
WHAT MAKES BIG DATA DIFFERENT FROM A DATA QUALITY ASSESSMENT PERSPECTIVE?

Practical Challenges for Data and Information Quality Research

Michael Kläs, Adam Trendowicz, Andreas Jedlitschka

Parts of this work are funded by



Gefördert durch:
 Bundesministerium
für Wirtschaft
und Energie
aufgrund eines Beschlusses
des Deutschen Bundestages

ODQ2015

30 March 2015, Garching, Germany

© Fraunhofer IESE



Open / Big Data Quality – The Motivation



- **Data quality (DQ)**
 - Research topic for several decades
 - Business intelligence raised awareness

- **Big data (BD)**
 - Existing DQ means not directly transferable
 - Still largely unclear how to assess DQ of BD

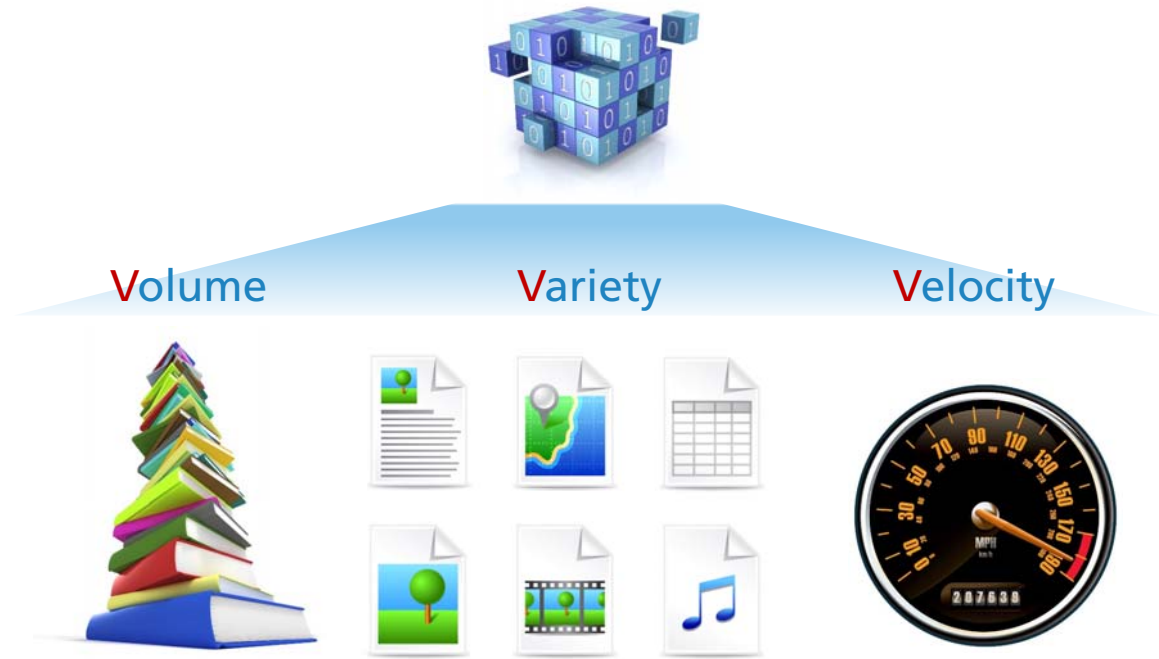
- **Open data (OD)**
 - shows similar properties as other BD

2

© Fraunhofer IESE



What makes BIG Data Different? – The 3 Vs [Laney2001]



3

Data Quality Assessment – A Conceptual Model

 Goals / Quality Needs


Analyze the impact of the voter's age on the choice of a specific political party

 Model of Data Properties

... Validity of DATE entries

 Metrics

M1: % „birth date” entry has correct structure, M2: % ELECTION_DATE - “birth date” >= 18 years

