo journaling and  
documenting

Service blueprinting

Journey mapping

Ride-alongs

Process mapping

Ethnographic field research  
and user observation

Usability testing

# What is **NOT** data science?



Service change



Small changes

Created by Daniil Polshin  
from Noun Project



Use existing data

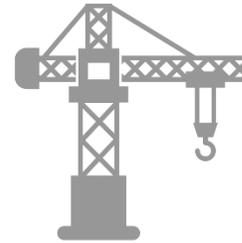
Created by Arthur Shlain  
from Noun Project

✗ Not



Academic research

Created by Rockicon  
from Noun Project



Major overhauls /  
service disruptions

Created by Hopkins  
from Noun Project



Collecting new  
data (mostly ;)

Created by Chameleon Design  
from Noun Project



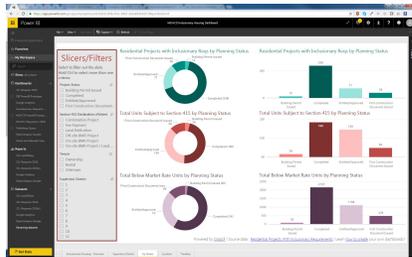
# The Truffle Pig

## Problem:

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

# You know you have a truffle pig problem if...

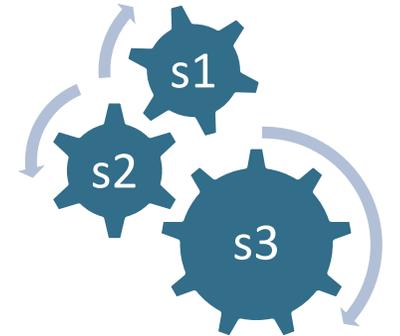
Can I have a dashboard?



Can you build a warehouse?



Can you automate this process?



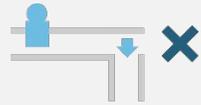
# Solution: The project typology



Find the needle in the haystack



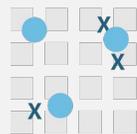
Prioritize your backlog



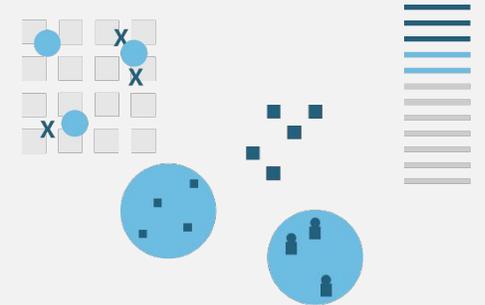
Flag "stuff" early



AB test something



Optimize your resources



Some combination



Something else...

# How it works: “Prioritize your backlog”



**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: Processing a giant backlog of property sales



**Sale price**

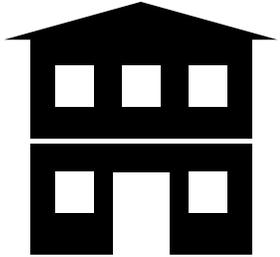


**Fair market value**

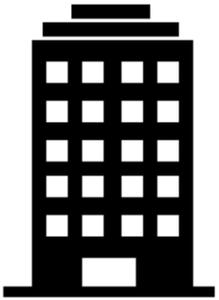




# Data Science



Condo

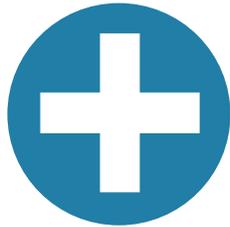


Multi-family



Single family home

## Property Characteristics



Date sold



Square feet



Location, Location,  
Location



**PREDICTED  
PRICE**





# Service Change

✓ **Quality checks** —

Accept as FMV  
and assess new  
tax amount

Sale price within range



Predicted Price



Sale price outside range

Sale price outside range



**Full appraisal**



## Results

1st model run reduced backlog 10%:

**\$239 M**

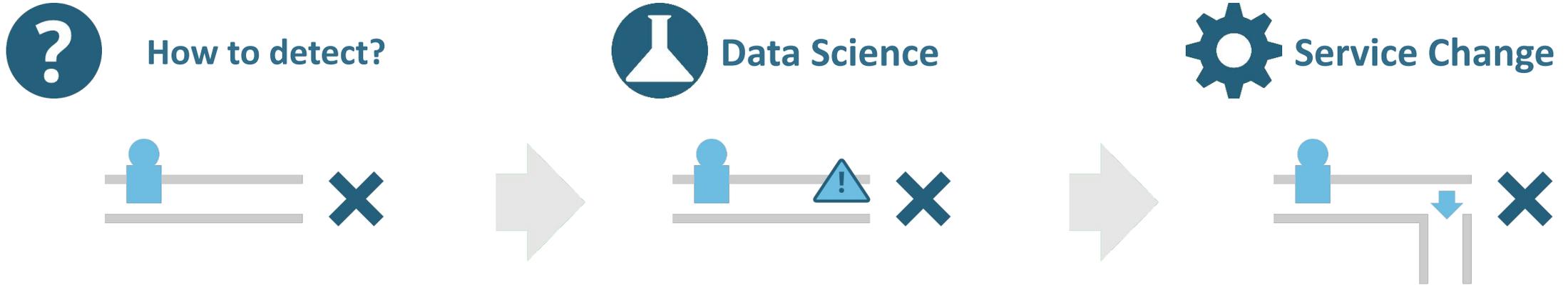
in roll value



**\$2.8 M**

in tax revenue

# Project Type: Flag “stuff” early



## Service Issue:

Hard to predict future condition which leads to reactive services

## Data Science Process:

Use historical and current data to create estimate ranges for potential outcomes

## Service Change:

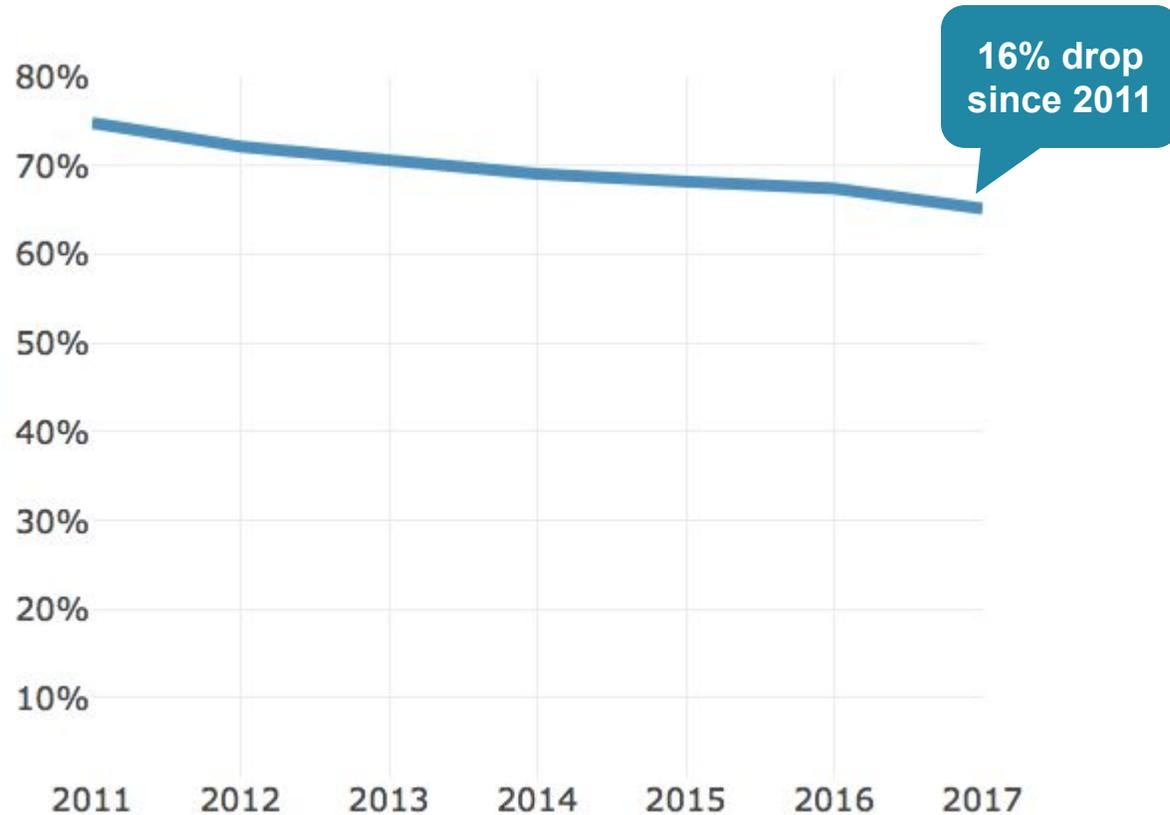
Use estimates to change and tailor intervention points

**Result:** Department provides pro-active early interventions

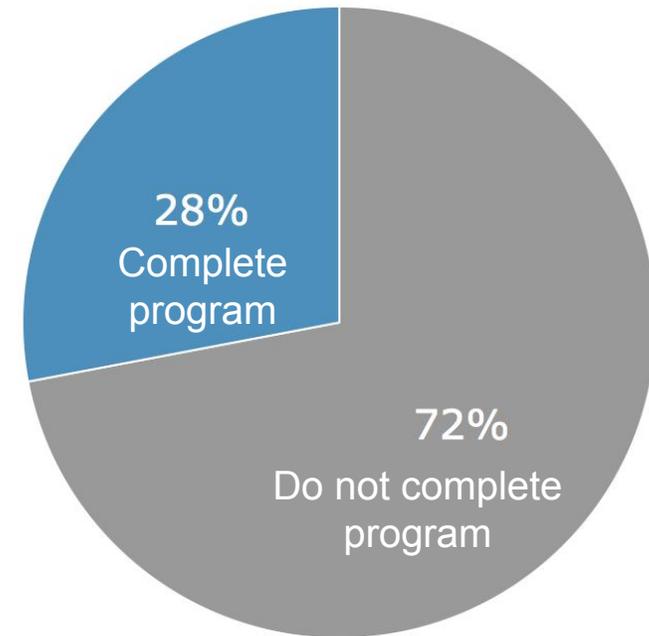


# Service Issue: Dropping out of WIC

% still in program at 1 year of age (retention)

