Software Division

Data Quality Measurements in a Hospital Data Warehouse

Richard E. Biehl, CSQE, CSSBB

Data Warehouse Architect
Data-Oriented Quality Solutions

Orlando, FL, USA rbiehl@doqs.com

Learning Objectives

1. Recognize quality issues that affect the viability of healthcare data warehouses.
2. Define overlapping mechanisms that can minimize data warehouse quality risks.
3. Describe why bad data can't always be corrected at the source applications.
4. Articulate criteria for balancing data quality errors against data availability needs.
5. Adapt example generic quality rules to specific local settings and needs.

Supplier-Customer Model (SIPOC)

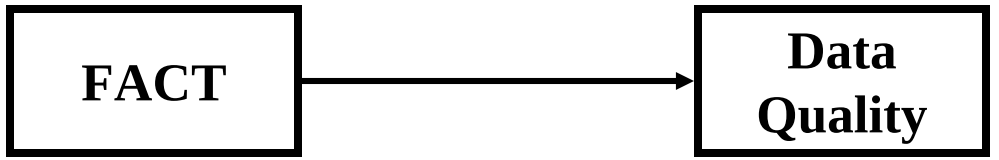
SUPPLIERS	INPUTS	REQ'TS	PROCESS	OUTPUTS	REQ'TS	CUSTOMERS
Standards Organizations	Metadata	Stable	Mount Sinai Data Warehouse	Data marts	Accessible	MSDW Users
		Complete			Controlled	
	Reference data	Stable			Compliant	
		Aligned			Appendable	
Mount Sinai Leadership	Data policies	Comprehensive		Data queries & reports	Controlled	
	MSDW resources	Available			Meaningful	
		Adequate			Clear	
Data Administration	Performance & control settings	Stable			Compliant	
		Optimal			Traceable	
Application Systems (includes appended data marts)	Metadata	Stable		Data issues & hypotheses	Relevant	Data Owners
		Complete			Timely	
	Reference data	Stable			Actionable	
		Aligned			Traceable	
	Factual data	Compliant		Performance & control data	Auditable	Data Administration
Data Owners	Reconciled data issues	Correct			Actionable	
		Authorized			Compliant	

Data Quality Measurement

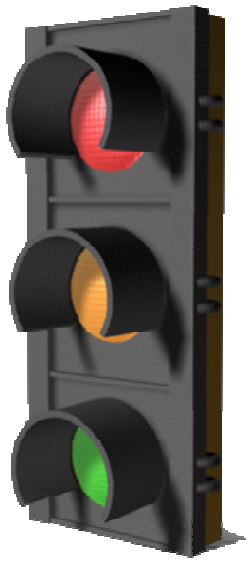
- Fact-level data quality indicators
- Generalized notification mechanism
- Hierarchy exception analysis
- Activity measurement error
- Time-series exception analysis
- Statistical process control

Fact-Level Data Quality Indicators

Every fact is assigned a data quality score.



