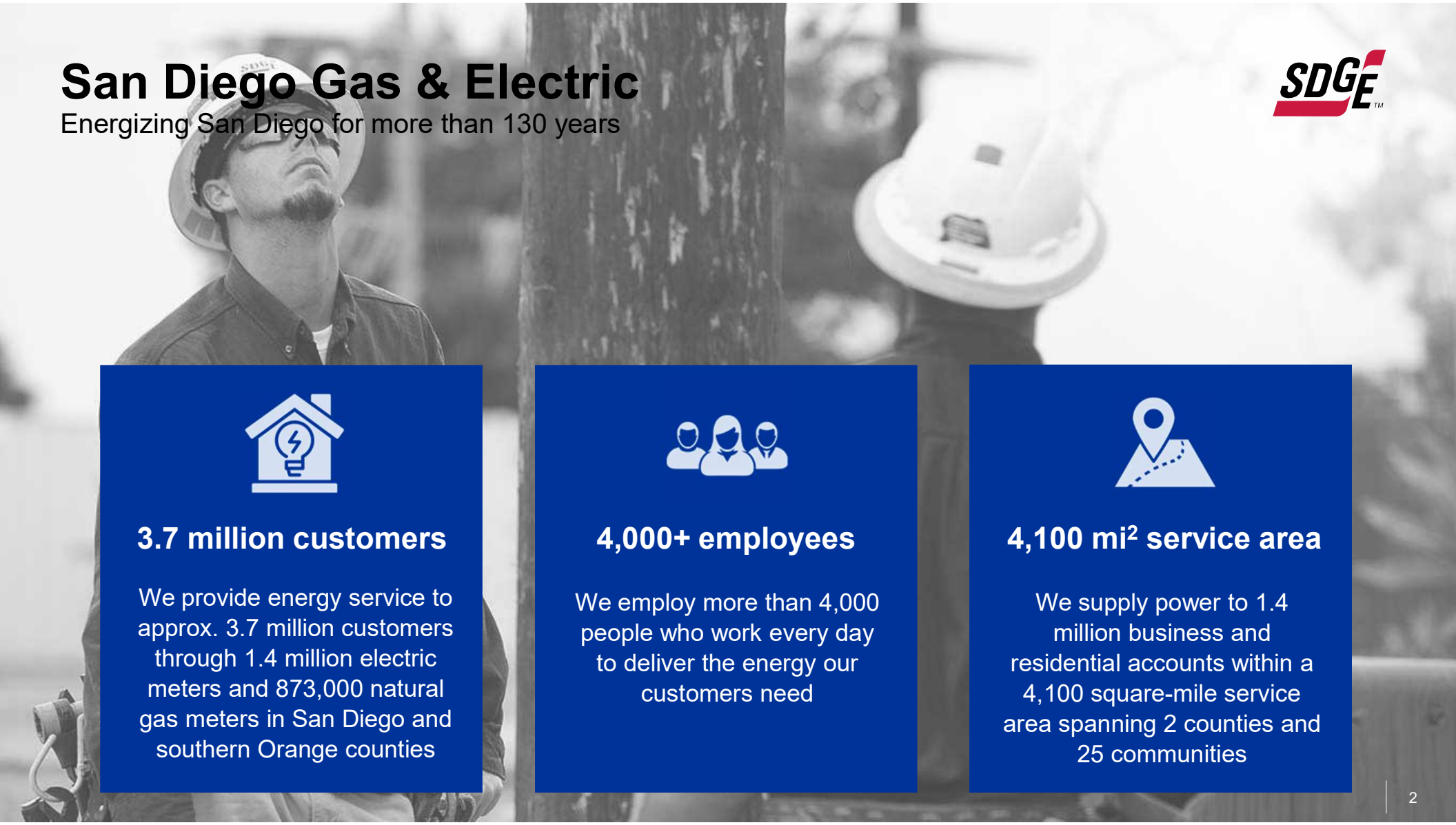

UA Summit®
May 10 - 12, 2022 | New Orleans, LA
Hosted by Entergy
Predicting Wildfires
From the First Data Scientist to Now
Presentation by:
Phi Nguyen, PhD
Senior Data Scientist @
San Diego Gas & Electric® Company

San Diego Gas & Electric

Energizing San Diego for more than 130 years



3.7 million customers

We provide energy service to approx. 3.7 million customers through 1.4 million electric meters and 873,000 natural gas meters in San Diego and southern Orange counties

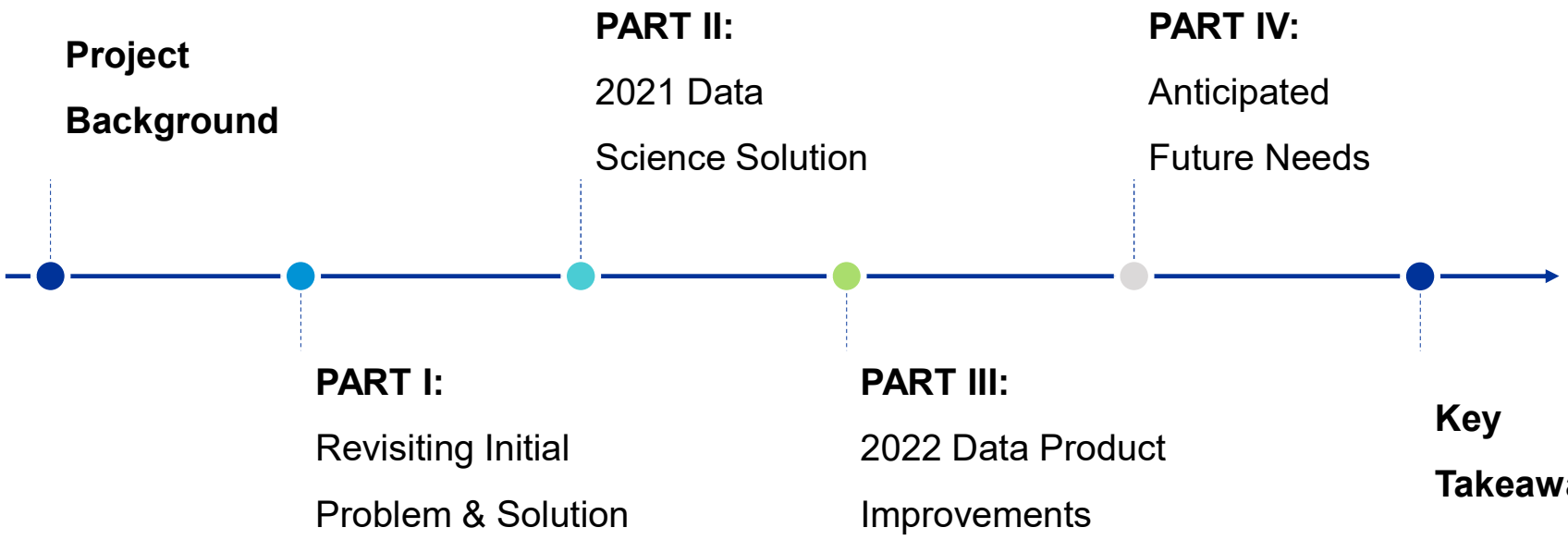
4,000+ employees

We employ more than 4,000 people who work every day to deliver the energy our customers need

4,100 mi^2 service area

We supply power to 1.4 million business and residential accounts within a 4,100 square-mile service area spanning 2 counties and 25 communities

Presentation Outline



Motivation

Can we use data to assist and inform difficult decisions?