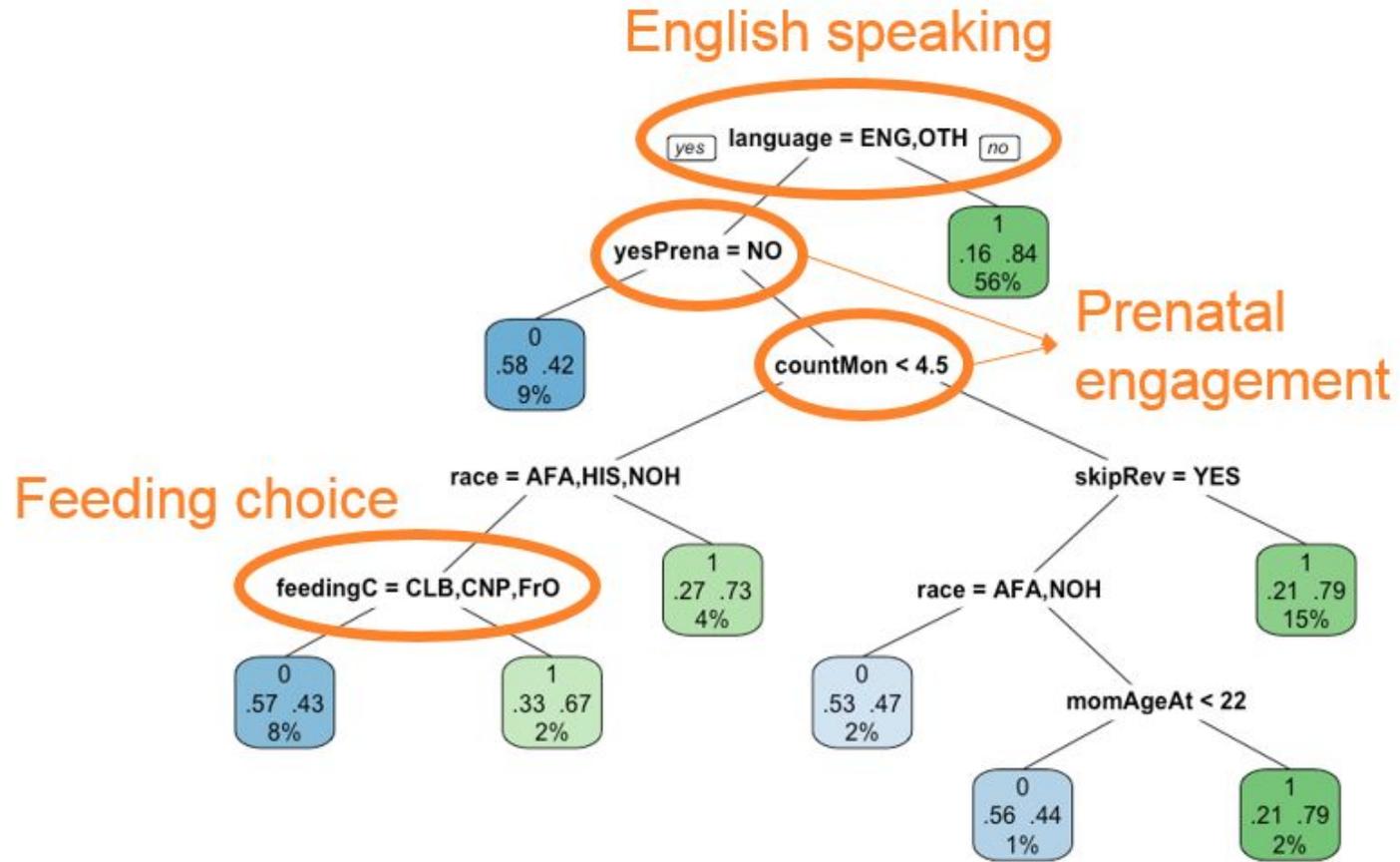


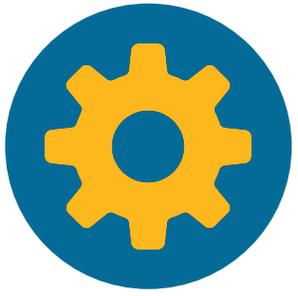
Completion status through age 5





# Data Science



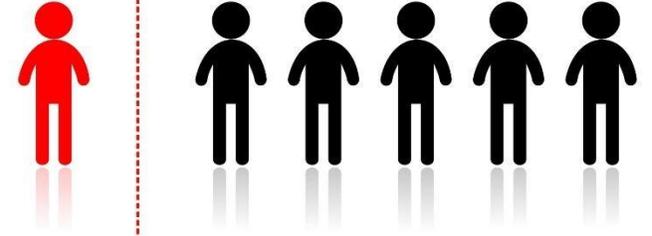


# Service Change

Qualitative



WHY



Anti-Stigma Campaign

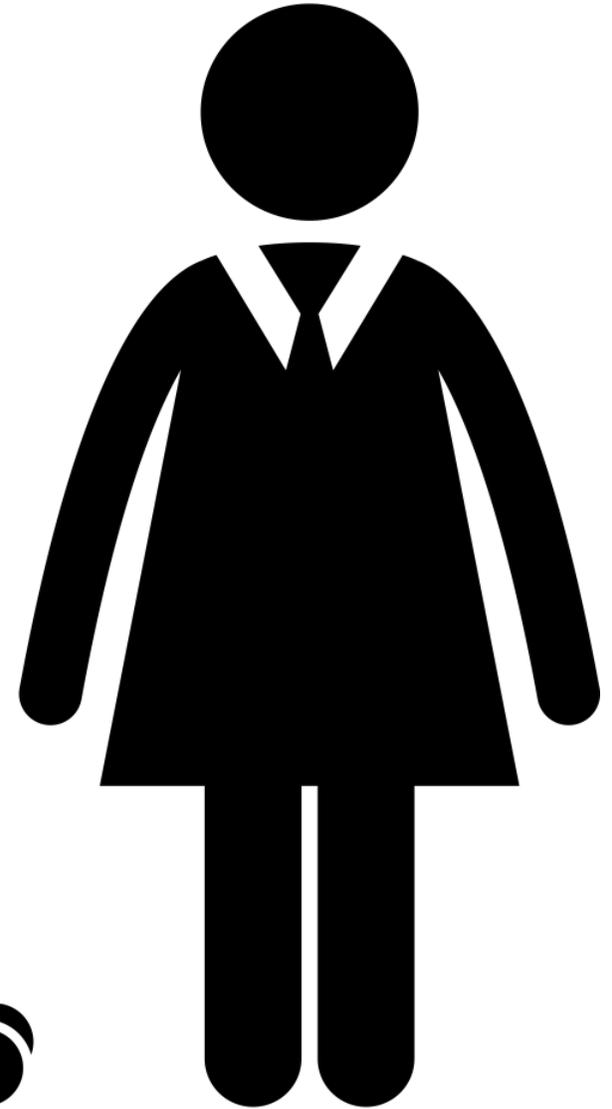


New Technology



Future Outreach Strategy

Please, please, please purty  
please do this project. We think it  
will be awesome!



# Invert the dynamic with an application process



DataScienceSF - Cohort 2

## DataScienceSF

Thank you for your interest in a pitching a project with DataScienceSF!

During cohort 2, we anticipate selecting 5-10 projects. If you are not selected, we will identify if we can help in other ways.

Before you apply, be sure to review our materials on what makes for a good data science project.

- [One page overview](#) of the DataScienceSF
- [Powerpoint deck](#) for DataScienceF
- [DataScienceSF website](#)

We will evaluate submissions based on the criteria outlined in the PowerPoint.

Please note the following:

- The goal of DataScienceSF is to use analytics to identify and implement a **service change** within your agency, not to generate a recommendation or report.
- DataScienceSF projects are designed to be completed and implemented within six months. Your department will present the results at the end of that time along with your cohort.

Submissions closed

[View a previous application or resume a draft](#)  
Submissions were due on Nov 22, 2017 at 7:00pm.

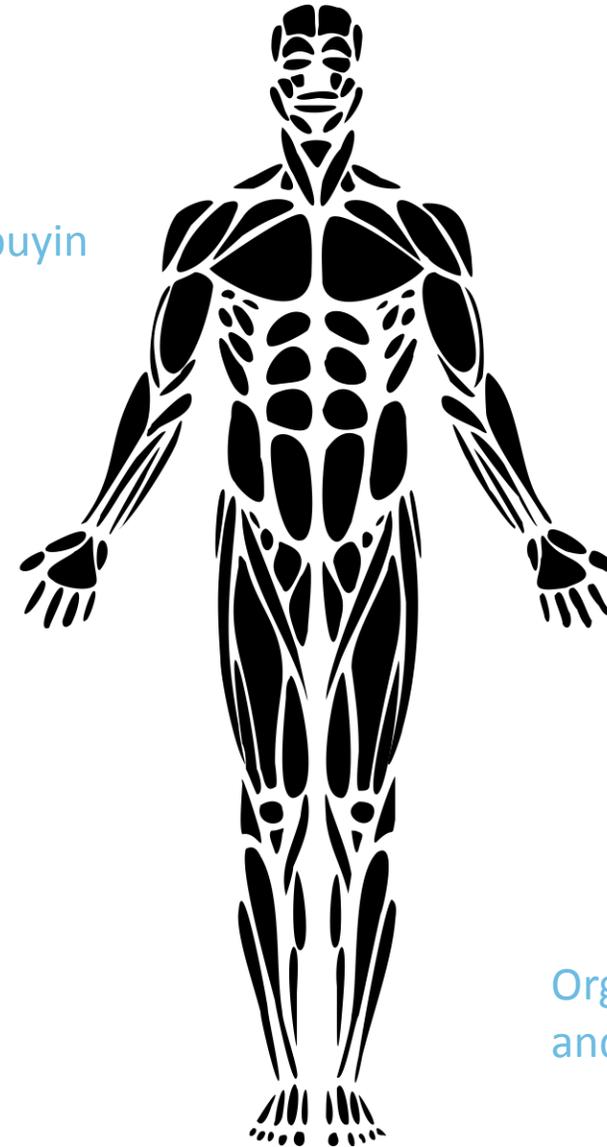
CONTACT EMAIL  
[blake\\_valenta@hks.harvard.edu](mailto:blake_valenta@hks.harvard.edu)