100-point Scale

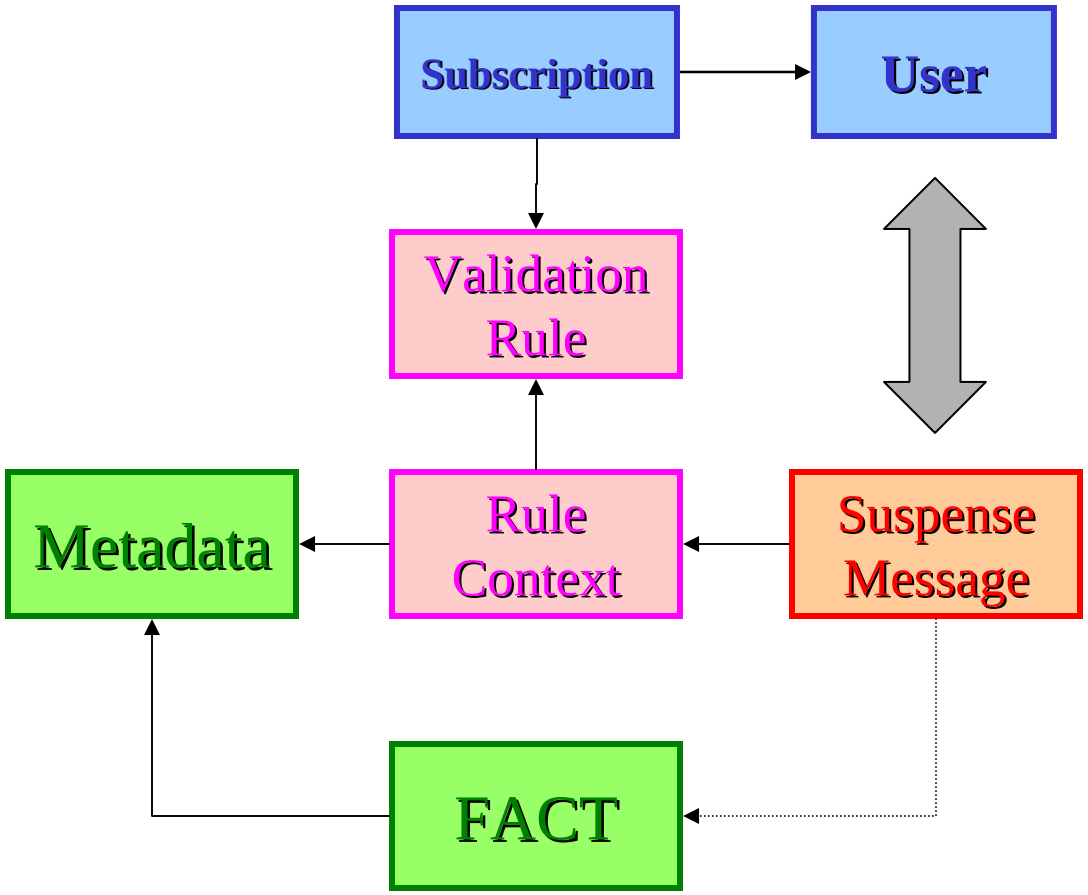


Red (0-30) – Significant problems known. Data must be explicitly requested, otherwise it is automatically omitted.

Yellow (31-70) – Problems suspected to be significant enough that data should not be used for certain analysis or reporting. Included unless requested to be excluded.

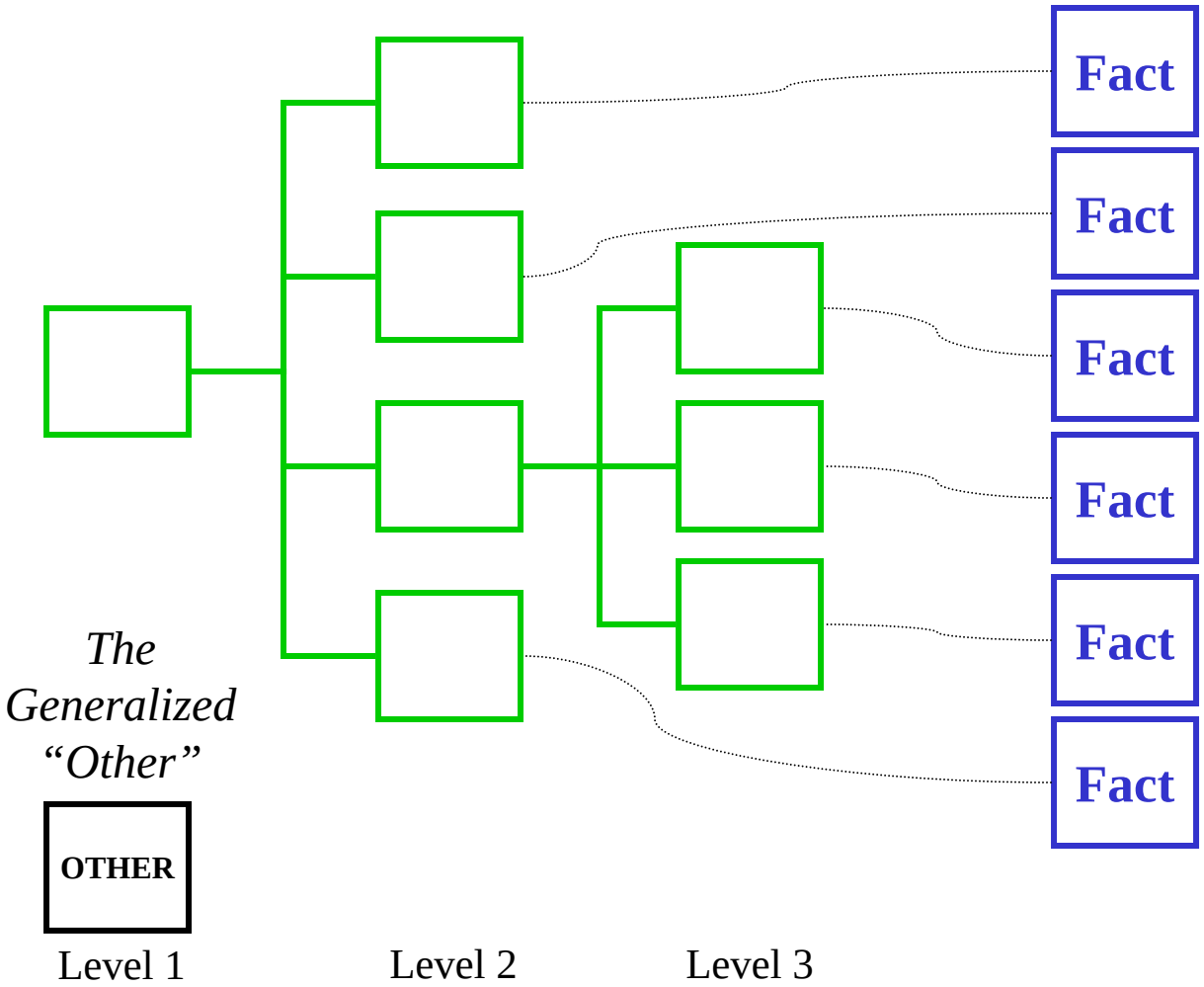
Green (71-100) – No significant problems.

