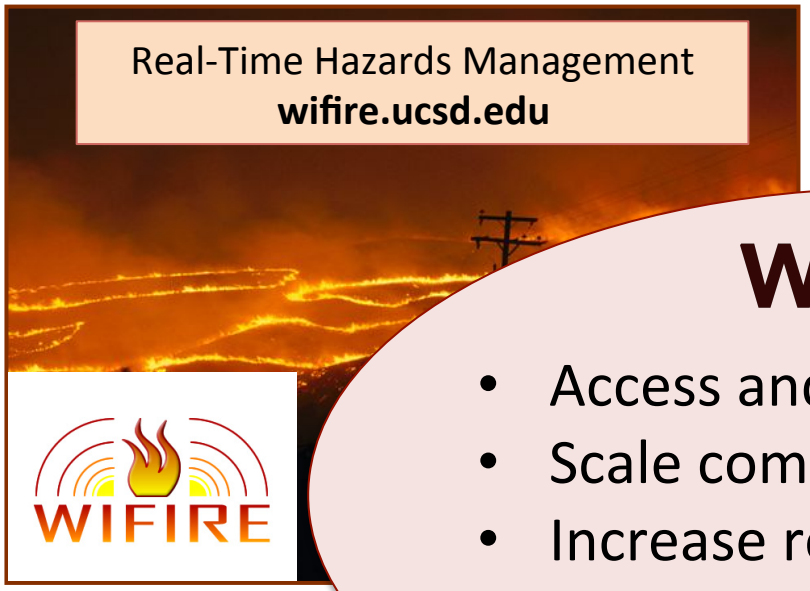
Dynamic Data-Driven Applications using Workflows as a Programming Model

… for Scalable and Reproducible Streaming and Steering

Dr. İlkay ALTINTAS
Chief Data Science Officer, San Diego Supercomputer Center (SDSC)
Founder and Director, Workflows for Data Science (WorDS) Center of Excellence
UC San Diego

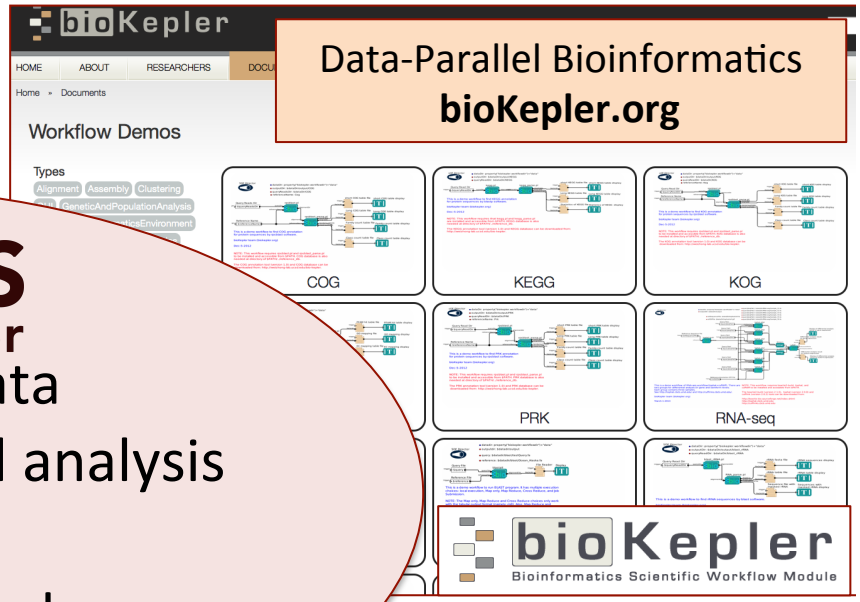
SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Computational and Data Science Workflows
- **Programmable** and **Reproducible Scalability** -



Real-Time Hazards Management
wifire.ucsd.edu

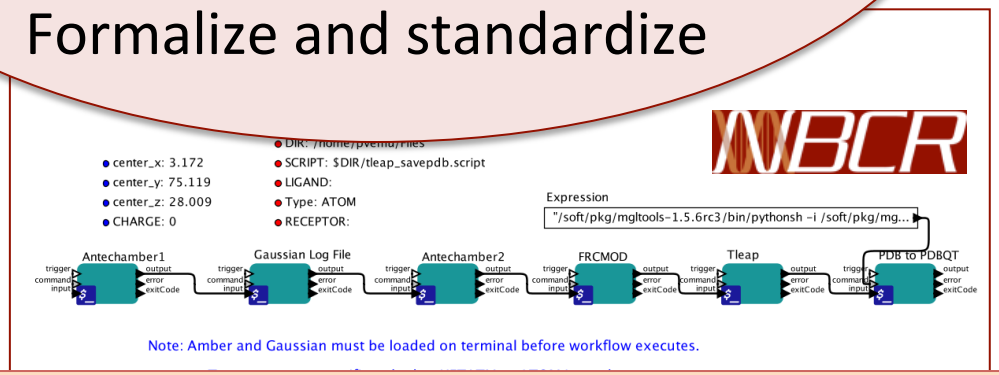
WIFIRE



Data-Parallel Bioinformatics
bioKepler.org

WorDS Center

- Access and query data
- Scale computational analysis
- Increase reuse
- Save time, energy and money
- Formalize and standardize



NBCR

kepler-project.org

Kepler

WorDS.sdsc.edu

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

Kepler is a Scientific Workflow System

www.kepler-project.org

- A cross-project collaboration
  … initiated August 2003
- 2.5 about to be released