DEADLINE  
Nov 22, 2017 at 7:00pm  
In your local timezone (GMT-7)

SHARE THIS

f t in g+

# Anatomy of a committed client



Problem that data science can solve

Leadership buyin

Appropriate project champion

Leads to a service change, not just an academic exercise

Organizational value and alignment

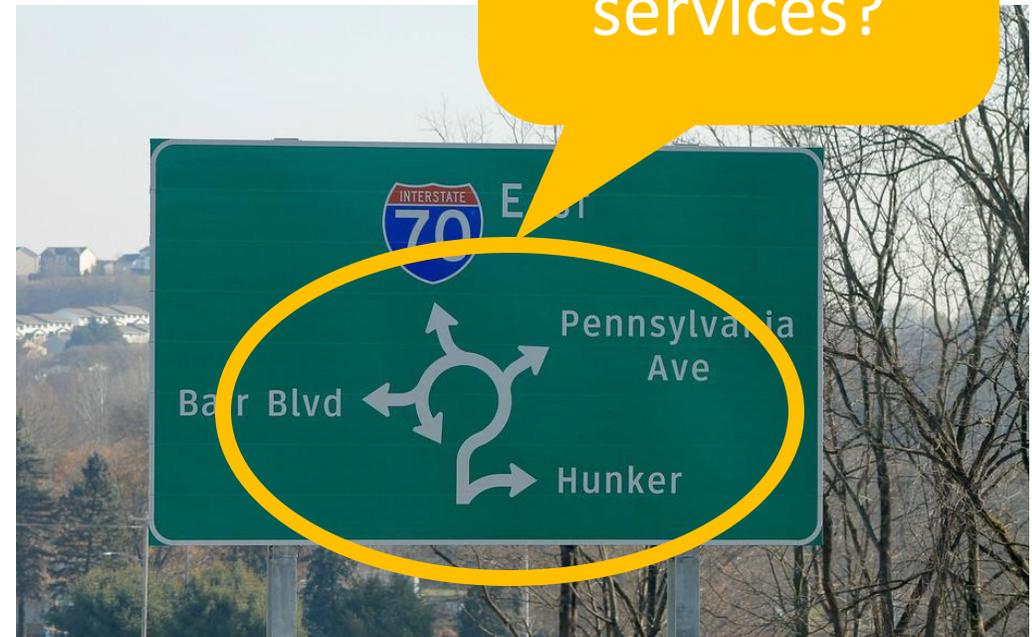


**Don't do this.**  
Sorry, can't help.

**Do this.**  
Provide gentle offramps.

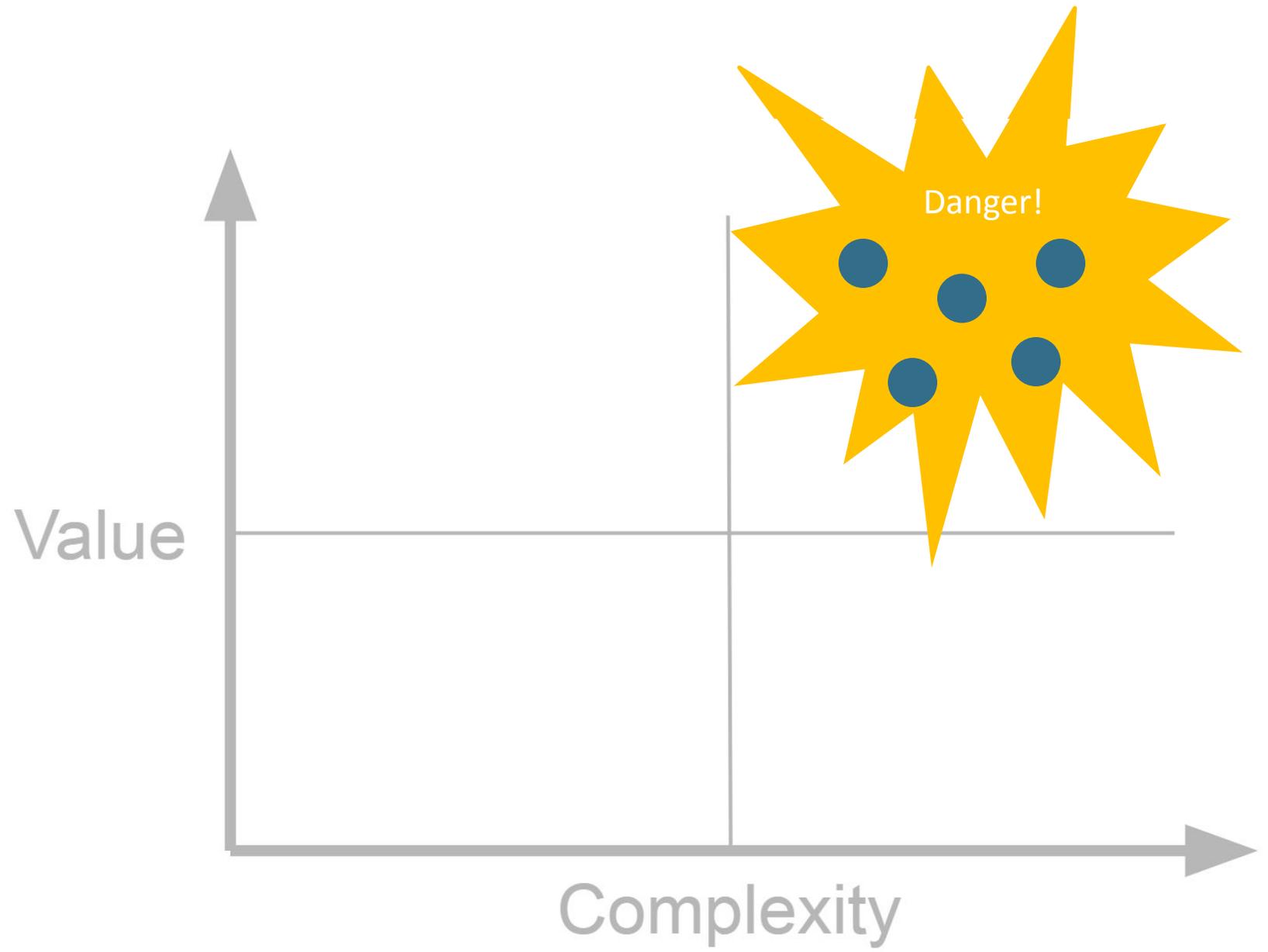


