

Workflow Algebra for Data Science

Eduardo Ogasawara

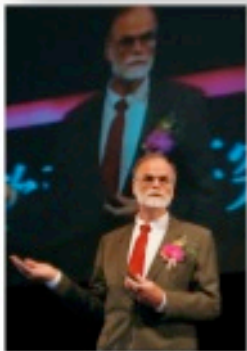
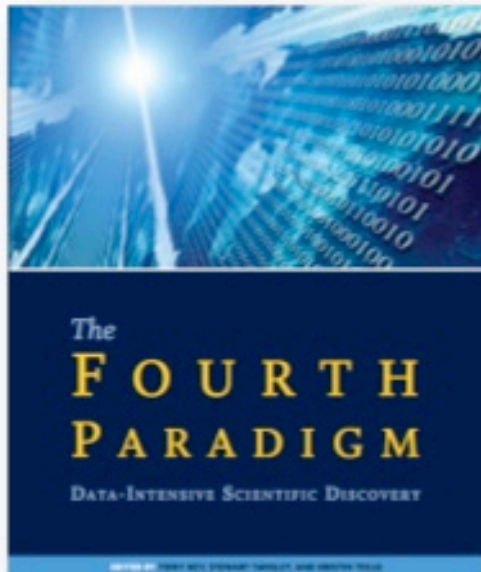
eogasawara@ieee.org



Federal Center of Technological Education
CEFET/RJ



The 4th Paradigm



Jim Gray on eScience: A Transformed Scientific Method

Based on the transcript of a talk given by Jim Gray
to the NRC-CSTB¹ in Mountain View, CA, on January 11, 2007²

Science Paradigms

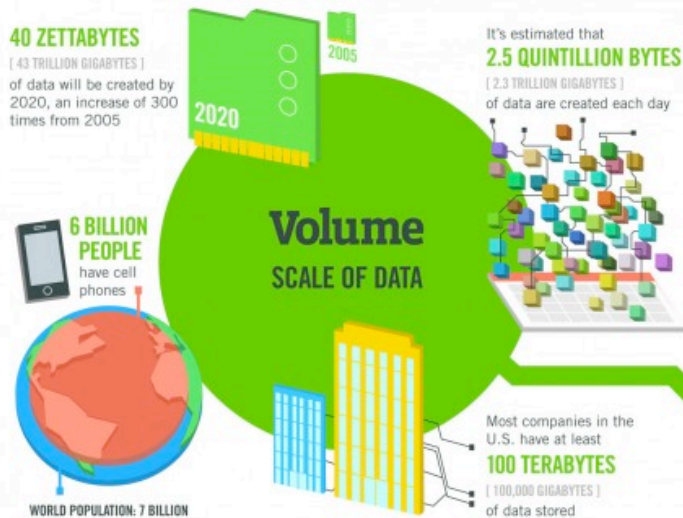
- Thousand years ago:
science was **empirical**
describing natural phenomena
- Last few hundred years:
theoretical branch
using models, generalizations
- Last few decades:
a **computational** branch
simulating complex phenomena
- Today: **data exploration** (eScience)
unify theory, experiment, and simulation
 - Data captured by instruments
or generated by simulator
 - Processed by software
 - Information/knowledge stored in computer
 - Scientist analyzes database/files
using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G \rho}{3} - \frac{\kappa c^2}{a^2}$$



- **CI, e-Science**
 - bioinformatics
 - ecoinformatics
 - geoinformatics
- **Big Data**
- **Data Science**
- **Information Science**
- **Digital Humanities ...**

Data (Big Data)



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States



As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES [161 BILLION GIGABYTES]



30 BILLION PIECES OF CONTENT are shared on Facebook every month



Variety DIFFERENT FORMS OF DATA

By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month



400 MILLION TWEETS are sent per day by about 200 million monthly active users



The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session



Velocity ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** - almost 2.5 connections per person on earth



Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure



1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions



27% OF RESPONDENTS

in one survey were unsure of how much of their data was inaccurate



Veracity UNCERTAINTY OF DATA

Poor data quality costs the US economy around

\$3.1 TRILLION A YEAR



Data Science - Definition

- Data science is the extraction of knowledge from data
- Models emerge from data
- Combines **Theory** and **Applied Computing**
- **Methods**
- **Data management**
- Applied Computing aids many domains
 - **Scientific**: biological, medical, astronomy ...
 - **Business**: enterprises and finance...
 - **Government**: economics and strategy...