Primary Use Case: Public Safety Power Shutoffs (PSPS)

2021 Thanksgiving Day PSPS

- Customers affected: **5,858**
- Longest outage duration: **42+ hours**

2007 Witch Fire in San Diego

- Acres burned: **240,000+ acres**
- Estimated cost: **$1B+**

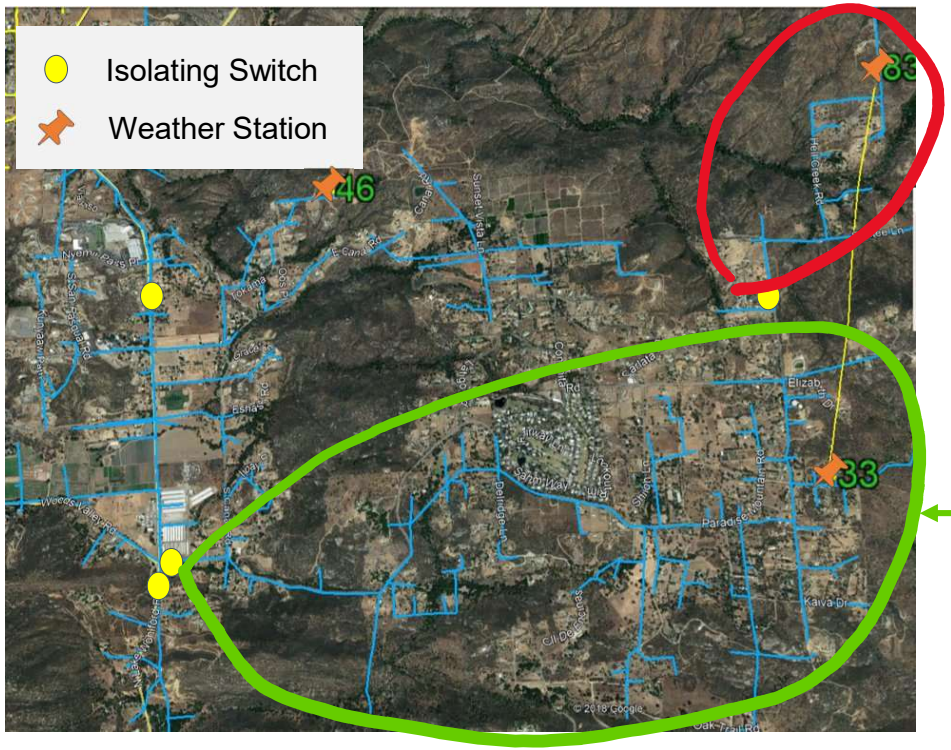
Part I

Rewinding the clock back to 2020…

Risk-Based Project Planning

Decisions considered at the segment-level

A **segment** is a collection of wires and structures between two isolation points, typically defined based on how SDG&E operates PSPS



30 Customers Served

470 Customers Served

For Operations:

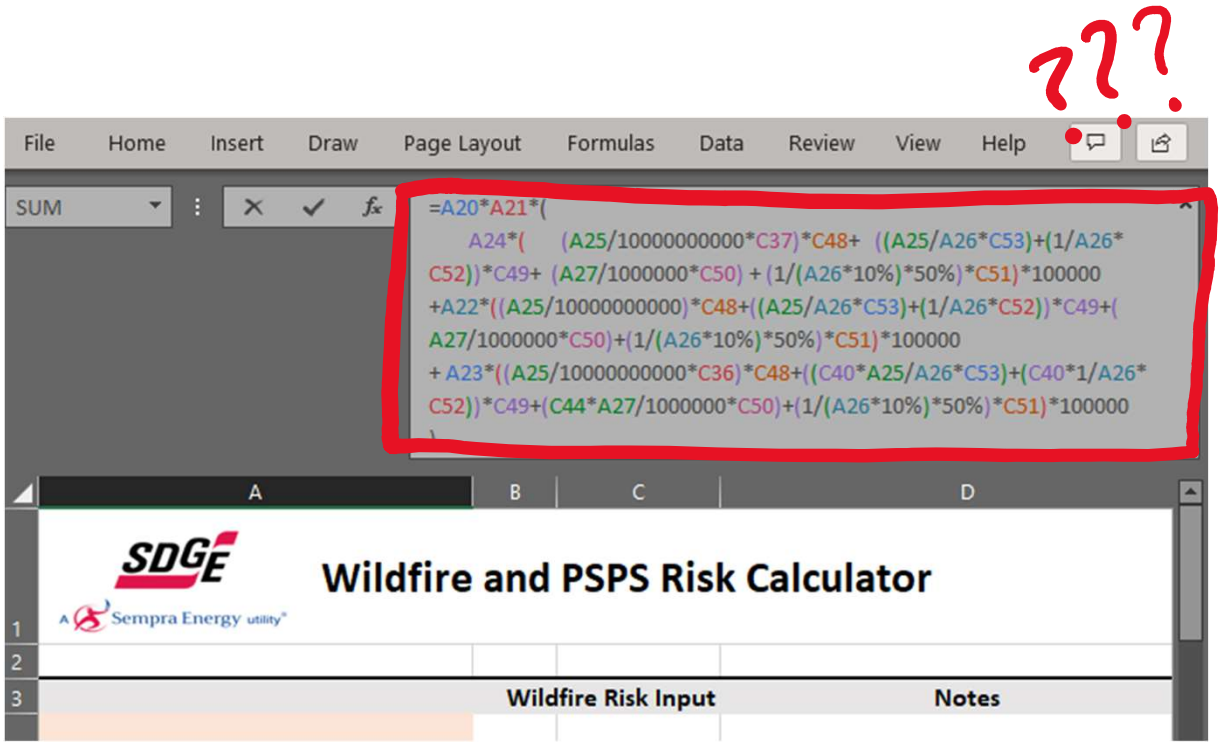
- Segments controlled during weather events by SCADA devices, which are strategically positioned
- Weather stations inform conditions around the segment
- Geospatial mapping adds additional situational awareness

For Calculations:

- It is assumed that each segment contains assets that are in similar risk conditions (weather, vegetation, etc.)
- Number of customers for each segment queried; calculations consider all downstream customers

Work Product: Spreadsheets

A powerful, user-friendly solution based on annual ignition rates



- For each segment, the system of records is used to calculate annual ignition rates:
  - Requires bringing in multiple datasets, usually as “tabs”
  - Pivot tables in MS Excel for calculations
- Expected Risk Score = likelihood x consequence of event
- PSPS risk mainly driven by customers that would be deactivated
- A ratio is calculated to weigh the benefits of de-energization

