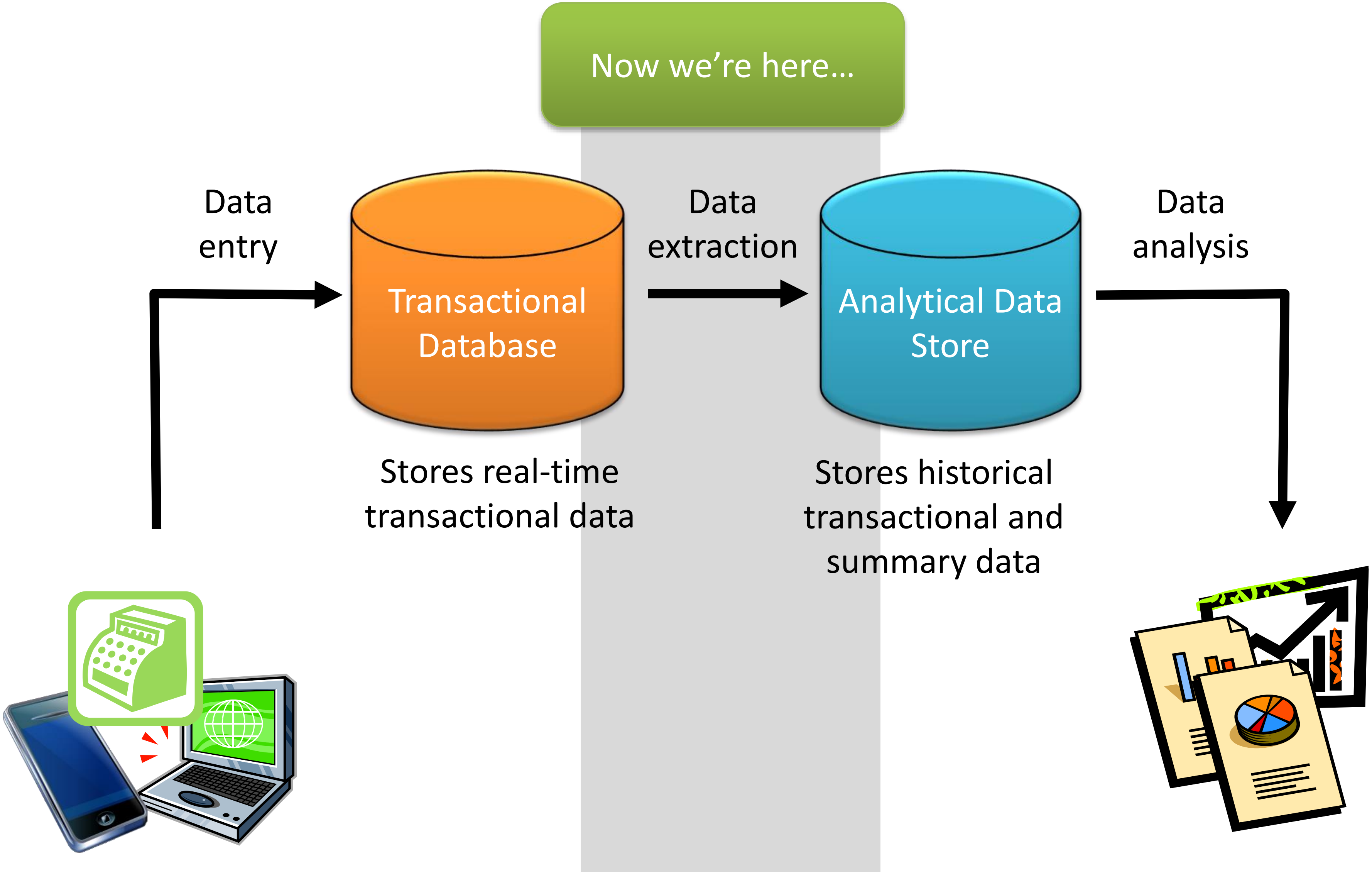
MIS2502:
Data Analytics
Extract, Transform, Load

