

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

Computational and Data Science Workflows

- Programmable and Reproducible Scalability -

Real-Time Hazards Management
wifire.ucsd.edu

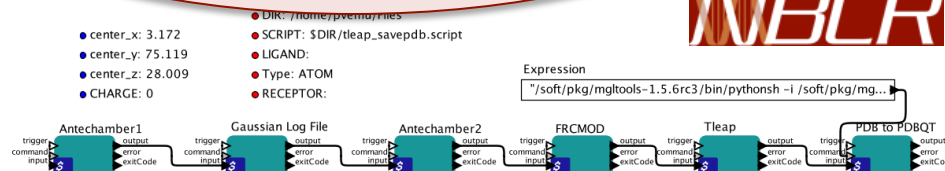


WorDS
Center

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

Data-Parallel Bioinformatics
bioKepler.org

kepler-project.org

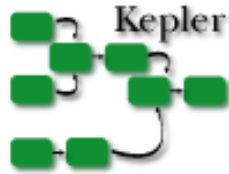


Note: Amber and Gaussian must be loaded on terminal before workflow executes.

WorDS.sdsc.edu

Scalable Automated Molecular Dynamics and Drug Discovery
nbcrc.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

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

Astrophysics, e.g., DIAPL

Image averaging with DIAPL

Nanotechnology, e.g., ANELLI

The screenshot shows two Kepler workflow windows. The left window, titled 'Image averaging with DIAPL', displays workflow parameters, utility functions, and a complex flowchart of tasks including 'Refine automated image selection' and 'Prepare bad pixel masks'. The right window shows a Kepler workflow for 'ANELLI' nanotechnology simulation, featuring tasks like 'ggsplit', 'fngl', 'anelli', and 'recover' connected by data links.

European Transport Simulator

Workflow parameters

Fusion, e.g., ITER

The screenshot shows a Kepler workflow for the 'European Transport Simulator'. It includes a 'Start up' section with initialization steps and a 'Time loop' section with a 'CONVERGENCE LOOP' and 'RUN THE TIME EVOLUTION' tasks. Parameters for 'General', 'Times', and 'ETS dimensions' are listed.

EP Model Parameters

Optimizer Parameters

- ConductRatioLV: '1 25 50 75 100'
- ConductBulk: '0.0001; 0.000075; 0.00005; 0.000025; 0.00001'
- ConductRatioRV: '1 25 50 75 100'
- ConductScar: '0.5 0.1 0.005 0.001'
- GridSpacing: '10 6 8'
- Delta: 1/4
- StoppingCriteria: 1/8
- MinParamValues: '0.5 0.0 -1.3'

Metagenomics, e.g., CAMERA

The screenshot shows a Kepler workflow for 'CAMERA' metagenomics analysis. It includes 'Chromospect Parameters' and a 'Time loop' with tasks for 'Chromospect', 'Constructing PHACCS', and 'Fetch Phacces output'. A note states: 'Chromospect service produces the contig spectra for the input sequence file, based on the parameters and their values provided by the user.'