Generalized Notification Mechanism



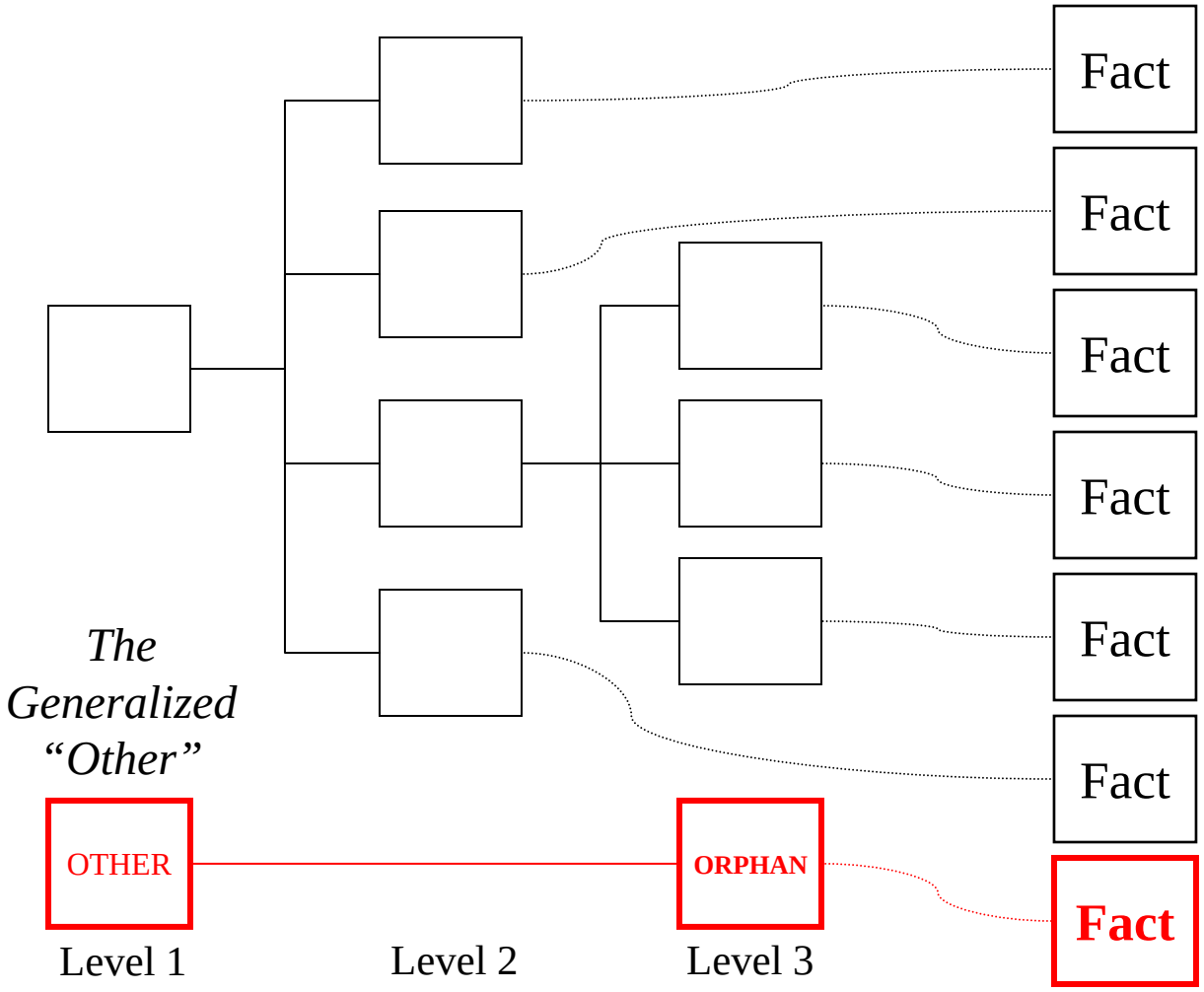
- Users subscribe to messages of interest.
- Subscriptions specify desired communication type and frequency.
- *Expert Workstation* allows for the reviewing and addressing of identified problems.
- Messages are tracked and aged by Data Administration.
- Many messages are eventually discharged due to low priority and lack of resources.

