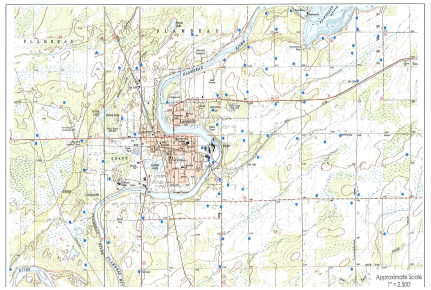


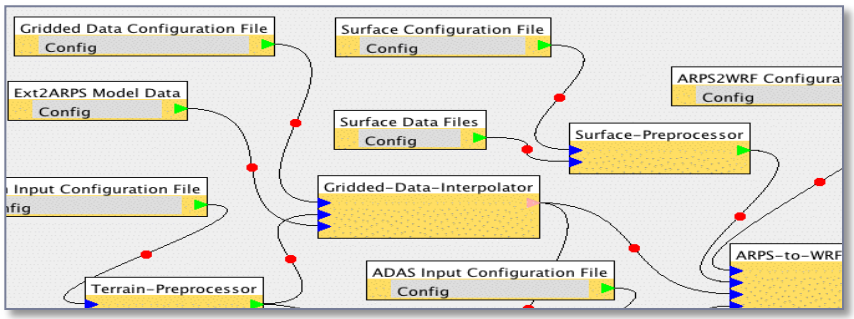
Multi-Dimensional Classification Model for Scientific Workflow Characteristics

Lavanya Ramakrishnan[1] and Beth Plale[2]

[1] Lawrence Berkeley Nat'l Lab, Berkeley, CA

[2] Data To Insight Center and School of Informatics and Computing, Indiana University Bloomington

DATA TO INSIGHT CENTER
INDIANA UNIVERSITY
Pervasive Technology Institute



Overview

- Motivation: workflow research often done in context of single or small number of science domains.
  - Results in limited understanding of performance and characteristics of scientific workflows
- Approach: classification model and survey of scientific workflow characteristics.
- Informal defn of workflow: set of *tasks* generally organized as a graph G=(V,E), task = V.
- Notion in talk: *executed footprint* of workflow. *Task* run as 5 parallel instances has count *5*

Classification Model

- Why? Helps classify or “bin” workflow types
  - Enable applicability of solutions
- Dimensions
  - Size,
  - Resource Usage,
  - Structural Pattern,
  - Data Pattern, and
  - Usage Scenarios
- Based on survey of workflows from different domains
  - Each workflow has been modeled using one of several workflow tools and/or through scripts

Dimensions 1 and 2: Size and Resource Usage

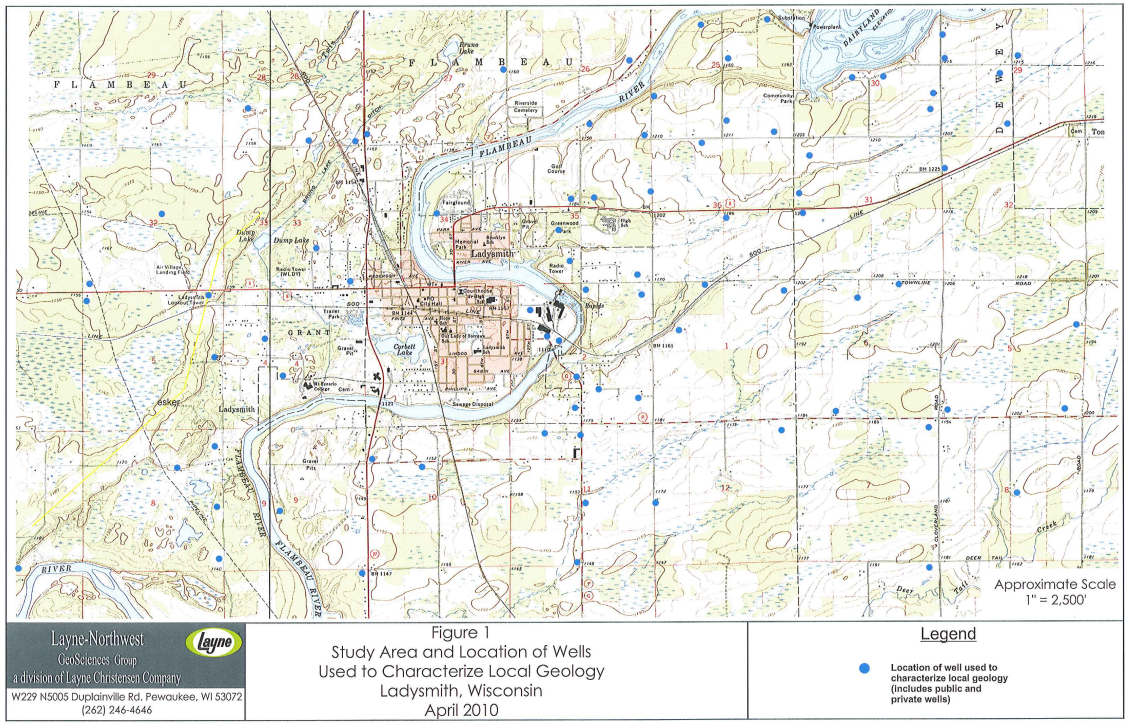
- Properties of *Size* dimension
  - *Total number of tasks* – total number of tasks in executed workflow. If task executes in parallel, each instance of task is counted.
  - *Number of parallel tasks* – defines maximum number of parallel tasks in any part of workflow
  - *Longest chain* – defines number of tasks in longest chain of workflow
- *Resource Usage dimension*
  - *Max task processor width* – max concurrent # processors required by workflow
  - *Computation time* – total computational time required
  - *Data sizes* – data size of workflow inputs, outputs, and intermediate products