 Operation-ization

R script mysql script Data quality tool

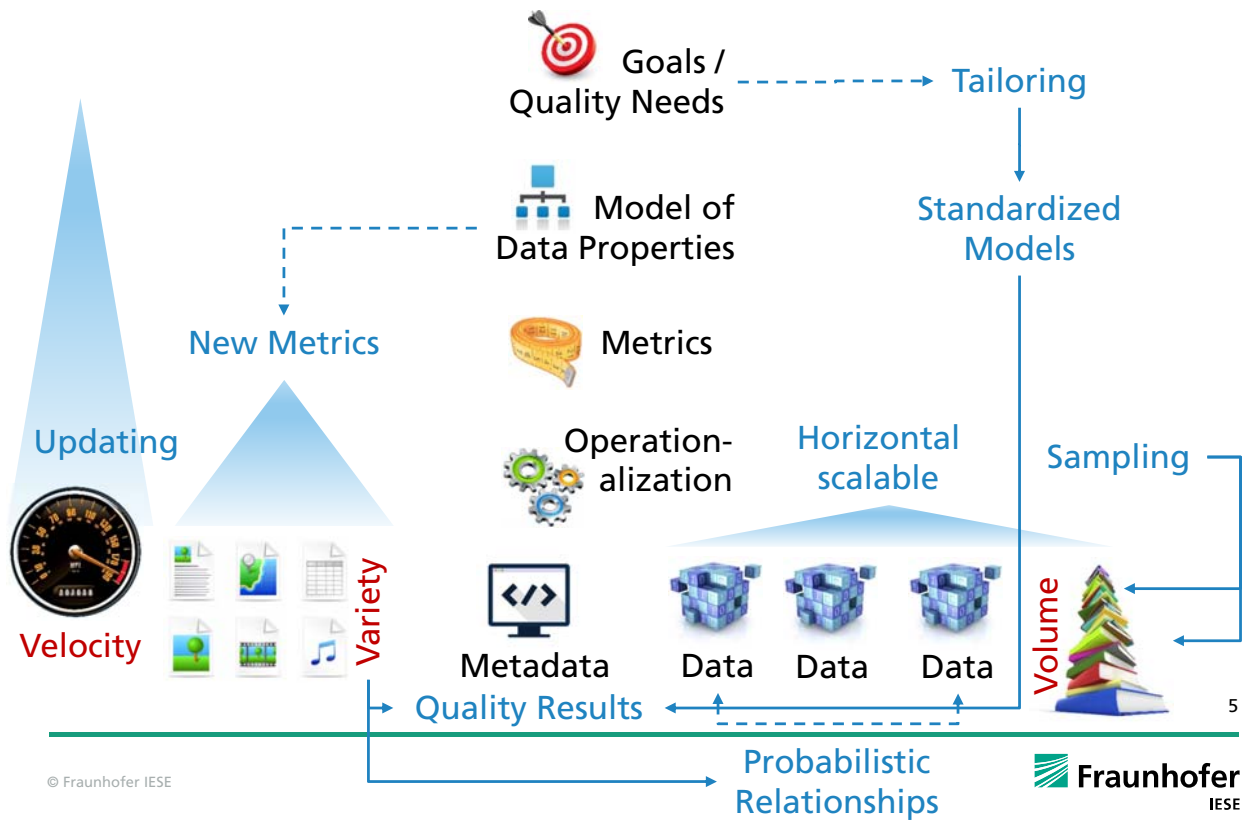
 Metadata  Data

Structure of “birth date” column: YYYY-MM-DD

..., 1957-03-21
..., 1970-30-05

4

Challenges for BIG Data Quality Assessments



Outlook

Adaptation of concepts developed for **software quality assessments** ?

- **Abstracting** from different implementations
- **Systematic adaptation** procedures
- **Incremental analysis** of quality

Contact

Michael Kläs

michael.klaes@iese.fraunhofer.de

Fraunhofer IESE

Fraunhofer-Platz 1
D-67663 Kaiserslautern
Germany



7