Alvin Zuyin Zheng
zheng@temple.edu
http://community.mis.temple.edu/zuyinzheng/

Where we are…



Now we're here…

Data entry

Transactional Database

Stores real-time transactional data

Data extraction

Analytical Data Store

Stores historical transactional and summary data

Data analysis

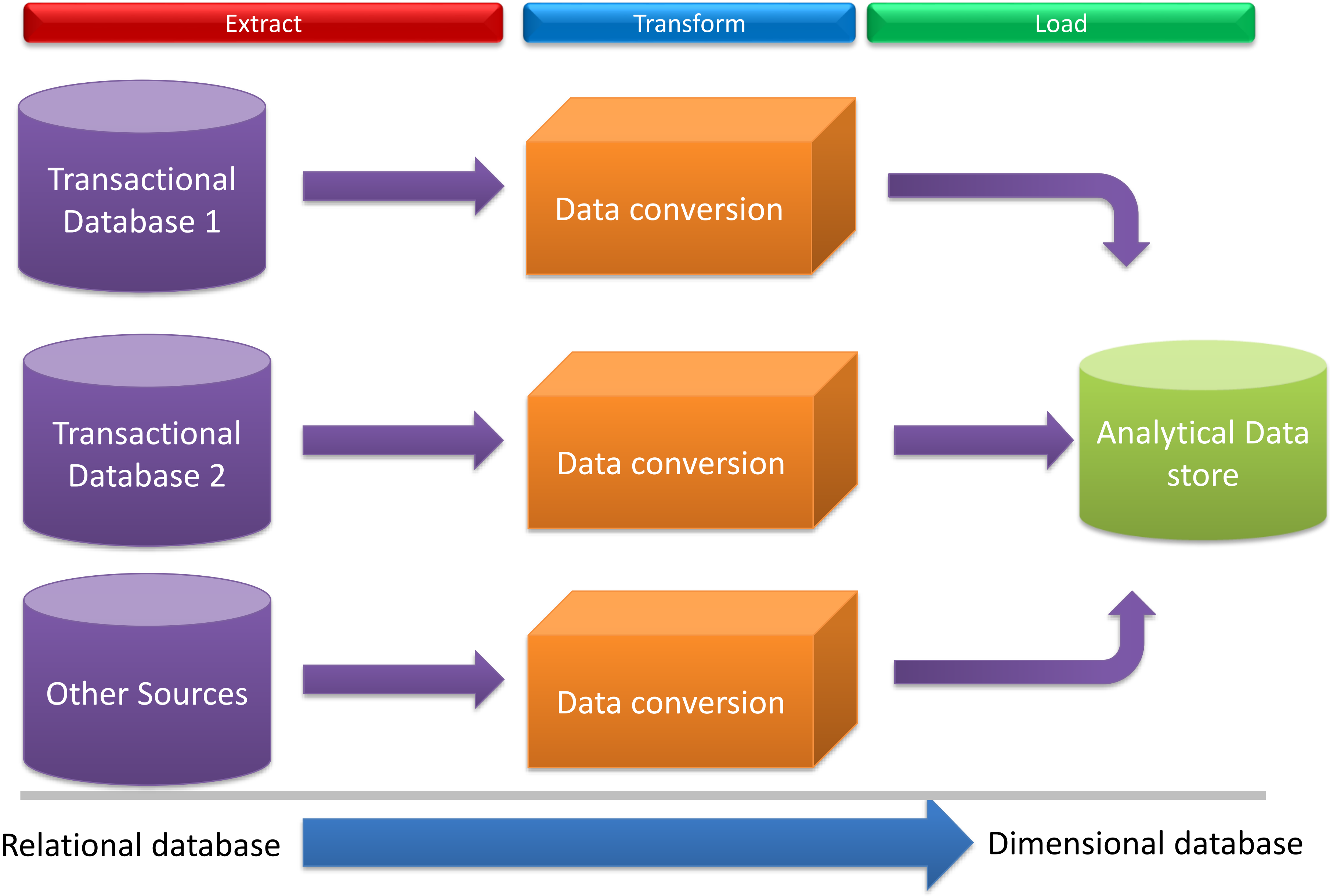
Extract, Transform, Load (ETL)

Extract data from the transactional database

Transform data into an analysis-ready format

Load it into the analytical data store

The Actual Process



ETL’s Not That Easy!

Data Consistency

- What if the data is in different formats?

e.g., **2017-03-01**

vs.

March 1st, 2017

Data Quality

- How do we know it’s correct?
- What if there is missing data?
- What if the data we need isn’t there?