Dimensions 3: Structural Pattern

- Captures dominant pattern seen in workflow
  - Sequential
    - consists of tasks that follow one after the other
  - Parallel
    - multiple tasks that can be run at same time
  - Parallel-split
    - one task's output feeds to multiple tasks
  - Parallel-merge
    - multiple tasks merge into one task
  - Parallel-split-merge
    - both parallel-merge and parallel-split
  - Mesh
    - task dependencies are interleaved

Dims 4 and 5: Data Pattern and Usage scenarios

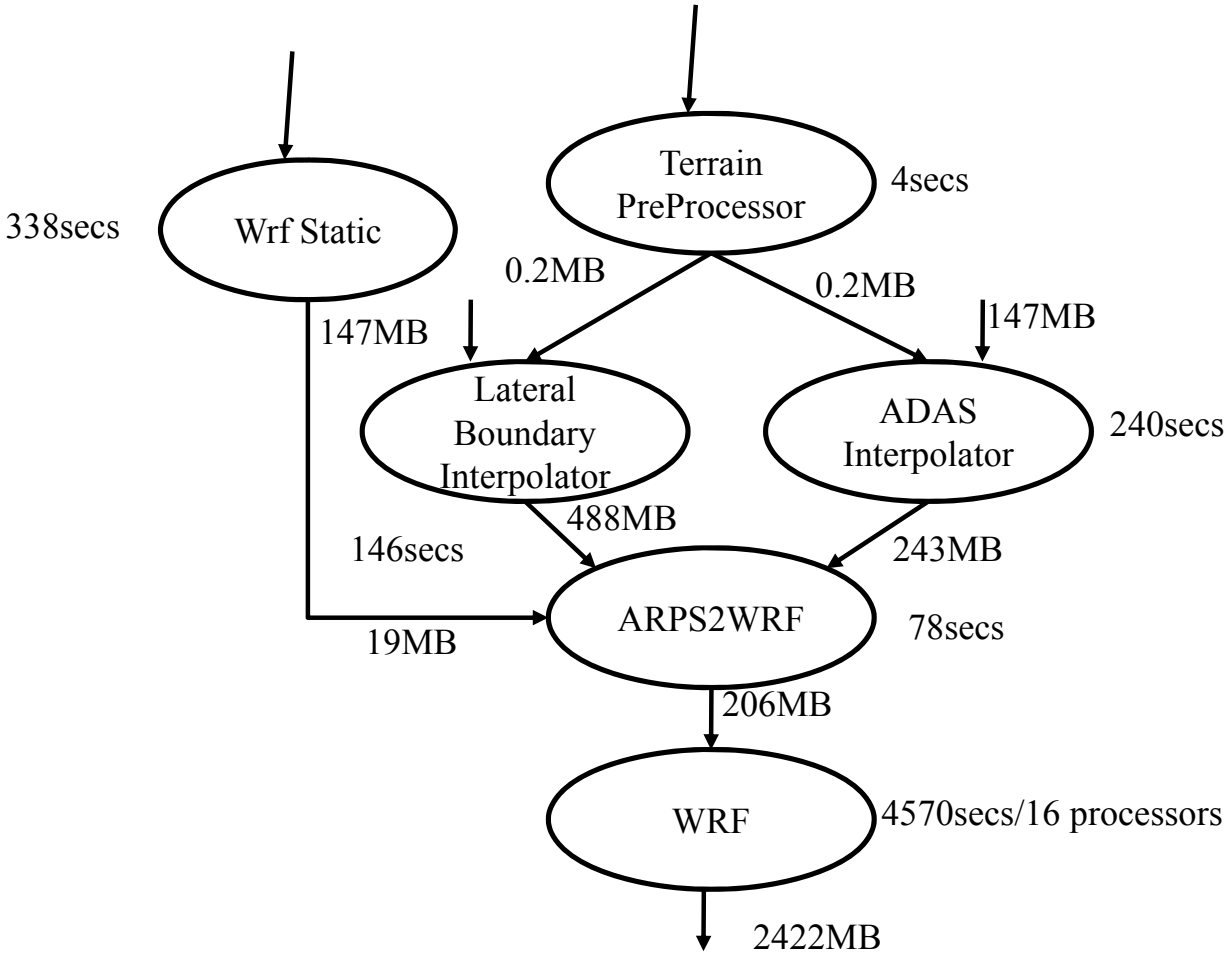
- Inputs, outputs, intermediate data, back end databases. Bioinformatics apps can be small data in, small data out, large result sets accessed from databases
  - Data reduction : size(output data) << size(input data)
  - Data production : size(output data) >> size(input data)
  - Data processing : size(input data) ≅ size(output data
- Usage scenarios
  - Interactive workflows – human-in-loop (i.e., steering)
  - Event-driven workflows – triggered by event
  - User constrained workflows – time or resource constrained



Workflow survey examples

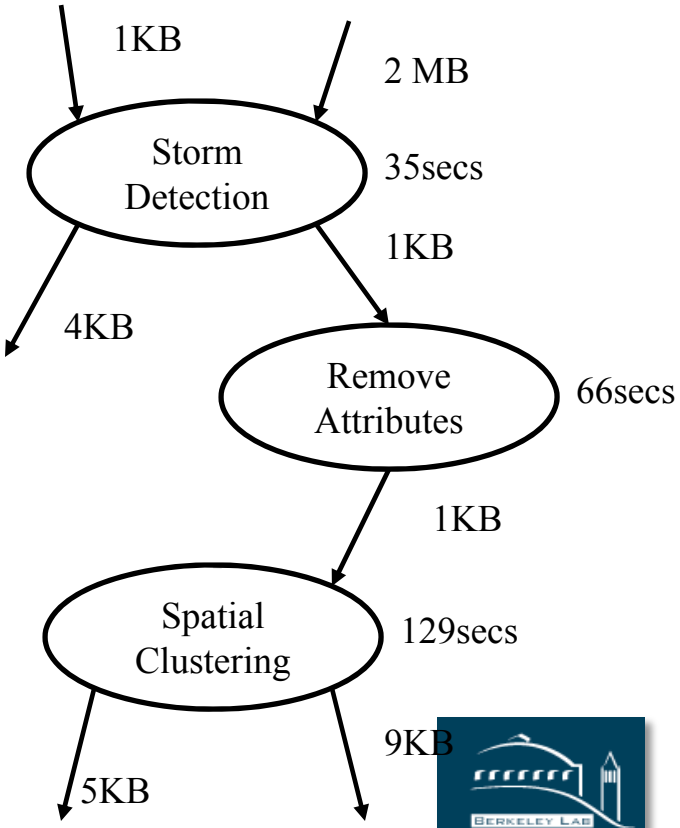
