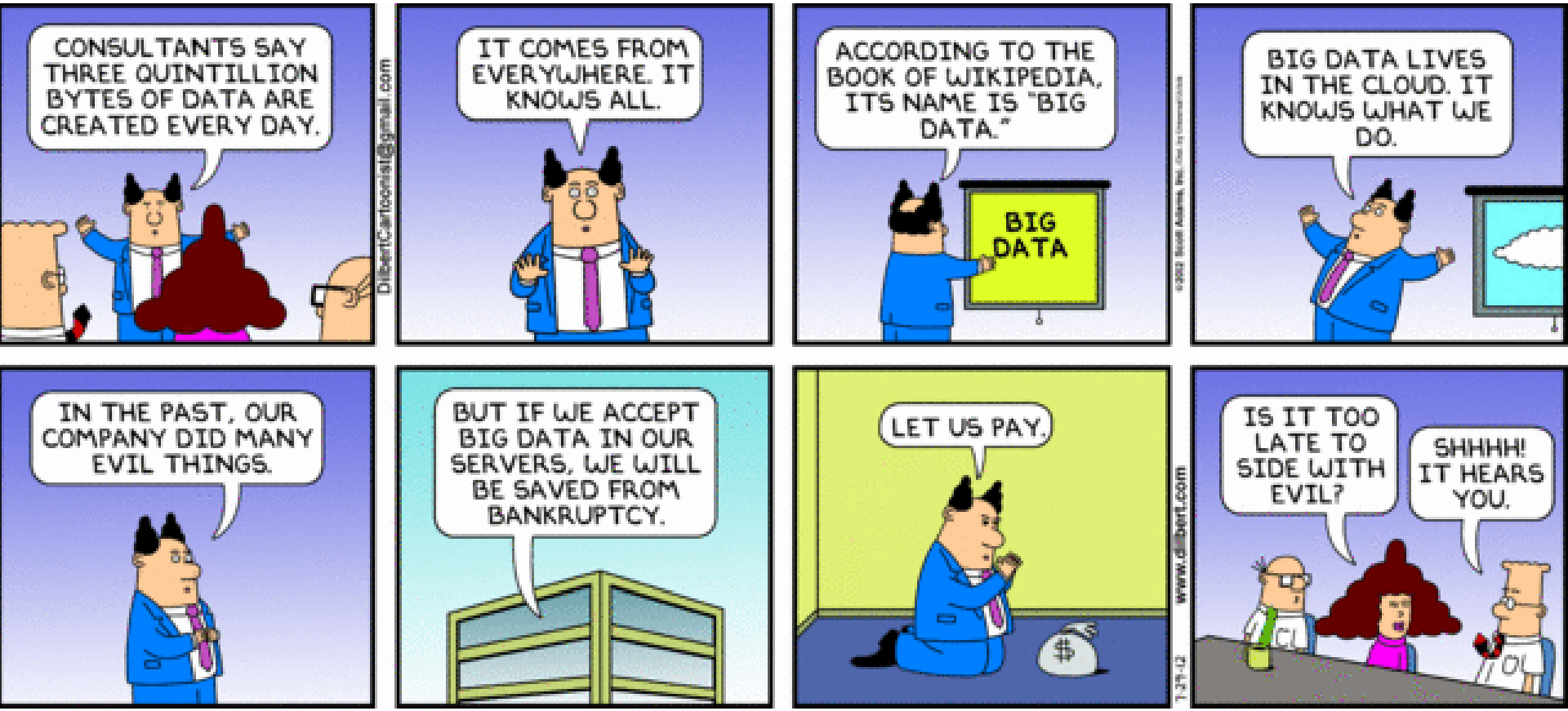
Starting Small with Big Data

Lessons in turning data into action for a smarter grid

Matthew Gee
Open Energy Efficiency
Center for Data Science and Public Policy
ACEEE-IE 2015


CONSULTANTS SAY THREE QUINTILLION BYTES OF DATA ARE CREATED EVERY DAY.
IT COMES FROM EVERYWHERE. IT KNOWS ALL.
ACCORDING TO THE BOOK OF WIKIPEDIA, ITS NAME IS "BIG DATA."
BIG DATA
BIG DATA LIVES IN THE CLOUD. IT KNOWS WHAT WE DO.
IN THE PAST, OUR COMPANY DID MANY EVIL THINGS.
BUT IF WE ACCEPT BIG DATA IN OUR SERVERS, WE WILL BE SAVED FROM BANKRUPTCY.
LET US PAY.
IS IT TOO LATE TO SIDE WITH EVIL?
SHHHH! IT HEARS YOU.
DilbertCartoonist@gmail.com
www.dilbert.com

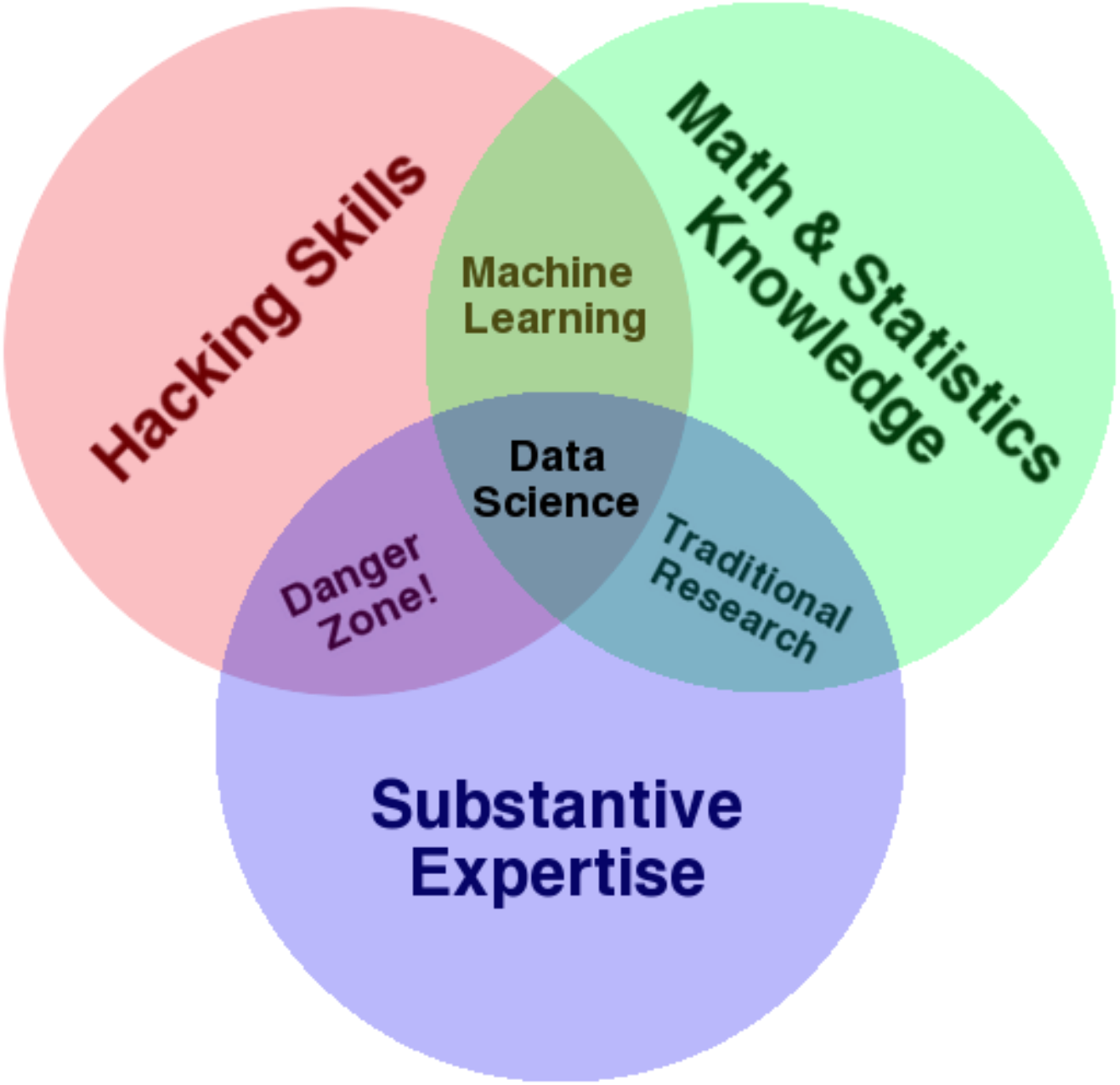

Think
BIG
Start
small

Goals

Everyone leaves understanding:

1. How data can be used to dynamically understand & improve programs and products.
2. How three real-world use cases demonstrate the feasibility of integrated data platforms and custom analytics helping connect, collect, analyze, and visualize program data.
3. What the biggest challenges and road blocks you'll run into in trying to do the same thing.
4. Where you can go to get help overcoming them.

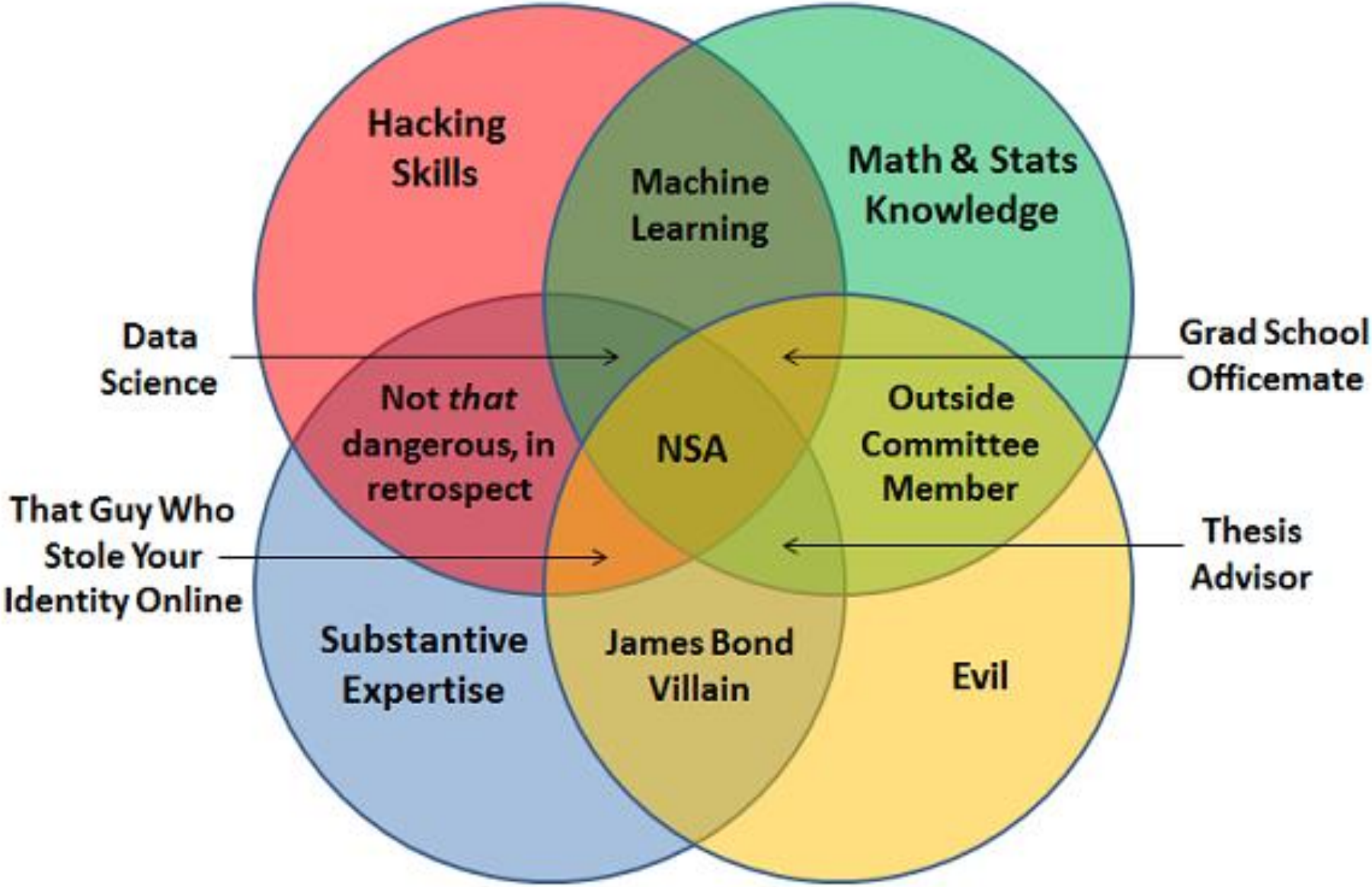

$\Delta x = x_f - x_i$
$\Delta V = V_f - V_i$
$E = mc^2$


Hacking Skills
Math & Statistics Knowledge
Substantive Expertise
Machine Learning
Data Science
Danger Zone!
Traditional Research

Google

facebook

NATIONAL SECURITY AGENCY
UNITED STATES OF AMERICA


Hacking Skills
Machine Learning
Math & Stats Knowledge
Data Science
Grad School Officemate
Not *that* dangerous, in retrospect
NSA
Outside Committee Member
That Guy Who Stole Your Identity Online
Thesis Advisor
Substantive Expertise
James Bond Villain
Evil


The Eric & Wendy Schmidt
Data Science for Social Good
Summer Fellowship 2013
THE UNIVERSITY OF CHICAGO
Argonne NATIONAL LABORATORY
THE HARRIS SCHOOL
PUBLIC POLICY | THE UNIVERSITY OF CHICAGO

300+ Social Sector Organizations

From Data To Action
The Human Process of Data Science

Identify a Target Action

Define the Critical Question

Understand Available Data

Select Appropriate Methods

Choose Toolkit for prototyping

Four Typical Data Science Tasks

Description

- What patterns are there in local business hiring in the three zip codes my organization servers?
- What individual and community attributes relate most strongly with the outcomes I care about?
- Are there patterns in 311 call data that I can use to understand community needs?
- What are the patterns in power quality throughout the grid?
- What are the most important variables in my big dataset?
- What groups exist in the data?

