DATA7001
INTRODUCTION TO DATA SCIENCE

Module 3 Is my Data Fit for Use

Module Topics

- **What is Data Quality**

- Data Exploration
  - Discovering and understanding the quality characteristics of the data through exploratory techniques

- Data Transformation
  - Transforming the data through cleaning, curating, repairing

- Data Enrichment
  - Enriching the data through data imputation and integration

Any Problems?

misleading

inaccurate

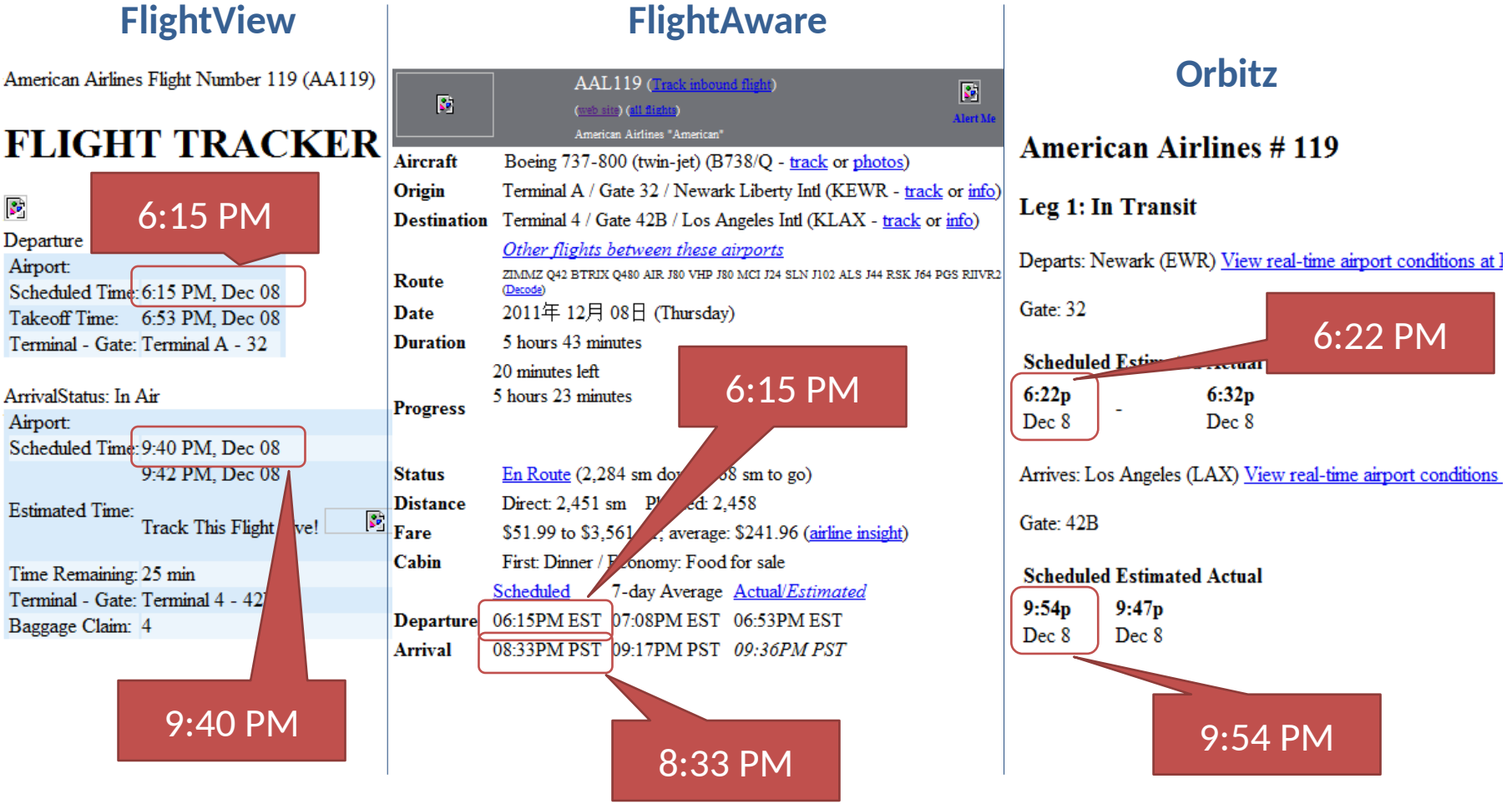
Title	Journal	Vol	Issue	Date	Pages
Insect Motion Detectors Matched to Visual Ecology	Nature	382	6586	1999	Pp 63-66
Information systems success: The quest for the dependent variable	Information Systems Research	3	1	NULL	60-95