Tim Evanson, <https://www.flickr.com/photos/tim-evanson/10000000000/>

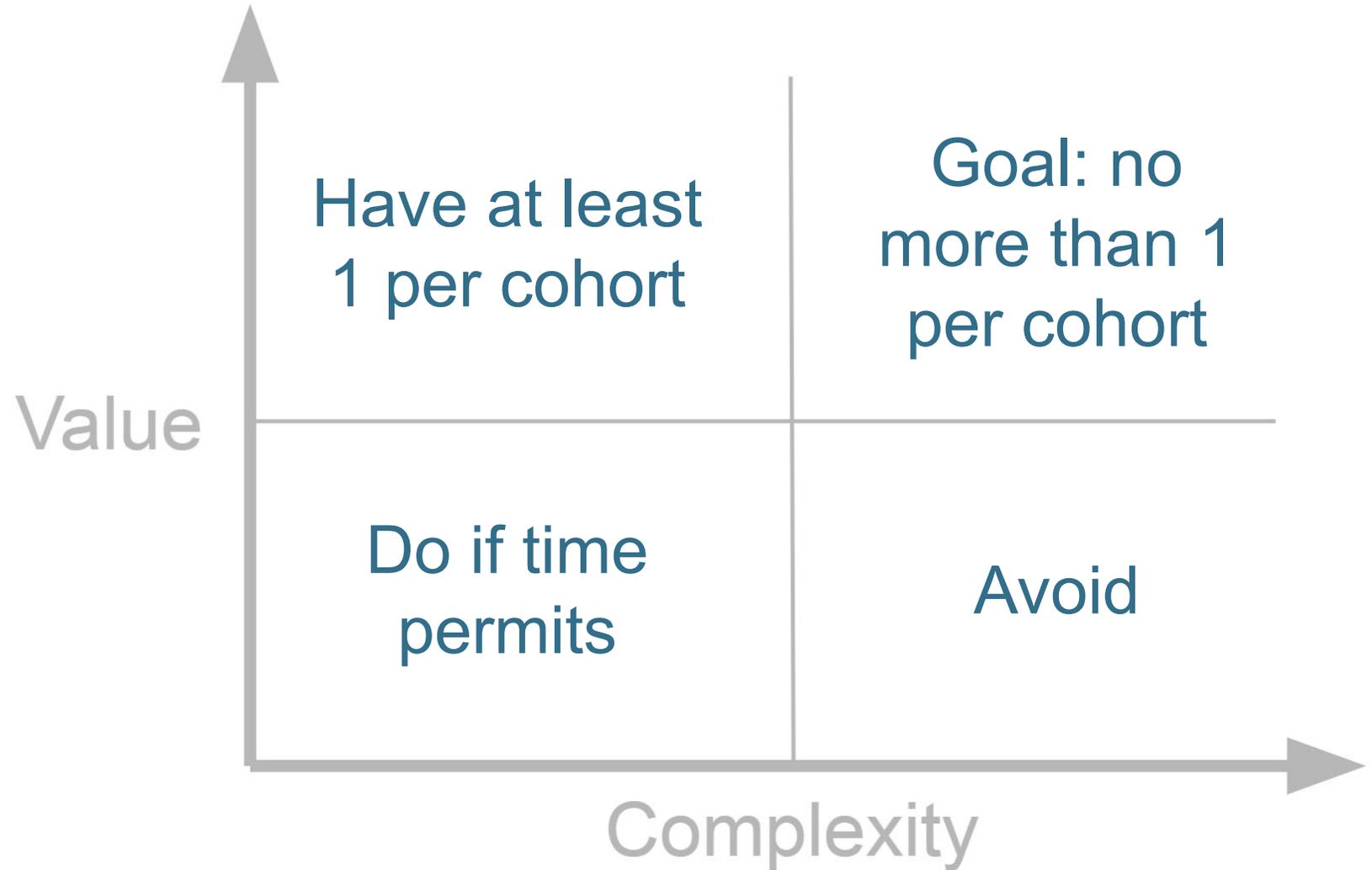


Jon Dawson, <https://www.flickr.com/photos/jmd41280/27040060139/>

**Score and  
balance your  
portfolio**



**Score and  
balance your  
portfolio**

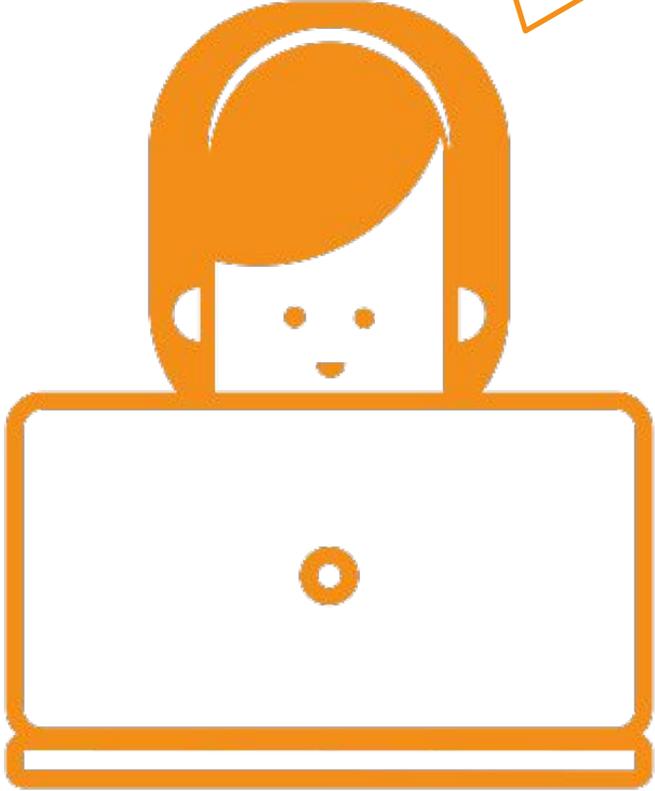




**Delivering**  
your data science pipeline

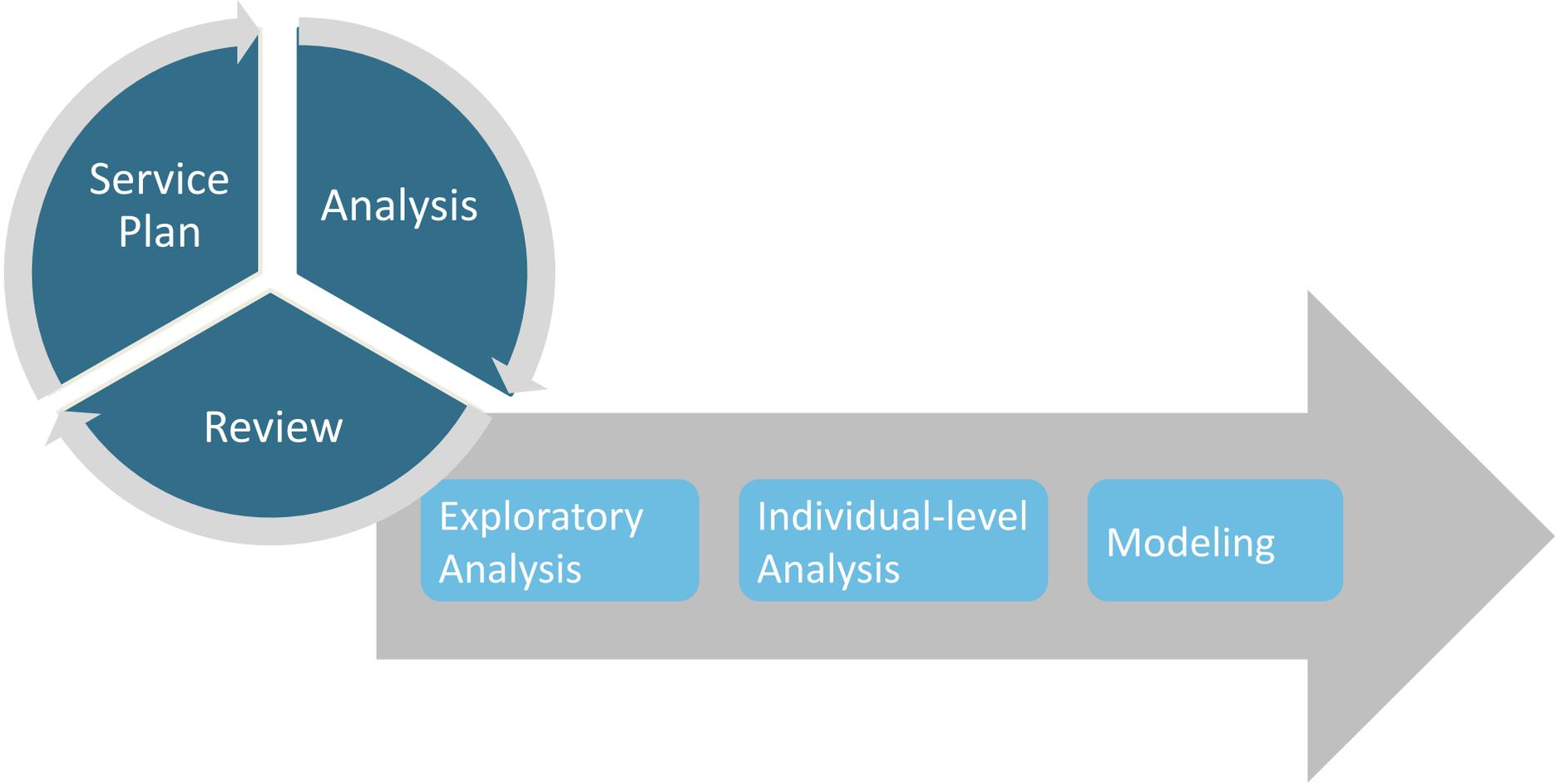


Hi! Haven't seen you in a while. Look what I did!

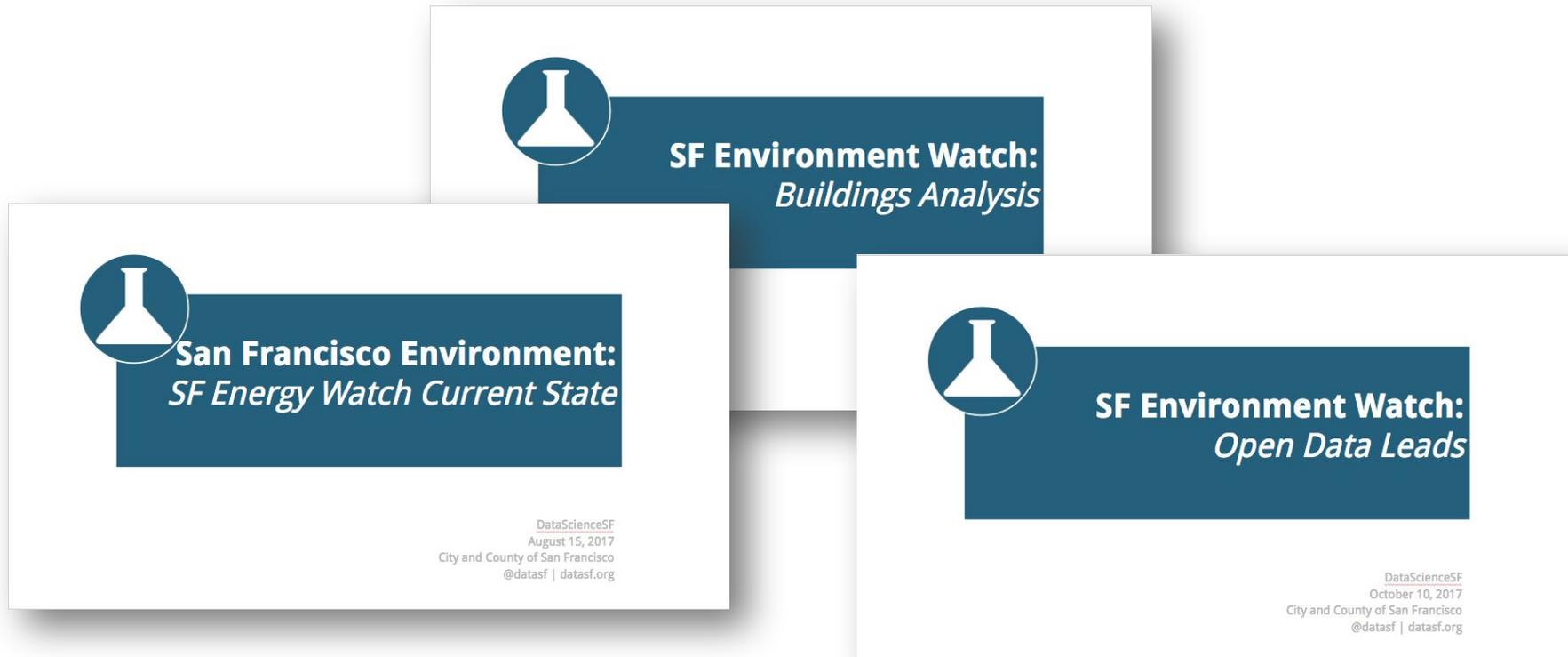


Hmmm...not what I was expecting...

