Les données scientifiques et les problématiques particulières liées à leur qualité

Laure Berti-Equille

IRD, UMR ESPACE DEV

laure.berti@ird.fr

Classification

Données d'observation

collectées à un instant, nécessitant un apparat descriptif conséquent (conditions, méthodologie, équipement, ...). Indissociables d'un contexte donné et uniques et impossibles à reproduire. A conserver de façon pérenne: neuroimagerie, concentration de phytoplanctons, cliché astronomique, données climatologiques, données d'enquête, séquence de gênes, ...

Données expérimentales

obtenues à partir d'équipements suivant une méthodologie bien définie. Potentiellement reproductible, mais à des coûts parfois prohibitifs. La conservation dépend des investissements engagés dans leur production et de leur possible reproductibilité : chromatogrammes, cinétique chimique, ...

Données computationnelles ou de simulation

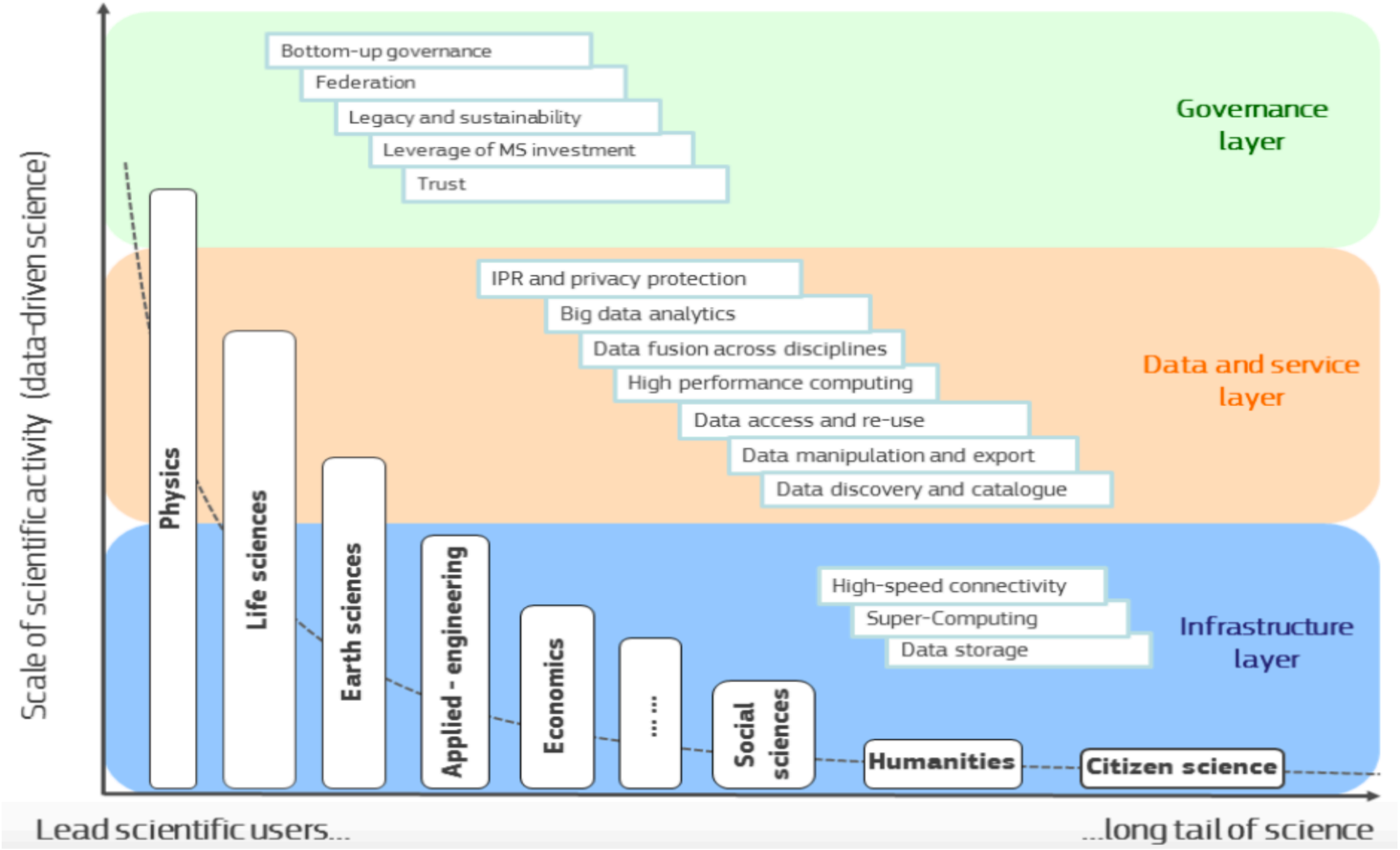
issues de simulations à partir de modèles informatiques. Potentiellement reproductibles si le modèle informatique est correctement documenté : modèles de simulation sismique, modèles météorologiques, modèle économique, ...

