TITLE:

The Power of Integrated Cyberinfrastructure for Scientific, Societal and Educational Impact: Lessons Learned from the WIFIRE Project

ABSTRACT:

The new era of data science is here. Our lives as well as any field of business and society are continuously transformed by our ability to collect meaningful data in a systematic fashion and turn that into value. These needs not only push for new and innovative capabilities in composable data management and analytical methods that can scale in an anytime anywhere fashion, but also require methods to bridge the gap between applications and compose such capabilities within solution architectures.

Existing cyberinfrastructure provides powerful components that can be utilized as building blocks to translate the newest advances into impactful solution architectures that can transform science, society and education. However, any solution architecture today depends on the effective collaboration of a multi-disciplinary data science team, not only with humans but also with analytical systems and infrastructure which are inter-related parts of the solution. Focusing on collaboration and communication between people, and dynamic, predictable and

programmable interfaces to systems and scalable infrastructure from the beginning of any activity is critical.

This talk will overview some of our recent work on building dynamic data driven cyberinfrastructure and impactful application solution architectures that showcase integration of a variety of existing technologies and collaborative expertise. In particular, the lessons learned from the development of the NSF WIFIRE cyberinfrastructure will be summarized. The WIFIRE project builds an end-to-end cyberinfrastructure for real-time and data-driven simulation, prediction, and visualization of wildfire behavior. WIFIRE's real-time data products and modeling services are routinely accessed by fire research and emergency response communities for modeling as well as the public for situational awareness. Existing sustainability efforts and public/private partnerships as a result of this demand will also be discussed.

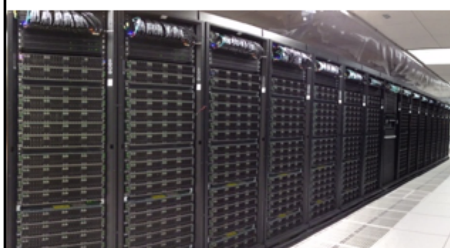
A little about me…

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”

I ♥ BIG DATA
SAN DIEGO SUPERCOMPUTER CENTER at UC SAN DIEGO

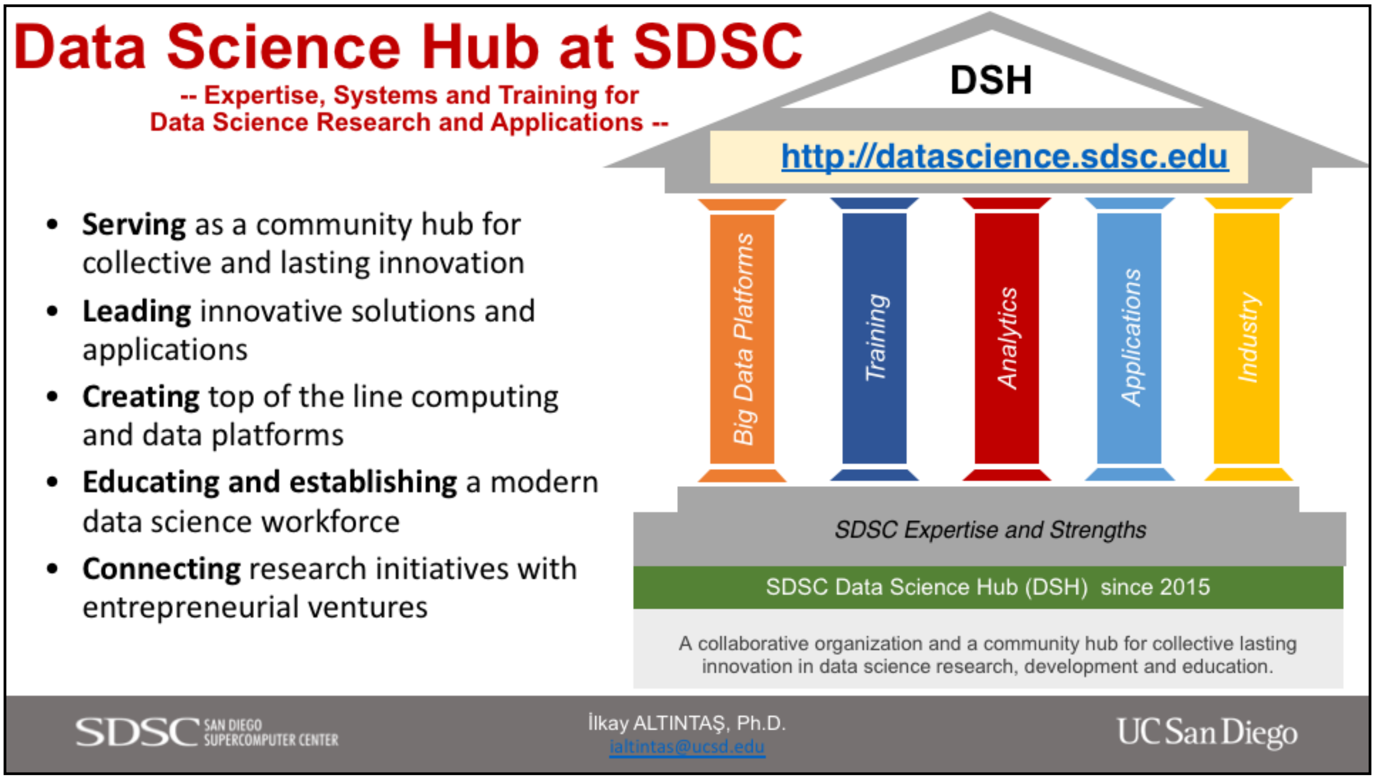


Recent Innovative Architectures

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

SDSC SAN DIEGO SUPERCOMPUTER CENTER

UC San Diego



Serving as a community hub for collective, lasting innovation in data science research and expertise.

Leading innovative solutions and applications for data-driven research, analytics, and development.

Creating top of the line computing and data platforms for the data science community.

Educating and establishing a modern data science workforce that can help drive innovation in public and private organizations.

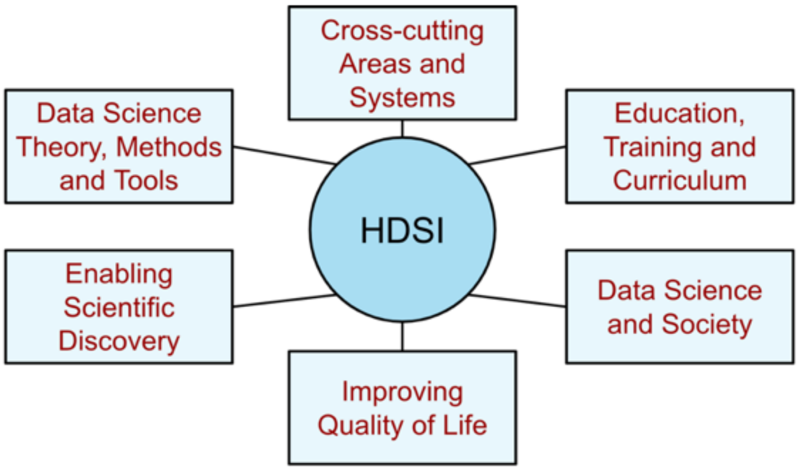
Connecting data science research initiatives with entrepreneurial ventures and potential industry commercialization.

MATCHPOINT is a matching sevice provided by the Data Science Office at SDSC. Its goal is to match Domain experts and students with Methods experts and students to create interdisciplinary working teams that can take on challenging problems in data science and analytics.

The Data Science Business Forum is an interdisciplinary research gateway connecting member companies with other companies and SDSC faculty and staff focused on a common research agenda.

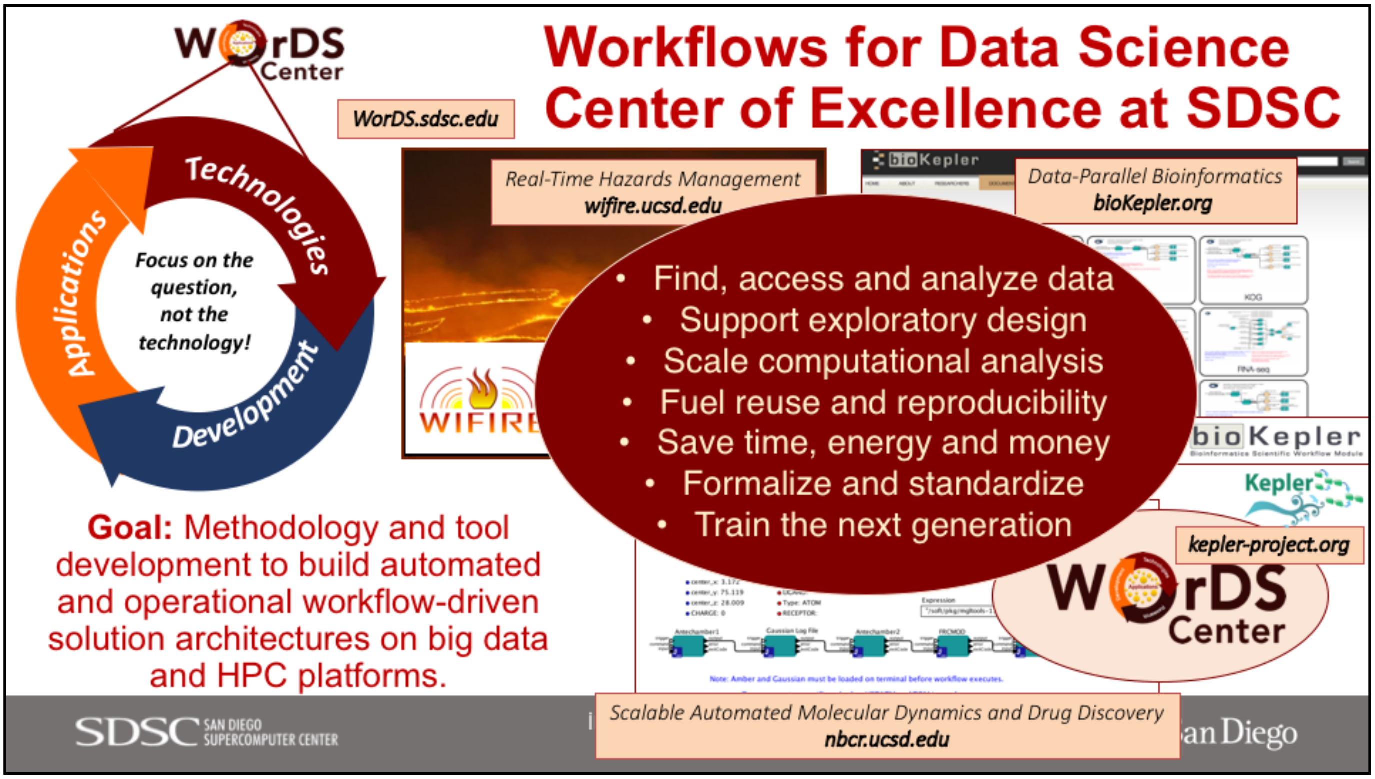
Halıcıoğlu Data Science Institute

https://datascience.ucsd.edu



An academic unit that provides a home for curating the growth in Data Sciences as a discipline

