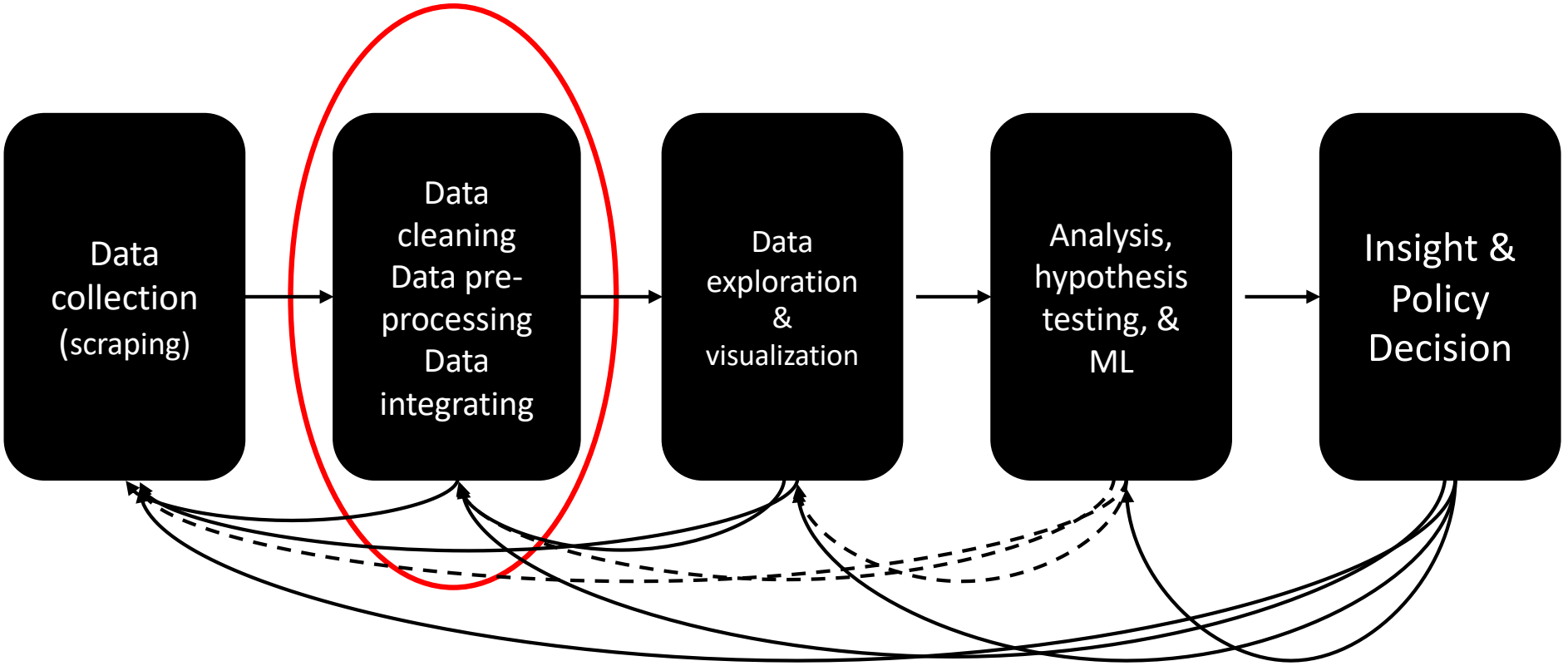
HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Data integration and preprocessing

Outline

- Data integration
  - Introduction
  - Current approaches
  - Apache Nifi
  - Hand-ons Apache Nifi
- Data preprocessing
  - Introduction
  - Data quality
  - Data preprocessing steps
  - Hand-ons Openrefine

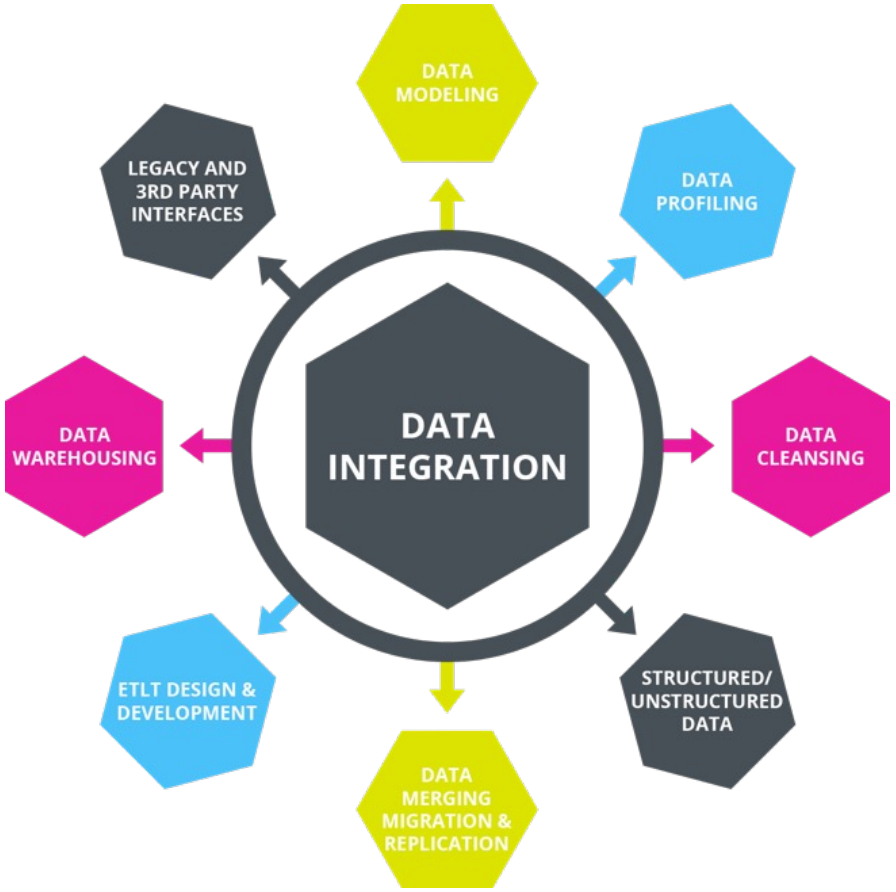
Recall: insight-driven DS methodology



Data integration

Data integration

- Provide uniform access to data available in multiple, autonomous, heterogeneous and distributed data sources
  - Uniform
  - Access to
  - Multiple
  - Autonomous
  - Heterogeneous
  - Distributed
  - Data Sources



Why data integration

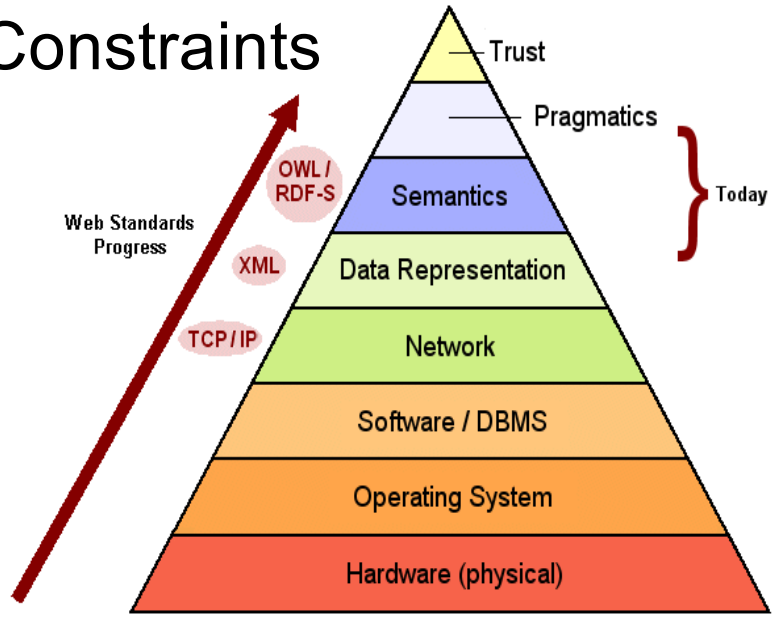
- To facilitate information access and reuse through **a single information access point**
- Data from different complementing information systems is **to be combined to gain a more comprehensive basis** to satisfy the need
  - Improve decision making
  - Improve customer experience
  - Increase competitiveness, Streamline operations
  - Increase productivity
  - Predict the future

Data integration challenges

- Physical systems
  - Various hardwares, standards
  - Distributed deployment
  - Various data format
- Logical structures
  - Different data models
  - Different data schemas
- Business organization
  - Data security and privacy
  - Business rules and requirements
  - Different administrative zones in the business organization