Challenges

Model Maturation & Complexity

- Implementing a multivariate approach
- Assessing statistical significations
- Performing sensitivity analyses
- Limitation of data sources (size and type)

Reproducibility & Documentation

- Version controlling and versioning granularity
- Error-prone
- Reverse-engineering equations or forgotten quantities
- Data security and privacy

Scalability & Automation

- More frequent model outputs
- Acquiring latest values
- Accessing data sources & managing “pipelines”
- Managing memory constraints

Part II

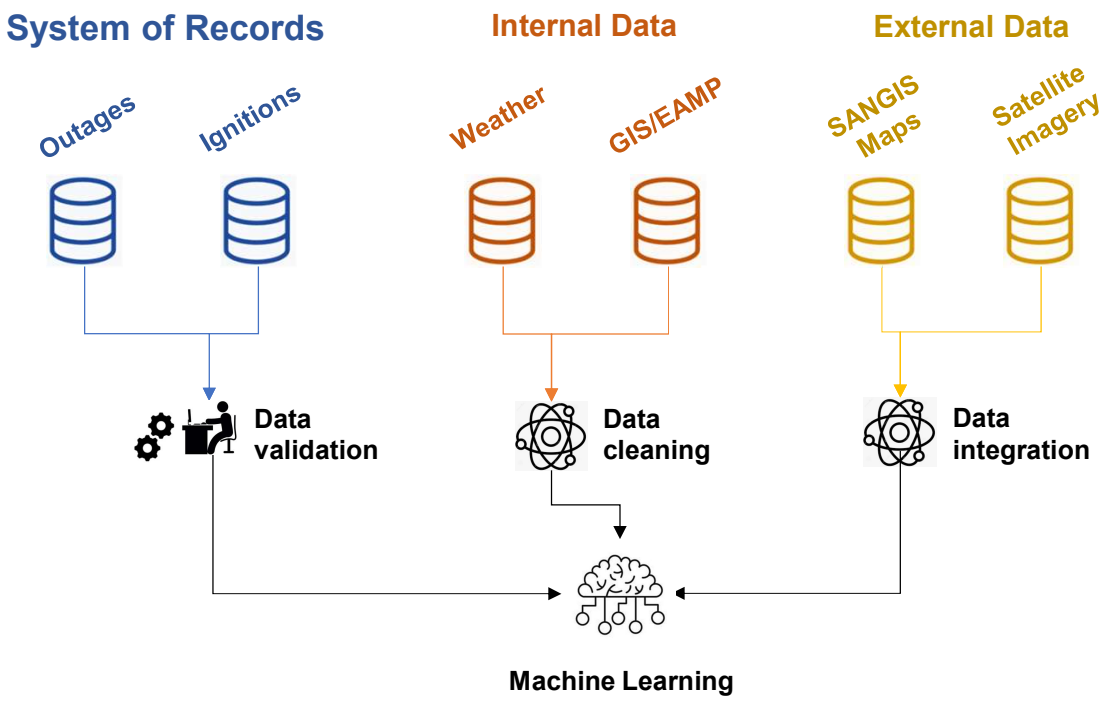
Reframing the problem in 2021

Vision & Anticipated Needs

Forward-Looking Improvements	Data Science Translation
More targeted analytics	Higher granularity in the dataset; potentially “big data”
Numbers “backed by data”	Training sets and statistical modeling
Most recent calculations	Automation, connecting to data sources
Expand to entire service territory	Larger datasets, potentially requires parallel computing
Improved documentation	Version control
Consistency with other programs	Central “source of truth”, data mesh

Model Training

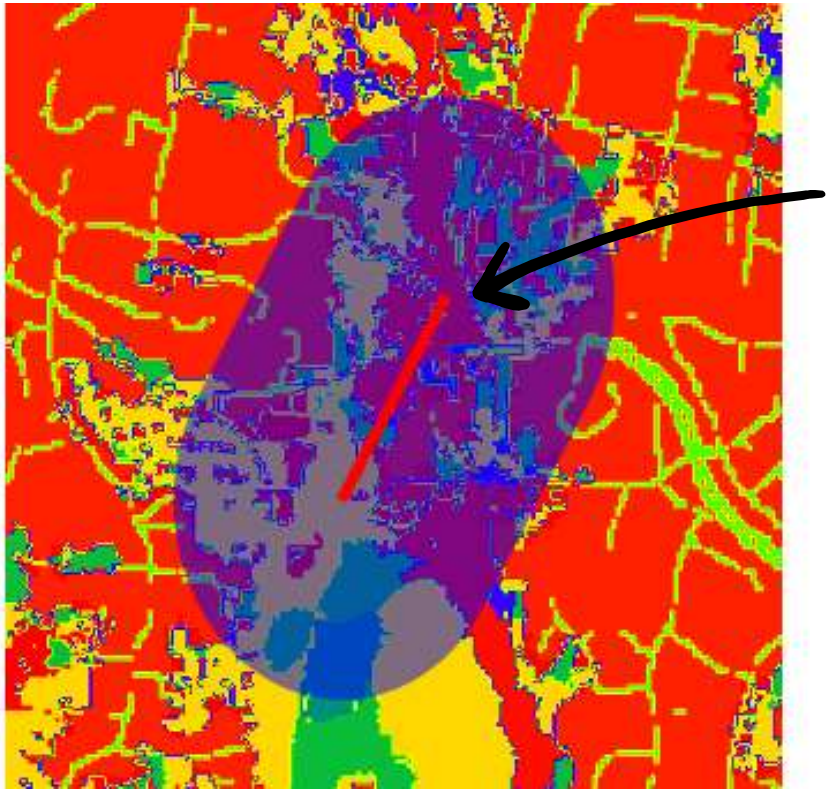
Data sources + machine learning



- **System of records** used for model training:
  - Human-recorded values present challenges
  - Highly imbalanced dataset (generally “rare” occurrences)
  - Collected for entirely different purposes than modeling
  - Varying levels of granularity
- **Weather data** critical for correlations with wind gust (primary decision factor)
- **Asset information** historically focused on geospatial analysis:
  - Excellent geolocations
  - Improvements needed on asset characteristics
  - Not trivial to join/map the data