Données dérivées ou compilées

Issues du traitement, de la combinaison ou de la réorganisation de données brutes, pour les rendre plus lisibles ou les présenter sous une forme canonique : imagerie IRM, fouille de texte, bases de données intégrées, résumés

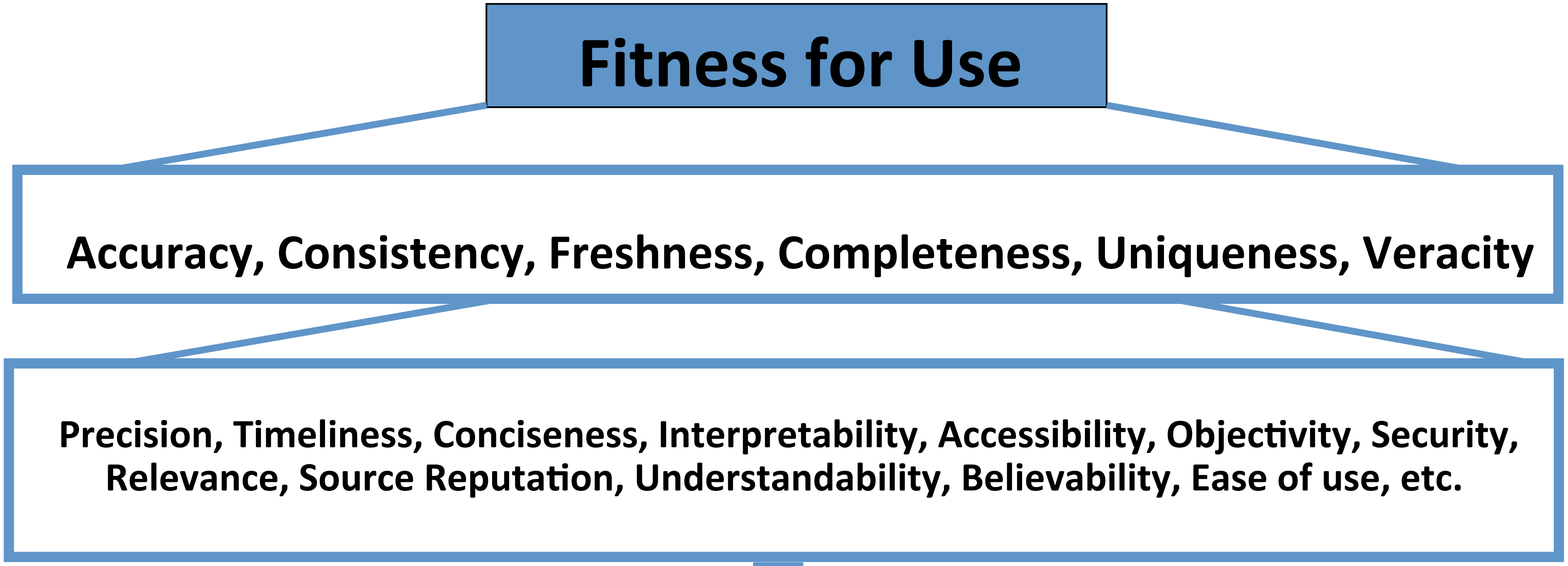
Source: Rapport de R. Gaillard, 2014, p18, citant la NSF et le RIN (Research Information Network)

