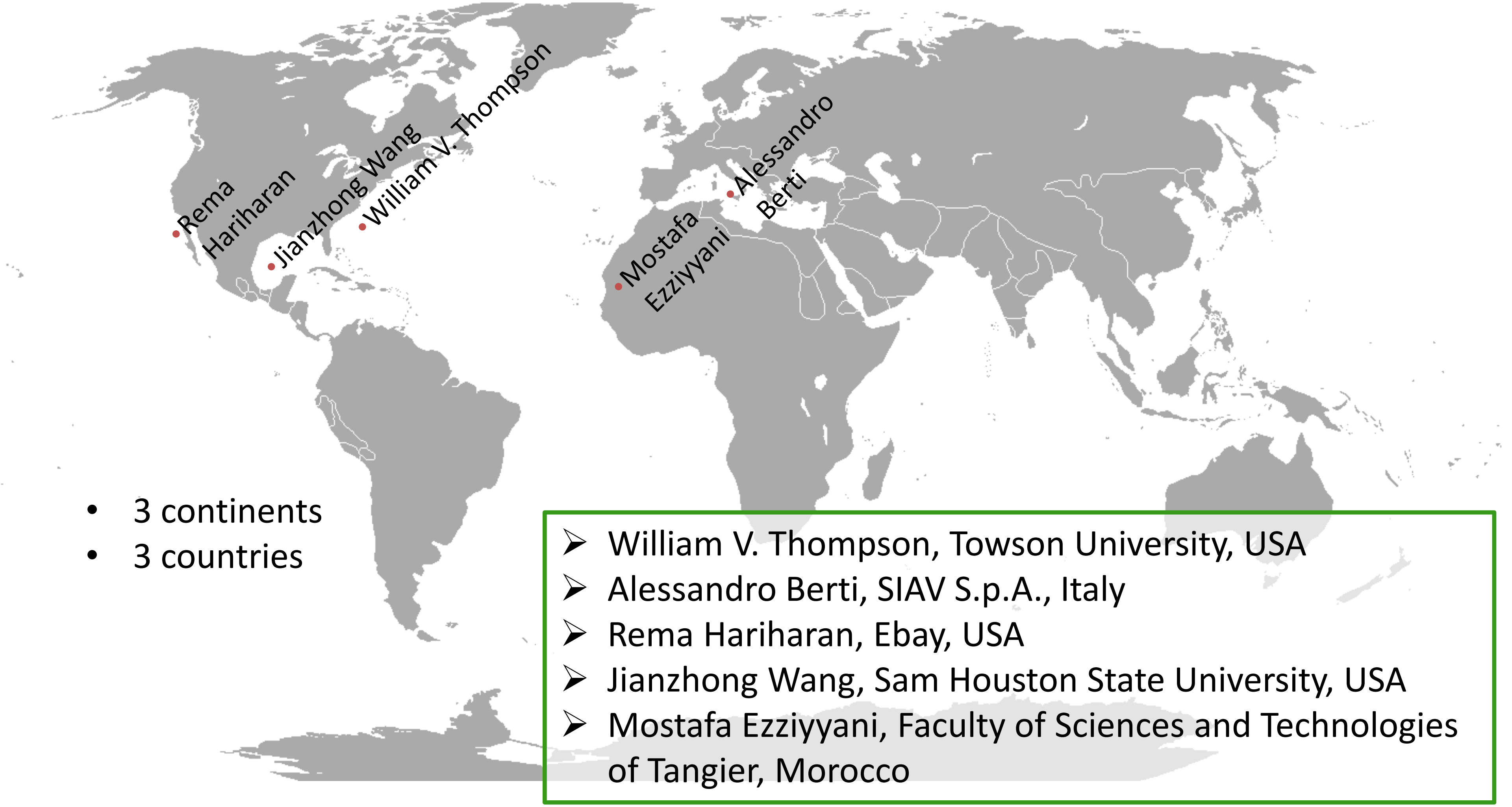
Panel on ICIW /IMMM / DATASETS

Trustable Computation in SQL and NoSQL Data Models

Dirk Labudde

Sonntag, 5. Juni 2016

The Penalists



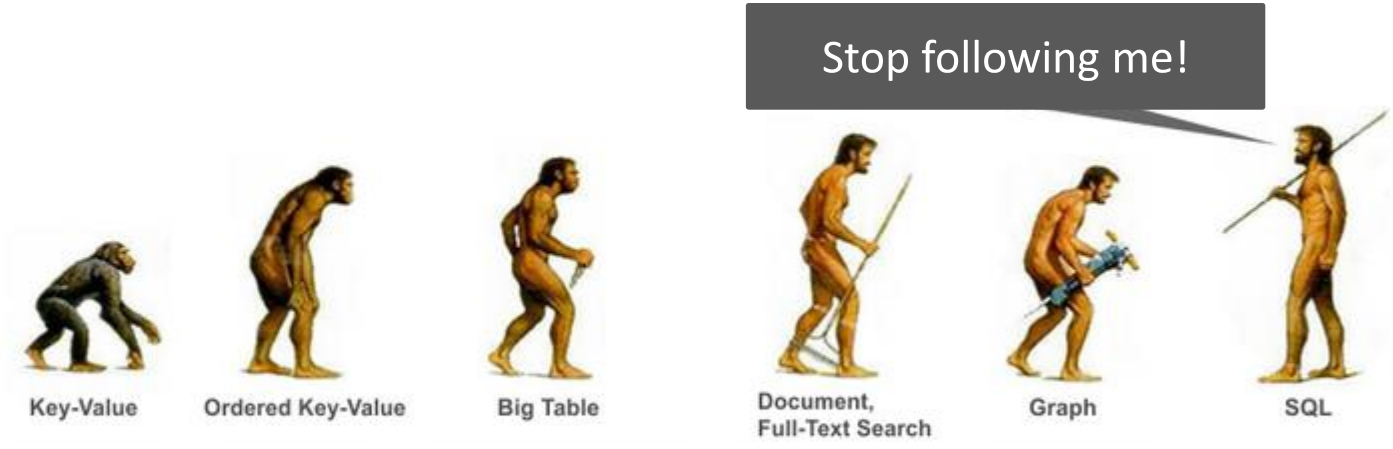
- 3 continents
- 3 countries

- ➢ William V. Thompson, Towson University, USA
- ➢ Alessandro Berti, SIAV S.p.A., Italy
- ➢ Rema Hariharan, Ebay, USA
- ➢ Jianzhong Wang, Sam Houston State University, USA
