REQUISITES FOR A WELL FOUNDED BI PROJECT

Paolo Ceravolo

paolo.ceravolo@unimi.it

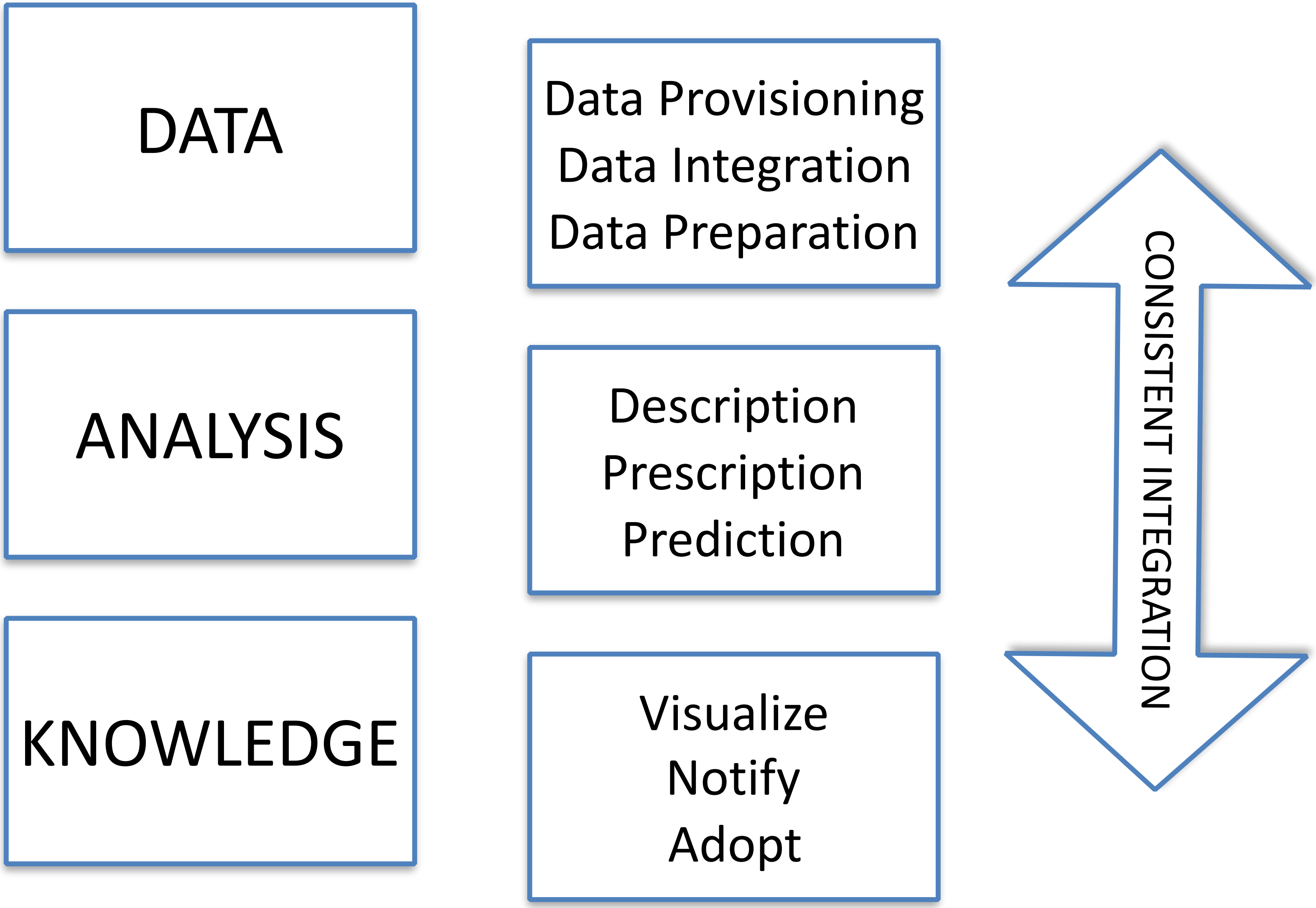
Dipartimento di Informatica - Sede di Crema