Feature Engineering

Creating geospatial quantities correlated with wildfire ignition



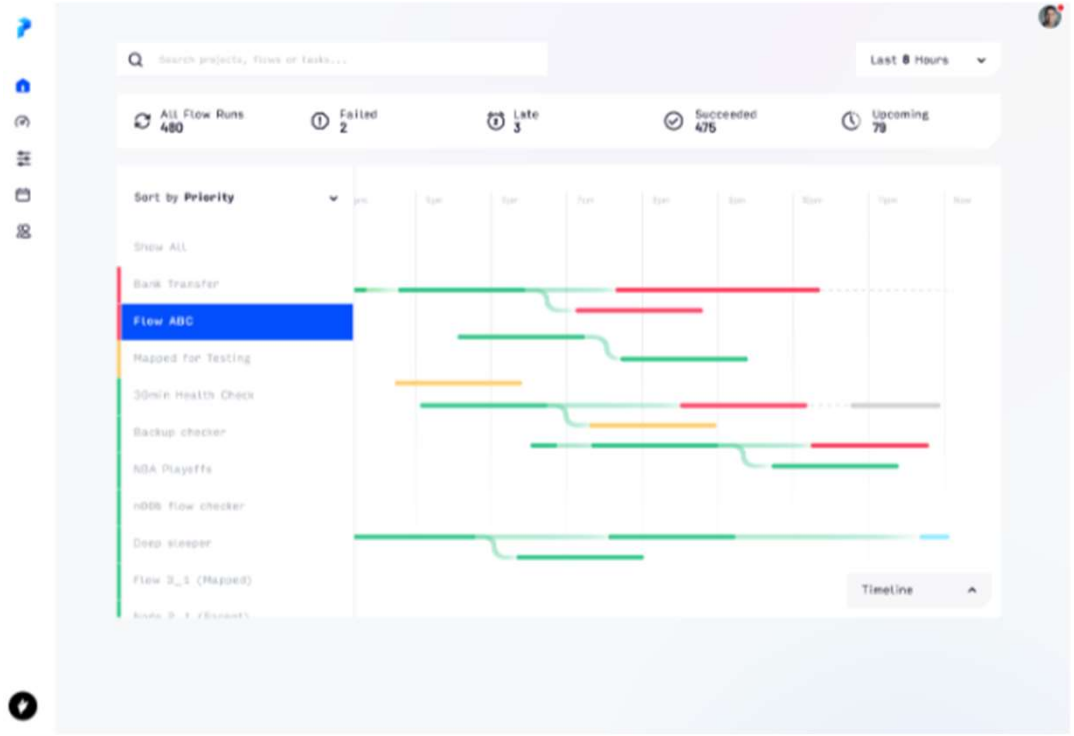
- 9m resolution fuel sources map with 30+ fuel types provided annually for SDG&E service territory
- **Quantification:**
  - Wire spans are overlayed with the raster map
  - Buffers created around the spans to calculate the fraction of each fuel type around the span
  - Area processing to calculate fraction of each fuel type within a buffer
- Buffer size can be experimented with and optimized
- Each outage is mapped to one or more wire spans in our inventory

Summary of Approaches

Model	Algorithm	SME	Python Library
Conductor Failure	Linear regression (log-log)	Electric District Operations	statsmodels
Balloon Contact	Logistic regression	n/a	statsmodels
Animal Contact	Extreme gradient boosted trees	n/a	xgboost
Vegetation Contact	Empirical	Vegetation Management	n/a
Vehicle Contact	Extreme gradient boosted trees	Electric District Operations	xgboost
Wire Ignition	Ensemble decision trees (random forest)	Fire Science	pycaret
Pole Ignition	Ensemble decision trees (random forest)	Fire Science	sklearn

Inference Pipeline in Python

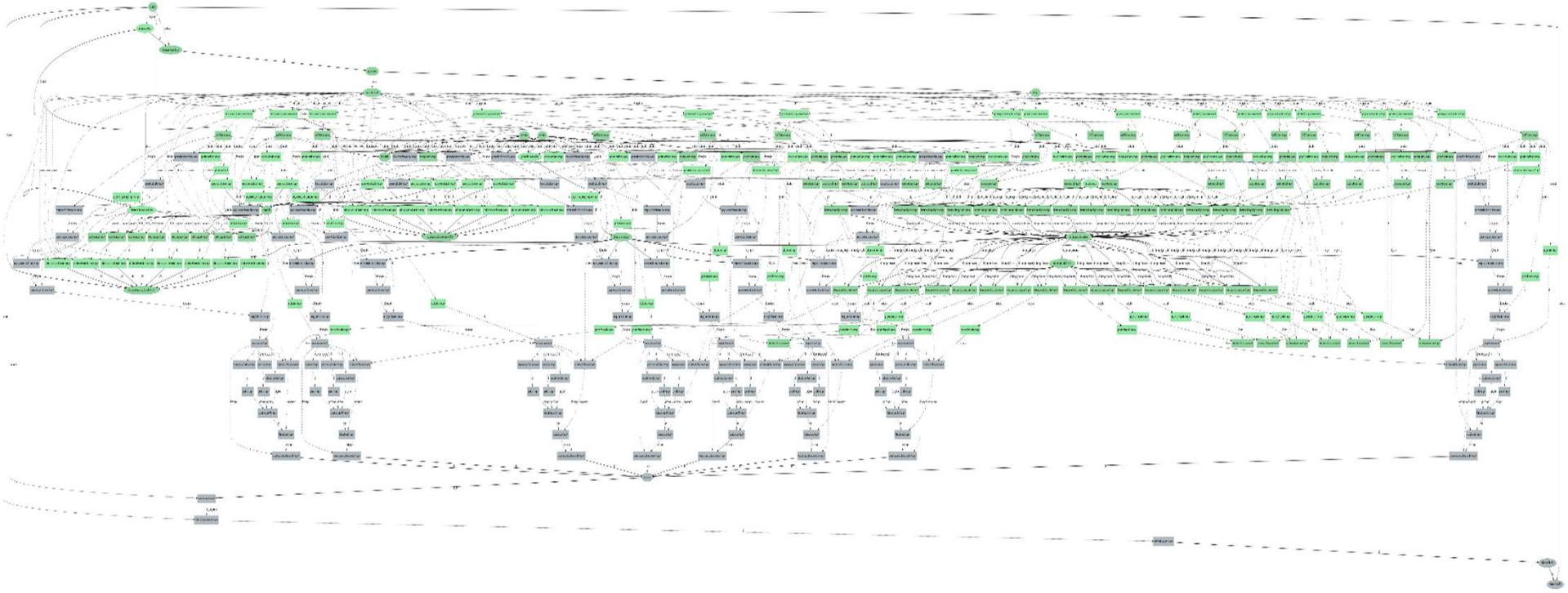
Prefect – Open-Source Framework



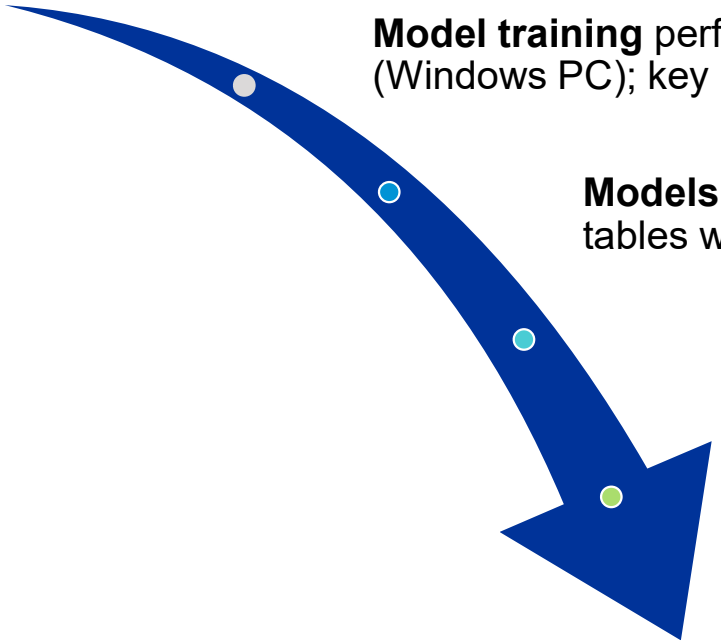
**For demonstration purposes only*

- Generating model outputs requires sometimes-complex data pipelines that must be carefully managed
- Pipeline also considered a Directed Acyclic Graphs (DAG), consisting of:
  - tasks
  - task dependencies
  - other metadata
- A framework to define DAGs allows the project to scale, make changes reliably and more easily tracks errors and interim outputs

In practice, DAGs can get pretty ugly!