incomplete

incomplete

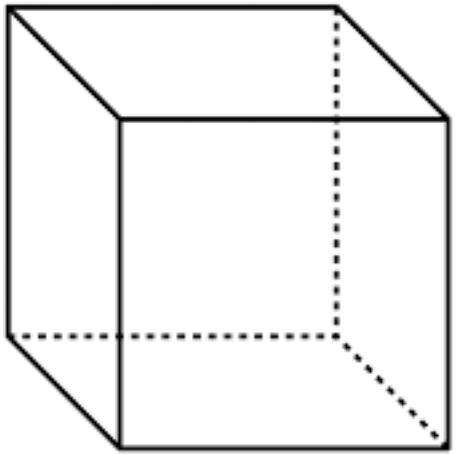
invalid

Any Problems?



Quality Dimensions

- Software (it-CISQ.org)
  - Security
  - Reliability
  - Efficiency
  - Maintainability
  - …?
- Computer System
  - Throughput
  - Response time
  - Availability
  - …?



Dimension:

A central notion in a Quality Domain

"a measureable extent of a particular kind"

Improvements Don’t Always Add Up?

- When improving some dimension of quality, need to consider the possible downstream effects

- Story: Youtube “Feather”

  > https://blog.chriszacharias.com/page-weight-matters

Service quality dimensions [Russell and Taylor 2003]

Dimension	Definition
Time & Timeliness	Customer wait time, On-time completion
Completeness	Customers get all they ask for
Courtesy	Treatment by employees
Consistency	Same level of service for all customers
Accessibility and convenience	Ease of obtaining service
Accuracy	Performed correctly every time
Responsiveness	Reaction to special circumstances or requests

Product quality dimensions [Garvin 1987]

Dimension	Definition
Performance	The product's primary operating characteristic (such as acceleration, braking distance, steering, and handling of an automobile)
Features	The ``bells and whistles'' of a product (such as power option and a tape or CD deck of a car)
Reliability	The probability of a product's surviving over a specified period of time under stated conditions of use
Conformance	The degree to which physical and performance characteristics of a product match pre-established standards
Durability	The amount of use one gets from a product before it physically deteriorates or until replacement is preferable
Serviceability	The speed, courtesy, and competence of repair
Aesthetics	How a product looks, feels, sounds, tastes, or smells
Perceived quality	The subjective assessment of quality resulting from image, advertising, or brand names.

What is Data Quality?

- Degree to which data can be used for its intended purpose

- Degree to which data accurately represents the real-world

Dimensions of Data Quality

Completeness

Missing points on a trajectory

Accuracy

Postcode “4107” rather than “4017”

Freshness

Old telephone number

Consistency

ITEE vs. Information Technology and Electrical Engineering

…

Wand &Wang (1996) Anchoring Data Quality Dimensions in Ontological Foundations. Comm. ACM 39/11. 1996

Shazia Sadiq, Naiem Khodabandehloo Yeganeh and Marta Indulska. 20 years of data quality research: Themes, trends and synergies. Proceedings of: The 22nd Australasian Database Conference (ADC 2011). Perth, WA, Australia, (1-10). 17-20 January 2011.