- ➢ Mostafa Ezziyyani, Faculty of Sciences and Technologies of Tangier, Morocco

Welcome to Spain – to Valencia – to the panel

SQL and/or NoSQL

Trustable Computation in SQL and NoSQL Data Models



structured storage

2009 by Johan Oskarsson

Non relational Approach

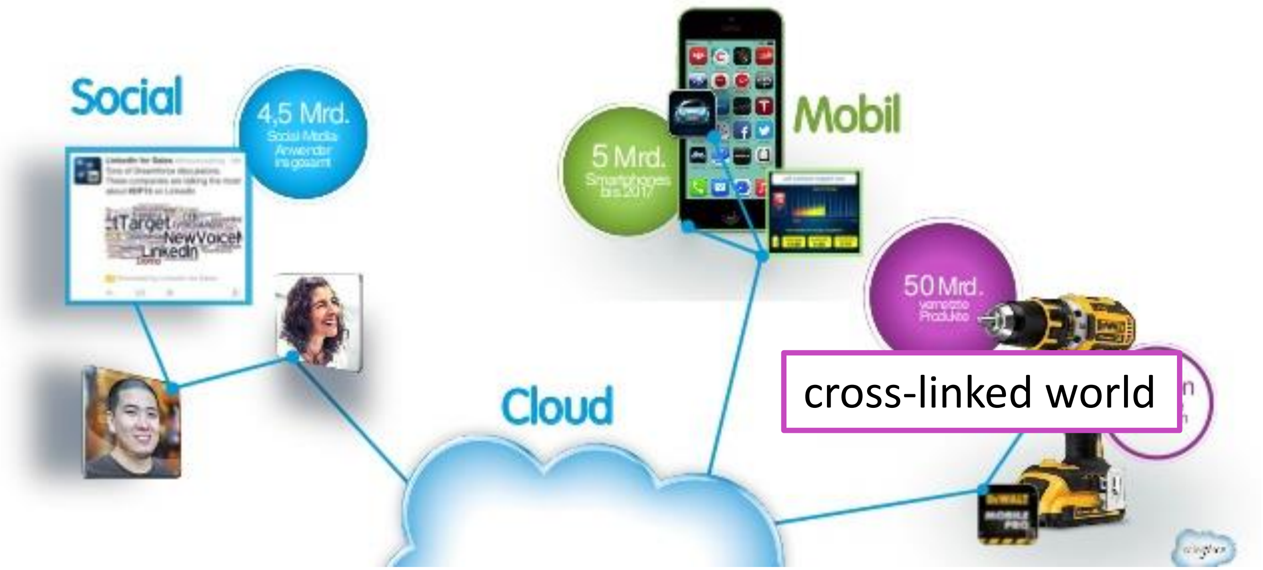
Not only SQL

🚫 relational Approach

Reason for the paradigm change?

Is that a paradigm change?