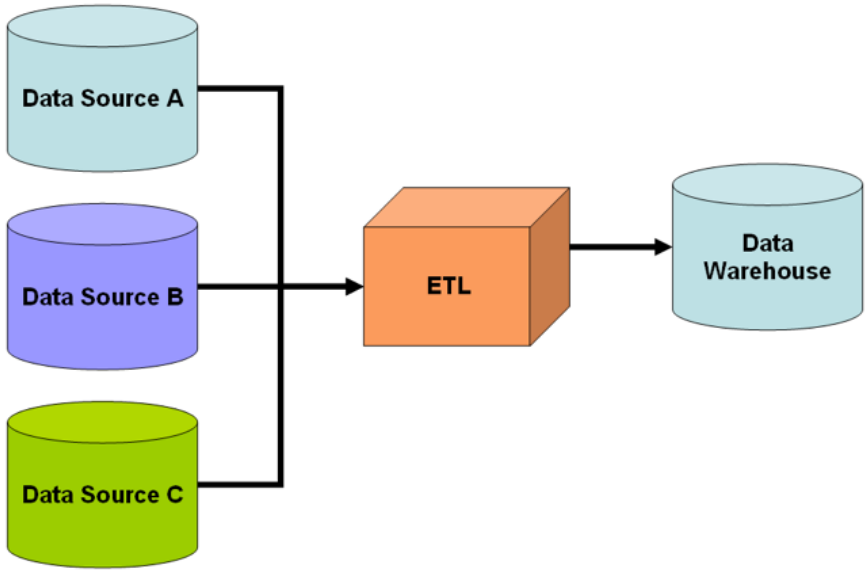
Kinds of Heterogeneity

- Hardware and Operating Systems
- Data Management Software
- Data Models, Schemas and Semantic
- Middle-ware
- User Interfaces
- Business Rules and Integrity Constraints



Current approaches

- Data Warehouse
  - Realize a common data storage approach
  - Data from several operational sources (OLTP) are extracted, transformed, and loaded (ETL) into a data warehouse
  - Analysis, such as OLAP, can be performed on cubes of integrated and aggregated data



ETL process

- 70-80% of BI (DI or DW) project is reliable ETL process
- ETL = Extract – Transform – Load
- Extract
  - Get the data from source system as efficiently as possible
- Transform
  - Perform calculations on data
- Load
  - Load the data in the target storage

Why is ETL (System) Important?

- Adds **value** to data
  - Removes mistakes and corrects data
  - Documented measures of confidence in data
  - Captures the flow of transactional data
  - Adjusts data from multiple sources to be used together (conforming)
  - Structures data to be usable by BI tools
  - Enables subsequent business / analytical data procesing

ETL market

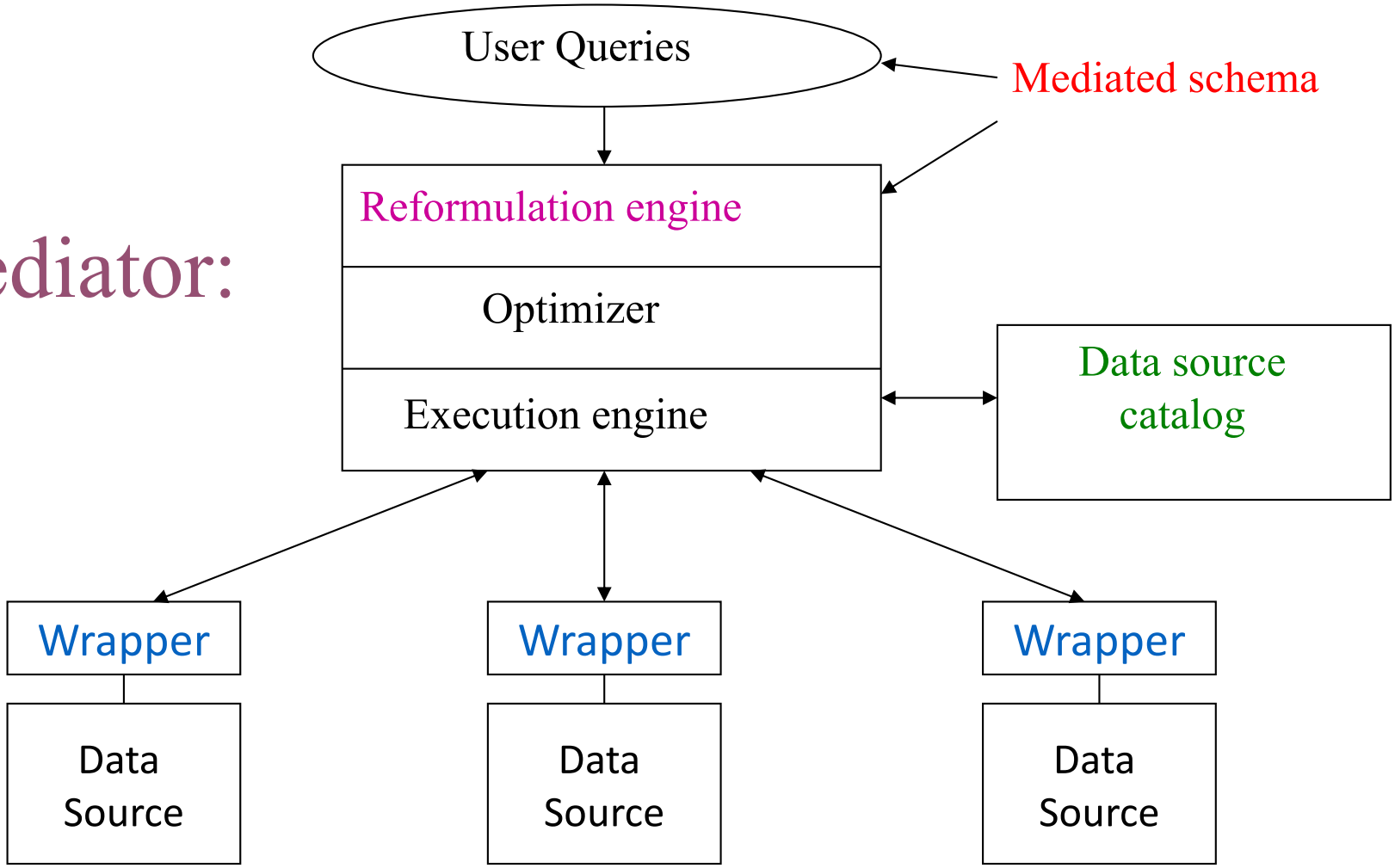

CHALLENGERS
LEADERS
Informatica
IBM
Microsoft
SAP
Oracle
SAS
Talend
Information Builders
Syncsort
Actian
Adeptia
Cisco
Denodo
NICHE PLAYERS
VISIONARIES
ABILITY TO EXECUTE
COMPLETENESS OF VISION
As of July 2015

Problems with DW approach

- Data has to be cleaned – different formats
- Needs to store all the data in all the data sources that will ever be asked for
  - Expensive due to data cleaning and space requirements
- Data needs to be updated periodically
  - Data sources are autonomous – content can change without notice
  - Expensive because of the large quantities of data and data cleaning costs

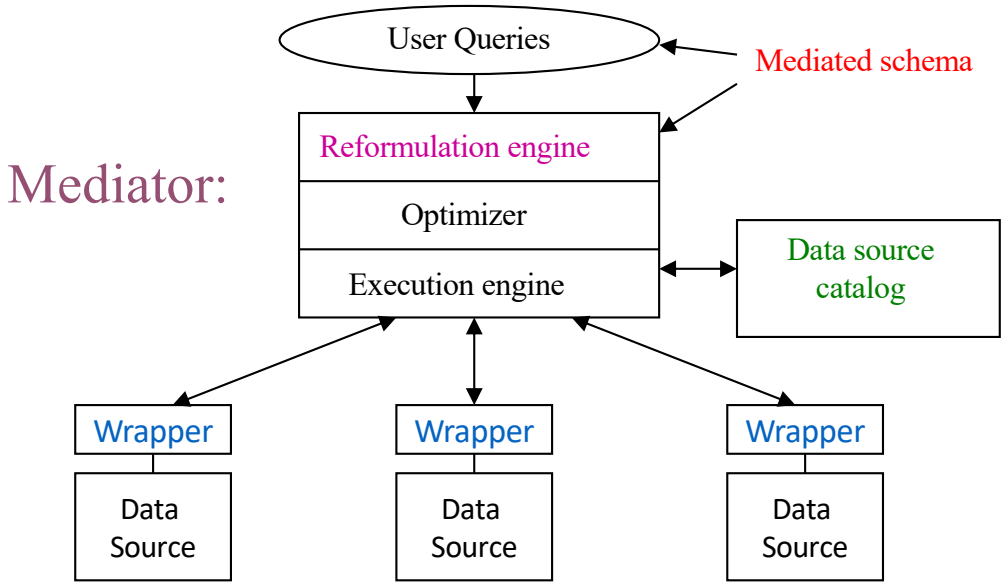
Virtual integration approach

Mediator:



Virtual integration: Architecture

- Leave the data in the data sources
- For every query over the mediated schema
  - Find the data sources that have the data (probably more than one)
  - Query the data sources
  - Combine results from different sources if necessary



