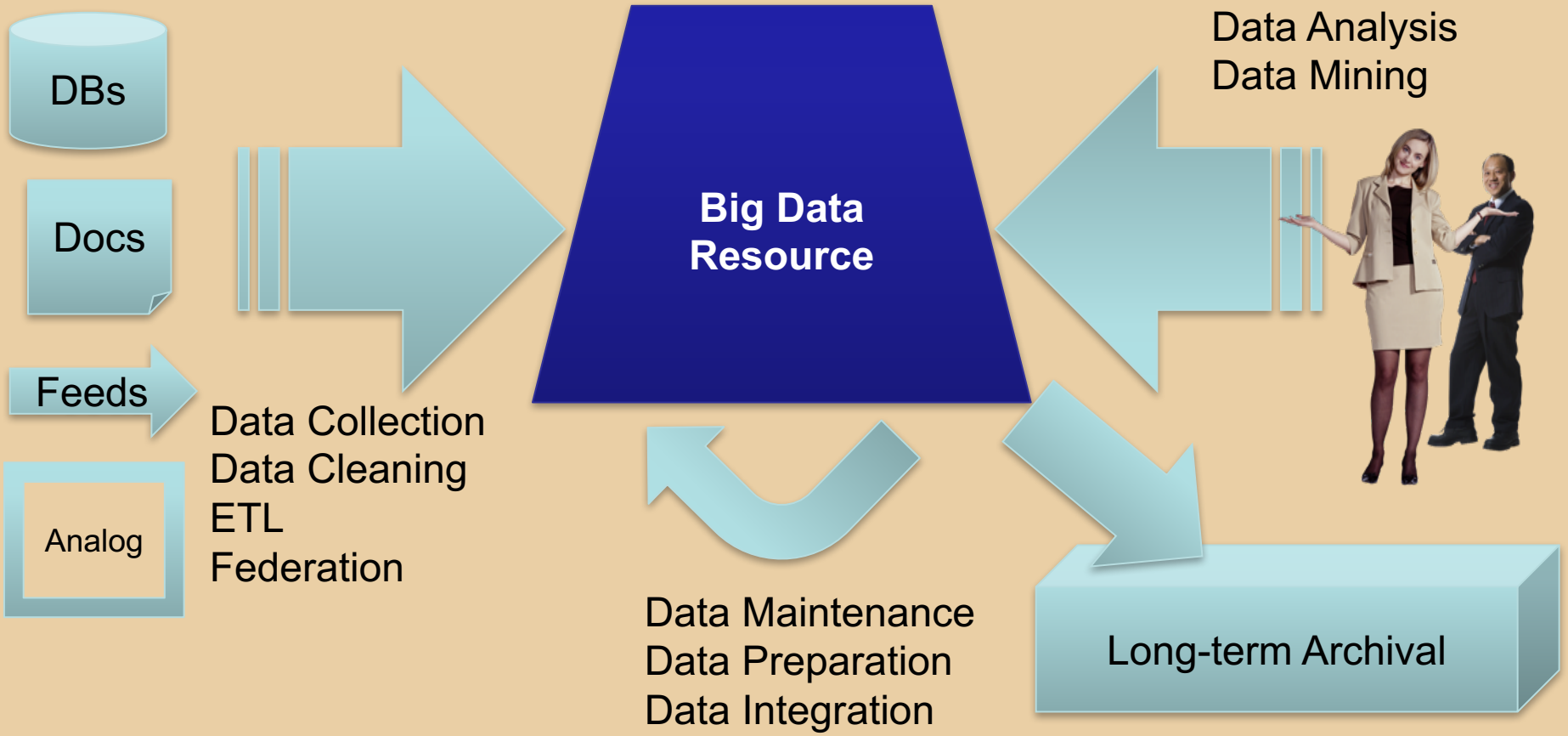
IT University of Copenhagen

Big Data Management
Introduction to Data Quality

Björn Þór Jónsson



Big data is not a product, but a collection of **processes**

Big Data + Data Quality

- Big data: all about the V's ☺
  - Size: huge **volume** of data from multiple sources
  - Speed: dynamic data, collected and analyzed at high **velocity**
  - Complexity: huge **variety** of data and sources
- Goal: to extract significant **value** from big data
- Key issue: data quality
  - Raw data is often of questionable **veracity**
  - How do we obtain high quality information?