Data Science - Related Fields



Data Science - Toolbox

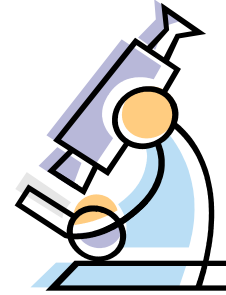
- Software Engineering practices
 - Software Reuse
 - Configuration Management: Git
- Data Analytics
 - R
 - Python
 - Orange
- Parallel Computing
 - Hadoop (Map-Reduce)
 - Pig-Latin
 - Spar

Typical scenario for Data Science experiments

This scenario demands the execution of many programs as a chain of activities, *i.e.*, workflow



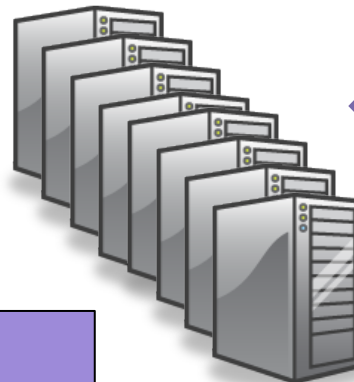
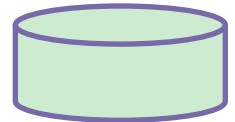
1. Data collection



2. Data analyzed by program A



3. Large Volume of Data Produced ...



4. ...which need to be processed by programs X,Y,Z in a HPC environment



5. Results are analyzed by program V

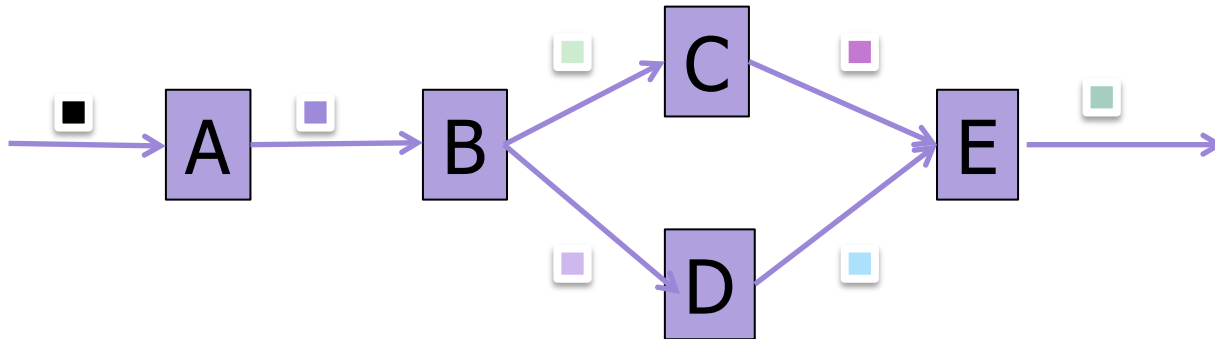
Computational intensive
Data intensive
Many programs executed one after another...
Steering
Reproducibility

Workflow – WfMC definition

- The automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules

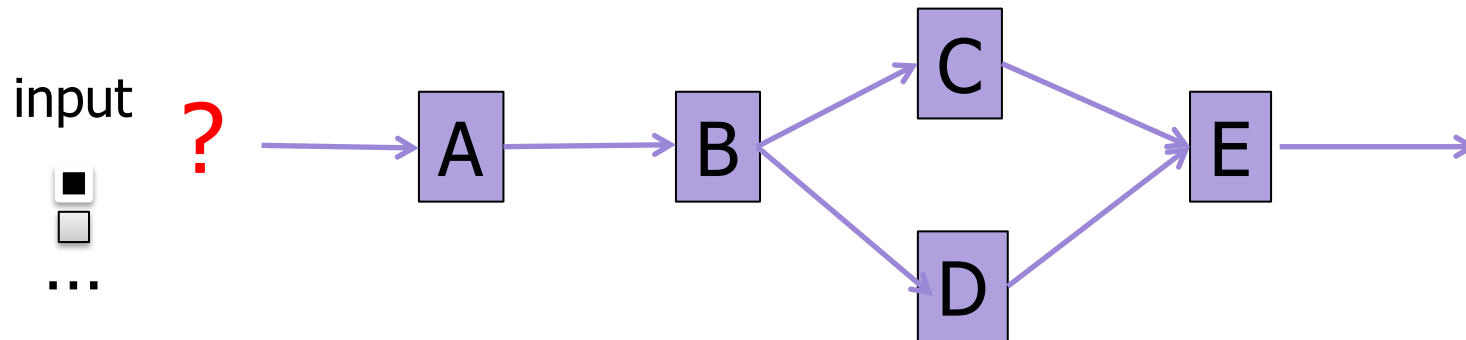
Workflow in the context of Data Science

- The automation of activities during which data is passed from one activity to another for processing establishing a data dependency between them



