Wetenschappelijk Onderzoek- en Documentatiecentrum
Ministerie van Veiligheid en Justitie

On Utilizing Data Analysis in Practice

Sunil Choenni

Juni 2017

Content

- Introduction
- Challenges in Data Collection
- Challenges in Interpreting of the result of Data Analysis
- Two strategies for interpretations
- Conclusions

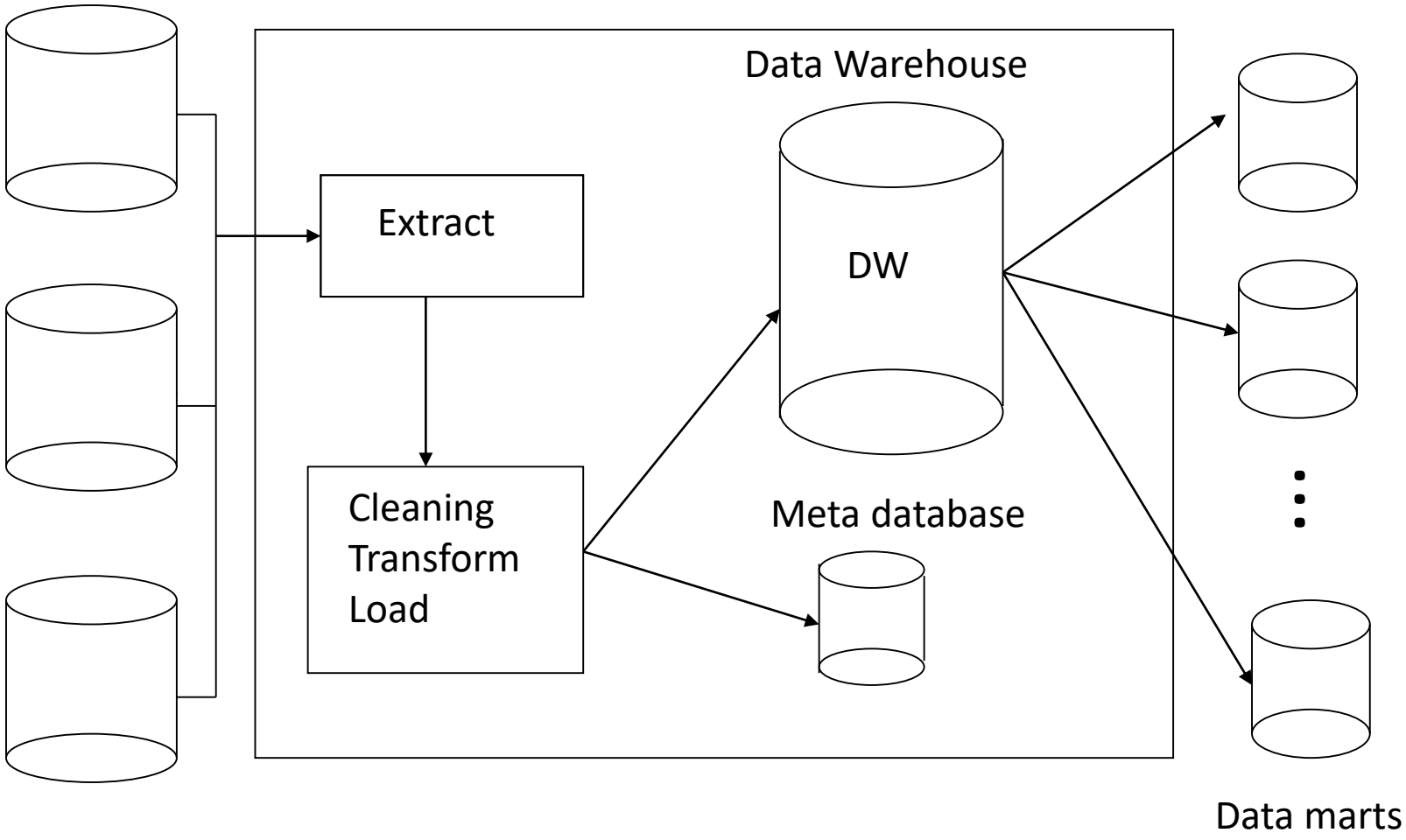
Data Collection: Perfect Data

- Closed World Assumption
- Complete data set

Data Analysis

- Induction
- Deduction

Architectures I



Challenges Data Collection: Uncertainty

- Legacy Data: poor documentation
- Null Values
- Evolving environment/semantics
- Interoperability