- Builds upon the open-source Ptolemy II framework

Ptolemy II: A laboratory for investigating design

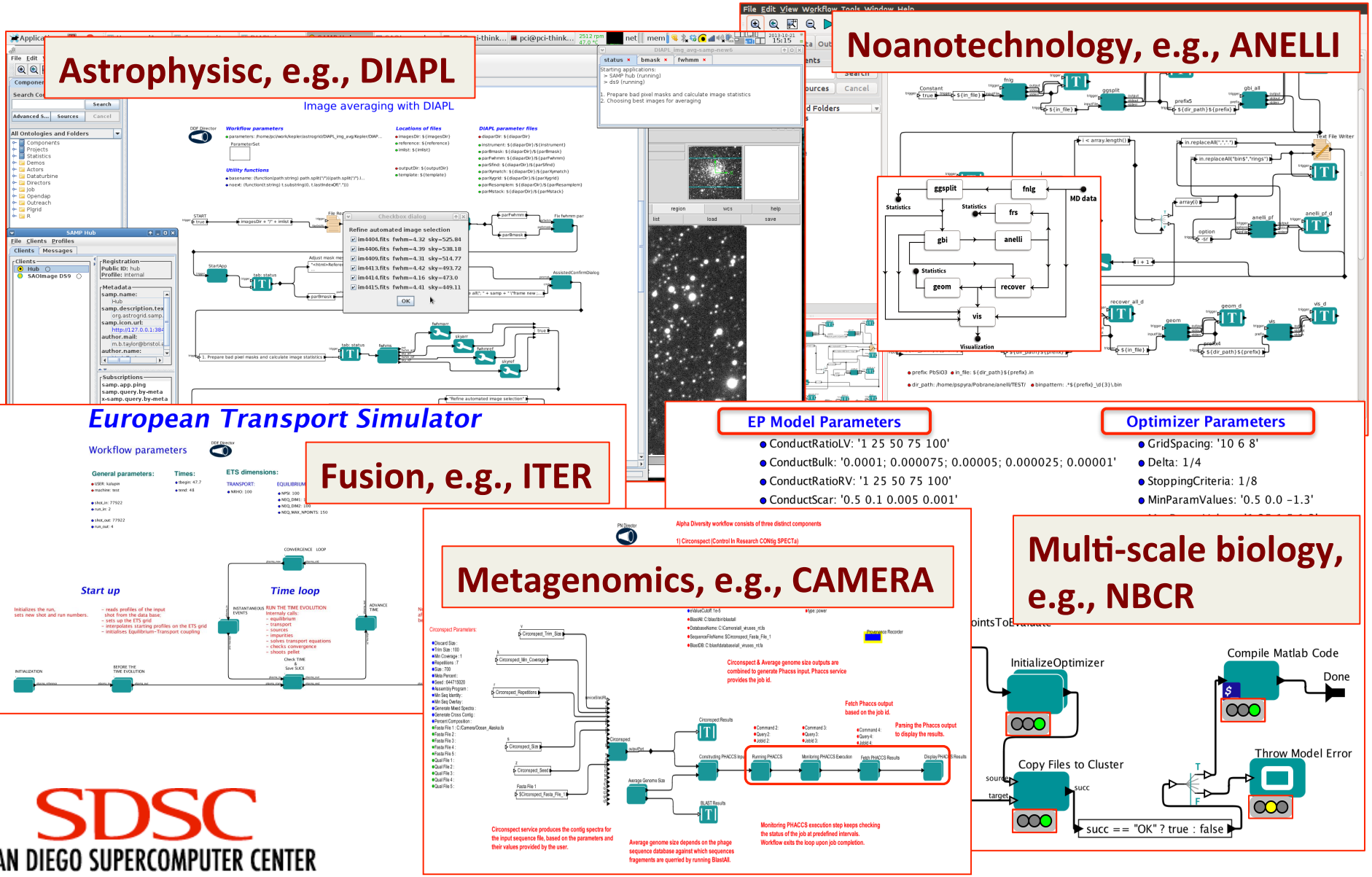
KEPLER: A problem-solving environment for workflow management

---

KEPLER = "Ptolemy II + X" for Scientific Workflows

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Kepler can be applied to problems in different scientific disciplines: some here and many more…



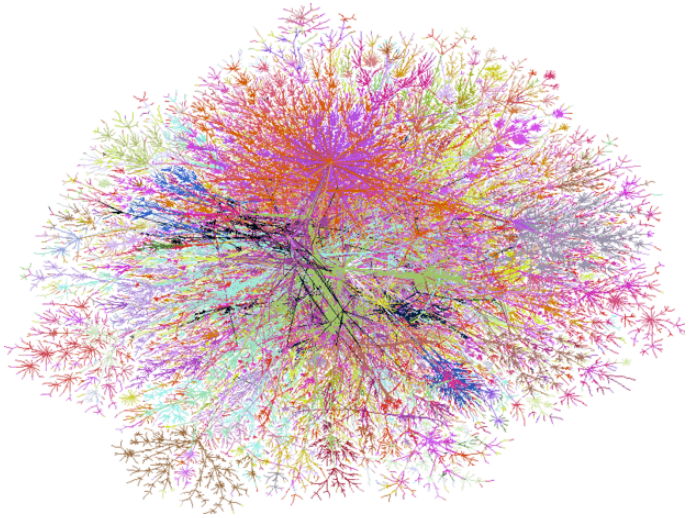
SDSC
SAN DIEGO SUPERCOMPUTER CENTER

So,
how can we use Kepler workflows in the context of big data applications?

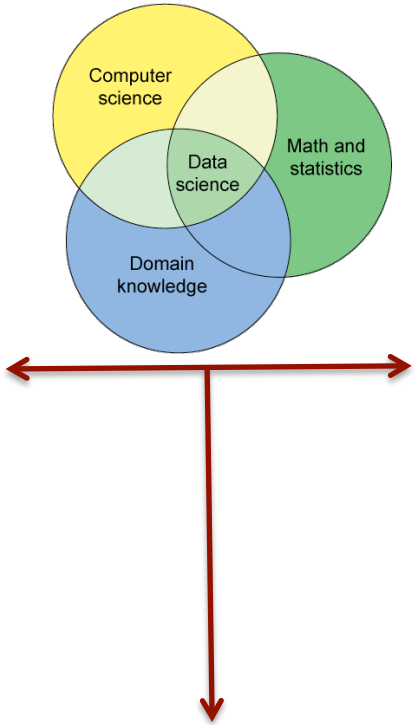
… while coupling all scales computing computing within a reusable solution…

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Streaming Data is Changing Data Science



BIG DATA



ON-DEMAND COMPUTE

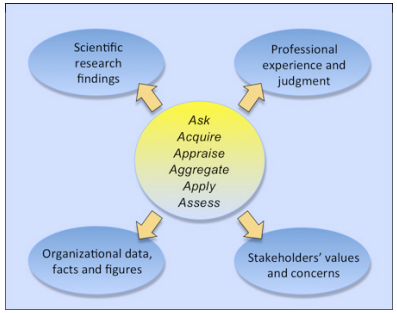
Allows for data-enabled decision making at scale

Many current and future applications with dynamic and measurable impact!

