Workflow Algebra for Data Science

Eduardo Ogasawara

eogasawara@ieee.org

Federal Center of Technological Education
CEFET/RJ

Data Science

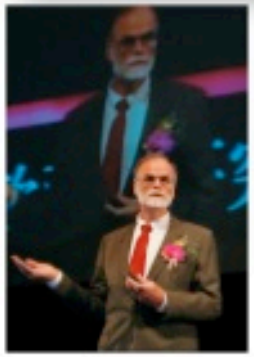
Science

The 4th Paradigm

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

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

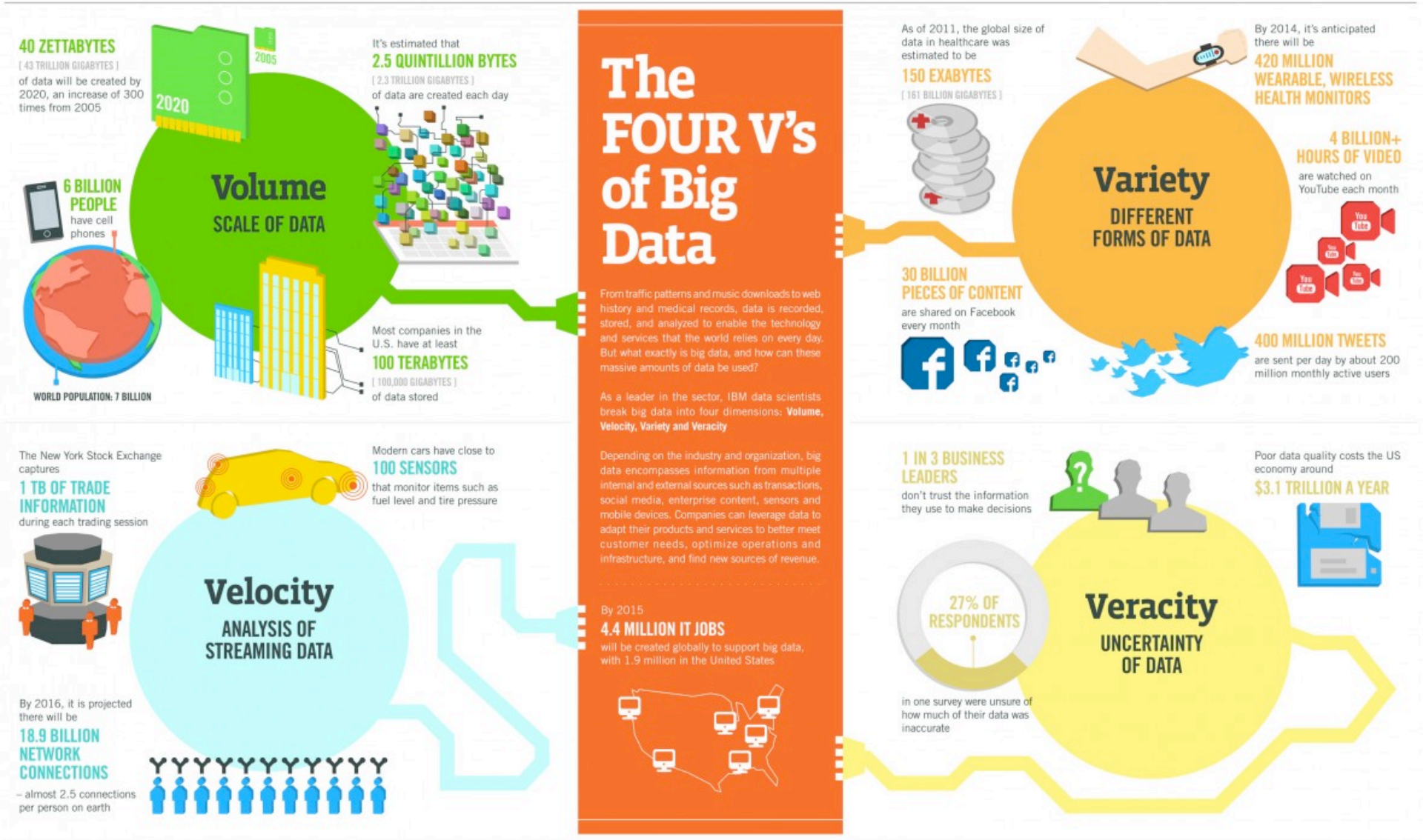


Jim Gray on eScience:
A Transformed Scientific Method

Based on the transcript of a talk given by Jim Gray to the NRC-CSTB[1] in Mountain View, CA, on January 11, 2007[2]