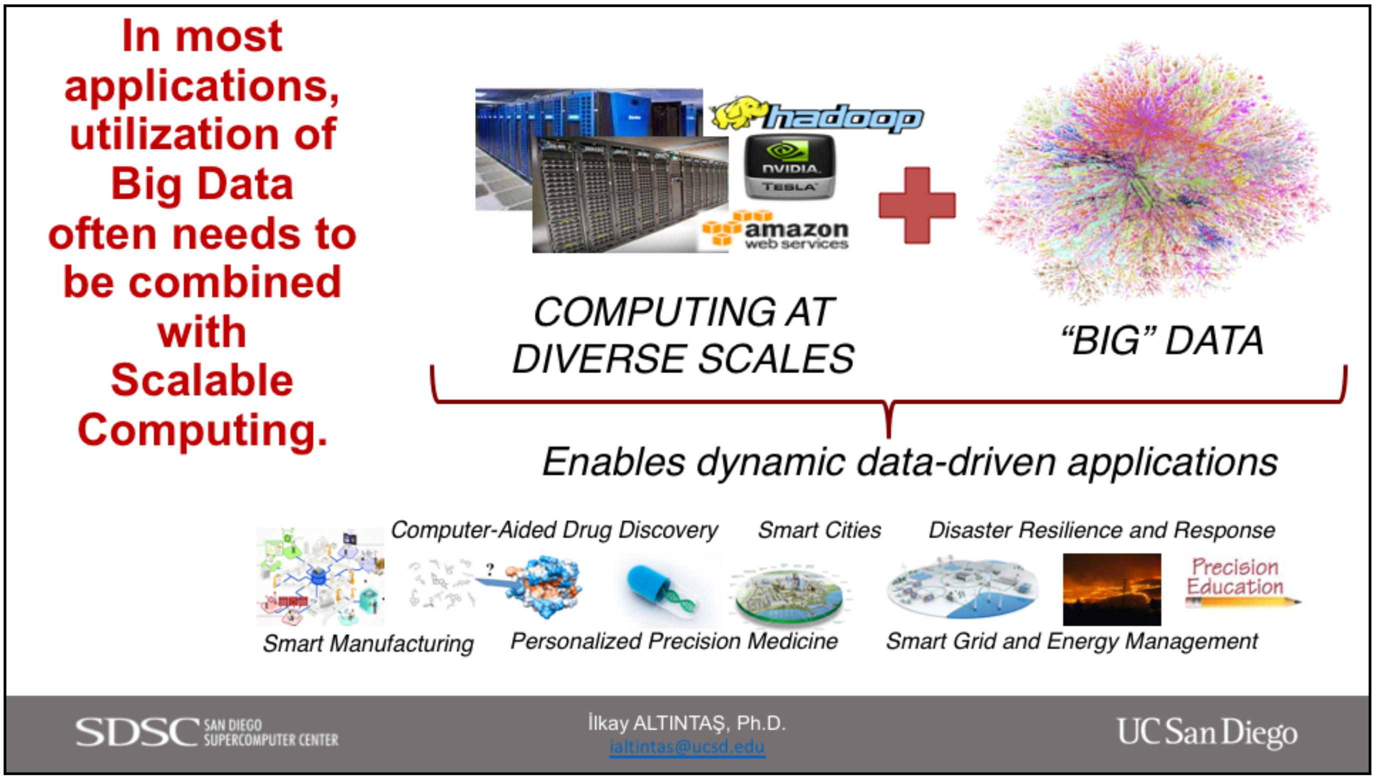
- Deep expertise organized into clusters
- Manage new engagements to seed, cultivate and grow the practice of Data Sciences



Workflows for Data Science Center of Excellence at SDSC
WorDS Center
WorDS.sdsc.edu
Technologies
Applications
Development
Focus on the question, not the technology!
Real-Time Hazards Management
wifire.ucsd.edu
WIFIRE
Data-Parallel Bioinformatics
bioKepler.org
bioKepler
Find, access and analyze data
Support exploratory design
Scale computational analysis
Fuel reuse and reproducibility
Save time, energy and money
Formalize and standardize
Train the next generation
Kepler
kepler-project.org
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
SDSC SAN DIEGO SUPERCOMPUTER CENTER

Common Theme…

"Big" Data, Computational Science, Data Science, Cyberinfrastructure, and Their Applications



Today's computing has diverse workload characteristics spanning high-performance computing, high-throughput computing and big data analytics.

The traditional supercomputing applications are stronger than ever on their way to embrace exascale computing capacity.

As our ability to collect data in real-time from internet-of-things has improved, the demand to process such data at scale has increased requiring big data processing capabilities.

We observe a growing number of applications including smart cities, precision medicine, energy management and smart manufacturing, that require a combination of advanced data analytics with traditional modeling and simulations.

In addition, thanks to the advances in chip design, most scientific codes are ported for special environments like GPUs.

There is also an increasing demand for computing from scientific disciplines like social sciences which weren't traditionally seen as supercomputing disciplines.

In fact, there is no domain of science and engineering today that can't take advantage of big data and computing.

A challenge for today's computing architectures is the ability to respond to such heterogeneous needs and lowering the barriers to computing for long tail researchers as well as supporting the most cutting edge computing applications.

Some supercomputer systems operational at the moment has responded to such needs, including the NSF-funded Comet at the San Diego Supercomputer Center, which was designed as a converged HPC, HTC and BDA platform.

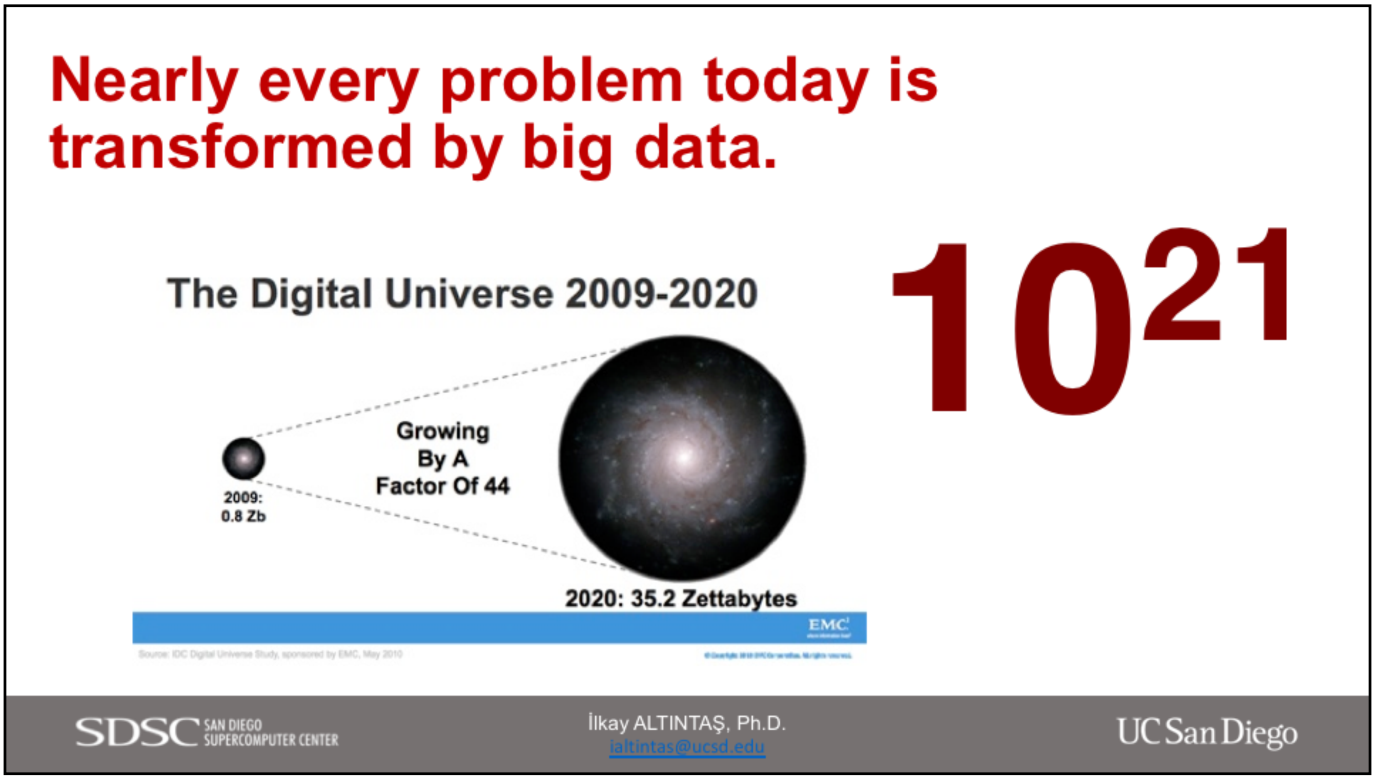
Such heterogeneous capability in computing brings with it the need for software systems that can coordinate applications across different scales of computing, data and networking needs.

A number of software innovations like cluster virtualization, Singularity containers and gateways have enabled system software be more portable and lowered the barriers for many more to take advantage of computing.

However, there is still a need at the converged application level to enable communications with data and computing middleware, while optimizing resources and dynamically adapting to the changes in application workflows.

Even though workflow systems have been a part of the HPC and HTC ecosystem for task coordination and management, we are just discovering their potential for managing dynamic-data driven applications and decision support using advances in big data platforms and on-demand computing systems.

They provide an ideal programming model for deployment of computational and data science applications on all scales of computing and provide a platform for system integration of data, modeling tools and computing while making the applications reusable and reproducible.



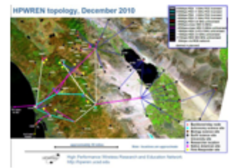
According to a report and predictions by an IDC report sponsored by a big data company called EMC, digital data will grow by a factor of 44 until the year 2020.

This is a growth from .8 Zettabytes in 2009 to 35.2 Zettabytes. A Zettabye is 1 trillion gigabytes, that is 10 to the power of 21. The effects of it will be huge! Think of all the time, cost, and energy that will be used to store and make sense of such an amount of data. The next era will be Yottabytes (10 to the power 24) and Brontobytes (10 to the power 27) which is really hard to imagine for most of us at this time…

This is also what we call data at an astronomical scale. The choice of Milky Way Galaxy in the middle of the circle is not just for esthetics. That is what we would see if we were to scale out 10 to the 21 time into the universe. It is like we are generating a galaxy of data. Cool, isn't it?