Domain	Project	Website	Tool
Weather and Ocean Modeling	Linked Environments for Atmospheric Discovery (LEAD) TeraGrid Science Gateway	http://portal.lead.project.org	xbaya, GPEL, Apache ODE
	Southeastern Coastal Ocean Observing and Prediction Program (SCOOP)	http://www.renci.org/focusareas/disaster/scoop.php	[Scripts]
	North Carolina Floodplain Mapping Program		[Scripts]
Bioinformatics and Biomedical	North Carolina Bioportal, TeraGrid Bioportal Science Gateway	http://www.renci.org/focusareas/biosciences/motif.php	Taverna
	MotifNetwork	http://www.motifnetwork.org/	Taverna
	National Biomedical Computation Resource (NBCR), Avian Flu Grid, Pacific Rim Application and Grid Middleware Assembly	http://nbcr.sdsc.edu/ http://gemstone.mozdev.org http://www.pragma-grid.net/ http://avianflugrid.pragma-grid.net/ http://mgltools.scripps.edu/	Kepler, Gemstone, [Scripts] and Vision
	cancer Biomedical Informatics Grid (caBIG)	http://www.cagrid.org/	Taverna
Astronomy	Pan-STARRS	http://pan-starrs.ifa.hawaii.edu/public/, http://www.psisc.org/	
Neutron Science	Spallation Neutron Source (SNS), Neutron Science TeraGrid Gateway (NSTG)	http://neutrons.ornl.gov/	