Data-driven Science



Source Francis André CNRS, 2016 :
https://anfdonnees2016.sciencesconf.org/data/pages/ANF_RENATIS_2016_FANDRE_1.pdf

Data Quality:
A multidimensional definition

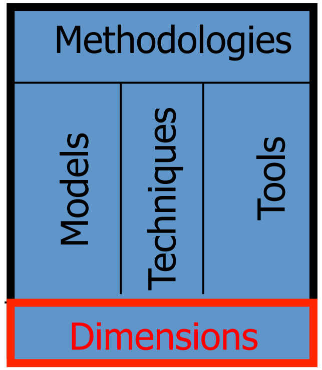


Fitness for Use

Accuracy, Consistency, Freshness, Completeness, Uniqueness, Veracity

Precision, Timeliness, Conciseness, Interpretability, Accessibility, Objectivity, Security, Relevance, Source Reputation, Understandability, Believability, Ease of use, etc.

Up to 179 dimensions



Methodologies

Models

Techniques

Tools

Dimensions

Categories of Data Quality Problems

Input Data Type
Continuous
Nominal (string)
Categorical
Binary
Multimedia (text, AV, image)
Hybrid

Relationship between Data Instances
Structural (record)
Sequential
Graph-based
Temporal
Spatial
Spatio-Temporal

Nature
Missing data
Atypical data
Duplicate Data
Inconsistent Data

Cardinality
Single-Point
Collection

Detection Referential
Model
Data Distribution
Constraint
Data Pattern

Categories of Data Quality Problems

Input Data Type
Continuous
Nominal (string)
Categorical
Binary
Multimedia (text, AV, image)
Hybrid

Relationship between Data Instances
Structural (record)
Sequential
Graph-based
Temporal
Spatial
Spatio-Temporal