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

Data (Big Data)



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

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…

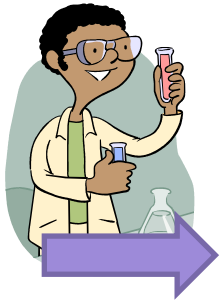
Data Science - Related Fields

Data Science - Toolbox

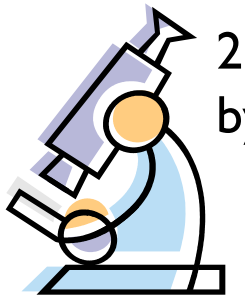
- Software Engineering practices
  - Software Reuse
  - Configuration Management: Git
- Data Analytics
  - R
  - Python
  - Orange
- Parallel Computing
  - Hadoop (Map-Reduce)
  - Pig-Latin
  - Spar

Typical scenario for Data Science experiments

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

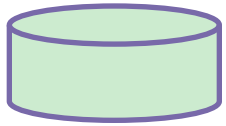


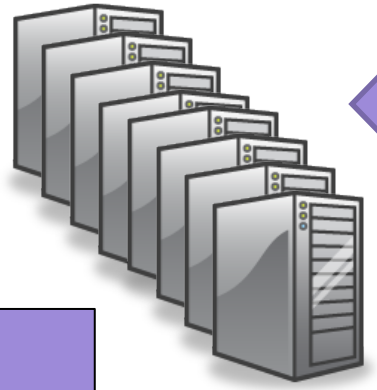
1. Data collection



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

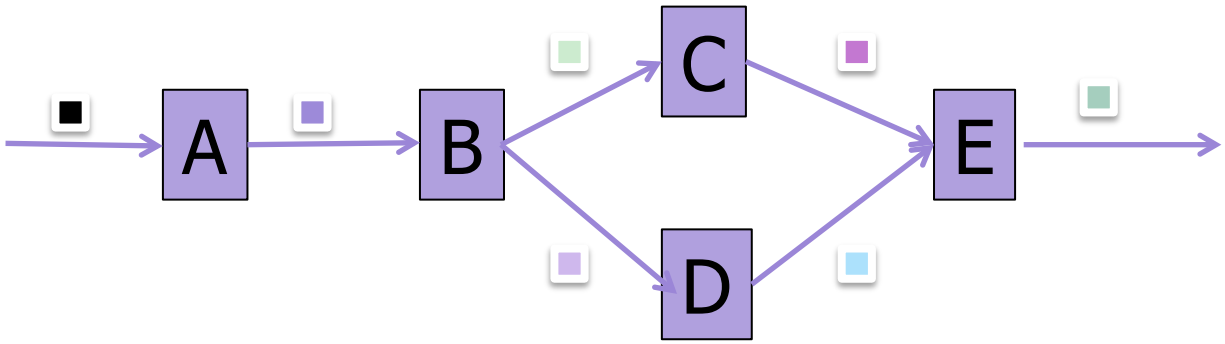
Computational intensive
Data intensive
Many programs executed one after another...
Steering
Reproducibility