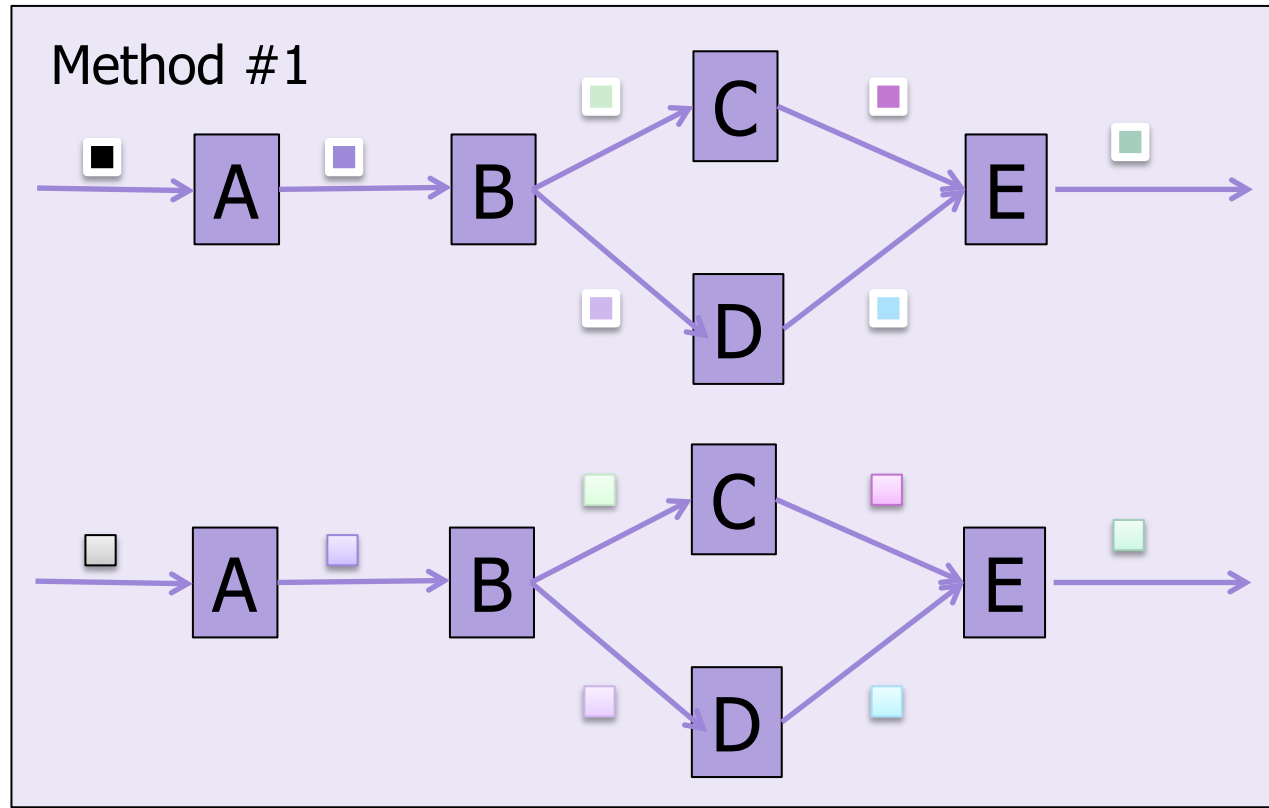
- Synonyms for workflow: Dataflow, Pipeline

Exploratory Analysis



What is the best way to execute it?