Dimensions of Data Quality

Data Completeness:

1) A measure of the availability and comprehensiveness of data compared to the total data universe or population of interest. [McGilvray, 2008]
2) A record exists for every Real-World Object or Event the Enterprise needs to know about. [English, 2009]
3) Quality of having all data that existed in the possession of the sender at time the data message was created. [ISO, 2012]
4) Completeness refers to the degree to which values are present in a data collection, as for as an individual datum is concerned, only two situations are possible: Either a value is assigned to the attribute in question or not. In the latter case, null, a special element of an attribute's domain can be assigned as the attribute's value. Depending on whether the attribute is mandatory, optional, or inapplicable, null can mean different things. [Redman, 1997]
5) Determined the extent to which data is not missing. For example, an order is not complete without a price and quantity.[Gatling et al, 2007]

+ 13 more …

A classification of data quality dimensions

User Independent

- Completeness of mandatory attributes
- Completeness of optional attributes
- Precision
- Business rules compliance
- Meta-data compliance
- Uniqueness
- Non-redundancy
- Semantic consistency
- Value consistency
- Format consistency
- Referential integrity

User Dependent

- Completeness of records
- Data volume
- Continuity of data access
- Data maintainability
- Data awareness
- Ease of data access
- Data punctuality
- Data access control
- Data timeliness
- Data freshness
- Accuracy to reference source
- Accuracy to reality
- Standards and regulatory compliance
- Statistical validity
- Source quality
- Objectivity
- Traceability
- Usefulness and relevance
- Understandability
- Appropriate presentation
- Interpretability
- Information value

good enough ≠ good data

Jayawardene, Vimukthi, Sadiq, Shazia and Indulska, Marta (2013). The curse of dimensionality in data quality. In Proceedings of the 24th Australasian Conference on Information Systems (ACIS 2013). Melbourne, VIC, Australia, (1-11). 4-6 December, 2013.

Whose problem is data quality?

- ❑ Management problem
- ❑ IT problem
- ❑ Computational/ Statistical problem
- ❑ All of the above

Ownership

- Who owns the data?
  - Possession, Responsibility, Power, or Control
- Who is liable if data is faulty
  - New legislation such as Data Transparency Act (DATA) - http://www.datacoalition.org
  - Open data initiatives - https://data.qld.gov.au
- Who profits from the value of data assets
  - How do you monetize data?

Consider a large distribution company (LDC) that acquires two other distribution establishments, which will now form part of LDC operations as subsidiaries while maintaining their individual brandings.

Each of the subsidiaries may have its own partner suppliers along with item catalogs.

Consider the case that there is a large overlap of business with a particular supplier group, which may put LDC into a favorable bargaining position to negotiate significant discounts.

However, data differences do not reveal this position, and thus directly impact on the bottom line for LDC

1. Create a reference (synonym) table for suppliers
2. Load supplier data from all subsidiaries into the reference table
3. Use *matching* techniques to identify potential overlaps
4. Extract a master table for suppliers – represents a single version of truth
5. Retain original representations – represent multiple versions of truth
6. Allow access for subsidiaries to reference master data in all new (or update) transactions involving supplier data
7. Ensure data managers are accountable for continued master data checks
8. Introduce a periodic monitoring scheme

Consider a large distribution company (LDC) that acquires two other distribution establishments, which will now form part of LDC operations as subsidiaries while maintaining their individual brandings.

Each of the subsidiaries may have its own partner suppliers along with item catalogs.

Consider the case that there is a large overlap of business with a particular supplier group, which may put LDC into a favorable bargaining position to negotiate significant discounts.