Challenges

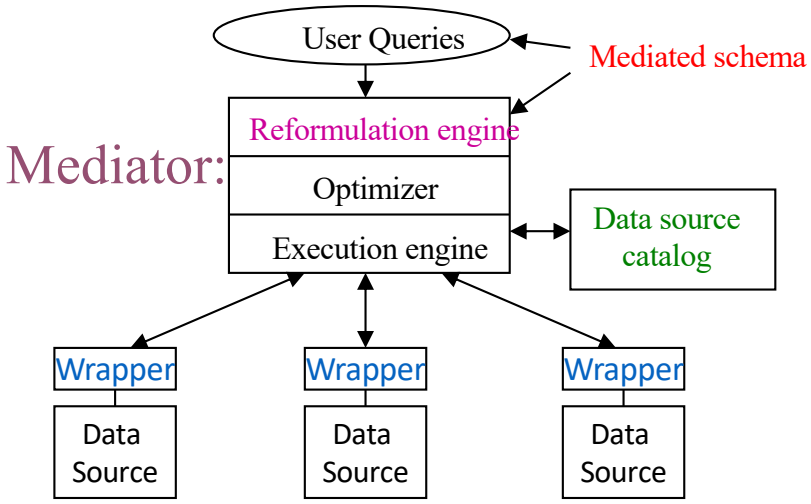
- Designing a single mediated schema
  - Data sources might have different schemas, and might export data in different formats
- Translation of queries over the mediated schema to queries over the source schemas
- Query Optimization
  - No/limited/stale statistics about data sources
  - Cost model to include network communication cost
  - Multiple data sources to choose from

Challenges (2)

- Query Execution
  - Network connections unreliable – inputs might stall, close, be delayed, be lost
  - Query results can be cached – what can be cached?
- Query Shipping
  - Some data sources can execute queries – send them sub-queries
  - Sources need to describe their query capability and also their cost models (for optimization)
- Incomplete data sources
  - Data at any source might be partial, overlap with others, or even conflict
  - Do we query all the data sources? Or just a few? How many? In what order?

Wrappers

- Sources export data in different formats
- Wrappers are custom-built programs that transform data from the source native format to something acceptable to the mediator



HTML

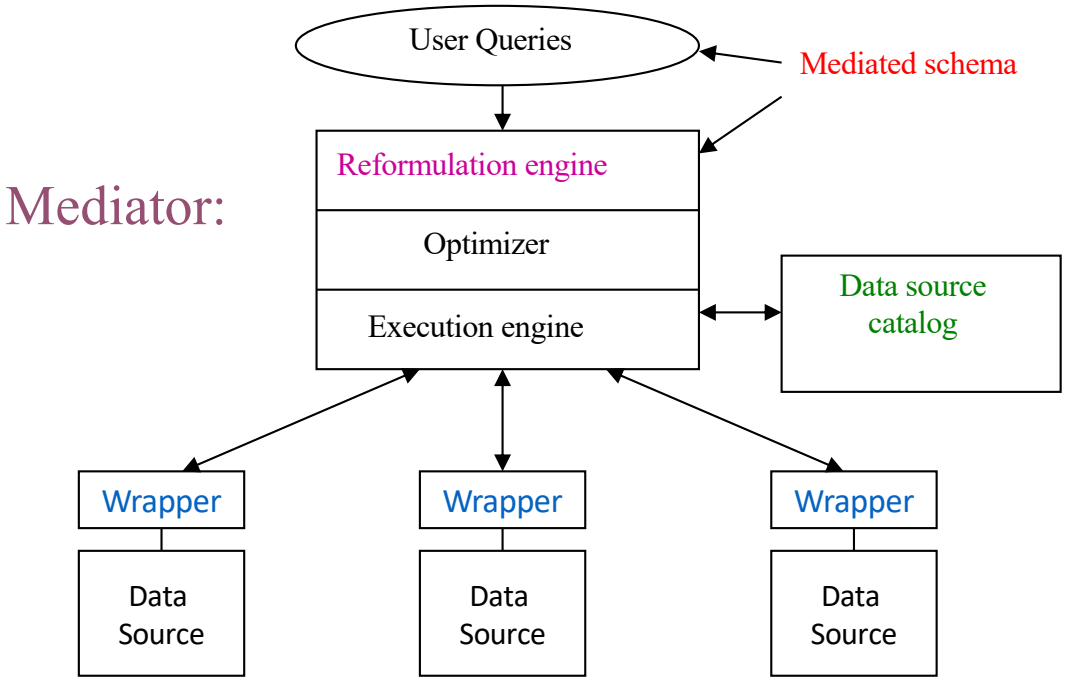
```
<b> Introduction to DB </b>
<i> Phil Bernstein </i>
<i> Eric Newcomer </i>
 Addison Wesley, 1999
```

XML

```
<book>
<title> Introduction to DB </title>
<author> Phil Bernstein </author>
<author> Eric Newcomer </author>
<publisher> Addison Wesley </publisher>
<year> 1999 </year>
</book>
```

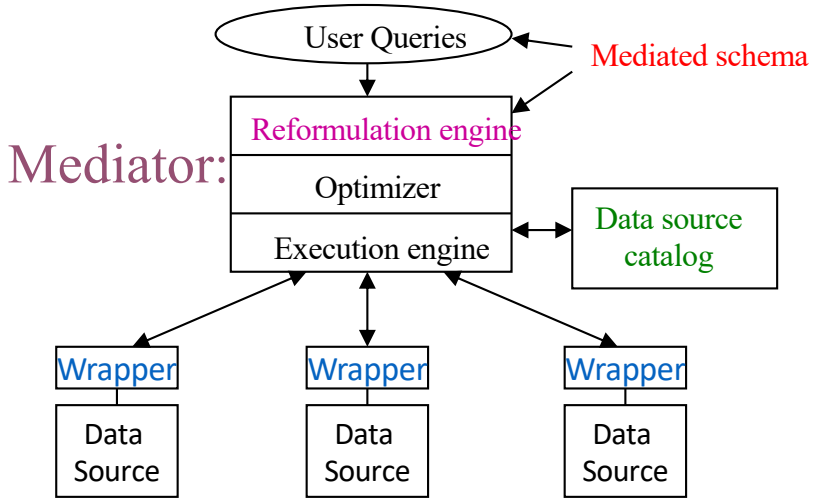
Wrappers(2)

- Can be placed either at the source or at the mediator
- Maintenance problems
  - have to change if source interface changes



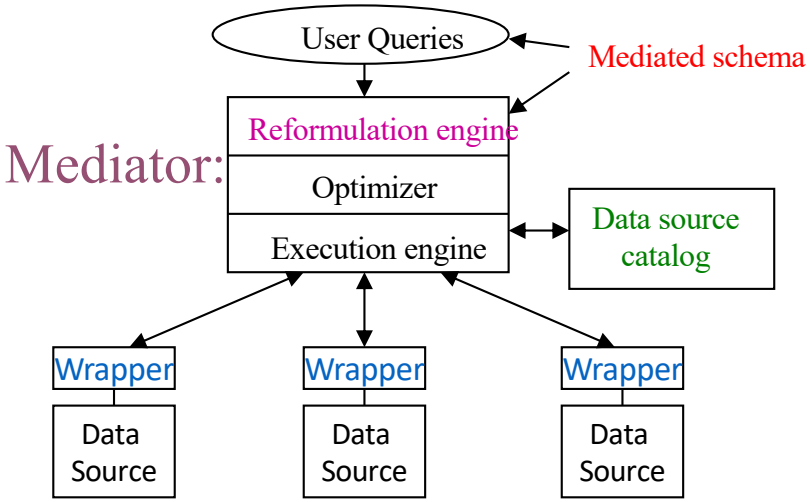
Data Source Catalog

- Contains meta-information about sources
  - Logical source contents (books, new cars)
  - Source capabilities (can answer SQL queries)
  - Source completeness (has all books)
  - Physical properties of source and network
  - Statistics about the data (like in an RDBMS)
  - Source reliability
  - Mirror sources
  - Update frequency



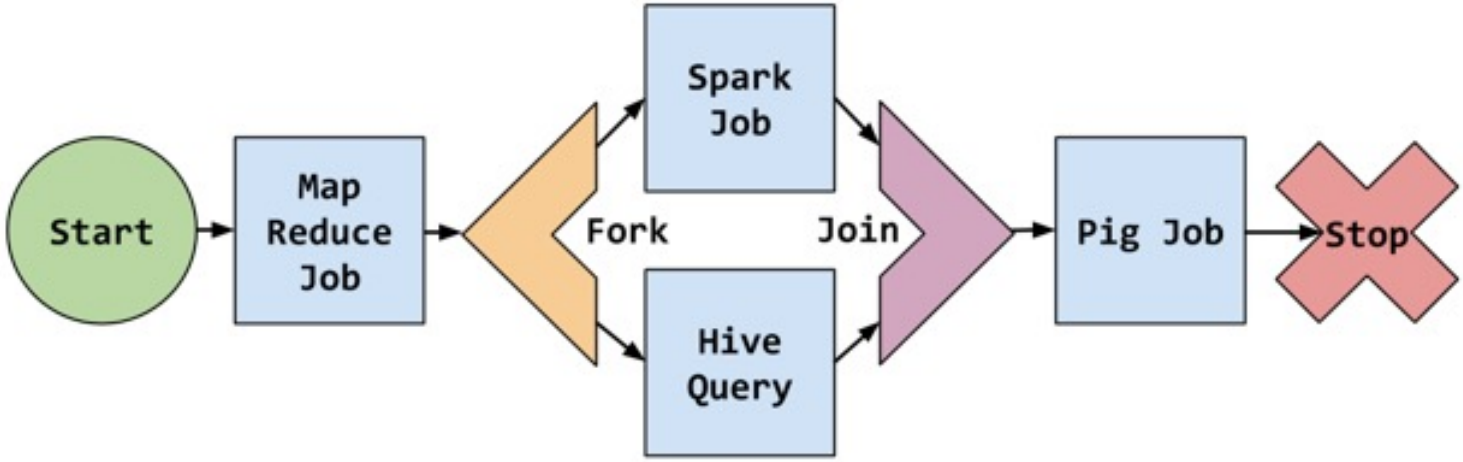
Schema mediation

- Users pose queries over the mediated schema
- The data at a source is visible to the mediator is its local schema
- Reformulation: Queries over the mediated schema have to be rewritten as queries over the source schemas



