Workflow-Driven Geoinformatics Applications and Training in the Big Data Era

İlkay ALTINTAŞ, Ph.D.

Chief Data Science Officer, San Diego Supercomputer Center

Division Director, Cyberinfrastructure Research, Education and Development

Founder and Director, Workflows for Data Science Center of Excellence

SDSC SAN DIEGO SUPERCOMPUTER CENTER

UC San Diego

SAN DIEGO SUPERCOMPUTER CENTER at UC San Diego
Providing Cyberinfrastructure for Research and Education

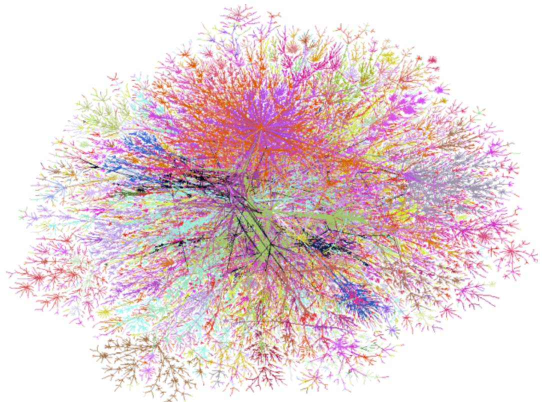
- Established as a national supercomputer resource center in 1985 by NSF
- A world leader in HPC, data-intensive computing, and scientific data management
- Current strategic focus on “Big Data”, “versatile computing”, and “life sciences applications”

Recent Innovative Architectures

- ***Gordon:*** *First Flash-based Supercomputer for Data-intensive Apps*
- ***Comet:*** *Serving the Long Tail of Science*

Data Science Today is Both a Big Data and a Big Compute Discipline

COMPUTING AT SCALE



BIG DATA

Requires:

- *Data management*
- *Data-driven methods*
- *Scalable tools for dynamic coordination and resource optimization*
- *Skilled interdisciplinary workforce*

Enables dynamic data-driven applications

New era of data science!