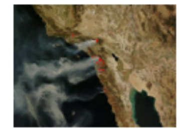
Example: Geospatial Big Data

- 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**

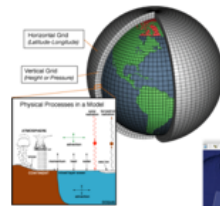


Real-time sensors

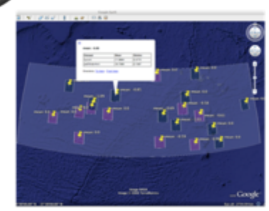
Drone imagery



Satellite imagery



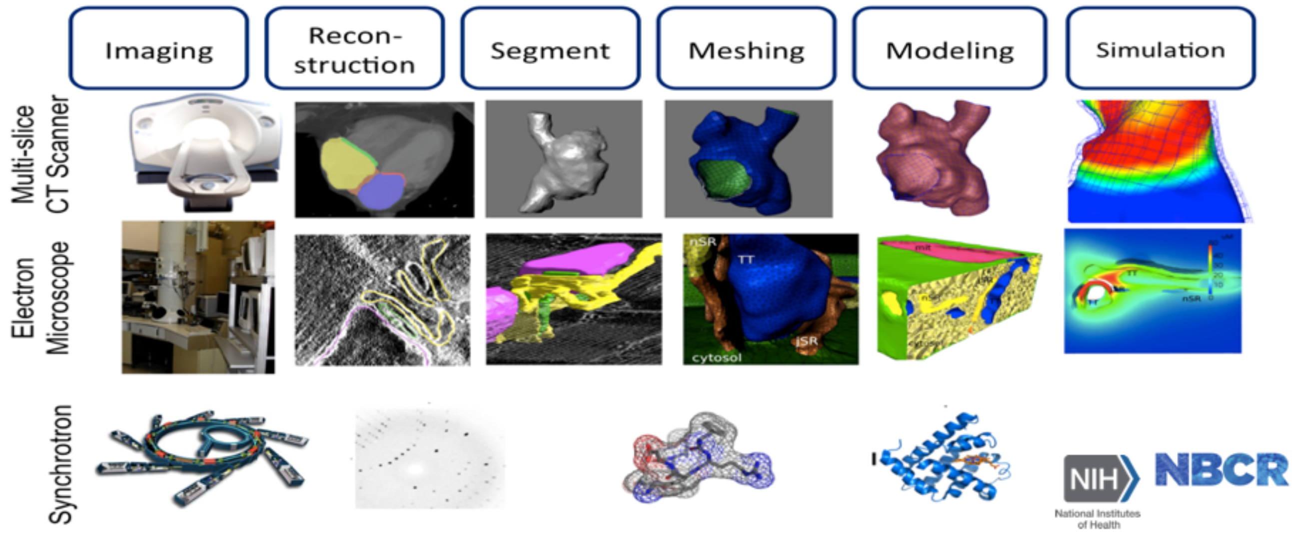
Weather forecast

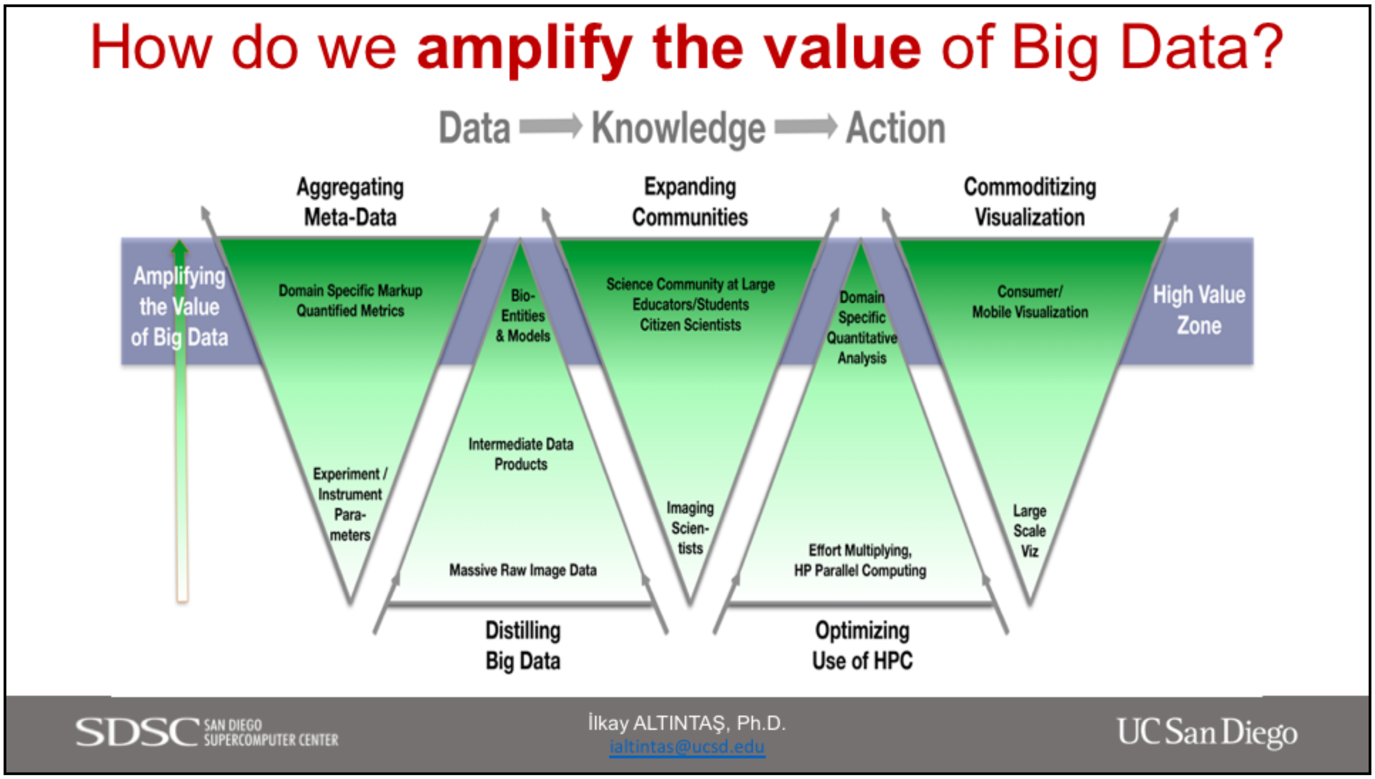


Sea Surface Temperature Measurements

Example: Biomedical Big Data

http://nbcr.ucsd.edu





Any GIS researcher using big data needs to be able to reduce the size and complexity of big data while adding more meaning and value to it with richer metadata, which in turn reduces computational complexity and leads the way for search, access, retrieval and analysis at different levels of scale including mobile devices to large-scale big data and computing systems.

And our challenges isn't just to manage the data, but to try to see how EVERYTHING IS CONNECTED. Finding the connections between the kinds of datasets we've discussed has the potential to lead to interesting discoveries.

Such an endeavor requires, proper use of

- Data management
- Data-driven methods
- Scalable tools for dynamic coordination and scalable execution and
- Skilled interdisciplinary workforce

****

Amplifying the Value of Data Related to **X**

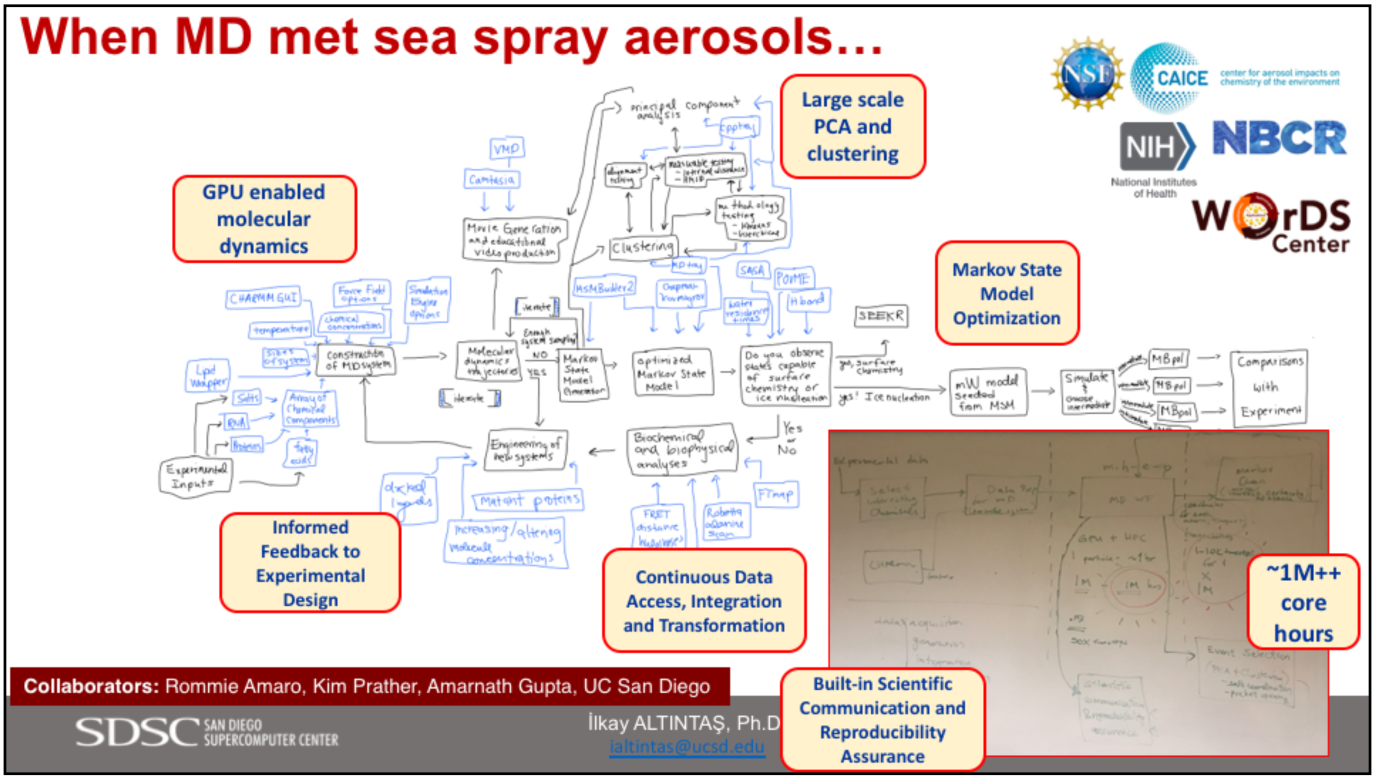
➡

Benefit **Y** for Science, Business, Society or Education

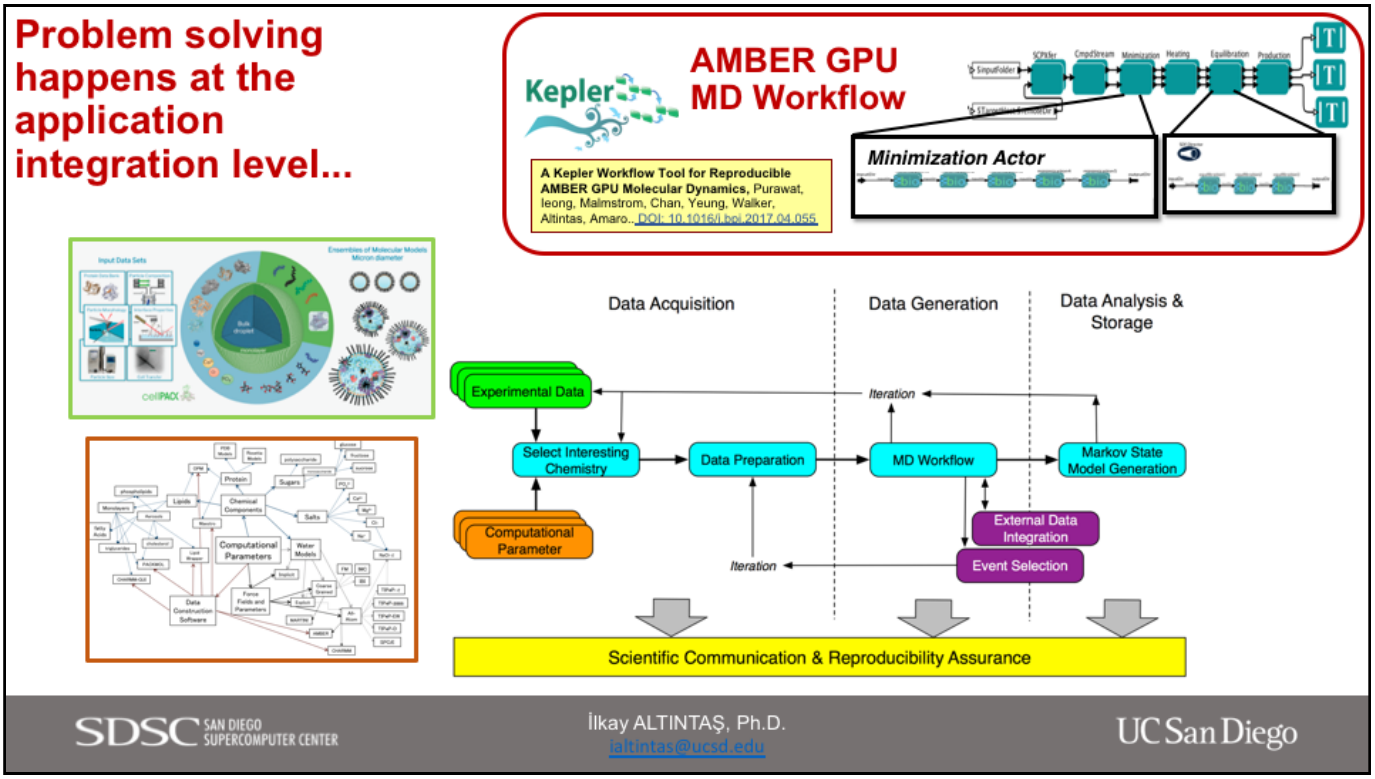
We need to focus on the problems to solve.

Problem solving often requires application integration.

İlkay ALTINTAŞ, Ph.D.
ialtintas@ucsd.edu



Different experimental datasets are being integrated to create a molecular level SSA

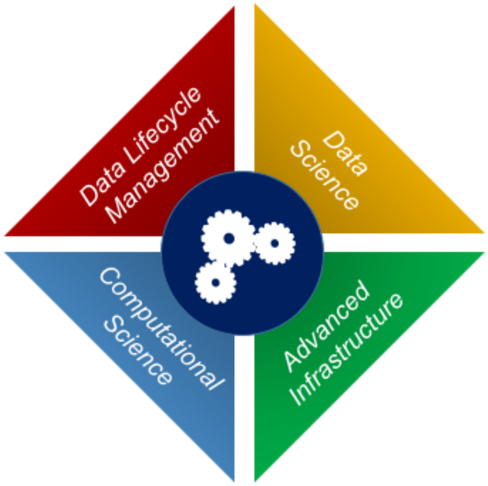


Different experimental datasets are being integrated to create a molecular level SSA

Application integration requires the expertise of a collaborative team.

