Introducing Data Science

in Legacy Organizations

Joy Bonaguro | Statewide Chief Data Officer | State of CA

NVP
NewVantage Partners

Big Data and AI Executive Survey 2020

Executive Summary of Findings

Data-Driven Business Transformation

Connecting Data/AI Investment to Business Outcomes

With a Foreword by Thomas H. Davenport and Randy Bean

NewVantage Partners LLC

www.newvantage.com

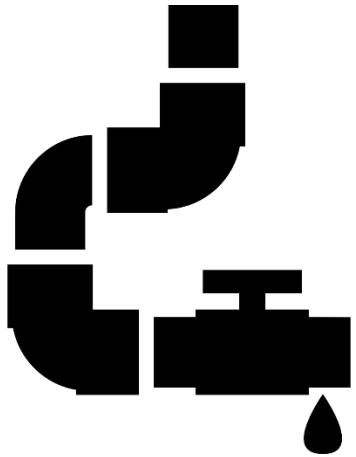
Boston | New York | San Francisco | Austin | Raleigh

NVP
NewVantage Partners
Boston | New York | San Francisco | Austin | Raleigh
www.newvantage.com

"Companies ... focus on the **supply side** for data and technology, **instead of increasing demand** for them by business executives and employees.

Big Data and AI Executive Survey 2020, NewVantage Partners LLC

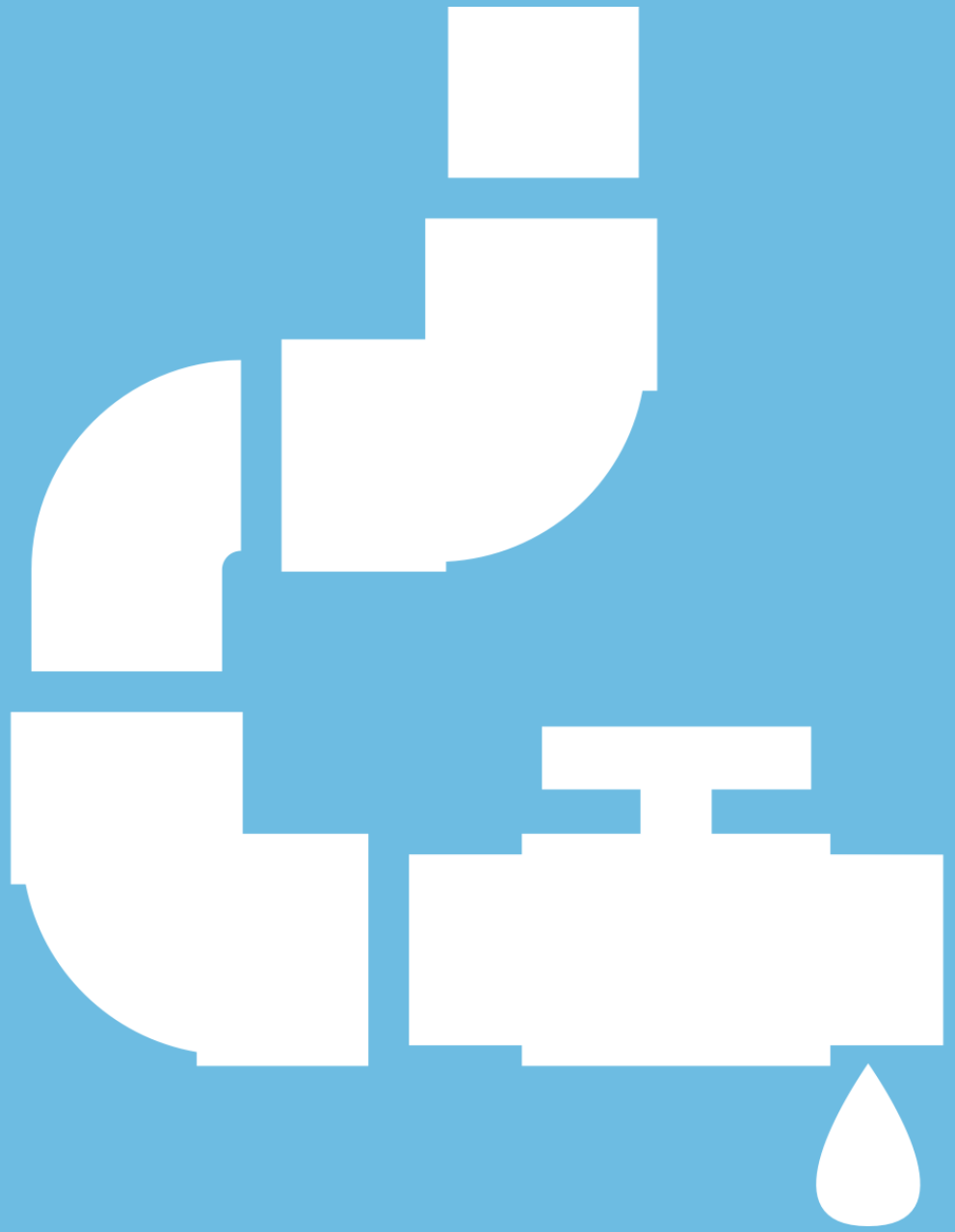
Objective: Demonstrate power of data science and eventually embed in culture and fabric



Develop

Deliver

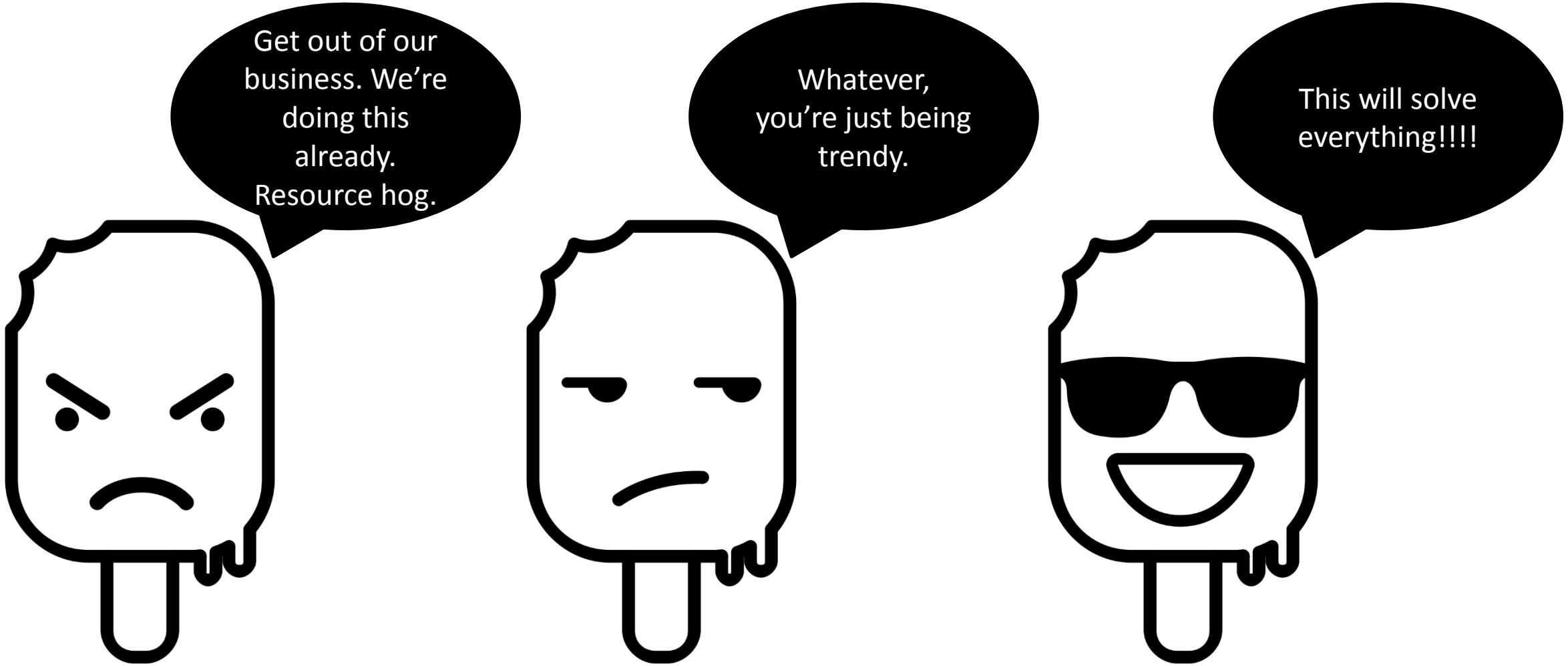
Celebrate



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 | Tools | User Experience Research

Sentiment analysis

Time series analysis

Data mining

Multilevel modeling

Missing data imputations

Classification and clustering

Survival analysis

Pattern recognition

Principal component and factor analysis

AB testing

Machine learning

Forecasting

Propensity score matching

Logistic, multinomial and multiple linear regression techniques

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?

✓ This

Service change

Small changes

Created by Danil Polshin from Noun Project

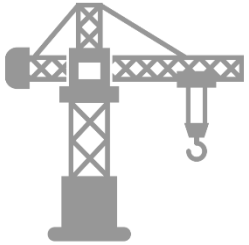
Use existing data

Created by Arthur Shlain from Noun Project

✕ Not that

Academic research

Created by Rockicon from Noun Project



Major overhauls / service disruptions

Created by Hopkins from Noun Project

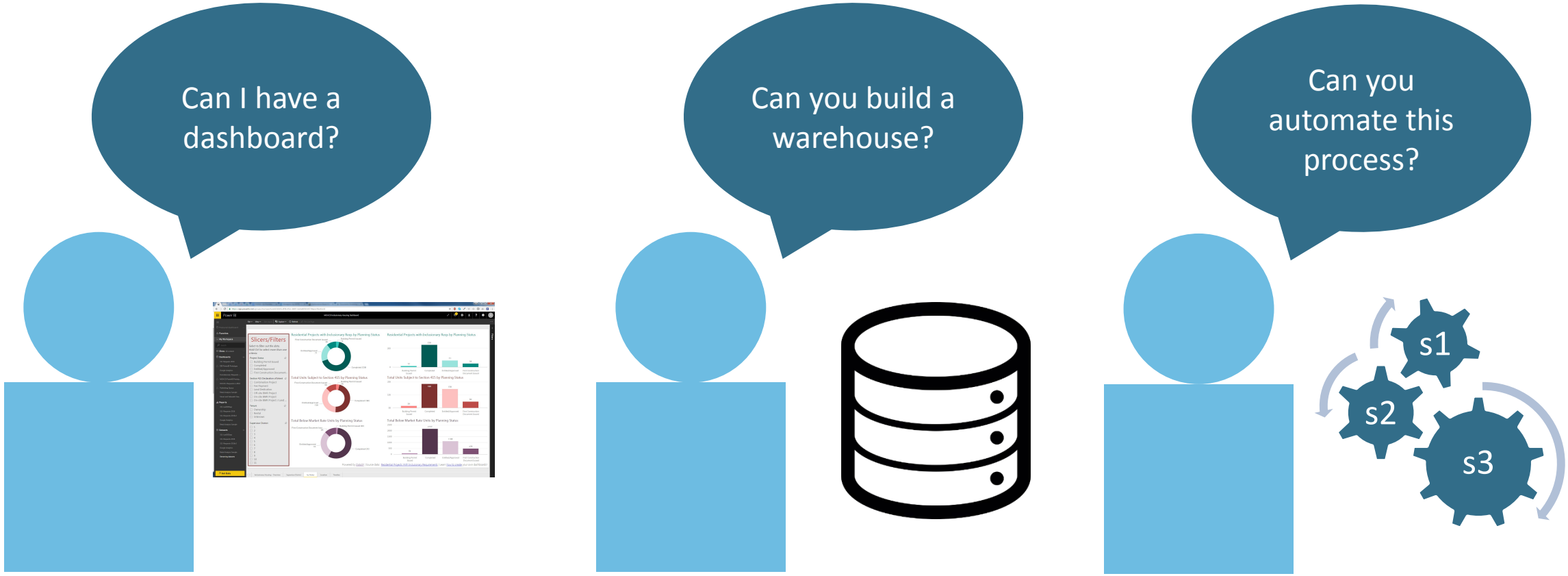
Collecting new data (mostly ;)