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Disaster Resilience, Response and Management

Marketing and Recommender Systems

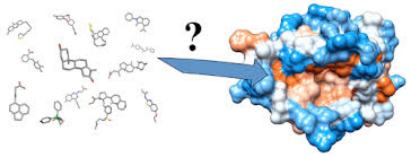


Evidence-Based Management

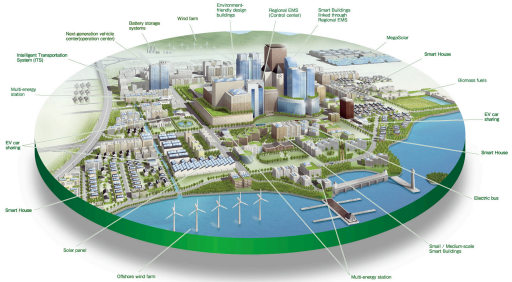
Smart Manufacturing

Smart Grid and Energy Management

Computer-Aided Drug Discovery



Personalized Precision Medicine

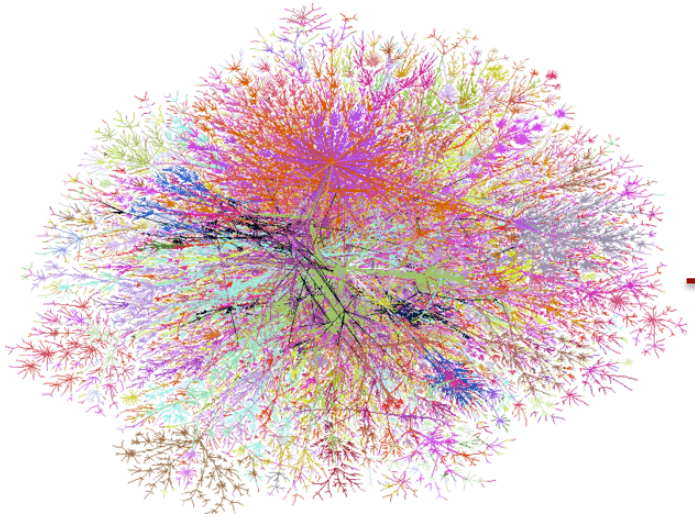


Smart Cities, Urgent Computing and Crowd Management

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

APPLICATION-SPECIFIC KNOWLEDGE and QUESTIONS

?



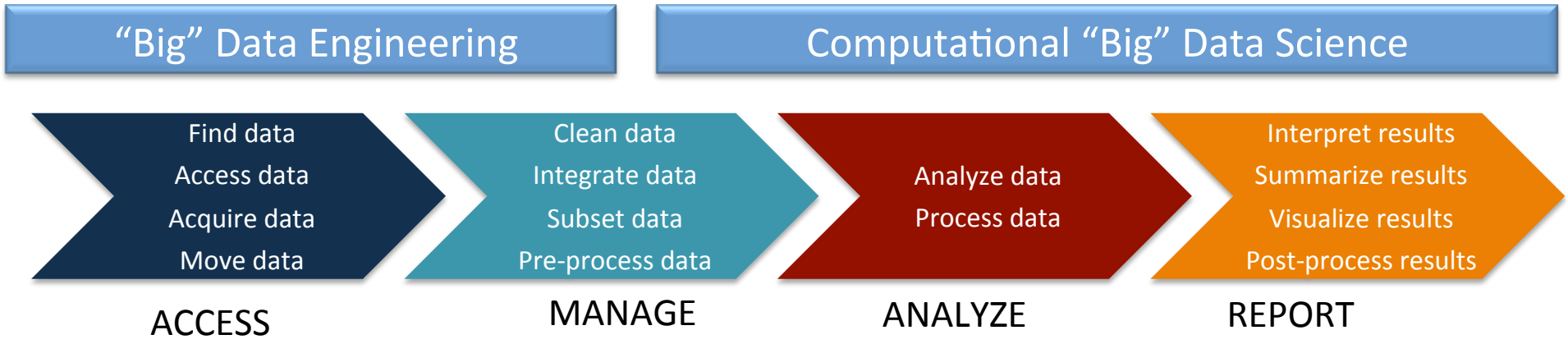
BIG DATA

ON-DEMAND COMPUTE

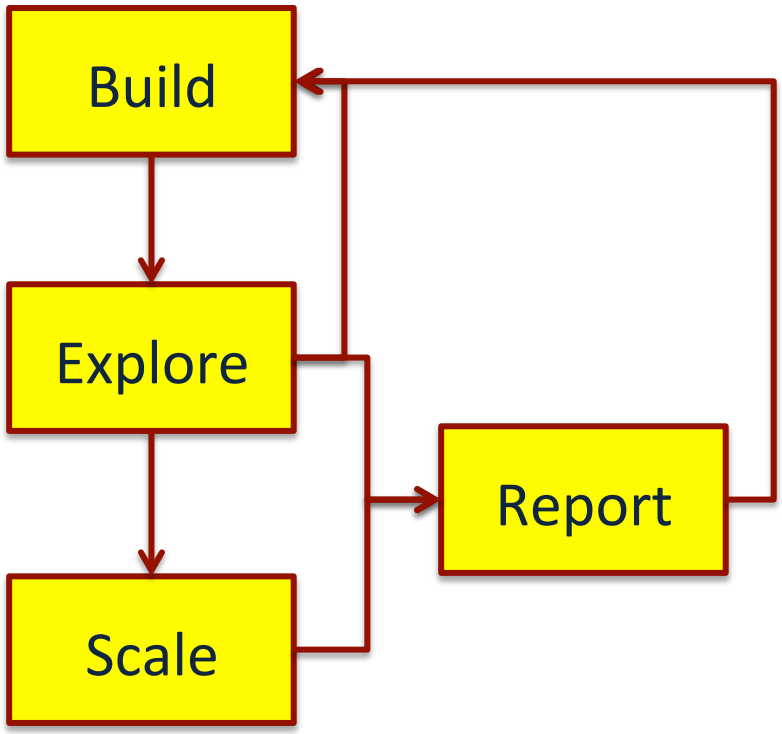
Allows for data-enabled decision making at scale, using statistics, data mining, graph analytics, computational models, etc.

Requires support for experimental work by a multidisciplinary group of experts and dynamic scalability on many platforms!