Difference between Faith and Trust?



cross-linked world

Reason for the paradigm change?

cross-linked world

The modern World in
your smartphone

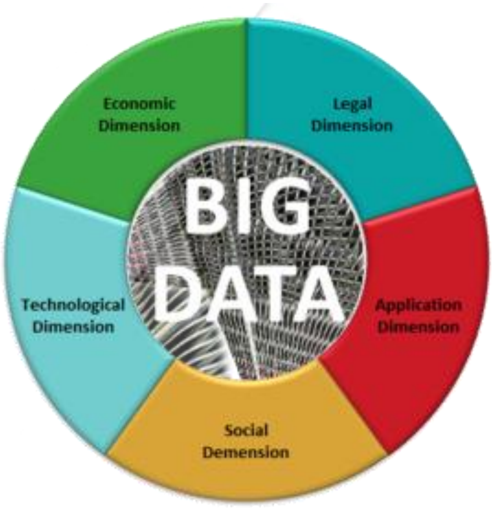
BigData

- ➢ New organization of data?
- ➢ New data modelling ?
- ➢ New data models ?

BigData

BigData Definition **- Recognition**, **Storages and Analysis of mounds of differently structured data** by specific methods.

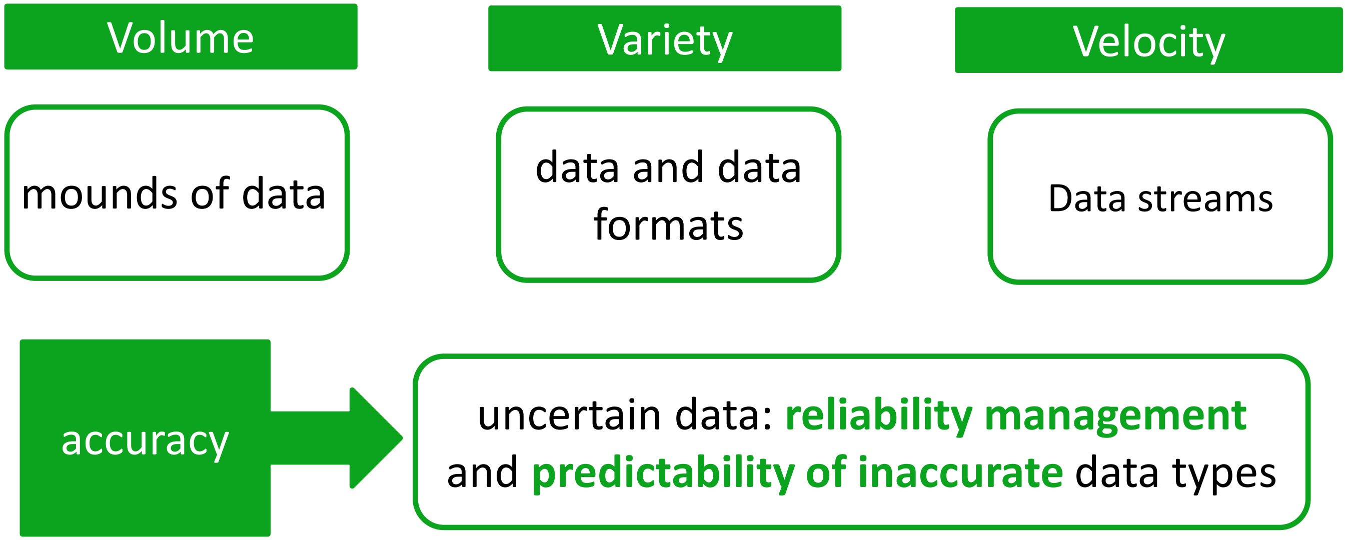
V3-equiation:
Components of Big Data Volume, Variety und Velocity