Thursday 14.00 - 15.00

Outline of the course

1. Definition of BPI
2. Overview on BI methods
3. Focus on Process Mining algorithms
4. **Requisites for a well founded BI project**
5. Exercise: review BPI works

Business Intelligence



DATA

ANALYSIS

KNOWLEDGE

Data Provisioning
Data Integration
Data Preparation

Description
Prescription
Prediction

Visualize
Notify
Adopt

CONSISTENT INTEGRATION

Validation

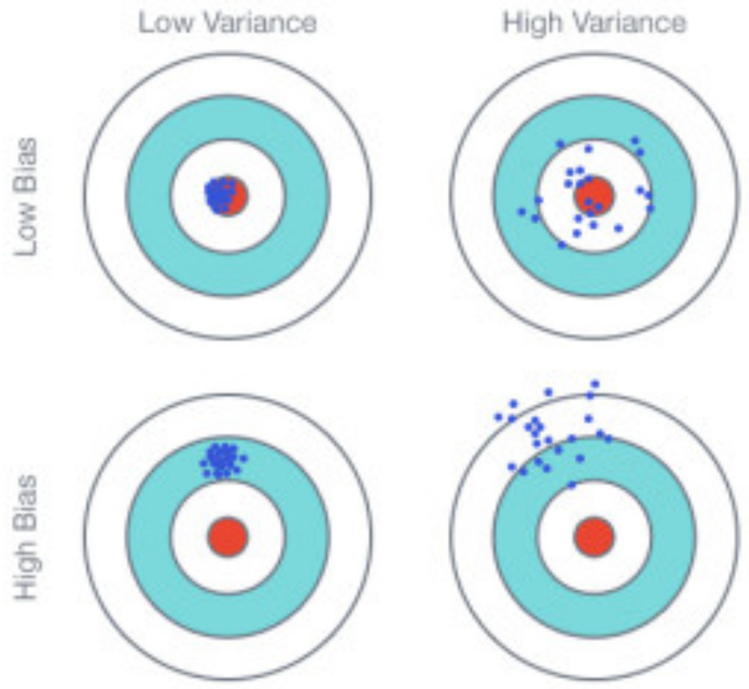
In Data Mining validation is required to assess the quality of a learning stage

Overfitting and Underfitting are both undesirable behaviours that must be countered. Both lead to poor predictions on new data sets

They both depends on:

- **Bias** is originated from erroneous assumptions in the learning process
- **Variance** is originated from sensitivity to fluctuations