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

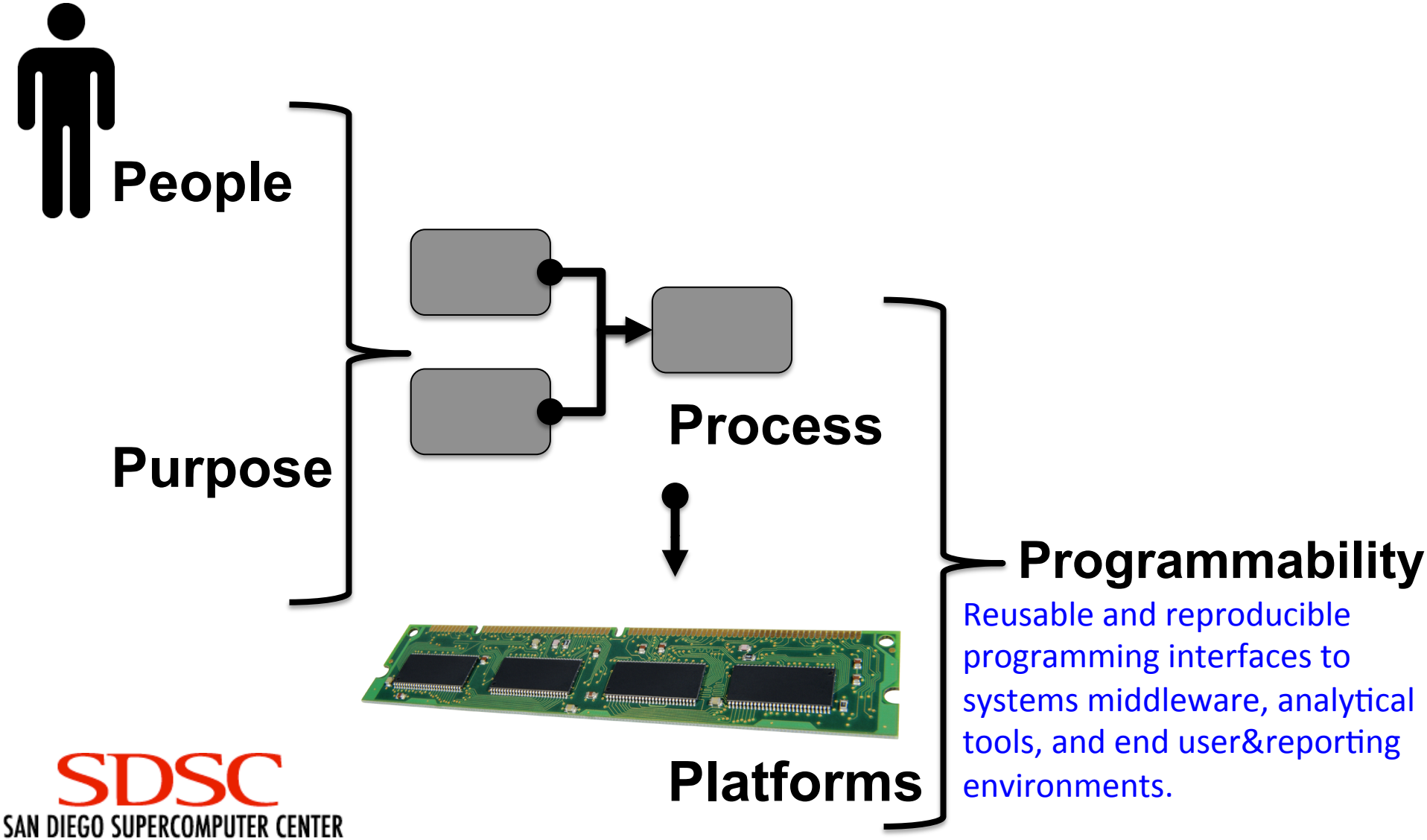


Many ways to look at the process… not every step is automatable!



SDSC
SAN DIEGO SUPERCOMPUTER CENTER

The scalable Process should be Programmable!

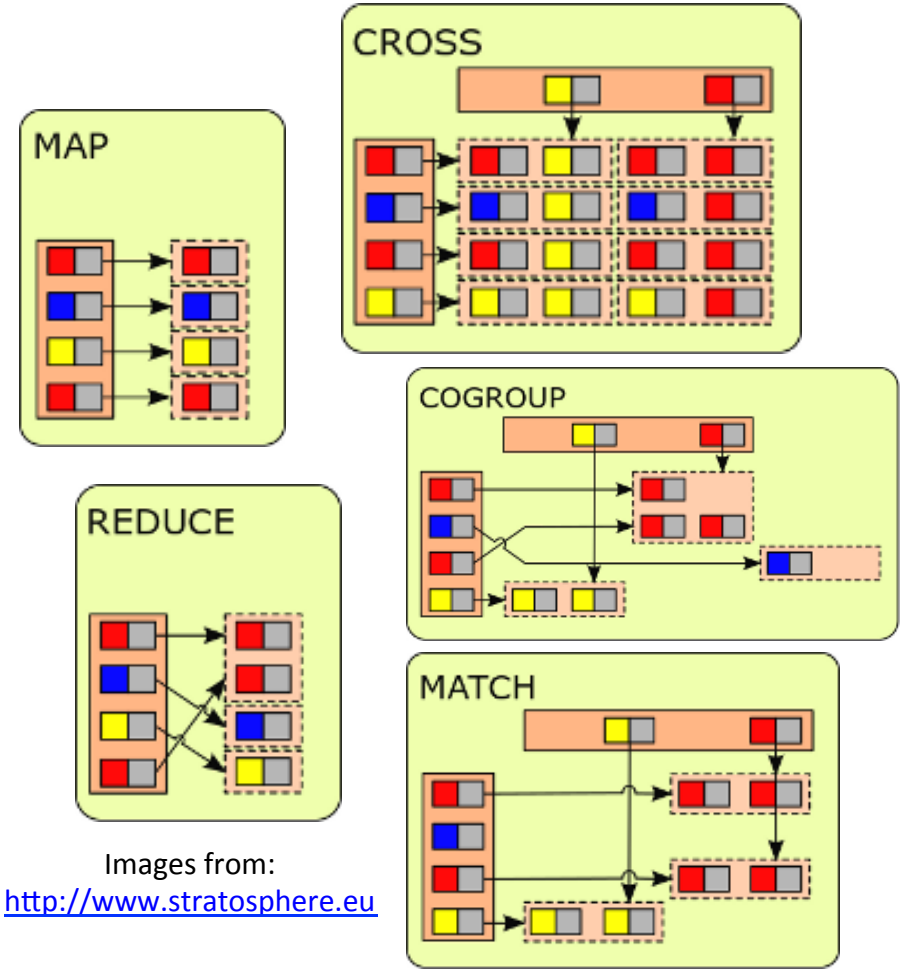


SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Many challenges ahead, but there are things we can do today to close the loop on data science and systems!

- We need to start thinking data science as a whole process today!
- Some R&D challenges:
  - What is programmability of the process at the boundary of data, computing and analytical systems?
  - How can we keep the process accountable, useful and, to a large extent, automatable?
  - How can we achieve the best performance through the process?
  - How can we report through the process?