Data Quality by the Numbers

Impact of poor data quality?

- Erroneous data costs US businesses $600 billion/year
- In DW projects, data cleaning takes 30-80% of time and budget
- Data quality tools market is growing at 16% annually, way over the 7% average for other IT segments

How much data is erroneous?

- Enterprise data error rates: average of 1-5%, some > 30%
- Only 1/3rd of XML Web documents with XSD/DTD are valid, 14% even lack well-formedness

IT University of Copenhagen

Barna Saha, Divesh Srivastava: Data Quality: The “Other” Face of Big Data, Tutorial @ ICDE 2014

Data Error Types
Validity

- Data-Type Constraints
- Range Constraints
- Mandatory Constraints (non-null)
- Unique Constraints (keys)
- Set-membership constraints (labels, codes)
- Foreign-key constraints
- Regular expression patterns

- Cross-field validation
  - Sum of fields must add up to 100%
  - Date of exit cannot be earlier than date of entry
- Similar concerns may apply across records!
  - Many records should add up to 100%
  - One records contains an (incorrect) count of many records (denormalised)

Data Error Types
Accuracy

Id	Title	Director	Year	#Remakes	LastRemakeYear
1	Casablanca	Weir	1942	3	1940
2	Dead poets society	Curtiz	1989	0	NULL
3	Rman Holiday	Wylder	1953	0	NULL
4	Sabrina	null	1964	0	1985

Fig. 1.1. A relation `Movies` with data quality problems

Data Error Types
Completeness

Id	Title	Director	Year	#Remakes	LastRemakeYear
1	Casablanca	Weir	1942	3	1940
2	Dead poets society	Curtiz	1989	0	NULL
3	Rman Holiday	Wylder	1953	0	NULL
4	Sabrina	null	1964	0	1985

Fig. 1.1. A relation `Movies` with data quality problems

Data Error Types
Currency

Id	Title	Director	Year	#Remakes	LastRemakeYear
1	Casablanca	Weir	1942	3	1940
2	Dead poets society	Curtiz	1989	0	NULL
3	Rman Holiday	Wylder	1953	0	NULL
4	Sabrina	null	1964	0	1985

Fig. 1.1. A relation `Movies` with data quality problems

Data Error Types
Consistency

Id	Title	Director	Year	#Remakes	LastRemakeYear
1	Casablanca	Weir	1942	3	1940
2	Dead poets society	Curtiz	1989	0	NULL
3	Rman Holiday	Wylder	1953	0	NULL
4	Sabrina	null	1964	0	1985

Fig. 1.1. A relation `Movies` with data quality problems

Detect Data Errors?

Id	Title	Director	Year	#Remakes	LastRemakeYear
1	Casablanca	Weir	1942	3	1940
2	Dead poets society	Curtiz	1989	0	NULL
3	Rman Holiday	Wylder	1953	0	NULL
4	Sabrina	null	1964	0	1985