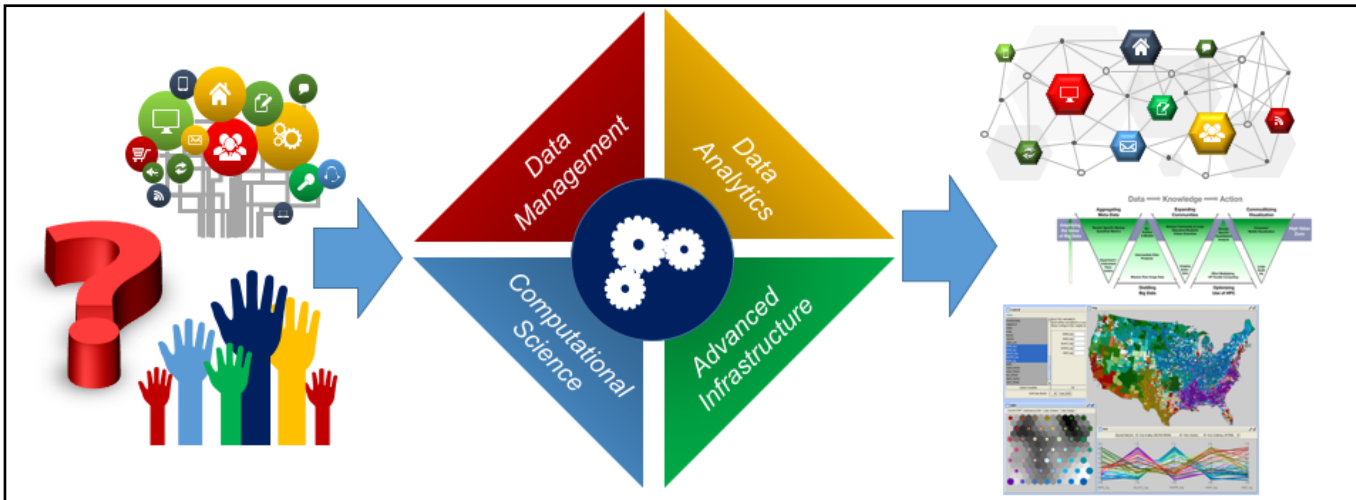
- Multi-disciplinary scientific and technological expertise
- Integration of many scales of computing
- Integration of big and small experimental datasets
  - Can be historical or real time
- Usage of individual or community developed legacy tools
- Methods to manage and interpret data
- Modeling and simulation
- Gateways for visualization and dashboard
- Long term active and passive storage needs

İlkay ALTINTAŞ, Ph.D.
ialtintas@ucsd.edu



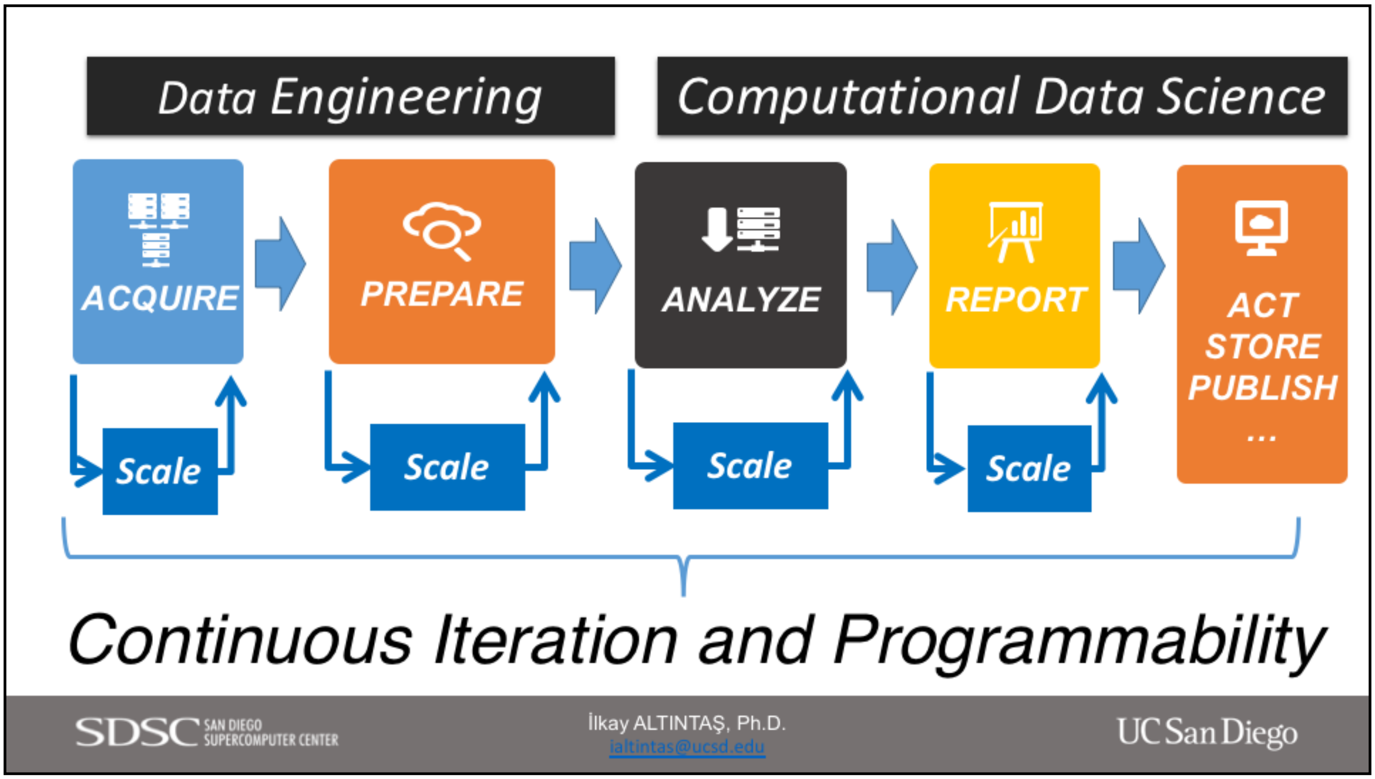
Data-driven problem solving requires:

- Heterogenous systems
- Data management
- Data-driven methods
- Scalable tools for dynamic coordination and resource optimization
- Skilled interdisciplinary workforce
- Collaborative culture and tools that enable groups to communicate



A Typical Collaborative Data Science Ecosystem

İlkay ALTINTAŞ, Ph.D.
ialtintas@ucsd.edu



However, at the end of the day, the scalable Process should be Programmable through utilization of reusable and reproducible programming interfaces to systems middleware, analytical tools, visualization environments, and end user reporting environments.

Create an Ecosystem that Enables Needs and Best Practices

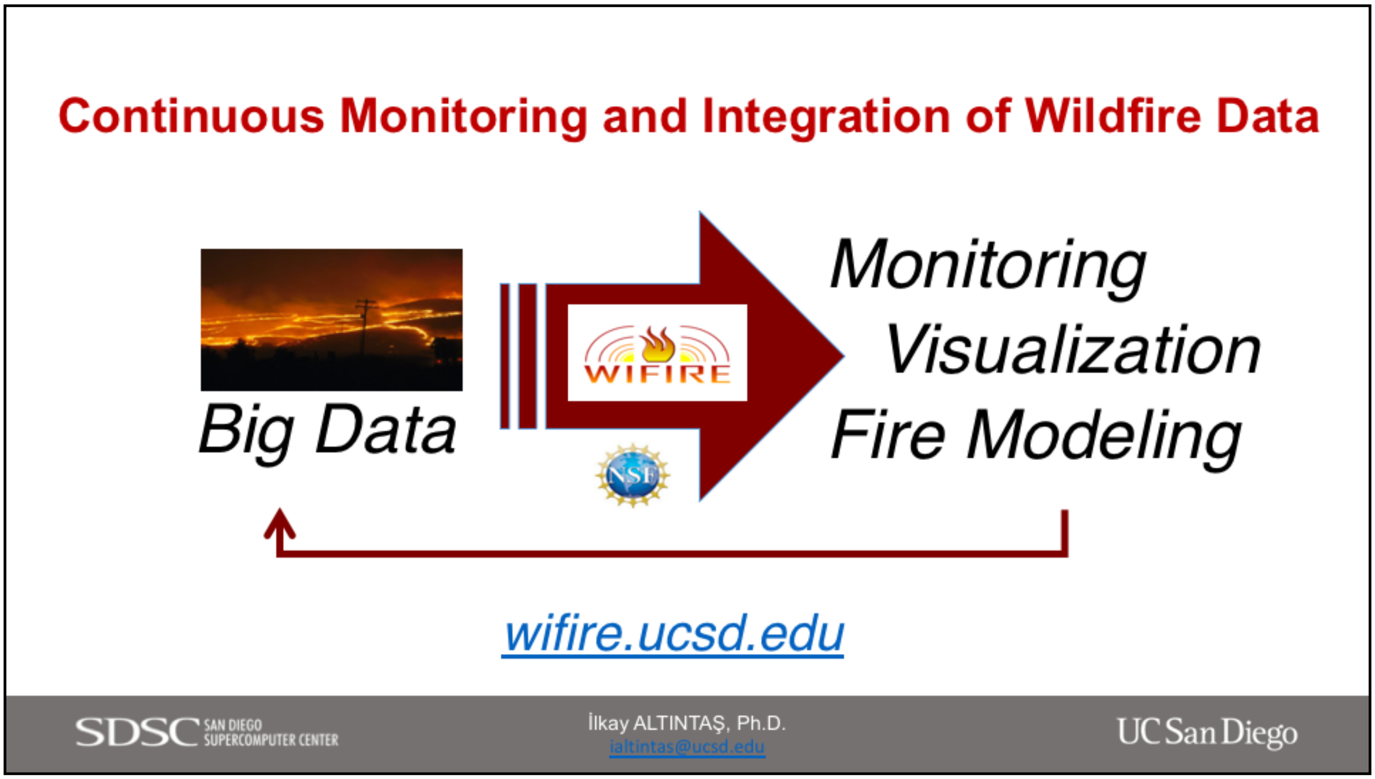
- data-driven
- scalable
- dynamic
- process-driven
- collaborative
- accountable
- reproducible
- interactive
- heterogeneous
- includes many different expertise

Amplifying the Value of Data Related to **X**

Benefit **Y** for Science, Business, Society or Education

What if X was wildfires?

İlkay ALTINTAŞ, Ph.D.
ialtintas@ucsd.edu



An example to such use of workflows is the NSF-funded WIFIRE project I lead.

In WIFIRE, we built a scalable cyberinfrastructure that can utilize any high end computing, cloud and big data platform for dynamic big data-driven fire modeling and prediction. The approach here is to use real-time data to learn about the dynamics of fire behaviour and environment, and assimilate what we learn into a fire model to adapt to the changes in the situation over time.

Workflows were used for system integration and dynamic application scalability in WIFIRE. All the data, models and computing systems being used were there before WIFIRE, but a programmable system integration that can match the application needs were lacking. The expressive workflow models enabled such computing capabilities to become available to fire response, research and planning communities.

In fact, the system is being used as we speak as a situational awareness tool for the Sand fire in LA County.

WIFIRE represents a wide range of applications where real-time big data can be assimilated with modeling and simulation tools for better situational awareness and dynamic decision support.

The word dynamic here represents the ability to adapt to the changes in the environment. The urgency of time to solution is application dependent.

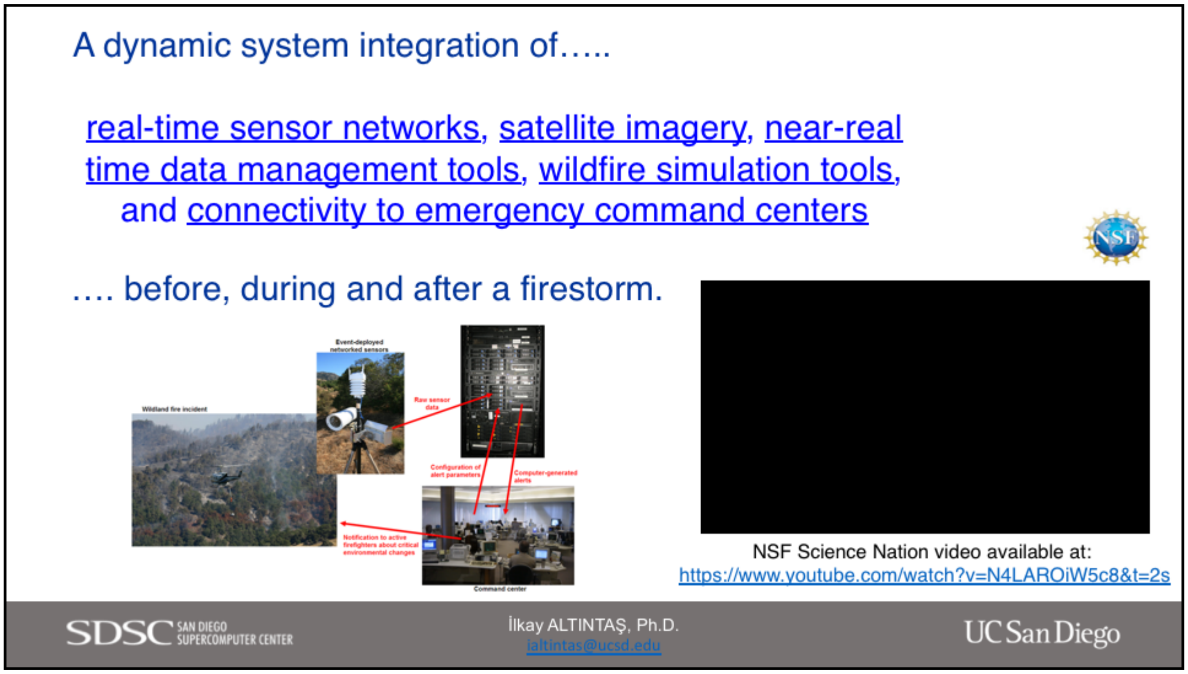
While it is in the order of minutes for a wildfire, it can be less for traffic or energy load management in a smart city and more for a personalized cancer treatment

scenario.