5 levels of bigdata

BigData

4 Dimensions of BigData



Volume

mounds of data

Variety

data and data formats

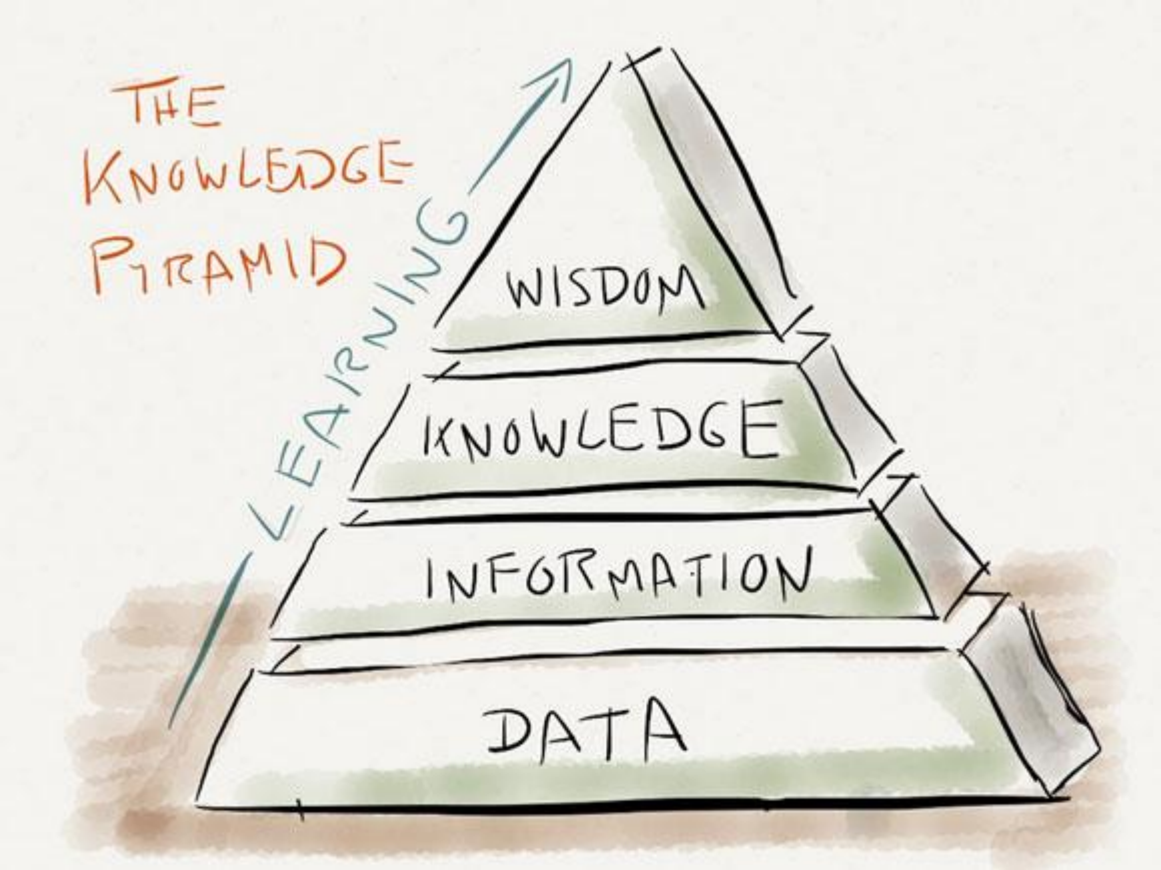
Velocity

Data streams

accuracy

uncertain data: **reliability management** and **predictability of inaccurate** data types