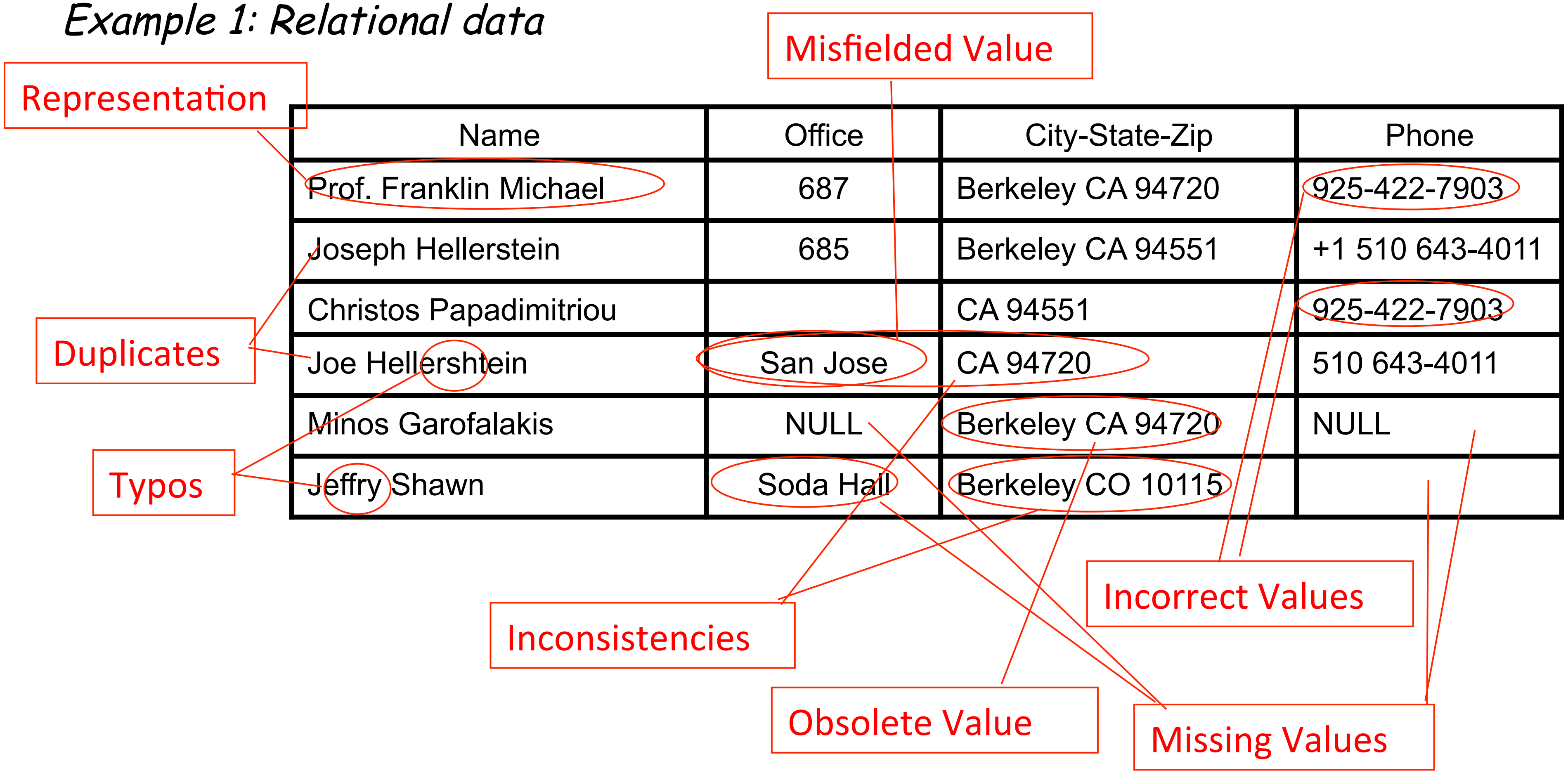
Nature
Missing data
Atypical data
Duplicate Data
Inconsistent Data

Cardinality
Single-Point
Collection

Detection Referential
Model
Data Distribution
Constraint
Data Pattern

Data Quality Problems

Example 1: Relational data



Name	Office	City-State-Zip	Phone
Prof. Franklin Michael	687	Berkeley CA 94720	925-422-7903
Joseph Hellerstein	685	Berkeley CA 94551	+1 510 643-4011
Christos Papadimitriou		CA 94551	925-422-7903
Joe Hellershtein	San Jose	CA 94720	510 643-4011
Minos Garofalakis	NULL	Berkeley CA 94720	NULL
Jeffry Shawn	Soda Hall	Berkeley CO 10115	

Representation

Misfielded Value

Duplicates

Typos

Incorrect Values

Inconsistencies

Obsolete Value

Missing Values

Categories of Data Quality Problems

Input Data Type
Continuous
Nominal (string)
Categorical
Binary
Multimedia (text, AV, image)
Hybrid

Relationship between Data Instances
Structural (record)
Sequential
Graph-based
Temporal
Spatial
Spatio-Temporal