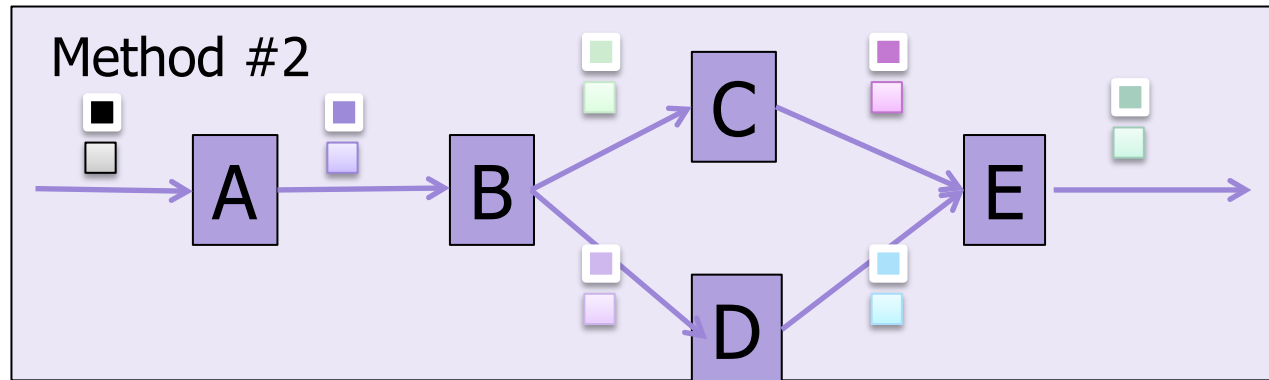
Exploratory Analysis



input



What is the best way to execute it?



Challenges related to Workflows for Data Science

- Workflow modeling
 - What is an activity?
 - What is data?
 - Simple
 - Uniform
 - Agnostic of computing infra-structure
- Workflow execution
 - Processing: Clusters, Grids, Clouds
 - Storage: Shared disk, Non-shared disk
 - Homogeneous, Heterogeneous, and Hybrid models
 - Parallel execution
 - Data Stage in / Stage out

Inspiration on Databases

- What is good about queries?
 - Declarative
 - Query processing is isolated from Query specification
 - Query processing is isolated from Hardware
 - Relational Algebra!
- What is good about transactions?
 - Queries execute in parallel
 - ACID properties

Workflow Algebra

- Workflow Representation
 - Activity
 - Data Model
- Workflow Execution Model
- Workflow Optimization Process
- Workflow Execution Plan

Data Model

- Relation are defined as a set of tuples that contains both primitive datatypes (integer, float, string, date, etc) and complex object (blobs, clobs)
- $\mathcal{R} = (\text{Serie: String; CaseStudy: String; TrainData: Blob; TestData: Blob})$
- Example of $R(\mathcal{R})$

<u>Serie</u>	<u>CaseStudy</u>	trainData	TestData
A	U-125	U-125T.DAT	U-125V.DAT
A	U-127	U-127T.DAT	U-127V.DAT
B	U-129	U-129T.DAT	U-129V.DAT

Relational Algebra Operators

- Basic Operators
 - Selection: σ
 - Projection: π
 - union: \cup
 - difference: $-$
 - Cross product: \times
 - Join: \bowtie
 - Aggregation: Γ
 - Division: \div

Relational Algebra Expressions

- Algebraic Expression E: $\pi_{A_1, A_2, \dots, A_n}(\sigma_P(r_1 \times r_2 \times \dots \times r_m))$
- Given E_1 and E_2 relational algebra expressions; the following expressions are also algebraic expressions:
 - $E_1 \cup E_2$
 - $E_1 - E_2$
 - $E_1 \times E_2$
 - $\sigma_p(E_1)$, p is a predicate over attributes of E_1
 - $\pi_s(E_1)$, S is a list of attributes of E_1

Activities

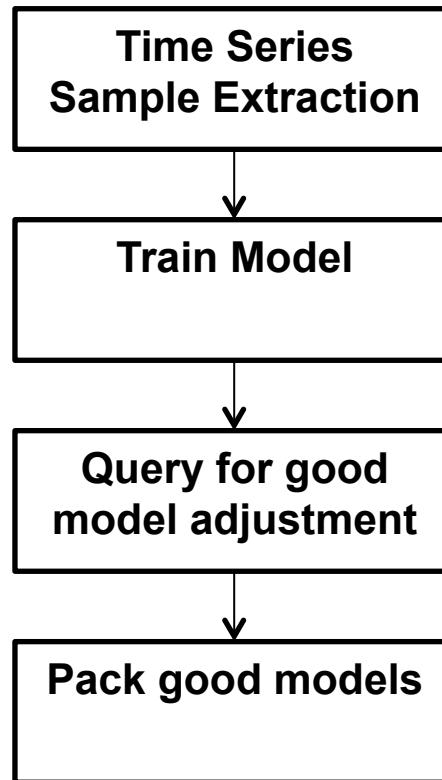
- A triple that specifies:
 - Schema \mathcal{R} for input relation
 - Schema \mathcal{S} for output relation
 - Program P that consumes a set of tuples to produce a set of tuples
- Notation: $Y \langle \mathcal{R}, \mathcal{S}, P \rangle$