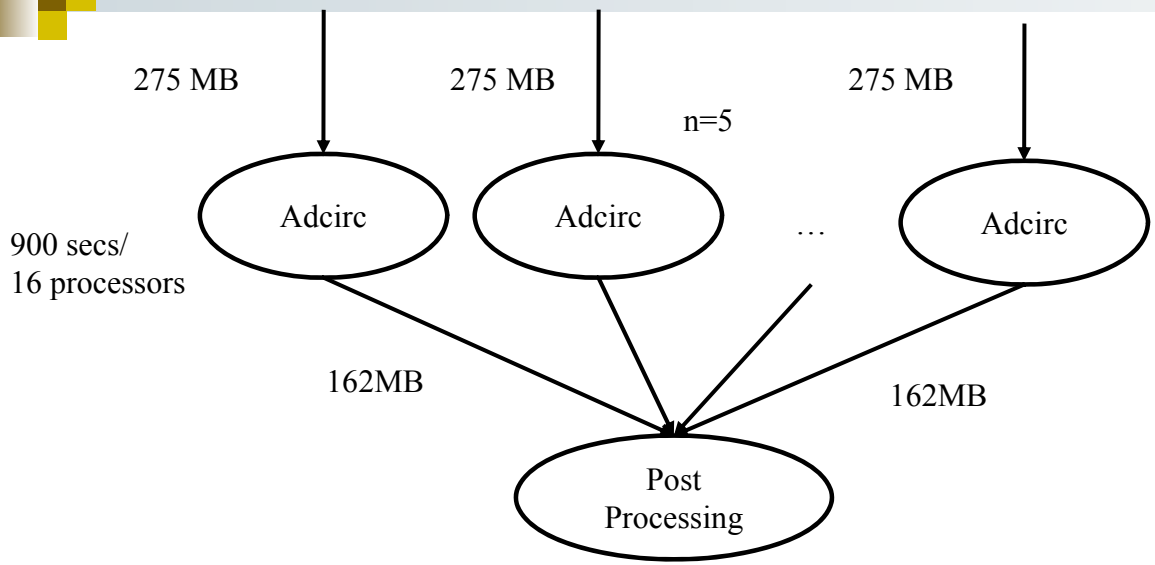
LEAD II workflows



Weather forecast workflow

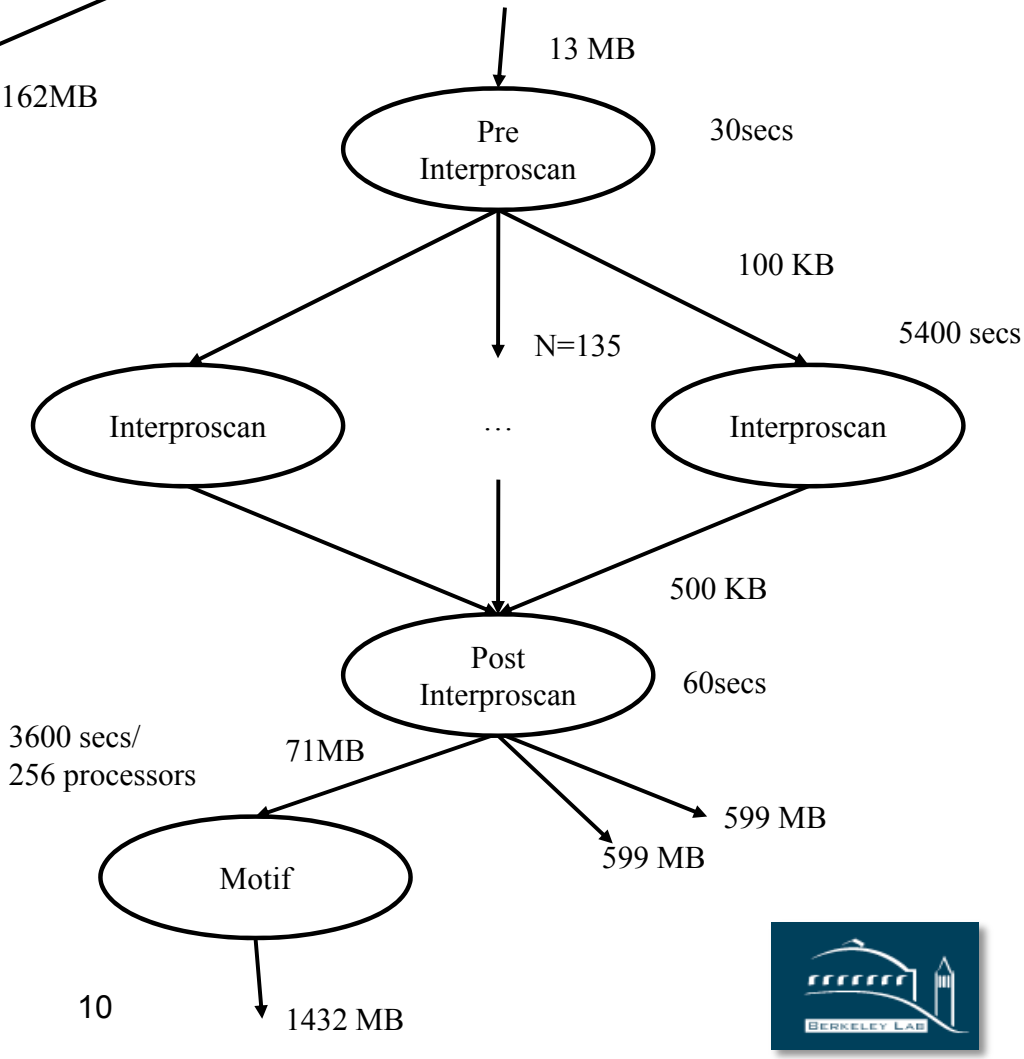
Data mining workflow