Example Challenge: Programmability of Distributed Data Parallel Execution

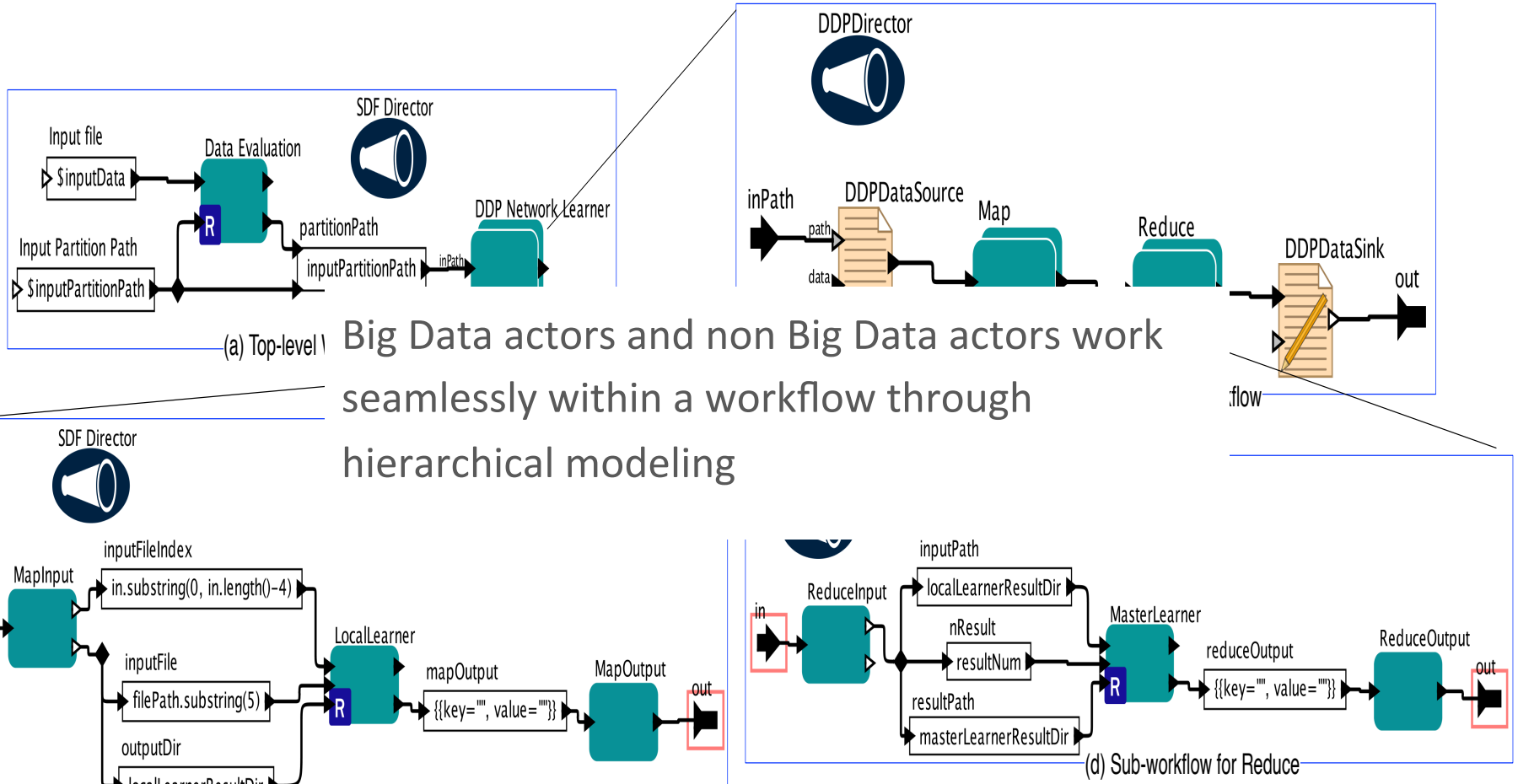


Images from:
http://www.stratosphere.eu

- How to easily apply the Big Data Patterns in a workflow?
- How to parallelize legacy tools for Big Data?
- How to pick pattern(s) each specific task/tool?
- How to run the same process on top of different Big Data engines, such as Hadoop, Spark and Stratosphere (Apache name: Flink)?

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

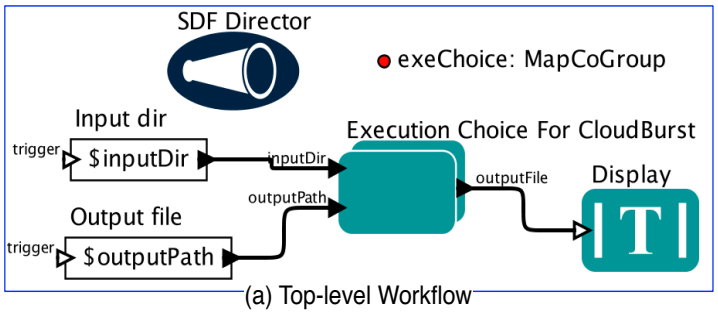
Kepler Workflow for Bayesian Network Learning Application



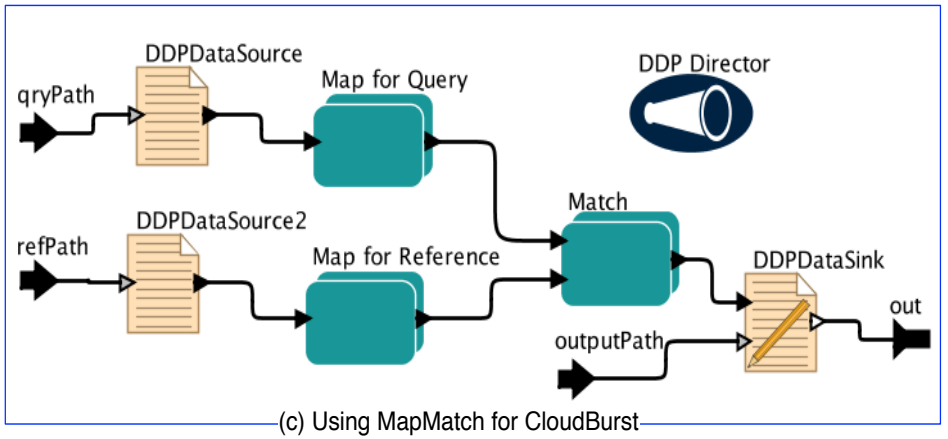
Big Data actors and non Big Data actors work seamlessly within a workflow through hierarchical modeling