After the success of wifire, and based on initial prototypes for life sciences, smart city, grid and manufacturing applications on the Comet system, SDSC decided to make convergence of IoT streaming data and versatile computing a strategic priority going forward. We are on the vanguard of designing and operating more versatile and capable computing systems, and we anticipate future NSF systems will benefit from the hardware technologies the NCSI will engender.

Situational awareness of and preparedness for wildfires heavily relies on understanding their Direction and Rate of Spread (RoS).

How do we Better Predict Wildfire Behavior?



Make sensor data useful

Large dimension to levels ingestible by analytical and visual platforms

Combine real-time data with physical models

Data-driven predictive and preventive capabilities

Risk assessment, training and dissemination using developed tools

Both municipal and firefighting

ICCS 2015 Best Paper

CENIC 2018 Innovations in Networking Award for Experimental Applications

WIFIRE

An Integrated Cyberinfrastructure for Scalable Data-Driven Monitoring, Dynamic Prediction and Resilience of Wildfires

wifire.ucsd.edu

Best Data-Intensive System (End User focused) — HPCwire Editors' Choice Awards 2014

Best Application of Big Data in HPC — HPCwire Readers' Choice Awards 2014

Best Application of Big Data in HPC — HPCwire Editors' Choice Awards 2014

Ilkay Altintas[1], Jessica Block[2], Daniel Crawl[1], Raymond de Callafon[3], Hans-Werner Braun[1], Michael Gollner[4], Larry Smarr[2], Jurgen Schulze[2] and Arnaud Trouve[4]

[1]San Diego Supercomputer Center, University of California San Diego, U.S.A.
[2]Qualcomm Institute, University of California San Diego, U.S.A.
[3]Dept. of Mechanical and Aerospace Engineering, University of California San Diego, U.S.A.
[4]Fire Protection Engineering Dept., University of Maryland, U.S.A.

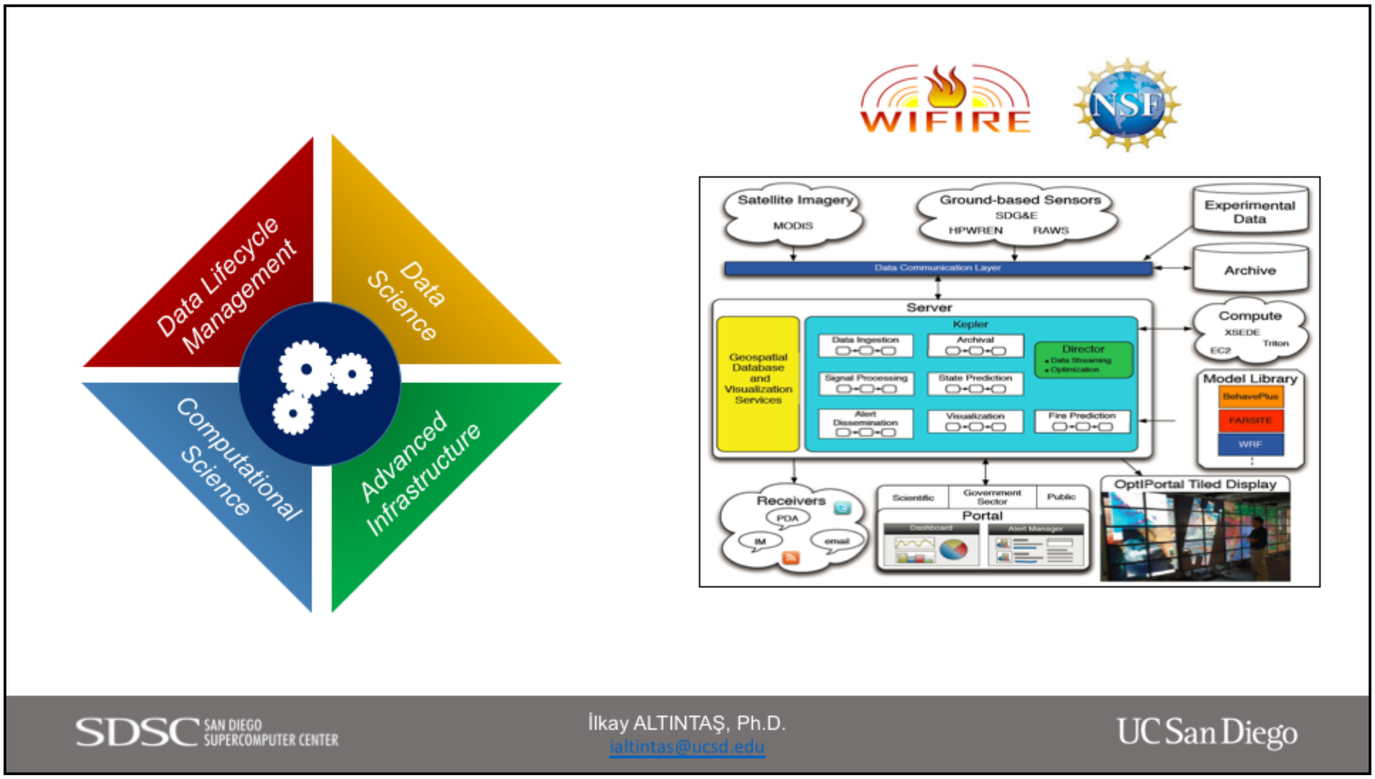
THINK GLOBAL ACT LOCAL

NSF-1331615

SDSC SAN DIEGO SUPERCOMPUTER CENTER

İlkay ALTINTAŞ, Ph.D.
ialtintas@ucsd.edu

UC San Diego



An example to such use of workflows is the NSF-funded WIFIRE project I lead.

In WIFIRE, we built a scalable cyberinfrastructure that can utilize any high end computing, cloud and big data platform for dynamic big data-driven fire modeling and prediction. The approach here is to use real-time data to learn about the dynamics of fire behaviour and environment, and assimilate what we learn into a fire model to adapt to the changes in the situation over time.

Workflows were used for system integration and dynamic application scalability in WIFIRE. All the data, models and computing systems being used were there before WIFIRE, but a programmable system integration that can match the application needs were lacking. The expressive workflow models enabled such computing capabilities to become available to fire response, research and planning communities.

In fact, the system is being used as we speak as a situational awareness tool for the Sand fire in LA County.

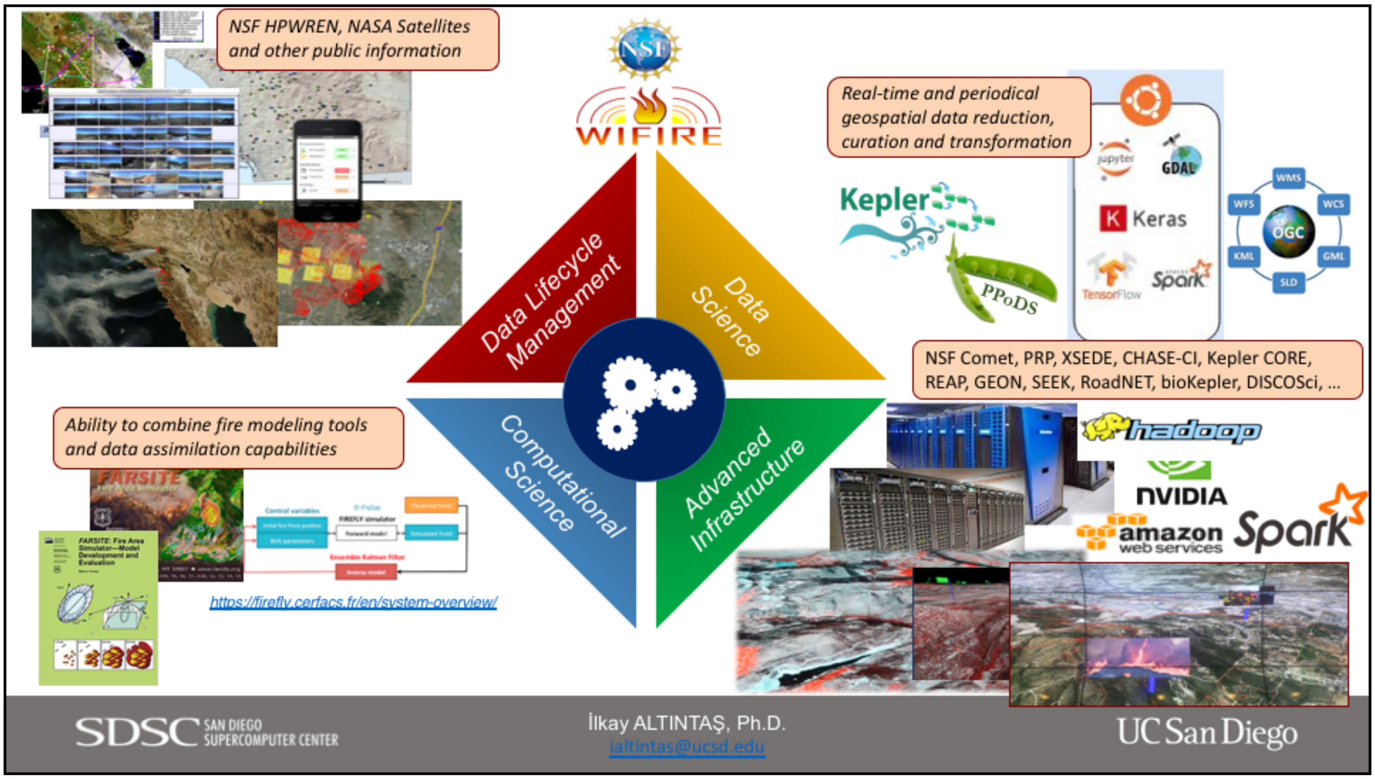
WIFIRE represents a wide range of applications where real-time big data can be assimilated with modeling and simulation tools for better situational awareness and dynamic decision support.

The word dynamic here represents the ability to adapt to the changes in the environment. The urgency of time to solution is application dependent.

While it is in the order of minutes for a wildfire, it can be less for traffic or energy load management in a smart city and more for a personalized cancer treatment

scenario.

After the success of wifire, and based on initial prototypes for life sciences, smart city, grid and manufacturing applications on the Comet system, SDSC decided to make convergence of IoT streaming data and versatile computing a strategic priority going forward. We are on the vanguard of designing and operating more versatile and capable computing systems, and we anticipate future NSF systems will benefit from the hardware technologies the NCSI will engender.



An example to such use of workflows is the NSF-funded WIFIRE project I lead.

In WIFIRE, we built a scalable cyberinfrastructure that can utilize any high end computing, cloud and big data platform for dynamic big data-driven fire modeling and prediction. The approach here is to use real-time data to learn about the dynamics of fire behaviour and environment, and assimilate what we learn into a fire model to adapt to the changes in the situation over time.

Workflows were used for system integration and dynamic application scalability in WIFIRE. All the data, models and computing systems being used were there before WIFIRE, but a programmable system integration that can match the application needs were lacking. The expressive workflow models enabled such computing capabilities to become available to fire response, research and planning communities.

In fact, the system is being used as we speak as a situational awareness tool for the Sand fire in LA County.

WIFIRE represents a wide range of applications where real-time big data can be assimilated with modeling and simulation tools for better situational awareness and dynamic decision support.

The word dynamic here represents the ability to adapt to the changes in the environment. The urgency of time to solution is application dependent.