# Bring your client with you: Communicate analysis phases and what to expect – don't jump to the model



Every meeting is your team delivering a product...  
...usually a digestible deck delivered at the meeting  
...weekly or biweekly (scheduled early)



# Budget time for user and context research and user testing...

...otherwise you may end up with something that doesn't solve their problem

...and sometimes the best model isn't implementable

...or you may deliver a product that is not usable

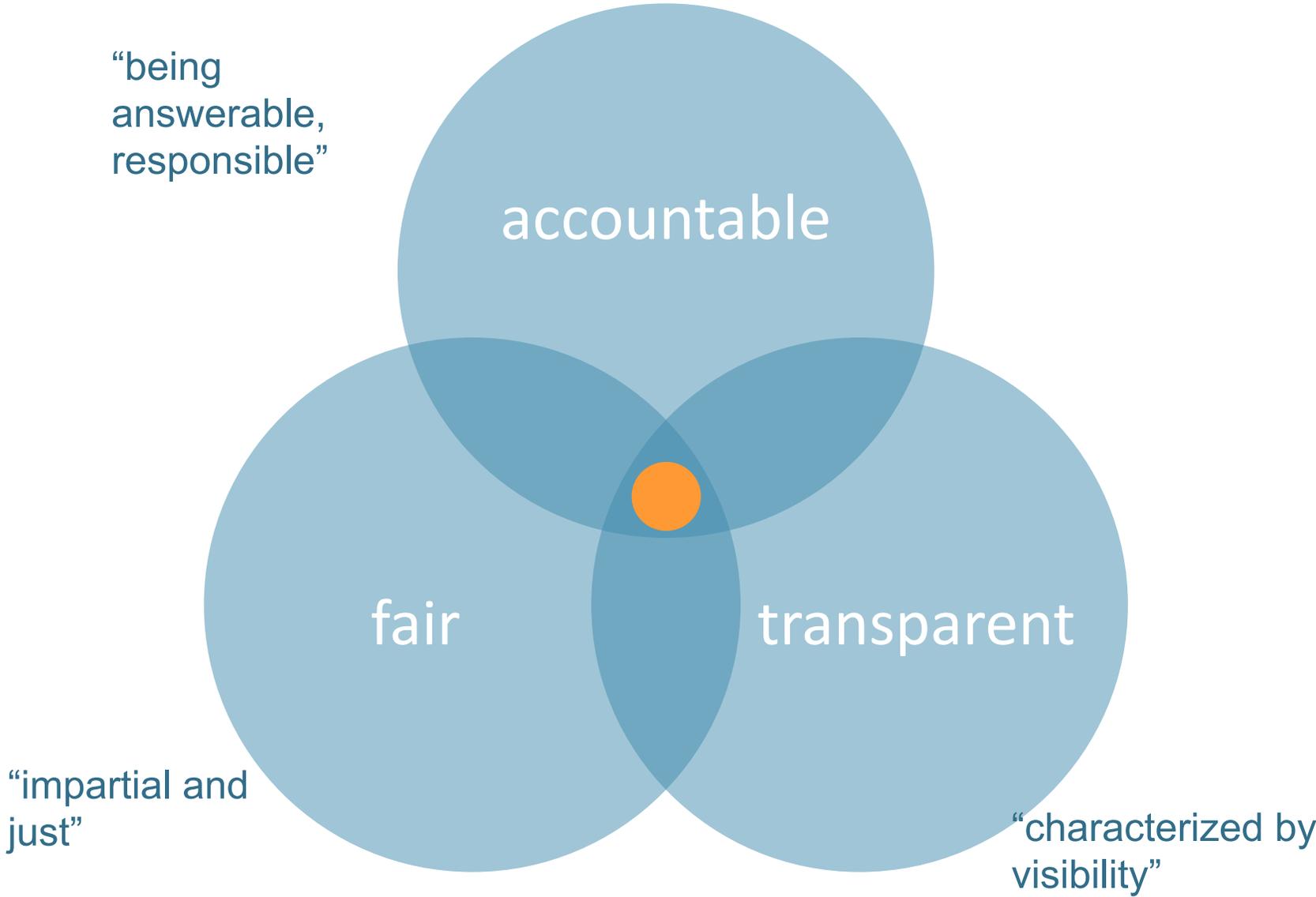
## Playbook: Context & User Research

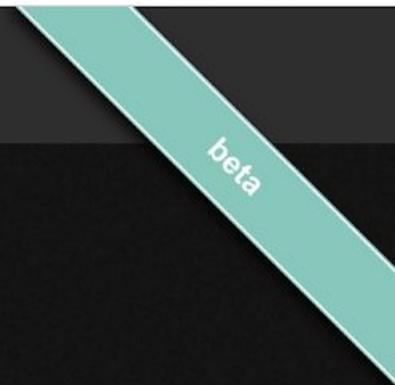
Playbook: Context & User Research	1
Purpose	2
Overview of steps	2
Step 1. Draft research plan	2
Step 2. Conduct secondary research	3
Step 3. Conduct user research	3
Step 3.1 Conduct research activity	3
Typical Research Activities	3
Who to research with	4
Step 3.2 Analyze	4
Step 3.3 Document	4
Artifacts	4
Writing user needs	4
Step 3.4 When to stop	5
Step 4. Share research	5
Appendices	5
Readings & Online Resources	5
Online	5
Books	5
UX Techniques: Table of strengths & Weaknesses	5
Stakeholders	7
Potential Stakeholders	7
Questions to ask	8

Squeeze, don't wring



# Ethics & Algorithms a practical toolkit



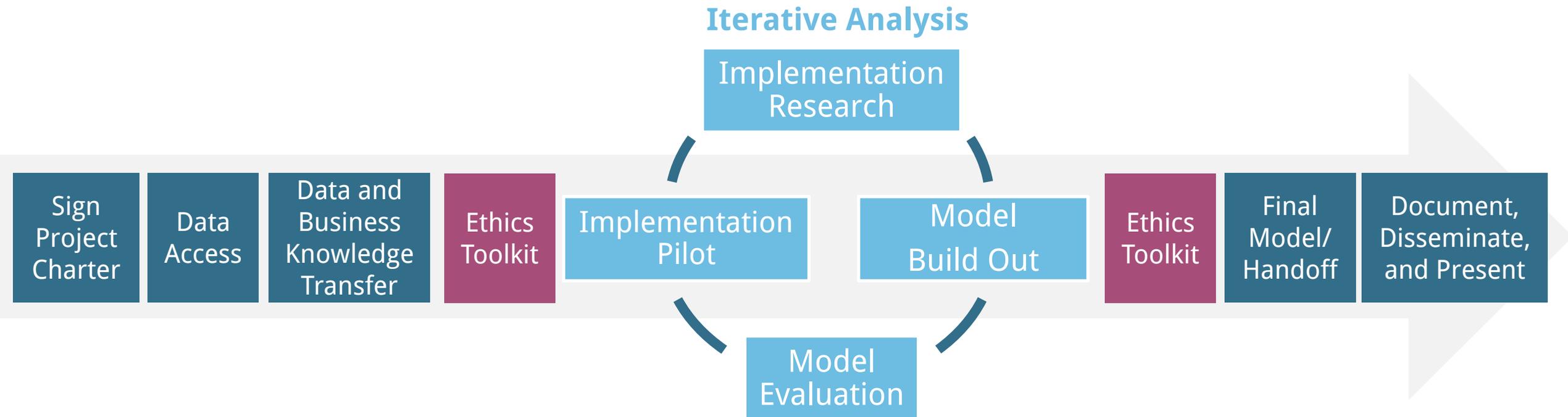


# Ethics & Algorithms Toolkit

A risk management framework for governments (and other people too!)



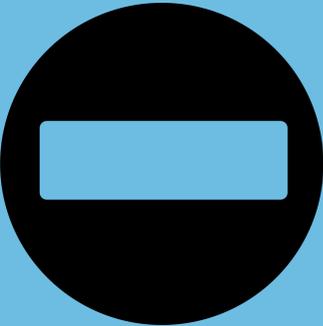
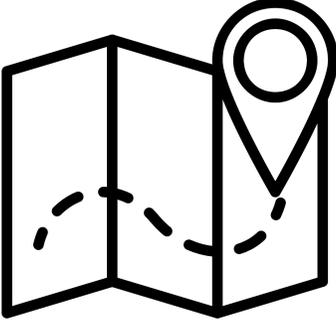
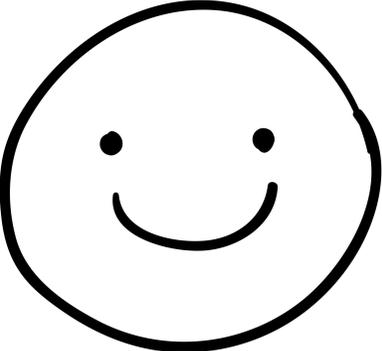
# Toolkit became a core part of the project cycle