Created by Chameleon Design from Noun Project

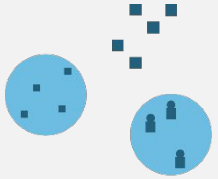
The Truffle Pig Problem:

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

You know you have a truffle pig problem if…



Solution: The project typology

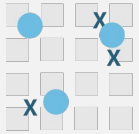


Find the needle in the haystack

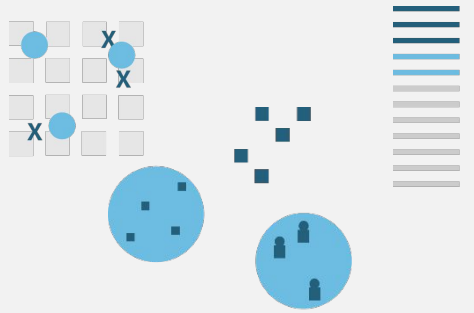
Prioritize your backlog

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

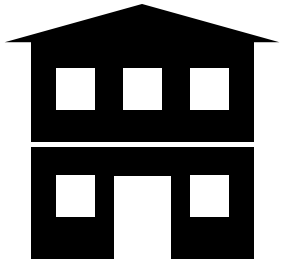
Service Issue: Processing a giant backlog of property sales

Sale price

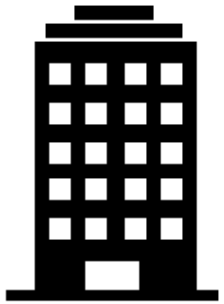
?

Fair market value

Data Science



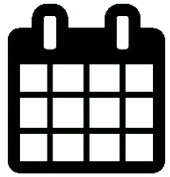
Condo



Multi-family

Single family home

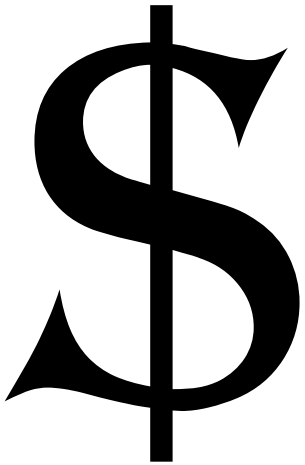
Property Characteristics



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

How to detect?

Data Science

Service Change

Service Issue:
Hard to predict future condition which leads to reactive services

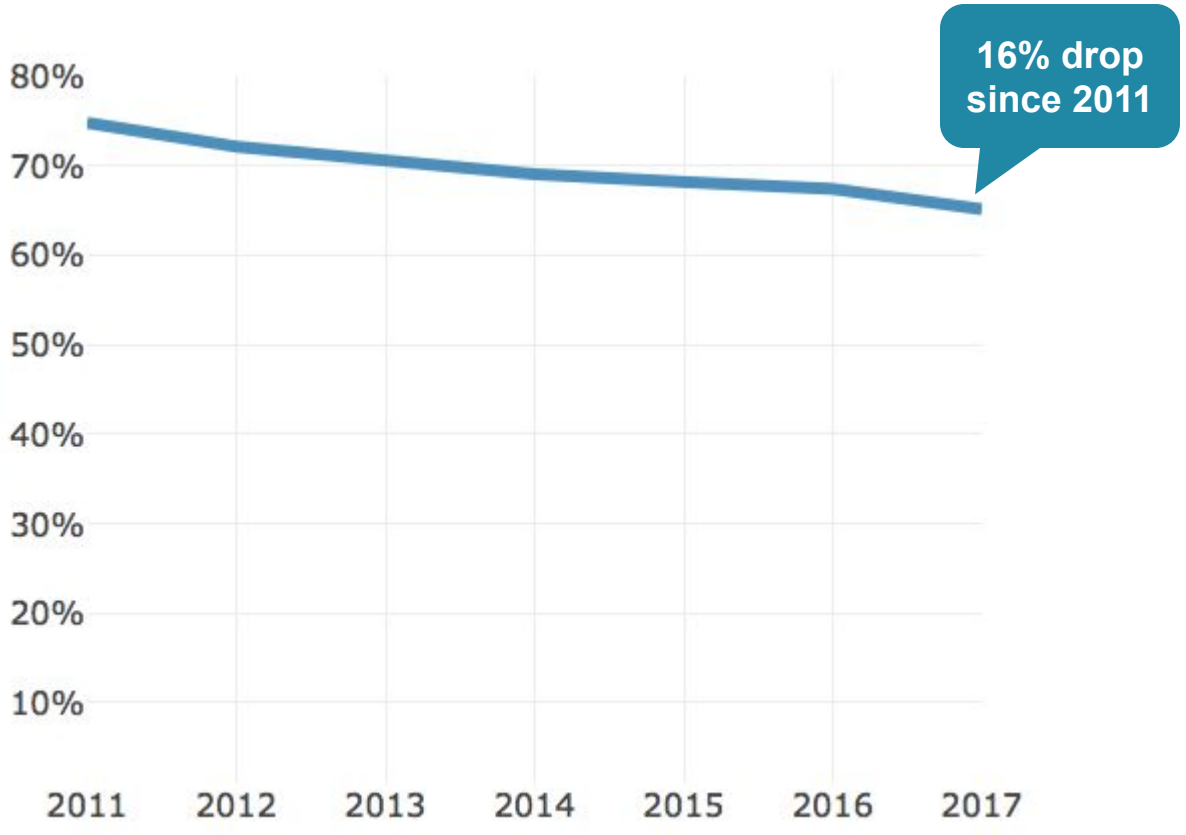
Data Science Process:
Use historical and current data to create estimate ranges for potential outcomes

Service Change:
Use estimates to change and tailor intervention points

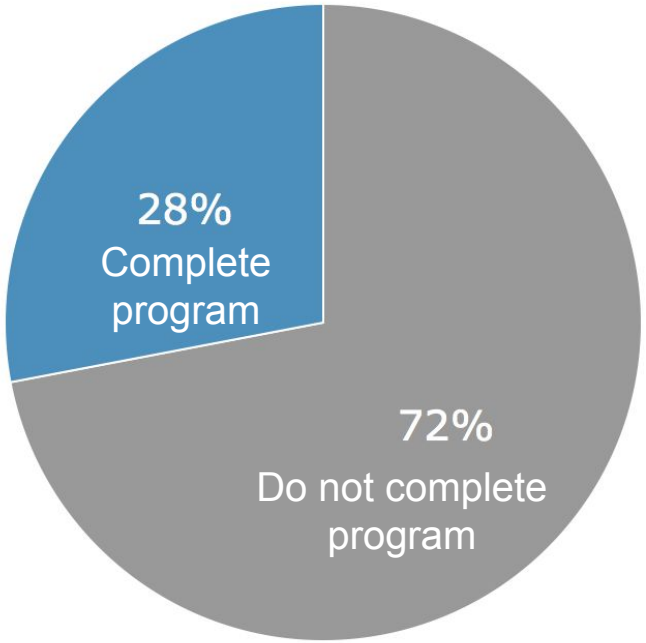
Result: Department provides pro-active early interventions

Service Issue: Dropping out of WIC

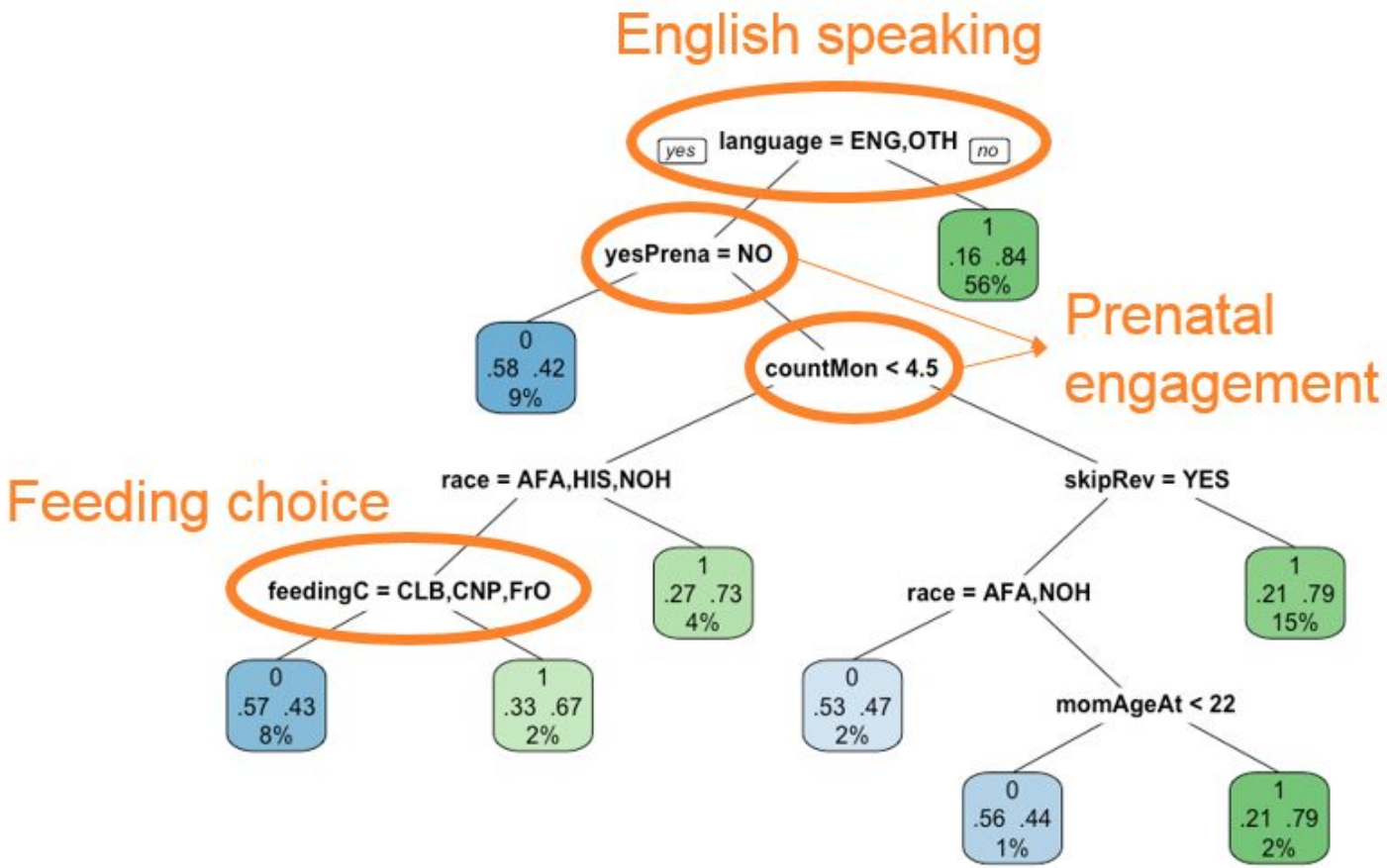
% still in program at 1 year of age (retention)