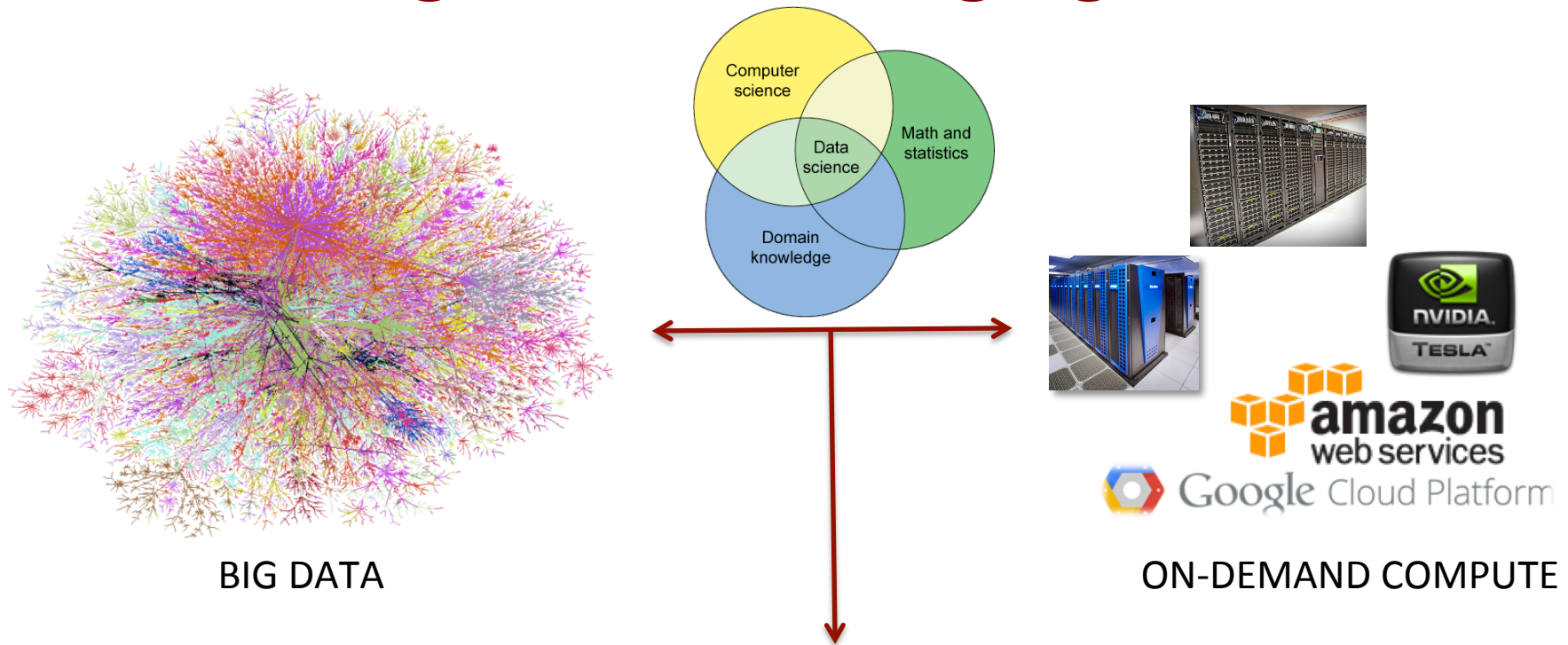
Multi-scale biology, e.g., NBCR

The screenshot shows a Kepler workflow for 'NBCR' multi-scale biology simulation. It includes tasks for 'InitializeOptimizer', 'Copy Files to Cluster', 'Compile Matlab Code', and 'Throw Model Error'. A conditional task is defined as 'succ == "OK" ? true : false'.

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

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

Streaming Data is Changing Data Science



Allows for data-enabled decision making at scale

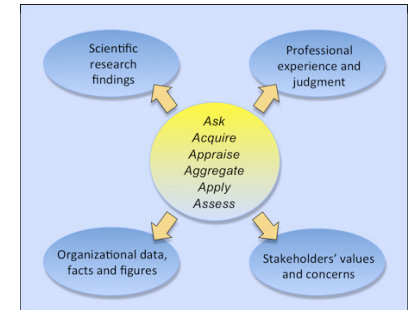
Many current and future **applications** with **dynamic** and **measurable** impact!



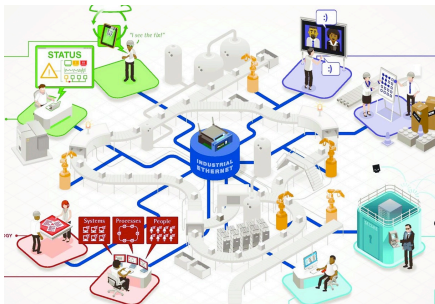
Photograph by Mark Thiessen

Disaster Resilience, Response and Management

Marketing and Recommender Systems



Evidence-Based Management

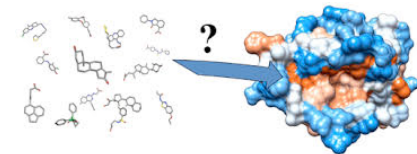


Smart Manufacturing

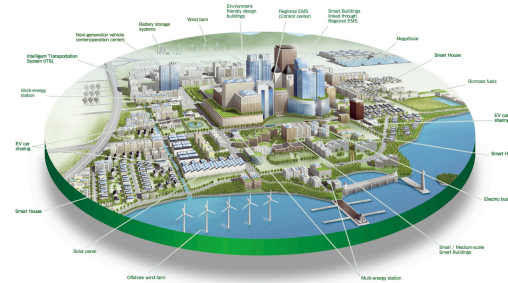


Smart Grid and Energy Management

Computer-Aided Drug Discovery

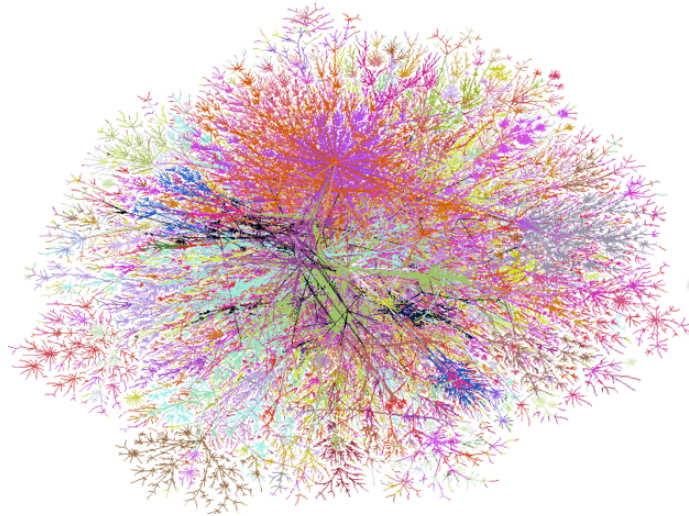


Personalized Precision Medicine

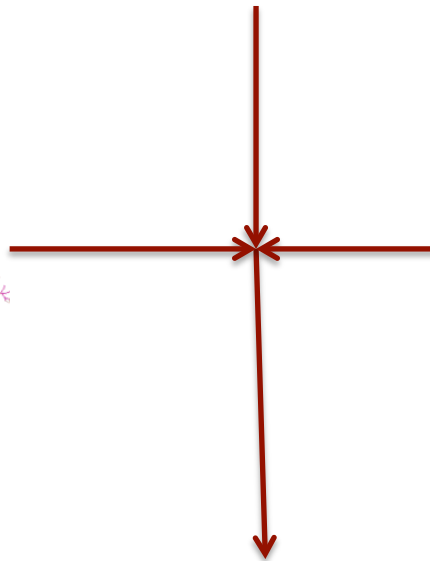


Smart Cities, Urgent Computing and Crowd Management

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.