Data … Information … Knowledge



System understanding

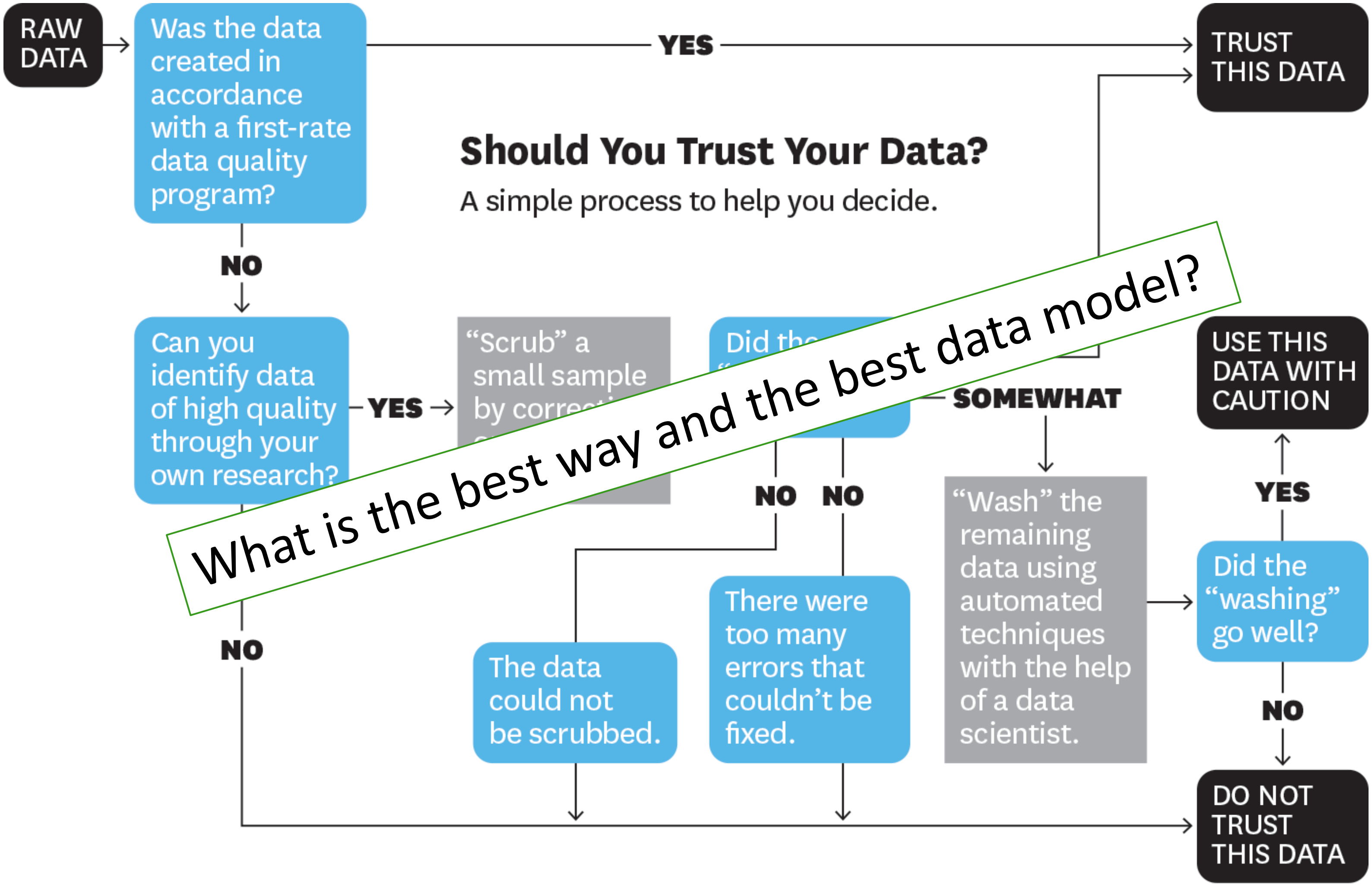
Models Modeling

relationship

Data, counts, numbers
the lowdown

Modeling Models <–> real System

Faith or Trust



RAW DATA

Was the data created in accordance with a first-rate data quality program?

YES

TRUST THIS DATA

Should You Trust Your Data?

A simple process to help you decide.

NO

Can you identify data of high quality through your own research?

YES

“Scrub” a small sample by correcti

Did th

SOMEWHAT

USE THIS DATA WITH CAUTION

NO

NO

“Wash” the remaining data using automated techniques with the help of a data scientist.

YES

Did the “washing” go well?

NO

The data could not be scrubbed.

There were too many errors that couldn’t be fixed.

DO NOT TRUST THIS DATA

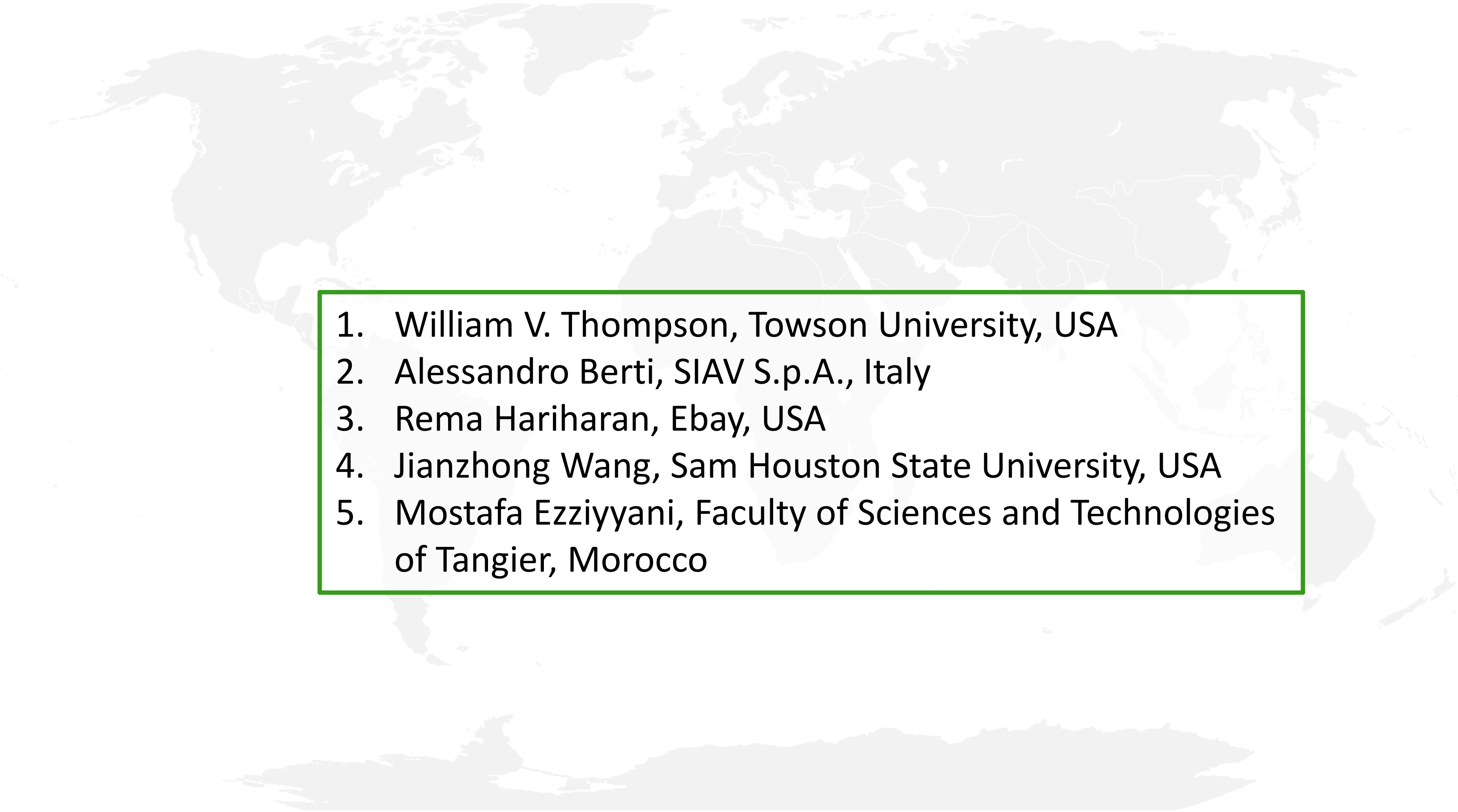
What is the best way and the best data model?