Completion status through age 5

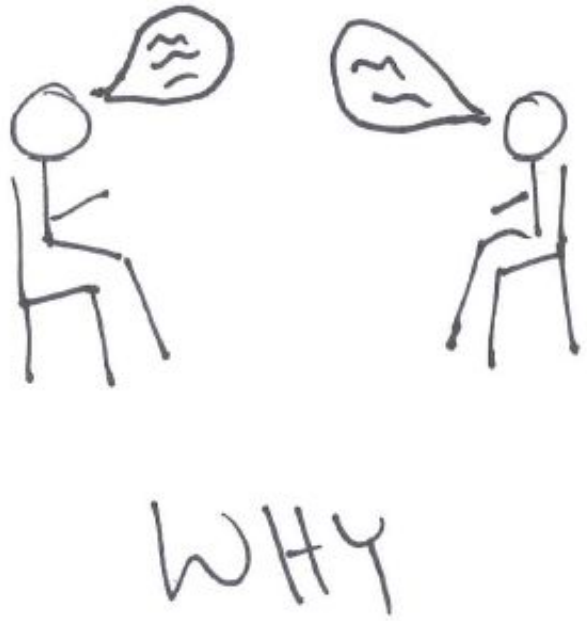


Data Science



Service Change

Qualitative



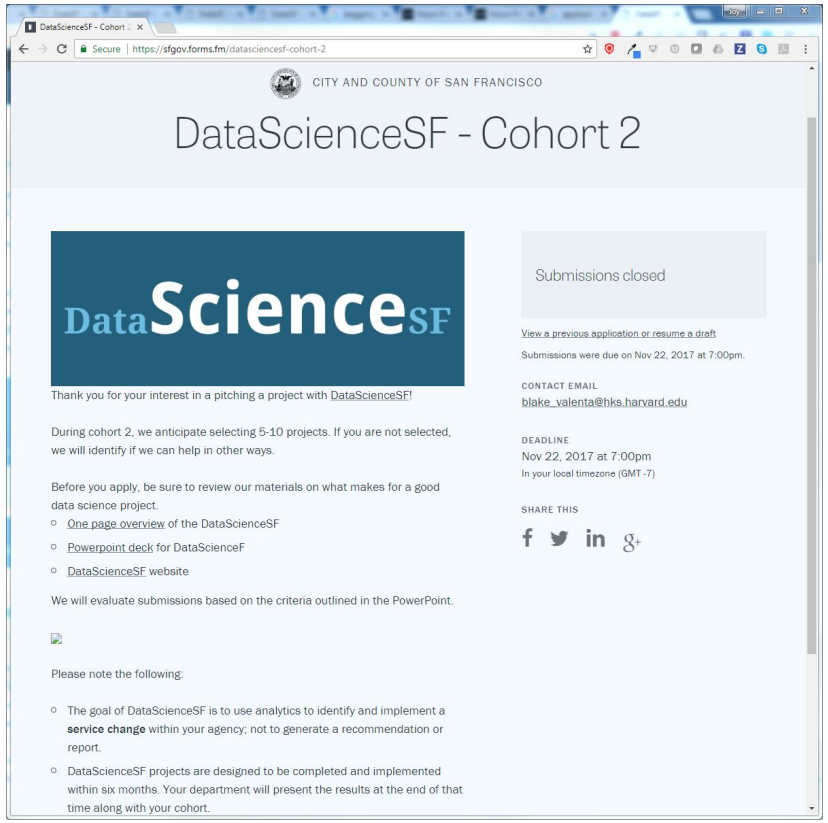
Anti-Stigma Campaign

New Technology

Future Outreach Strategy

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

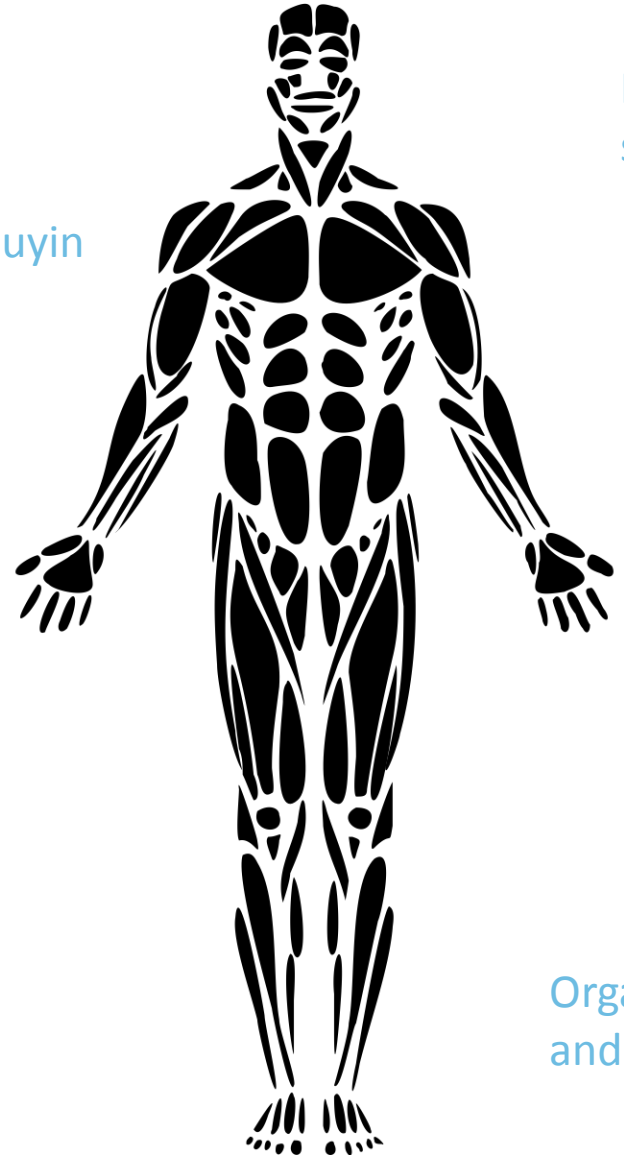
Invert the dynamic with an application process



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

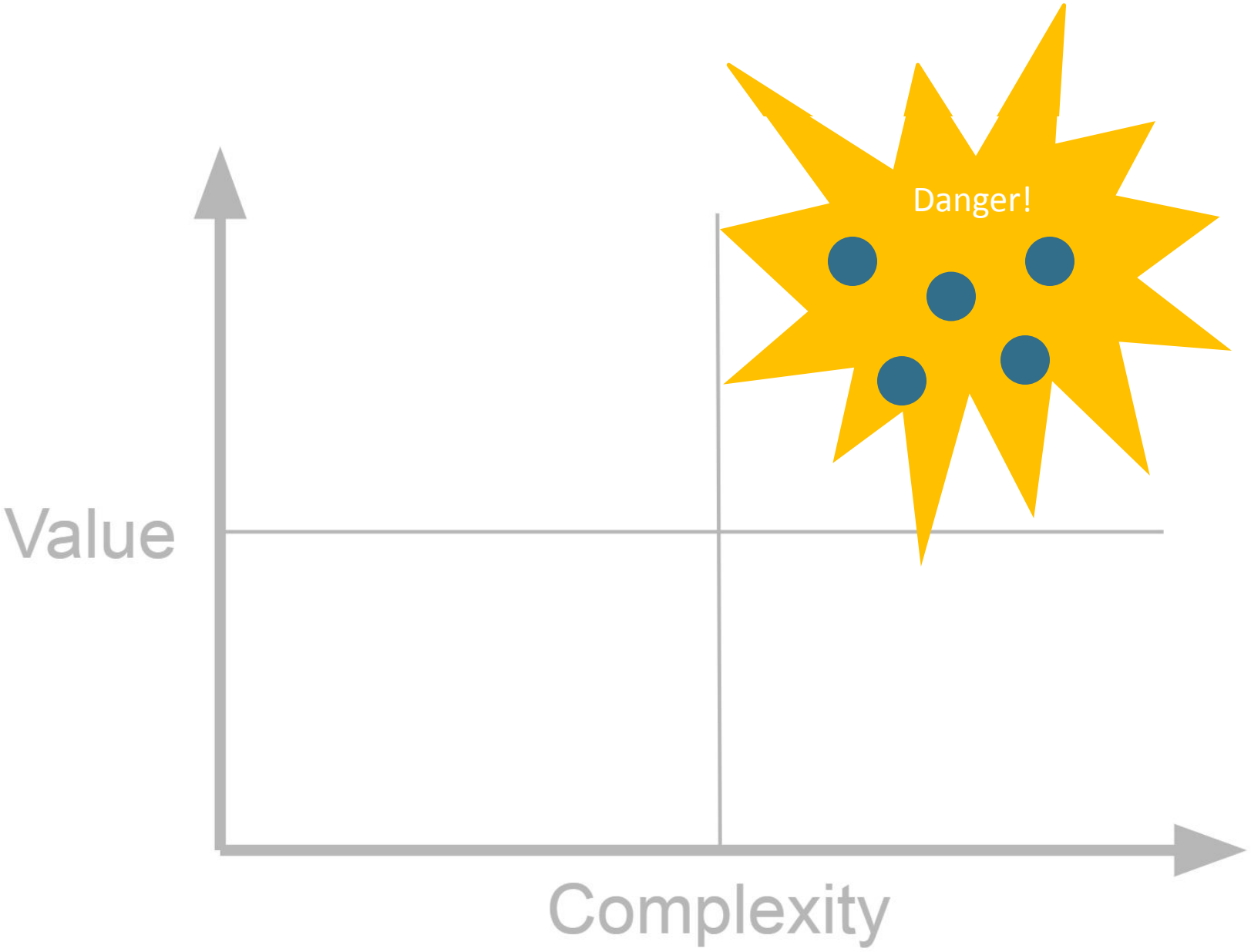
Don’t do this.
Sorry, can’t help.