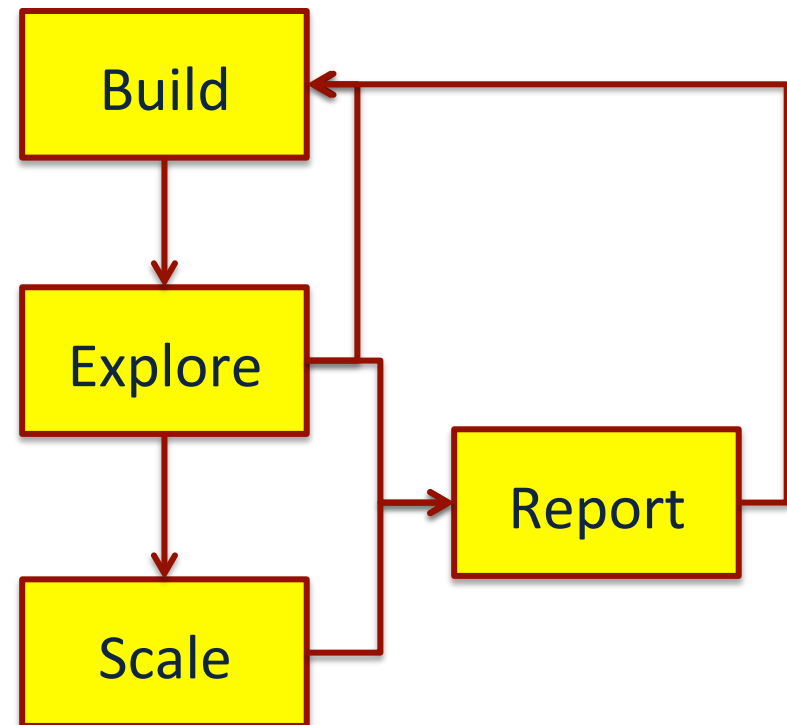
Requires support for **experimental work** by a multidisciplinary group of experts and **dynamic scalability** on many platforms!