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

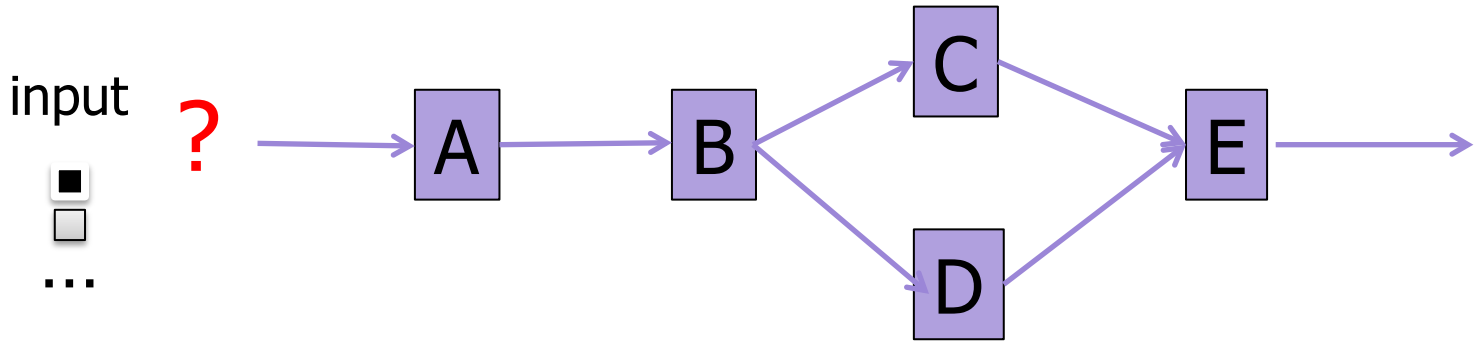
Workflow in the context of Data Science

- 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

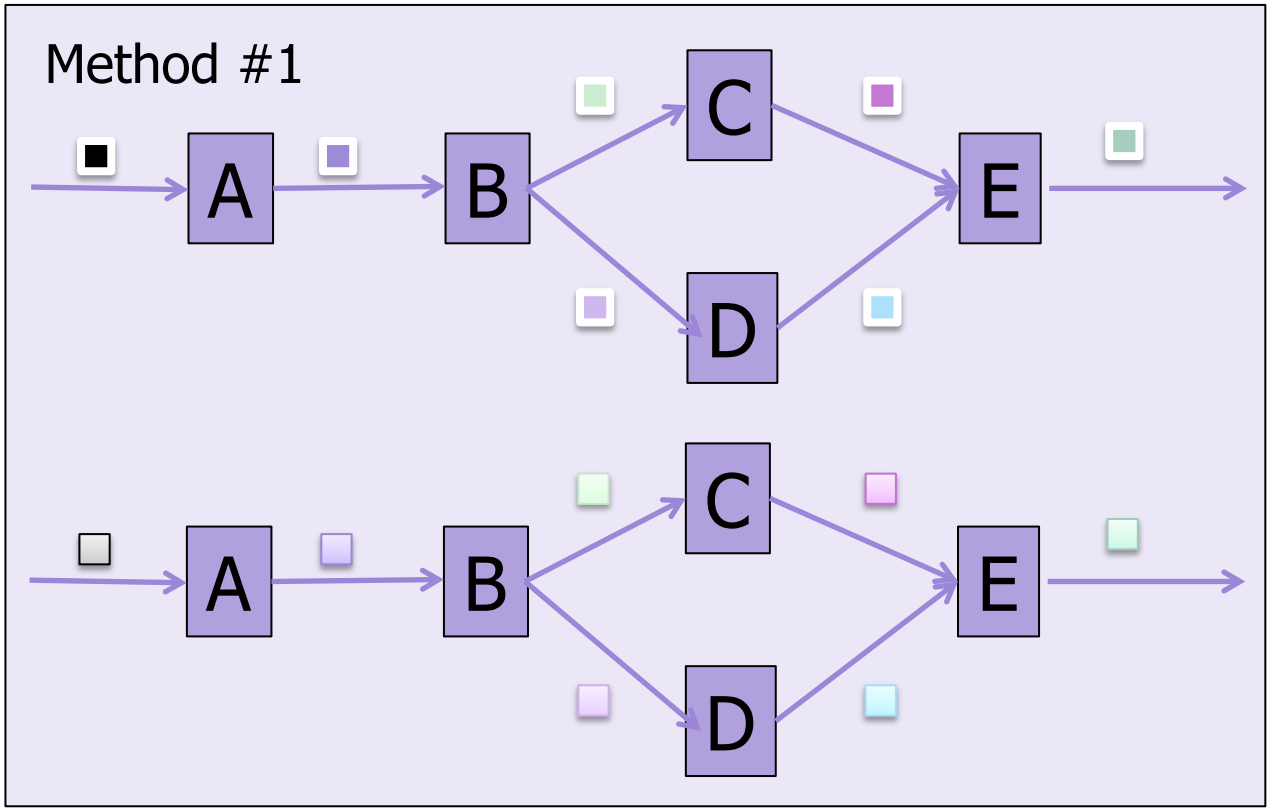


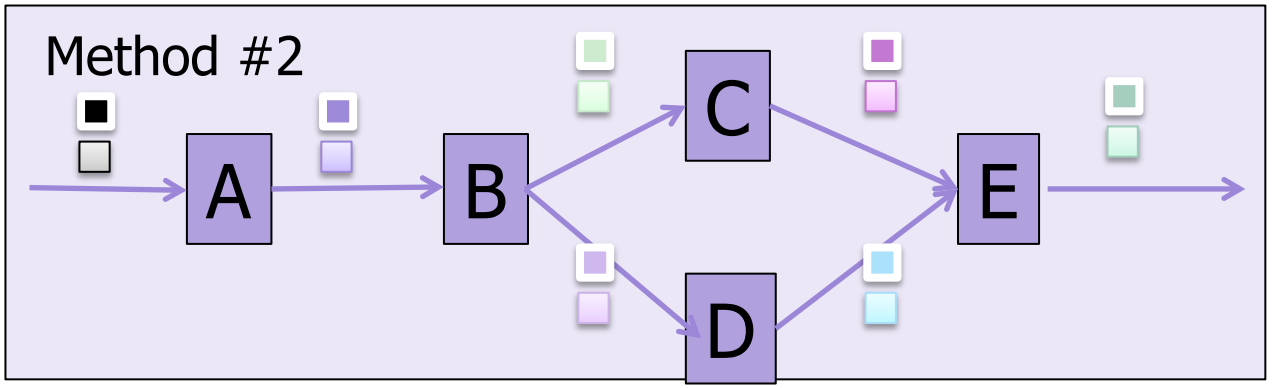
What is the best way to execute it?

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?

- What is good about transactions?

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)

- $\mathcal{R}$ = (Serie: String; CaseStudy: String; TrainData: Blob; TestData: Blob)

- Example of R($\mathcal{R}$)

Serie	**CaseStudy**	**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: ∪
  - difference: −
  - Cross product: ×
  - Join: ⋈
  - Aggregation: Γ
  - Division: ÷