Hierarchy Exception Analysis



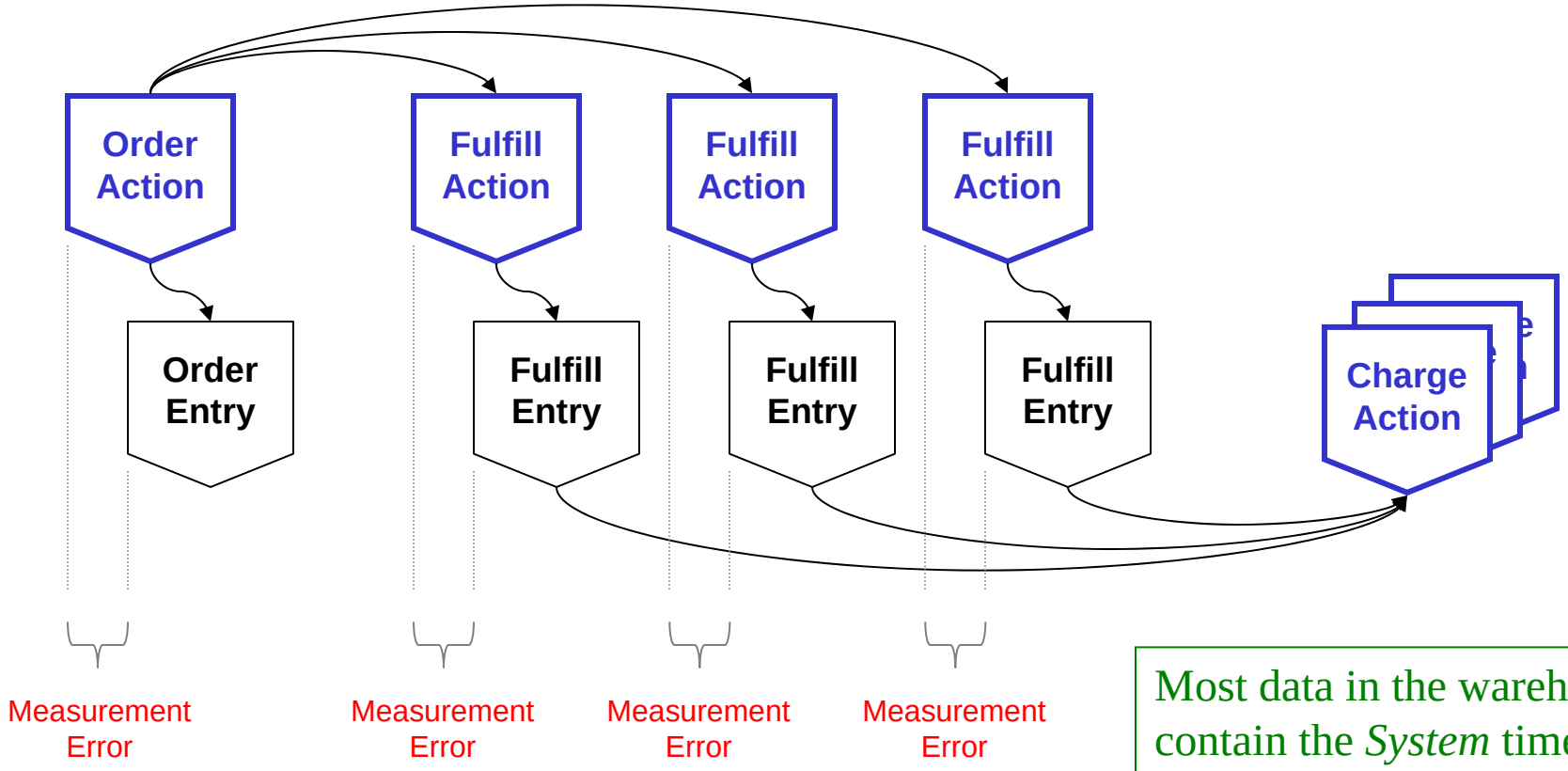
- Hierarchies support navigation of the data along each dimension.
- Every fact should be connected to the hierarchy at some point.
- Aggregations along the hierarchy form new facts with their own metadata.