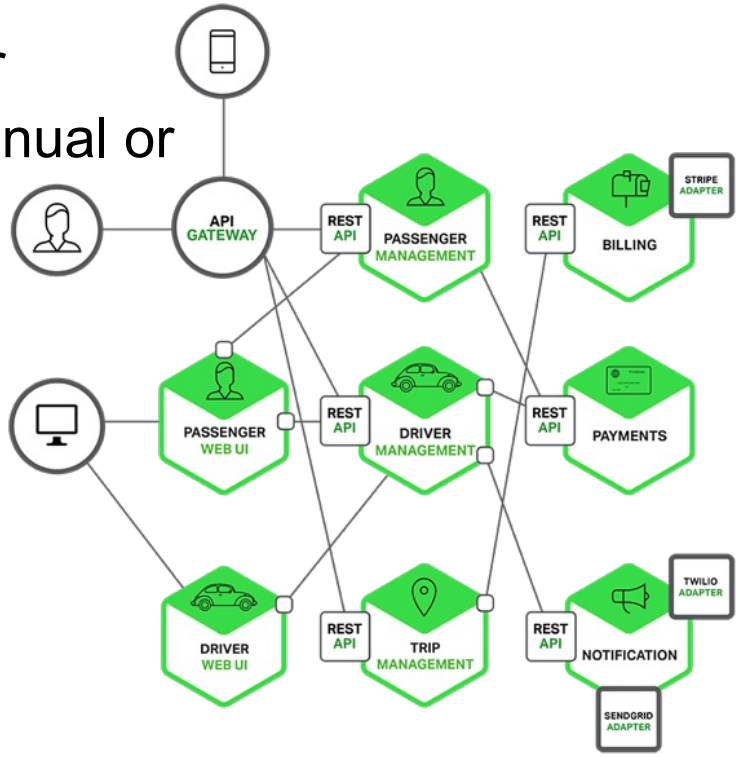
Workflow management approach

- Represent an integration-by-application approach
- Allow to implement business processes where each single step is executed by a different application or user
- Support modeling, execution, and maintenance of processes that are comprised of interactions between applications and human users



Web service approach

- Performs integration through software components (web services) that support machine-to-machine interaction by XML-based messages
- Depending on offered integration functionality either represent
  - a uniform data access approach, or
  - a common data access for later manual or
  - application-based integration



Apache Nifi

Introduction

What is dataflow?

What is NiFi?

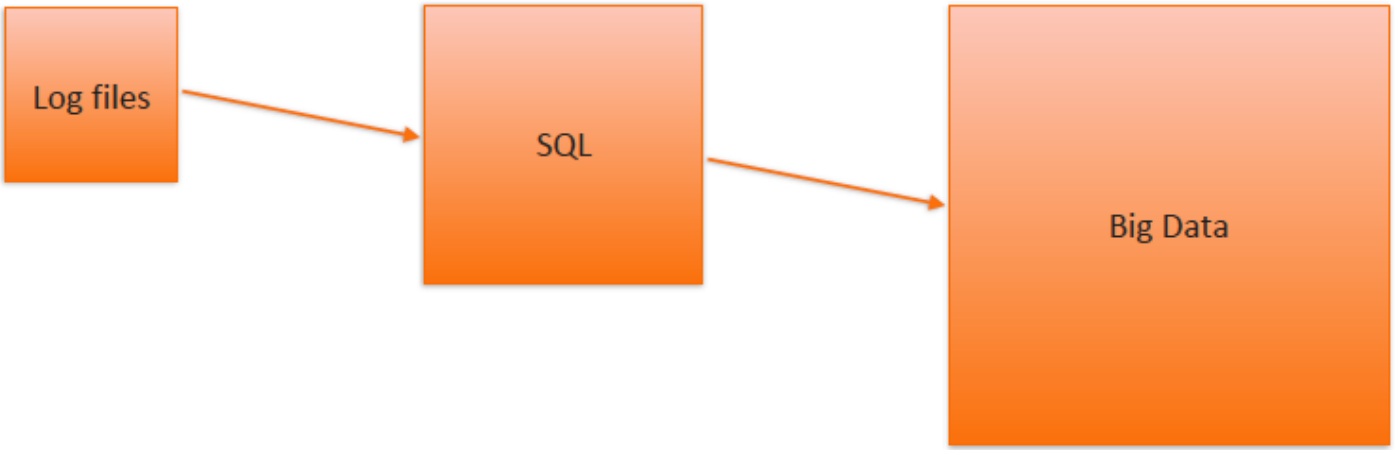
What's next?

What is dataflow?

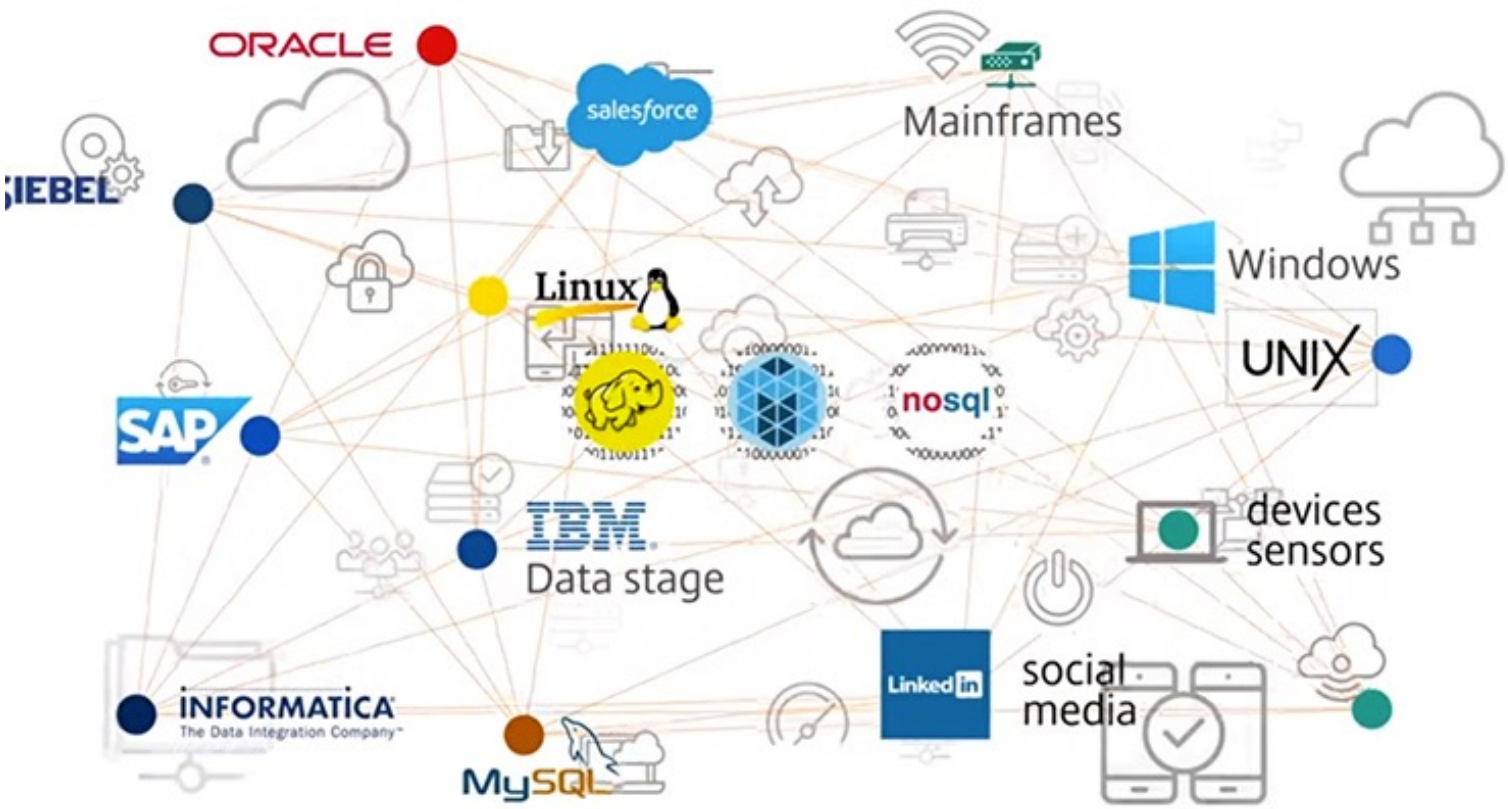
- Moving some content from A to B
- Content could be any bytes
  - Logs
  - HTTP
  - XML
  - CSV
  - Images
  - Video