Relational Algebra Expressions

- Algebraic Expression E: $\pi_{A_1, A_2, \ldots, A_n}(\sigma_P(r_1 \times r_2 \times \ldots \times r_m))$
- Given E_1 and E_2 relational algebra expressions; the following expression 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<$\mathcal{R}$, $\mathcal{S}$, *P*>

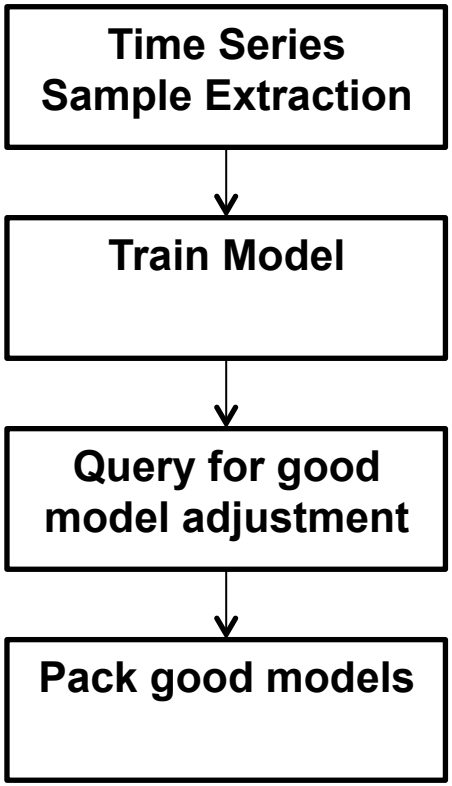
Workflow Operators

- Program operators
  - Map (1:1)
  - SplitMap (1:n)
  - Reduce (n:1)

- Relational operators
  - Query (vector of relations: relation)

[1]Ogasawara 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 ← SplitMap(Y, I)

R

Serie	Datasets
A	SerieA.txt
B	SerieB.txt

S ← SlipMap(sample-extraction, R)

S

Serie	CaseStudy	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 ← Map(Y, I)