Hierarchy Exception Analysis



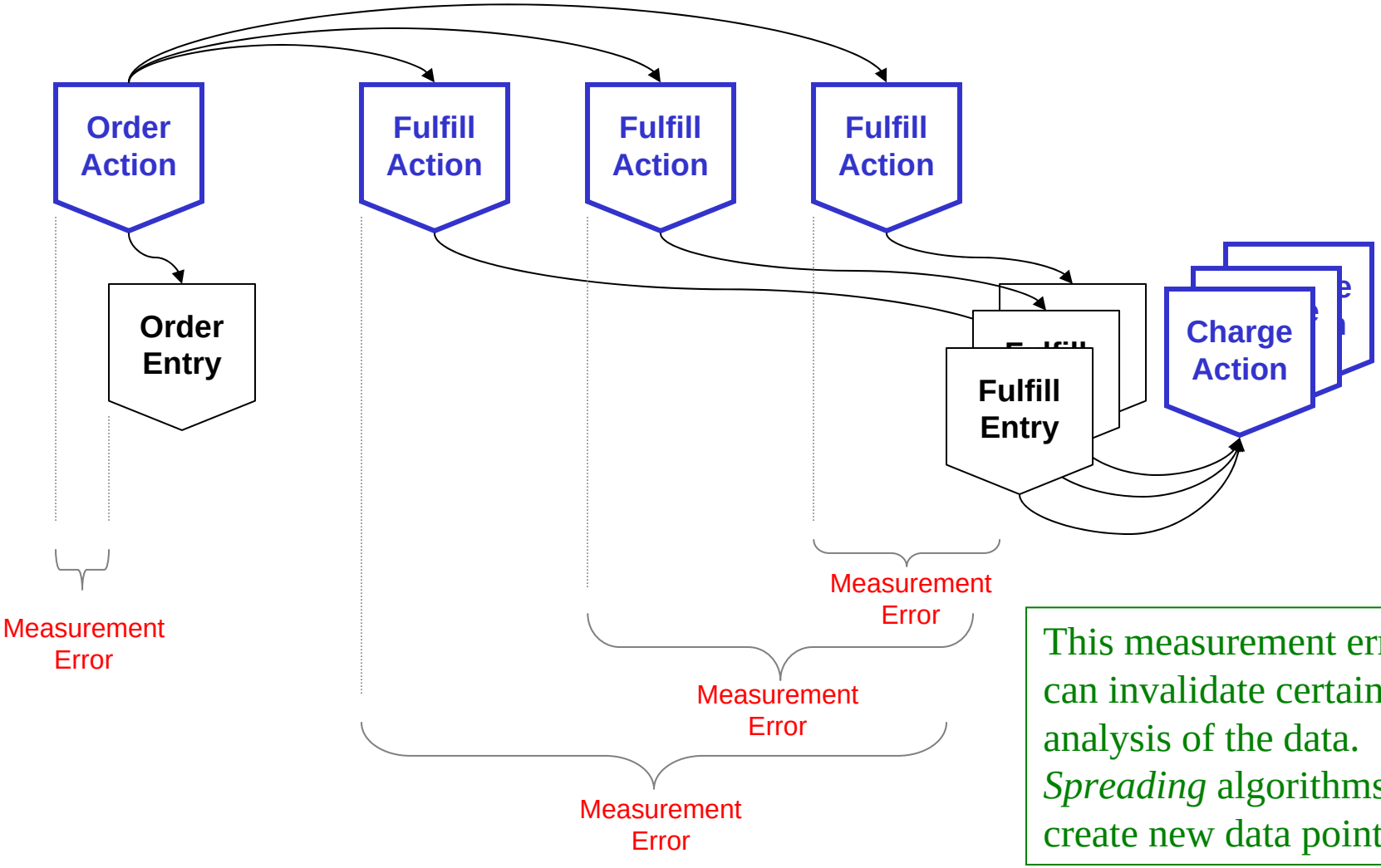
- If a hierarchy becomes incomplete, navigation and aggregation become erroneous.
- When orphans are created in a dimension, we don't know where they belong in the hierarchy.
- Orphaned facts can become lost, unless managed.