Tim Evanson, https://www.flic 5

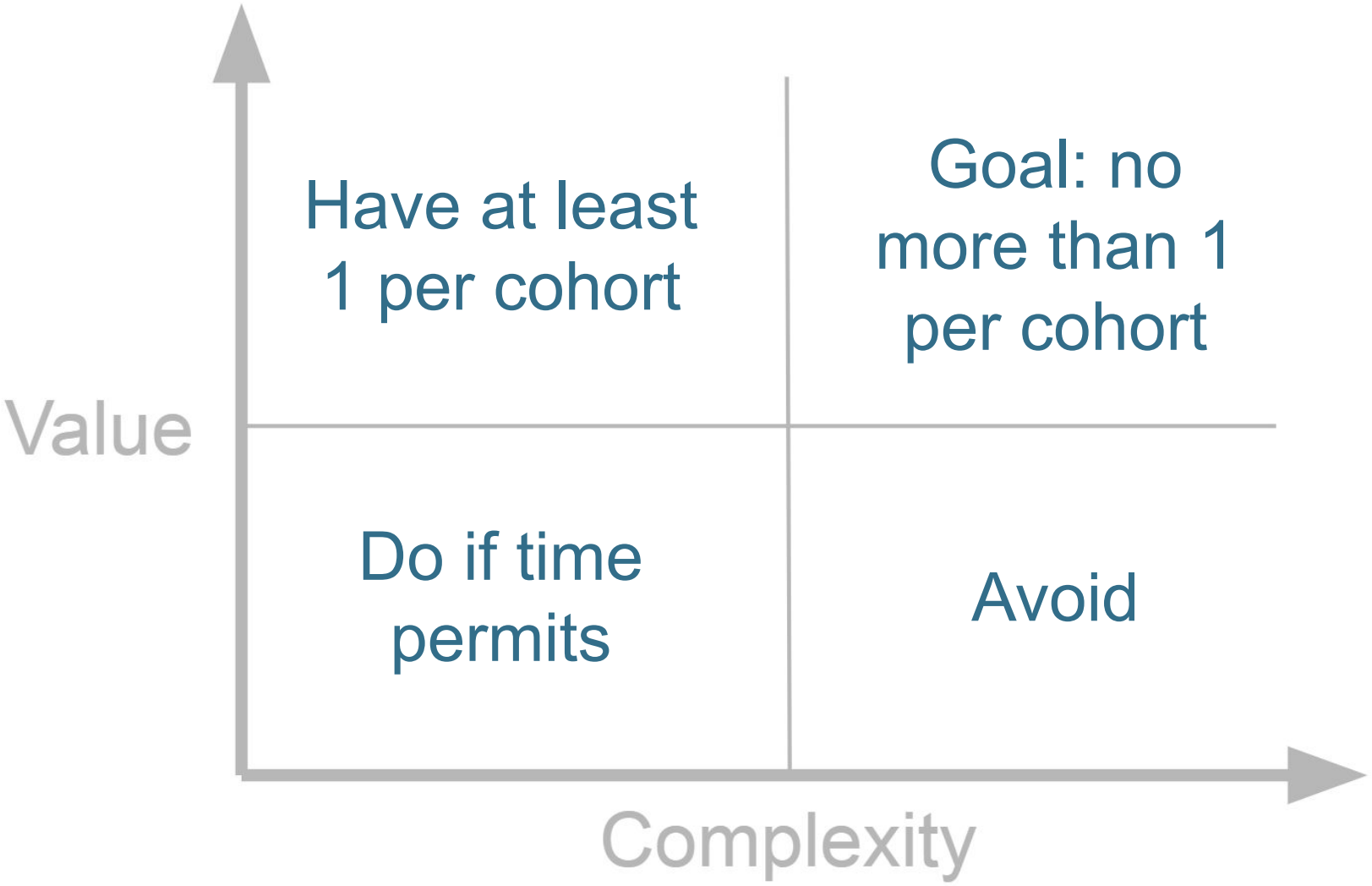
Do this.
Provide gentle offramps.

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

Score and balance your portfolio



Score and balance your portfolio

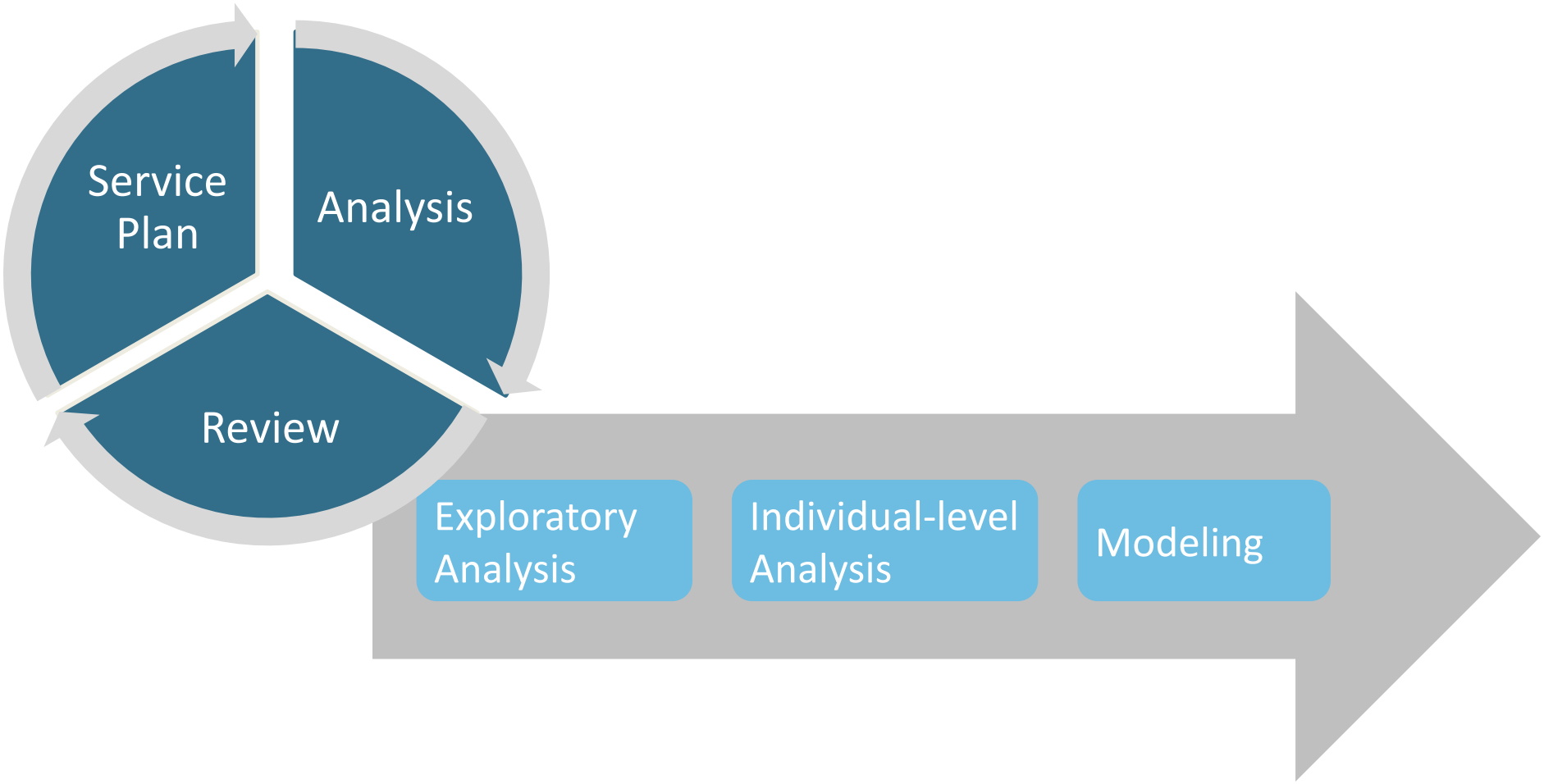


Delivering

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

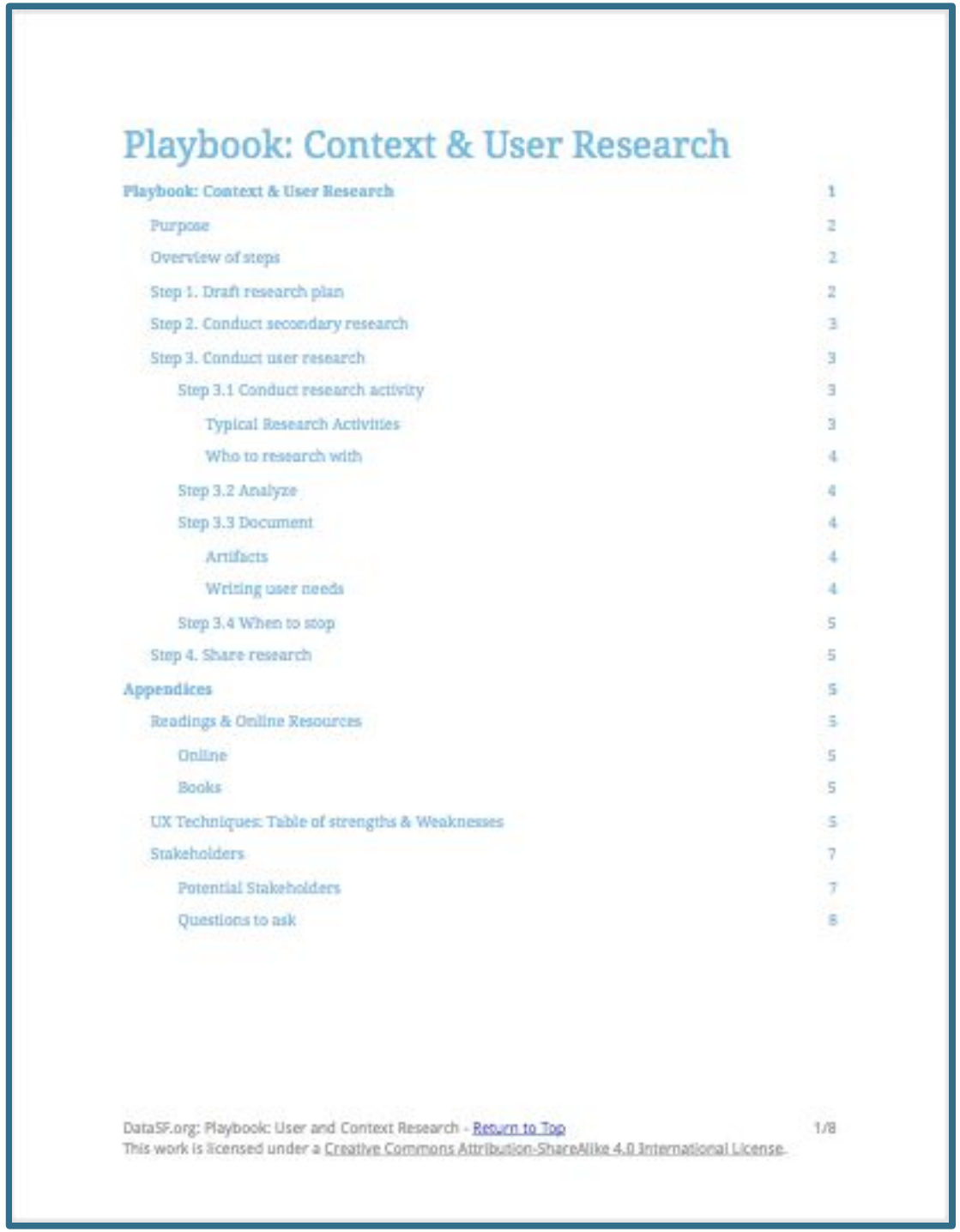
...weekly or biweekly (scheduled early)