While it is in the order of minutes for a wildfire, it can be less for traffic or energy load management in a smart city and more for a personalized cancer treatment

scenario.

After the success of wifire, and based on initial prototypes for life sciences, smart city, grid and manufacturing applications on the Comet system, SDSC decided to make convergence of IoT streaming data and versatile computing a strategic priority going forward. We are on the vanguard of designing and operating more versatile and capable computing systems, and we anticipate future NSF systems will benefit from the hardware technologies the NCSI will engender.

Using workflows for process integration and maps as an enabling interface.

Kepler

https://kepler-project.org/

SECURITY

PROVENANCE

DATA LIFECYCLE MANAGEMENT

ENABLING INTERFACES
e.g., Gateways and other online tools for research and teaching

WORKFLOW MANAGEMENT
e.g., Application Integration, Coordination, Optimization, Communication, Reporting

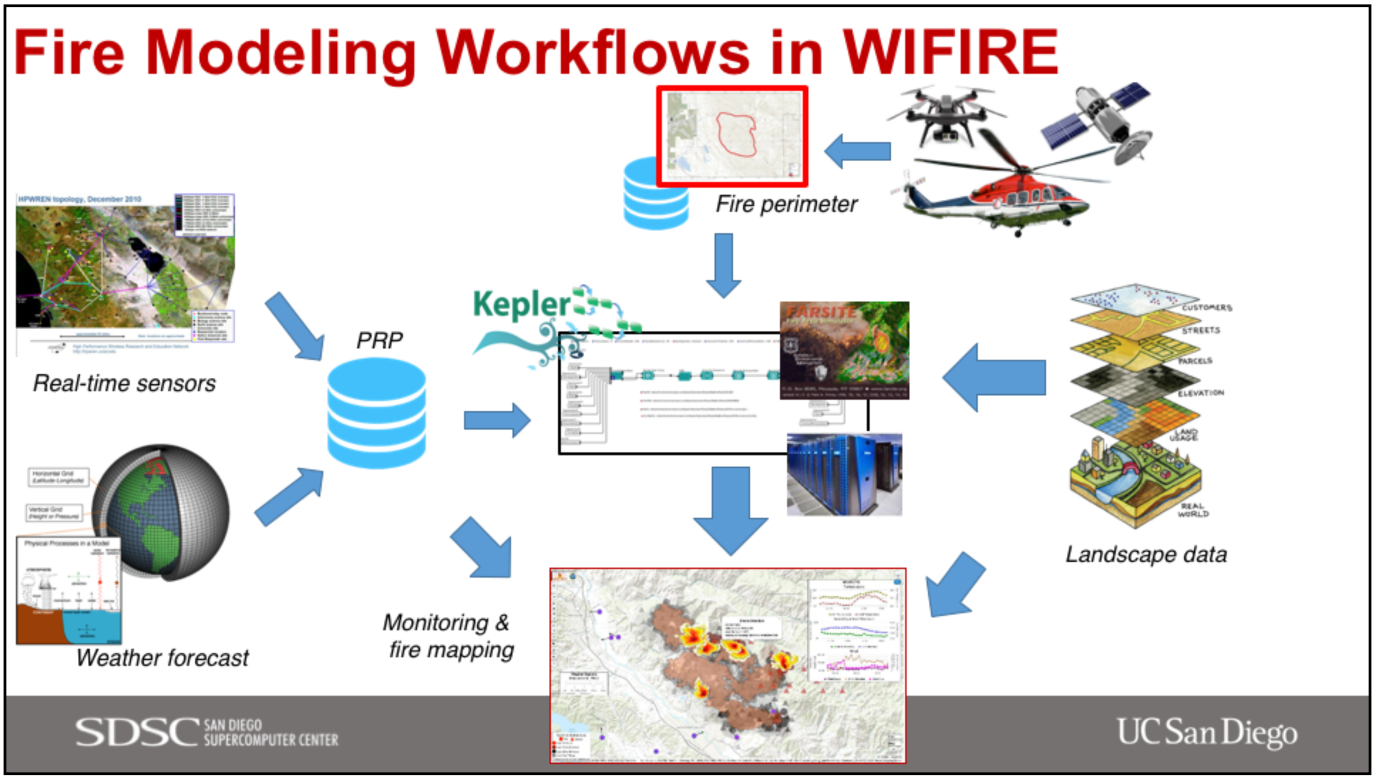
COMPOSABLE SERVICES
e.g., Model and Data Archives, Deep Learning, Analytics, HPC, Training, Notebooks

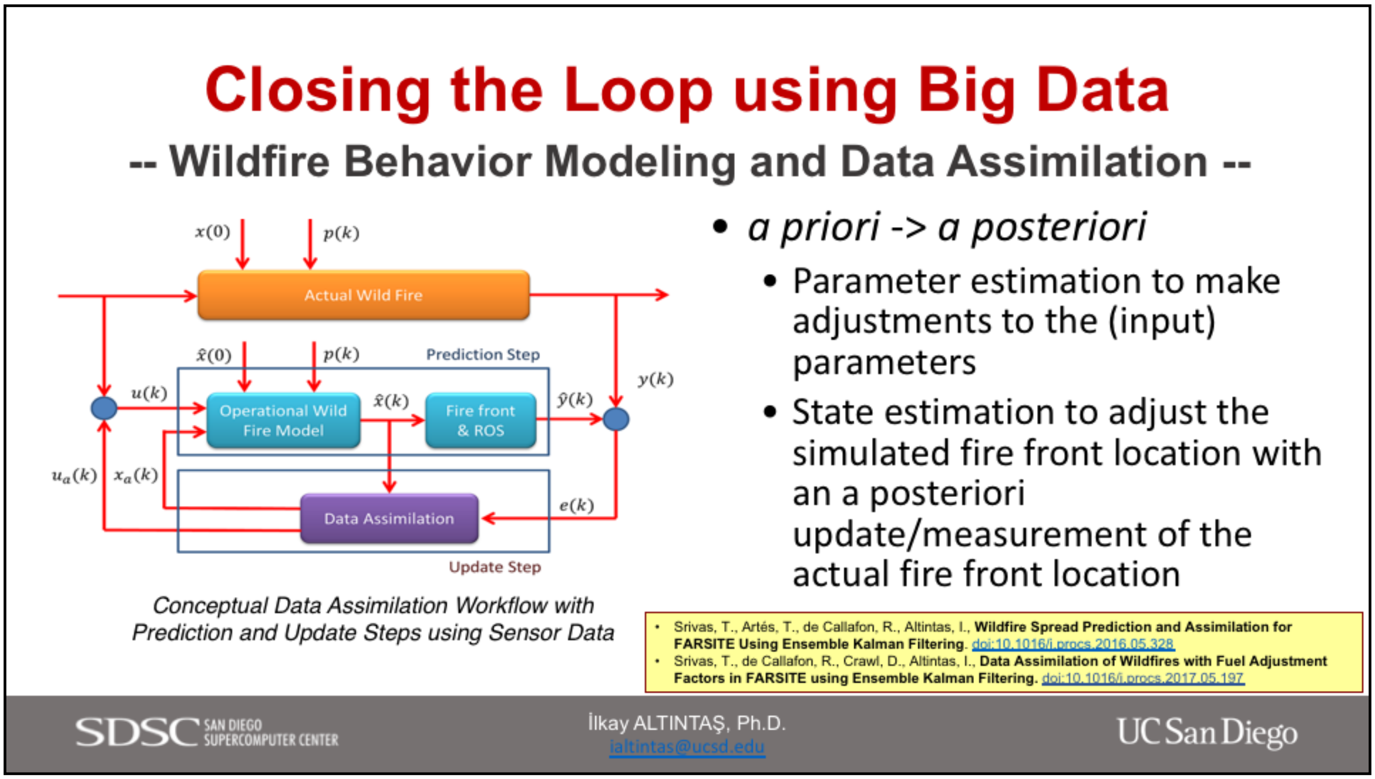
RESOURCE MANAGEMENT
e.g., Kubernetes Container Cloud

COMPOSABLE SYSTEMS
e.g., GPU, CPU, Big Data, Neuromorphic, Networks, Storage, …

COLLABORATION MANAGEMENT

İlkay ALTINTAŞ, Ph.D.
ialtintas@ucsd.edu


Fire Modeling Workflows in WIFIRE
Fire perimeter
Real-time sensors
PRP
Kepler
FARSITE
Landscape data
CUSTOMERS
STREETS
PARCELS
ELEVATION
LAND USAGE
REAL WORLD
Weather forecast
Monitoring &
fire mapping
SDSC SAN DIEGO SUPERCOMPUTER CENTER
UC San Diego



Spatial information on topography, fuel content and moisture along with regional weather and wind input is used to drive the semi-empirical models. This allows the fire spread model to adopt a regional scale perspective and simulate a wildfire as a propagating front for long time periods under heterogeneous conditions of terrain, fuel parameters, and weather conditions.

Srivas, T., de Callafon, R., Crawl, D., Altintas, I., Data Assimilation of Wildfires with Fuel Adjustment Factors in FARSITE using Ensemble Kalman Filtering, In Proceedings of the Workshop on Data-Driven Computational Sciences (DDCS) at the 17th International Conference on Computational Science (ICCS 2017), 2017.

Srivas, T., Artés, T., de Callafon, R., Altintas, I., Wildfire Spread Prediction and Assimilation for FARSITE Using Ensemble Kalman Filtering, In the Data-Driven Computational Sciences Workshop at the 16th International Conference on Computational Science (ICCS 2016). doi:10.1016/j.procs.2016.05.328

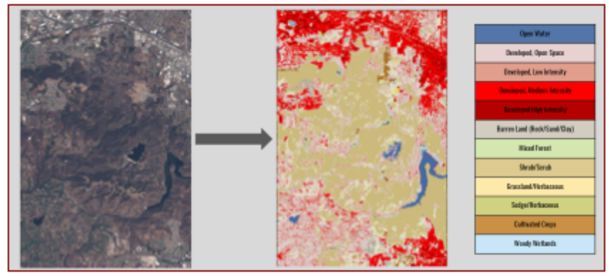
Some Machine Learning Case Studies

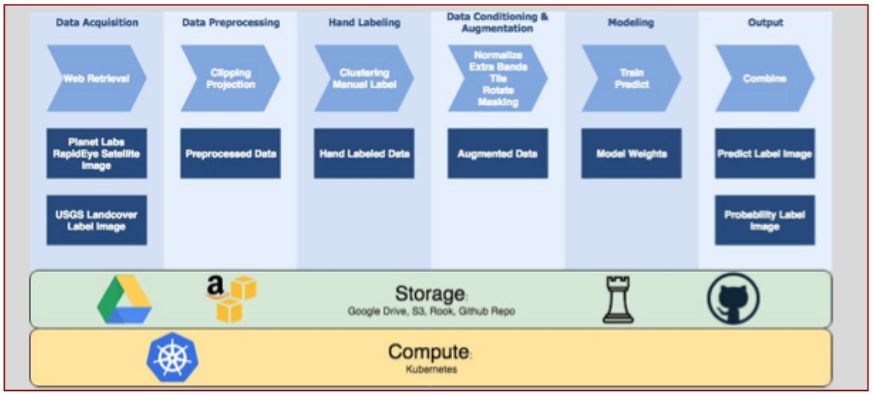
- Smoke and fire perimeter detection based on imagery
- Prediction of cific to location
- Prediction of ather history
- NLP for under n radio communicati
- Deep learning h resolution fuel maps
- Classification e fuel maps (using Planet Labs satellite data)

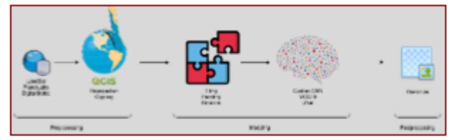
All require periodic, dynamic and programmatic access to data!