However, data differences do not reveal this position, and thus directly impact on the bottom line for LDC

Algorithms and Methods

1. Create a reference (synonym) table for suppliers
2. Load supplier data from all subsidiaries into the reference table
3. Use *matching* techniques to identify potential overlaps
4. Extra master table for suppliers – a single version of truth
5. inal representations – multiple versions of truth
6. Allow access for subsidiaries to reference master data in all new (or update) transactions involving supplier data
7. Ensure data managers are accountable for continued master data checks
8. Introduce a periodic monitoring scheme

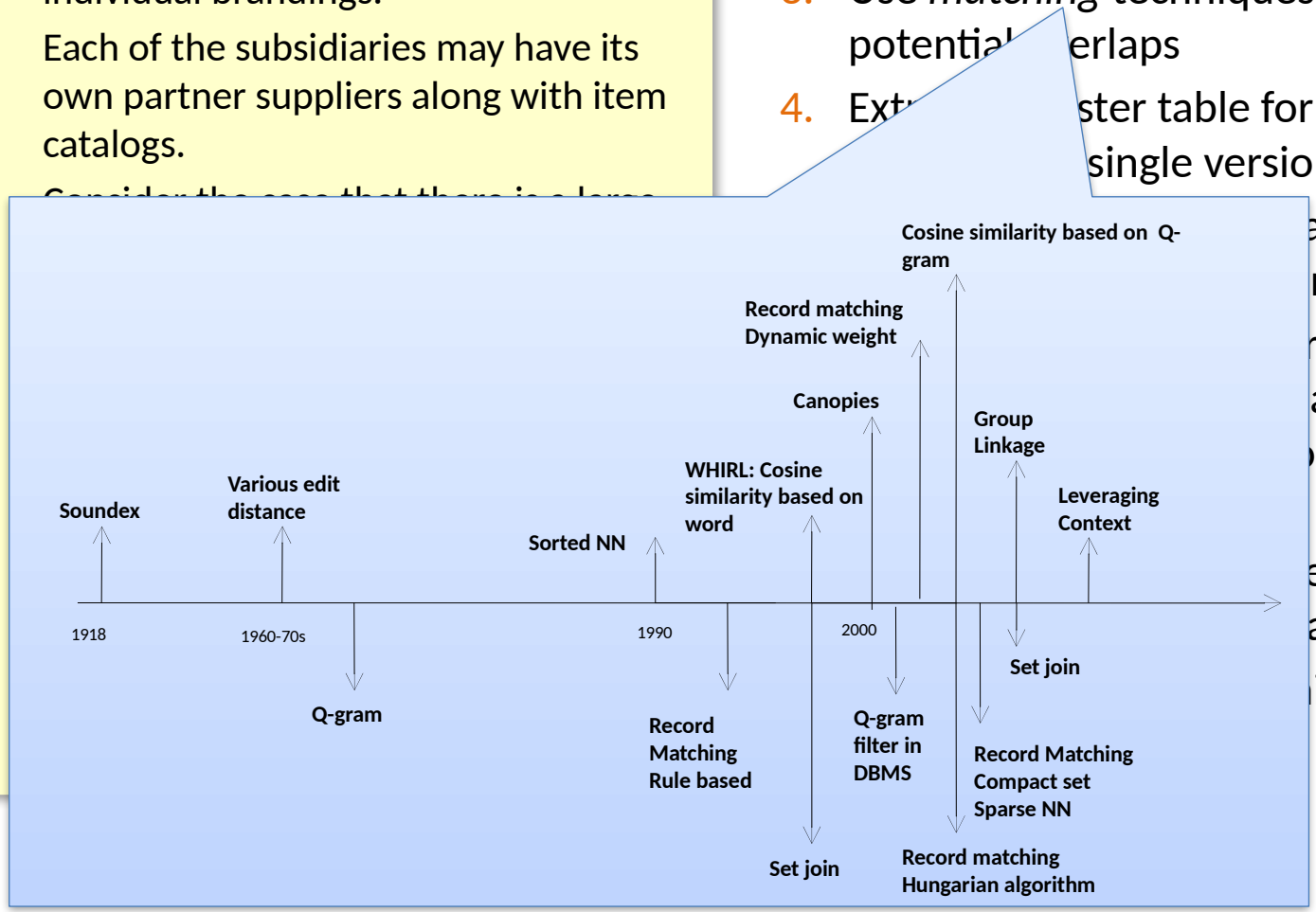
Consider a large distribution company (LDC) that acquires two other distribution establishments, which will now form part of LDC operations as subsidiaries while maintaining their individual brandings.