Prediction

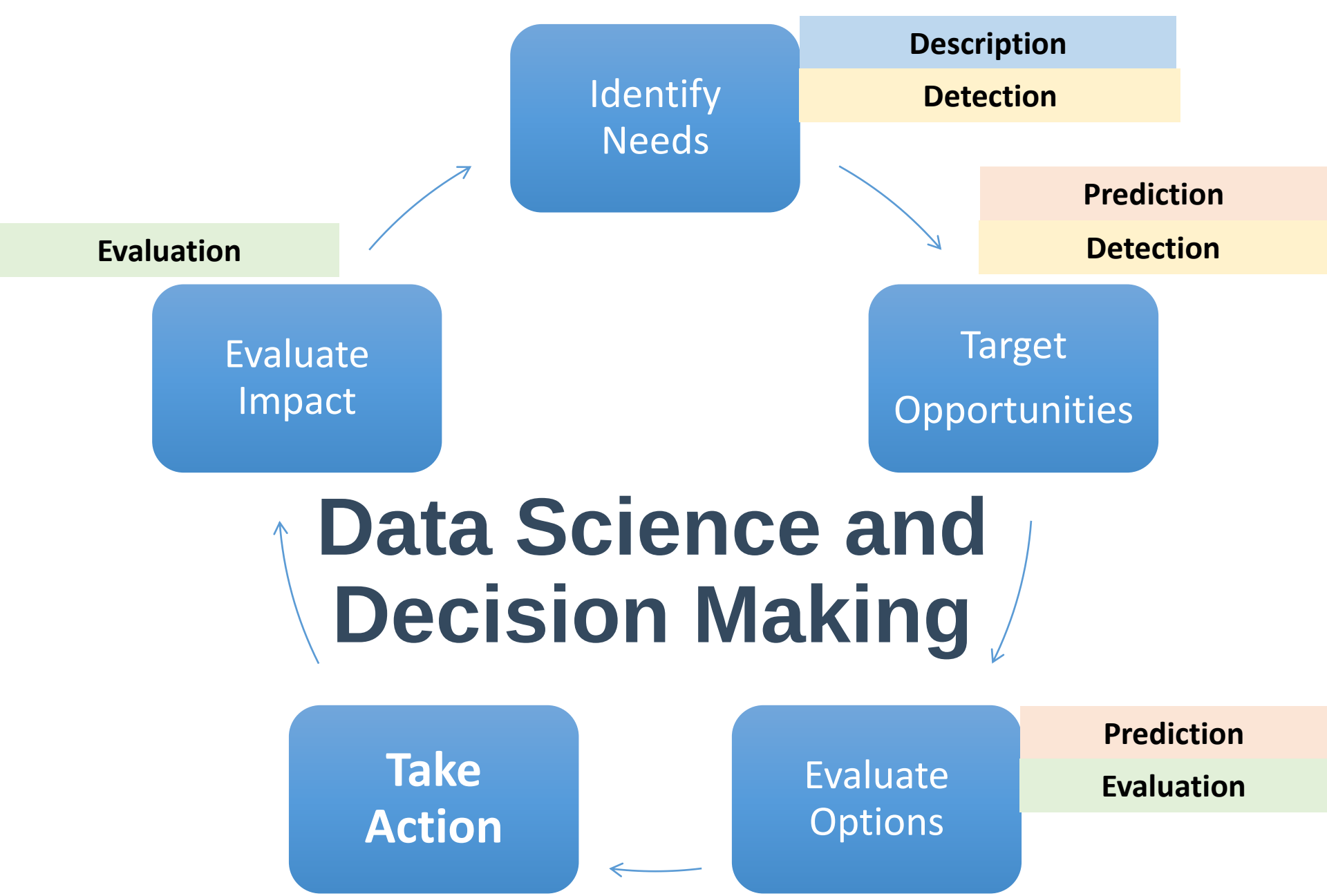
- Who is most at risk in the population I serve?
- Which homes are most likely to have lead in them?
- Which of my students are most likely to drop out of school next year?
- Which buildings have the biggest energy efficiency potential?
- Which patient is most likely to have a heart attach in the next 3 days?
- Which group does this thing belong to?
- Can I predict a number that I care about?

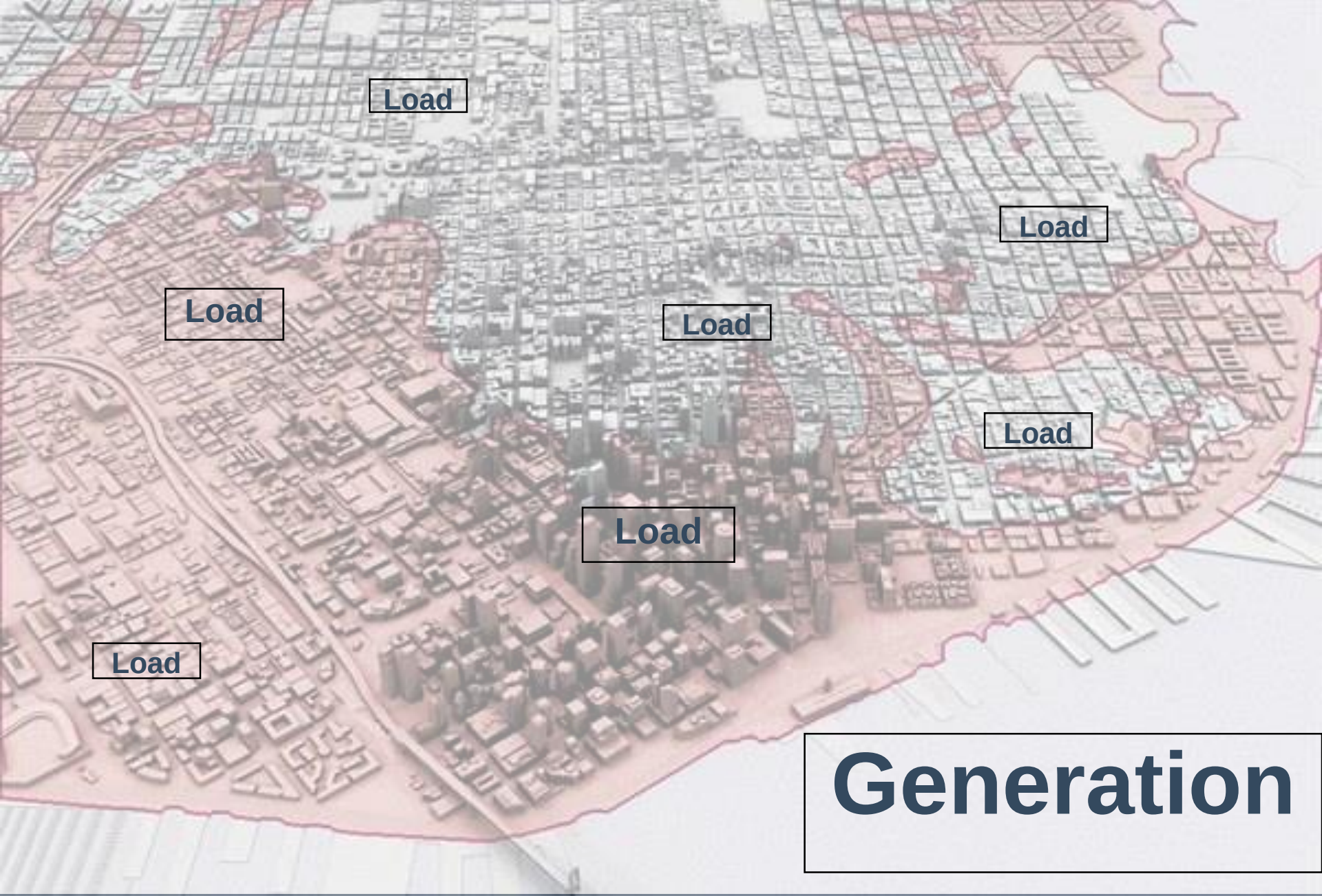
Detection

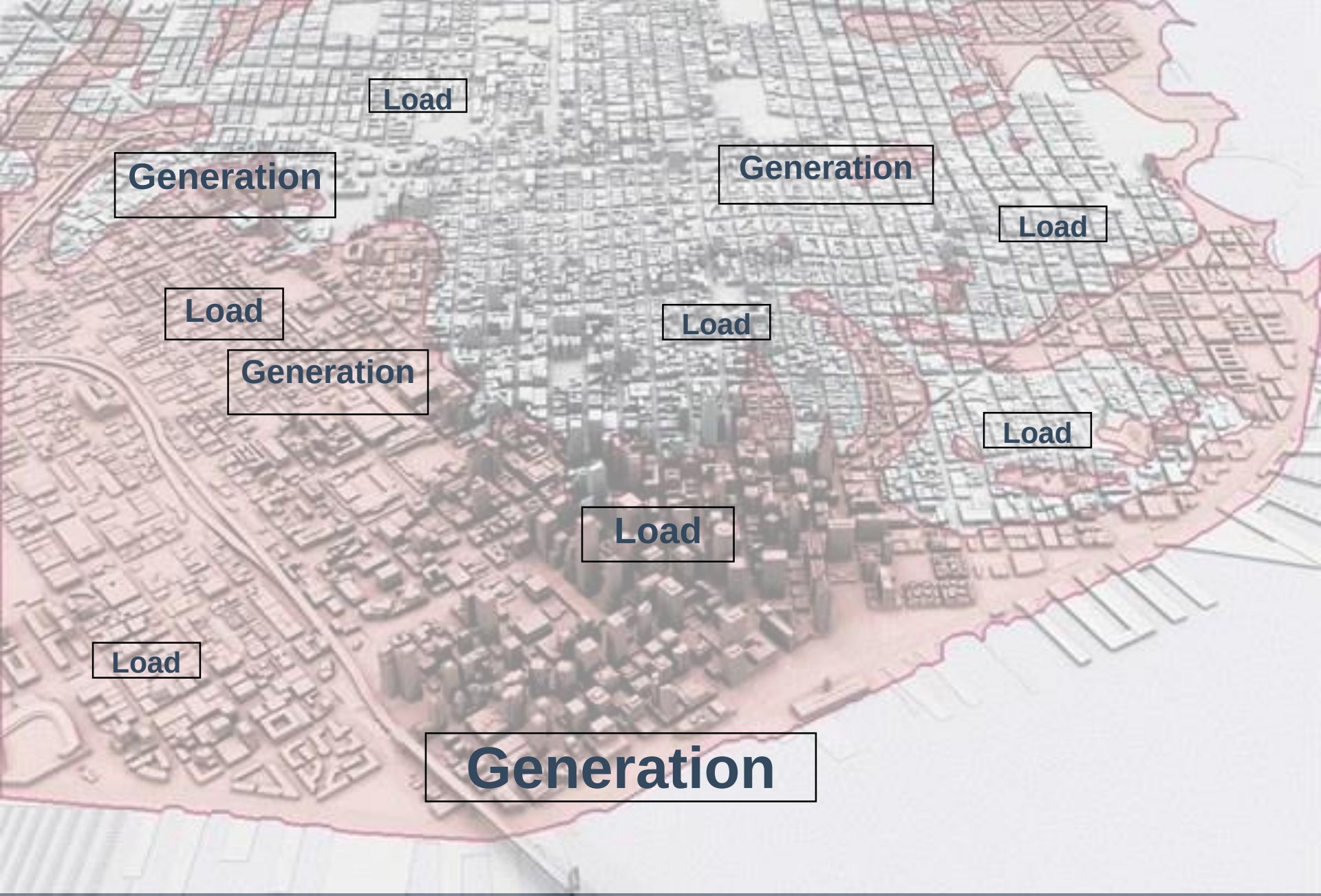
- What are the most important variables in my big dataset?
- When is one of my students falling below trend?
- Where are there unexpected changes in use that might indicate my community is struggling?
- Where and when is something unexpected happening?
- How can I find the needle in the haystack?