Connecting data points is easy

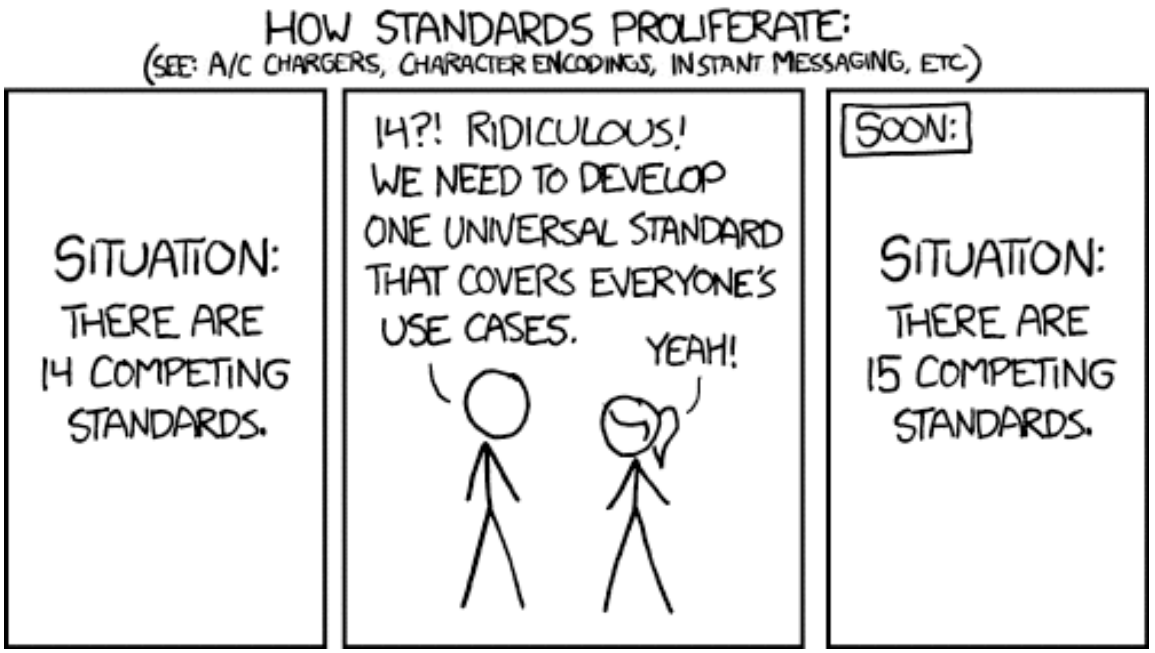
- Simple enough to write a process
  - Bash/Ruby/Python
  - SQL proc
  - etc.



This approach doesn't scale



Moving data effectively is hard



Standards: http://xkcd.com/927/

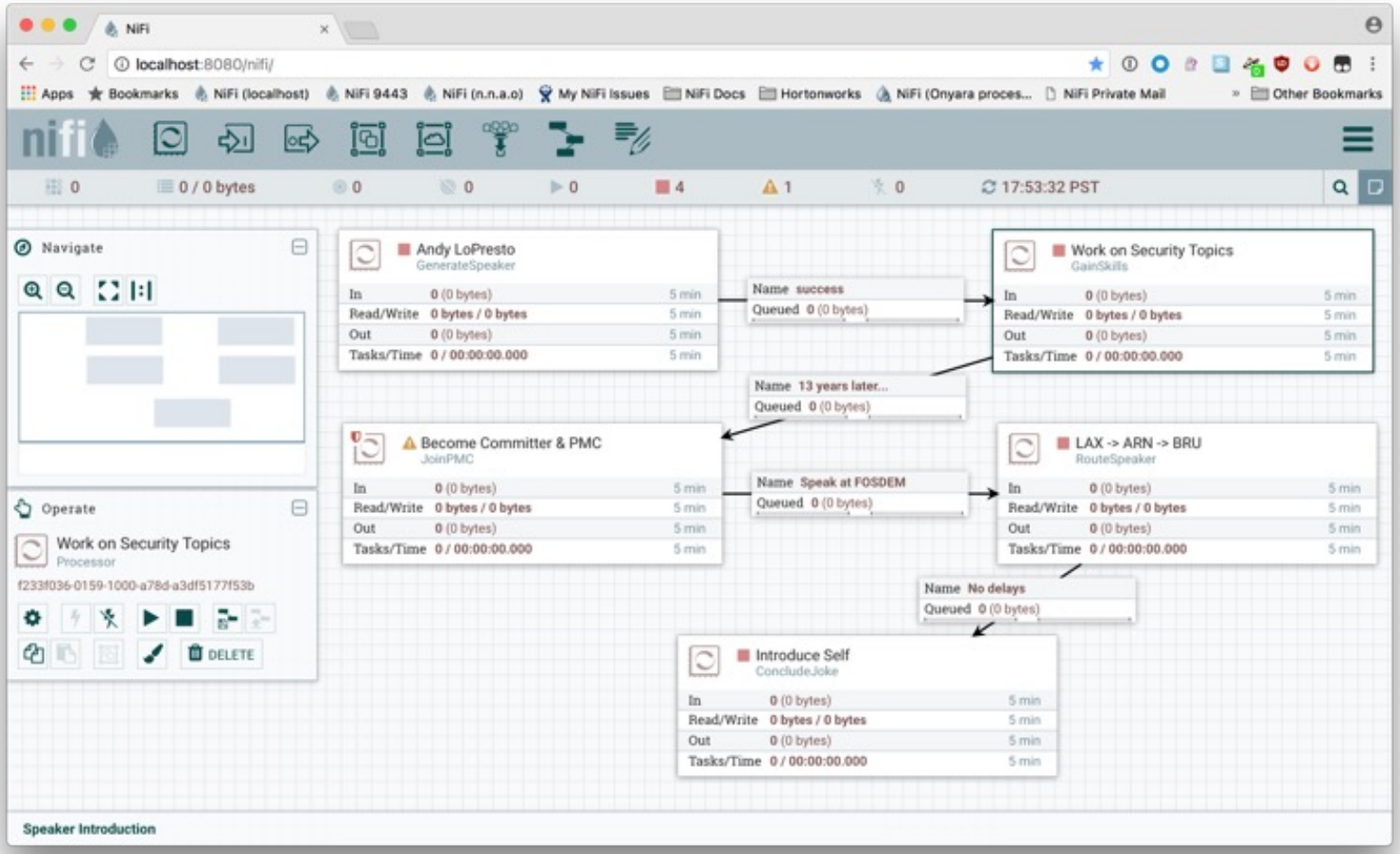
NiFi is based on Flow Based Programming (FBP)

FBP term	Niffi Term	Description
Information Packet	FlowFile	Each object moving through the system.
Black Box	FlowFile Processor	Performs the work, doing some combination of data routing, transformation, or mediation between systems.
Bounded Buffer	Connection	The linkage between processors, acting as queues and allowing various processes to interact at differing rates.
Scheduler	Flow Controller	Maintains the knowledge of how processes are connected, and manages the threads and allocations thereof which all processes use.
Subnet	Process Group	A set of processes and their connections, which can receive and send data via ports. A process group allows creation of entirely new component simply by composition of its components.