SAN DIEGO SUPERCOMPUTER CENTER

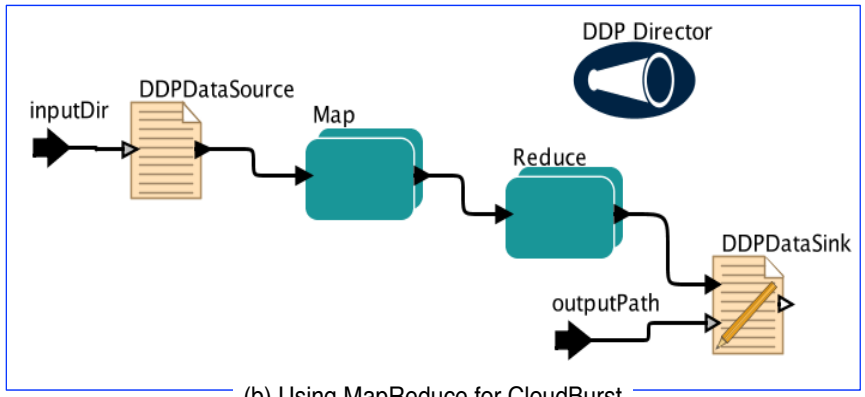
Kepler Workflow for a Bioinformatics Application (CloudBurst)