Smart Manufacturing

Computer-Aided Drug Discovery

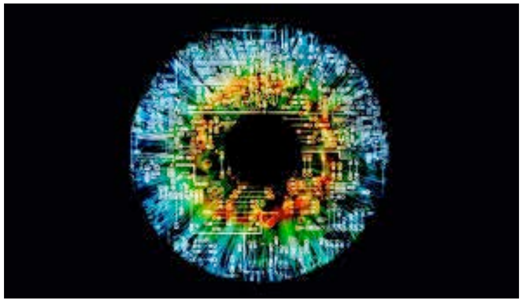
Personalized Precision Medicine

Smart Cities

Smart Grid and Energy Management

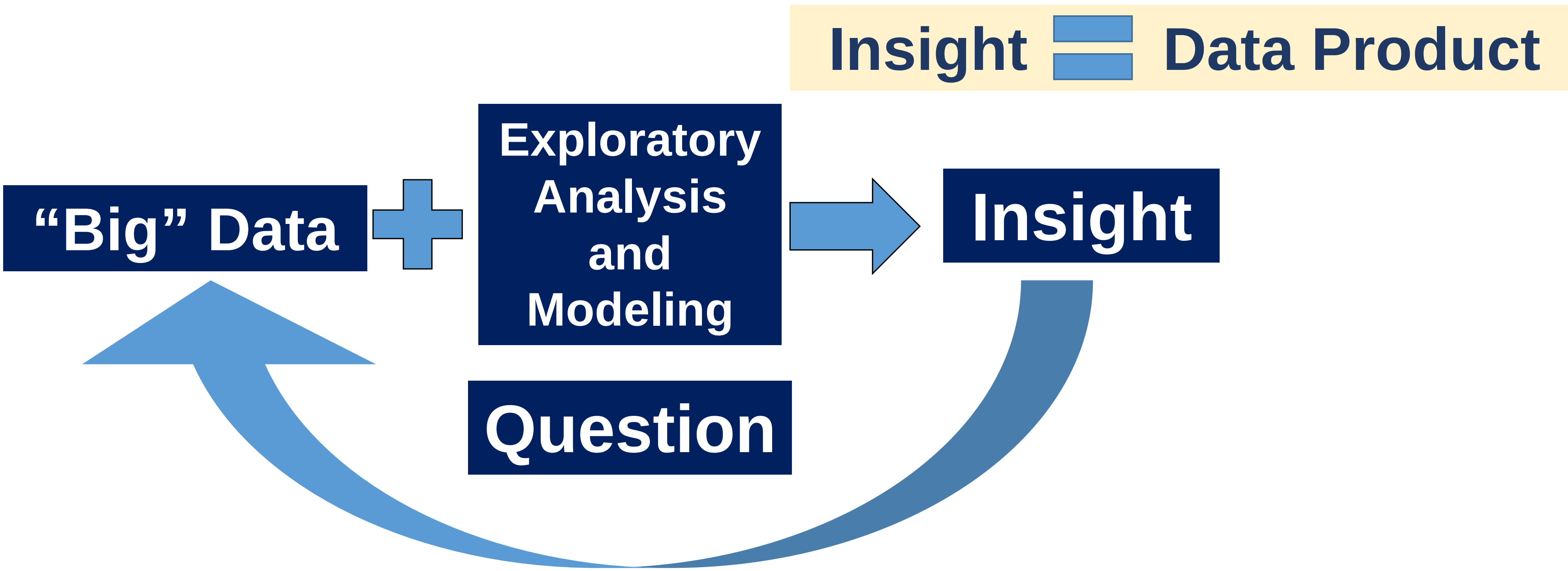
Disaster Resilience and Response

Needs and Trends
for the New Era Data Science

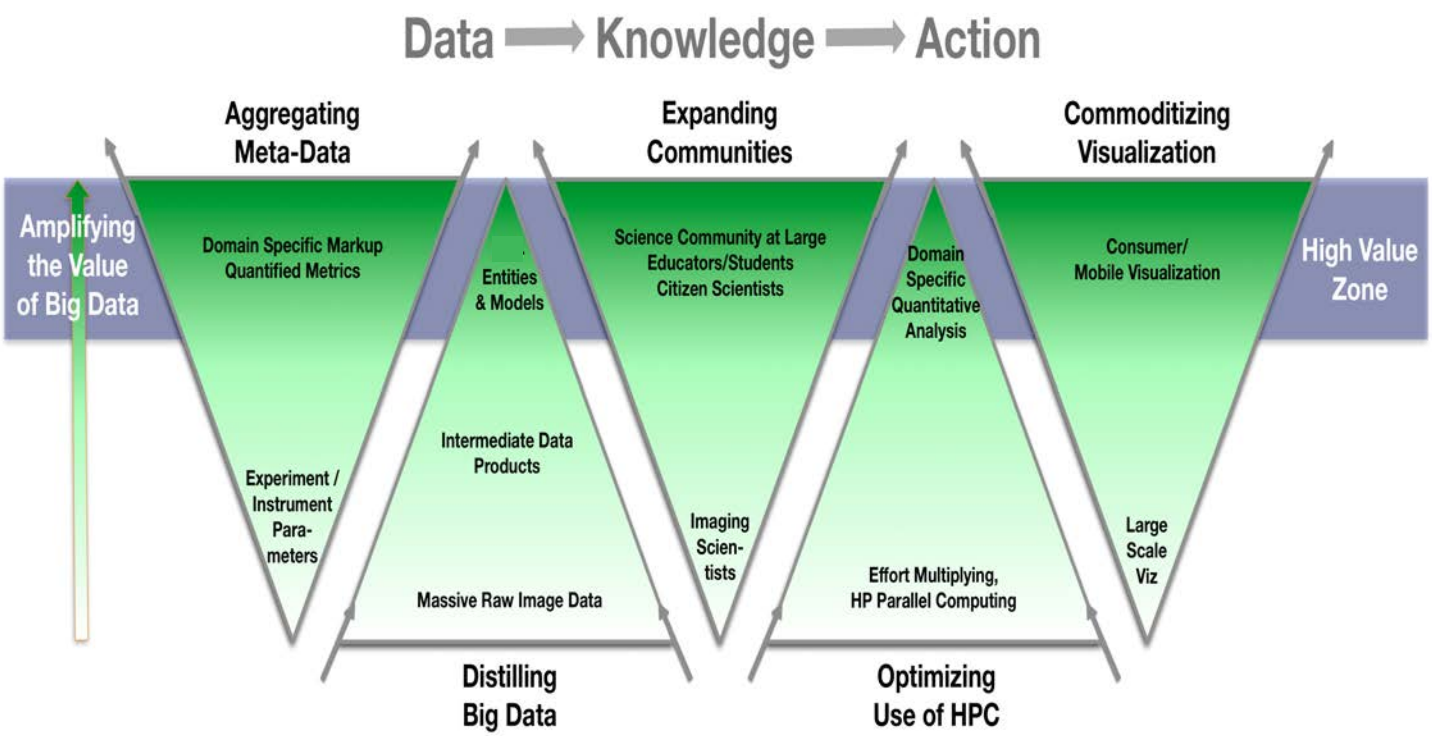


Ultimate Goal

How does successful data science happen?

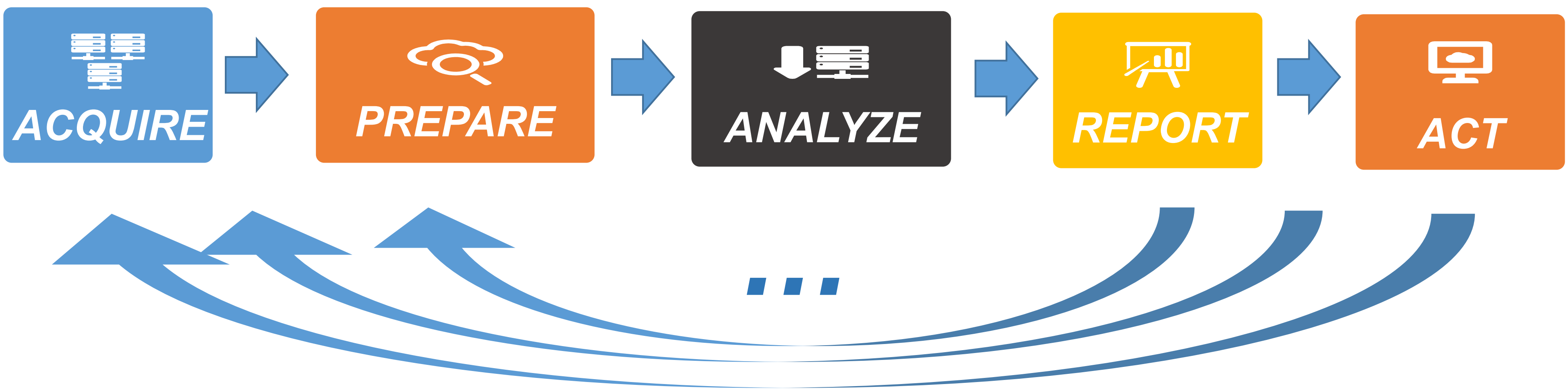


Insights amplify the value of data…



…, but there are many ways to get to insights.