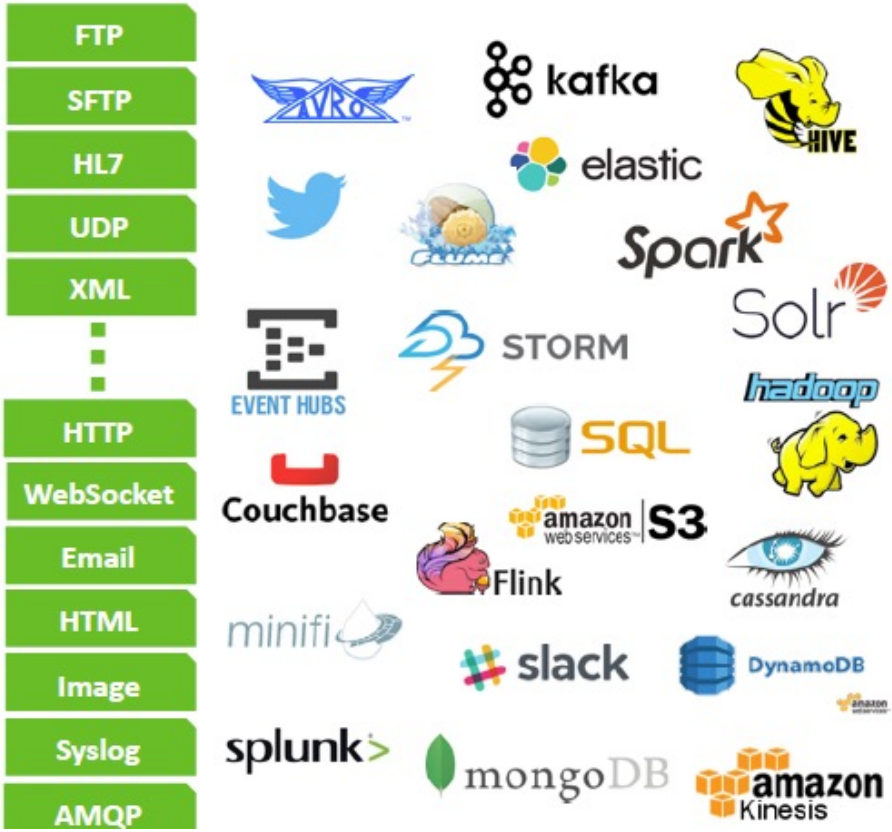
Nifi key features

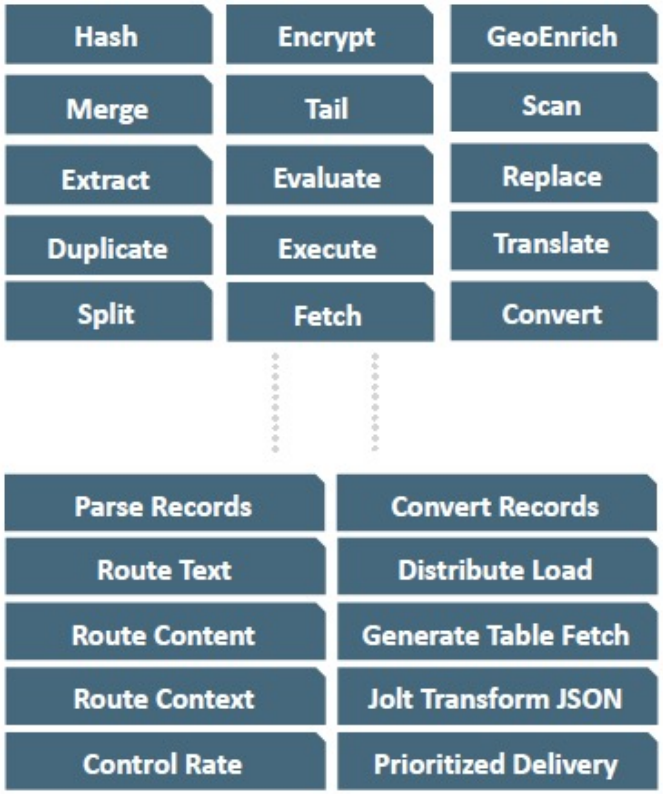
- Guaranteed delivery
- Prioritized queuing
- Flow specific QoS
  - Latency vs. throughput
  - Loss tolerance
- Data provenance
- Supports push and pull models
- Recovery/recording a rolling log of fine-grained history
- Visual command and control
- Flow templates
- Pluggable, multi-tenant security
- Designed for extension
- Clustering of Nifi instances

User Interface



Ecosystem Integration: 260+ Processors, 48 Controller Services





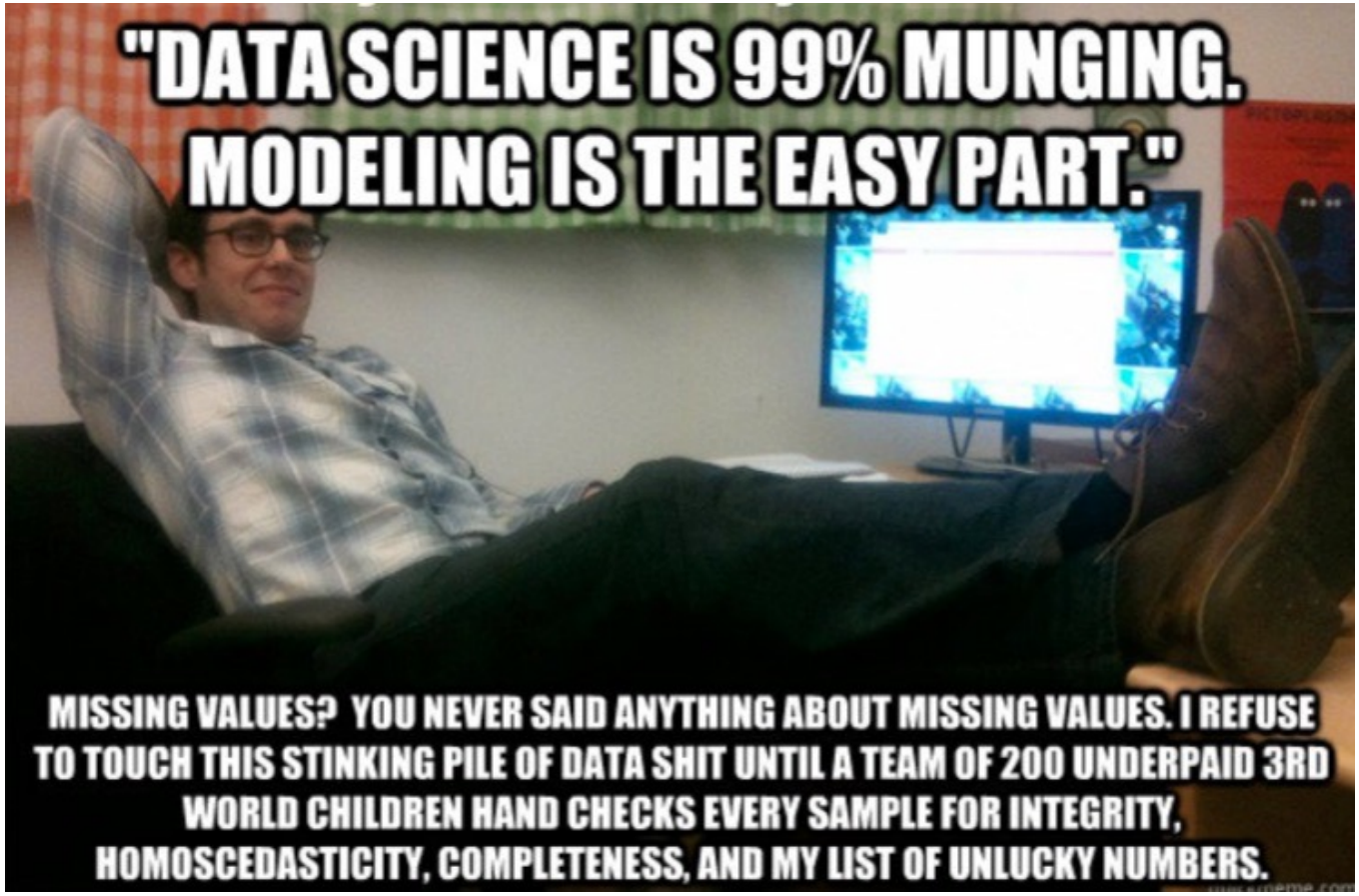
Data cleaning

Why data cleaning?