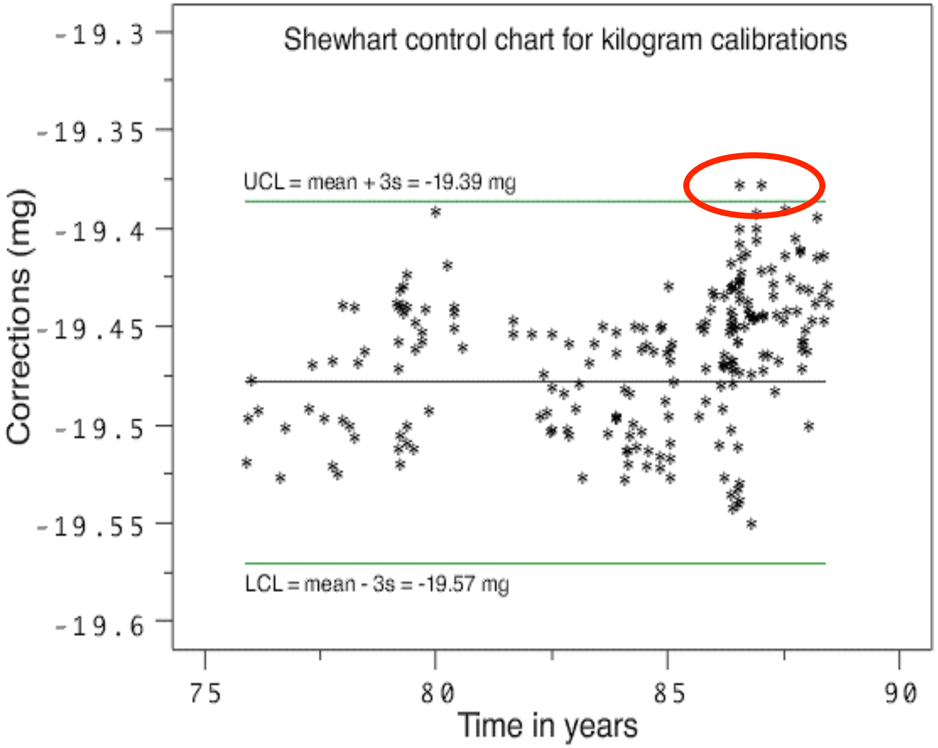
Nature
Missing data
Atypical data
Duplicate Data
Inconsistent Data

Cardinality
Single-Point
Collection

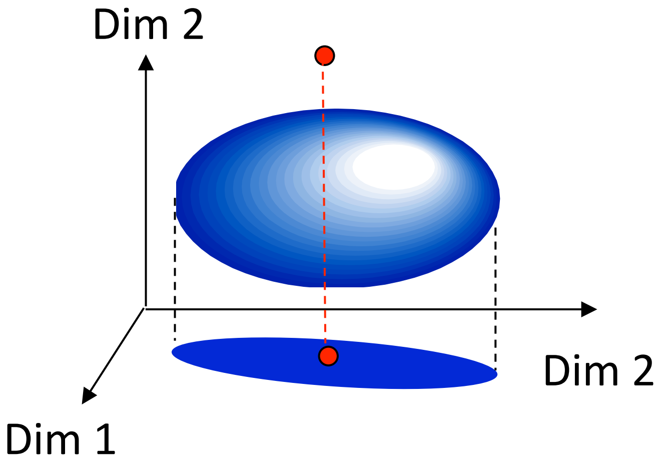
Detection Referential
Model
Data Distribution
Constraint
Data Pattern

Data Quality Problems

Example 2: Bivariate and multivariate outliers



(http://www.itl.nist.gov/div898/handbook/mpc/section3/mpc3521.htm)



Categories of Data Quality Problems

Input Data Type
Continuous
Nominal (string)
Categorical
Binary
Multimedia (text, AV, image)
Hybrid

Relationship between Data Instances
Structural (record)
Sequential
Graph-based
Temporal
Spatial
Spatio-Temporal