- **At data or model level!**

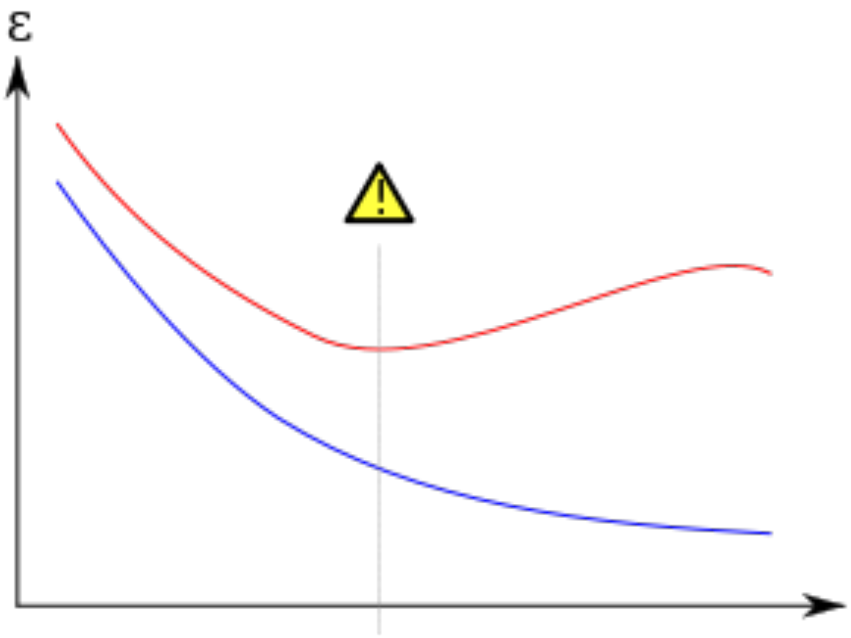


Validation

- Overfitting occurs when a statistical model or machine learning algorithm captures the noise of the data
  - Intuitively, overfitting occurs when the model or the algorithm fits the data too well, it is an excessively complicated model
  - Specifically, overfitting occurs if the model or algorithm shows low bias but high variance
- Underfitting occurs when a statistical model or machine learning algorithm cannot capture the underlying trend of the data
  - Intuitively, underfitting occurs when the model or the algorithm does not fit the data well enough, it is an excessively simple model
  - Specifically, underfitting occurs if the model or algorithm shows low variance but high bias

Validation

- How to detect them?
  - Underfitting – Validation error and training error are both high
  - Overfitting – Validation error is high while training error is low
  - Good fit – Validation error is low, and only slightly higher than the training error
- How to reduce Bias and Variance?
  - cross-validation
  - early stopping
  - model comparison



Consistency at the Design Time

The challenge is ensuring consistency among these different models

Clearly *deviant behaviour* is common and unavoidable:

- Errors in the implementation
- Partners not conforming to the agreements
- Deviation not depending to the business partners

It is then crucial to control

- Making things observable
- Having a methodology to regain consistency

Consistency at the Data Level

- Data Quality is a multidisciplinary field aimed at assessing data’s fitness to their intended uses in operations, decision making and planning [Juran 1981] or in data acquisition, transformation and processing [Kohavi et al. 2004]
- A Data Quality Dimension is a recognised term used by data management professionals to describe a feature of data that can be measured or assessed against defined standards in order to determine the quality of data
- In a structured process that combines a broad set of data analysis applications, like BI, Data Quality must be verified at each input-out stage

Consistency at the Data level

- At syntactic level data quality refers to the assessment of properties that can be validated by internal observations:
  - completeness: checking the presence of each mandatory attribute
  - integrity: checking the internal identifiers applying to a data set
  - consistency: checking compatibility between entries
  - validity: validating the entries against data type or a set of well-defined valid values
  - maintainability: estimating the extend to which data is easy to manipulate, manage and make interoperable with other data formats
  - timeliness: comparing timestamps against each other to validate their operational consistency

Consistency at the Data Level

- At semantic level data quality refers to the assessment of properties that can be validated by external observations:
  - accuracy: checking entries values against a third party data, typically considered at a higher level of reliability
  - coverage: validating the ability of representing every meaningful state of the represented real world system, avoiding ambiguity, incompleteness, and unmeaningly

Consistency at the Data Level

- At pragmatic level data quality refers to the assessment of properties related to fitness to intended uses:
  - relevance: validating the capability in enabling users to achieve the intended uses
  - usability: validating the effectiveness, efficiency, and satisfaction in enabling users to achieve the intended uses
  - currency and decay: checking availability and decay for use in the time frame in which they are expected
  - risks: estimating risks for the business process and the operational environment

Consistency at the Data Level

- How to use these dimensions?
  - Identify which data items need to be assessed for data quality, typically this will be data items critical to the Business Intelligence operations especially dimension impacting on reporting and knowledge acquisition
  - Assess which data quality dimensions to use and their associated weighting
  - For each data quality dimension, define values or ranges representing good and bad quality data. Note, that as a data set may support multiple requirements, a number of different data quality assessments may need to be performed