**What is the scope of the impact?**

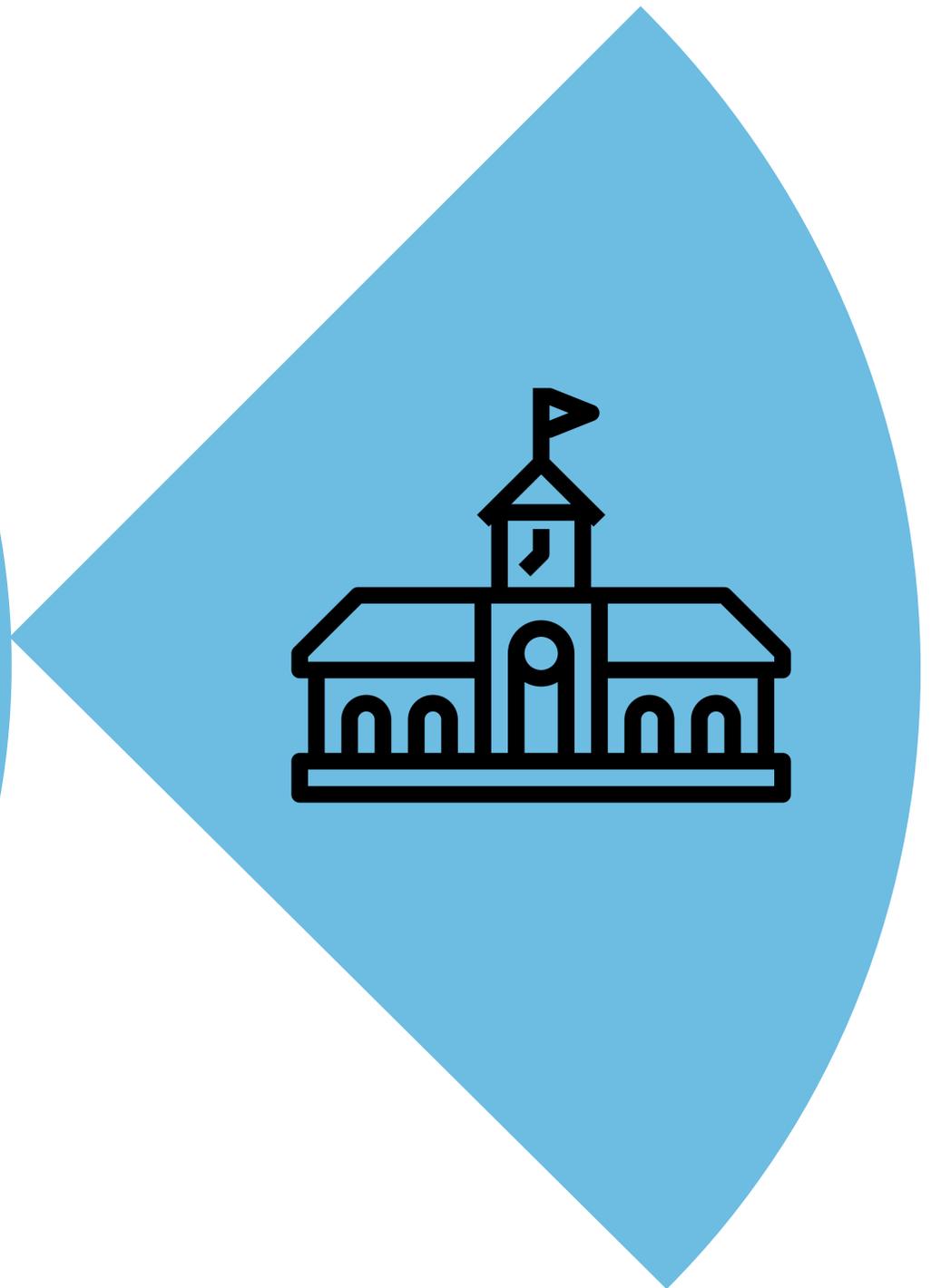
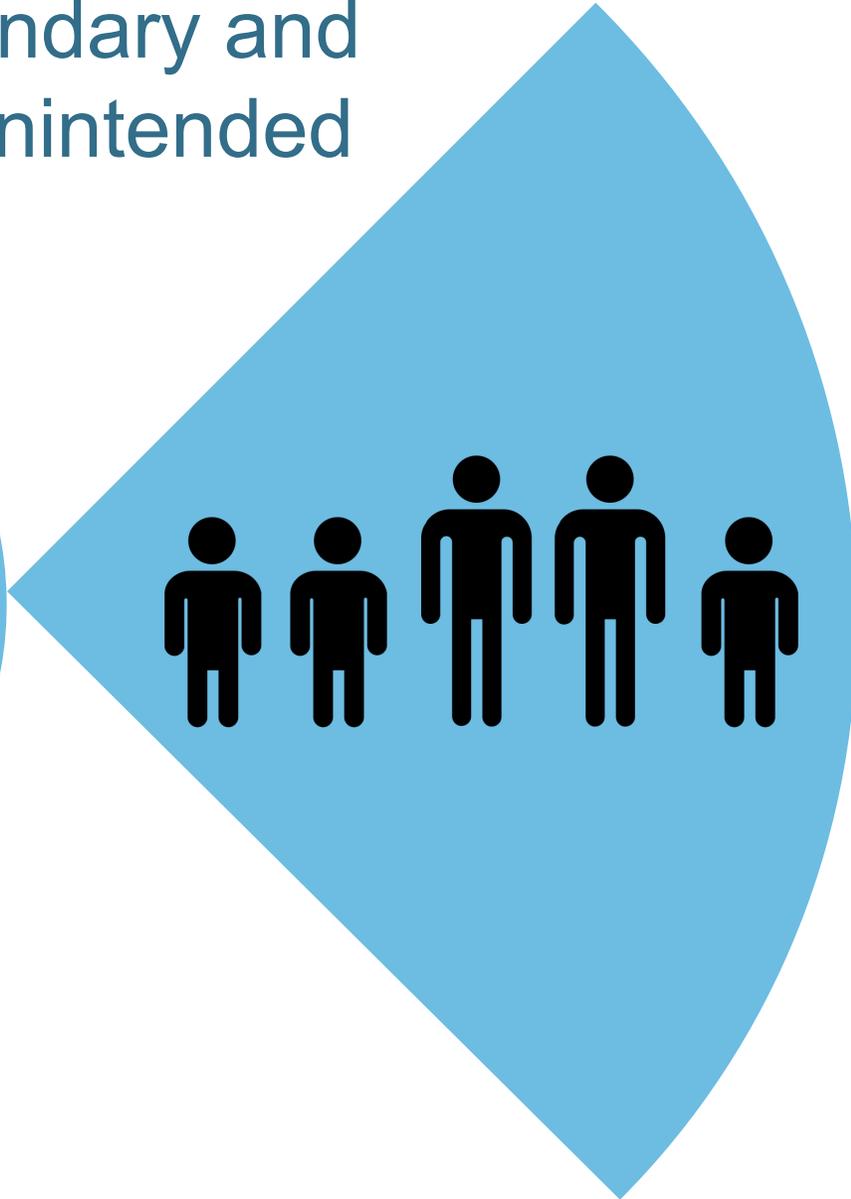
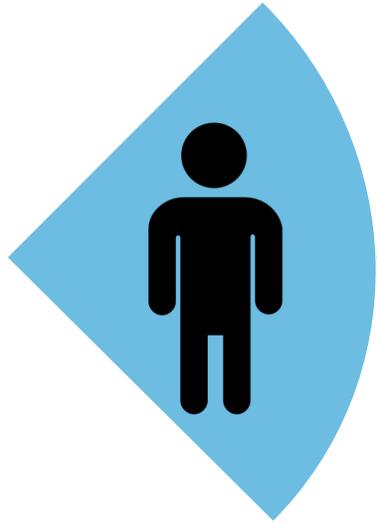
Who, what,  
where is  
impacted?



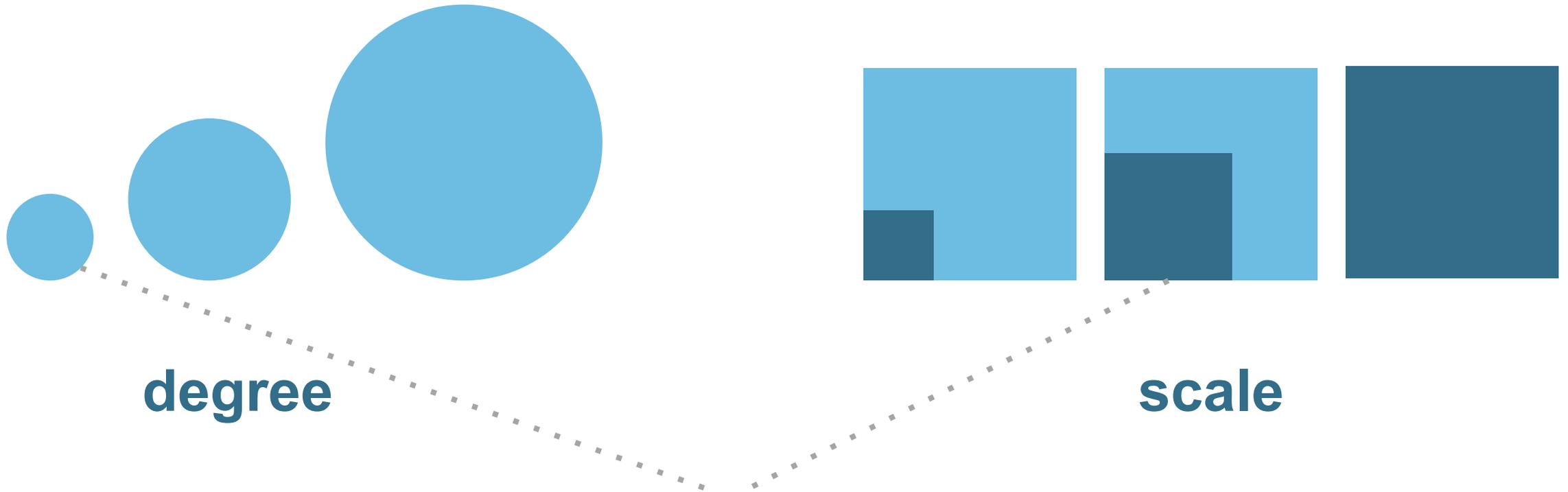
And in what way?



Primary, secondary and  
unexpected/unintended



# Scope of impact: a function of degree and scale

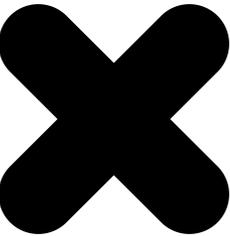


Minor degree for a substantial number of people  
= limited/narrow scope



**Is your use of the data appropriate?**

Consistent and compatible with original purpose?



Caution!!!

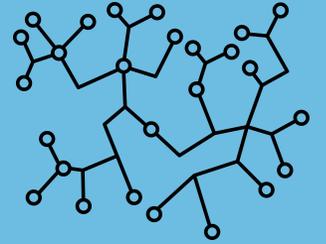
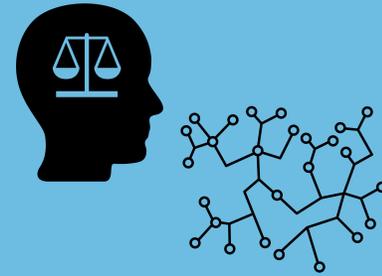
How would people react?