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

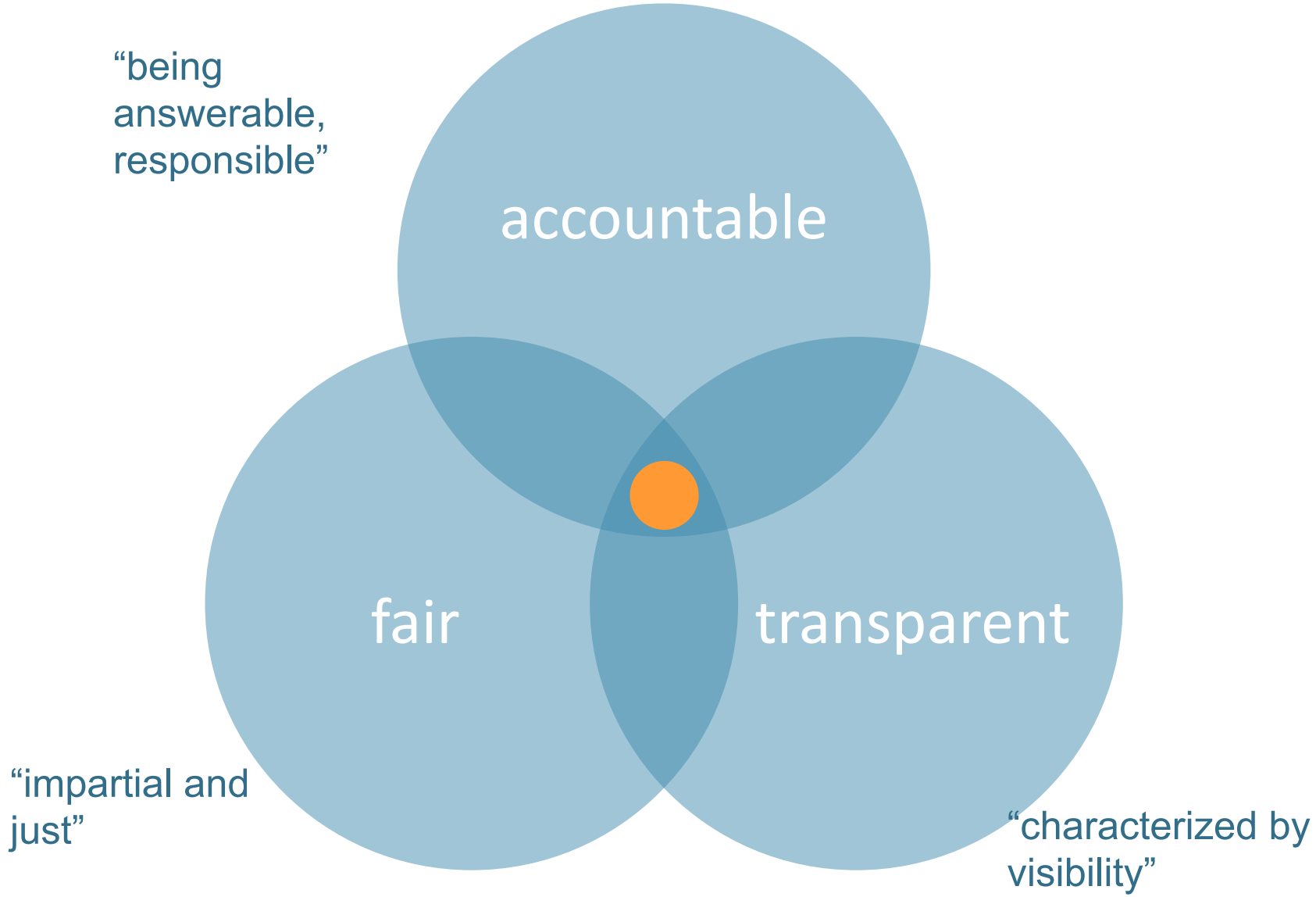
...or you may deliver a product that is not usable



Playbook: Context & User Research

Squeeze, don’t wring

Ethics & Algorithms a practical toolkit



DataSF

GovEx

HARVARD Kennedy School
ASH CENTER
for Democratic Governance and Innovation

Data Community DC

Ethics & Algorithms Toolkit

Not Secure | ethicstoolkit.ai

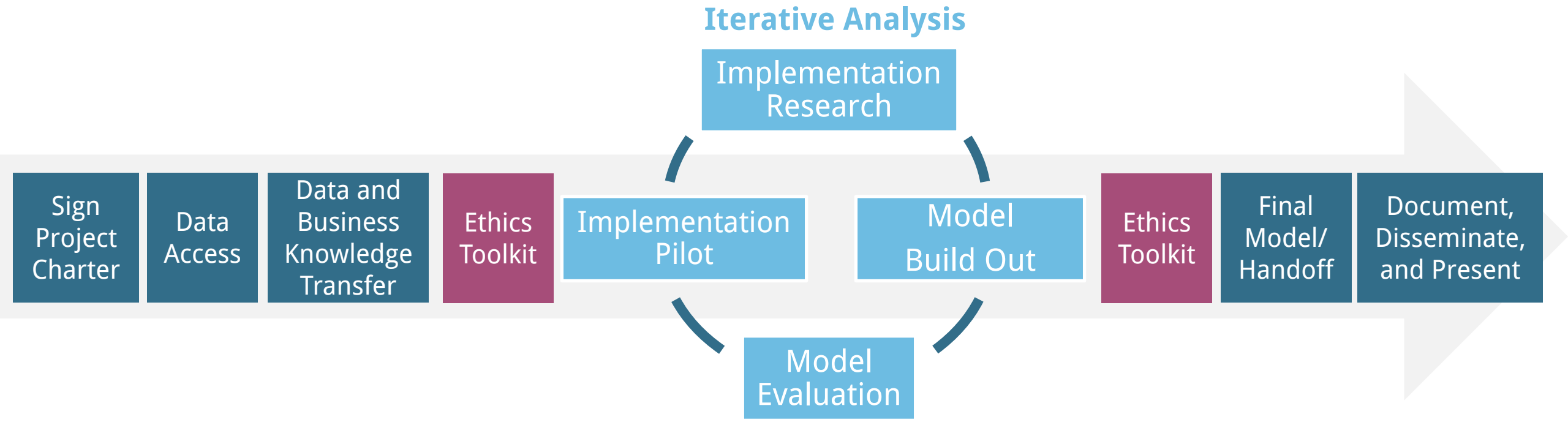
TOP BACKGROUND TEAM THE TOOLKIT FEEDBACK MEDIA LICENSE

beta

Ethics & Algorithms Toolkit

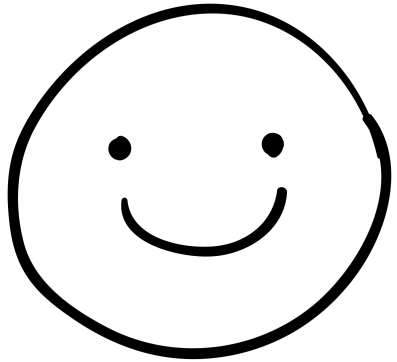
A risk management framework for governments (and other people too!)

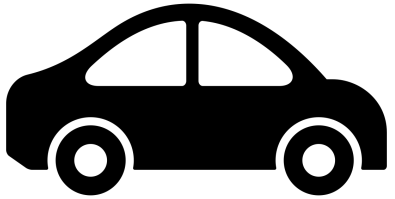
Toolkit became a core part of the project cycle

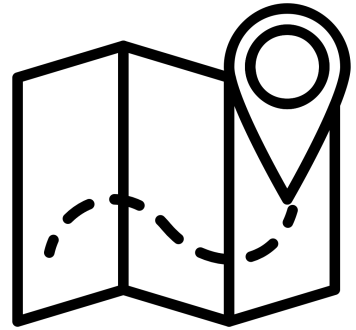


What is the scope of the impact?

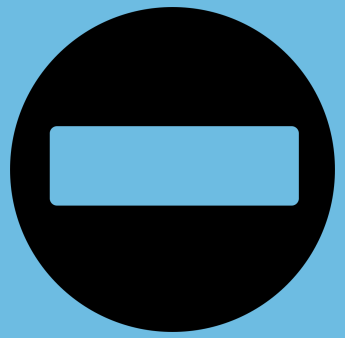
Who, what, where is impacted?