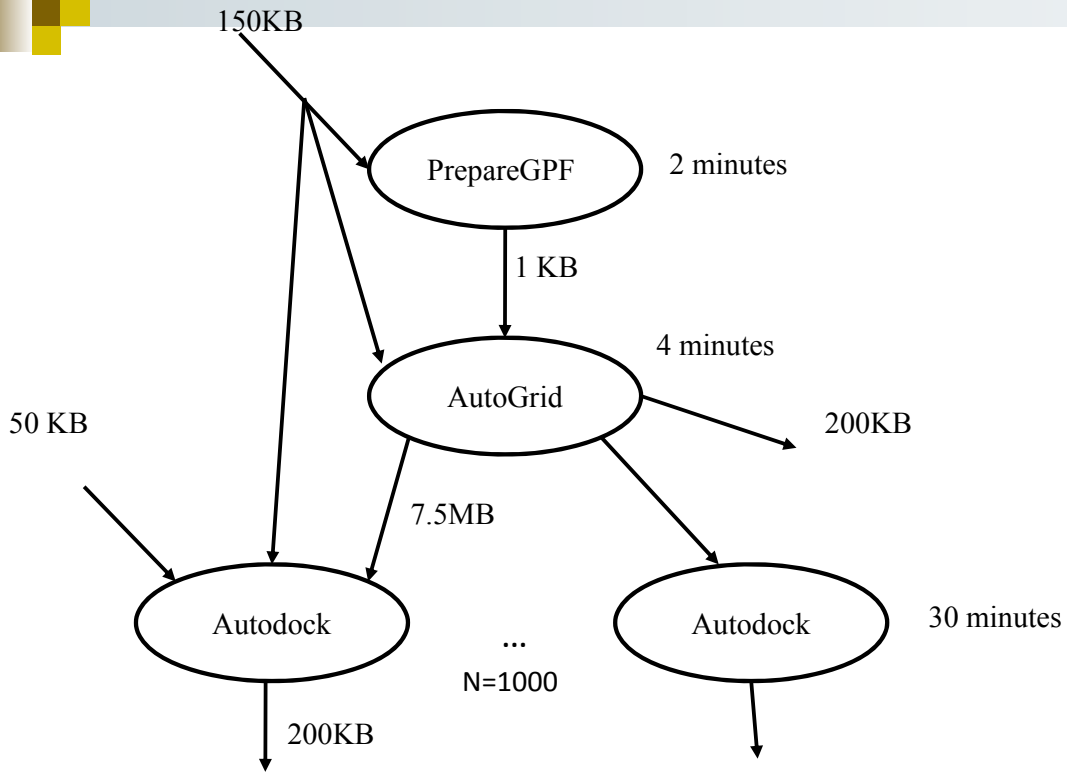


SCOOP workflow: arriving wind data triggers ADCIRC model used for storm-surge prediction during hurricane season