SOURCE THOMAS C. REDMAN

https://hbr.org/2015/10/can-your-data-be-trusted

© HBR.ORG

The Penalists



1. William V. Thompson, Towson University, USA
2. Alessandro Berti, SIAV S.p.A., Italy
3. Rema Hariharan, Ebay, USA
4. Jianzhong Wang, Sam Houston State University, USA
5. Mostafa Ezziyyani, Faculty of Sciences and Technologies of Tangier, Morocco

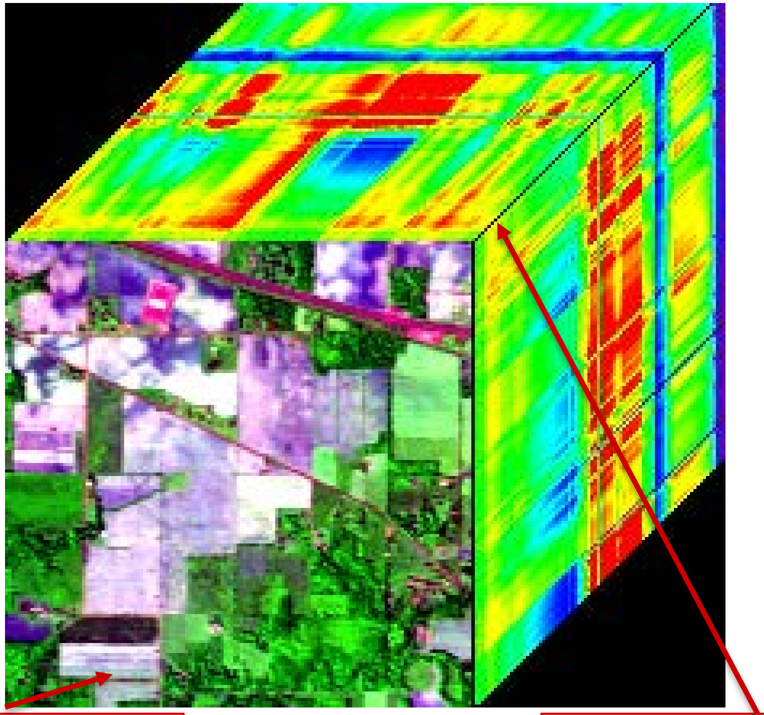
Computation and Processing on Image and Hyperspectral Image Data

Jianzhong Wang

Sam Houston State University

Examples of Image Data Basis

Face Images
(from UMIST
Face Database)



Band image

HIS cube

Spectral curve

Typical pre-processing

- Data Standardization
- Noise Removal.
- Dimensionality Reduction.
- Sparse representation
- Data Compression