Nature
Missing data
Atypical data
Duplicate Data
Inconsistent Data

Cardinality
Single-Point
Collection

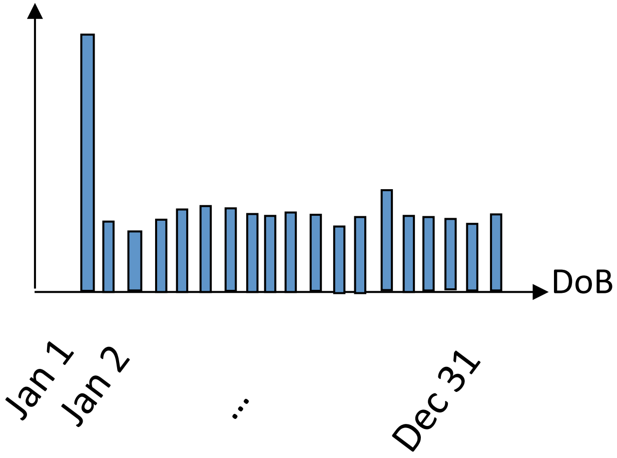
Detection Referential
Model
Data Distribution
Constraint
Data Pattern

Data Quality Problems

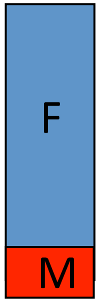
Example 3: Disguised missing data

The data values exist, satisfy the syntactical or domain constraints (inliers) but are erroneous. Potentially detectable with the data distribution that doesn' t conform to an expected model

e.g., 30% of the population is born on January 1rst



e.g., 10% patients in obstetrical emergency are male



Domain knowledge is required !

Categories of Data Quality Problems

Input Data Type
[✓ Continuous]
Nominal (string)
Categorical
Binary
Multimedia (text, AV, image)
Hybrid

Relationship between Data Instances
Structural (record)
Sequential
Graph-based
[✓ Temporal]
Spatial
Spatio-Temporal

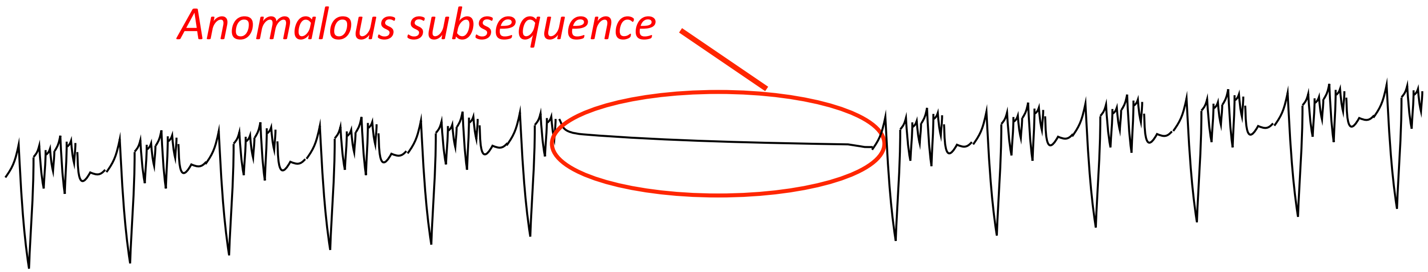
Nature
Missing data
[✓ Atypical data]
Duplicate Data
Inconsistent Data

Cardinality
Single-Point
[✓ Collection]

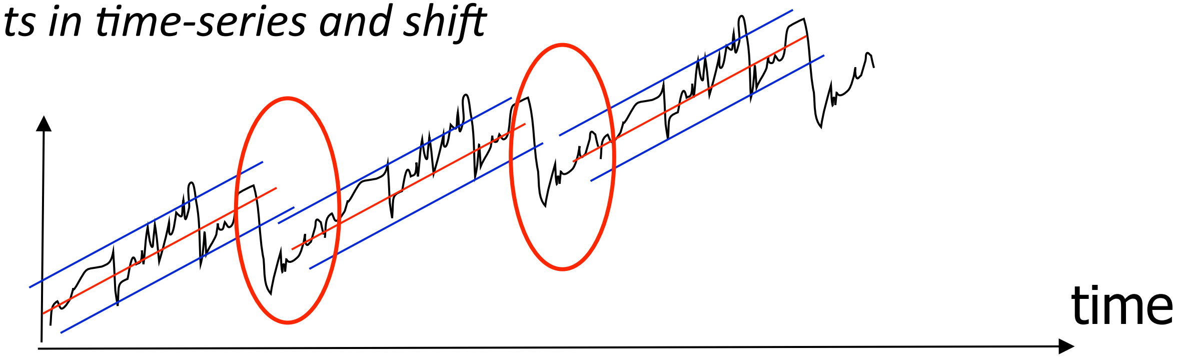
Detection Referential
Model
Data Distribution
Constraint
[✓ Data Pattern]

Data Quality Problems

Example 4: Time-Dependent Anomalies



Example 5: Deviants in time-series and shift



Domain knowledge is required !