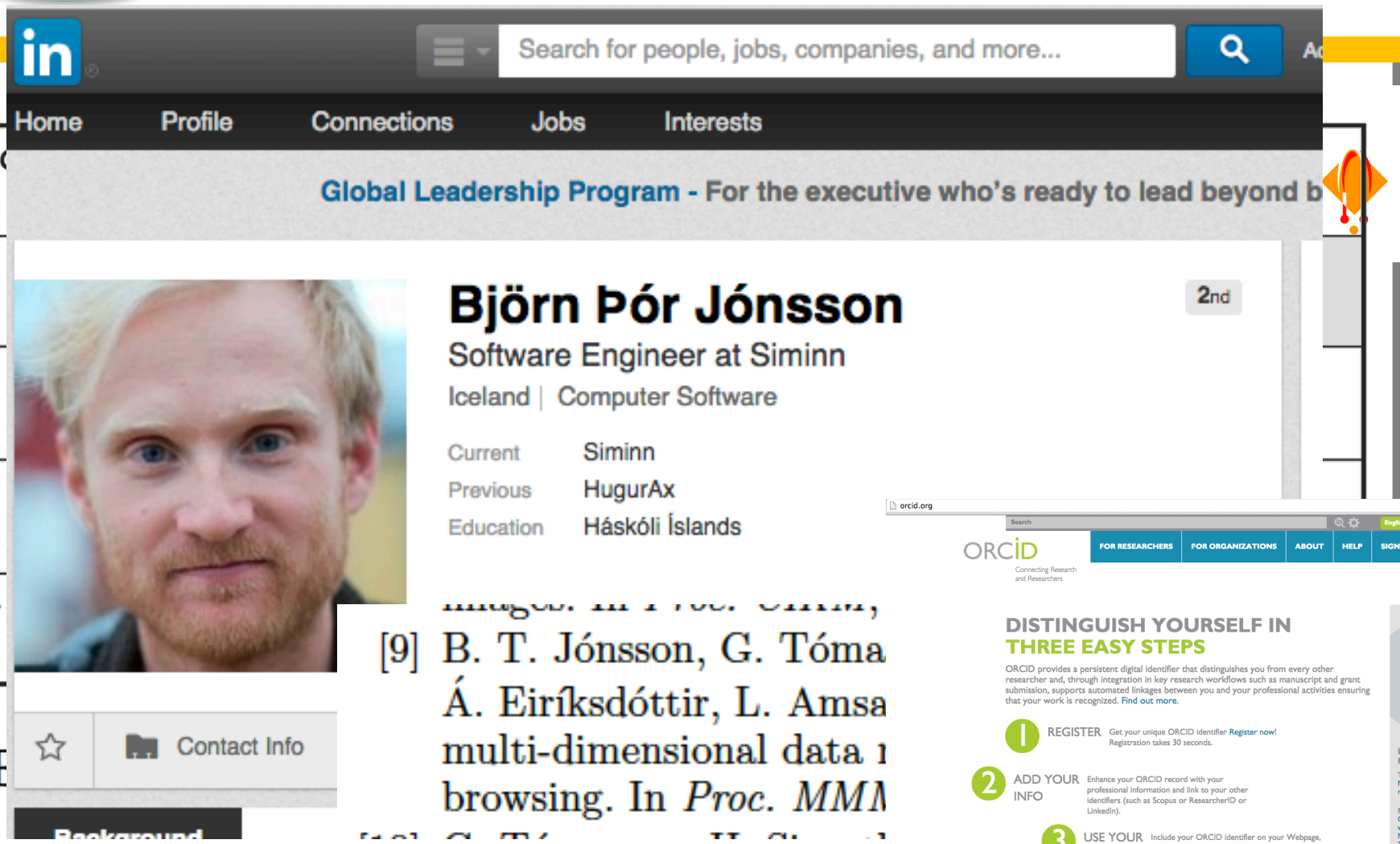
Fig. 1.1. A relation `Movies` with data quality problems

Correct Data Errors?

Id	Title	Director	Year	#Remakes	LastRemakeYear
1	Casablanca	Weir	1942	3	1940
2	Dead poets society	Curtiz	1989	0	NULL
3	Rman Holiday	Wylder	1953	0	NULL
4	Sabrina	null	1964	0	1985

Fig. 1.1. A relation `Movies` with data quality problems

Avoid Data Errors?



How to Achieve Data Quality?

Small Data

Specify all domain knowledge as integrity constraints on data

- Reject updates that do not preserve integrity constraints
- Works well when the domain is well understood and static

Big Data

Integrity constraints cannot be specified a priori

- Data diversity → complete domain knowledge is infeasible
- Data evolution → domain knowledge quickly becomes obsolete