Approach: Focus on Process and Team Work



Create an Ecosystem that Enables Needs and Best Practices

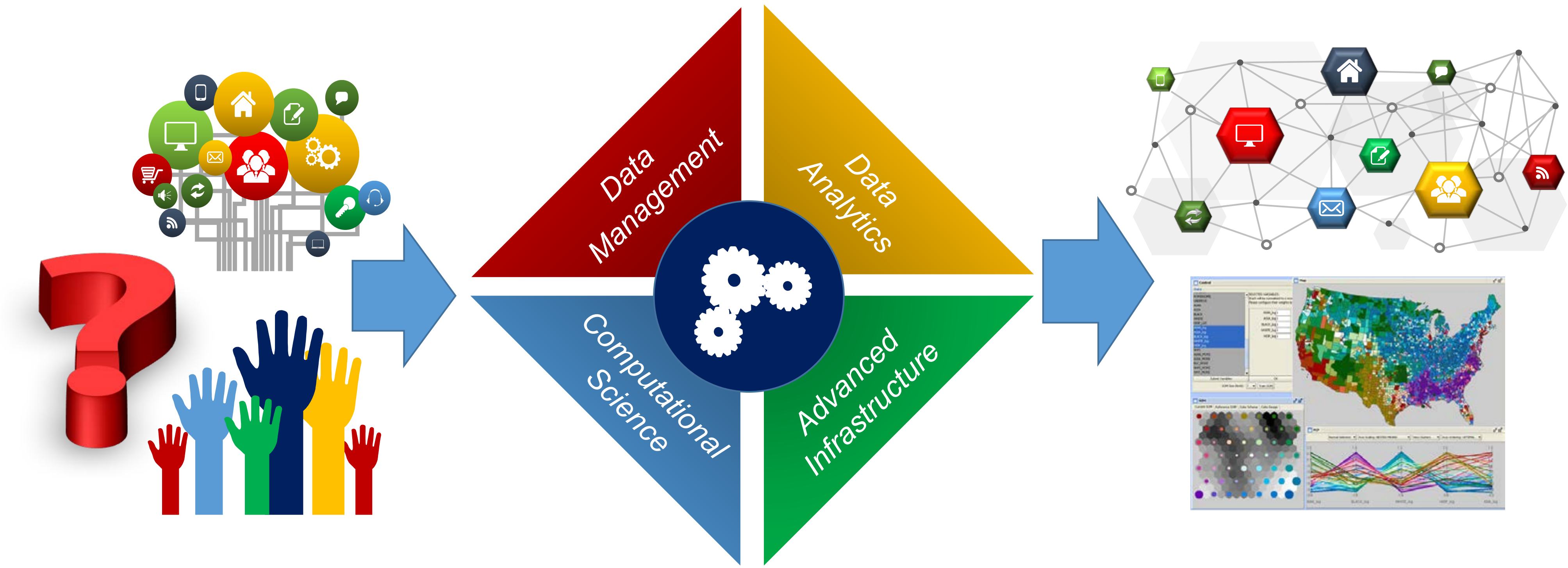
ACQUIRE → PREPARE → ANALYZE → REPORT → ACT

- data-driven
- dynamic
- process-driven
- collaborative

- accountable
- reproducible
- interactive
- heterogeneous

?

What would it such an ecosystem look like?

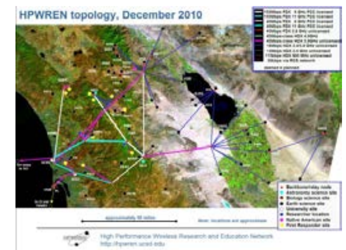


Creating a Collaborative Data Science Ecosystem on top of Advanced Infrastructure

What are some challenges specific to atmospheric sciences?

Geospatial Big Data

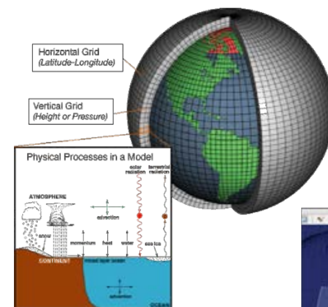
- Flood of new data sources and types
  - Needs new data management, storage and analysis methods
  - Too big for a single server, fast growing data **volume**
  - Requires special database structures that can handle data **variety**
  - Too continuous for analysis at a later time, with increasing streaming rate, i.e., **velocity**
  - Varying degrees of uncertainty in measurements, and other **veracity** issues
  - Provides opportunities for scientific understanding at different scales more than ever, i.e., potential high **value**



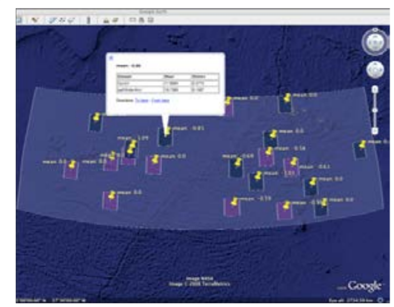
Real-time sensors

Drone imagery

Satellite imagery



Weather forecast



Sea Surface Temperature Measurements

The ‘scalability’ bottleneck

- Resources needed for geospatial big data (e.g., satellite imagery) analysis exceed current capabilities, especially in an on-demand fashion
- Cloud computing is an attractive on-demand decentralized model
  - Need new scheduling capabilities
    - on-demand access to a shared configurable resources
    - programmable networks, servers, storage, applications, and services
  - Need ability to easily combine users environment and community tools together in a scalable way
    - Various tools with different computing scalability needs
  - Cost!!!

The 'sensor data' bottleneck

- Data streaming in at various rates
- "Big Data" by definition in its volume, variety, velocity and viscosity
  - Need to improve veracity and add value by providing provenance- and standards-aware on-the-fly archival capabilities
  - QA/QC and automate (real-time) analysis of streaming data before it is even archived.
  - Often low signal-to-noise ratio requiring new methods
- Need for integration of new streaming data technologies

The "workforce" bottleneck

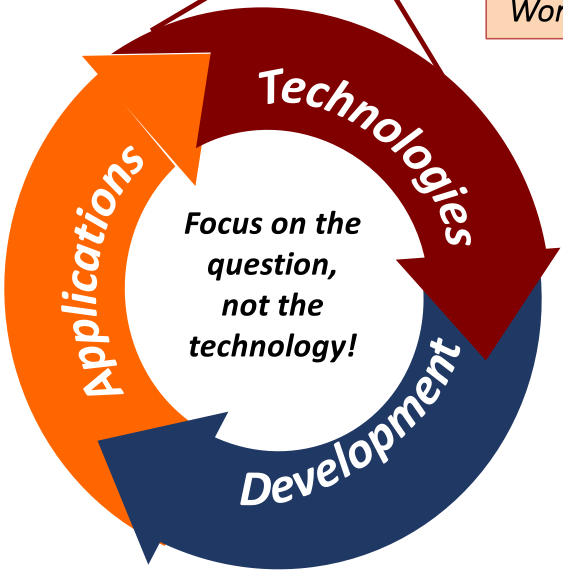
- Geospatial data processing requires a lot of expertise
  - GIS, domain expertise, data engineering, scalable computing, machine learning, …
- No open geospatially enabled big data science education platform
- Teach not just technical knowledge, but collaborative work culture and ethics