Workflow Operators

- Program operators
 - Map (1:1)
 - SplitMap (1:n)
 - Reduce (n:1)
- Relational operators
 - Query (vector of relations: relation)

¹Ogasawara et al., 2011, An Algebraic Approach for Data-Centric Scientific Workflows, Proceedings of the VLDB Endowment.

Workflow for time series forecasting model adjusting



Split Map (1:n)

$O \leftarrow \text{SplitMap}(Y, I)$

R

<u>Serie</u>	Datasets
A	SerieA.txt
B	SerieB.txt

$S \leftarrow \text{SlipMap}(\text{sample-extraction}, R)$

s

<u>Serie</u>	<u>CaseStudy</u>	trainData	TestData
A	U-125	U-125T.csv	U-125V.csv
A	U-127	U-127T.csv	U-127V.csv
B	U-129	U-129T.csv	U-129V.csv
B	U-131	U-131T.csv	U-131V.csv

Map(1:1)

$O \leftarrow \text{Map}(Y, I)$

S

<u>Serie</u>	<u>CaseStudy</u>	<u>trainData</u>	<u>TestData</u>
A	U-125	U-125T.csv	U-125V.csv
A	U-127	U-127T.csv	U-127V.csv
B	U-129	U-129T.csv	U-129V.csv
B	U-131	U-131T.csv	U-131V.csv

$T \leftarrow \text{Map}(\text{train}, S)$

T

<u>Serie</u>	<u>CaseStudy</u>	<u>trainData</u>	<u>trainedModel</u>	<u>adjErr</u>	<u>TestData</u>
A	U-125	U-125T.csv	U-125.model	0.1	U-125V.csv
A	U-127	U-127T.csv	U-127.model	0.17	U-127V.csv
B	U-129	U-129T.csv	U-129.model	0.15	U-129V.csv
B	U-131	U-131T.csv	U-131.model	0.25	U-131V.csv

Reduce (n:1)

$O \leftarrow \text{Reduce}(Y, \{\text{Atr}\}, I)$

U

<u>Serie</u>	<u>CaseStudy</u>	<u>trainData</u>	<u>trainedModel</u>	<u>adjErr</u>	<u>TestData</u>
A	U-125	U-125T.csv	U-125.model	0.1	U-125V.csv
A	U-127	U-127T.csv	U-127.model	0.17	U-127V.csv
B	U-129	U-129T.csv	U-129.model	0.15	U-129V.csv

$V \leftarrow \text{Reduce}(\text{CompressRD}, \{\text{Serie}\}, U)$

V



<u>Serie</u>	<u>Pack</u>
A	SerieA.zip
B	SerieB.zip

Query Activity ($[Relation]:Relation$)

$O \leftarrow \text{Query}(E, [R])$

T

<u>Serie</u>	<u>CaseStudy</u>	<u>trainData</u>	<u>trainedModel</u>	<u>adjErr</u>	<u>TestData</u>
A	U-125	U-125T.csv	U-125.model	0.1	U-125V.csv
A	U-127	U-127T.csv	U-127.model	0.17	U-127V.csv
B	U-129	U-129T.csv	U-129.model	0.15	U-129V.csv
B	U-131	U-131T.csv	U-131.model	0.25	U-131V.csv