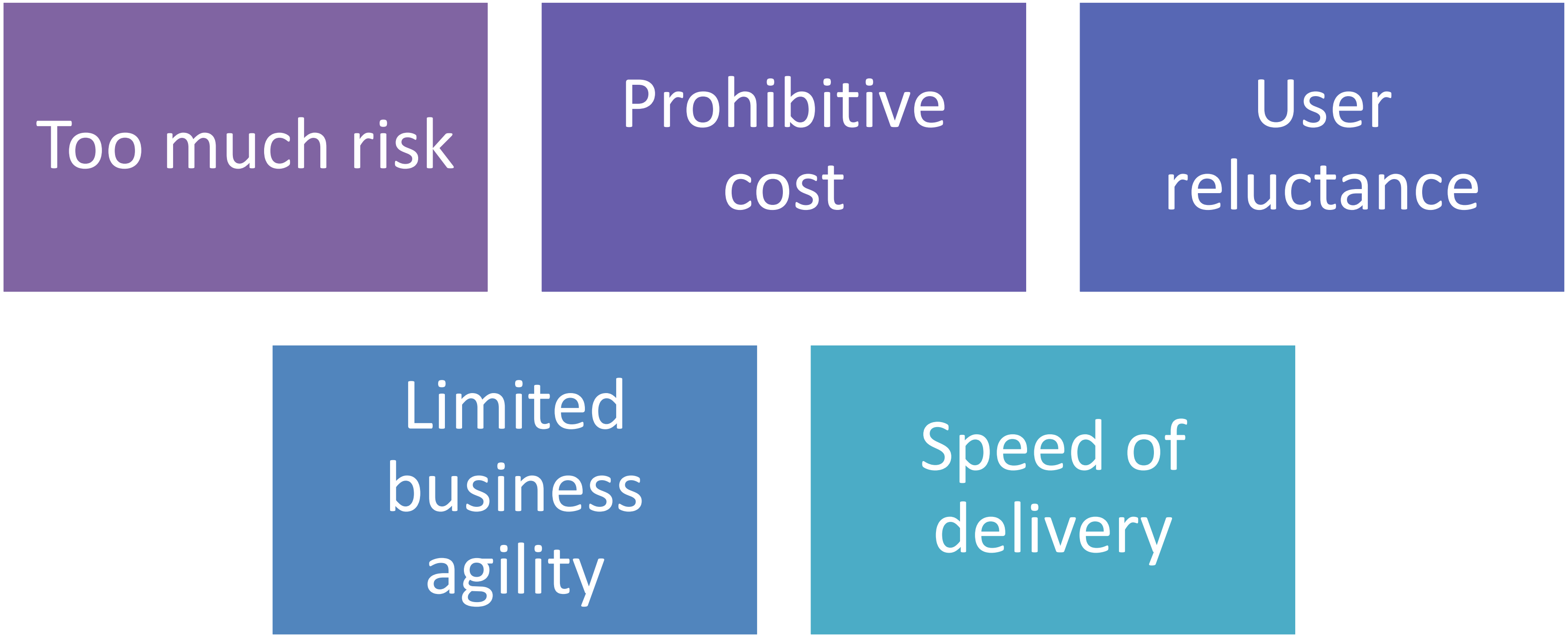
Data Consistency:
The Problem with Legacy Systems

- An IT infrastructure evolves over time
- Systems are created and acquired by different people using different specifications

This can happen through:

- Changes in management
- Mergers & Acquisitions
- Externally mandated standards
- Generally poor planning

Why Not Replacing Legacy Systems?



- Too much risk
- Prohibitive cost
- User reluctance
- Limited business agility
- Speed of delivery

https://www.onbase.com/~/media/Files/hyland/whitepaper/wp_trouble-with-legacy-systems.pdf

Problems with Data Consistency

The same data element stored in different **formats**

- Social Security number (123-45-6789 versus 123456789)
- Date (10/9/2015 versus 9/10/2015)

Redundant data across the organization

- Customer record maintained by accounts receivable and marketing

Different **naming** conventions

- "Management Information Systems" versus "MIS" verus "Man. Info. Sys."

Different **unique identifiers** used

- AccessNet account versus Temple ID

What are the problems with each of these

?

What’s the big deal?