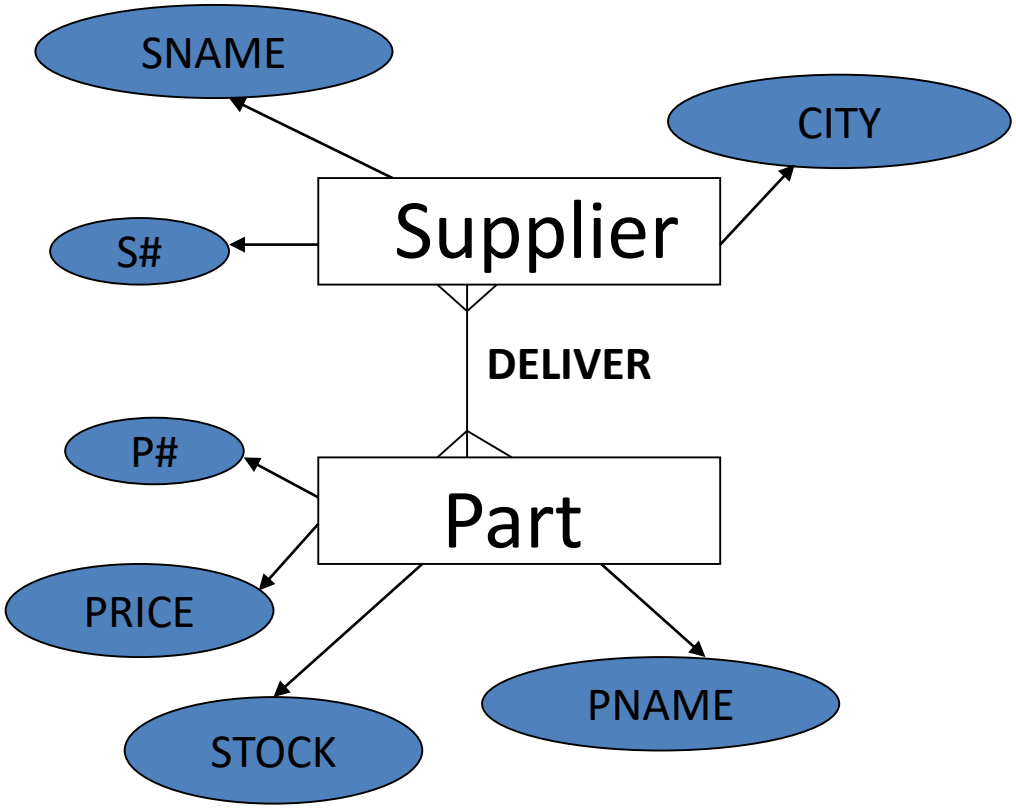
Exploit Domain Knowledge -→ Uncertainty

Semantic Level: Example

Stored birth-place of an offender is USSR
In 1991, USSR is divided in several countries.
Should stored birth-place be updated to Russia or be left as it is?

Leaving as it is:
How many offenders were born in Russia?
Today, USSR does not reflect a real-place
DQ(country) in the past was fine but today poor.
So,DQ degradation

Challenges Data Collection: Incompleteness



On the Interpretation of Outcome of DA

System realities may deviate from true reality

- Databases may contain large amounts of data that were collected and stored in the (distant) past.
- When legacy databases are used, the results obtained through analysis, (e.g.data mining) do not always hold in the real world of today
- The results may hold for the past at the time the data was gathered

Analysis of the complaints over the last 35 years of the National Ombudsman resulted in

Males who are well educated living in urban areas → higher chance to lodge complaint

On the Interpretation of Outcome of DA

Statistical Truths

young men driving in leased cars -→

80% chance to be involved in car accidents

Simplification: Jones has 80% chance to be involved in a car accident

frequentist approach: p=relative frequency

- Clones
- Number of drives

On the Interpretation of Outcome of DA

Statistical Truths

Simplification: Jones has 80% chance to be involved in a car accident

Subjective approach: p = quantified judgement

- Prior probability (may include “frequentist approach”)
- Interpretation different for receiver and probability generating entity

Summary

- Uncertainty wrt truthfulnes of data
- Incompleteness of data
- Legacy data sets and notion of probability

Two strategies to deal with outcome: Strategy 1

- Consider the outcome of DA as a central body of evidence and extract a hypothesis

  e.g. Jones is a risky driver
- Search for evidences that weaken hypothesis

  e.g. Jones is a cautious man
- If enough evidences are found to weaken the hypothesis then hypothesis is rejected