Operations



Model training performed on company-approved laptops (Windows PC); key challenge was *data access*

Models packaged with `.pkl` file; in some cases, static tables were produced by models and used as outputs

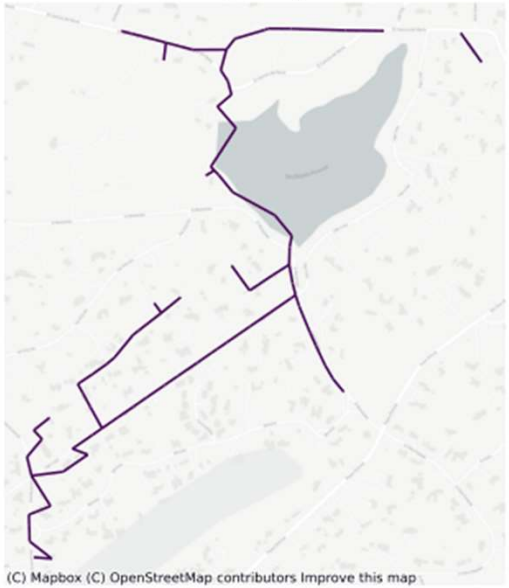
Inference pipeline run on virtual “Wintel” machine by a single user (~1 week process)

Outputs stored on virtual machine hard disk and transferred as needed through SharePoint

Model Outputs in Real-Time

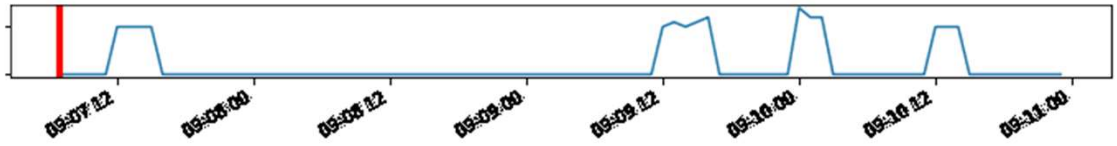
High temporal and spatial granularity for better targeting

Failure Probability Model



Ignition Likelihood Model

Wind Gust



**For demonstration purposes only*

- Models are specifically created to be responsive to geography and wind gust, since these are the primary decision parameters
- Models determine failure and ignition probabilities at the asset level
- Combined with weather forecasts, SDG&E can anticipate when and where high-risk events will occur

Application & Usage

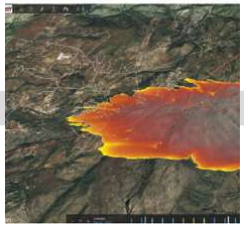
Assisting PSPS decision-making

- Pre-event analysis for areas at potential risk of de-energization
- Information provided to EOC during situational awareness updates
- Integration into PSPS decision dashboard via “Conductor Risk Index”
- Post-event reporting

7-10 days ahead

Forecasts indicate extreme weather; SDG&E begins predictive modeling to assess potential impact

3 days ahead



72-Hour: EOC Activation; High-level weather briefing to provide two full days of daylight for pre-patrols

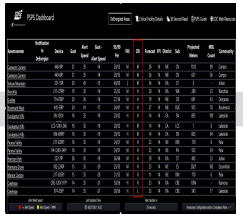
2 days ahead

48-Hour: Receive circuit list from Meteorology; Continue pre-patrol and gather damage assessments

1 day ahead

24-Hour: Placement of Observers

Day of power shutoff



Weather arrives: De-energization

Power restored

Weather passes: Post-patrol and re-energization

Retrospective & Reflection

Some key takeaways from 2021

Computing Resources

Most existing systems got the job done, but forward-looking platforms enable growth

Documentation

Code version control improved the granularity and organization of changes to modeling logic

Product Requirements

Could have benefitted from clearer requirements and better product development practices

Interdepartmental Coordination

Most time was spent chasing, understanding, or cleaning data that is owned and managed by other departments

Organizational Change

Close coordination with subject matter experts and data owners was vital to project success

Wow-factor

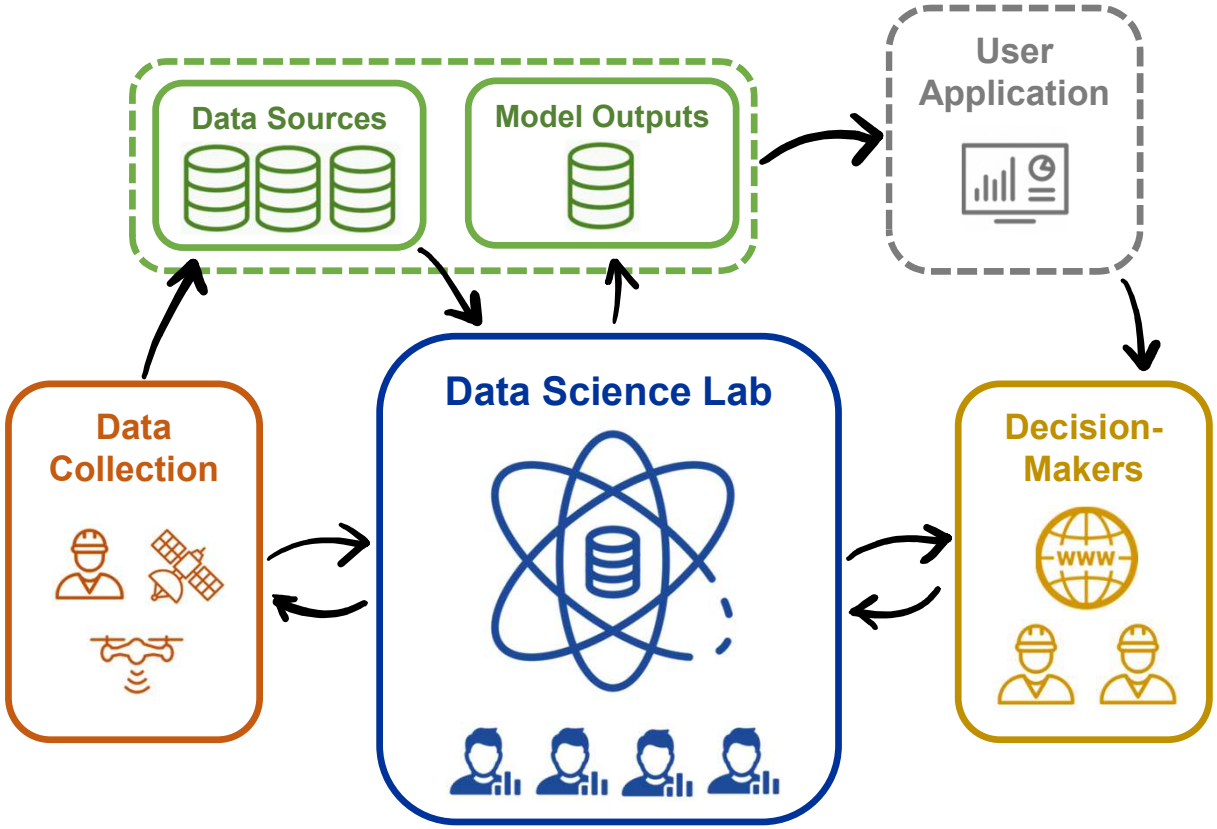
End-product appealed mainly to data enthusiasts; stakeholders expressed desire for better visualizations