Activity Measurement Error



Most data in the warehouse contain the *System* time of the transactions that created them, not the *Actual* time of the clinical actions.

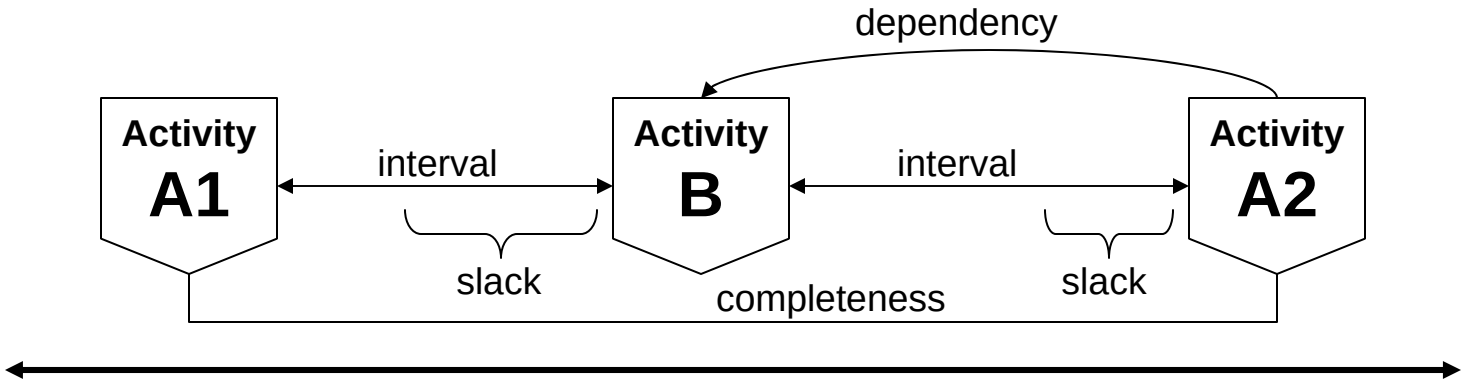
Activity Measurement Error



This measurement error can invalidate certain analysis of the data. *Spreading* algorithms create new data points.

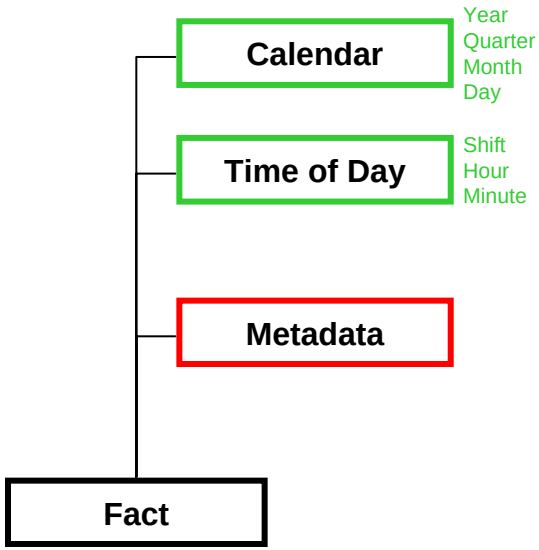
Time-Series Exception Analysis

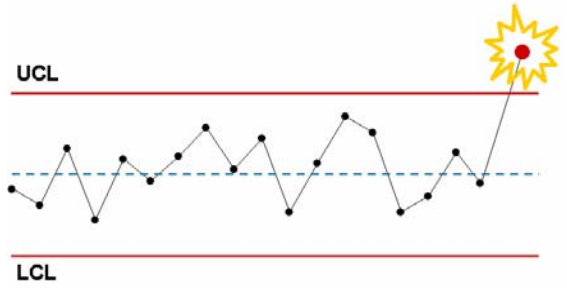
- Completeness Analysis
  – Are data for all the action steps there?
- Dependency Analysis
  – Are required previous events there?
- Interval Analysis
  – How far apart in time were actions accomplished, and did they occur in the expected order?
- Critical Path Analysis
  – Was there slack time that could have allowed things to happen sooner?