Model Consistency

When we use the term ‘consistency’ we mean the absence of contradictions

A set of statements is considered to be consistent to a certain logical calculus if no formula $P \wedge \neg P$ can be derived from those statements by the rules of the calculus

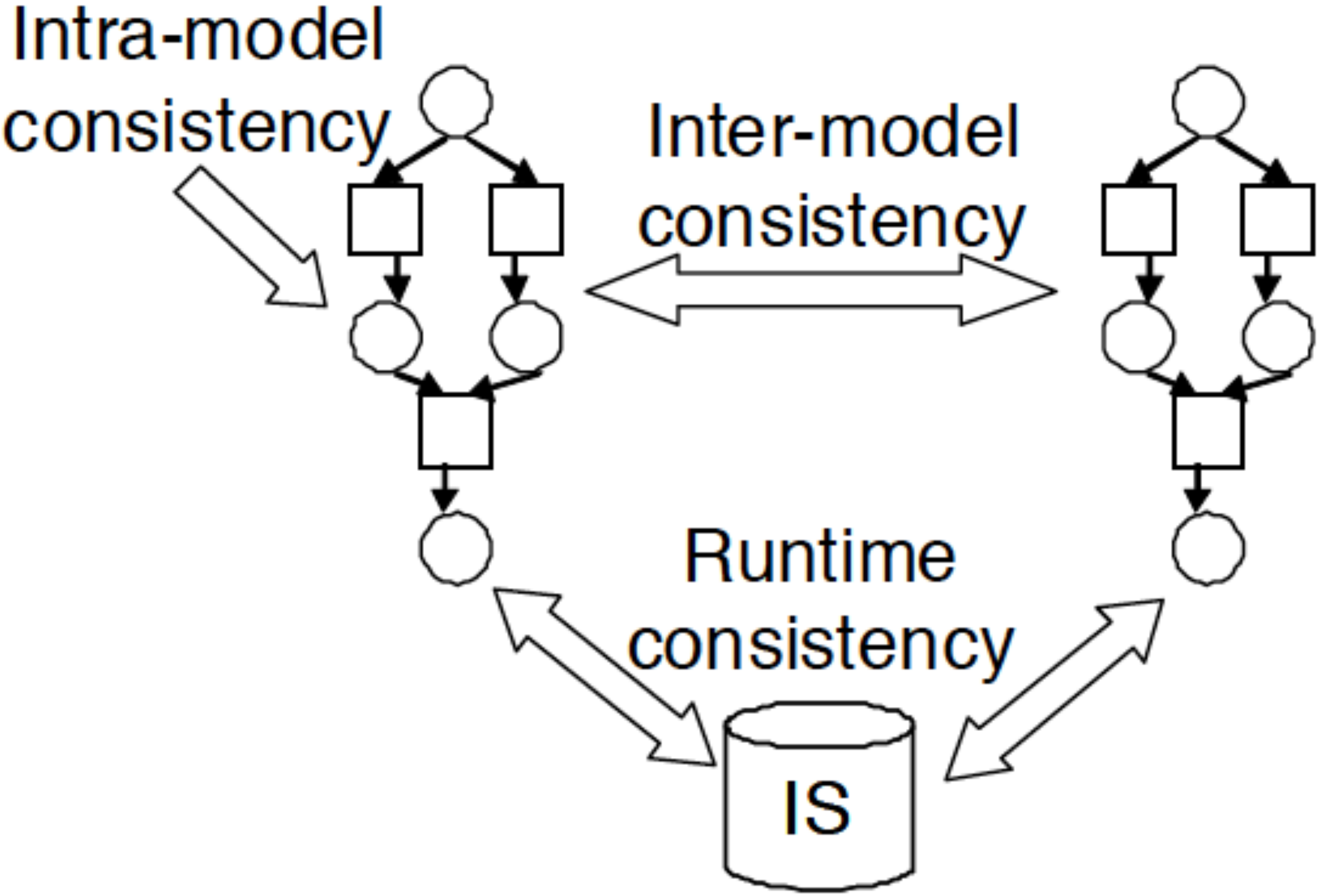
Model Consistency

Consistency can be assessed over:

- viewpoints: focus on a specific characteristics of the system
- components: focus on a sub-part of the system

- Consistency between Models taking different Viewpoint requires to identify the relations among the foci of these models
- Consistency between Models describing different Components requires to identify the relations among the sub-parts

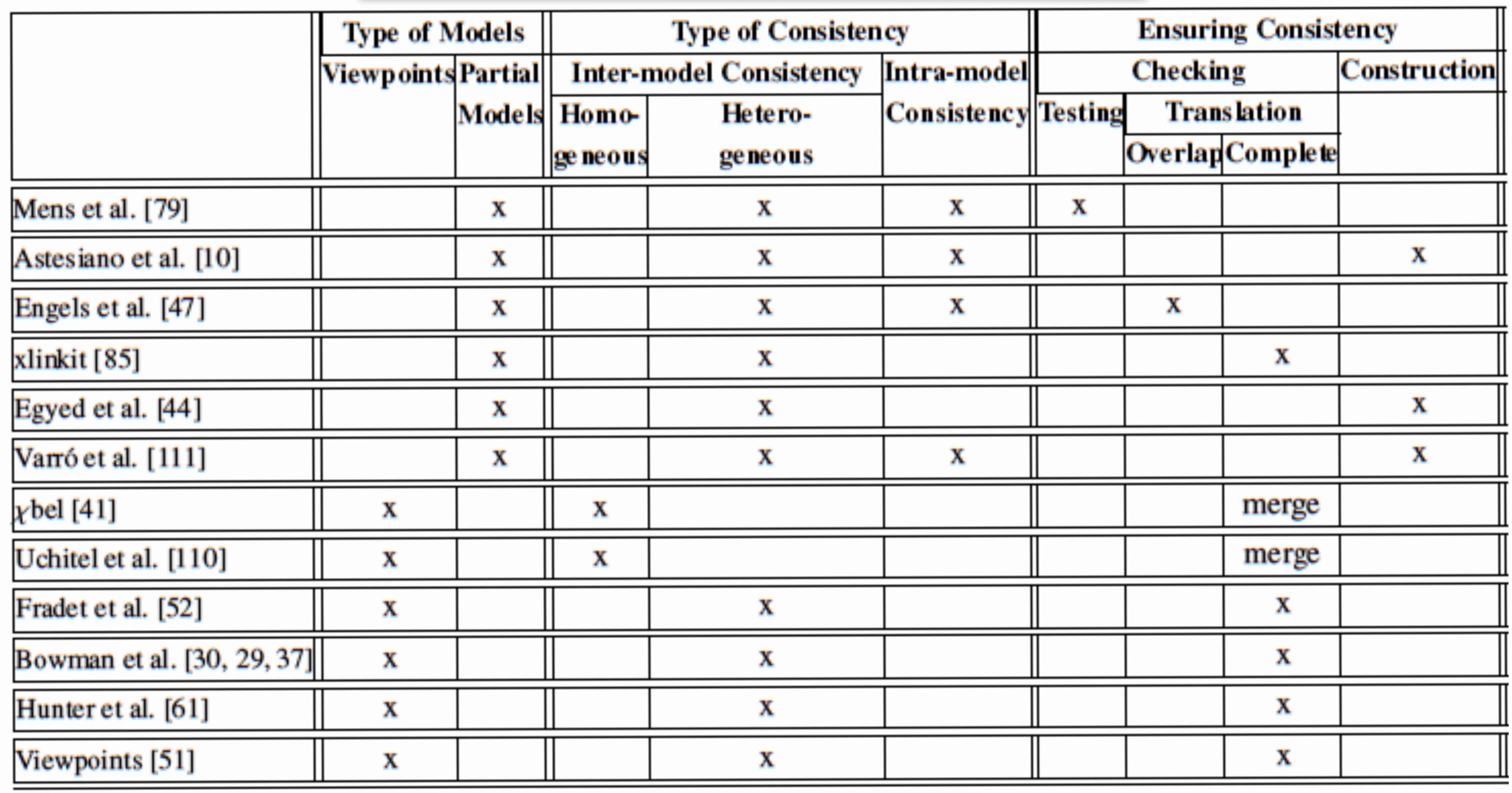
Model Consistency



Ensuring Model Consistency

- Consistency can be assessed by:
  - checking consistency
    - testing the models with some model checker
    - finding a translation with some semantically defined formalism
  - ensuring consistency
    - referring to a meta-model
- We have horizontal consistency for models with the same level of abstraction
- We have vertical consistency for model with different level of abstraction

SotA



	Type of Models		Type of Consistency			Ensuring Consistency			
	Viewpoints	Partial Models	Inter-model Consistency		Intra-model Consistency	Checking			Construction
			Homo-geneous	Hetero-geneous		Testing	Translation		
							Overlap	Complete	
Mens et al. [79]		x		x	x	x			
Astesiano et al. [10]		x		x	x				x
Engels et al. [47]		x		x	x		x		
xlinkit [85]		x		x				x	
Egyed et al. [44]		x		x					x
Varró et al. [111]		x		x	x				x
χbel [41]	x		x					merge	
Uchitel et al. [110]	x		x					merge	
Fradet et al. [52]	x			x				x	
Bowman et al. [30, 29, 37]	x			x				x	
Hunter et al. [61]	x			x				x	
Viewpoints [51]	x			x				x	

Bodenstaff, Lianne. "Managing dependency relations in inter-organizational models." (2010).

Knowledge Acquisition Consistency

It is important that domain and process experts are involved in the verification and validation of the results

It is essential they are involved during the previous mining stage, guiding the analysis to make sure that the results are useful for the organisation

Knowledge Acquisition Consistency