Using workflows to get there…

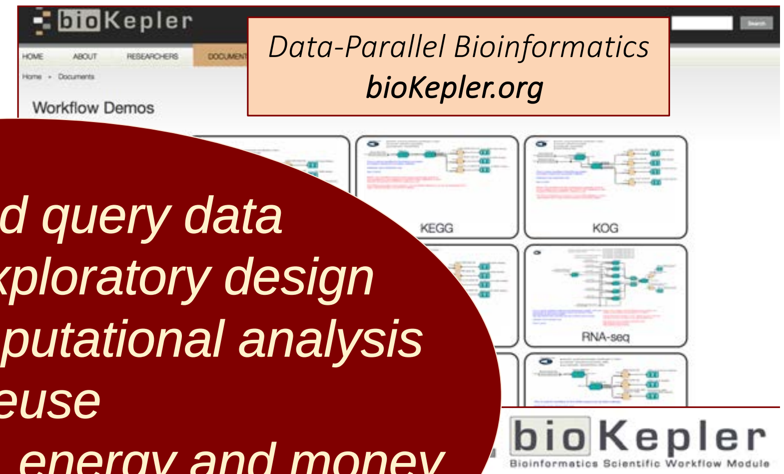
WorDS Center

Workflows for Data Science Center of Excellence at SDSC

WorDS.sdsc.edu



Real-Time Hazards Management
wifire.ucsd.edu

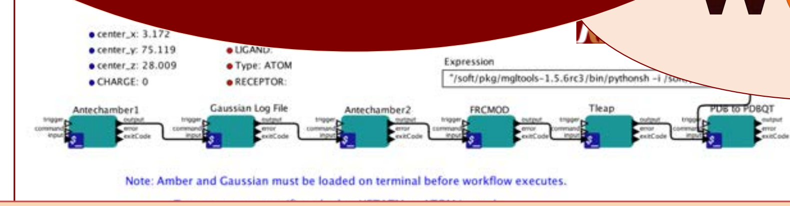


Data-Parallel Bioinformatics
bioKepler.org

- *Access and query data*
- *Support exploratory design*
- *Scale computational analysis*
- *Increase reuse*
- *Save time, energy and money*
- *Formalize and standardize*
- *Train*

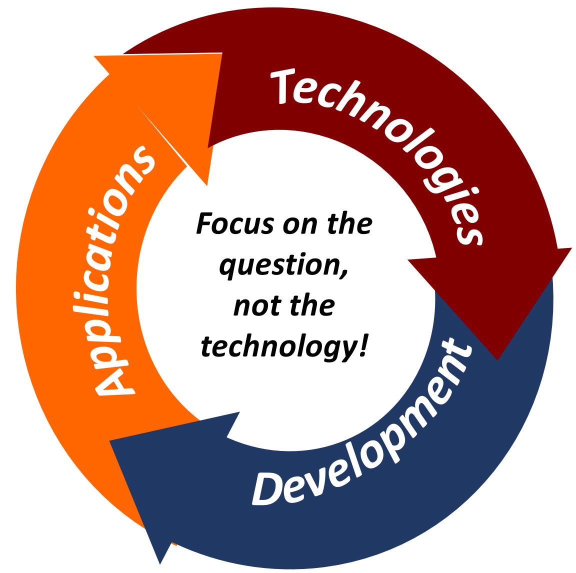
WorDS Center

Goal: Methodology and tool development to build automated and operational workflow-driven solution architectures on big data and HPC platforms.



Scalable Automated Molecular Dynamics and Drug Discovery
nbcr.ucsd.edu

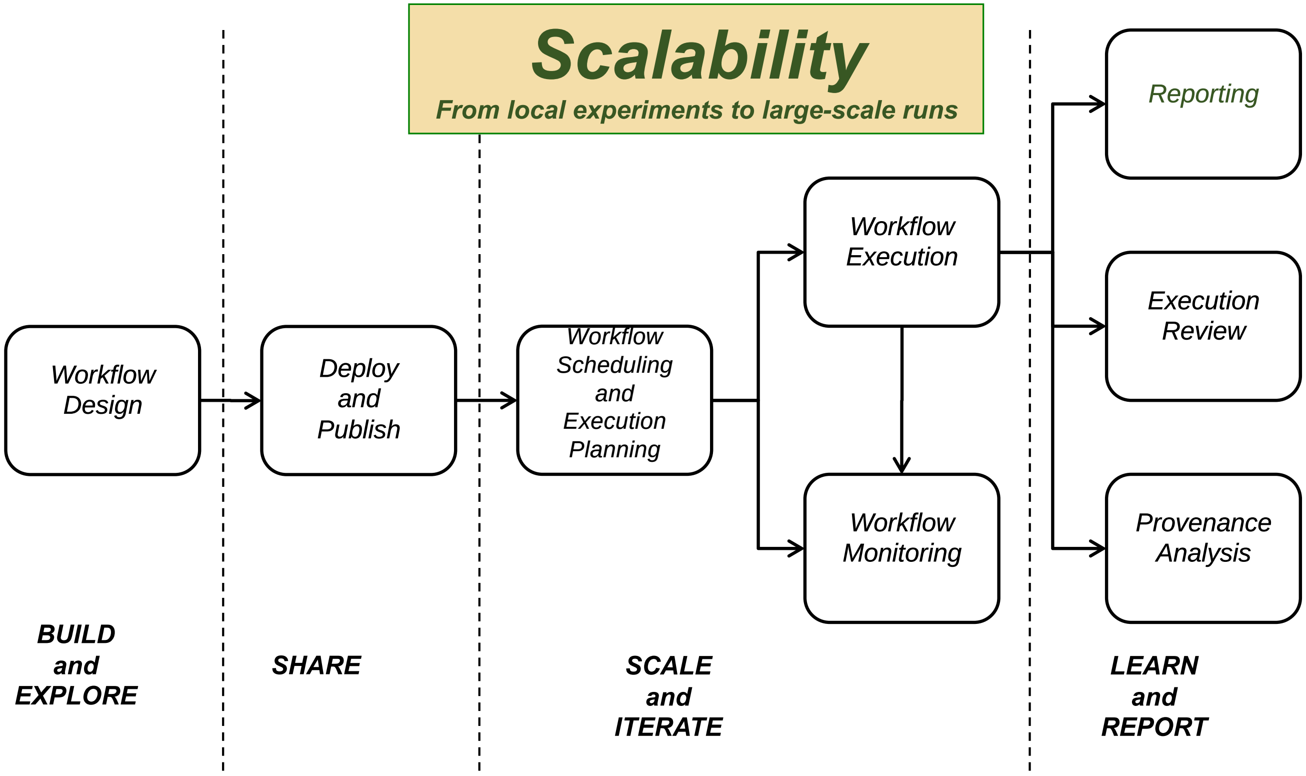
How can I get smart people to collaborate and communicate to analyze data and computing to generate insight and solve a question?