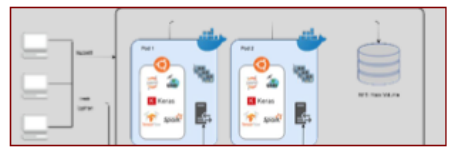
İlkay ALTINTAŞ, Ph.D.
ialtintas@ucsd.edu

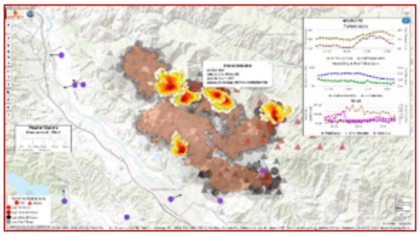
One Piece of the Puzzle: Vegetation Classification using Satellite Imagery





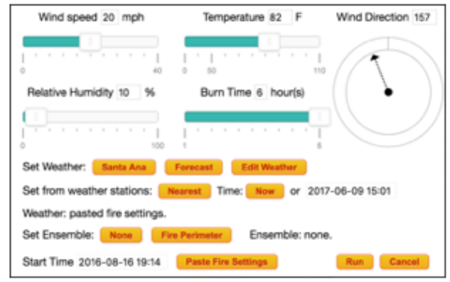




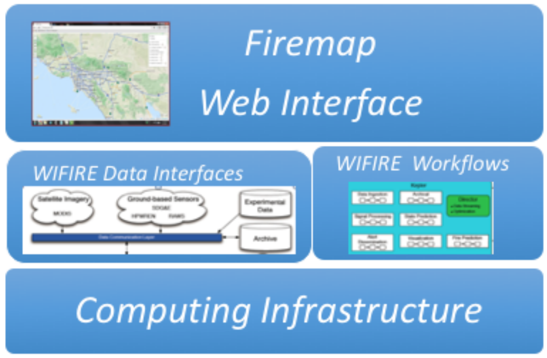


Firemap Tool

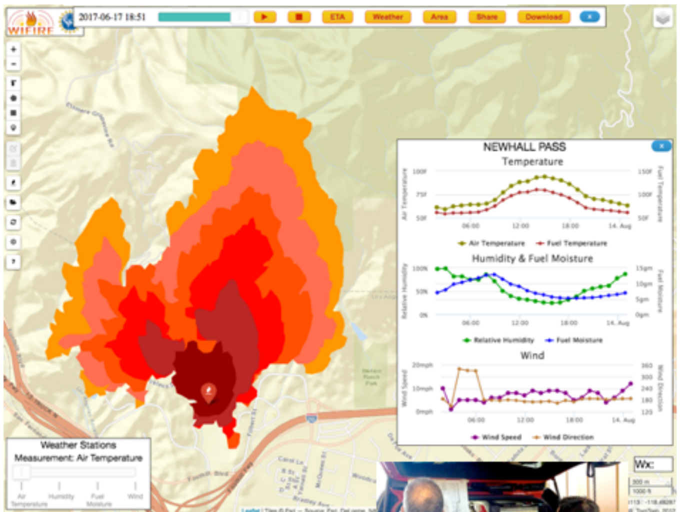
http://firemap.sdsc.edu



- A web-based GIS environment:
  - access information related to fire behavior
  - analyze what-if scenarios
  - model real-time fire behavior
  - generate sharable reports
- Powered by WIFIRE
- Developed in partnership with LAFD

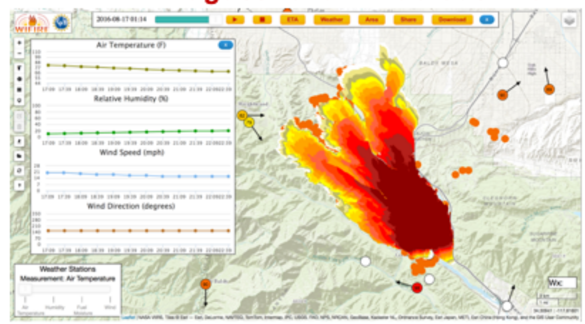


Data-Driven Fire Progression Prediction and Visualization



Collaboration with LA and SD Fire Departments
http://firemap.sdsc.edu

August 2016 – Blue Cut Fire



Tahoe and Nevada Bureau of Land Management Cameras

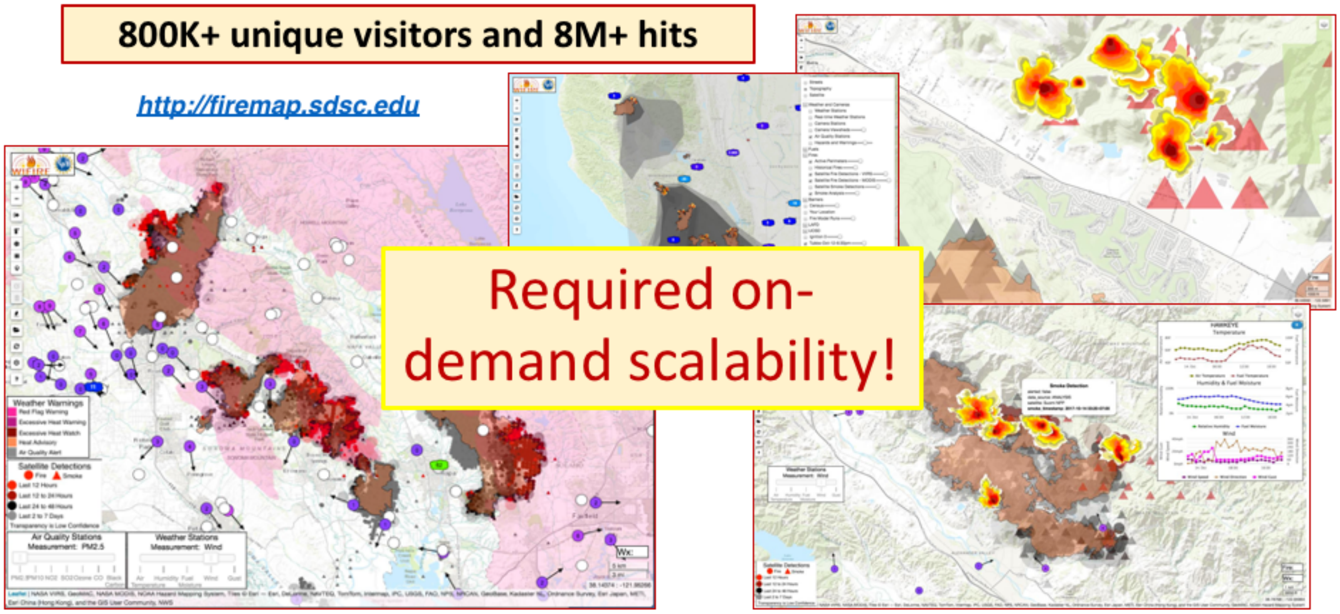

WIFIRE Team: It takes a village!
SDSC - Cyberinfrastructure, Workflows, Data engineering, Machine Learning, Information Visualization, HPWREN
Calit2/QI- Cyberinfrastructure, GIS, Advanced Visualization, Machine Learning, Urban Sustainability, HPWREN
UMD - Fire modeling
UCSD MAE - Data assimilation
SIO - HPWREN
• PhD level researchers
• Professional software developers
• 31 undergraduate students
• UC San Diego
• UC Merced
• UC Berkeley
• MURPA University
• University of Queensland
• 1 high school student
• 6 MSc and 10 MAS students
• 2 PhD students (UMD)
• 1 postdoctoral researcher
• Partners from fire departments
• Advisory board with diverse expertise and affiliations
SDSC SAN DIEGO SUPERCOMPUTER CENTER
İlkay ALTINTAŞ, Ph.D.
ialtintas@ucsd.edu
UC San Diego

What is the impact of such a system?

The Public Impact Example: CA Fires 10/2017 through 12/2017

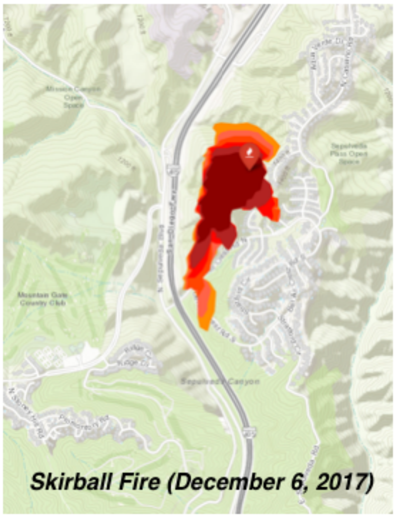
800K+ unique visitors and 8M+ hits

http://firemap.sdsc.edu

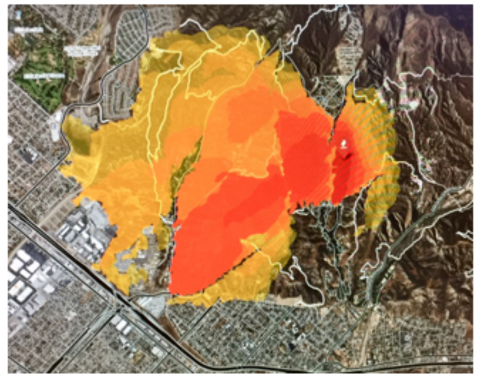


Required on-demand scalability!

The Public Impact Example: Collaboration with fire departments

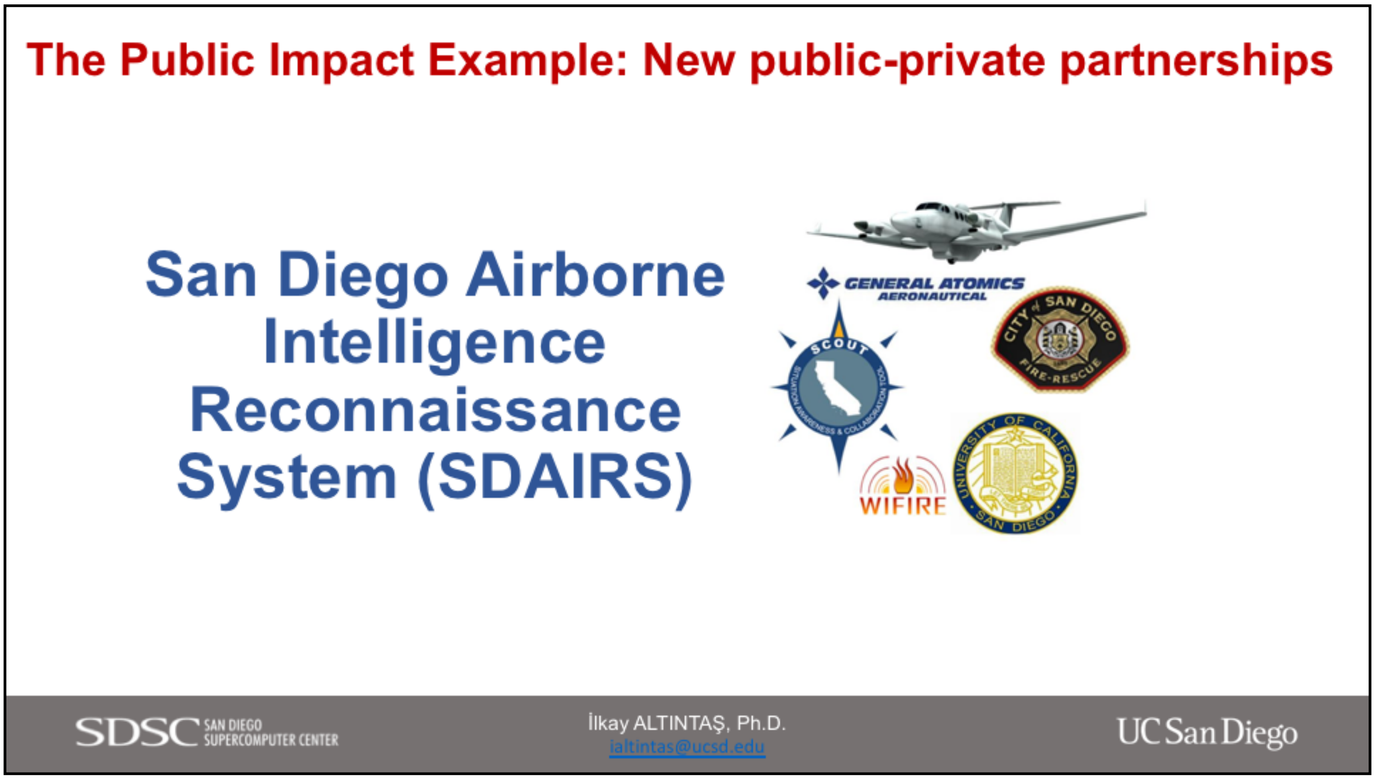


Skirball Fire (December 6, 2017)