MOTIF workflow: motif/domain analysis of genome collections as input sequences

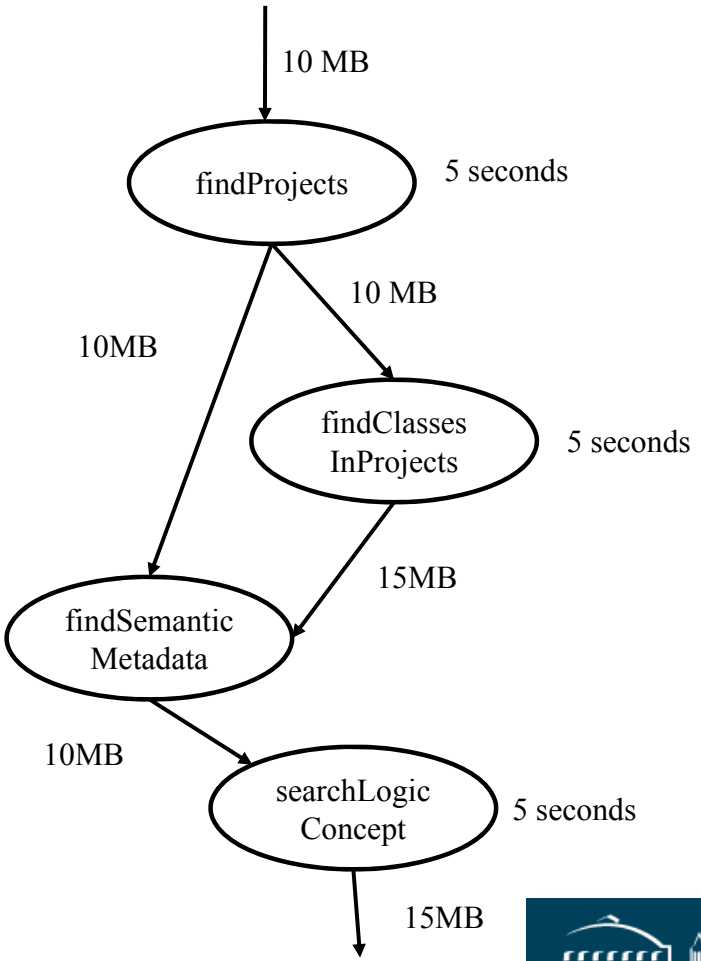