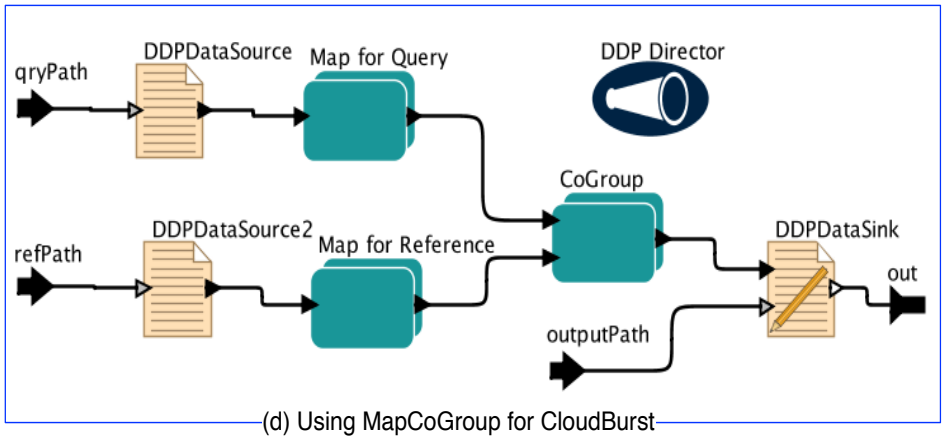
(a) Top-level Workflow



(c) Using MapMatch for CloudBurst



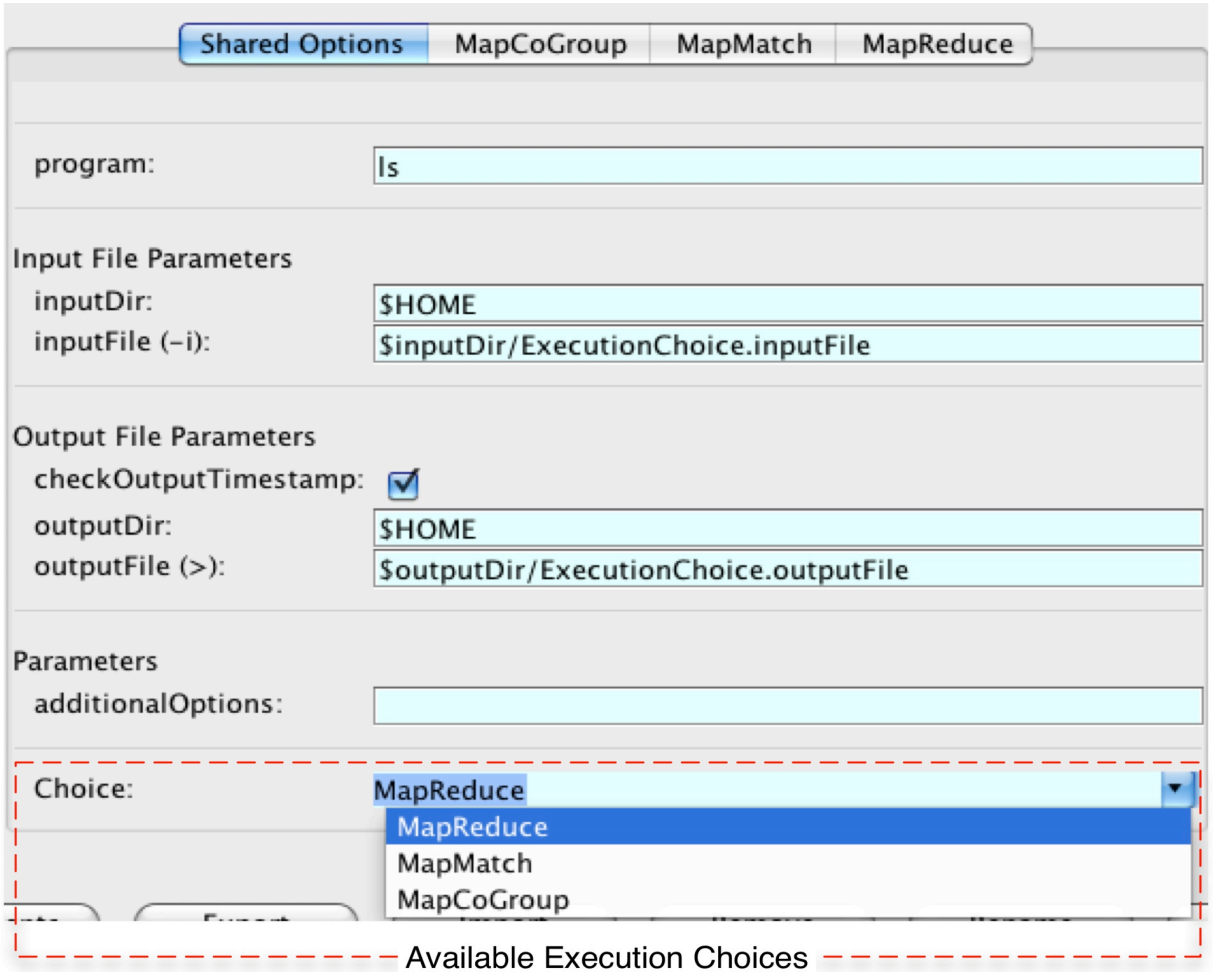
(b) Using MapReduce for CloudBurst



(d) Using MapCoGroup for CloudBurst

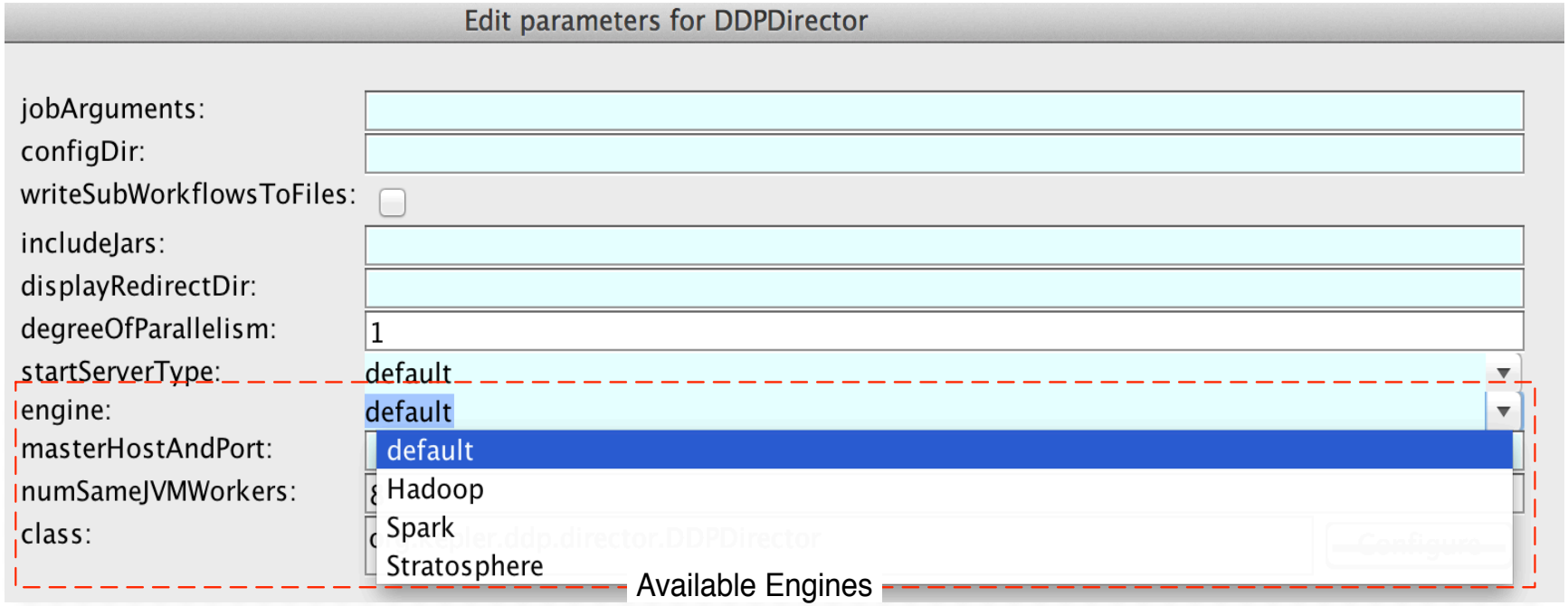
SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Execution Choices for CloudBurst Application



SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Engine Configuration of DDP Director



SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Legacy Tool Parallelization for Big Data

- Black-box approach (we use this approach)
  - Run a tool directly
  - Wrap the tool with Big Data techniques
  - Can quickly convert a tool into a parallelized one
- White-box approach
  - Investigate the source code of a legacy tool and try to re-implement it using Big Data techniques
  - Time-consuming
  - Often tightly-coupled with specific Big Data engine
  - Could find more parallel opportunities

Using Workflows and Cyberinfrastructure for Wildfire Resilience

- A Scalable Data-Driven Monitoring and Dynamic Prediction Approach -

WIFIRE

wifire.ucsd.edu

NSF

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

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

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

UC San Diego

University of Maryland 18 56

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

What is lacking in disaster management today is…

a dynamic system integration of real-time sensor networks, satellite imagery, near-real time data management tools, wildfire simulation tools, and connectivity to emergency command centers

…. before, during and after a firestorm.

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Research Questions

- 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

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

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

(WIFIRE)

wifire.ucsd.edu

Development of:

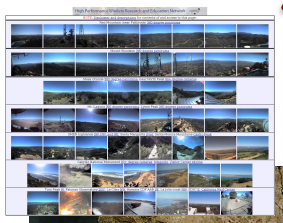
"cyberinfrastructure" for "analysis of large dimensional heterogeneous real-time sensed data" for fire resilience *before*, *during* and *after* a wildfire