S

Serie	CaseStudy	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

T ← Map(train, S)

T

Serie	CaseStudy	trainData	trainedModel	adjErr	TestData
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 ← Reduce(Y, {Atr}, I)

U

Serie	CaseStudy	trainData	trainedModel	adjErr	TestData
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 ← Reduce(CompressRD, {Serie}, U)

V

Serie	Pack
A	SerieA.zip
B	SerieB.zip

Query Activity ([Relation]:Relation)

O ← Query(E, [R])

T

Serie	CaseStudy	trainData	trainedModel	adjErr	TestData
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

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

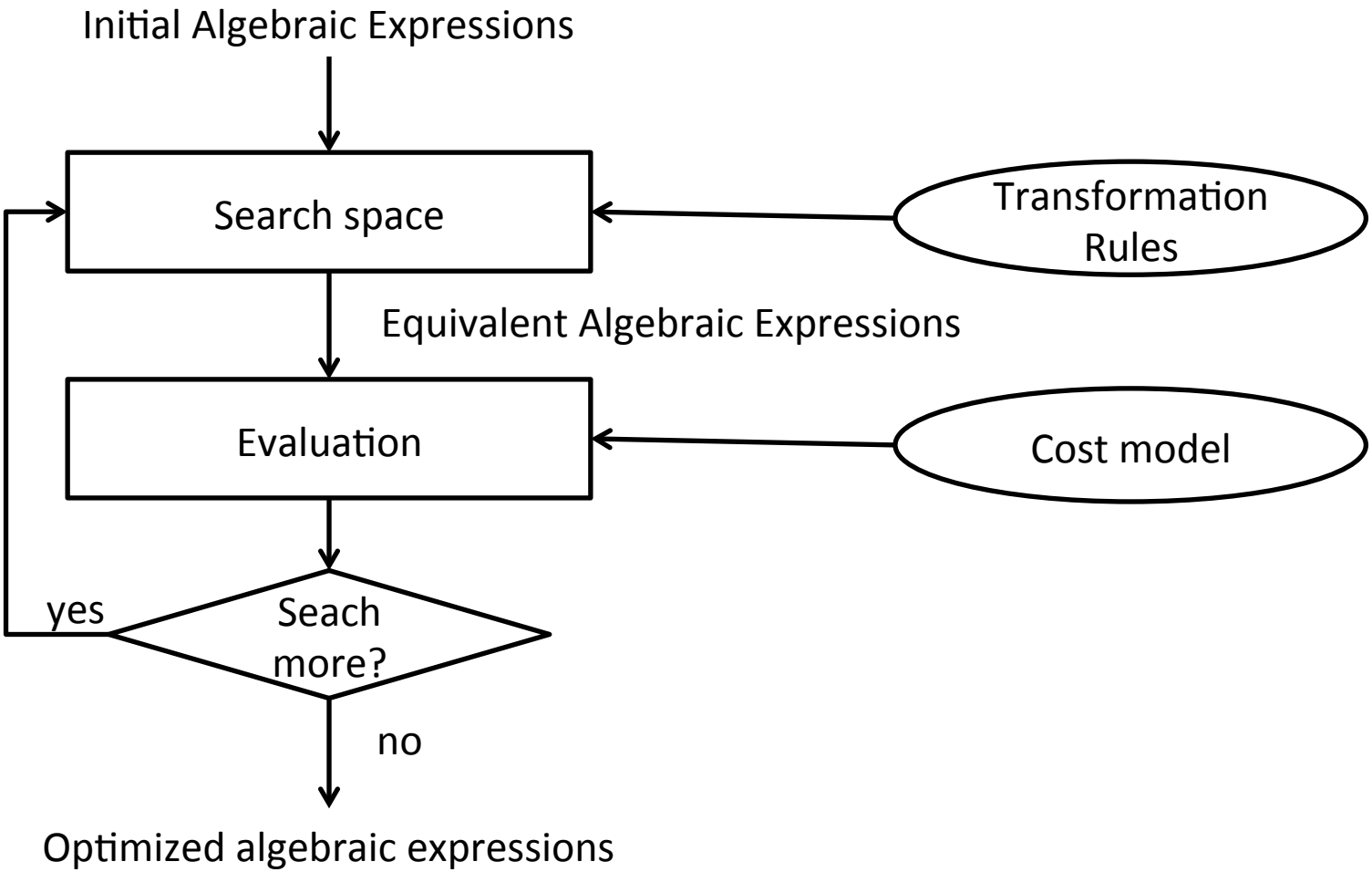
U

Serie	CaseStudy	trainData	trainedModel	adjErr	TestData
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