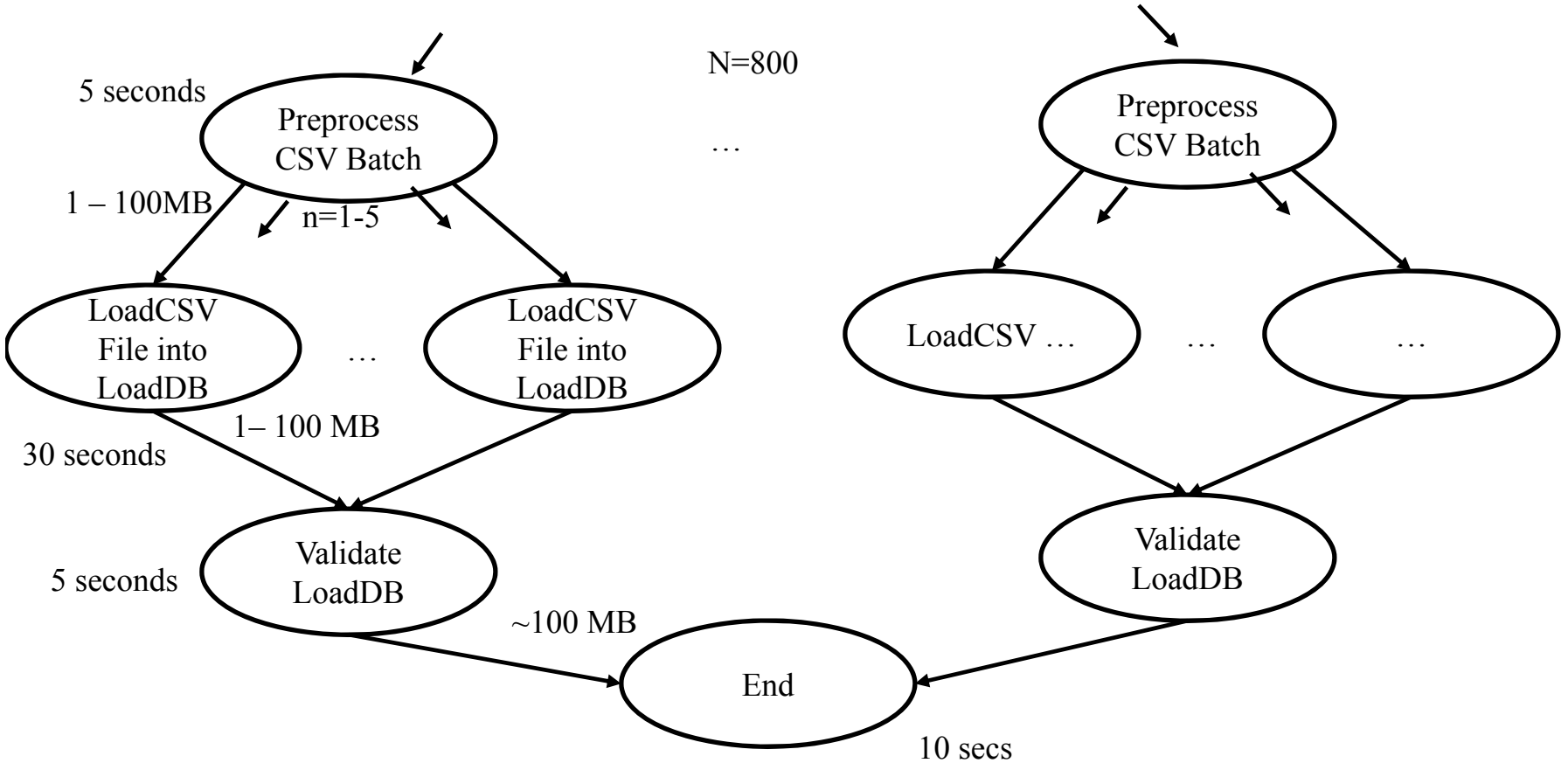
DATA TO INSIGHT CENTER
INDIANA UNIVERSITY
Pervasive Technology Institute