“Big” Data Engineering

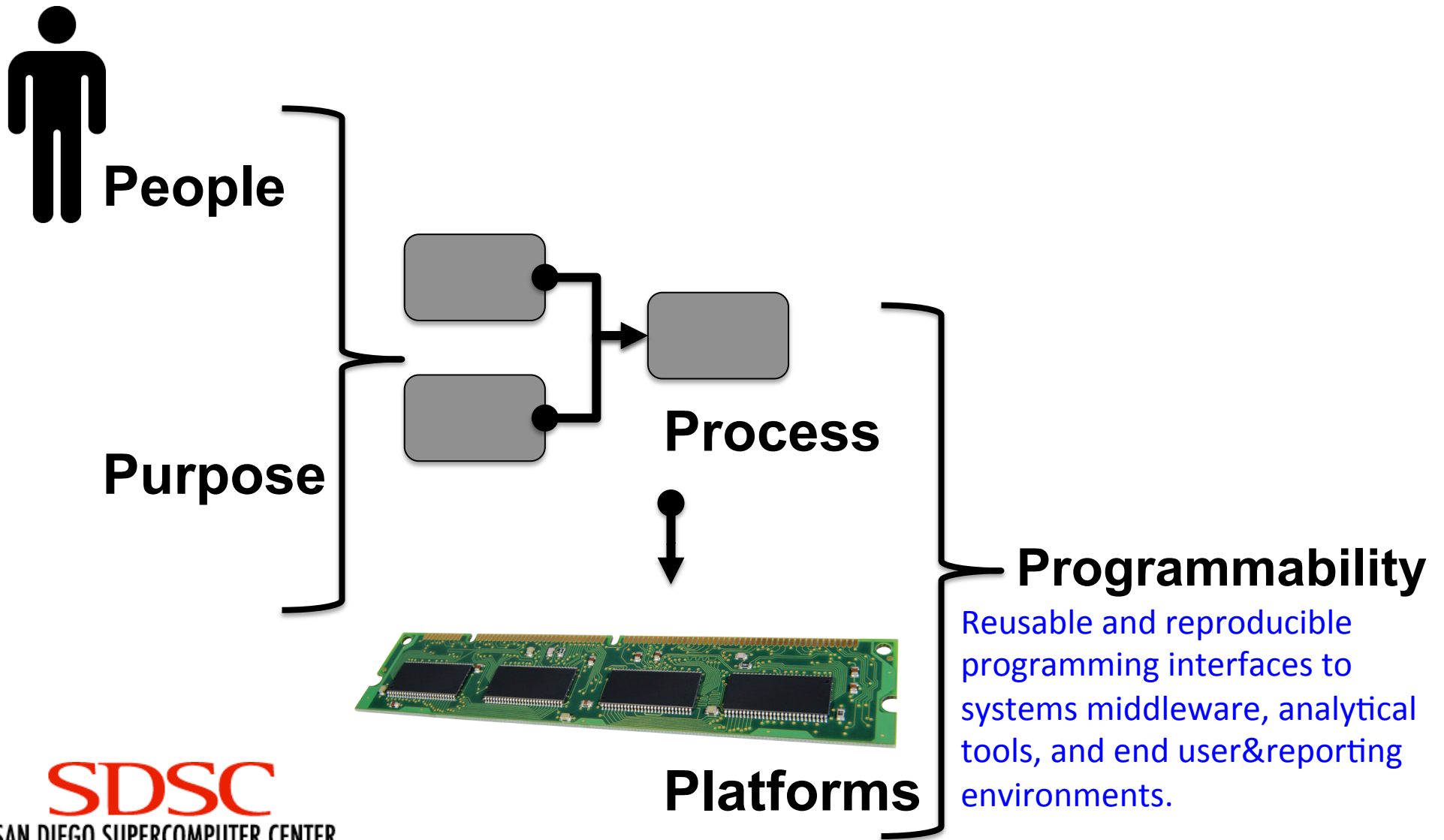
Computational “Big” Data Science



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



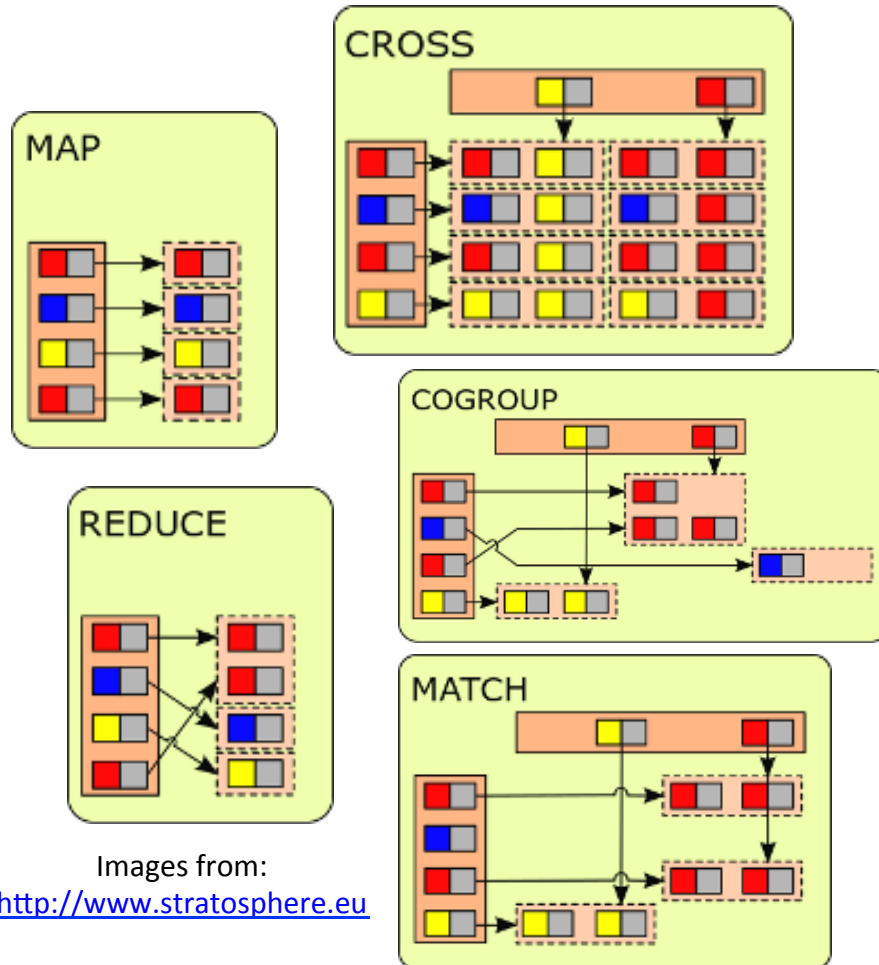
The scalable Process should be Programmable!



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?