Self-denying prohecy: true hypothesis might become false

Two strategies to deal with outcome: Strategy 2

- Consider the outcome of DA as a central body of evidence and extract a hypothesis

  e.g. Jones is a risky driver
- Search for evidences that strenghten hypothesis

  e.g. Jones is a reckless person
- If enough evidences are found to strengthen the hypothesis then hypothesis is accepted

Self-fulfilling prohecy: false hypothesis might become true

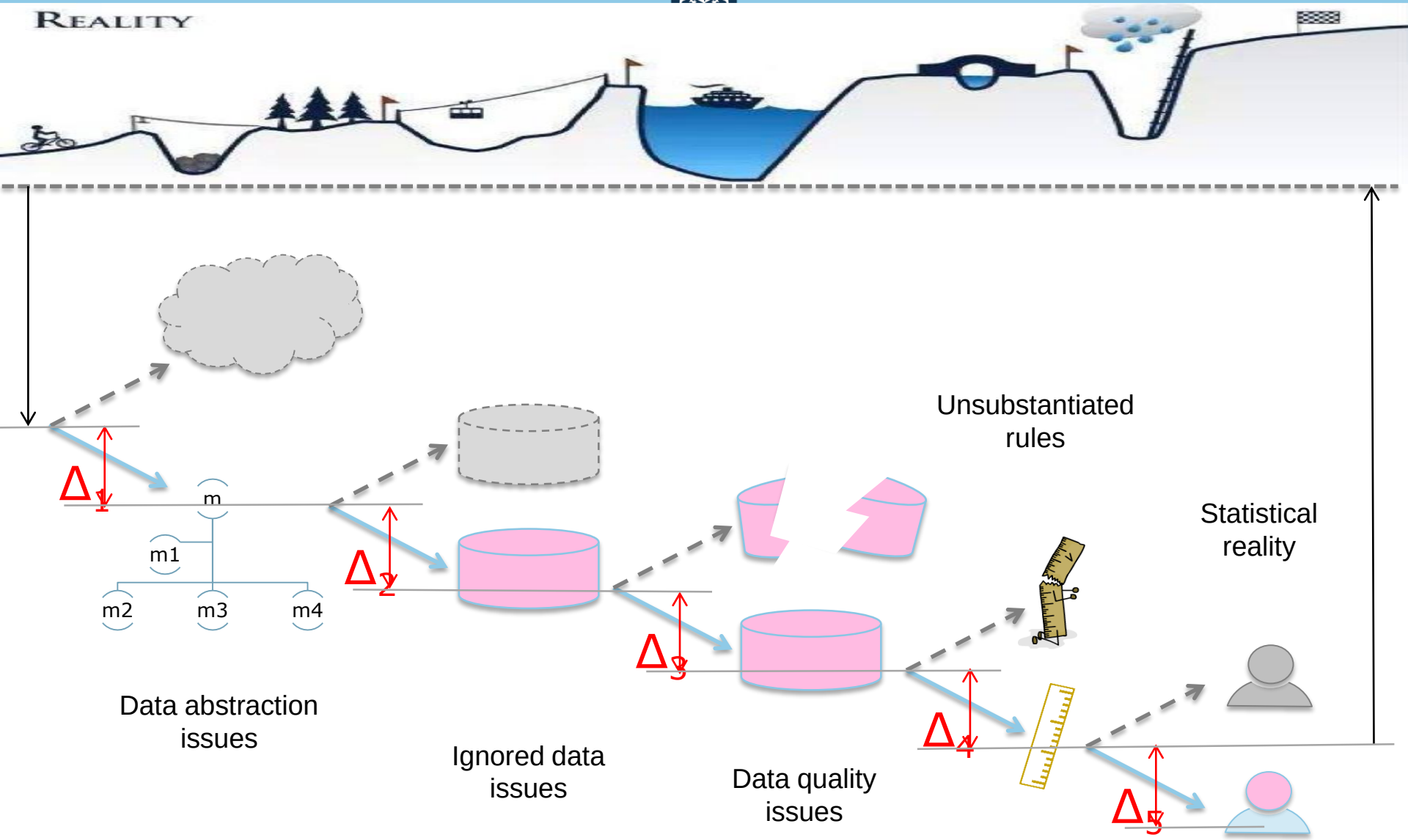
Which strategy to choose?

Depends on application

- impact of false positives and false negatives
- procedure to deal with false positives and false negatives
- tailor the strategy to application at hand

Strategy 1: self-denying; tends to reduce false positives

Strategy 2: self-fulfilling; tends to reduce false negatives


REALITY
m
m1
m2
m3
m4
Data abstraction issues
Ignored data issues
Unsubstantiated rules
Data quality issues
Statistical reality