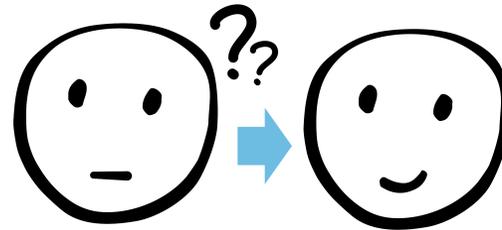
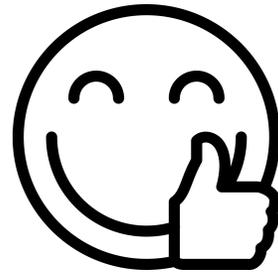


**What are the accountability risks?**

Level of automation?

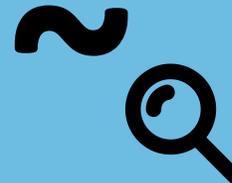


How explainable?

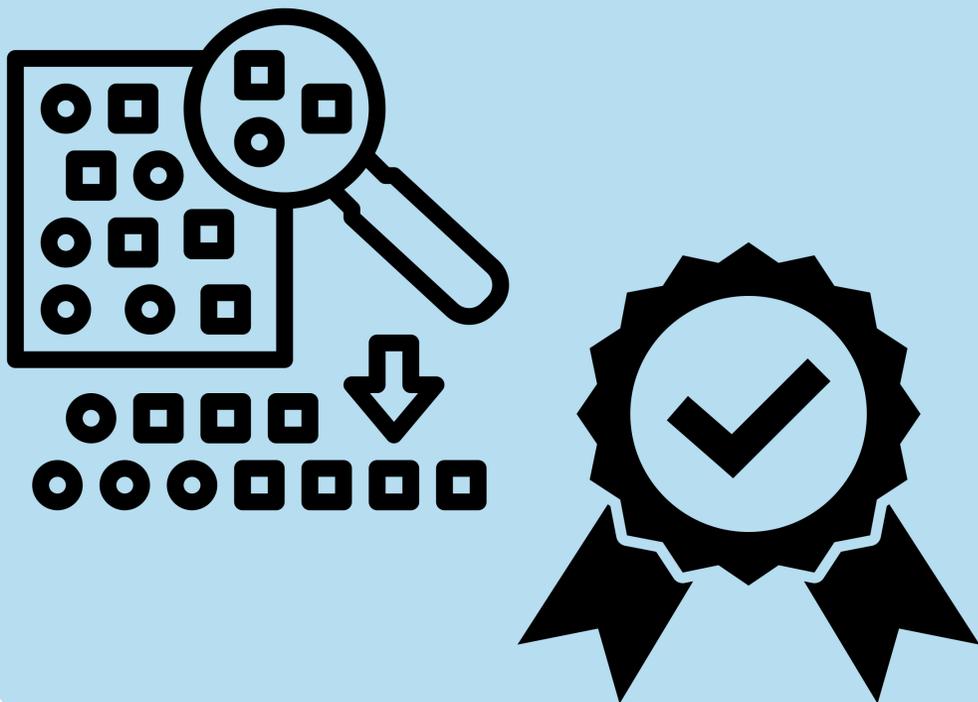


Some risk

Can we audit it?

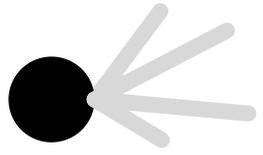


## Technical bias



## Historical bias





Inherently uncertain (technical + social)



Zero risk not possible



Policy and practice must balance  
risk with benefit





# Celebrating

**your data science pipeline**



Wrap it up with a bowtie. A demo day celebrates success, empowers your client and brings closure



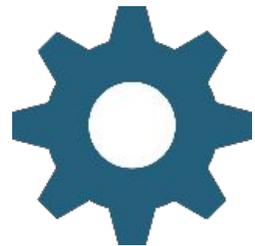
# Standardize your narrative structure



**Service question or issue**



**Data Science**

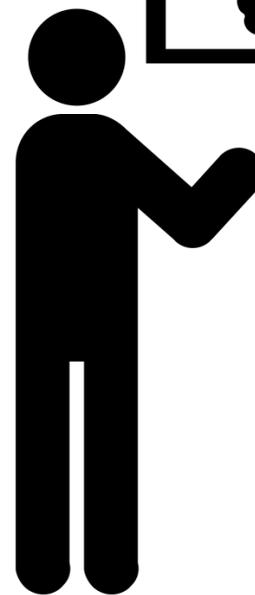


**Service Change**



**Results**

# Think Kindergarten, not death by deck



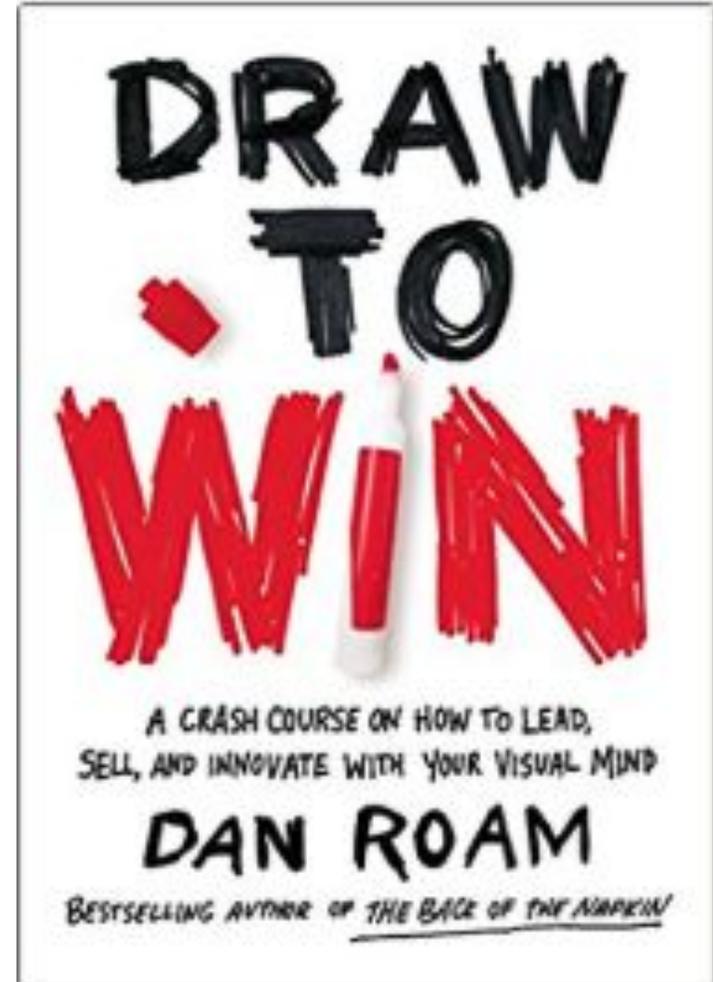
# Get visual but not in that data viz kinda way



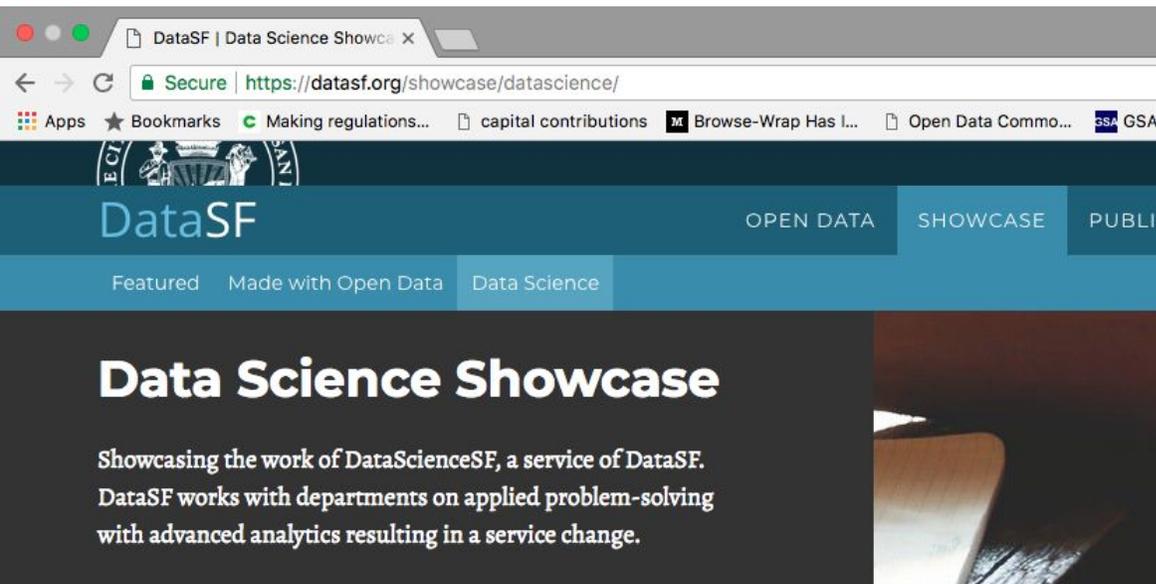
The Arts Commission's current formula was leading to some strange estimates. For example, it estimated that it would cost the same to conserve a piece made of a robust material and in good condition as a piece made of fragile material and in poor condition.

# More resources on storytelling

The screenshot shows a web browser window with the URL <https://www.lynda.com/Data-Science-tutorials/Storytelling-Data-Science/477450-2.html?srchtrk=index%3a1%0alink...>. The page title is "Learning Data Science: Tell Stories With Data". The main content area features a large video player with a play button and the text "Preview This Course". To the right, there are "Related Courses" listed, including "Big Data Foundations..." and "Data Science Foundations...". Below the video player, there are tabs for "Overview", "Transcript", and "View Offline". The "Overview" tab is selected, showing the author "Doug Rose" and the release date "2/19/2017". A skill level indicator shows "Appropriate for all" and a duration of "1h 17m". The "Contents" tab is also visible, showing a search bar and a list of topics: "Welcome 1m 40s", "1. Understand Stories", "Define a story 5m", "Spin a yarn 4m 39s", and "Weave a story together".