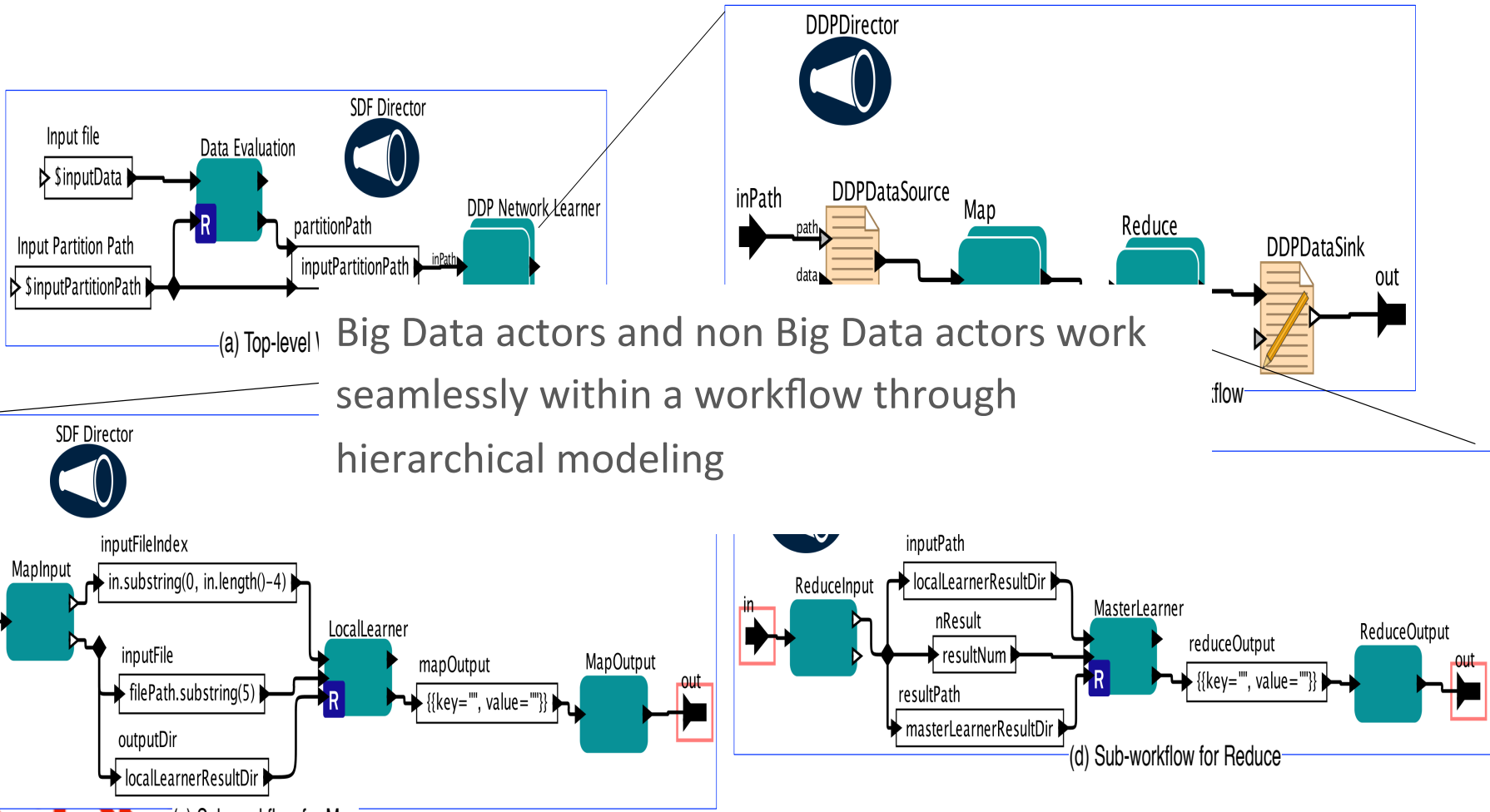
Example Challenge: Programmability of Distributed Data Parallel Execution