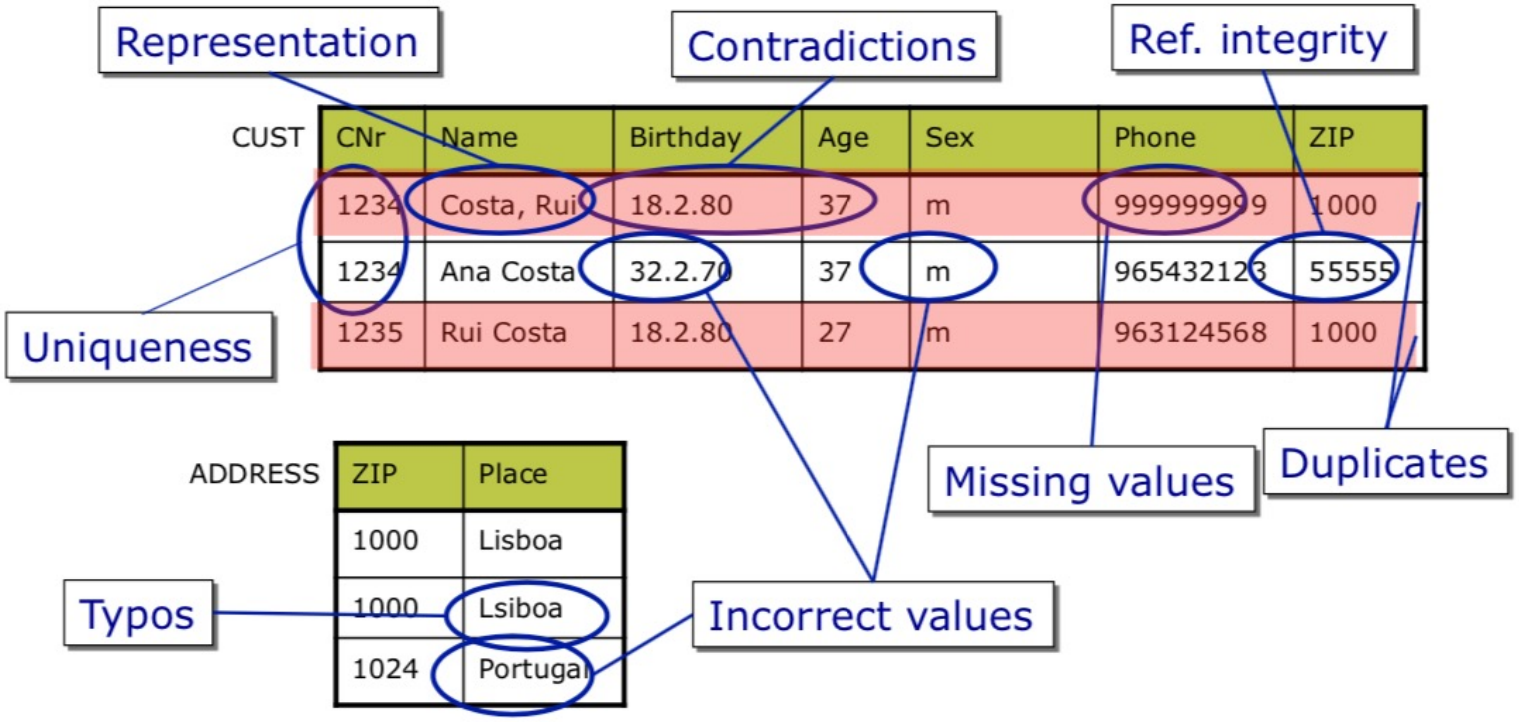
- Data in real world is dirty
- **Incomplete** (e.g name = “”)
  - Lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - Different considerations between the time when the data was collected and when it is analyzed
  - Human/hardware/software bugs
- **Noisy** (e.g. salary =‘-10k’)
  - Containing errors or outliers
  - Faulty data collection instruments
  - Human error at data entry
  - Error in data transmission
- **Inconsistent** (e.g., Age=“20” Birthday=“02/02/2000”)
  - Different data sources
  - Functional dependency violation
- Duplicate records also need data cleaning

Data preprocessing is costly

- Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse



Data quality problems

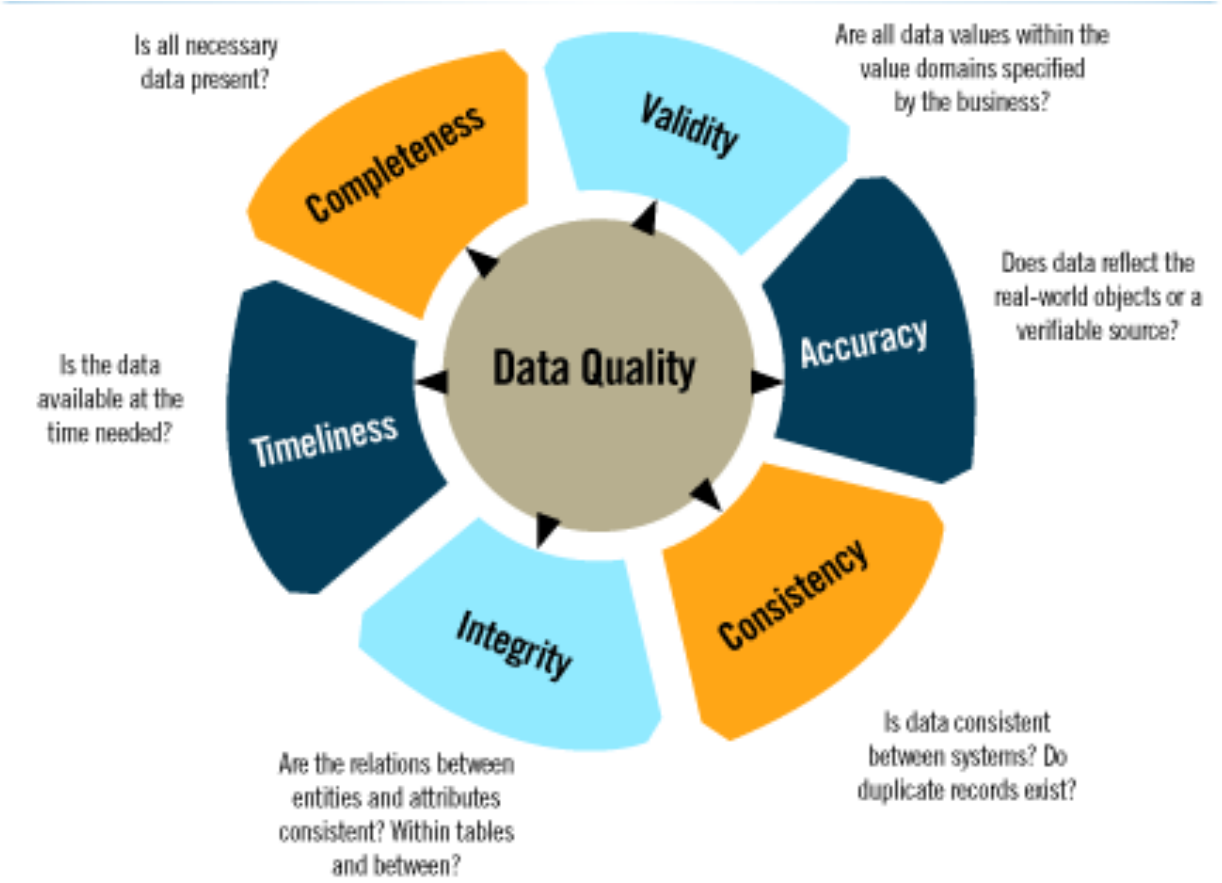


No quality data, no quality mining results

- Quality decisions must be based on quality data
  - e.g., duplicate or missing data may cause incorrect or even misleading statistics.

Data quality dimensions

- “Even though quality cannot be defined, you know what it is.” Robert Pirsig



Taxonomy of data quality problems

- Value-level
- Value-set (attribute/column) level
- Record level
- Relation level
- Multiple relations level

Value level

- Missing value: value not filled in a not null attribute
  - Ex:birthdate=‘’
- Syntax violation: value does not satisfy the syntax rule defined for the attribute
  - Ex:zipcode=27655-175;syntacticalrule:xxxx-xxx
- Spelling error
  - Ex:city=‘Lsboa’,insteadof‘Lisbon’
- Domain violation: value does not belong to the valid domain set
  - Ex:age=240;age:{0,120}

Value-set and record levels

- Value-set level
  - Existence of synonyms: attribute takes different values, but with the same meaning
    - Ex: emprego = ‘futebolista’; emprego = ‘jogador futebol’
  - Existence of homonyms: same word used with diff meanings
    - Ex: same name refers to different authors of a publication
  - Uniqueness violation: unique attribute takes the same value more than once
    - Ex: two clients have the same ID number
  - Integrity contraint violation
    - Ex: sum of the values of percent attribute is more than 100
- Record level
  - Integrity constraint violation
    - Ex: total price of a product is different from price plus taxes

Relation level

- Heterogeneous data representations: different ways of representing the same real world entity
  - Ex: name = ‘John Smith’; name = ‘Smith, John’
- Functional dependency violation
  - Ex: (2765-175, ‘Estoril’) and (2765-175, ‘Oeiras’)
- Existence of approximate duplicates
  - Ex: (1, André Fialho, 12634268) and (2, André Pereira Fialho, 12634268)!
- Integrity constraint violation
  - Ex: sum of salaries is superior to the max established

Multiple tables level

- Heterogeneous data representations
  - Ex: one table stores meters, another stores inches
- Existence of synonyms
- Existence of homonyms
- Different granularities: same real world entity represented with diff. granularity levels
  - Ex: age:{0-30,31-60,>60};age:{0-25,26-40, 40-65, >65}
- Referential integrity violation
- Existence of approximate duplicates
- Integrity constraint violation

What is data cleaning

- Tasks of Data cleaning
  - Fill in missing values
    - Value-level problem
  - Identify / correct noisy data
    - Value-level or record-level problem
  - Correct inconsistent data
    - Record-level problem, attribute-level problem, table-level problem, or multi-table level tables

Manage missing data

- Ignore the tuple: usually done when class label is missing (assuming the task is classification—not effective in certain cases)
- Fill in the missing value manually: tedious + infeasible?
- Use a global constant to fill in the missing value: e.g., “unknown”, a new class?!
- Use the attribute mean to fill in the missing value
- Use the attribute mean for all samples of the same class to fill in the missing value: smarter
- Use the most probable value to fill in the missing value: inference based such as regression, Bayesian formula, decision tree

Manage noisy data

- Binning method:
  - first sort data and partition into (equi-depth) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Clustering:
  - detect and remove outliers
- Regression
  - Use regression functions
- Semi automated
  - Computer and manual intervention

Manage inconsistent data

- Manual correction using external references
- Semi-automatic using various tools
  - To detect violation of known functional dependencies and data constraints
  - To correct redundant data