Programmability
Ease of use, iteration, interaction, re-use, re-purpose

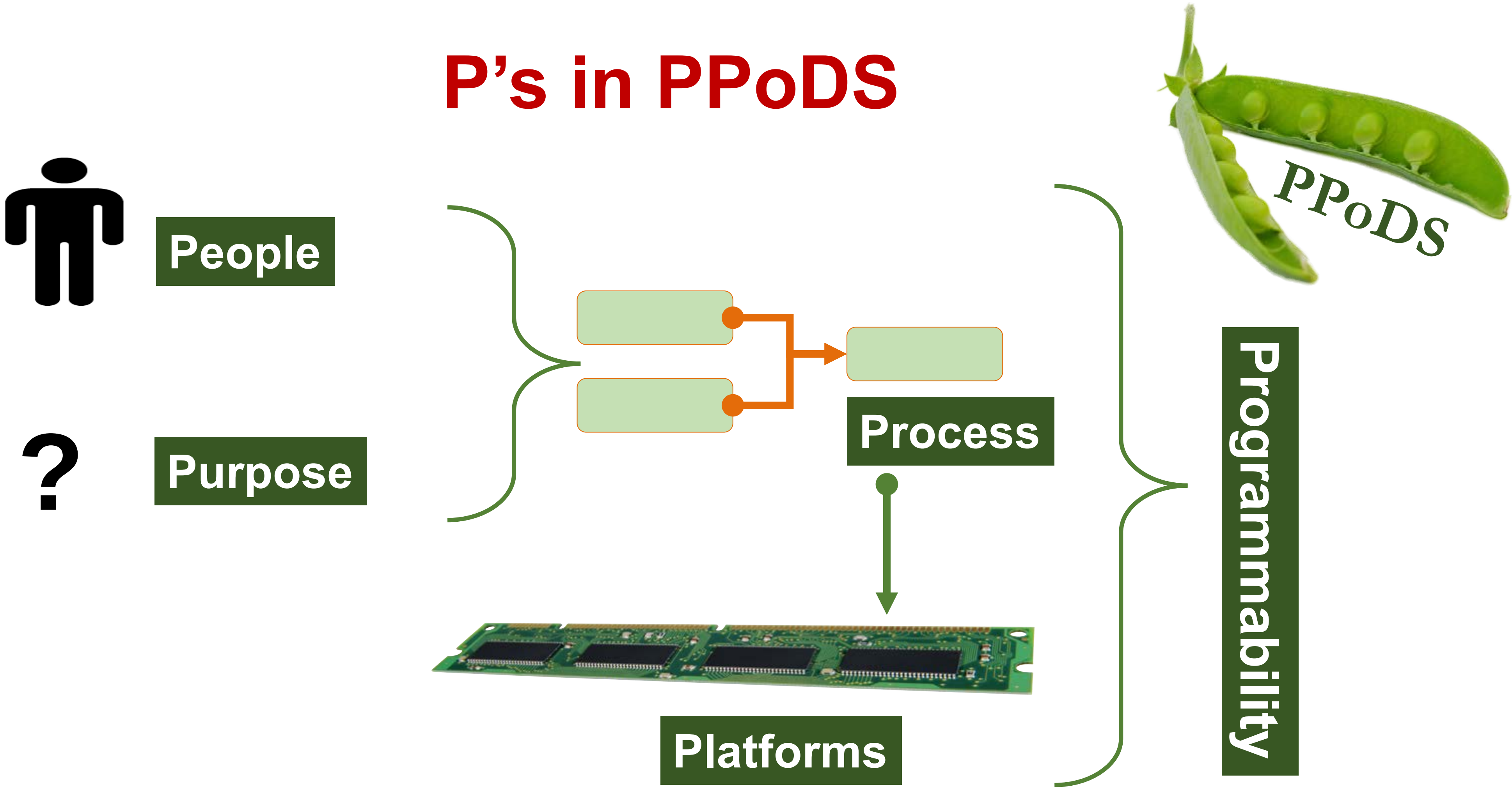
Reproducibility
Ability to validate, re-run, re-play

Scalability
From local experiments to large-scale runs



Process for Practice of Data Science

P's in PPoDS



Example: Using geospatial big data for wildfire predictions

WIFIRE: A Scalable Data-Driven Monitoring, Dynamic Prediction and Resilience Cyberinfrastructure for Wildfires

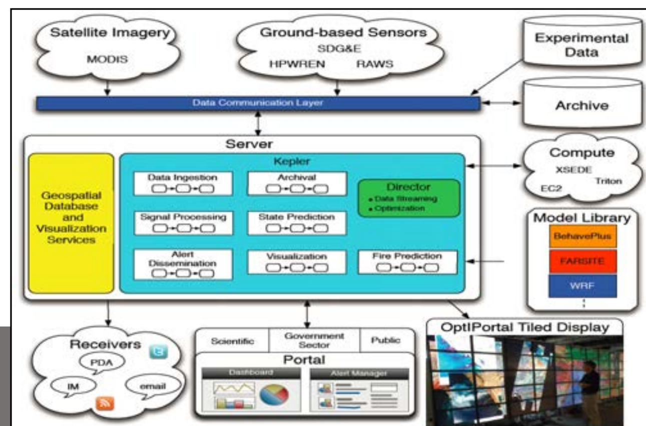
[*http://wifire.ucsd.edu*](http://wifire.ucsd.edu)