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

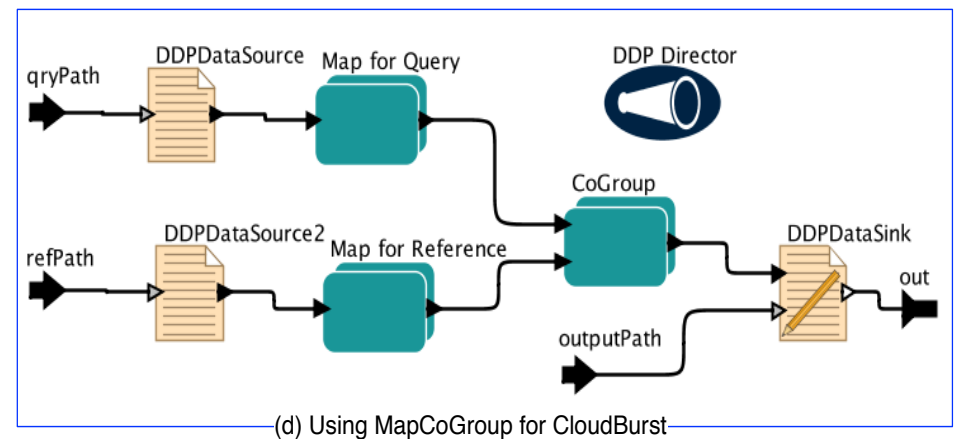
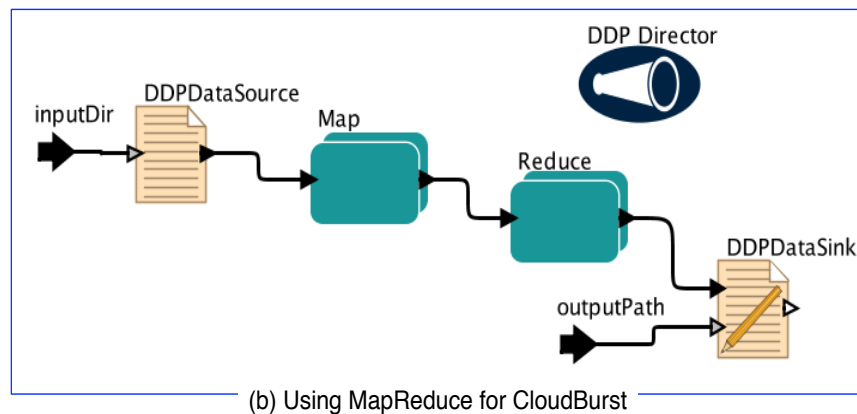
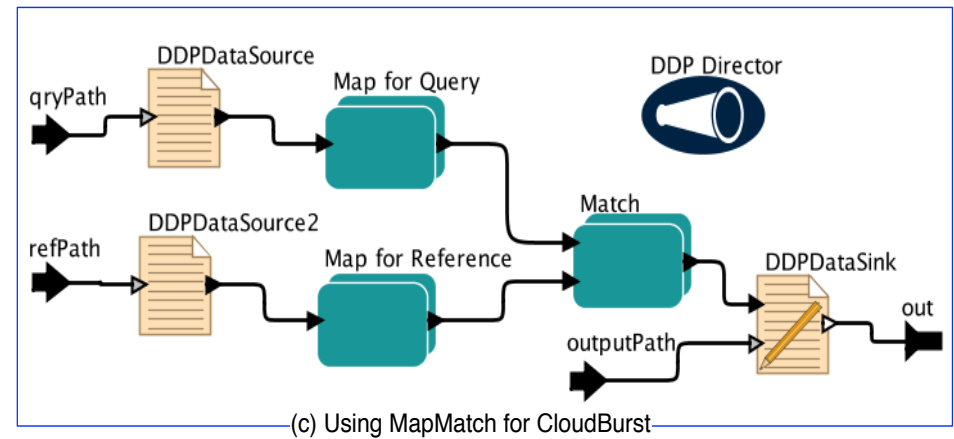
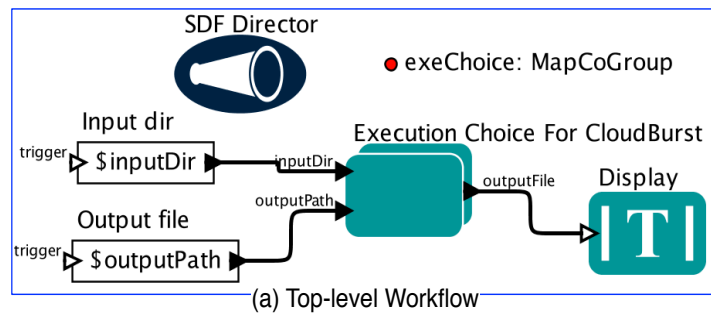
- How to **easily apply** the Big Data Patterns in a workflow?
- How to **parallelize legacy tools** for Big Data?
- How to **pick pattern(s)** each specific task/tool?
- How to run the same process on top of **different Big Data engines**, such as Hadoop, Spark and Stratosphere (Apache name: Flink)?

Kepler Workflow for Bayesian Network Learning Application



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

Kepler Workflow for a Bioinformatics Application (CloudBurst)



Execution Choices for CloudBurst Application

Shared Options | MapCoGroup | MapMatch | MapReduce

program: ls

Input File Parameters

inputDir: \$HOME

inputFile (-i): \$inputDir/ExecutionChoice.inputFile

Output File Parameters

checkOutputTimestamp:

outputDir: \$HOME

outputFile (>): \$outputDir/ExecutionChoice.outputFile

Parameters

additionalOptions:

Choice: MapReduce (selected), MapMatch, MapCoGroup

Available Execution Choices

Engine Configuration of DDP Director

Edit parameters for DDPDirector

jobArguments:	<input type="text"/>
configDir:	<input type="text"/>
writeSubWorkflowsToFiles:	<input type="checkbox"/>
includeJars:	<input type="text"/>
displayRedirectDir:	<input type="text"/>
degreeOfParallelism:	<input type="text" value="1"/>
startServerType:	<input type="text" value="default"/>
engine:	<input type="text" value="default"/>
masterHostAndPort:	<input type="text" value="default"/>
numSameJVMWorkers:	<input type="text" value="8"/>
class:	<input type="text" value="com.sdsc.ddp.director.DDPDirector"/>

Available Engines

- Hadoop
- Spark
- Stratosphere

Configure

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 -



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.