It is essential but ..

I CAN'T START THE PROJECT BECAUSE THE USER WON'T GIVE ME HIS REQUIREMENTS.
START MAKING SOMETHING ANYWAY. OTHERWISE WE'LL LOOK UNHELPFUL.
SO, OUR PLAN IS TO CLEVERLY HIDE OUR COMPETENCE.
YOU THINK TOO MUCH.
www.dilbert.com scottadams@aol.com
3/11/03 © 2003 United Feature Syndicate, Inc.

Knowledge Acquisition Consistency

Elicitation techniques

- Stakeholder analysis
- Analysis of existing systems or documentation, background reading
- Discourse analysis
- Task observation, ethnography
- Questionnaires
- Interviewing
- Brainstorming, focus groups
- Joint Application Design (JAD)
- Prototyping
- Pilot system
- Use cases and scenarios
- Risk analysis

Comparison of Data-Gathering Techniques

Technique	Good for	Kind of data	Plus	Minus
Questionnaires	Answering specific questions	Quantitative and qualitative data	Can reach many people with low resource	The design is crucial. Response rate may be low. Responses may not be what you want
Interviews	Exploring issues	Some quantitative but mostly qualitative data	Interviewer can guide interviewee. Encourages contact between developers and users	Time consuming. Artificial environment may intimidate interviewee
Focus groups and workshops	Collecting multiple viewpoints	Some quantitative but mostly qualitative data	Highlights areas of consensus and conflict. Encourages contact between developers and users	Possibility of dominant characters
Naturalistic observation	Understanding context of user activity	Qualitative	Observing actual work gives insight that other techniques cannot give	Very time consuming. Huge amounts of data
Studying documentation	Learning about procedures, regulations, and standards	Quantitative	No time commitment from users required	Day-to-day work will differ from documented procedures

Preece, Rogers, and Sharp "Interaction Design: Beyond human-computer interaction", p214