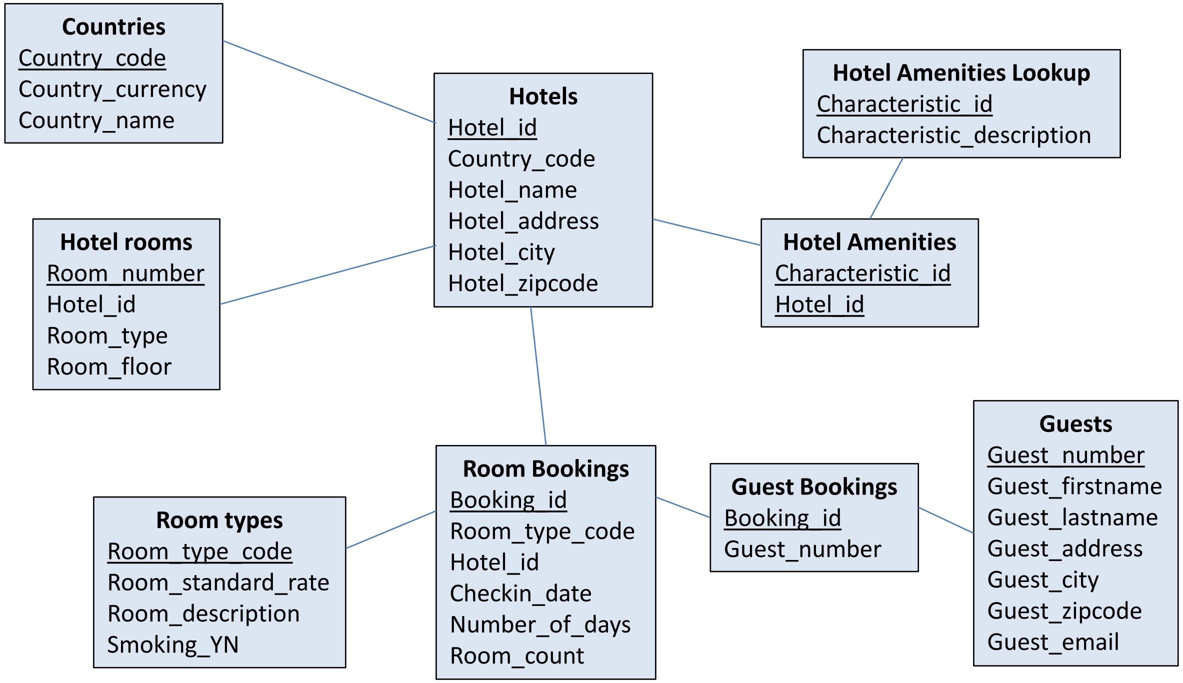
This is a fundamental problem for creating the analytical data store

We often need to combine information from several transactional databases

How do we know if we’re talking about the same customer or product?

Now think about this scenario

Hotel Reservation Database



Countries
Country_code
Country_currency
Country_name

Hotels
Hotel_id
Country_code
Hotel_name
Hotel_address
Hotel_city
Hotel_zipcode

Hotel Amenities Lookup
Characteristic_id
Characteristic_description

Hotel rooms
Room_number
Hotel_id
Room_type
Room_floor

Hotel Amenities
Characteristic_id
Hotel_id

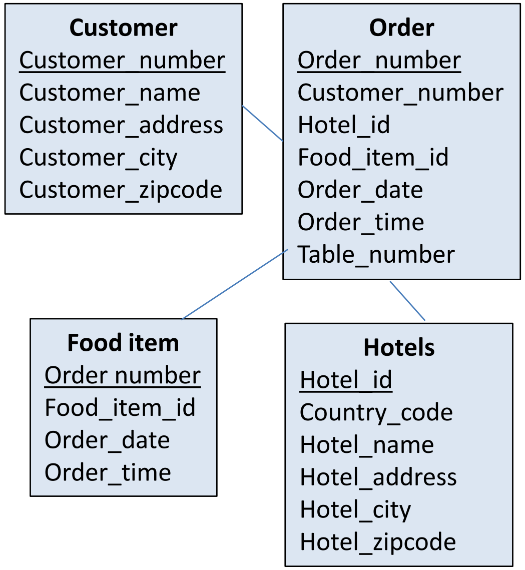
Guests
Guest_number
Guest_firstname
Guest_lastname
Guest_address
Guest_city
Guest_zipcode
Guest_email

Room Bookings
Booking_id
Room_type_code
Hotel_id
Checkin_date
Number_of_days
Room_count

Guest Bookings
Booking_id
Guest_number

Room types
Room_type_code
Room_standard_rate
Room_description
Smoking_YN

Café Database



Customer
Customer_number
Customer_name
Customer_address
Customer_city
Customer_zipcode

Order
Order_number
Customer_number
Hotel_id
Food_item_id
Order_date
Order_time
Table_number

Food item
Order number
Food_item_id
Order_date
Order_time

Hotels
Hotel_id
Country_code
Hotel_name
Hotel_address
Hotel_city
Hotel_zipcode

What are the differences between a “guest” and a “customer”?