Dynamic Evaluation

- Which information intervention works best in my community?
- Which educational campaigns drive adoption of preventative health practices?
- What will happen if I change the housing subsidy in my program?
- What survey format is most effective in getting people to respond?
- Which actions or interventions work and which ones should I try next?


Identify Needs
Description
Detection
Target Opportunities
Prediction
Detection
Evaluate Options
Prediction
Evaluation
Take Action
Evaluate Impact
Evaluation
Data Science and Decision Making


Load
Load
Load
Load
Load
Load
Load
Generation


Load
Generation
Generation
Load
Load
Load
Generation
Load
Load
Load
Generation

From this

PG&E

Southern California Gas Company

SOUTHERN CALIFORNIA EDISON
An *EDISON INTERNATIONAL* Company

SDGE

=

To this


PG&E


SolarCity


SUNRUN


Southern
California
Gas Company


SOUTHERN CALIFORNIA
EDISON
An EDISON INTERNATIONAL Company


SDGE

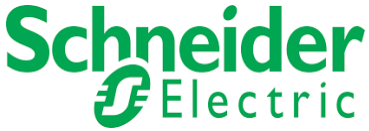

Bloomenergy


nrg

=


Johnson
Controls


AGENTIS


Schneider
Electric


nest


RENOVATEAMERICA


RENEWABLE FUNDING


wink


bidgely

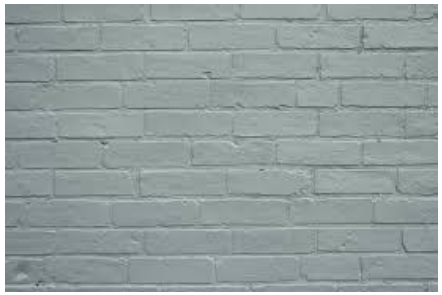

SEALED


OPOWER


TESLA

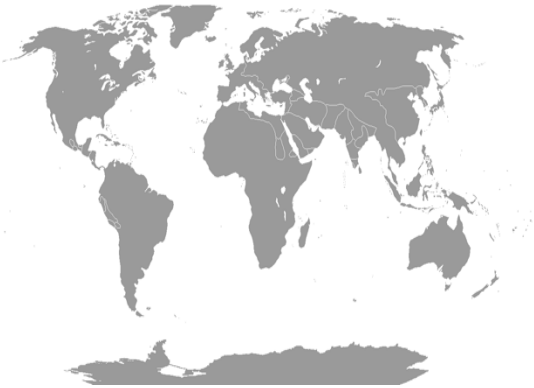

SEAL OF THE CITY AND COUNTY OF SAN FRANCISCO

What is Hard About Grid-scale Analytics



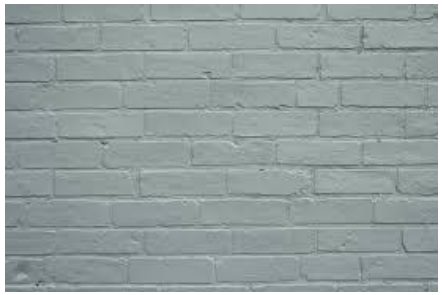
Data Is Siloed

Time Matters



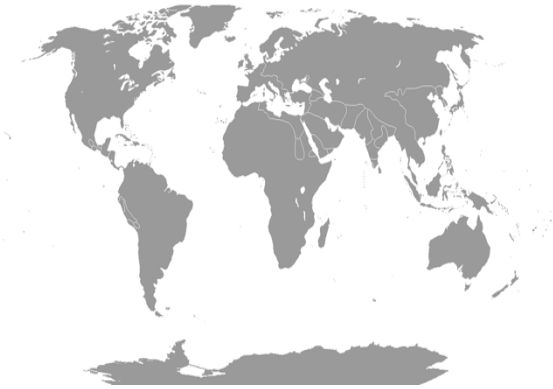
Place Matters

What Data Science is Good At



Integrated

Real-time



Locally Relevant

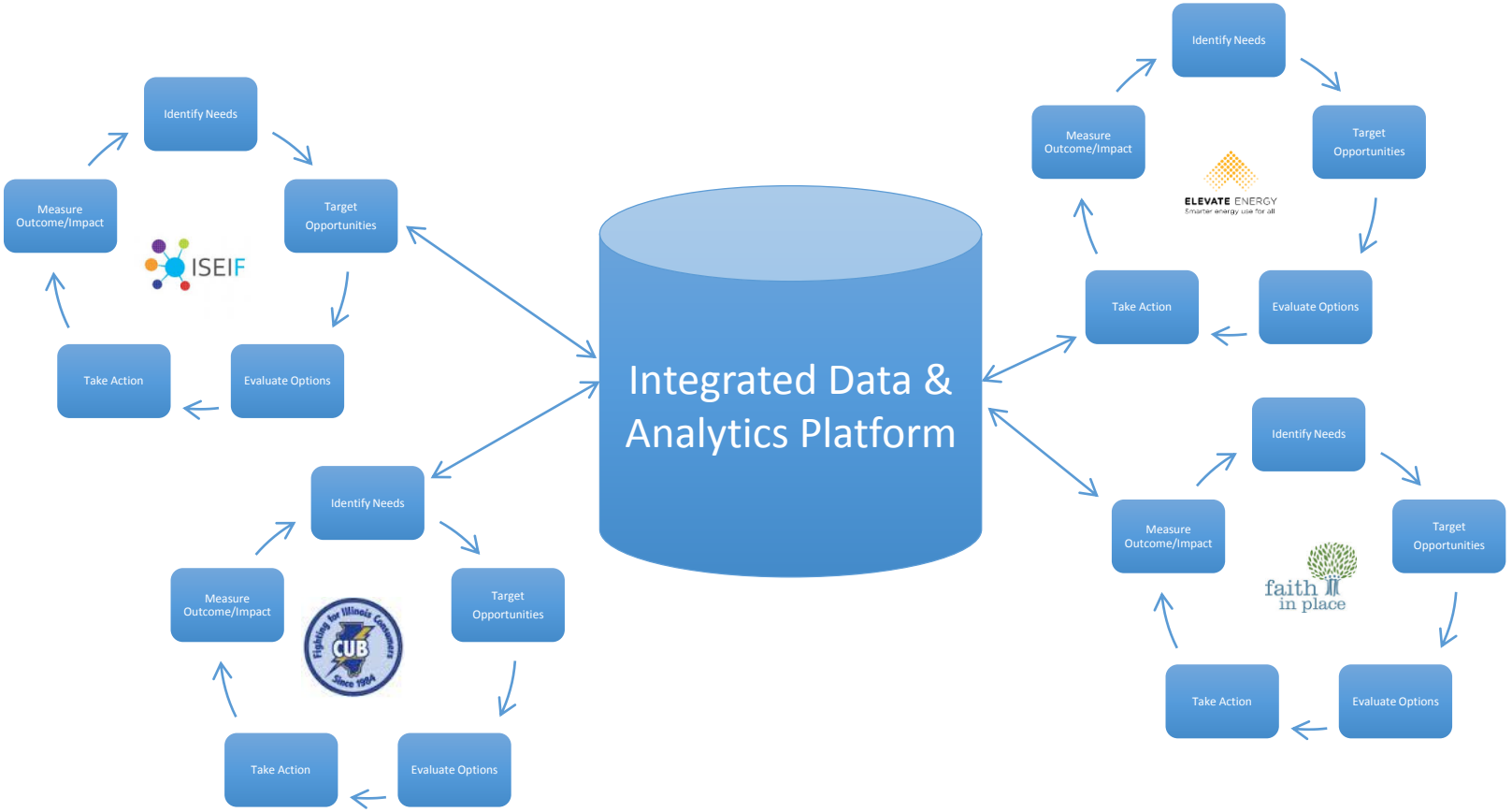
Big Use Case

Small Steps

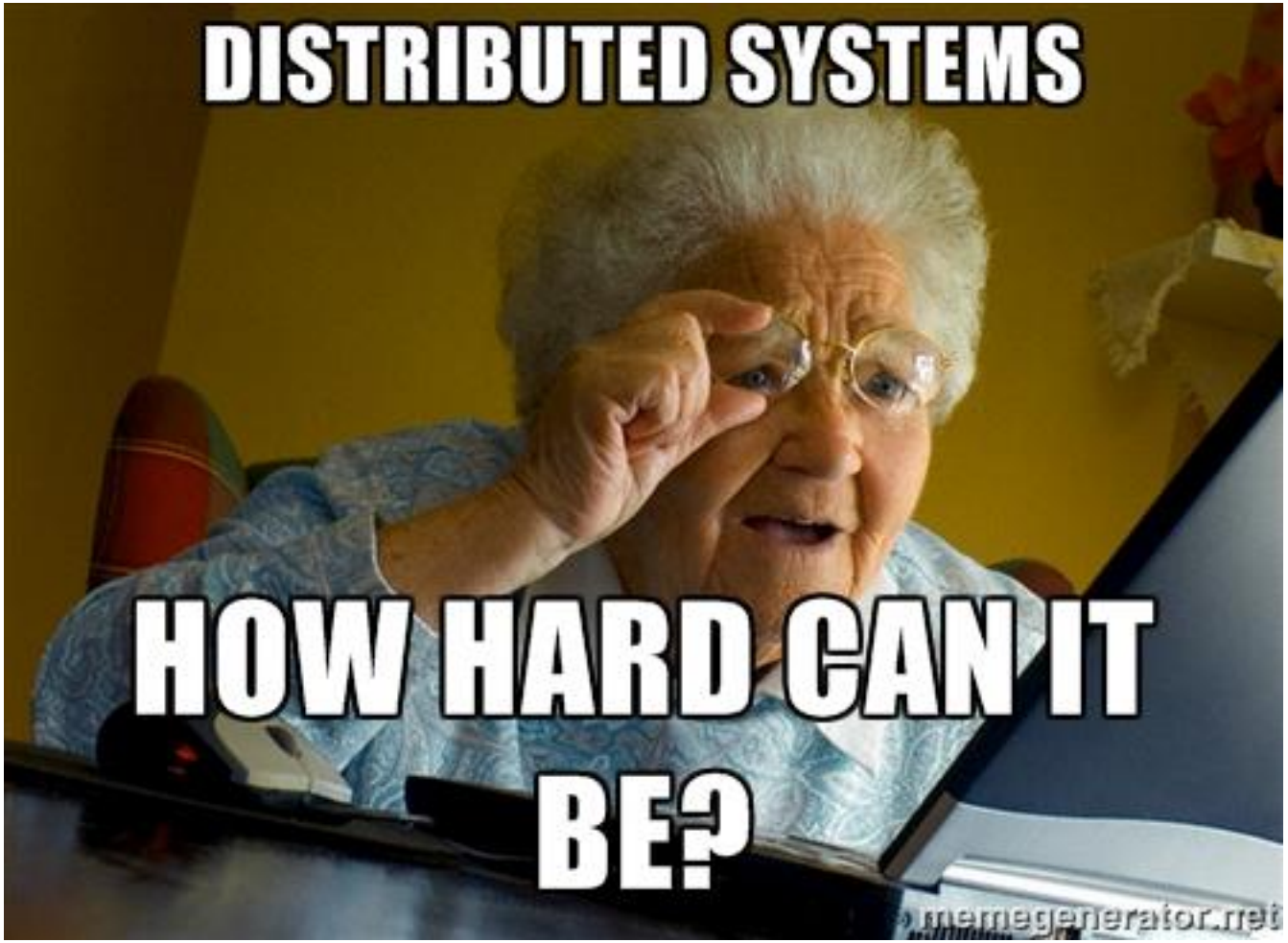
Distributed Decision Making

Shared Pool of Information for Distributed Coordination

Distributed Decision Making



Centralized Feedback


DISTRIBUTED SYSTEMS
HOW HARD CAN IT BE?
memegenerator.net


YOU CAN
DO IT!!
Troll.me

CalTRACK

Project Drivers

Stakeholder Feedback:

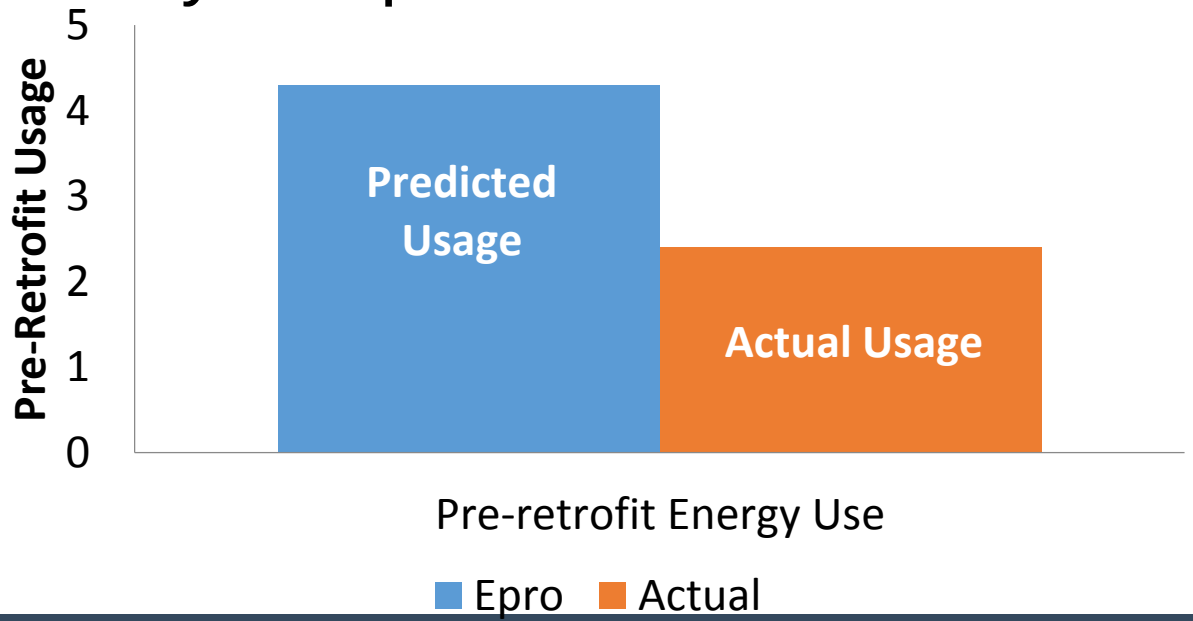
- Concerns about accuracy of energy savings predictions and how it impacts customer decision-making
- Contractors concerned about complexity of software and requested tools that model quickly and facilitate the job sales process

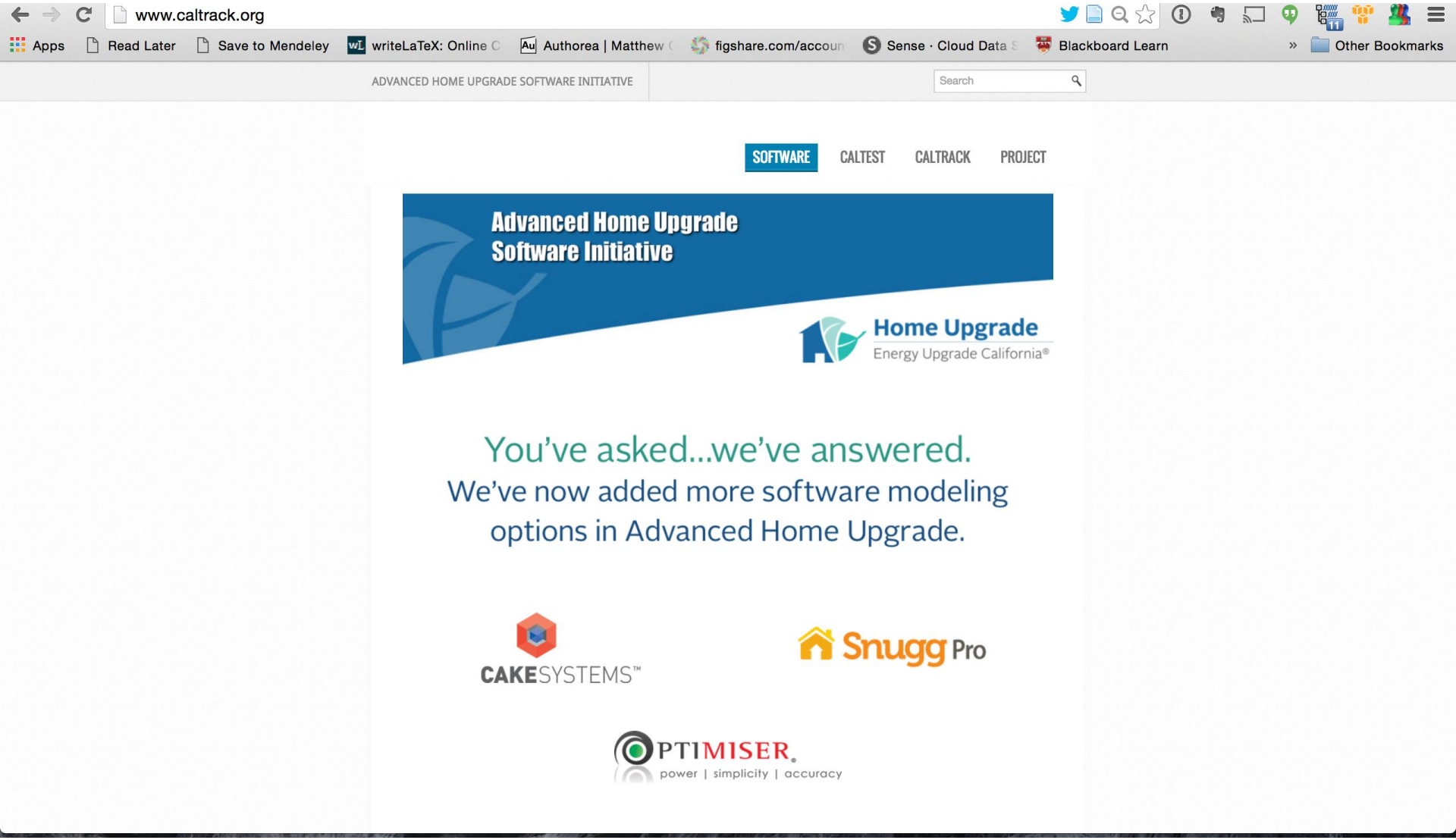
CPUC Direction

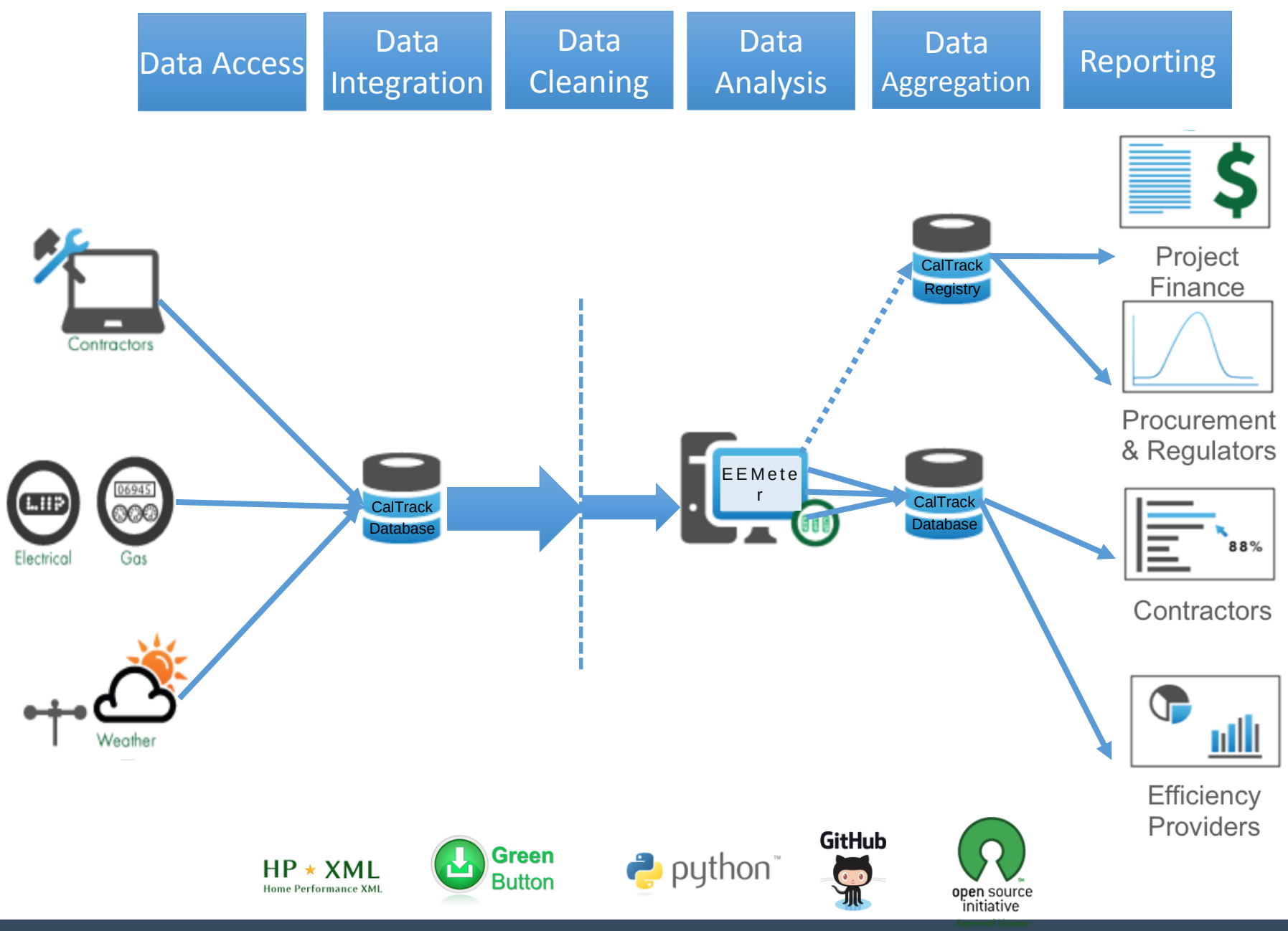
- “We direct Commission Staff and the IOUs to work collaboratively with the California Energy Commission and other Energy Upgrade California stakeholders to identify approaches to adequately **broaden allowable software under Energy Upgrade California** while containing costs required for needed Commission Staff Reviews.” (OP 61 D12-05-015)