Big Data

Monitoring

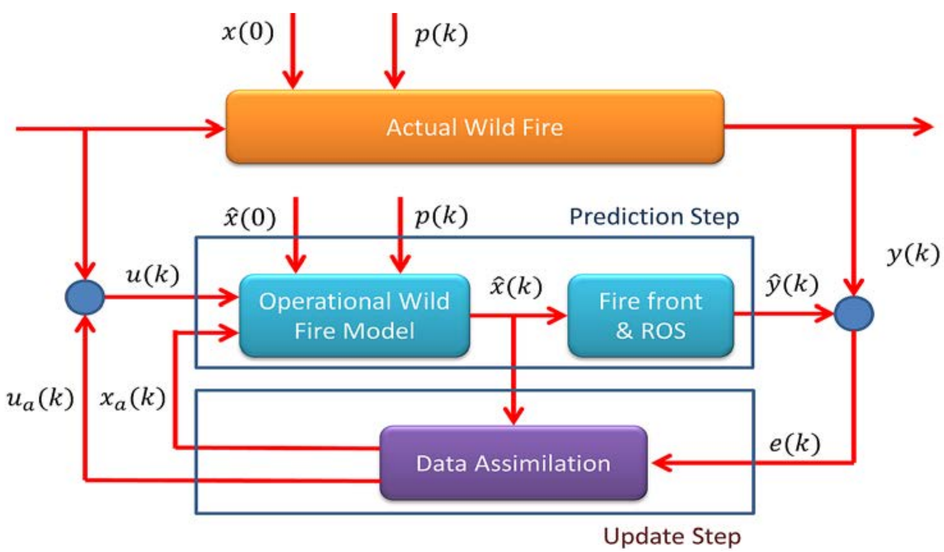
Visualization

Fire Modeling



Closing the Loop using Big Data

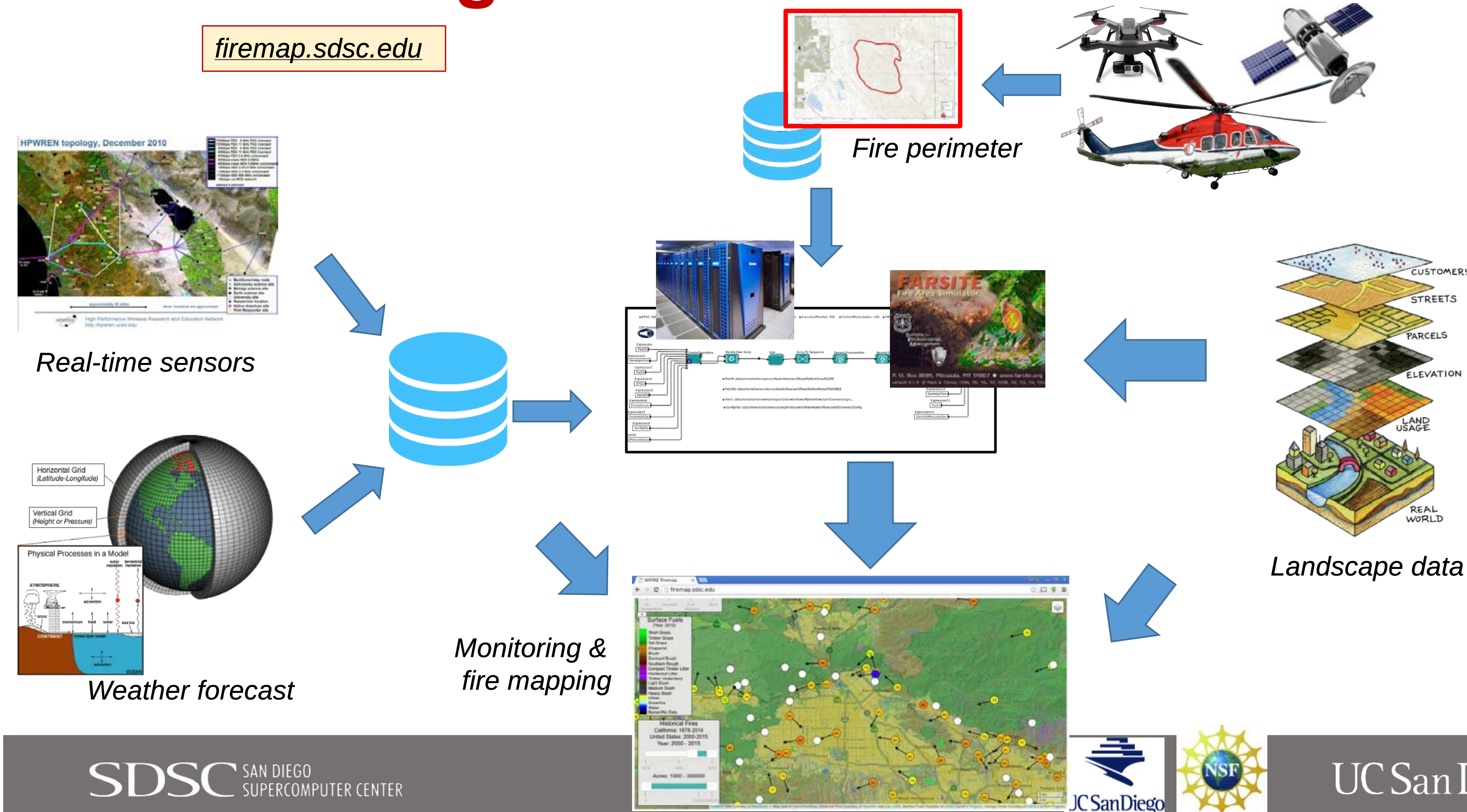
-- Wildfire Behavior Modeling and Data Assimilation --