Part III

Improving the solution in 2022

Data Product Development Planning

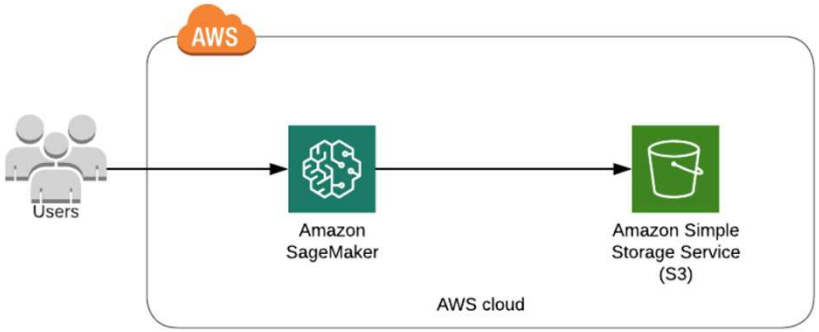
Specialization of features and product strategy



- Managing the product:
  - Strengthening stakeholder engagement
  - understanding and balancing various stakeholder needs
  - Better understanding of product requirements
- Clear product and feature definitions; clearer product improvement plan
  - Better understanding roles, skillsets, and areas to outsource
  - A result is that team does not have to “do it all”
- Core team focuses on building data products, not software

AWS Sagemaker Platform

Landing on a data science platform



Key Benefits

- Modern production model deployments
- Date lake connections
- Lower costs with on-demand workflow computing
- Highly scalable
- **A common framework**

Other Considerations

- Sacrifice flexibility of coding solutions
- Vendor lock
- Skills gap for widespread adoption
- Unlikely a “one-person job”
- **A common framework**

Vegetation Model – Overview

Build consensus around a technically rigorous, impactful and grounded solution

USGS Landsat 8

Surface Reflectance

- 30m resolution
- Multispectral images
- Reading every 16 days
- 2013-Present on AWS

SDG&E Tree Database

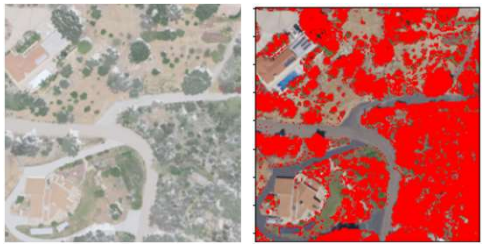
(Powerworkz)

- Tree species
- Tree height
- Human inspections

USDA NAIP

Aerial Imagery

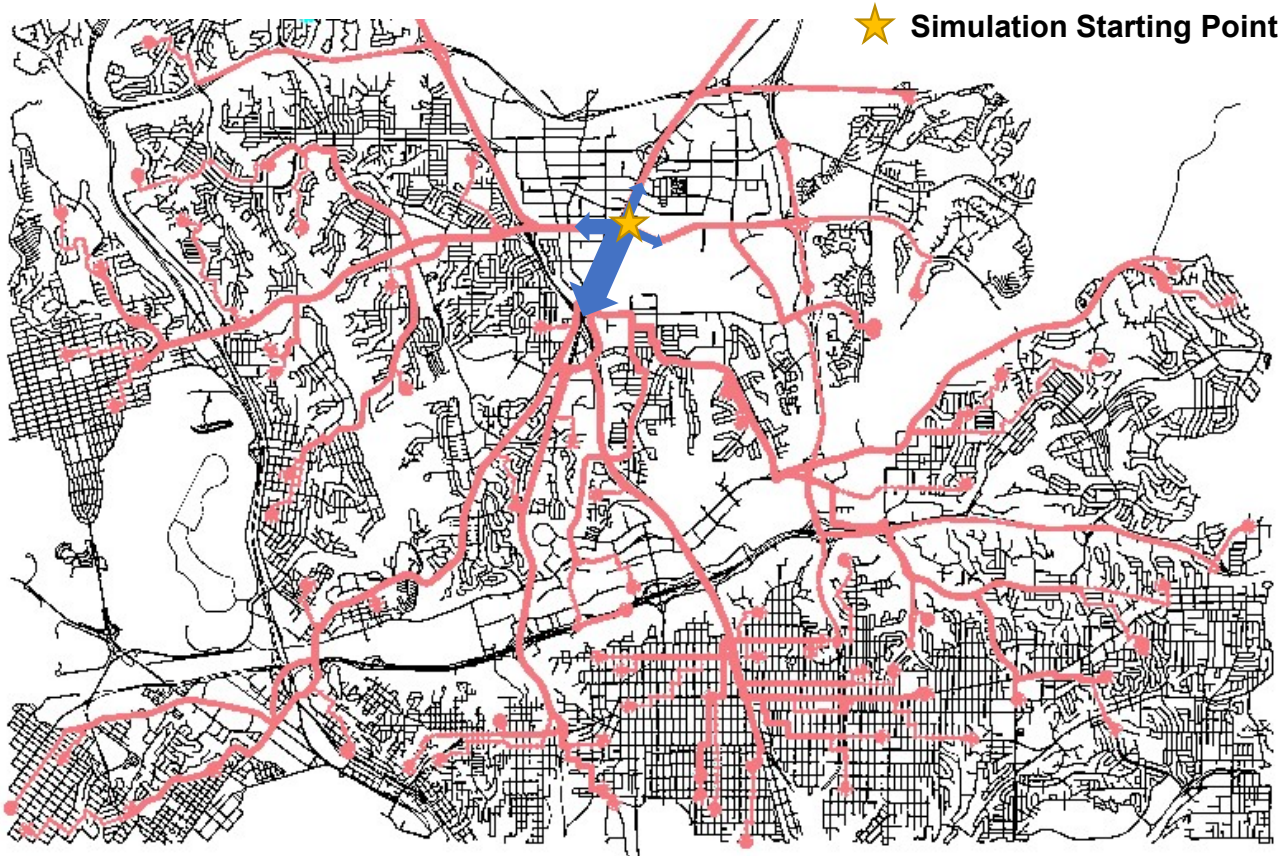
- 60cm resolution
- 4-band high-images
- Reading every 2 years
- 2012-2020 on AWS