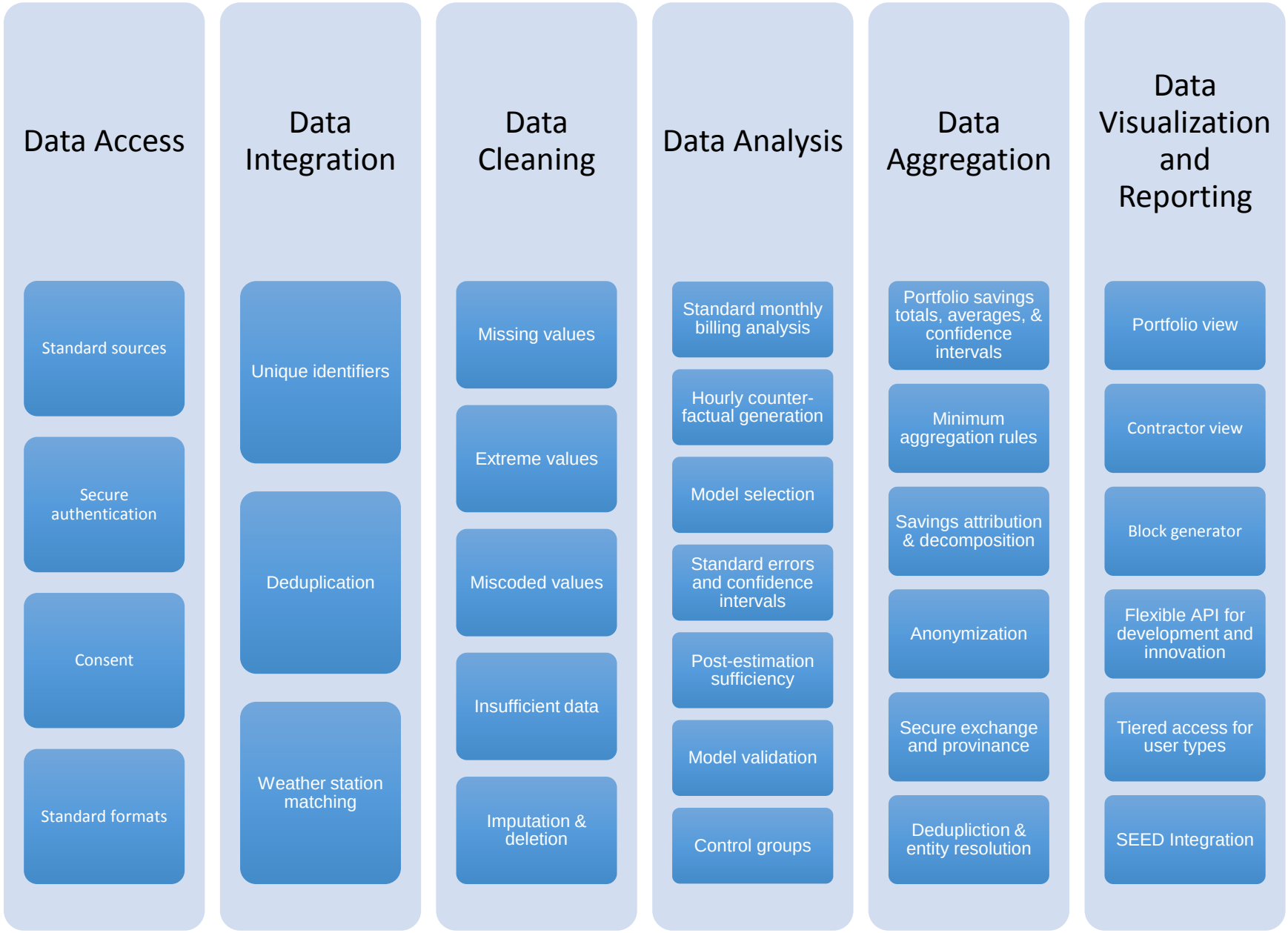
Existing Modeling Tool Over Predicts Savings

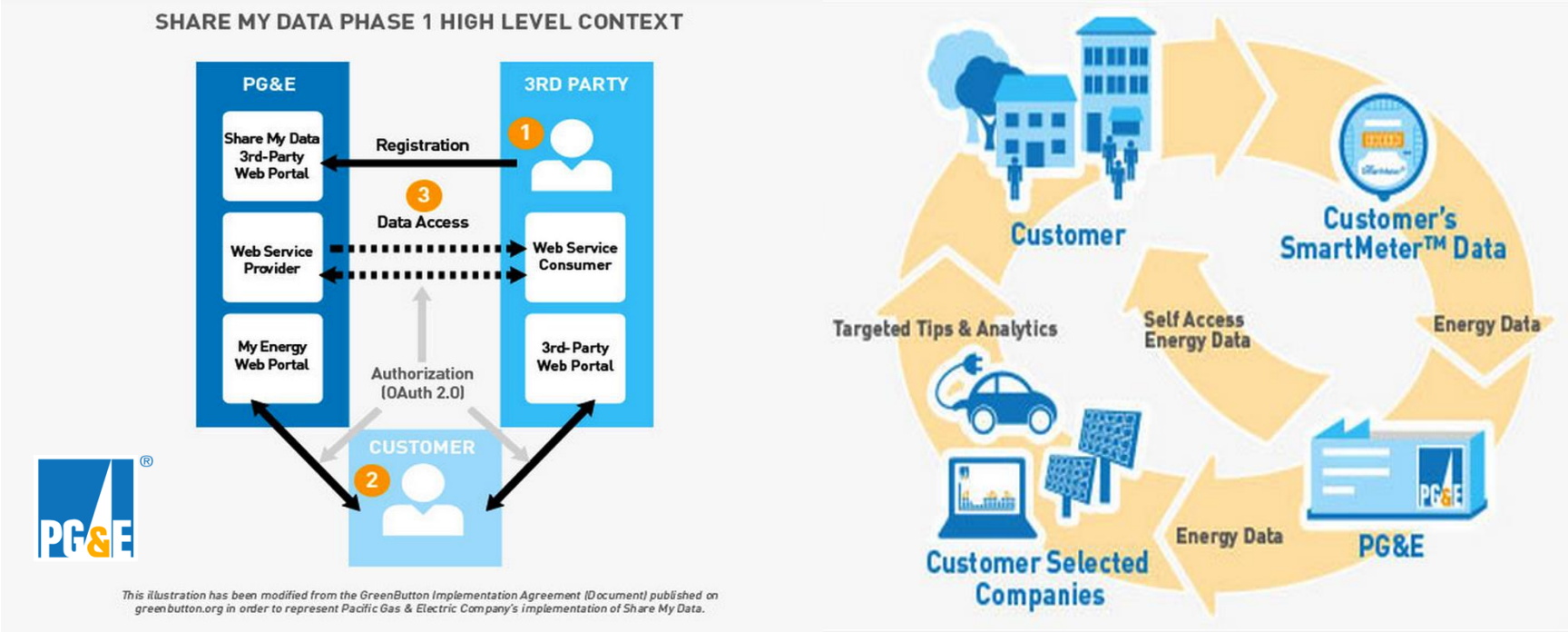
- Uncalibrated EnergyPro modeling over predicts baseline energy usage by 40-60%, leading to inflated projected savings
- Calibration is difficult and time-consuming in already complex sales environment




www.caltrack.org
Apps
Read Later
Save to Mendeley
writeLaTeX: Online C
Authorea | Matthew C
figshare.com/accoun
Sense · Cloud Data S
Blackboard Learn
Other Bookmarks
ADVANCED HOME UPGRADE SOFTWARE INITIATIVE
Search
SOFTWARE
CALTEST
CALTRACK
PROJECT
Advanced Home Upgrade
Software Initiative
Home Upgrade
Energy Upgrade California®
You've asked...we've answered.
We've now added more software modeling
options in Advanced Home Upgrade.
CAKESYSTEMS™
Snugg Pro
OPTIMISER
power | simplicity | accuracy


Data Access
Data Integration
Data Cleaning
Data Analysis
Data Aggregation
Reporting
Contractors
Electrical
Gas
Weather
CalTrack Database
EEMeter
CalTrack Registry
CalTrack Database
Project Finance
Procurement & Regulators
88%
Contractors
Efficiency Providers
HP XML
Home Performance XML
Green Button
python
GitHub
open source initiative


Data Access
Standard sources
Secure authentication
Consent
Standard formats
Data Integration
Unique identifiers
Deduplication
Weather station matching
Data Cleaning
Missing values
Extreme values
Miscoded values
Insufficient data
Imputation & deletion
Data Analysis
Standard monthly billing analysis
Hourly counter-factual generation
Model selection
Standard errors and confidence intervals
Post-estimation sufficiency
Model validation
Control groups
Data Aggregation
Portfolio savings totals, averages, & confidence intervals
Minimum aggregation rules
Savings attribution & decomposition
Anonymization
Secure exchange and provinance
Dedupliction & entity resolution
Data Visualization and Reporting
Portfolio view
Contractor view
Block generator
Flexible API for development and innovation
Tiered access for user types
SEED Integration


SHARE MY DATA PHASE 1 HIGH LEVEL CONTEXT
PG&E
3RD PARTY
Share My Data 3rd-Party Web Portal
Registration
1
3
Data Access
Web Service Provider
Web Service Consumer
My Energy Web Portal
3rd-Party Web Portal
Authorization (OAuth 2.0)
CUSTOMER
2
PG&E
This illustration has been modified from the GreenButton Implementation Agreement [Document] published on greenbutton.org in order to represent Pacific Gas & Electric Company's implementation of Share My Data.
Customer
Customer's SmartMeter™ Data
Energy Data
PG&E
Energy Data
Customer Selected Companies
Targeted Tips & Analytics
Self Access Energy Data

Openeemeter.org

This repository Search Explore Gist Blog Help goldenmatt

datamade / open-ee-meter

Unwatch 5 Unstar 2 Fork 1

branch: master open-ee-meter / data / processors / contractor_data_prep.py

cathydeng on Dec 22, 2014 add gross savings data prep for contractors, add contractor layout, c…

1 contributor

253 lines (211 sloc) 10.881 kb Raw Blame History

```
import pandas as pd
from pandas import to_datetime
import numpy
import json
import re
import os

projects = pd.read_csv( "build/merged.csv")
loc = pd.read_csv( "build/latlong_clean.csv")

merged = projects.merge(loc, on="zipcode")

contractor_dict = {
    'electricity_iou': {
        'contractor_names' : ['Contractor 12'],
        'actual_col' : 'weather_normalized_yearly_kwh_savings',
        'pred_col': 'predicted_yearly_kwh_savings',
        'hist_chunks' : [float(i)/2 for i in range(-8, 9)]
        },
    'gas_iou': {
        'contractor_names' : ['Contractor 12'],
        'actual_col' : 'weather_normalized_yearly_therm_savings',
        'pred_col': 'predicted_yearly_therm_savings',
```