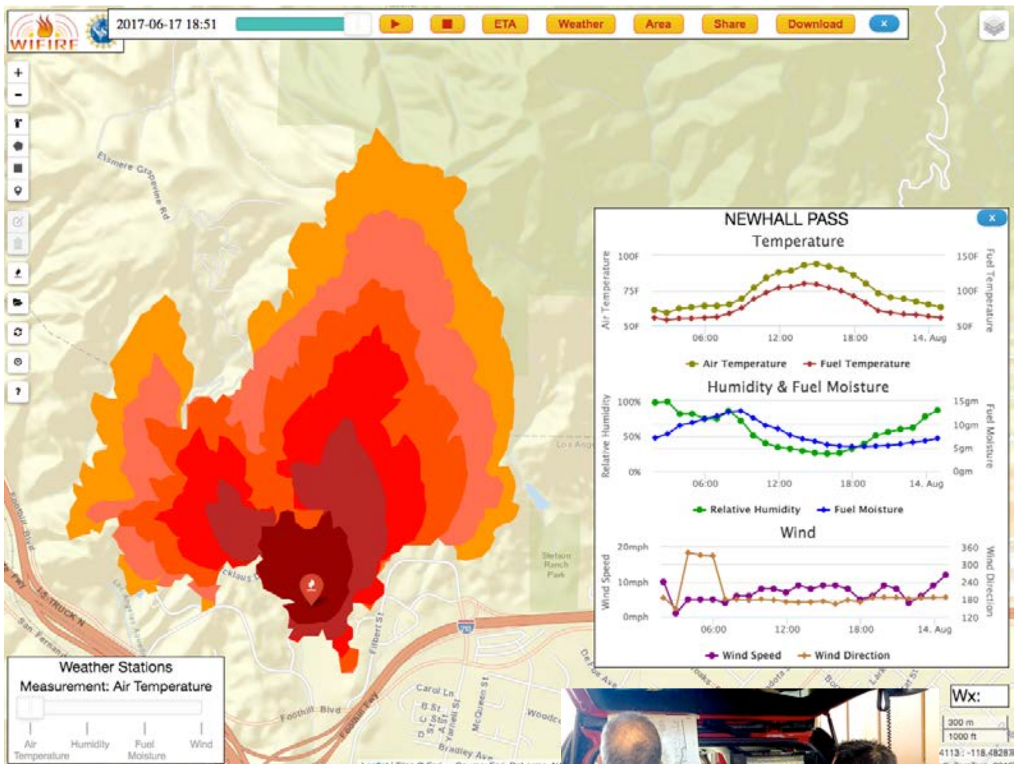
Conceptual Data Assimilation Workflow with Prediction and Update Steps using Sensor Data

- Computational costs for existing models too high for real-time analysis
- *a priori -> a posteriori*
  - Parameter estimation to make adjustments to the (input) parameters
  - State estimation to adjust the simulated fire front location with an a posteriori update/measurement of the actual fire front location

Fire Modeling Workflows in WIFIRE



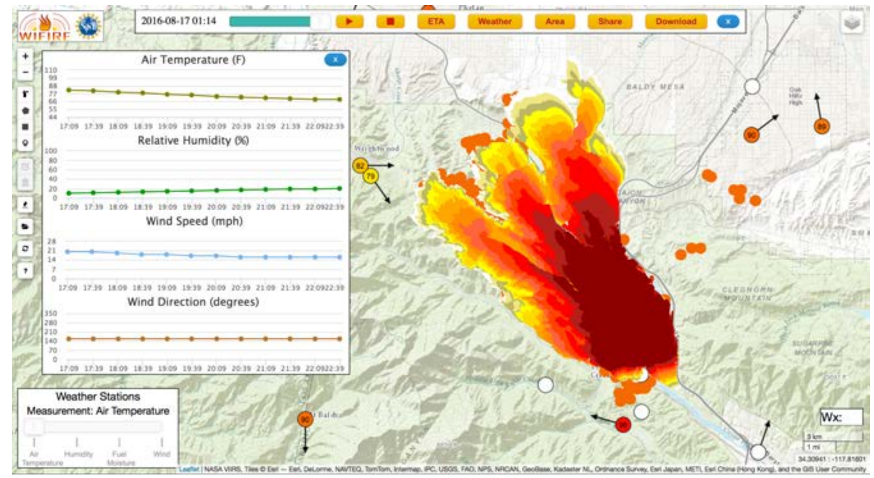
Data-Driven Fire Progression Prediction Over Three Hours

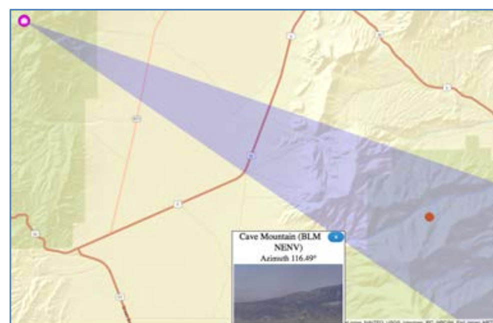


Collaboration with LA and SD Fire Departments

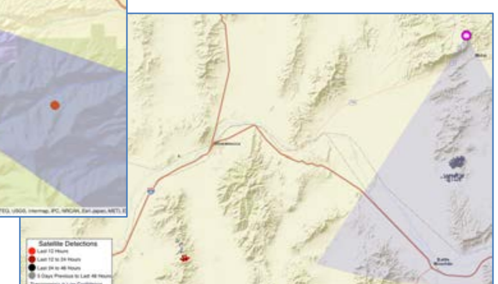
[*http://firemap.sdsc.edu*](http://firemap.sdsc.edu)

August 2016 – Blue Cut Fire





Tahoe and Nevada Bureau of Land Management Cameras: 20 cameras added with field-of-view

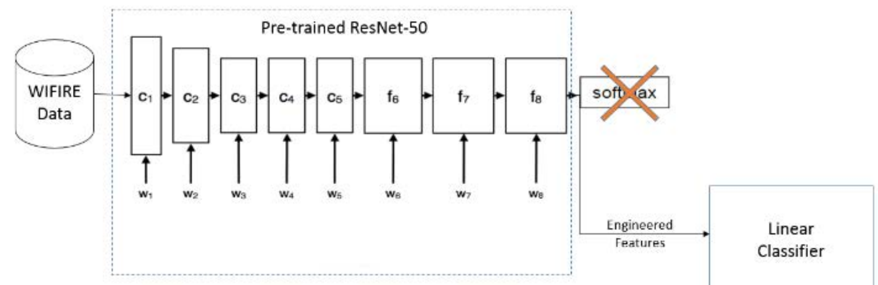


Some Machine Learning Case Studies

- Smoke and fire perimeter detection based on imagery
- Prediction of Santa Ana and fire conditions specific to location
- Prediction of fuel build up based on fire and weather history
- NLP for understanding local conditions based on radio communications
- Deep learning on multi-spectra imagery for high resolution fuel maps
- Classification project to generate more accurate fuel maps (using Planet Labs satellite data)