Main Tasks

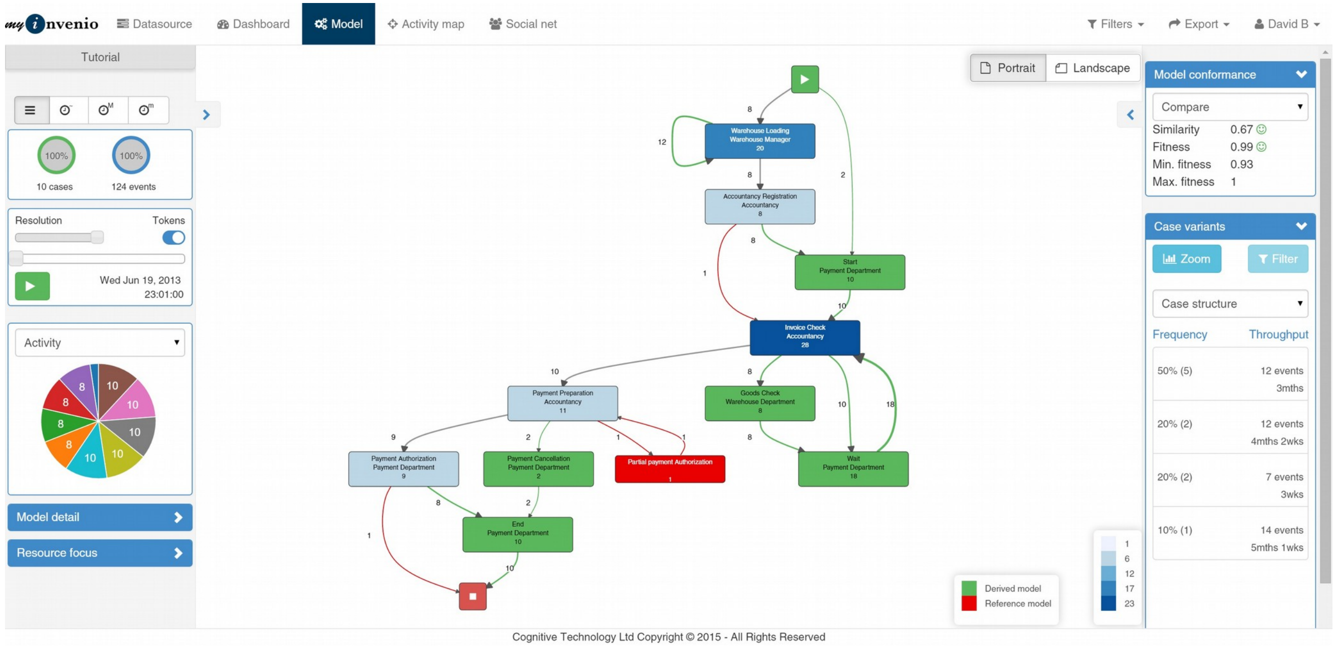
- Classification
  1. Machine learning
  2. Multiple classifier systems
- Feature Extraction
  1. Dimensionality reduction
  2. Machine learning
  3. Data tree and decision tree
  4. Harmonic Geometry: Data driven wavelets
- Target Detection
- Abnormal Detection

Trustable Computation in a Process Mining Conformance Checking framework

Alessandro Berti

Conformance Checking in Process Mining

- Checking pre-defined rules (about Control Flow and paths attributes) in process executions.
- The rules about Control Flow can be automatically discovered by Process Discovery techniques.



Wed Jun 19, 2013
23:01:00

Compare

Similarity	0.67 ☺
Fitness	0.99 ☺
Min. fitness	0.93
Max. fitness	1

Case structure

Frequency	Throughput
50% (5)	12 events 3mths
20% (2)	12 events 4mths 2wks
20% (2)	7 events 3wks
10% (1)	14 events 5mths 1wks

Cognitive Technology Ltd Copyright © 2015 - All Rights Reserved

NoSQL databases for Conformance Checking

- In NoSQL you can do fast documents storing and retrieval.
- Process Mining events and traces saved as documents in a NoSQL database.
- Indexing lets fast counts on paths, originators.
- Also, fast checking of rules can be done.