THE UNIVERSITY OF CHICAGO

IMPACT LAB

Utility 4

Total Projects: 22

Total Contractors: 7

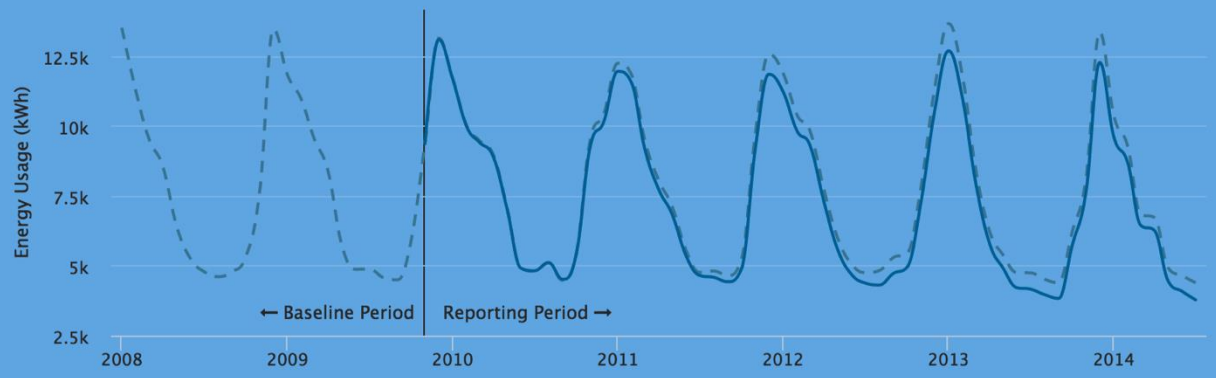
ELECTRICITY GROSS SAVINGS

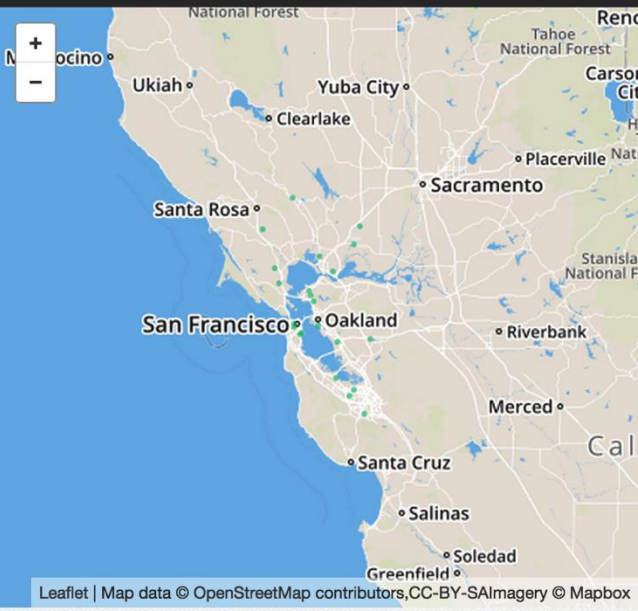
21,219 kWh

GAS GROSS SAVINGS

29,198 therm

Total Energy Usage Over Time





Electricity Savings - Block

REALIZATION RATE

85 %

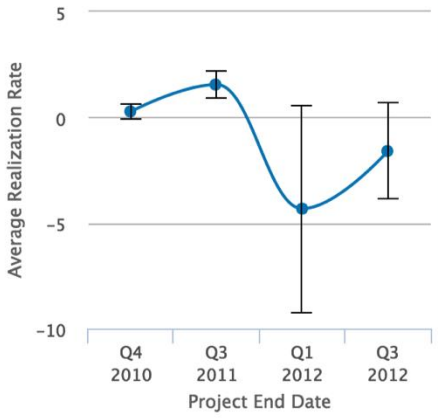
ACTUAL SAVINGS

21,219 kWh

PREDICTED SAVINGS

29,527 kWh

Average Electricity Realization Rate



Gas Savings - Block

REALIZATION RATE

62 %

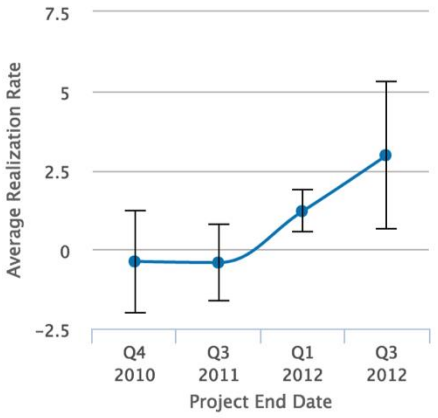
ACTUAL SAVINGS

29,198 therms

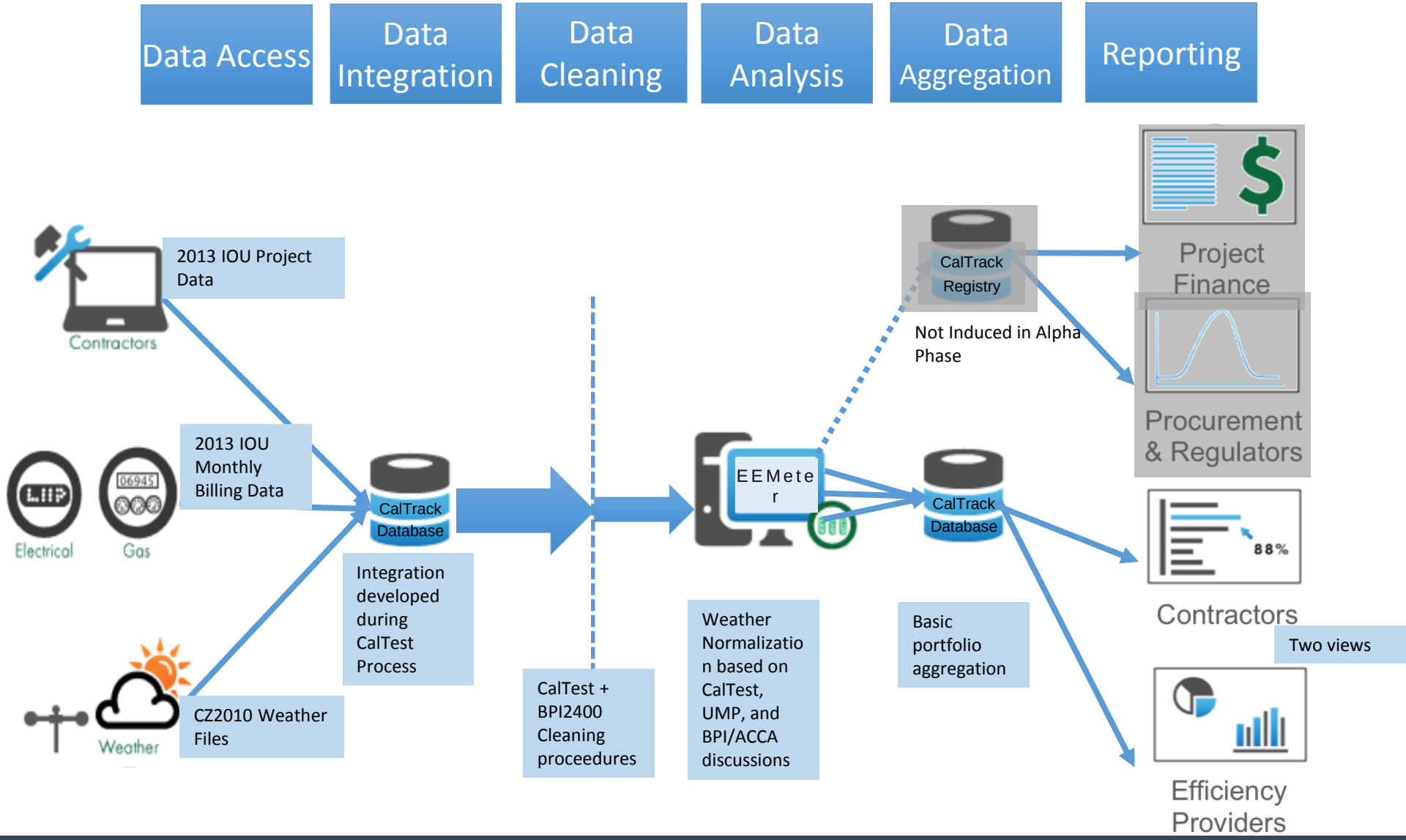
PREDICTED SAVINGS

44,411 therms

Average Gas Realization Rate



CalTrack Pilot: Alpha Version



Supporting New Regulations and Powering New Business Models

Current Regulation and Laws:

- EPA Clean Power Plan
- CA AB-32
- NY REV
- SB350 Goals increase EE in CA by 50%

Private Market Growth:

- Residential PACE in CA will hit $500M in EE in 2015
- WHEEL Project, first EE Loan Securitization
- Home Energy Management

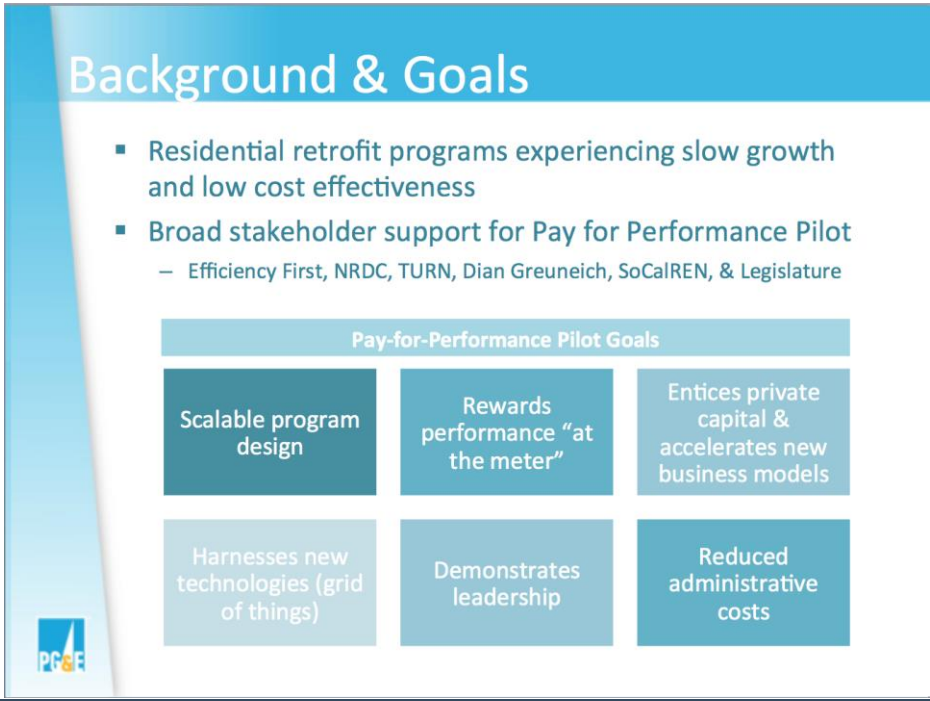
Shifting to Pay-4-Metered Performance

SB-350 / AB-802, signed into CA law in Oct 2015

- Increase of CA EE goals by 50%
- Redefines EE as normalized metered performance
- Removes regulatory barriers (code baseline, behavior, etc)
- Requires CPUC to run P4P pilots
- Implementation starts January 1st 2016

PG&E will launch Pay-4-Metered Performance Pilots for 2016:

- Open markets
- Savings based on EE Meter
- Pay for results



Paid on prediction in advance

Rebate based on modeled or deemed savings

Regulate progress through program QA/QC and questionable cost tests

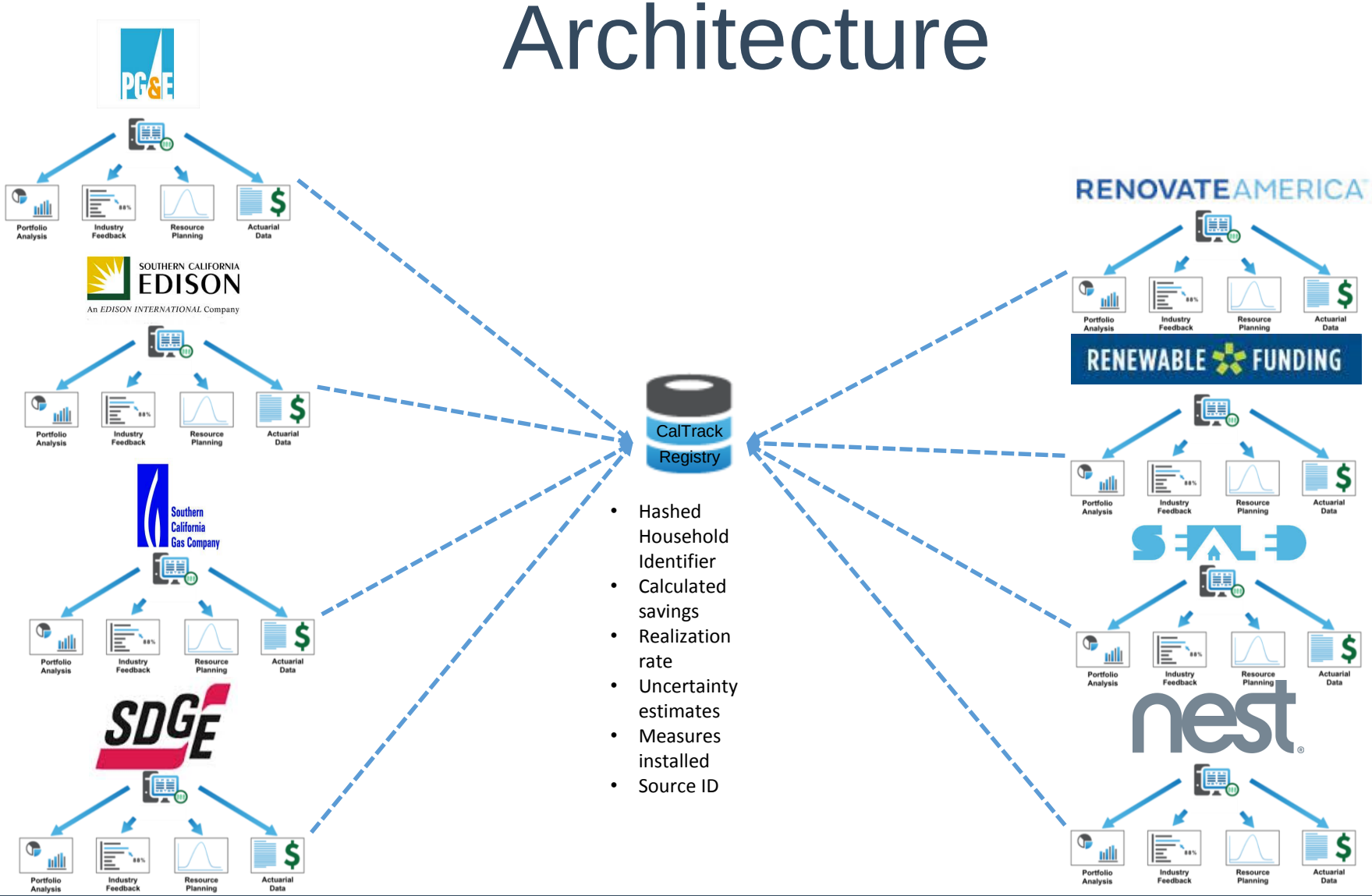
Paid on performance as delivered

Savings calculated at the meter

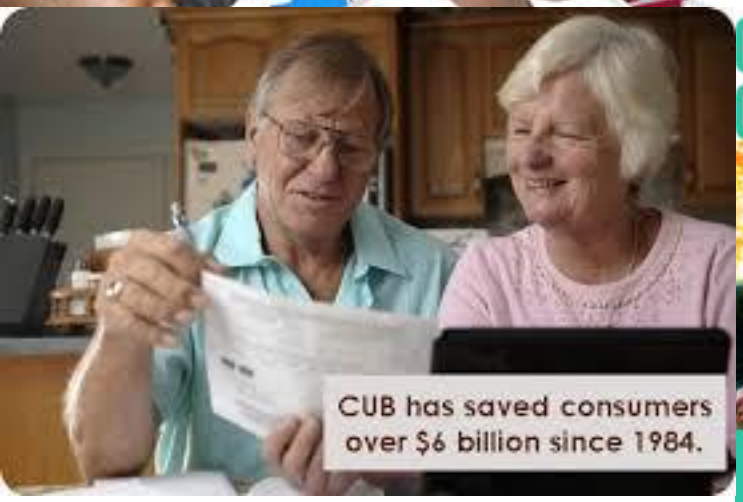
Align performance risk with the market

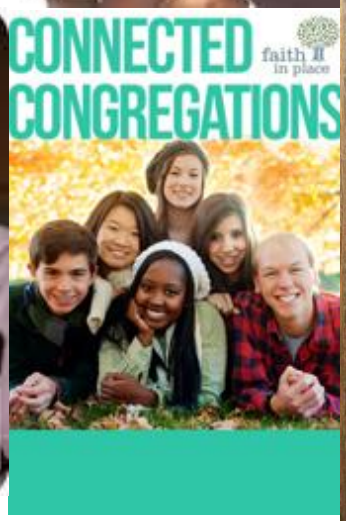
Scalable market-based procurement of EE as a resource, yielding competitive private markets innovating to deliver the most cost efficient and reliable energy savings.

CalTrack's Distributed Architecture



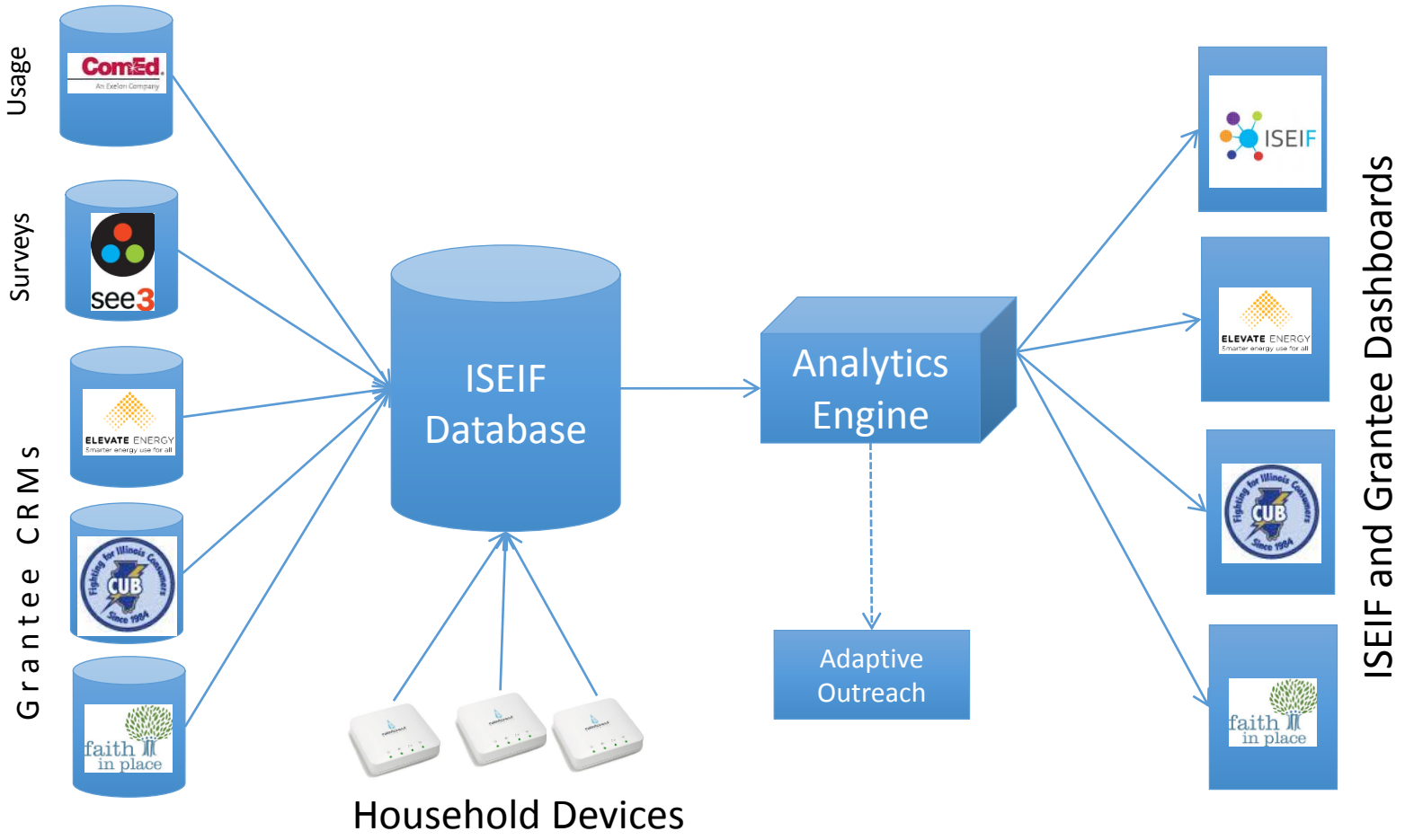

ISEIF


CUB has saved consumers
over $6 billion since 1984.


CONNECTED
CONGREGATIONS
faith in place


Time of Use
Agenda
Elevate Energy- who are we?
Energy Impact Illinois
Smart Grid + Meters
Next steps + Questions

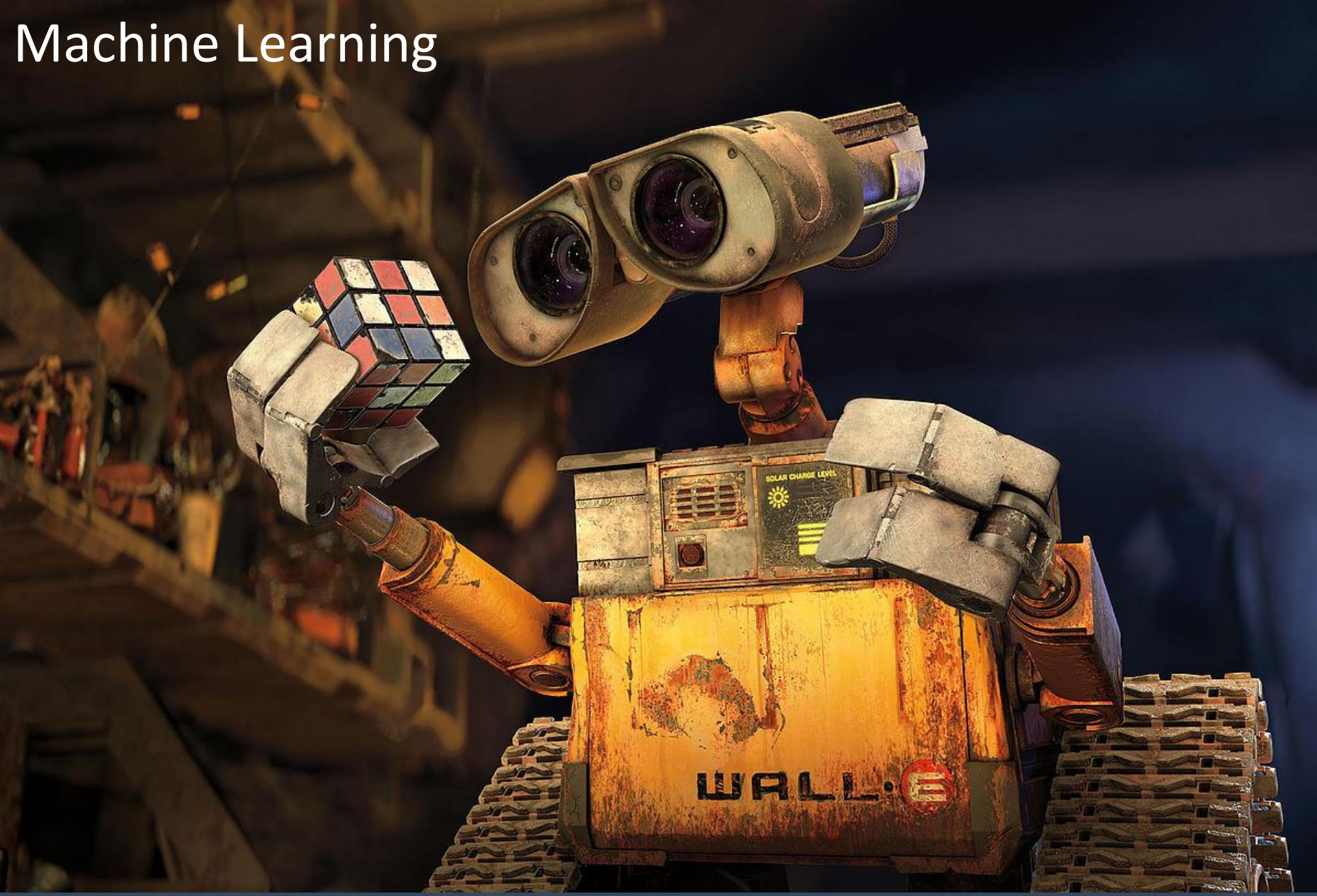
ISEIF System for the Analysis and Integration of Data (ISEIF-SAID)



Data Sources

- Surveys
  - Grantee Participant Surveys
  - See3 Participant Surveys
- Grantee Programmatic Data
  - Dates, times, and participant lists of events
  - Participant demographics
  - Text-based summaries and impressions of events
- Customer Usage Data
  - Monthly billing data
  - Hourly AMI data
  - 6-second interval data from hubs
- Community Data
  - Neighborhood demographics
  - Household locations (building footprints & voterfile)
  - Zip+4 load profiles
  - Smartgrid rollout shapefiles