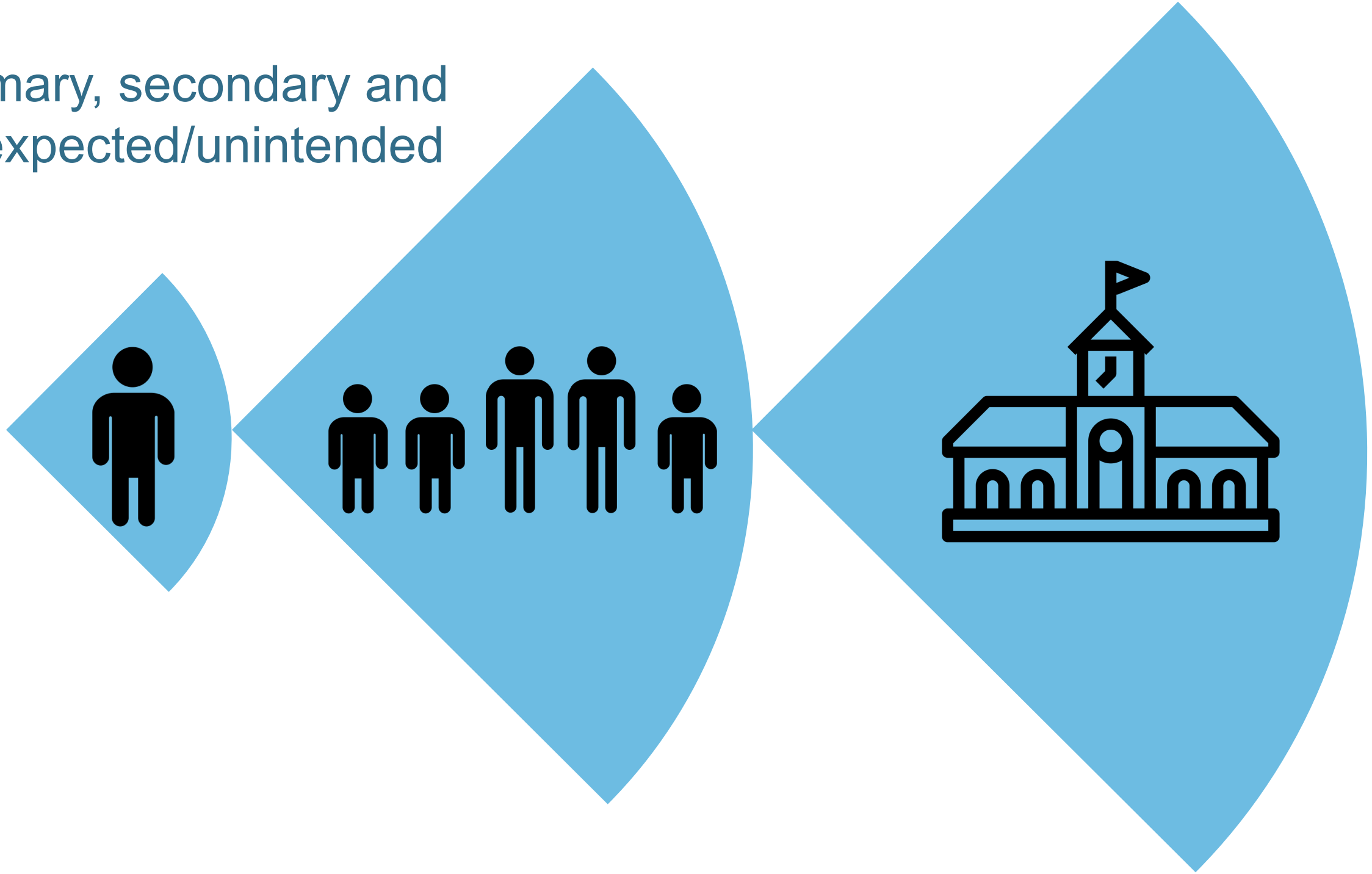




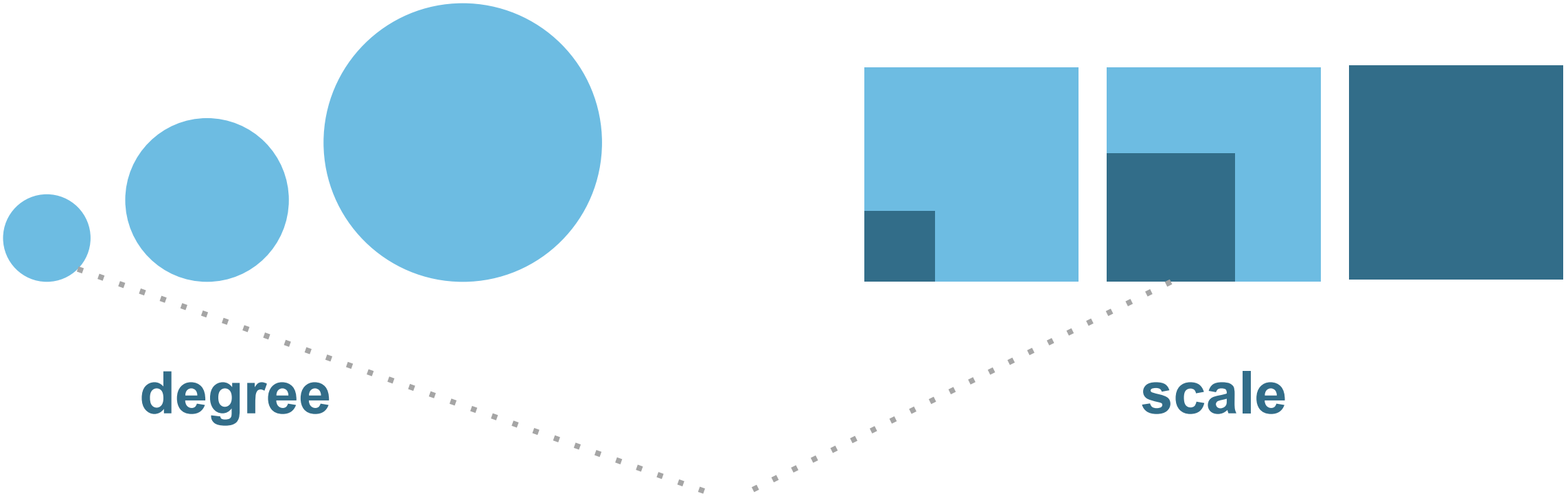
And in what way?



Primary, secondary and unexpected/unintended



Scope of impact:
a function of degree and scale

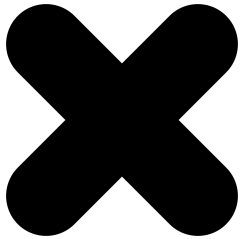


Minor degree for a substantial number of people
= limited/narrow scope

Is your use of the data appropriate?

Consistent and compatible with original purpose?

¯_(ツ)_/¯



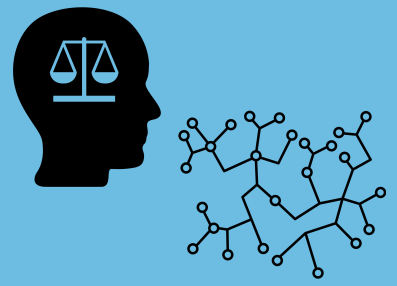
Caution!!!

How would people react?

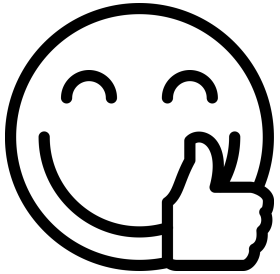
What are the accountability risks?

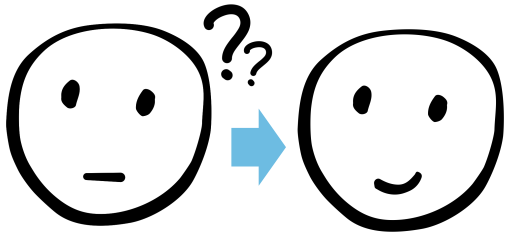
https://emersonsuite.blog/2017/02/08/are-you-accountable-for-what-you-ask/

Level of automation?



How explainable?

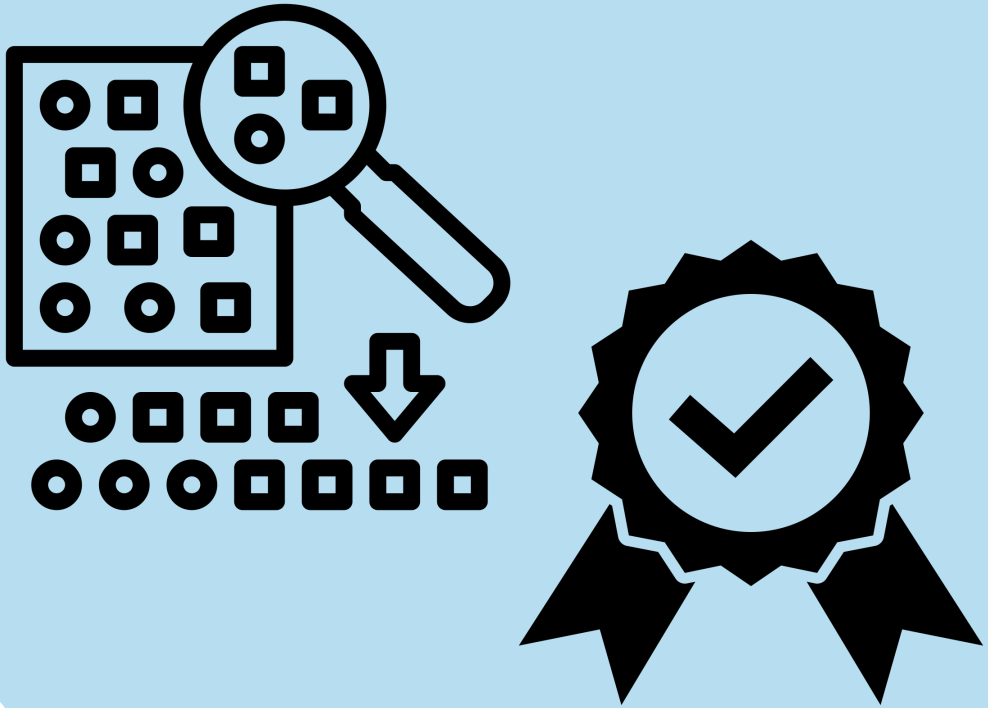




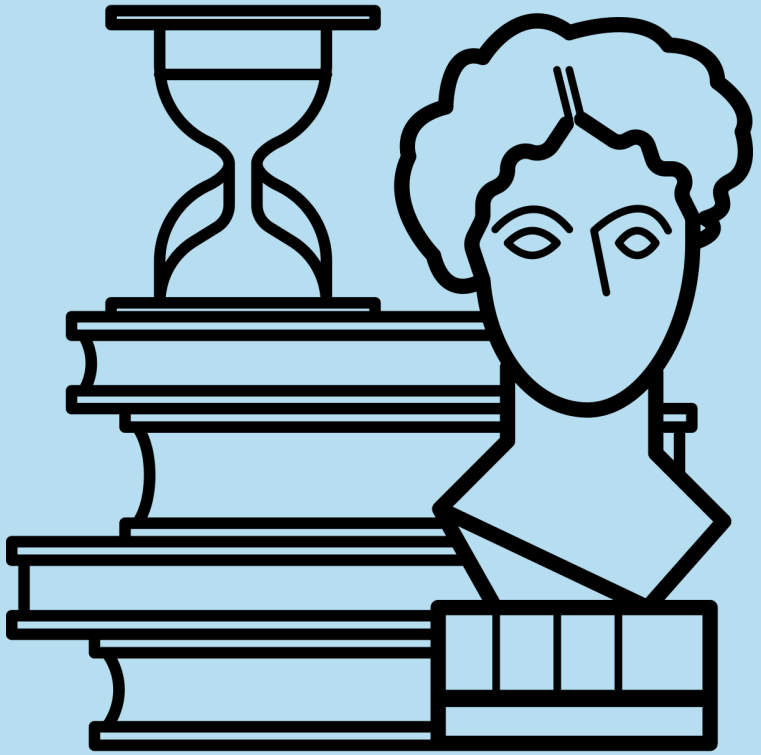
Some risk

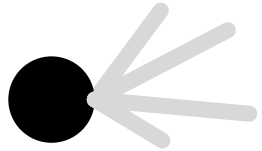
Can we audit it?