Apache Solr

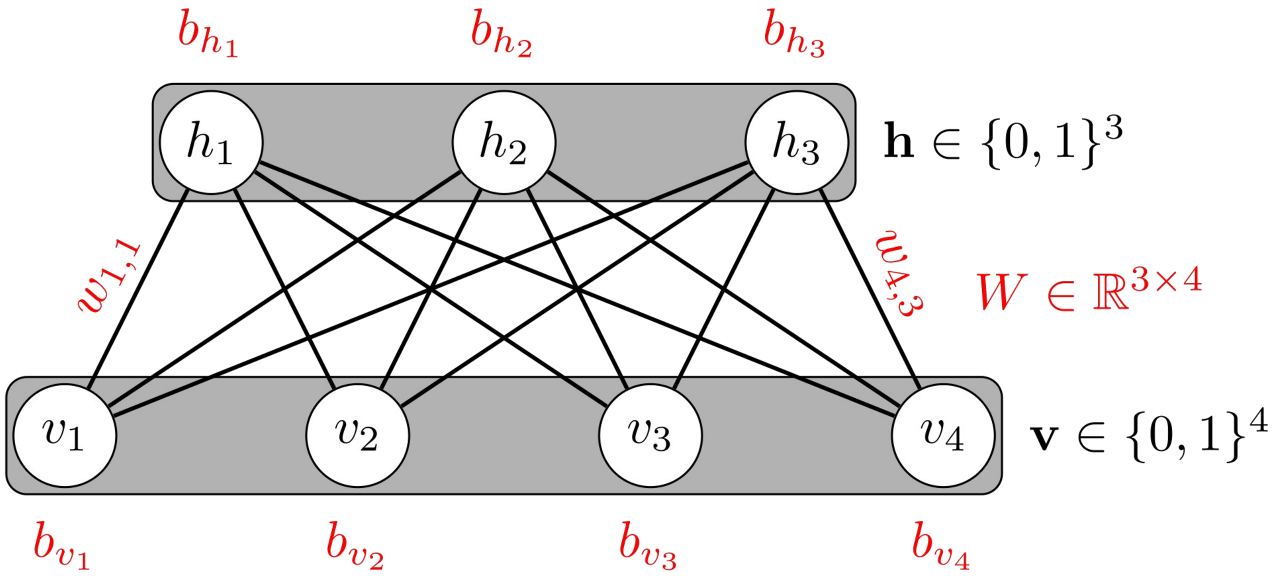
mongoDB®

Is Conformance Checking trustable

- No! Even if the considered information system follows strict rules, it may happen that fraudolent events get inserted following those rules but showing pretty different behavior than other events.
- Take as example a scientific paper peer review process: one may have fixed a single rule in which only some people (the reviewers) can do a «review», and other people (the researchers) can do a «submission». We can expect the reviewer to be different from the person who has done the submission, although there is not a clear rule about this in the considered information system.
- So, Conformance Checking on process models is not enough.

Rules discovery through Restricted Boltzmann Machines

- Restricted Boltzmann Machines trained on events to learn correlations between attributes.
- Higher-abstraction correlations can be learned by applying a Deep Belief network (multiple RBM).



Discovering significant event attributes

- Events/paths can be described by several attributes.
- With NoSQL you can fastly find event attributes and their count (so you can measure entropy).
- So you can find the suitable attributes to train a Restricted Boltzmann Machine.
- The activation of RBM hidden units is based on a probability value.

NoSQL and Security

Rema Hariharan
eBay

What drove the growth of our databases

- **What drove the Relational databases**
  - Atomicity
  - Consistency
  - Isolation
  - Durability
  - Guarantees CA from CAP theorem context (data consistent between nodes and as long as all nodes are online, users can read/write from any node and be sure that the data is the same on all nodes.)
  - Designed for vertical scaling.

- **What drives the NoSQL databases**
  - High Scalability – designed to run on clusters
  - Distributed Computing
  - Lower cost
  - Schema flexibility
  - Unstructured data
  - No complex relations
  - Guarantees AP/CP from CAP theorem context (nodes remain online even if they can't communicate with each other, re-syc data once the partition is resolved, no guarantee about all nodes having the same data)

Growth of NoSQL databases

- Triggered by Big data growth and prevalence of unstructured data.
- Need for Performance and scalability more than ACID properties.
- NoSQL expands into financial sector!!!
- Security is an afterthought!

Three stages at which security is needed

- **Dataflow protection**
  - DGI (Data Governance Initiative)
    - Apache Ranger (framework to enable, monitor and manage comprehensive data security across the Hadoop platform)
    - Apache Falcon (feed processing and feed management system)
- **At rest data encryption**
  - Not built into a lot of NoSQL dB
  - Version2.6 of Hadoop has encryption zones defined.
    - Prevents data from being moved around.
  - Computing on encrypted data (CryptDB), results decrypted by client.
- **Authentication**
  - Hadoop doesn’t have security mechanisms turned on by default.
  - Kerberos server can be used.

NoSQL Databases and Security

- Administrative user authentication – not enabled
- Plain text communication
- Cannot use external encryption tools like LDAP and Kerberos
- Lack of encryption support for data files
- Weak authentication between client and servers
- Overall security: Weak and inconsistent - wild west!

NoSQL contd.

- NoSQL was not designed with security as a priority, so the security layer needs to be added.
- NoSQL db are vulnerable to the same security risks as RDBMS
  - Encrypt sensitive fields
  - Apply authentication policies
  - Input validation

What some noSQL db are doing

- Oracle added transactional control over data written to node and Kerberos Authentication.
- Cassandra – transaction logging, automatic replication
- MongoDB – Master-slave replication
- NoSQL + RDBMS technology -> NewSQL

Anomaly Detection and Trust

- Use of Statistical methods to detect data abnormalities.
  - Rule based methods.
  - Methods to detect data with unusual combination of dimensions.
    - PCA dimensionality reduction + distance based methods
    - K-clustering
    - Simple Bolinger bands for known quantities
- We have the compute power to do this.

Summary

- NoSQL was designed for Performance and Scalability, security is an afterthought
- To store information securely, we need Confidentiality, Integrity and Availability. – not a package inclusion in NoSQL.
- Performance and Scalability at odds with Security
- Confidentiality and Integrity have to be provided by the application
  - Will certainly work (MarkLogic Server accredited with DOD)
  - No Read up and No Write Down (Government)
  - Application bugs
- Detailed Anomaly detection methods are being used to add trust to NoSQL data. Open sourced tools and methods are add-ons that can be tailored to individual needs.