Categories of Data Quality Problems

Input Data Type
Continuous
Nominal (string)
Categorical
Binary
Multimedia (text, AV, image)
Hybrid

Relationship between Data Instances
Structural (record)
Sequential
Graph-based
Temporal
Spatial
Spatio-Temporal

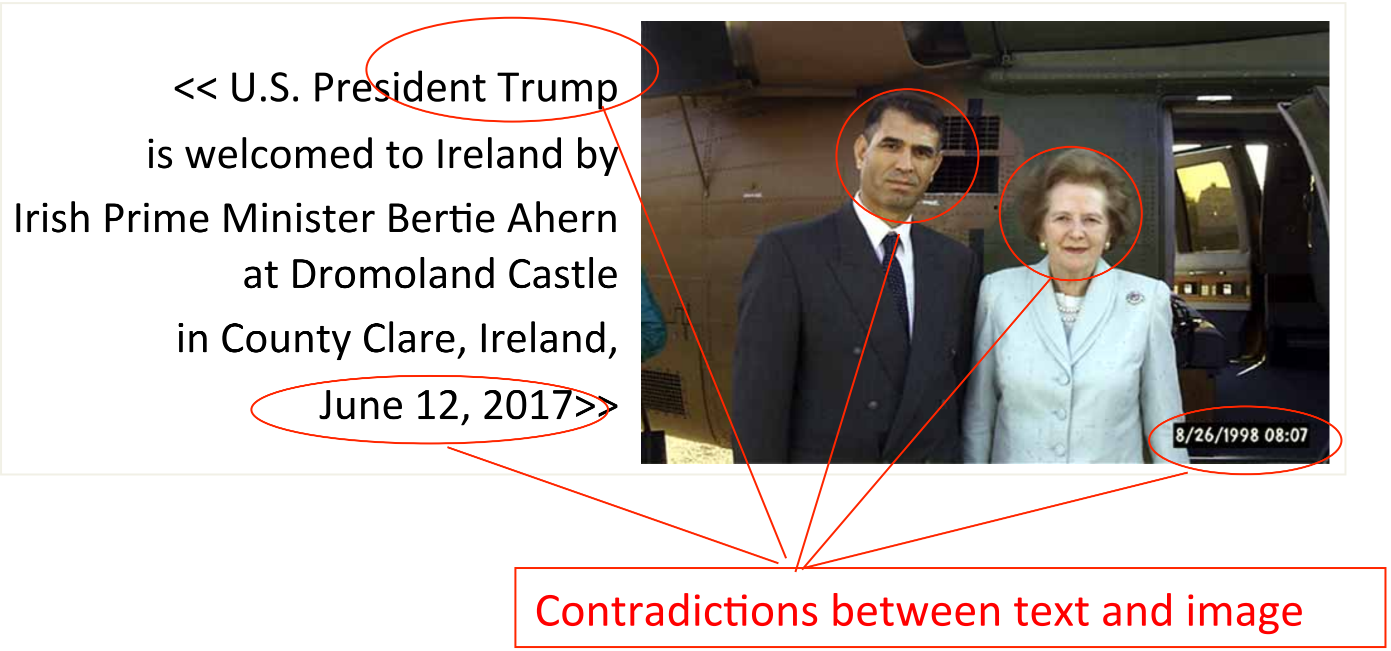
Nature
Missing data
Atypical data
Duplicate Data
Inconsistent Data

Cardinality
Single-Point
Collection

Detection Referential
Model
Data Distribution
Constraint
Data Pattern

Data Quality Problems

Example 6. Where was D. Trump Bush in June 2017?