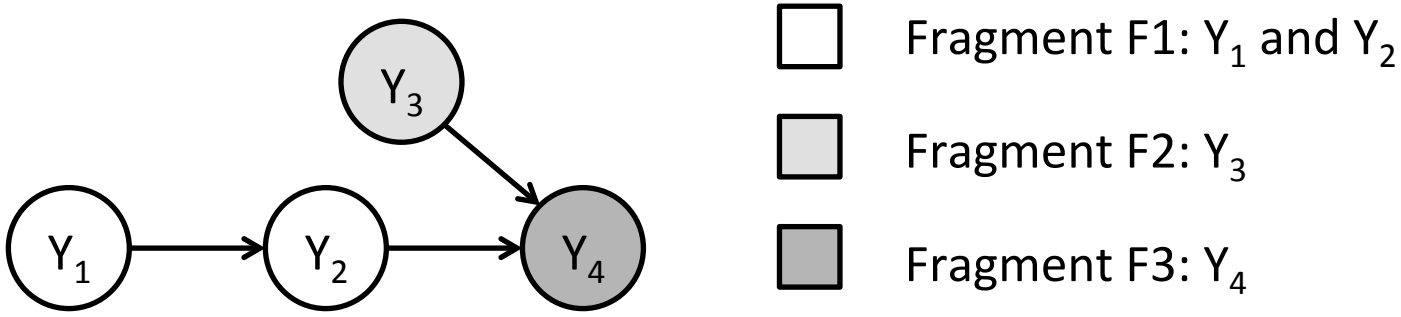
- S ← SlipMap(sample-extraction, R)
- T ← Map(train, S)
- $U \leftarrow Query(\sigma_{adjErr<0.2}(Q_0), Q[T])$
- V ← Reduce(CompressRD, {Serie}, U)

Workflow Optimization Process



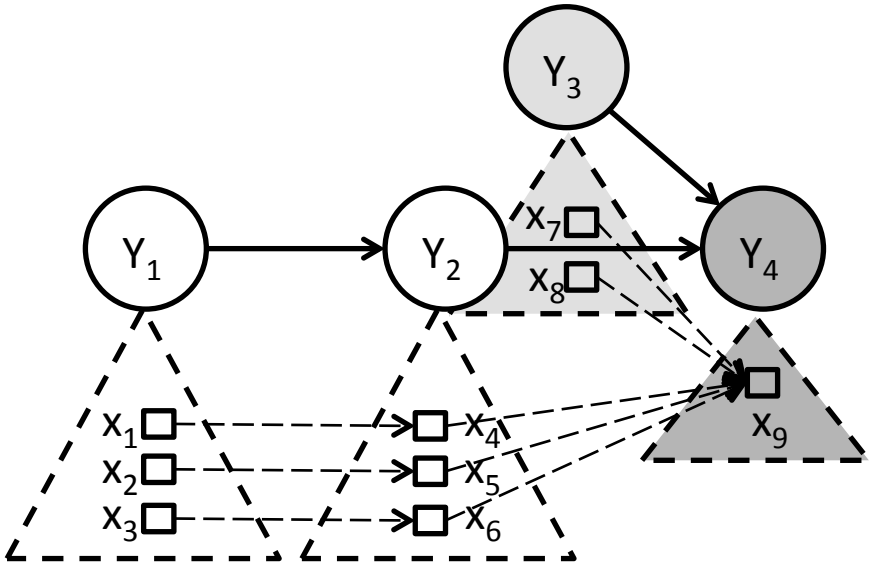
Workflow Fragments

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



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:
{<x_1, x_4>, <x_2, x_5>, <x_3, x_6>}

FAF:
{<x_1>,<x_2>,<x_3>,
<x_4>,<x_5>,<x_6>}

Algebraic Transformations

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