TensorFlow

Computationally Modeling Traffic

Traffic congestion during evacuations lead to wildfire egress risk



**For demonstration purposes only*

- Routing algorithms using network analysis techniques can be used to simulate traffic patterns
- Millions of simulations with specified constraints, such as areas blocked by wildfire, produce a probabilistic travel model (cloud computing)
- Dynamic weather-based fire growth creates changes and updates travel routes iteratively
- Coordination with Fire Science team at SDG&E to create realistic evacuation models

Traffic Congestion Demo

Identifying areas of highest risk

**For demonstration purposes only*

- In preliminary work, we created an evacuation zone and assessed areas of congestion based on escape routes as the zone grew

Note: *custom tools built in-house allow full flexibility of parameters and their values, assumptions, etc.*

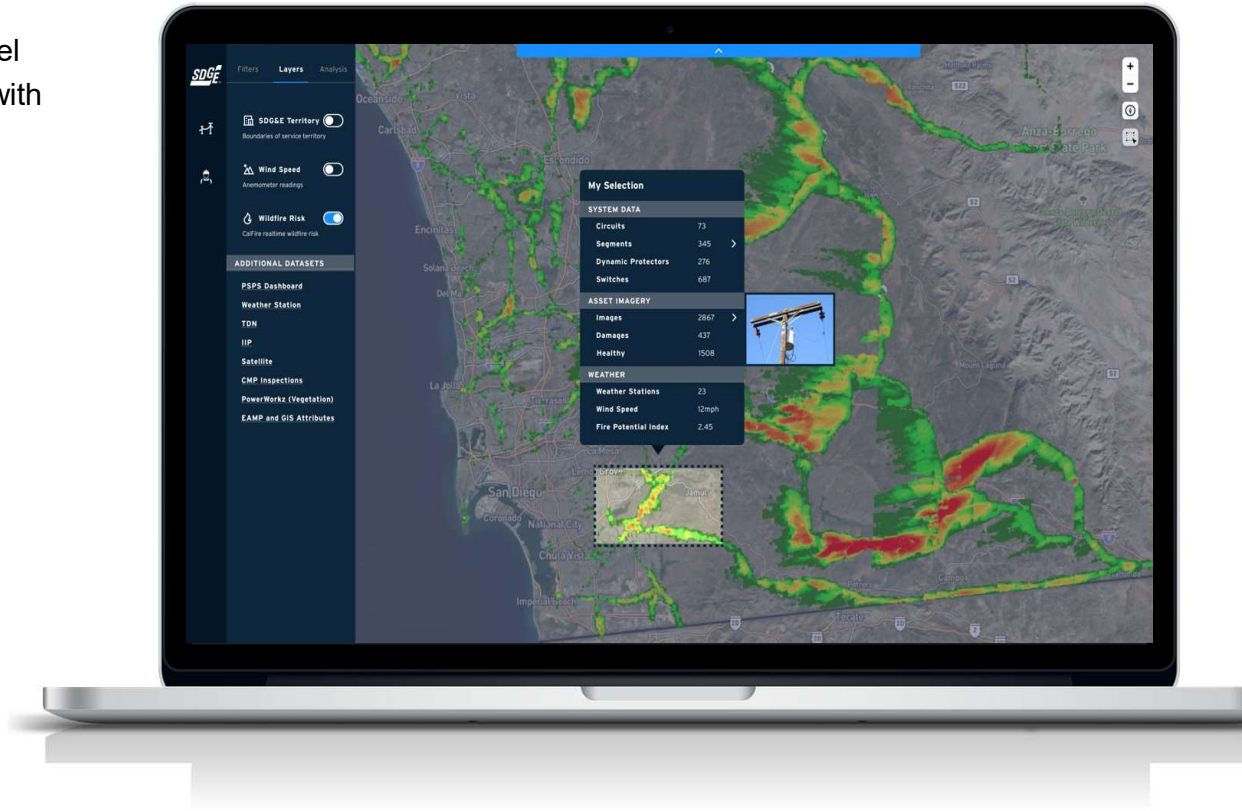
- Throughout each step, we can assess, track and analyze results
- A simulation of our entire territory is now technically feasible and will be a useful tool for risk-based planning

State-of-the-Art UI/UX

Real-Time Dashboards & Interactive Geospatial Visualizations

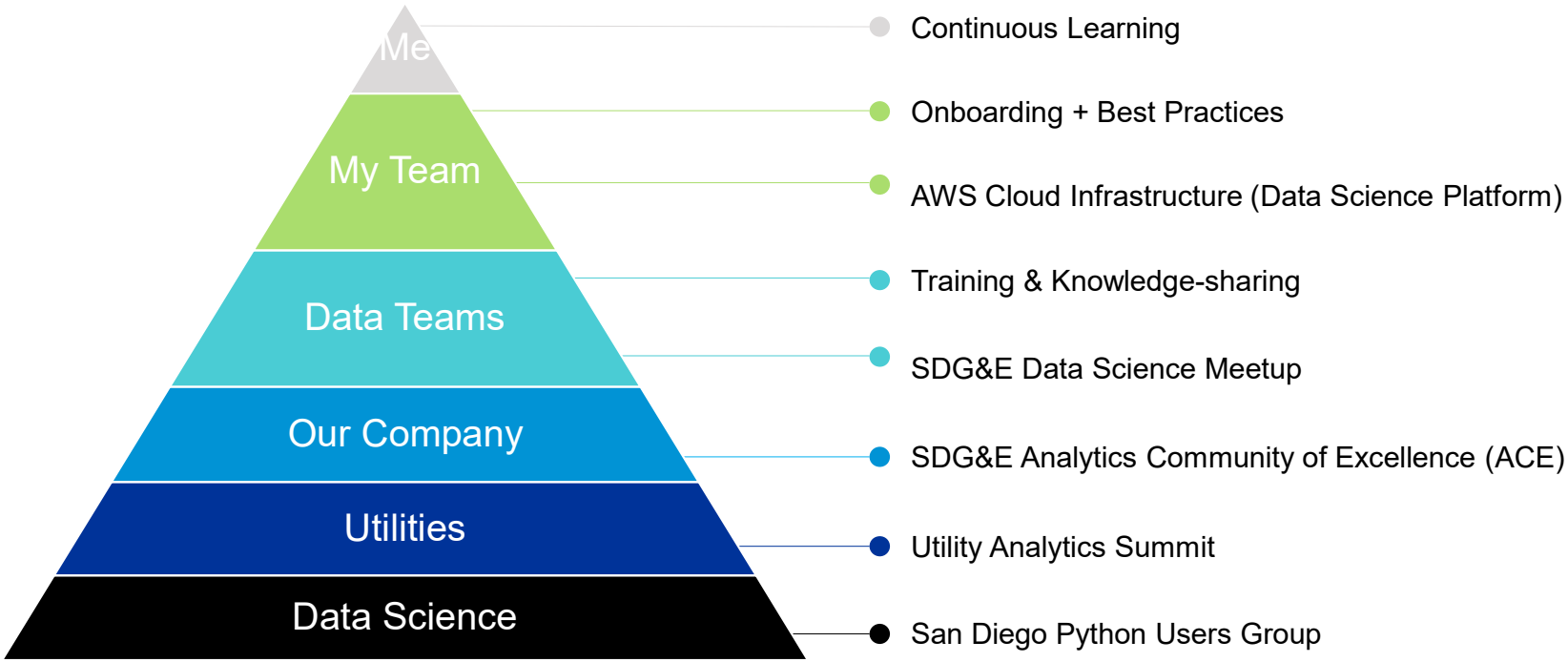
Provide a user-centric application for interactive with model outputs and other risk factors. Complements model data with external information for context and situation awareness. Additional data may include:

- External data sources
  - Satellite images
  - Drone images
- Asset records
  - Asset characteristics
  - Inspection history
  - Work orders
- Outage history
- Network tracing
- Customer data



Organizational Engagement

Strengthening the foundations for the success of data science at SDG&E



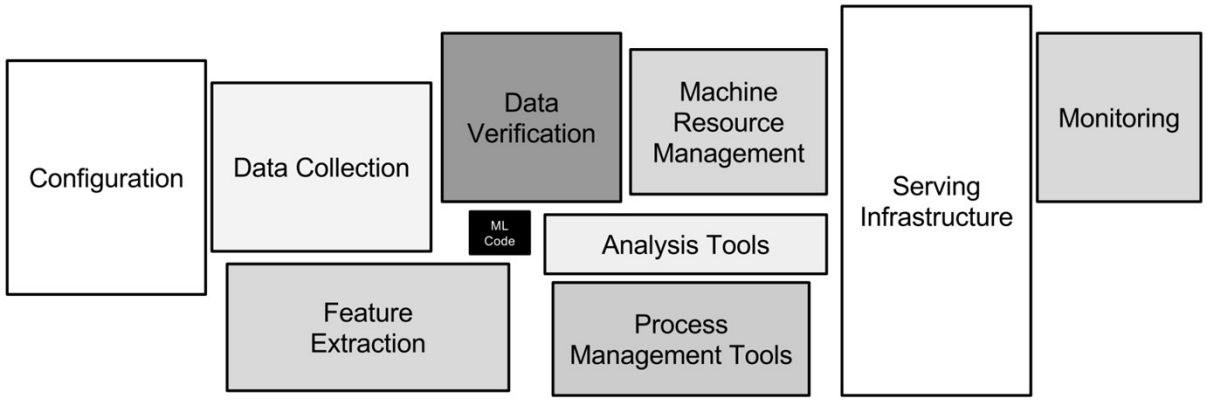
Part IV

Anticipating the future

Technical Debt

That which must always be repaid

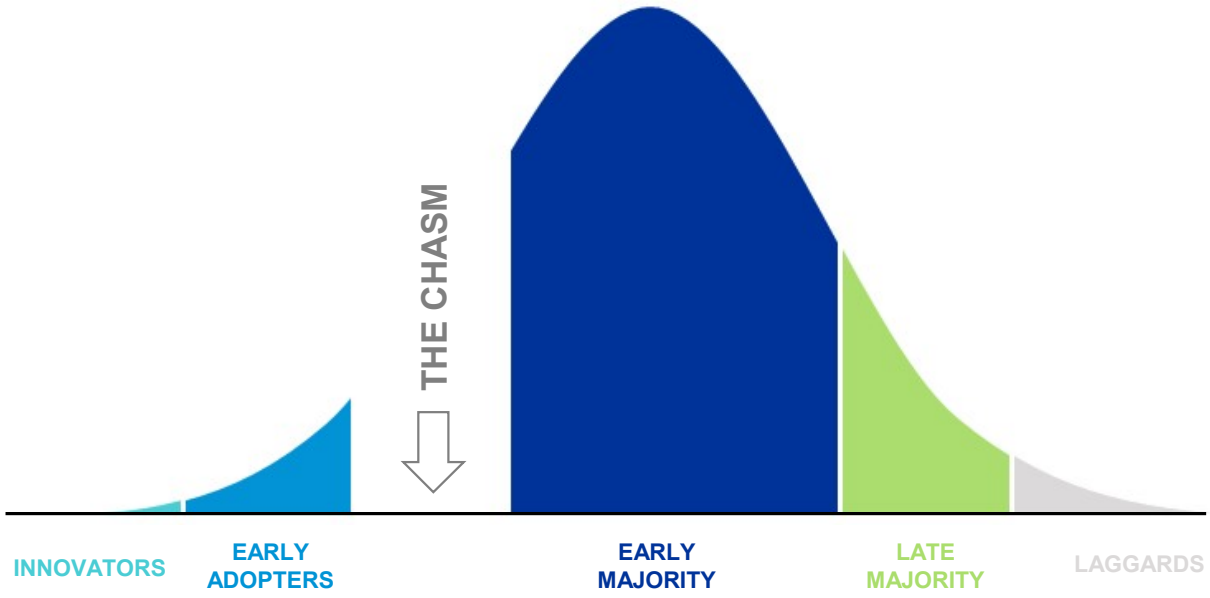
Early data science projects and proof of concepts must find a difficult balance between developing a solid infrastructure and quickly delivering a minimally viable product



Google (2015). Hidden Technical Debt in Machine Learning Systems. NIPS. 2494-2502.

The Chasm

Widespread adoption of data-driven, risk-based products



Key phrases:

- “Intrapreneurship”
- “Customer development”
- Data science evangelism
- Change management

Closing thoughts

Lessons learned & opinionated takeaways

Data sources can be tricky for modeling. Start with simple, informative models and build from there. Focus on model usefulness. Key philosophy: *"all models are wrong, some are useful"*

Work towards a **data science platform**. Spend the time to determine what fits your team and commit to it. This will make all technical considerations easier

If **hiring** your first data scientist, consider hiring a "builder". Look for enthusiasm, self-learning, attitud and seniority. Key philosophy: *"programming as a skill, not a specialty"*

Find your right balance between working on the minimum viable product and data infrastructure. The focus should be on the **data product**, not the software. Identify product requirements early and adjust frequently

Project success is highly dependent on **organizational maturity**. Focus on building a culture. Coalition building within the utility is a core necessity; very good interdepartmental relations a must

Always **version control** your code. Please just do it. It helps in every stage of development and should be part of the growth plan for all young teams

Thank you!

Questions?

Phi Nguyen, PhD

Senior Data Scientist,

San Diego Gas & Electric