Machine Learning



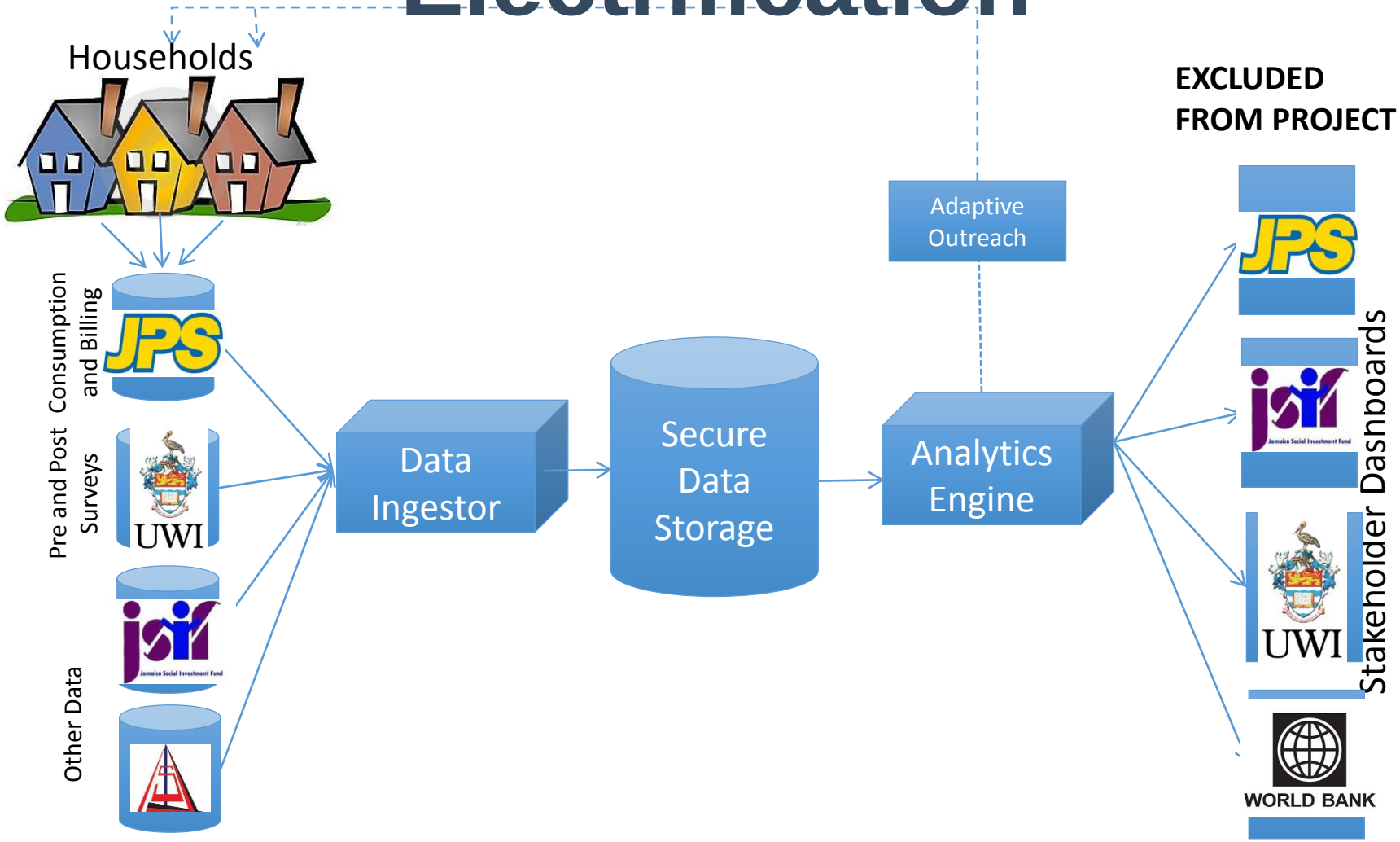
Models & Methods

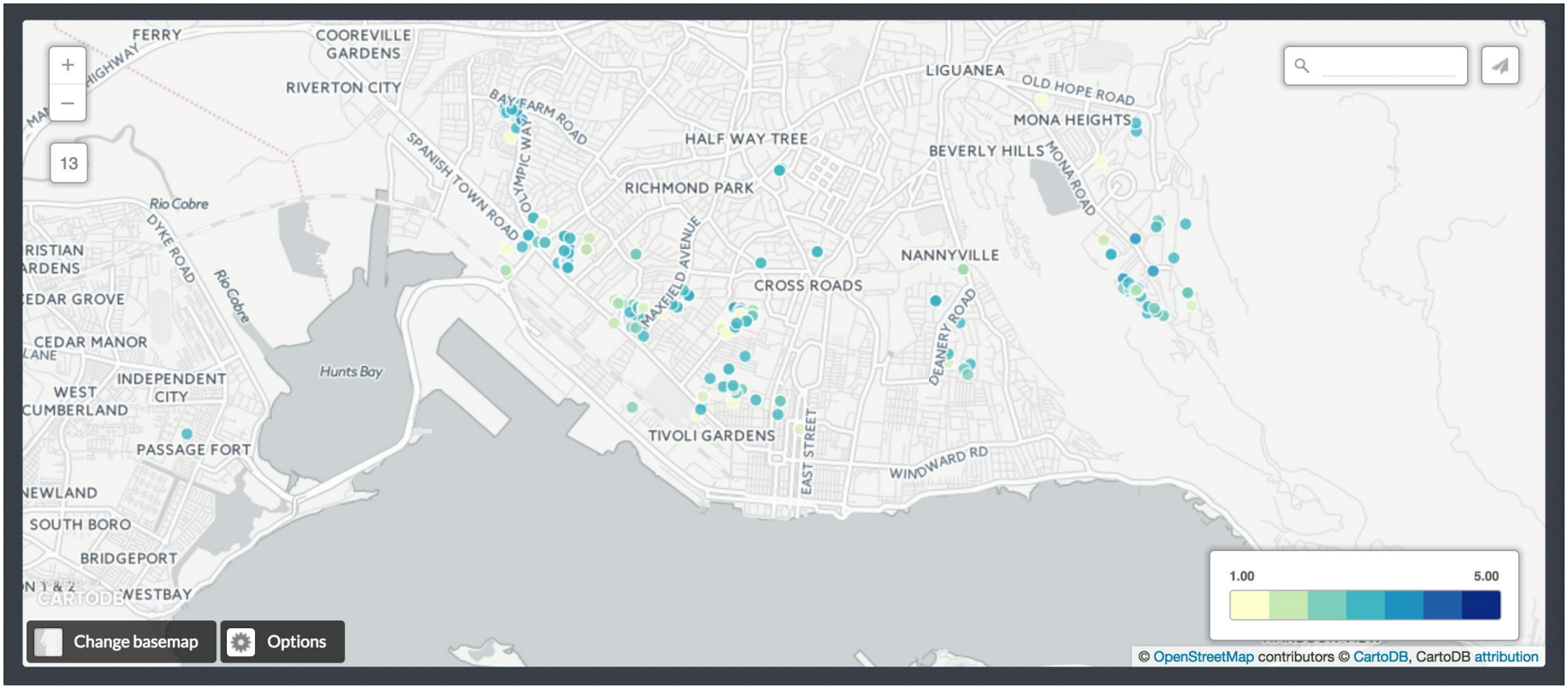
- Targeting
  - Contactability model
  - Persuadability model
- Enrollment
  - Intervention effects model
    - Nonparametric matching methods
  - Within-intervention attribute effects model
    - A/B testing
    - Multi-armed bandit
- Program Persistence
  - Survival analysis
- Program Effects
  - Energy usage effects model
    - OpenBaseline
    - Baysian structural time series
- Network effects
- Neighborhood effects
  - Neighborhood load profile model
    - Difference in difference

Starting Simple
Knowing where to outreach is happening & should happen

Urban Electrification Programs in the Developing World

Jamaican Urban Electrification




FERRY
COOREVILLE GARDENS
RIVERTON CITY
BAY FARM ROAD
SPANISH TOWN ROAD
OLYMPIC WAY
HALF WAY TREE
RICHMOND PARK
LIGUANEA
OLD HOPE ROAD
MONA HEIGHTS
BEVERLY HILLS
MONA ROAD
NANNYVILLE
CROSS ROADS
MAXFIELD AVENUE
DEANERY ROAD
Rio Cobre
DYKE ROAD
Rio Cobre
CEDAR GROVE
CEDAR MANOR
INDEPENDENT CITY
WEST CUMBERLAND
Hunts Bay
PASSAGE FORT
NEWLAND
SOUTH BORO
BRIDGEPORT
WESTBAY
TIVOLI GARDENS
EAST STREET
WINDWARD RD
13
1.00
5.00
Change basemap
Options
© OpenStreetMap contributors © CartoDB, CartoDB attribution

In conclusion…

Big Data Doesn’t Have to Mean Big Budget

Identify a critical question, find key datasets, and gets started with the simplest (possibly free & open source) tools.

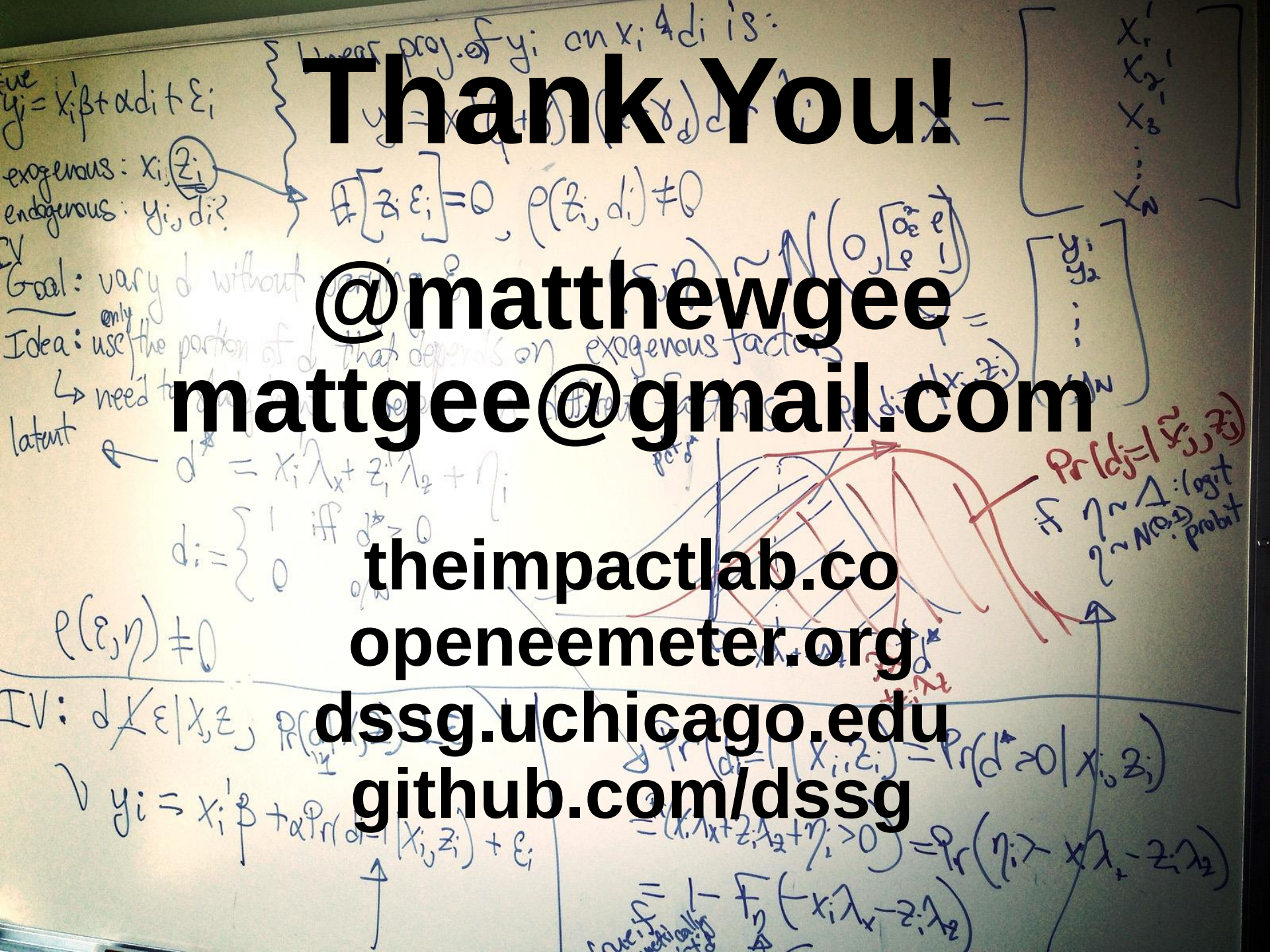
The critical skills for the big data future look like the essential skills of the little data past

- Ability to connect people and organizations
- Ability to build and lead cross-sector coalitions
- Ability to sell the big vision
- P-values: patience, persistence, perseverance
- Seeing the opportunities through the risks


The Eric & Wendy Schmidt
Data Science for Social Good
Summer Fellowship 2013
THE UNIVERSITY OF CHICAGO
Argonne NATIONAL LABORATORY
THE HARRIS SCHOOL
PUBLIC POLICY | THE UNIVERSITY OF CHICAGO

“Make no little plans; they have no magic to stir men’s blood and probably themselves will not be realized. Make big plans; aim high in hope and work.”

Daniel Burnham


Thank You!
@matthewgee
mattgee@gmail.com
theimpactlab.co
openeemeter.org
dssg.uchicago.edu
github.com/dssg