Is there any way to know if a customer of the café is staying at the hotel?

Solution: “Single view” of data

- The entire organization understands a unit of data in the same way
- It’s both a business goal and a technology goal

but it’s really more this…

…than this

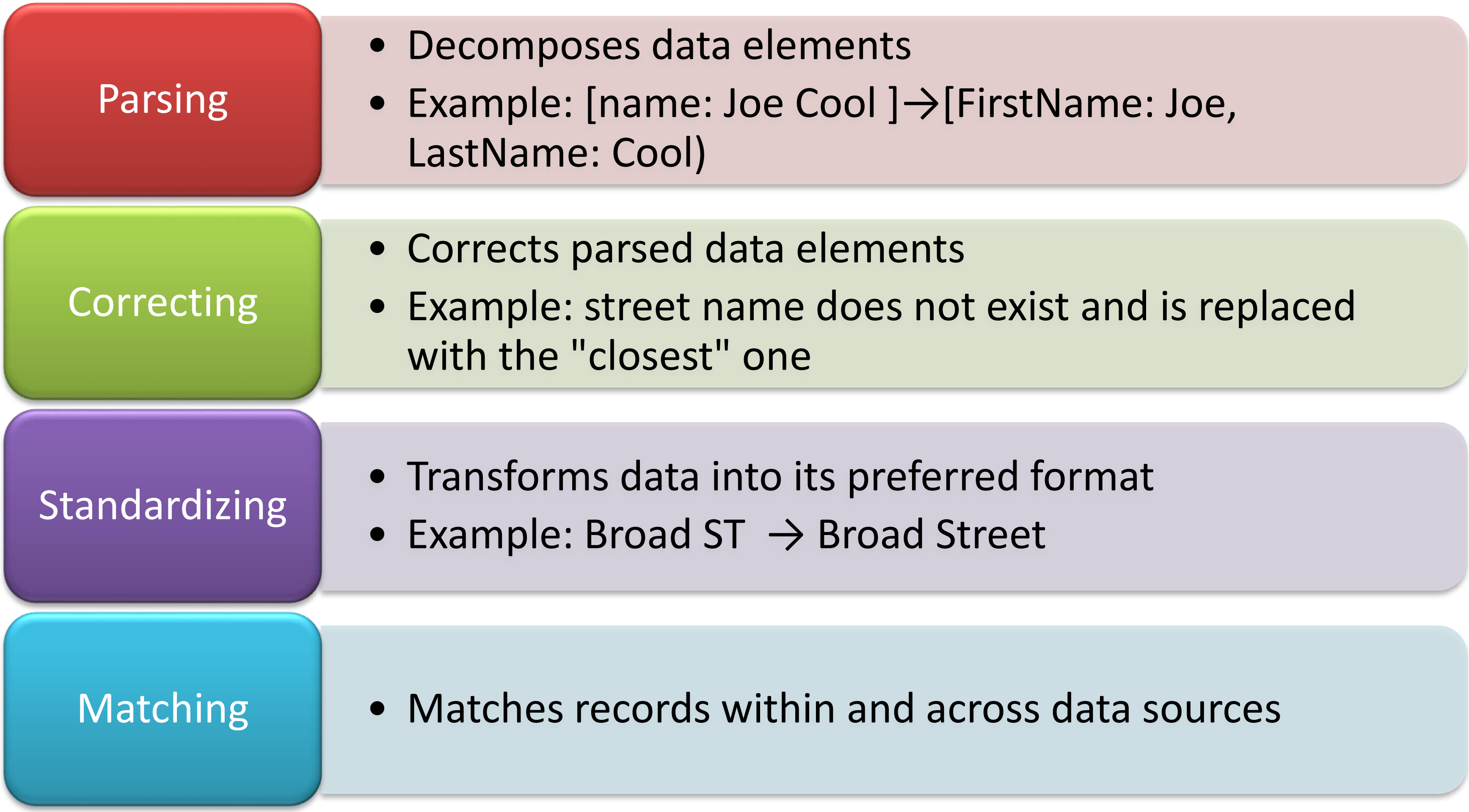
Organizational issues

Why might there be resistance to data standardization?

Is it an option to just “fix” the transactional databases?

If two data elements conflict, who’s standard “wins?”