Each of the subsidiaries may have its own partner suppliers along with item catalogs.

1. Create a reference (synonym) table for suppliers
2. Load supplier data from all subsidiaries into the reference table
3. Use *matching* techniques to identify potentia[illegible]erlaps
4. Ext[illegible]ster table for suppliers – [illegible]single version of truth

[illegible]ations –
[illegible]ns of truth
[illegible]ries to
[illegible]all new (or
[illegible]lving supplier

[illegible]e accountable
[illegible]a checks
[illegible]itoring scheme



Consider a large distribution company (LDC) that acquires two other distribution establishments, which will now form part of LDC operations as subsidiaries while maintaining their individual brandings.

Each of the subsidiaries may have its own partner suppliers along with item catalogs.

Consider the case that there is a large overlap of business with a particular supplier group, which may put LDC into a favorable bargaining position to negotiate significant discounts.

However, data differences do not reveal this position, and thus directly impact on the bottom line for LDC

1. Create a reference (synonym) table for suppliers
2. Load supplier data from all subsidiaries into the reference table
3. Use *matching* techniques to identify potential overlaps
4. Extract a master table for suppliers – represents a single version of truth
5. Retain original representations – represent multiple versions of truth
6. Allow access for subsidiaries to reference master data in all new (or update) transactions involving supplier data
7. Ensure data managers for continued master d
8. Introduce a periodic monitoring scheme

Technology Infrastructure

Consider a large distribution company (LDC) that acquires two other distribution establishments, which will now form part of LDC operations as subsidiaries while maintaining their individual brandings.

Each of the subsidiaries may have its own partner suppliers along with item catalogs.

Consider the case that there is a large overlap of business with a particular supplier group, which may put LDC into a favorable bargaining position to negotiate significant discounts.

However, data differences do not reveal this position, and thus directly impact on the bottom line for LDC

1. Create a reference (synonym) table for suppliers
2. Load supplier data from all subsidiaries into the reference table
3. Use *matching* techniques to identify potential overlaps
4. Extract a master table for suppliers – represents a single version of truth
5. Retain original representations – represent multiple versions of truth
6. Allow access for su… reference master … update) transactio… …ier data
7. Ensure data managers are accountable for continued master data checks
8. Introduce a periodic monitoring scheme

Data Governance

Total Data Quality

Organizational	Development of data quality objectives for the organization and **strategies** to establish the people, processes, policies, and standards required to manage and ensure the data quality objectives are met
Architectural	The **technology** landscape required to deploy developed data quality management processes, standards and policies
Computational	Effective and efficient **methods** & techniques required to meet data quality objectives

Develop the capacity to understand how the quality of data affects the quality of the insight we derive from it

J. Stoyanovich; and F. Suchanek. *The elephant in the room: getting value from Big Data*. Available: http://wp.sigmod.org/?p=1519 (accessed on 24/2/2015)

Data Quality

- Poor quality data costs …
  - "$3 trillion to US government"
  - "$611 billion to US business for customer data alone"

You have to start with a very basic idea: **Data is super messy**, and data cleanup will always be literally 80% of the work. In other words, data is the problem.

"If you take something like LinkedIn in the early days, let's say, there were 4,000 variations of how people said they worked at IBM — IBM, IBM Research, Software Engineer, all the abbreviations, etc.," says Patil.

First US Chief Data Scientist at the White House

http://www.datablueprint.com/cost-of-bad-data-by-the-numbers/

How can you accelerate the time to value from big data in the presence of data quality problems?

Find out if your data is fit for use

- Data Exploration
  - Discovering and understanding the quality characteristics of the data through exploratory techniques
- Data Transformation
  - Transforming the data through cleaning, curating, repairing
- Data Enrichment
  - Enriching the data through data imputation and integration