$U \leftarrow \text{Query}(\sigma_{adjErr < 0.2}(Q_0), Q[T])$

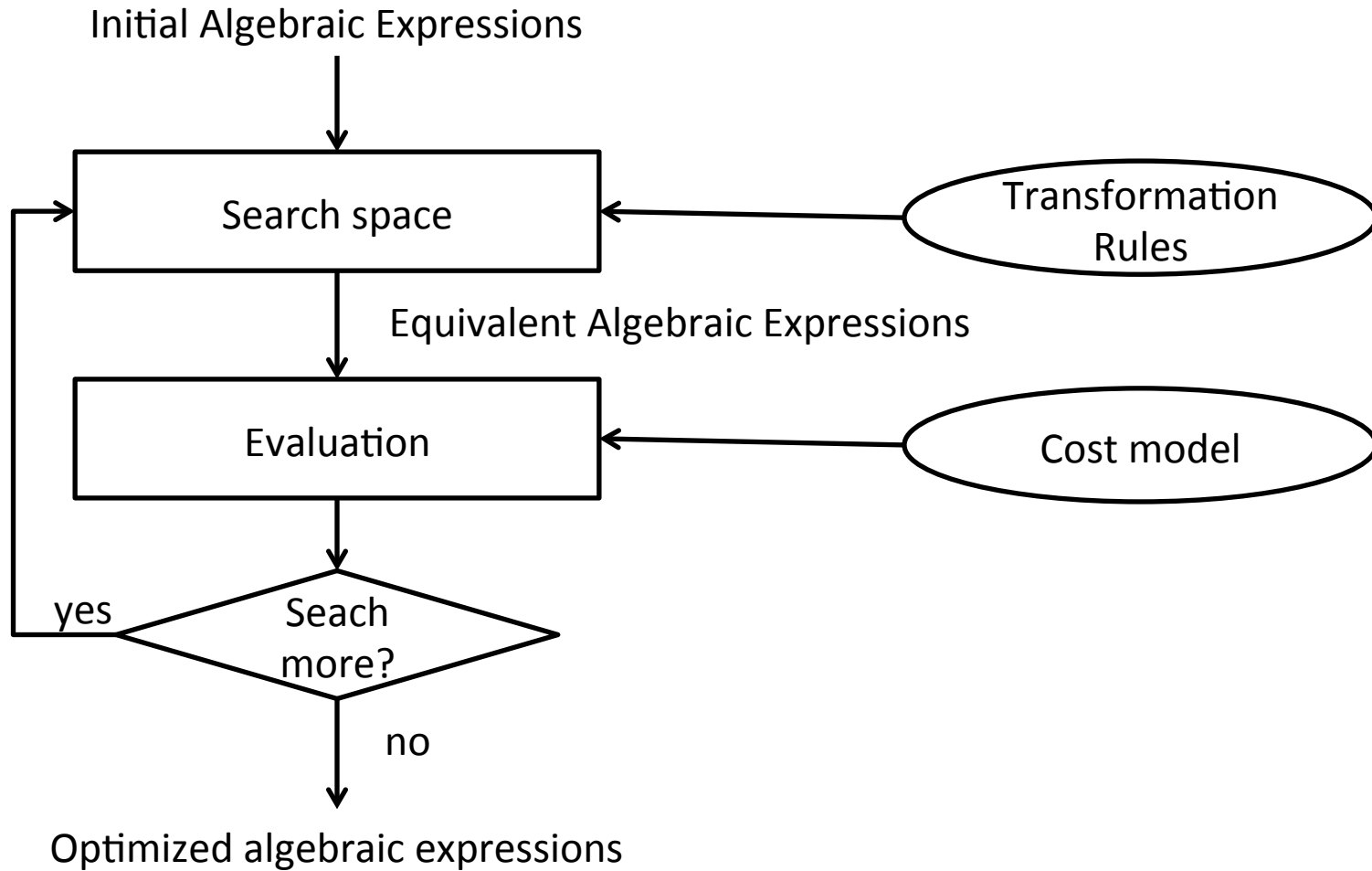
U

<u>Serie</u>	<u>CaseStudy</u>	<u>trainData</u>	<u>trainedModel</u>	<u>adjErr</u>	<u>TestData</u>
A	U-125	U-125T.csv	U-125.model	0.1	U-125V.csv
A	U-127	U-127T.csv	U-127.model	0.17	U-127V.csv
B	U-129	U-129T.csv	U-129.model	0.15	U-129V.csv