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

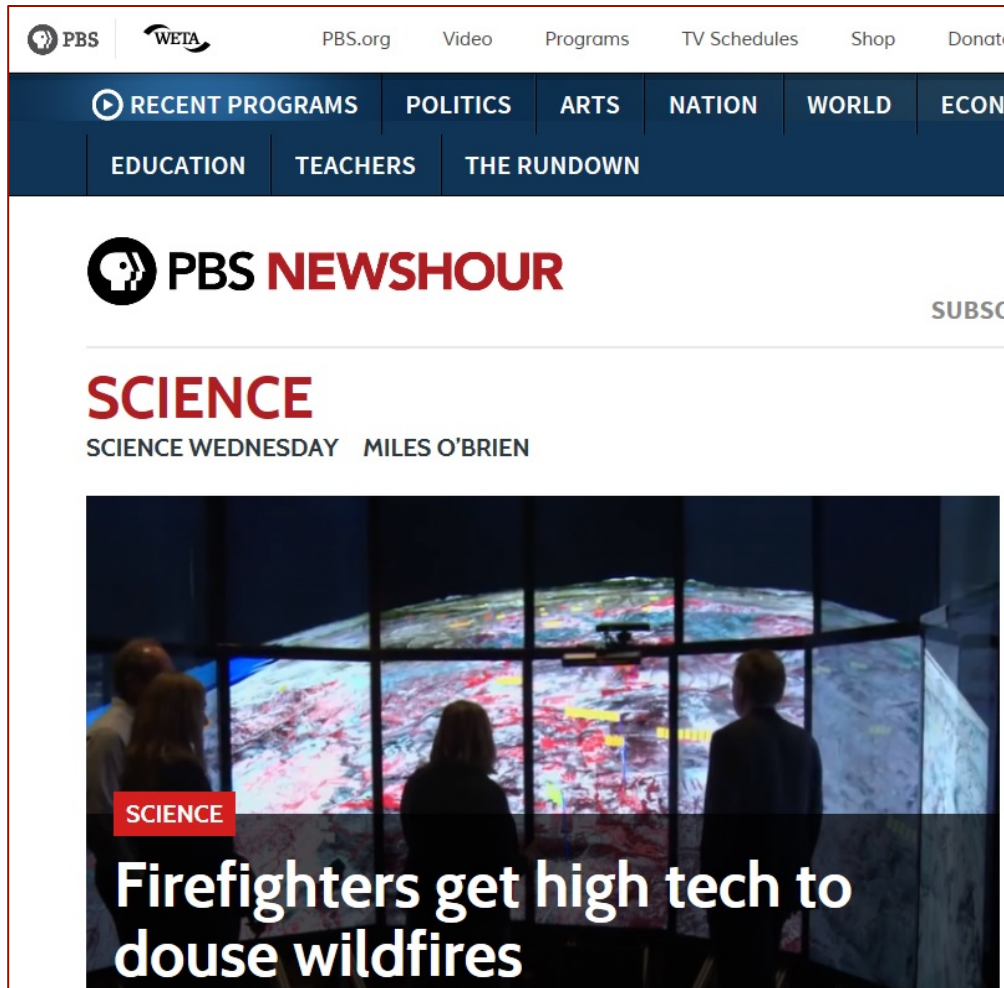
A Scalable Data-Driven Monitoring, Dynamic Prediction and Resilience Cyberinfrastructure for Wildfires

(**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

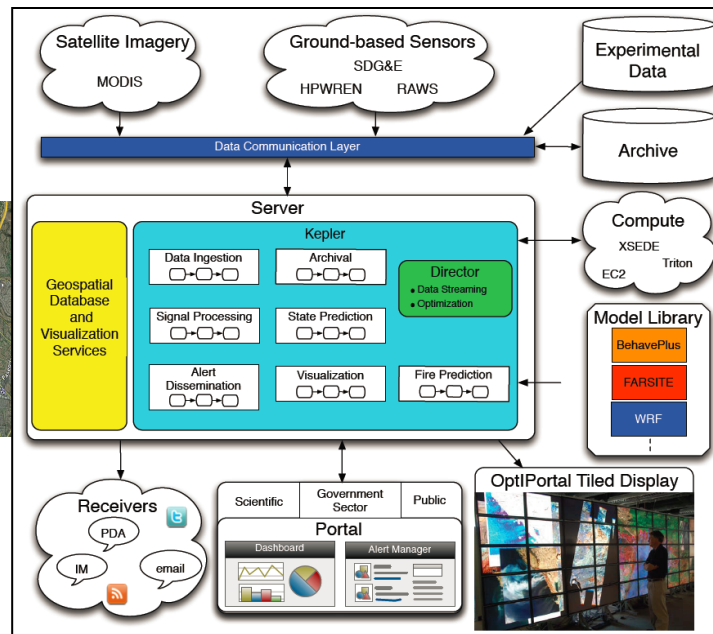
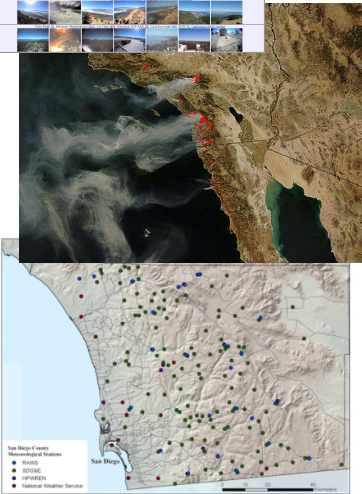
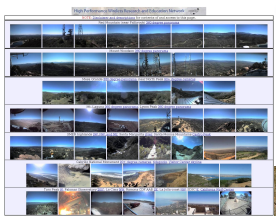