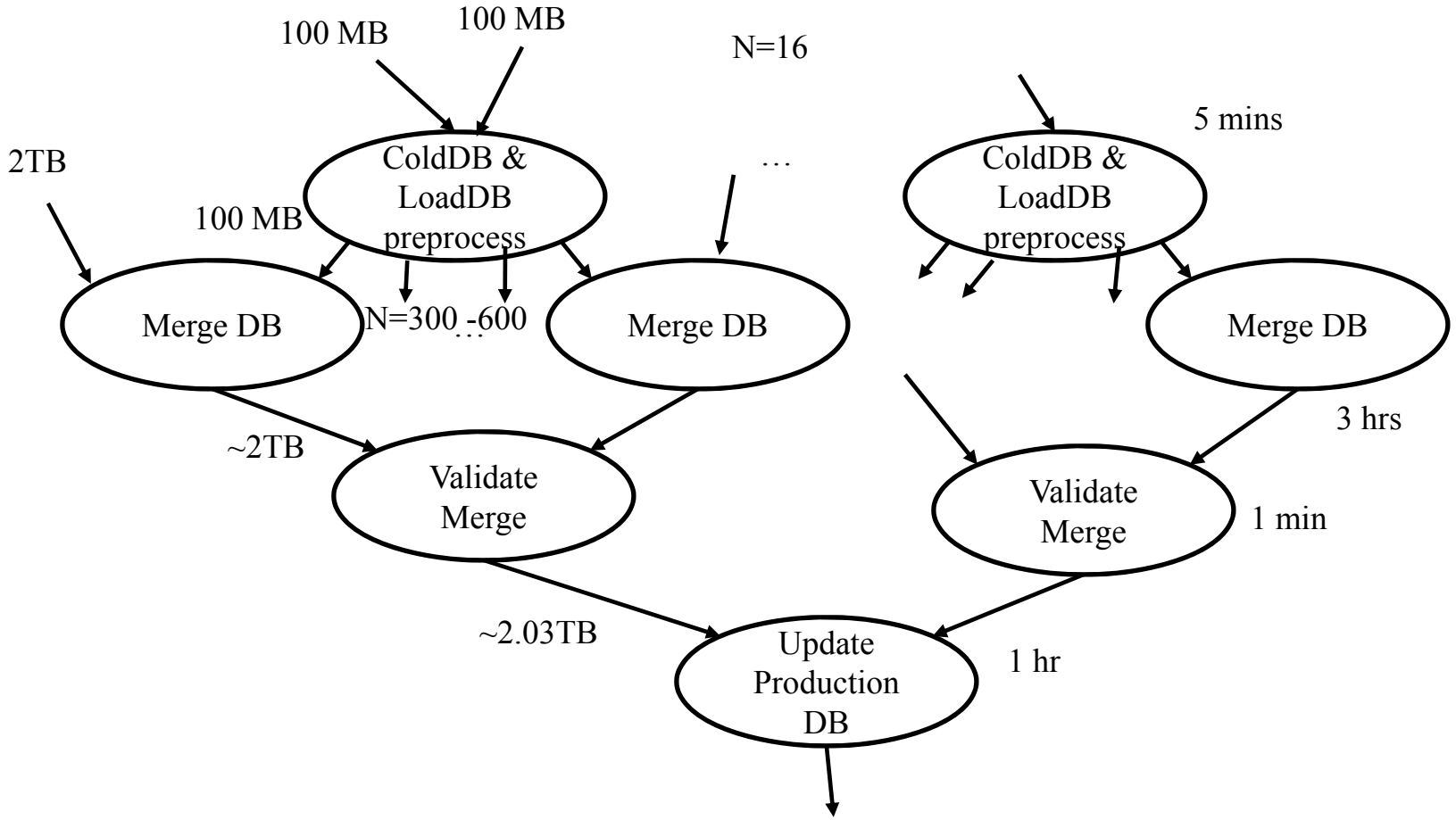
Avian Flu workflow: used in drug design to study interaction of drugs with environment; has large fan-out at end.

Cancer Data Standards Repository: queries concepts related to an input context





PanSTARRS workflow PSLoad: data arriving from PS1 telescope is processed and staged in relational databases each night.



PanSTARRS workflow PSMerge: each week, the production databases that astronomers query are updated with the new data staged during the week.

Conclusion

The proposed workflow classification model helps us think about workflows and workflow systems. As such, we hope it can help others and serve as a foundation for next generation workflow technologies.

Thanks!