Workflow as Algebraic Expressions

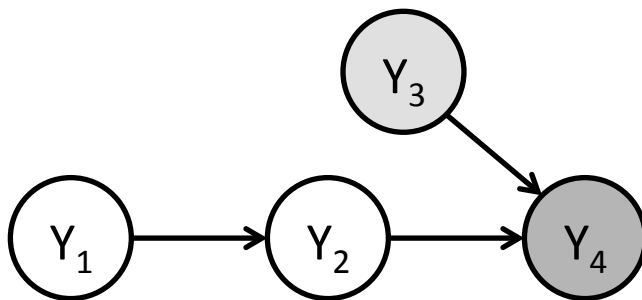
- $S \leftarrow \text{SlipMap}(\text{sample-extraction}, R)$
- $T \leftarrow \text{Map}(\text{train}, S)$
- $U \leftarrow \text{Query}(\sigma_{\text{adjErr} < 0.2}(Q_0), Q[T])$
- $V \leftarrow \text{Reduce}(\text{CompressRD}, \{\text{Serie}\}, U)$

Workflow Optimization Process



Workflow Fragments

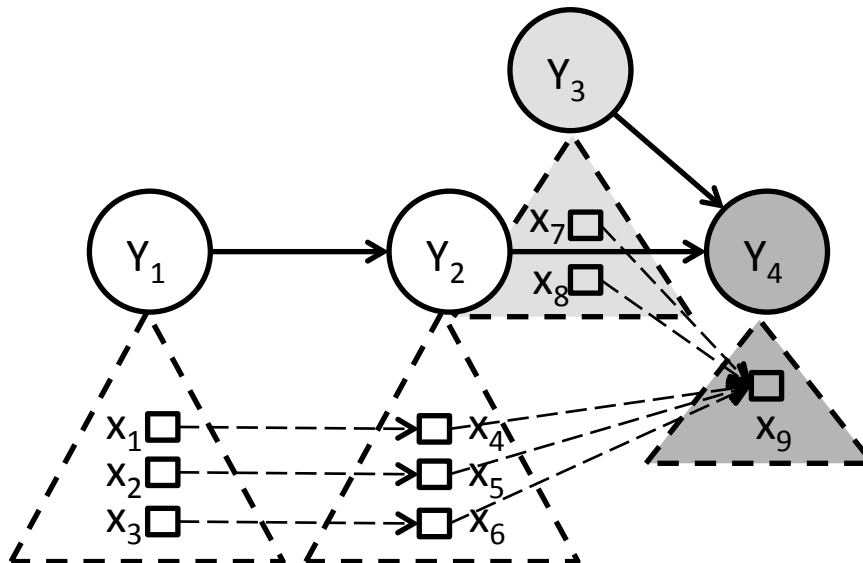
- A fragment F of a workflow is a subset F of the activities of a workflow W :
 - either F is an unitary set
 - or $\forall Y_j \in F, \exists Y_i \in F \mid (Dep(Y_i, Y_j)) \vee (Dep(Y_j, Y_i))$.



- Fragment F1: Y_1 and Y_2
- ◻ Fragment F2: Y_3
- ◼ Fragment F3: Y_4

Execution Strategies: Pipeline versus Materialization

- Pipeline(P) partitions a set of activations in a fragment into a complete list of dependent activations.
Instantiation of relations are based on pipelines.
- Materialization (M) partitions a set of activations in a fragment into a complete list of independent activations ordered by activity dependence.
Partial relations are fully instantiated before executing.



FTF:

$\{ \langle X_1, X_4 \rangle, \langle X_2, X_5 \rangle, \langle X_3, X_6 \rangle \}$

FAF:

$\{ \langle X_1 \rangle, \langle X_2 \rangle, \langle X_3 \rangle, \langle X_4 \rangle, \langle X_5 \rangle, \langle X_6 \rangle \}$

Algebraic Transformations

- Goal reduce number of tuples in intermediate relations
- Ex.: Query(E, [Map(Y, R),S])
- **Distributive**
 - $\text{Query}(E, [\text{Map}(Y, R), S]) = \text{Query}(E_2, [\text{Query}(E_1, [\text{Map}(Y, R)], S])$
- **Push down-selection**
 - $\text{Query}(E_2, [\text{Query}(E_1, [\text{Map}(Y, R)], S]) = \text{Query}(E_2, \text{Map}(Y, \text{Query}(E_1, [R]), S)$

Workflow Execution Plan

- Produce Algebraic Transformation
- Fragment Workflow
- Assign Execution Strategies for Workflow Fragments

Workflow Execution

- Wait for Sagitarii presentation...

Workflow Algebra for Data Science

Eduardo Ogasawara

eogasawara@ieee.org



Federal Center of Technological Education
CEFET/RJ