Big Data



Monitoring Visualization Fire Modeling



Modeling

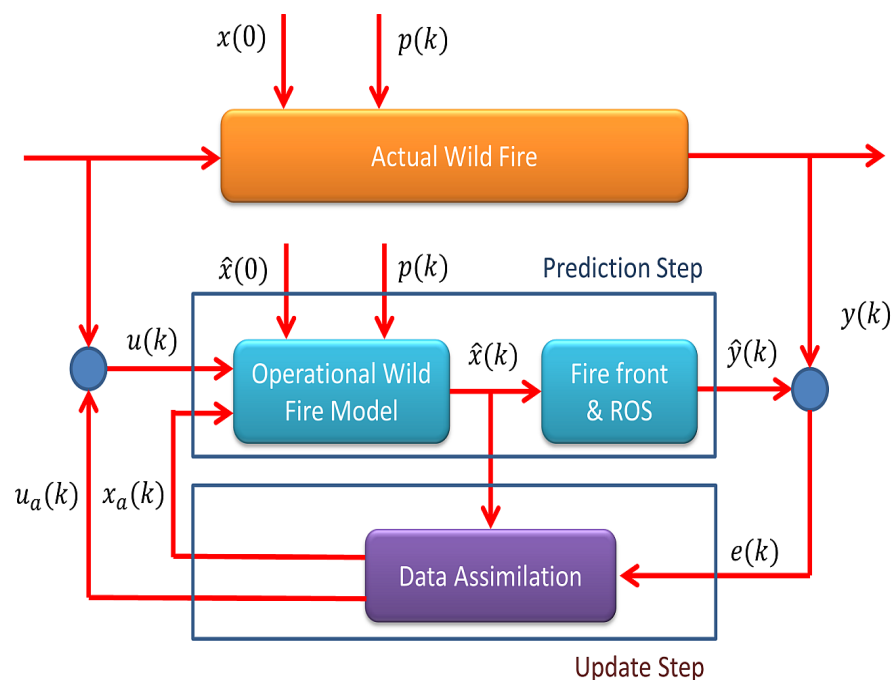
More accurate situational awareness using data

-- Data to Modeling in WIFIRE --

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

Putting it all together!

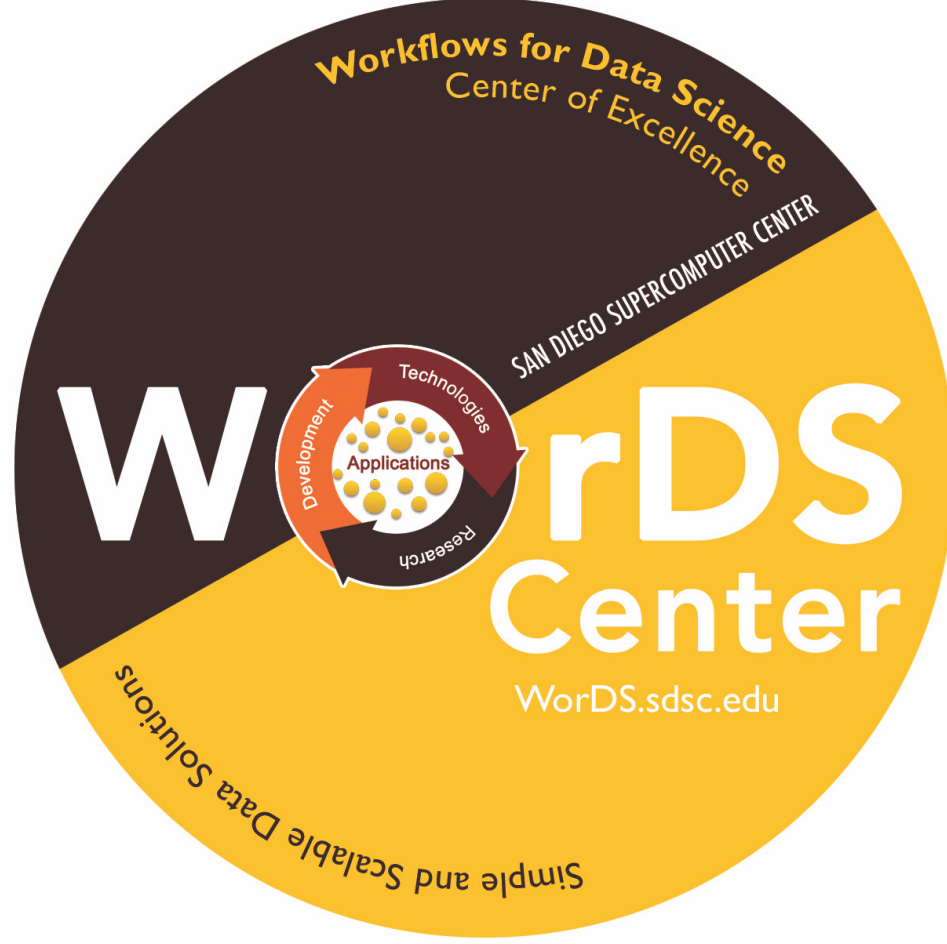
-- Wildfire Behavior Modeling and Data Assimilation --



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

Questions?



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