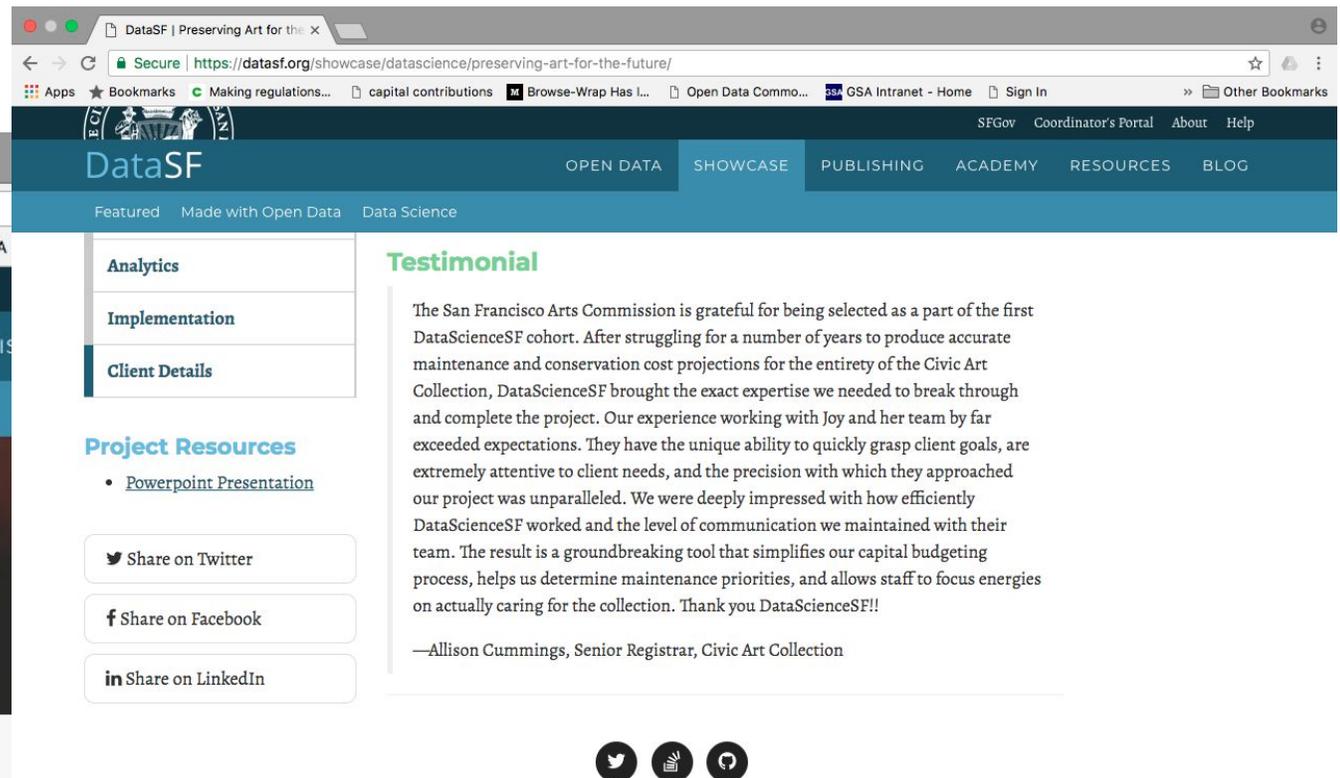


# Showcase your work so potential clients can discover it



**Data Science Showcase**

Showcasing the work of DataScienceSF, a service of DataSF. DataSF works with departments on applied problem-solving with advanced analytics resulting in a service change.



**DataSF** OPEN DATA SHOWCASE PUBLISHING ACADEMY RESOURCES BLOG

Featured Made with Open Data Data Science

**Analytics**

**Implementation**

**Client Details**

**Project Resources**

- [Powerpoint Presentation](#)

Share on Twitter

Share on Facebook

Share on LinkedIn

**Testimonial**

The San Francisco Arts Commission is grateful for being selected as a part of the first DataScienceSF cohort. After struggling for a number of years to produce accurate maintenance and conservation cost projections for the entirety of the Civic Art Collection, DataScienceSF brought the exact expertise we needed to break through and complete the project. Our experience working with Joy and her team by far exceeded expectations. They have the unique ability to quickly grasp client goals, are extremely attentive to client needs, and the precision with which they approached our project was unparalleled. We were deeply impressed with how efficiently DataScienceSF worked and the level of communication we maintained with their team. The result is a groundbreaking tool that simplifies our capital budgeting process, helps us determine maintenance priorities, and allows staff to focus energies on actually caring for the collection. Thank you DataScienceSF!!

—Allison Cummings, Senior Registrar, Civic Art Collection

Twitter Facebook LinkedIn



## AB Testing Service and Results to Date

Learn about our AB testing workshop service and about completed experiments across the City!



## Eviction Alert System

In 2015, 24 different families living at 2 Emery Lane in San Francisco received eviction notices. Fortunately, the residents were more than just neighbors -



## Greening the City with Better Lighting

When was the last time you decided to upgrade your light bulbs? If you're like many folks, you wait until a light bulb



Learn more  
about this  
approach to  
managing  
data science  
in gov't



Part 1: How to solicit and select data science projects

Joy Bonaguro [Follow](#)  
Nov 15, 2019 · 5 min read

[Twitter](#) [LinkedIn](#) [Facebook](#) [Bookmark](#)

*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](#)

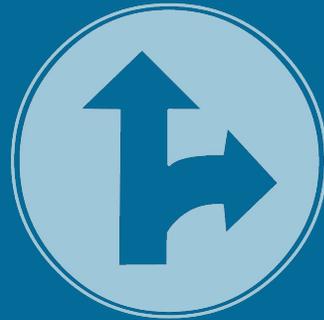


# Statewide Data Goals

*Equipping ourselves to navigate the data landscape*



**Build the data roads**  
*streamline data access*

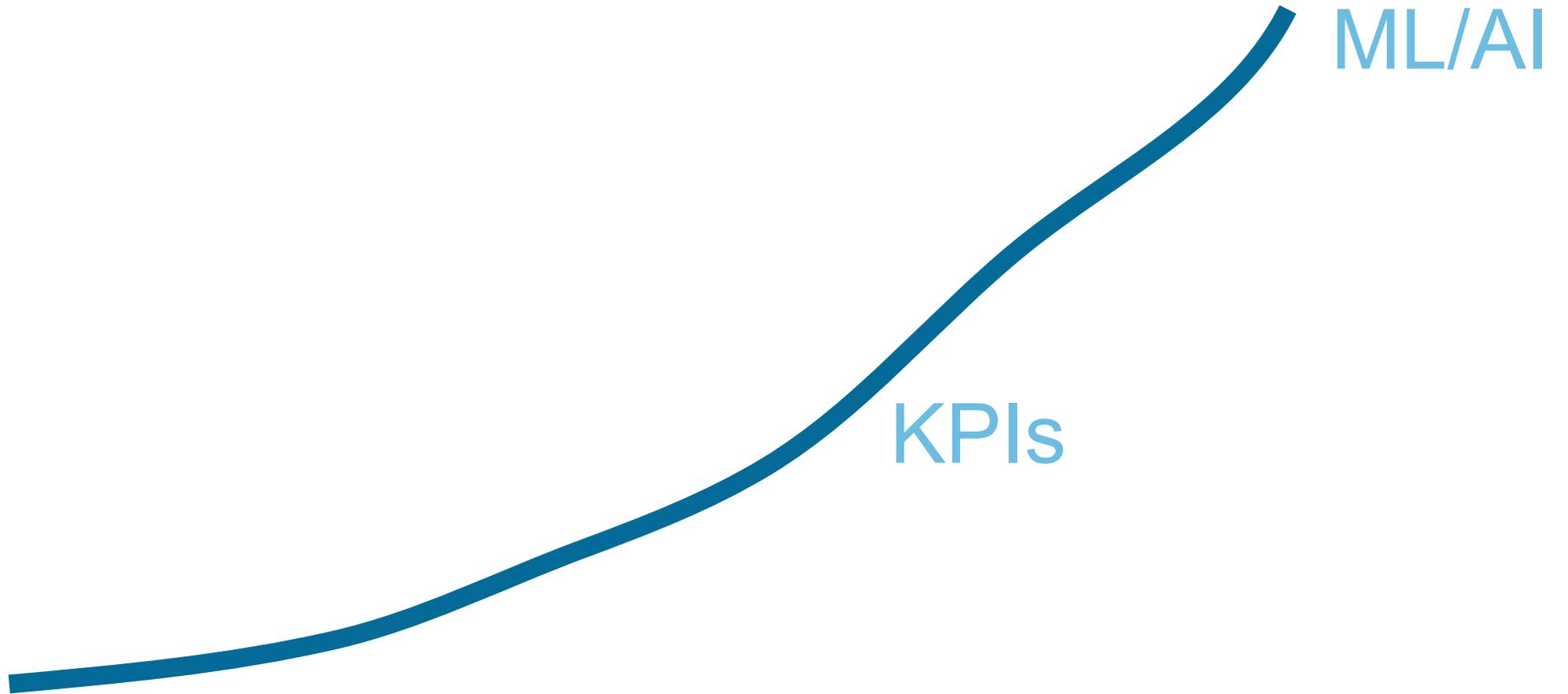


**Craft the rules of the road**  
*improve data management and  
governance*



**Boost the drivers**  
*spur data use and ability*

Data maturity models always assume up is best...



# 2 step plan to build data maturity

## 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.

## 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.



### Level 3: Data on Demand

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



# Thank you!

Questions / feedback / feelings /  
reactions / thoughts  
welcome!

[www.govops.ca.gov/caldata/](http://www.govops.ca.gov/caldata/)