Methodology for data cleaning

- Extraction of the individual fields that are relevant
- Standardization of record fields
- Correction of data quality problems at value level
  - Missing values, syntax violation, etc
- Correction of data quality problems at value-set level and record level
  - Synonyms, homonyms, uniqueness violation, integrity constraint violation, etc
- Correction of data quality problems at relation level
  - Violation of functional dependencies, duplicate elimination, etc
- Correction of data quality problems problems at multiple relations level
  - Referential integrity violation, duplicate elimination, etc
- User feedback
  - To solve instances of data quality problems not addressed by automatic methods
- Effectiveness of the data cleaning and transformation process must be always measured for a sample of the data set

Data pre-processing

Data pre-processing

- Data smoothing: remove noise from data
- Data aggregation: summarization
  - E.g. Average, median, etc.
- Data generalization: concept hierarchy climbing
- Data normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Data reduction:
  - Diminish the size of the data
- Attribute (feature) engineering
  - Data wrapping

Data normalization

- Min-max normalization: transformation maps the values of a variable to a new range [NewMin, NewMax]

$$x_i' = \frac{x_i - OriginalMin}{OriginalMax - OriginalMin} \times (NewMax - NewMin) + NewMin$$

- Z-score normalization: (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

- Normalization by decimal scaling: ensures the range is between -1 and 1
  - with n the number of digits of the maximum absolute value:

$$x_i' = \frac{x_i}{10^n}$$

Data reduction

- Goal: obtain a reduced representation of the data
  - much smaller in volume yet producing the same (or almost the same) data analysis results

- Dimensionality reduction
  - Feature selection (*e.g.* genetic algorithm)
  - Feature extraction (*e.g.* PCA analysis)

- Data Compression
  - Convert text to numbers
  - Data clustering

- Discretization
  - Convert continuous data to categories
  - *E.g.* age=1 if in [0,12]; 2 if in [12-25], etc…

Technical solutions for data cleaning and pre-processing

Many tools for pre-processing

- Open Refine
- Trifacta Wrangler
- Python libraries

My personal suggestions (all free)

Open Refine and Trifacta Wrangler are more user-friendly

With Python libraries, you are more in control

- Tableau (and its free version Tableau Public)
- TabDrake TIBCO
- Clarity
- Winpure Data
- LadderData
- Cleaner
- Cloudingo
- Reifier
- IBM Infosphere Quality Stage

Demo on OpenRefine

- OpenRefine was formerly Google Refine
- It is a desktop application, not a web-service
  - Your sensitive data is (supposedly) safe
- Demo
  - https://youtu.be/B70J_H_zAWM
- More information on
  - https://guides.library.illinois.edu/openrefine

Practicals using Python libraries

- 5-days challenge (on the google Teams)
  - Day 1: Handling missing values
  - Day 2: Scaling and normalization
  - Day 3: Parsing dates
  - Day 4: Character encodings
  - Day 5: Inconsistent Data Entry

- Libraries used:
  - Pandas
  - Numpy
  - Chardet
  - Datetime
  - Fuzzywuzzy
  - Seaborn
  - Scipy
  - Mlxtend.preprocessing
  - Matplotlib.pyplot

Summary

- You have learnt about
  - Decision Support Systems and data warehouses
  - Data integration
  - Data cleaning
  - Data pre-processing
- You have had demos / tutorials on
  - Microsoft SQL Server
  - Apache Nifi
  - Openrefine
- You will have practicals using Python libraries

Homework

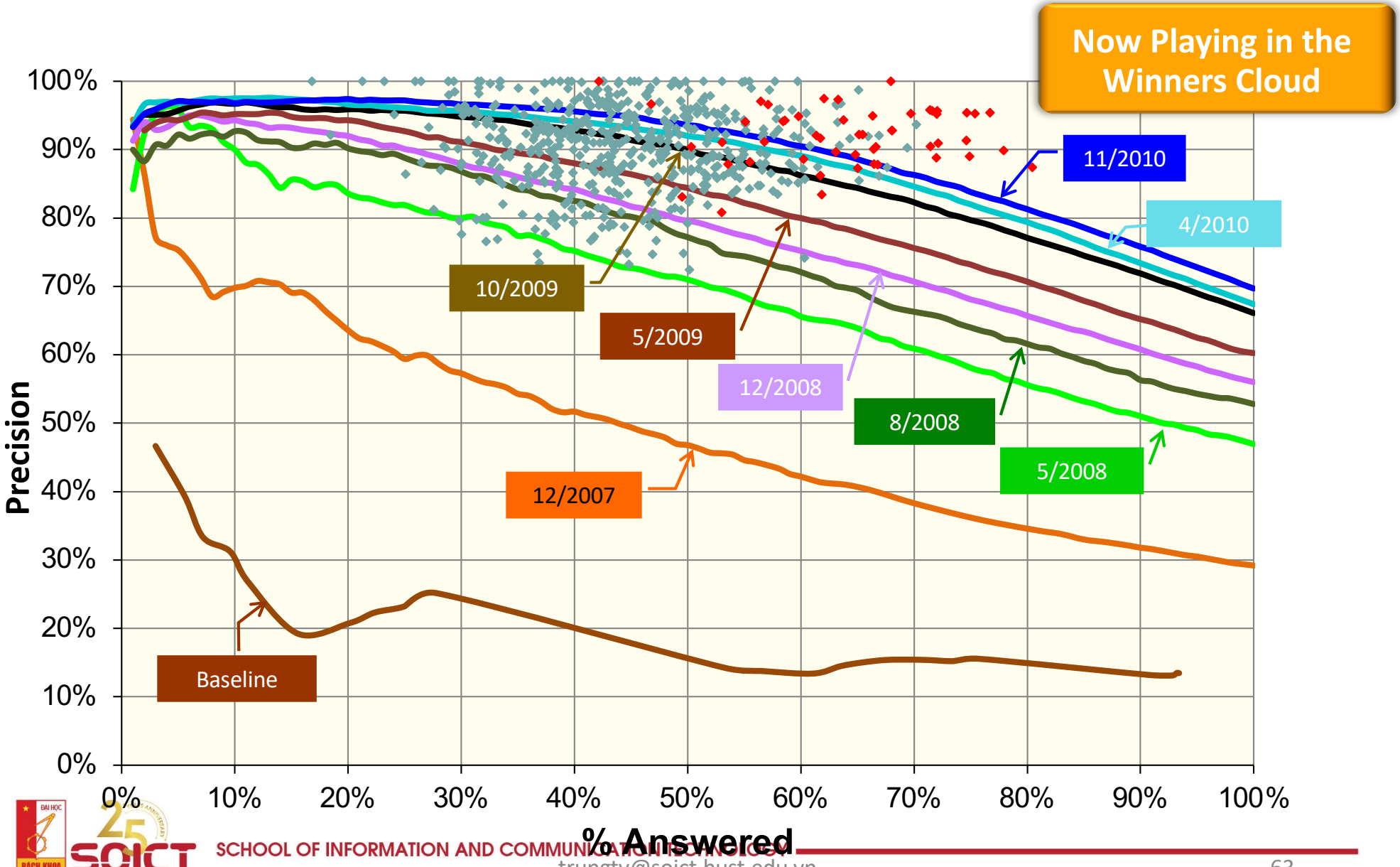
- Data integration / pre-processing
  - Access the demonstration website yourself and learn more about the NIFI flows:
  - https://youtu.be/MARazprrNYA
- Data cleaning
  - Learn some simple methods to remove outlier data
  - https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba
- Data pre-processing
  - Learn more about OpenRefine on
  - https://guides.library.illinois.edu/openrefine
- Practical / programming homework (5-day Challenge)
  - https://www.kaggle.com/rtatman/data-cleaning-challenge-handling-missing-values

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you for your attention!!!

soict.hust.edu.vn/ fb.com/groups/soict

DeepQA: Incremental Progress in Precision and Confidence 6/2007-11/2010



Appendix

Correlation analysis (numerical data)

- Correlation coefficient (also called Pearson's product moment coefficient)

$$r_{A,B} = \frac{\sum (A - \bar{A})(B - \bar{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum (AB) - n\bar{A}\bar{B}}{(n-1)\sigma_A \sigma_B}$$

- where n is the number of tuples, and are the respective means of A and B, σA and σB are the respective standard deviation of A and B, and Σ(AB) is the sum of the AB cross-product.
- If rA,B > 0, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.
- rA,B = 0: independent;
- rA,B < 0: negatively correlated

Correlation analysis (categorial data)

- X2 (chi-square) test

$$\chi^2 = \sum_i \sum_j \frac{(Observed - Expected)^2}{Expected}$$

- The larger the X2 value, the more likely the variables A, B are related (Observed is actual count of event (Ai,Bj))
- The cells that contribute the most to the X2 value are those whose actual count is very different from the expected count (based on totals)
- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

Chi-square calculation: An example

- X2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)
- It shows that like_science_fiction and play_chess are correlated in the group

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

	Play Chess	Don't play chess	Sum (Row)
Like science fiction	250(90)	200(360)	450
Don't Like science fiction	50(210)	1000(840)	1050
Sum (Column)	300	1200	1500