Data Transformation Steps



Parsing

- Decomposes data elements
- Example: [name: Joe Cool]→[FirstName: Joe, LastName: Cool)

Correcting

- Corrects parsed data elements
- Example: street name does not exist and is replaced with the "closest" one

Standardizing

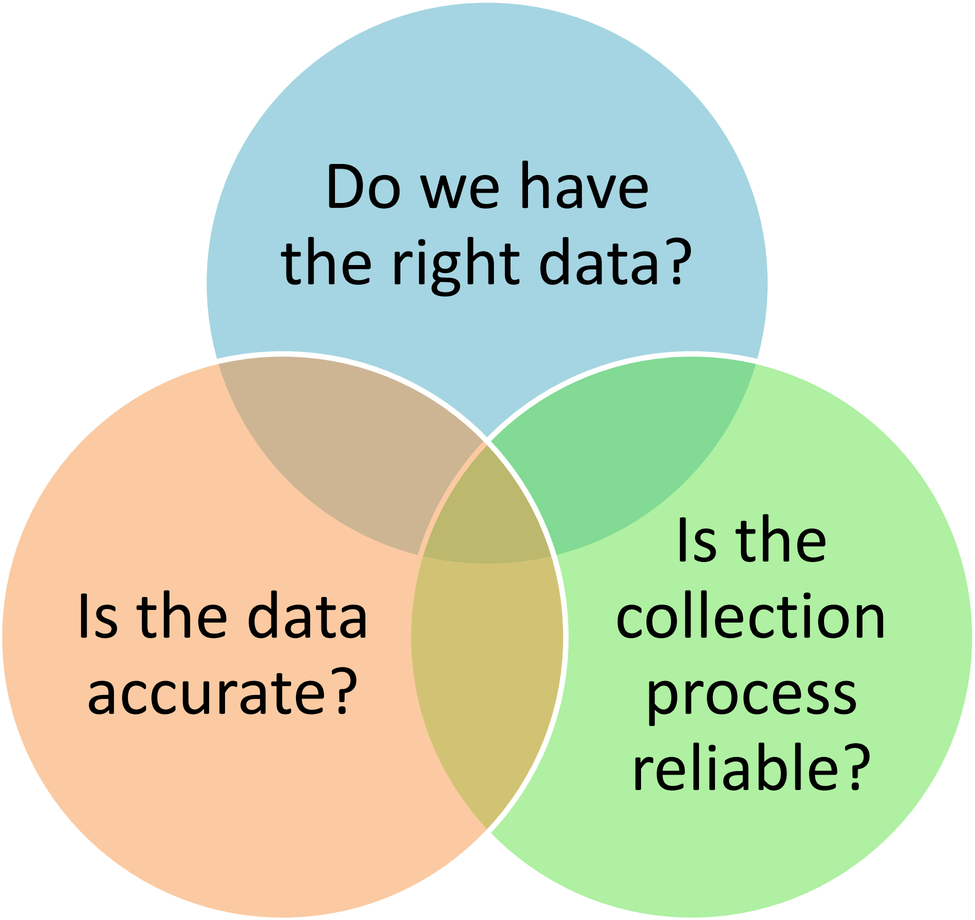
- Transforms data into its preferred format
- Example: Broad ST → Broad Street

Matching

- Matches records within and across data sources

Data Quality

The degree to which the data reflects the actual environment



Do we have the right data?

Is the data accurate?

Is the collection process reliable?

Finding the right data

Choose data consistent with the goals of analysis

Verify that the data really measures what it claims to measure

Include the analysts in the design process

Adapted from http://www2.ed.gov/about/offices/list/os/technology/plan/2004/ site/docs_and_pdf/Data_Quality_Audits_from_ESP_Solutions_Group.pdf

Ensuring accuracy

Know where the data comes from

Manual verification through sampling

Use of knowledge experts

Verify calculations for derived measures

Adapted from http://www2.ed.gov/about/offices/list/os/technology/plan/2004/site/docs_and_pdf/Data_Quality_Audits_from_ESP_Solutions_Group.pdf

Reliability of the collection process

Build fault tolerance into the process

Periodically run reports, check logs, and verify results

Keep up with (and communicate) changes

Adapted from http://www2.ed.gov/about/offices/list/os/technology/plan/2004/ site/docs_and_pdf/Data_Quality_Audits_from_ESP_Solutions_Group.pdf

Summary

- What is ETL? Why is it important?
  - Data consistency
  - Data quality
- Explain the purpose of each component (Extract, Transform, Load)

ETL Assignment

- We will perform the ETL process on an Excel workbook

- You will be:
  - **Extracting** the data from source worksheets.
  - **Transforming** the data using Excel formulas.
  - **Loading** the data into a new worksheet that contains a single set of combined data.