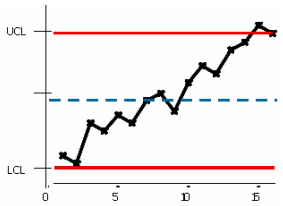


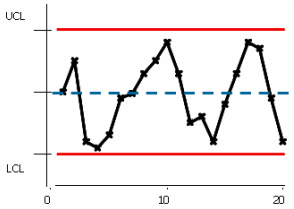
Statistical Process Control

- Control Analysis
  - Is the represented process under statistical control?
  - Are there multiple unexplained populations?
  - Are there outliers in the data?
  - Are there noteworthy trends occurring over time?









Standard Rules

Each FACT value shall:

- be present when required
- be the correct physical data type
- be in the expected unit of measure
- be the correct logical data type
  - Correct data range
  - Correct length
  - Correct codeset value
  - Desirable codeset value

Each DIMENSION reference shall:

- be context accessible
- be effective
- not be expired
- not be orphaned

Each TIMELINE interval shall:

- contain both action nodes
- be in the correct order
- minimize measurement error

Data Quality Informs Query Results

How many female patients with Type 2 Diabetes were treated with Glipizide and Metformin in 2005?

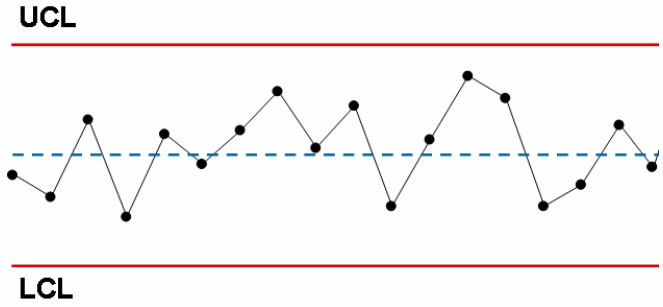
2,319

Data Quality

- 127 inferred medication orders from existence of billing data (included)
- 962 where Type 2 Diabetes was not the primary diagnosis for the encounter (included)
- 28 with multiple encounters in 2005 (included once)
- 845 included Glipizide, but not Metformin (excluded)
- 14 included Metformin but not Glipizide (excluded)
- 22 candidate patients with missing or erroneous gender value (excluded)

Data Quality Informs Query Results

What kind of variability exists between the initial incision times recorded in the Anesthesia and Surgical Manager systems for surgeries within the hospital over the past month?



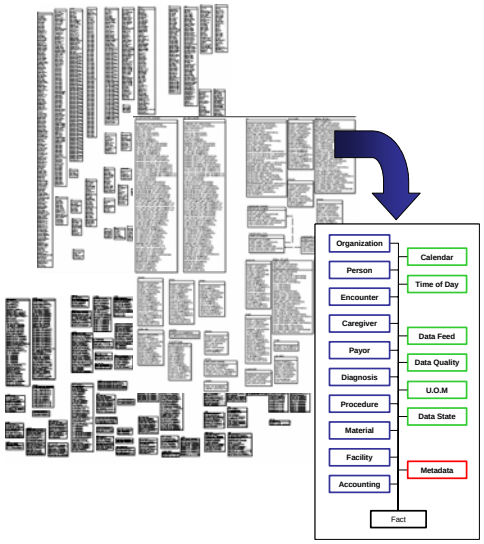
Data Quality

- 2,365 anesthesia-surgery events found (included)
- Anesthesia record not found for 12 surgical procedures (excluded)
- Incision times missing from 6 surgical procedure events (excluded)

Our Goal: Self-Aware Proactive Data

"Declare the past, diagnose the present, foretell the future. As to diseases, make a habit of two things — to help, or at least to do no harm."

Hippocrates, *Epidemics*, Bk. I, Sect. XI



"To wrest from nature the secrets which have perplexed philosophers in all ages, to track to their sources the causes of disease, to correlate the vast stores of knowledge, that they may be quickly available for the prevention and cure of disease – these are our ambitions."

Sir William Osler
(1849-1919)