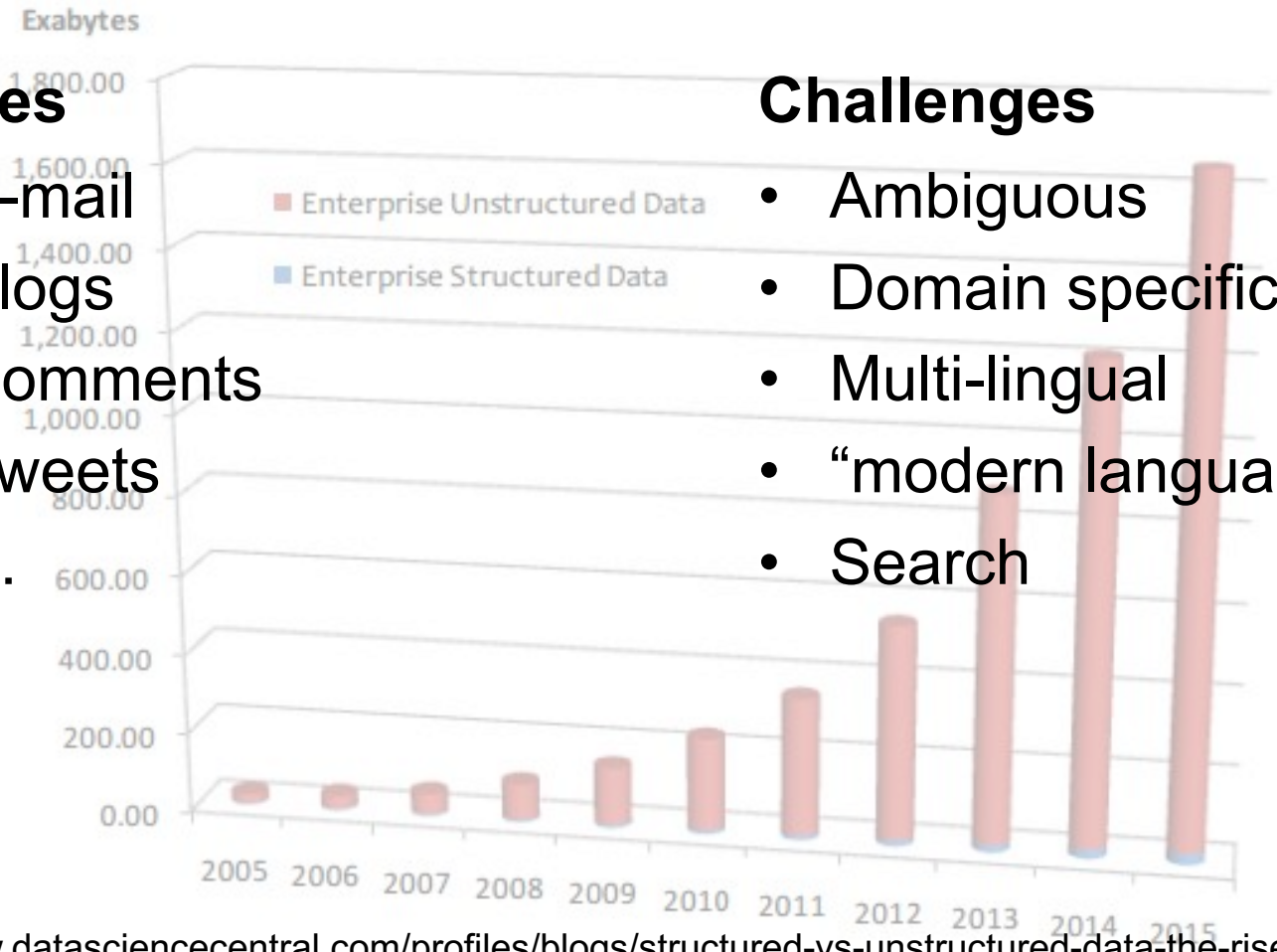
Data Quality: Lessons Learned

- Big data has considerable inconsistency
  - Even in domains where poor quality data can have big impact
  - Semantics ambiguity, out of date data, unexplainable errors
- Data sources often copy from each other
  - Copying can happen on erroneous data, spreading poor quality data

Unstructured Data?

Types

- E-mail
- Blogs
- Comments
- Tweets
- …



Challenges

- Ambiguous
- Domain specific
- Multi-lingual
- “modern language”
- Search

http://www.datasciencecentral.com/profiles/blogs/structured-vs-unstructured-data-the-rise-of-data-anarchy

Methods?

- Multi-lingual, ambiguous:
  - Text analysis, context
  - Translation
- Domain specific:
  - Autocoding – ontologies and classifications
    - e.g., disease ontologies
    - street names (st → street, ave → avenue, ...)
- Search: Term extraction → indexing

Data Correction Process

- Data auditing
- Workflow specification
- Workflow execution
- Post-processing and controlling

Data Quality vs Lambda Architecture

Data Cleaning vs Immutability

Storage Layer

- Apply corrective method and store result
- Pro: Can manually correct errors
- Con: WYSIWYGSW

Batch Layer

- Apply corrective methods again and again
- Pro: Better method → better results
- Con: Automated methods may not handle all errors

Take Away

- Data is often incorrect
  - Errors can be costly
- Various types of errors
  - Validity, Accuracy, Consistency, ...
- Need commitment and methodology to fix
  - On-the-fly vs permanent corrections