Technical bias



Historical bias





Inherently uncertain (technical + social)

Zero risk not possible

Policy and practice must balance risk with benefit

http://www.pbglifestyle.com/pbglm-blotter/boy-in-the-bubble-wrap/

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

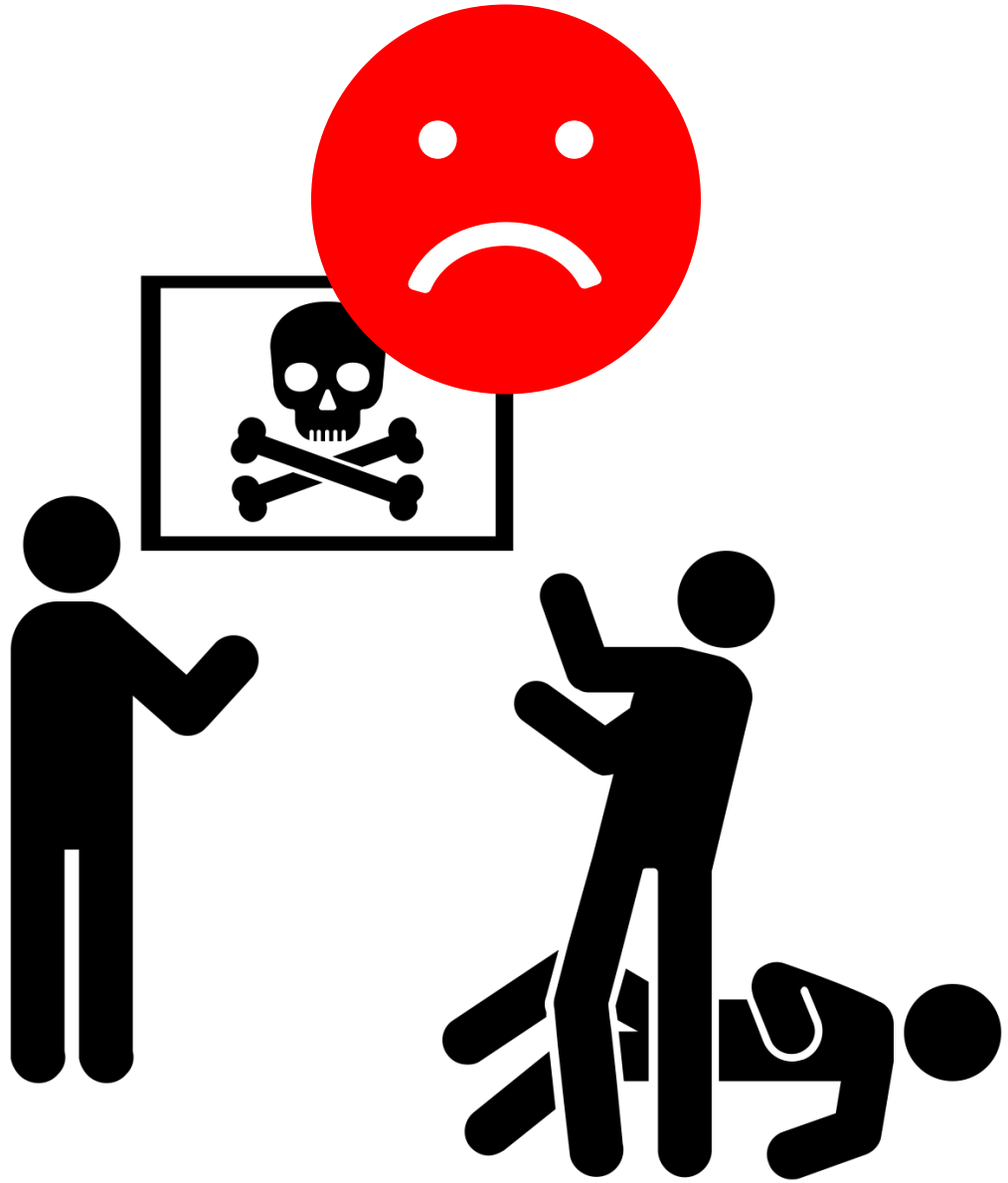
Service question or issue

Data Science

Service Change

Results

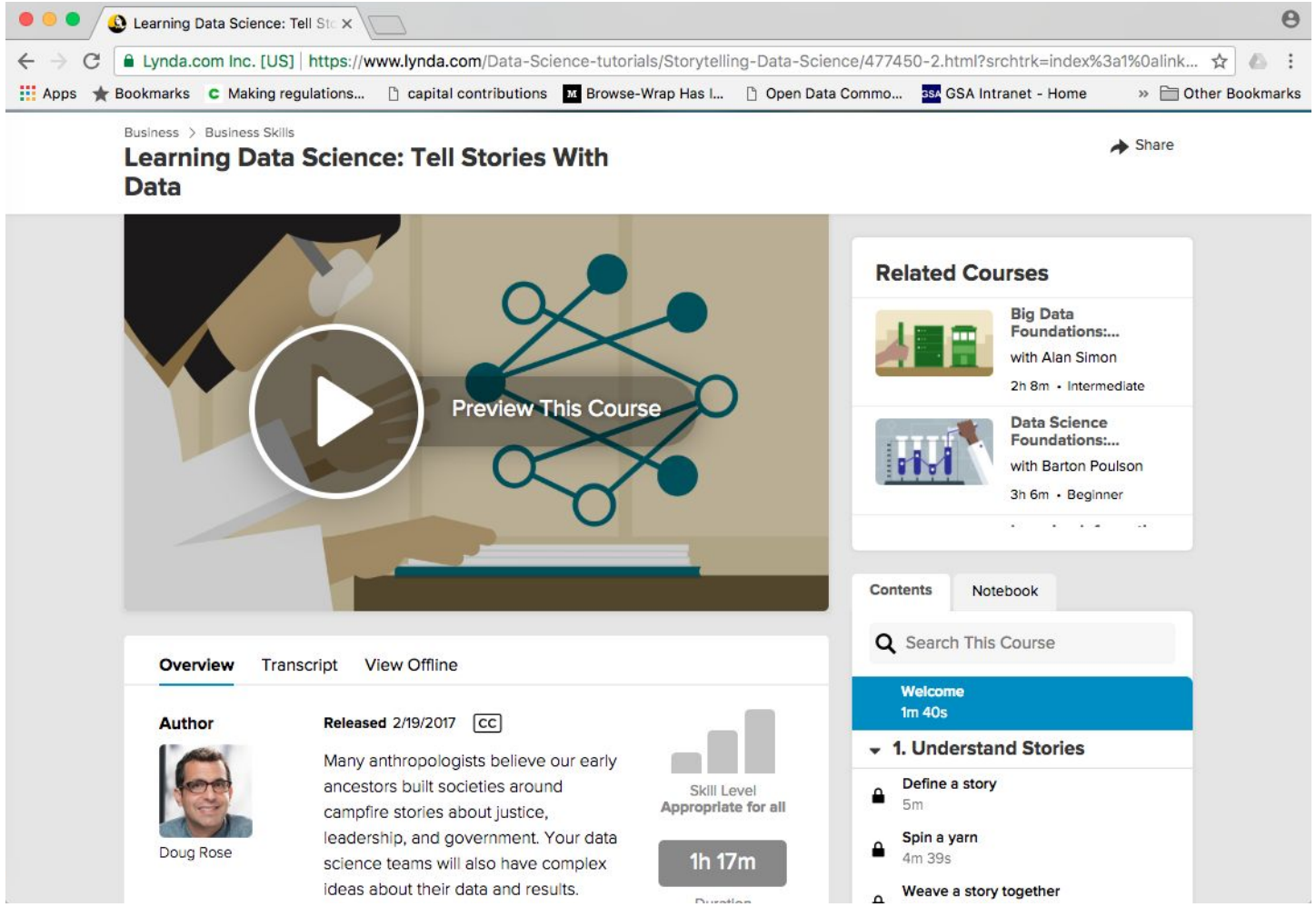
Think Kindergarten, not death by deck

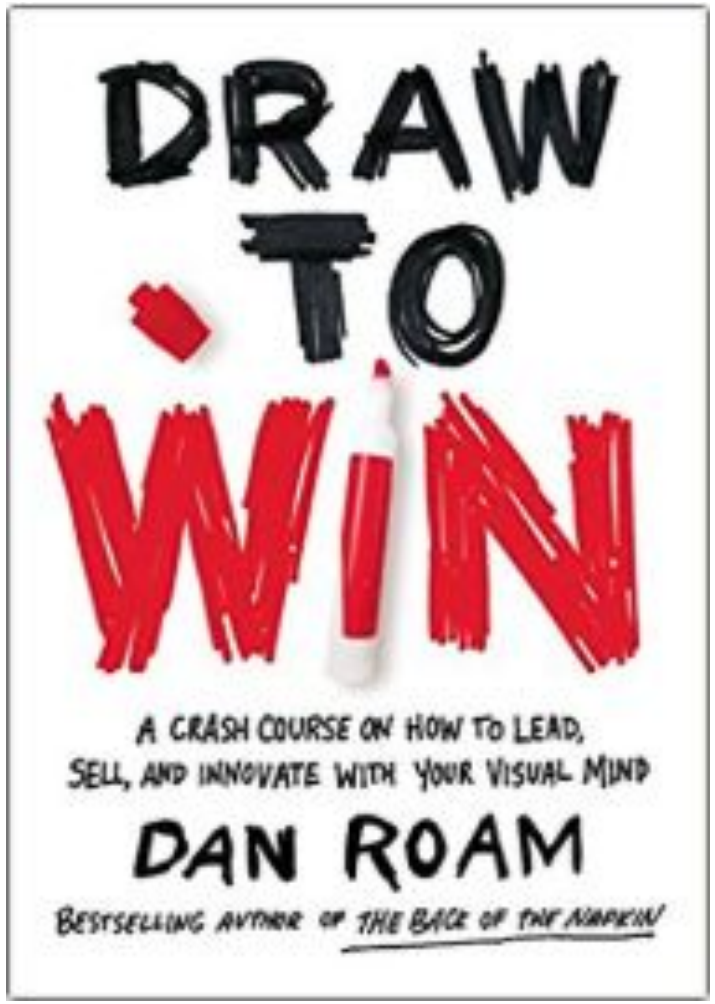


Get visual but not in that data viz kinda way

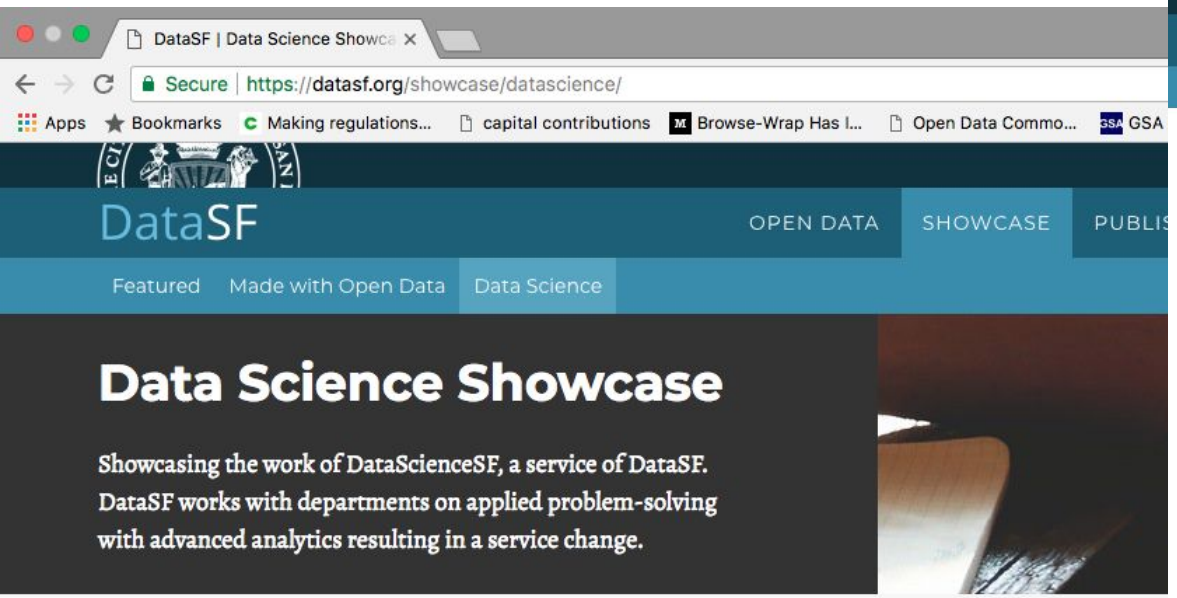
The Arts Commission's current formula was leading to some strange estimates. For example, it estimated that it would cost the same to conserve a piece made of a robust material and in good condition as a piece made of fragile material and in poor condition.