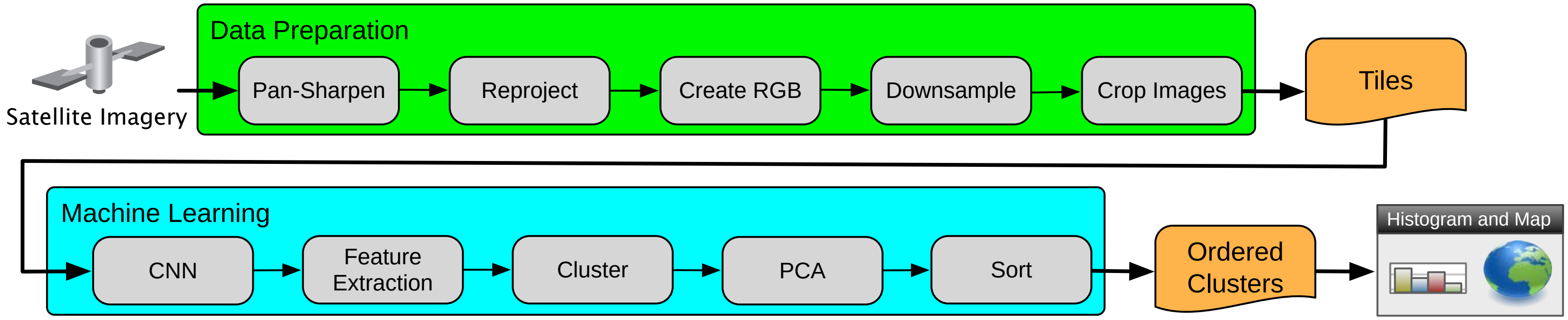
Classification project to generate more accurate fuel maps

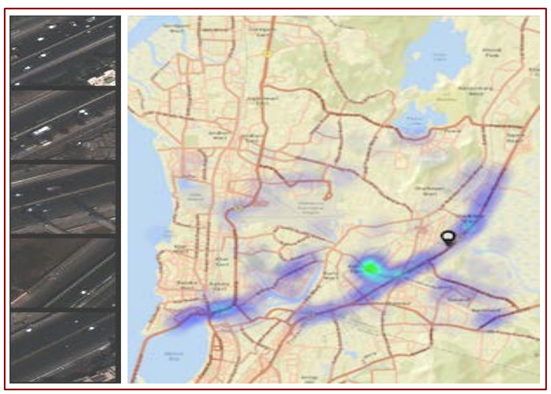
- Accurate and up-to-date fuel maps are critical for modeling wildfire rate of speed and potential burn areas.
- Challenge:
  - USGS Landfire provides the best available fuel maps every two years.
  - The WIFIRE system is limited by these potentially 2-year old inputs. Fuel maps created at a higher temporal frequency is desired.
- Approach:
  - Using high-resolution satellite imagery and deep learning methods, produce surface fuel maps of San Diego County and other regions in Southern California.
  - Use LandFire fuel maps as the target variable, the objective is create a classification model that will provide fuel maps at greater frequency with a measure of uncertainty.

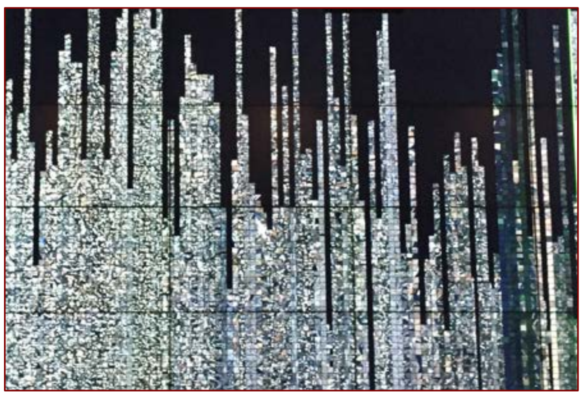


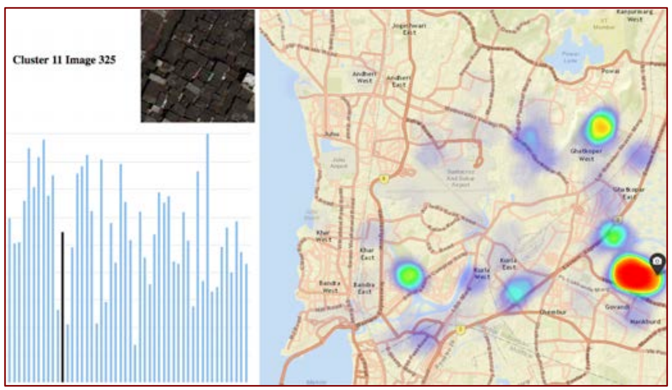
Cluster 1: Short Grass

Reused in Built Infrastructure and Demographic Analysis









Summary

- Geospatial big data has all the typical big data challenges
- Lessons learned from other disciplines to deal with these challenges should be applied
- Workflows can be used both for managing scalable coordination and training students and workforce
- Dynamic data-driven integration of machine learning, data assimilation and modeling is of potential use to many geo applications

WIFIRE Team: It takes a village!

SDSC - *Cyberinfrastructure, Workflows, Data engineering, Machine Learning, Information Visualization, HPWREN*

Calit2/QI*- Cyberinfrastructure, GIS, Advanced Visualization, Machine Learning, Urban Sustainability, HPWREN*

***UMD** - Fire modeling*

***UCSD MAE** - Data assimilation*

***SIO** - HPWREN*

- PhD level researchers
- Professional software developers
- 32 undergraduate students
  - UC San Diego
  - UC Merced
  - Monash University
  - University of Queensland
- 1 high school student
- 4 MSc and 5 MAS students
- 2 PhD students (UMD)
- 1 postdoctoral researcher

Questions?

WorDS Center
Workflows for Data Science Center of Excellence
San Diego Supercomputer Center
Simple and Scalable Data Solutions
WorDS.sdsc.edu

WorDS Director: Ilkay Altintas, Ph.D.
Email: altintas@sdsc.edu

Part of the presented work is funded by NSF, DOE, NIH, UC San Diego and various industry partners.