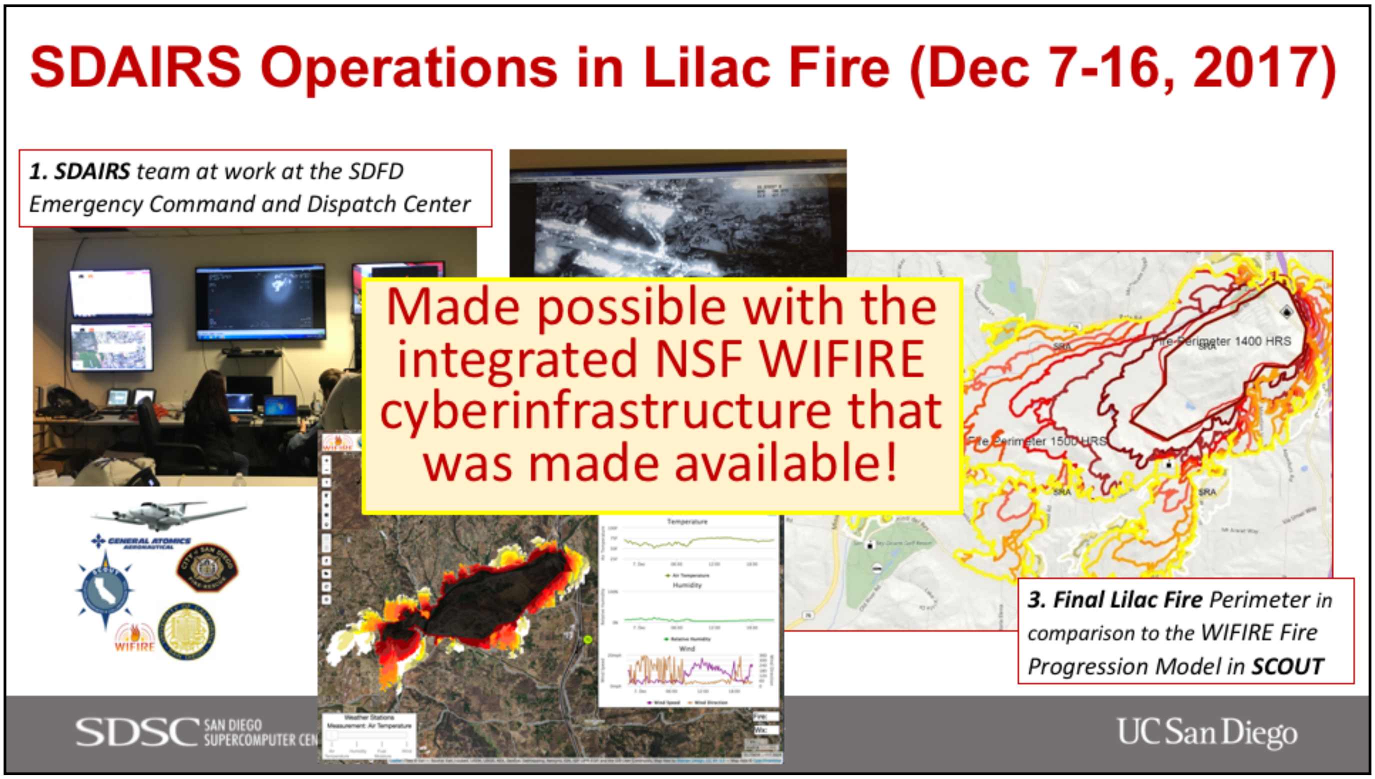
Creek Fire (December 5, 2017)

Some LA County Fires in December 2017



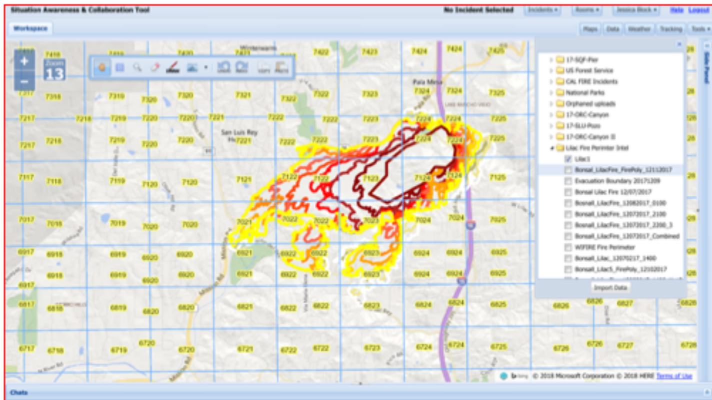
White – final perimeter

Colors – fire progression from wifire in half hour increments

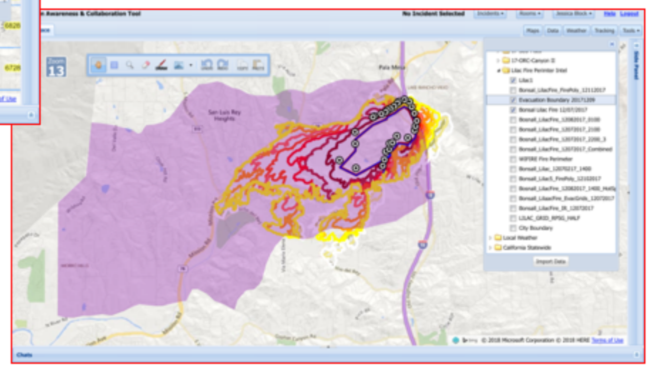


White – final perimeter

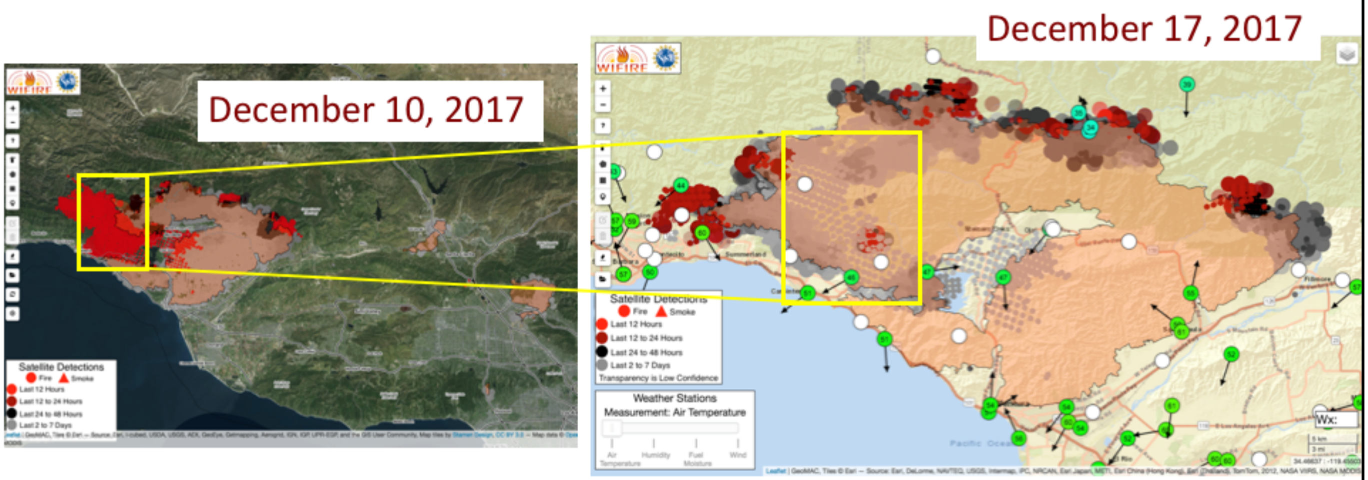
Colors – fire progression from wifire in half hour increments



Publishing the models in SCOUT enabled integration with operational data.



Thomas Fire: 12/04/2017- 01/12/2018



Real-time Satellite Detections During Thomas Fire: 12/04/2017- 01/12/2018

İlkay ALTINTAŞ, Ph.D.
ialtintas@ucsd.edu

Red is most recent, the colors change over time and become grey after 72 hours.

You can see 3-4 santa ana events over the course of the month just based on fire detections.

Detections are from VIIRS approximately every 12 hours.

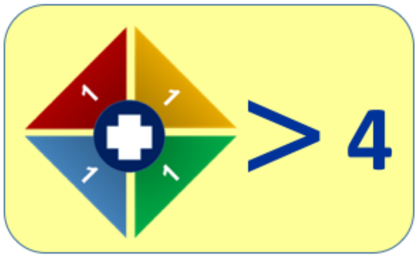
COLORS: up to 12 hours = gold, 12-24 = orangered, 24-48 = red, 2-3 days = black, over 3 days = gray

Educational Impact: Student and Firefighter Training

- Many opportunities for students to be a part of the solution
  - Integration of individual methods into the whole picture
  - Testing of alternative methods
- Teaching environmental data and fire behavior through visual exploration and analysis
- Access to integrated data for teaching and projects
  - Hundreds of thousands learners on our MOOC courses

- PhD level researchers
- Professional software developers
- 31 undergraduate students
  - UC San Diego
  - UC Merced
  - UC Berkeley
  - MURPA University
  - University of Queensland
- 1 high school student
- 6 MSc and 10 MAS students
- 2 PhD students (UMD)
- 1 postdoctoral researcher
- Partners from fire departments
- Advisory board with diverse expertise and affiliations

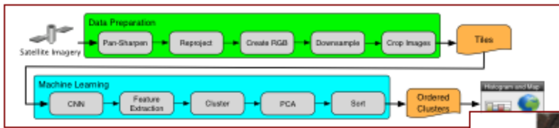
Scientific Impact: Beyond Fire Modeling

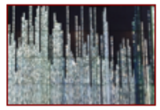


- Impact in many areas
  - Combined solution architecture
  - Fire modeling, ensembles and data assimilation
  - Geospatial data ingestion and mapping interfaces
  - Workflows, data services, machine learning and visualization
- WIFIRE Publications: https://wifire.ucsd.edu/research
- WIFIRE Data APIs: https://wifire.ucsd.edu/data-api
- WIFIRE system is built for expansion
  - Diverse geospatial data streams
  - Many geographical areas
  - Multiple hazards, including mudslides and floods
  - Application to other scientific and societal problems

İlkay ALTINTAŞ, Ph.D.
ialtintas@ucsd.edu

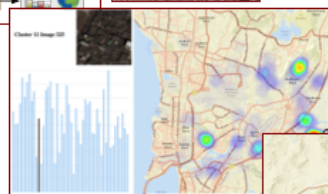
We would love to collaborate with you!





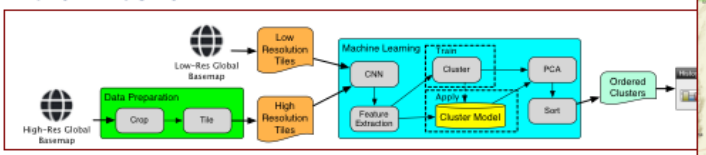
Demographics Analysis Using Satellite Images

BIG PIXEL INITIATIVE AT UC SAN DIEGO



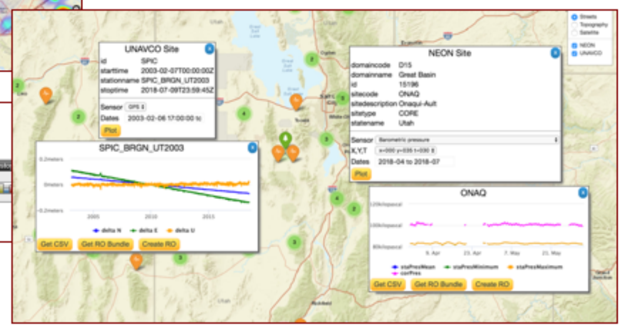
Locating Schools in Rural Liberia

unicef



ROHUB EXPERT SYSTEM UNAVCO neon Kepler

Finding and analyzing NEON and UNAVCO data in Kepler workflows, and publishing the whole analysis as reproducible research objects.



Final Points on Sustainability and Importance of Integrated Cyberinfrastructure

- Any integrated solution architecture depends on many prior CI investments
- Integration into solution architectures increases the impact of individual middleware and software
- Integrated multi-disciplinary engineering of CI needs to be funded separately
- Long term sustainability is still a challenge, and depends on getting users engaged from the beginning
  - WIFIRE Firemap is now supported through fire department subscriptions for hosting, user support, training and data upgrades
- Continued innovation requires sustained research funding
  - e.g., WIFIRE's long-term research still needs to be funded through research grants

Contact: ***Ilkay Altintas, Ph.D.***

Email: ialtintas@ucsd.edu

Questions?

The presented work is collaborative work with many wonderful individuals, and parts of it are funded by NSF, DOE, NIH, UC San Diego and various industry partners.