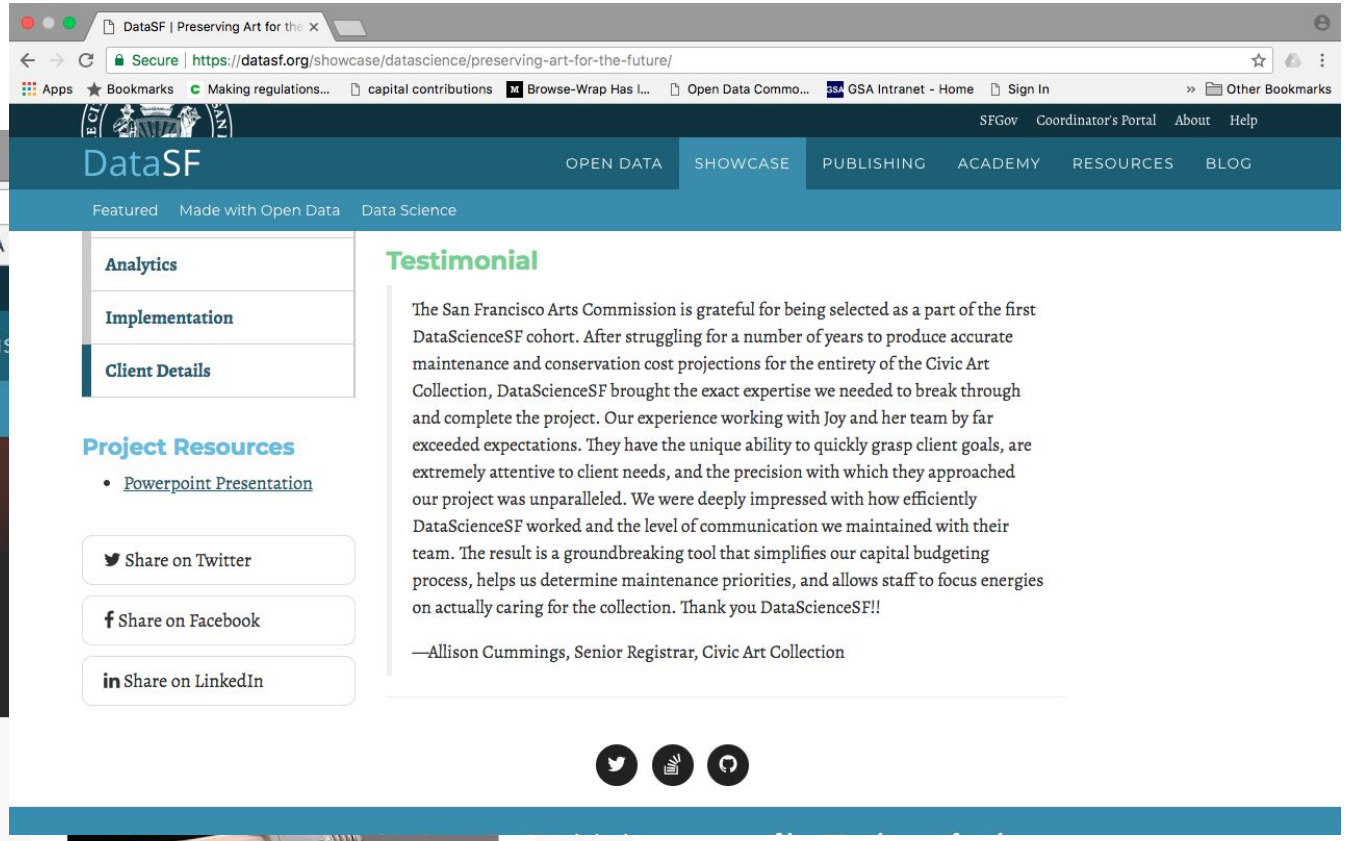
More resources on storytelling





Showcase your work so potential clients can discover it





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


Virtuous Solicitation Cycle
New clients
New Projects
Tell Story

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

Part 1: How to solicit and select

medium.com/@joybonaguro/part-1-how-to-solicit-and-select-data-science-projects-cc560e4c6aeb

Become a member Sign in Get started

soliciters_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 Follow
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*

Statewide Data Goals

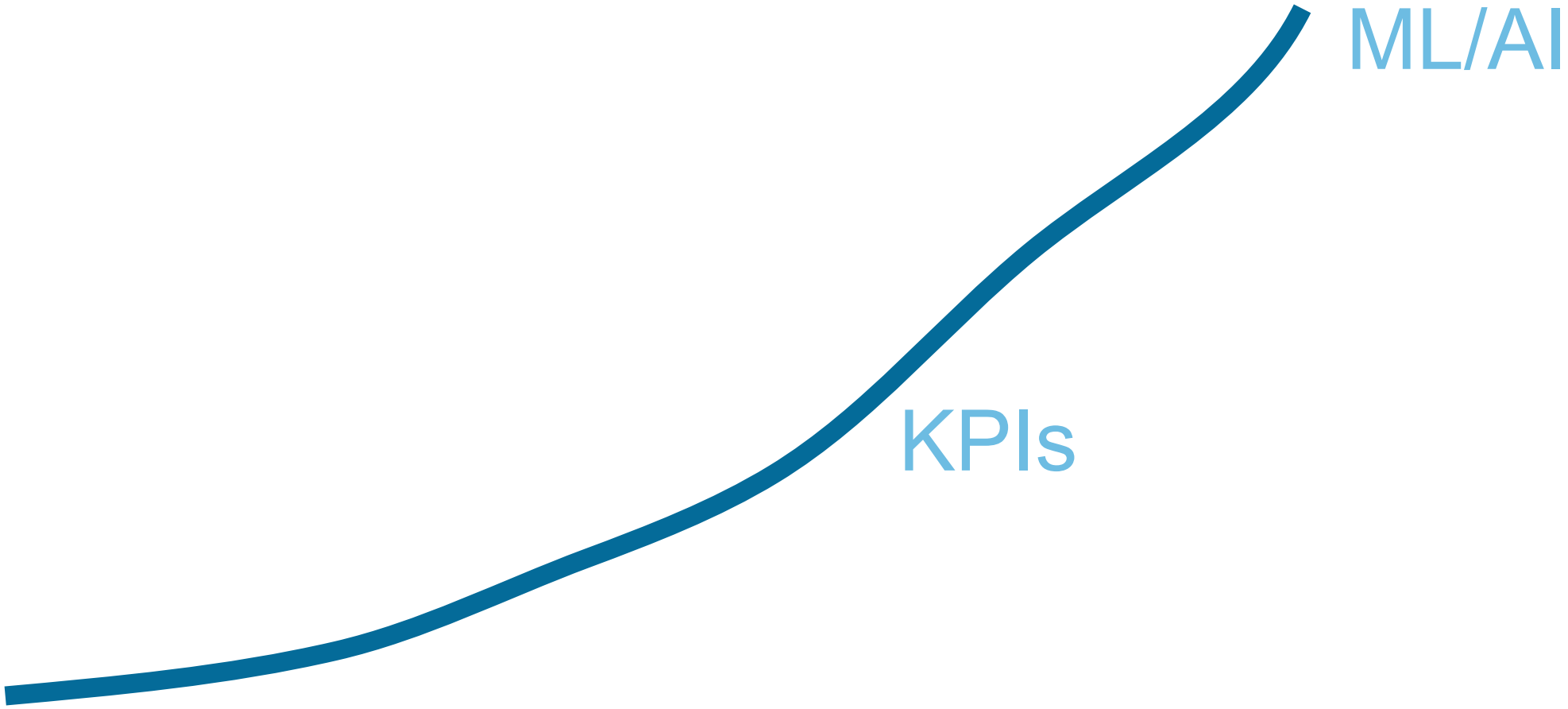
Equipping ourselves to navigate the data landscape

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/