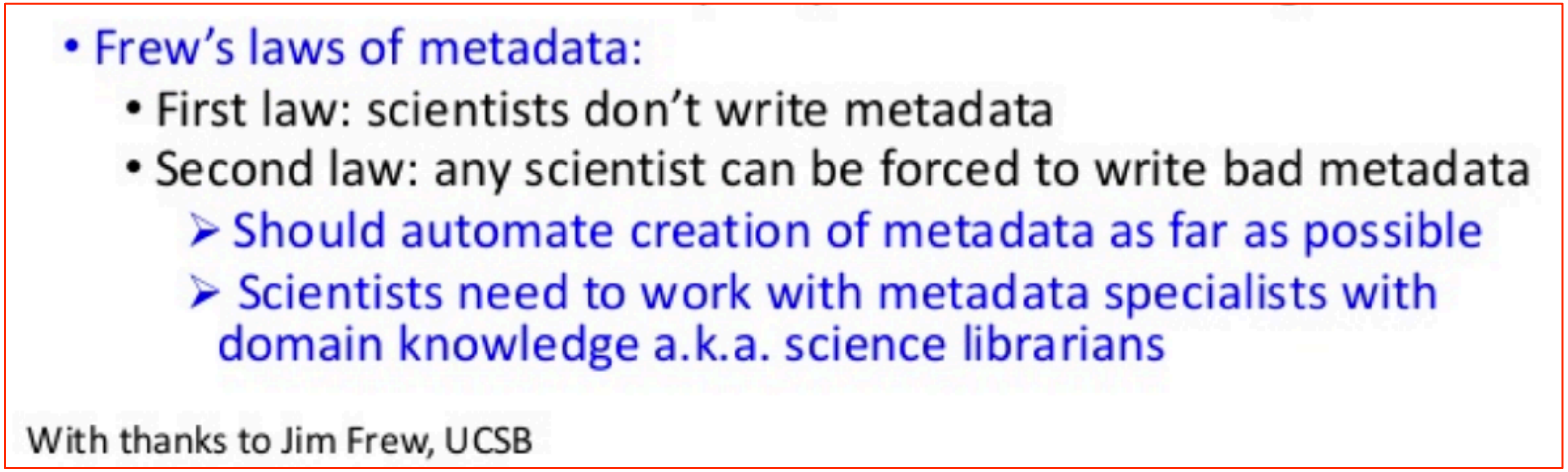
<< U.S. President Trump
is welcomed to Ireland by
Irish Prime Minister Bertie Ahern
at Dromoland Castle
in County Clare, Ireland,
June 12, 2017>>

Contradictions between text and image

Cross-modality inconsistency detection

Domain knowledge is required !

Data Quality Challenges for eScience (1)



- Frew's laws of metadata:
  - First law: scientists don't write metadata
  - Second law: any scientist can be forced to write bad metadata
    - ➢ Should automate creation of metadata as far as possible
    - ➢ Scientists need to work with metadata specialists with domain knowledge a.k.a. science librarians

With thanks to Jim Frew, UCSB

Main challenge:

How to capture the domain knowledge into DQ actionable constraints and indicators ?

Data Quality Challenges for eScience (2)

More “classical” challenges:

- **Research Methodology:** We need benchmarks
- **DB/IS Engineering**
  - Design patterns and “native” data and data quality management
- **DDL and DML Languages**
  - Declaration and management of data along with computed DQ indicators
  - Design and development of DQ-constrained query languages
- **Algorithms**
  - Generation of DQ metadata
  - Detection of error patterns and masking effect
  - UDF and approximation algorithms for DQ evaluation
  - Indexation of data with DQ metadata
  - Adaptive processing and optimization of queries with DQ UDAs