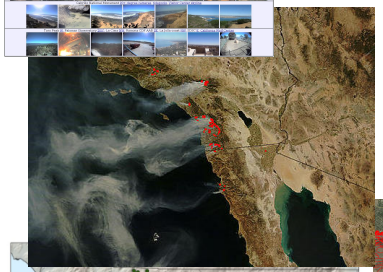
SDSC
SAN DIEGO SUPERCOMPUTER CENTER

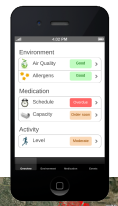
Big Data

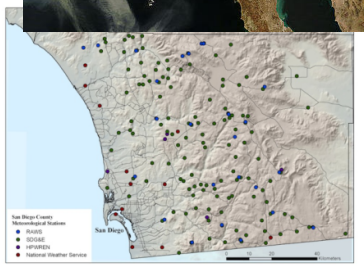
WIFIRE

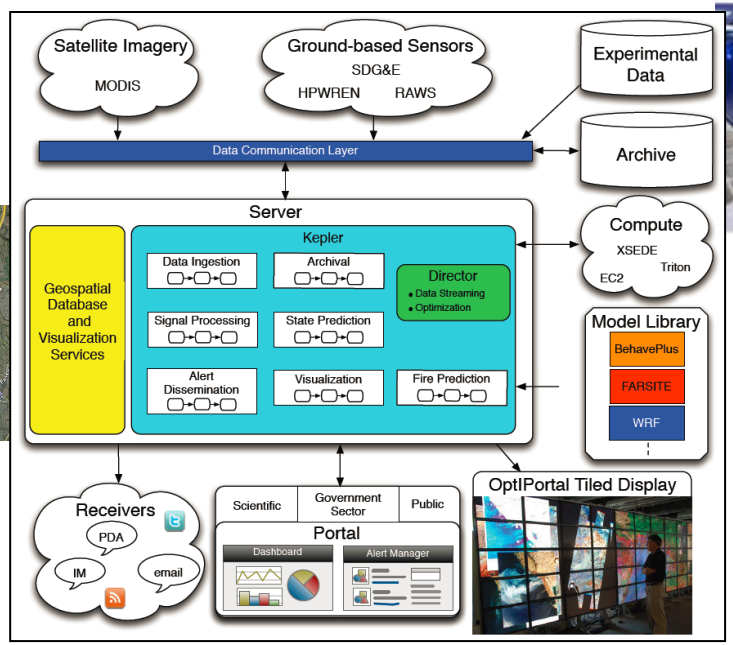
Monitoring
Visualization
Fire Modeling







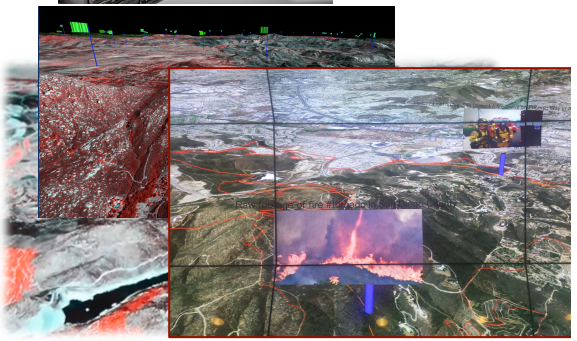




hadoop

NVIDIA TESLA

amazon web services



SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Modeling

More accurate situational awareness using data

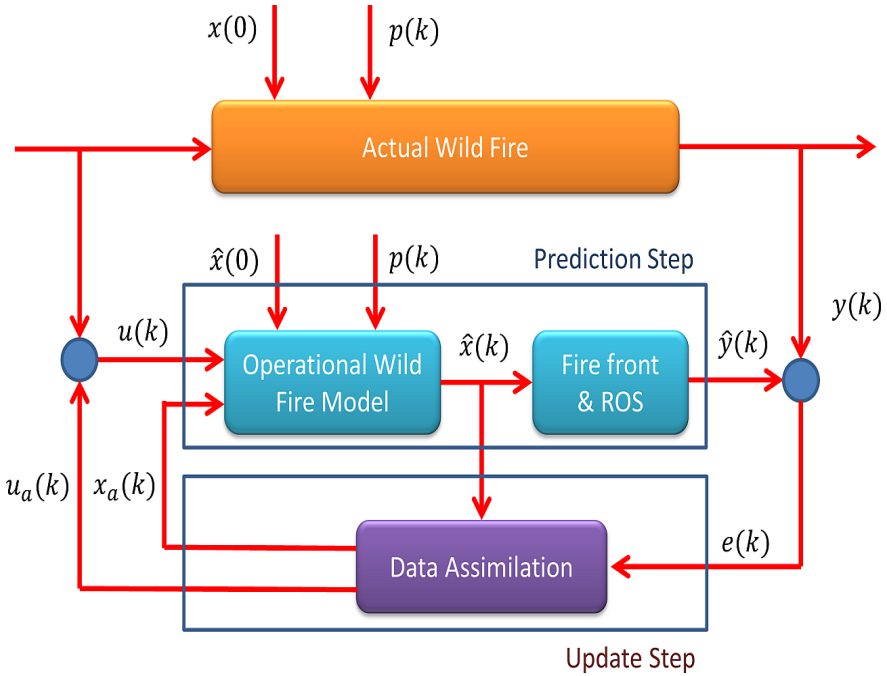
-- Data to Modeling in WIFIRE --

Real-time remote data –> Modeling, data assimilation and dynamic wildfire behavior prediction

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Putting it all together!

-- Wildfire Behavior Modeling and Data Assimilation --

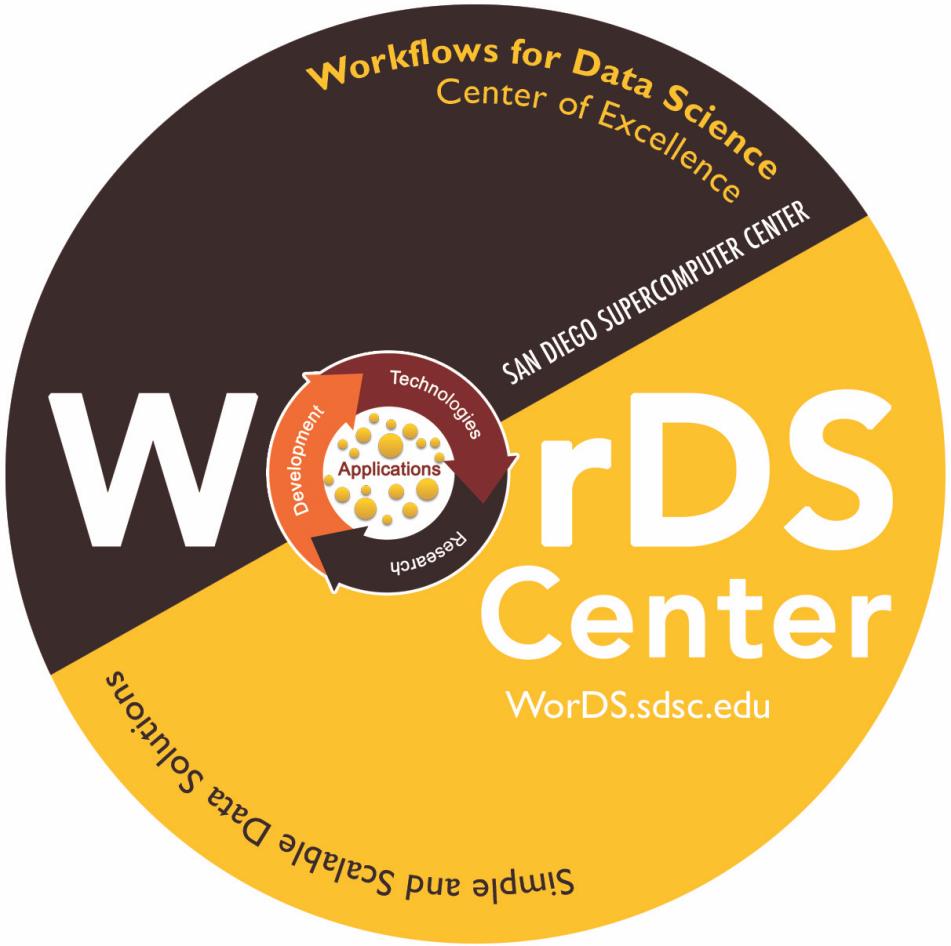


Conceptual Data Assimilation Workflow with Prediction and Update Steps using Sensor Data

- 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

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Questions?



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

SDSC
SAN DIEGO SUPERCOMPUTER CENTER