UNIVERSITY OF CALIFORNIA

Promoting Accuracy Through Data Quality: The UC Data Validation Framework

University of California Office of the President

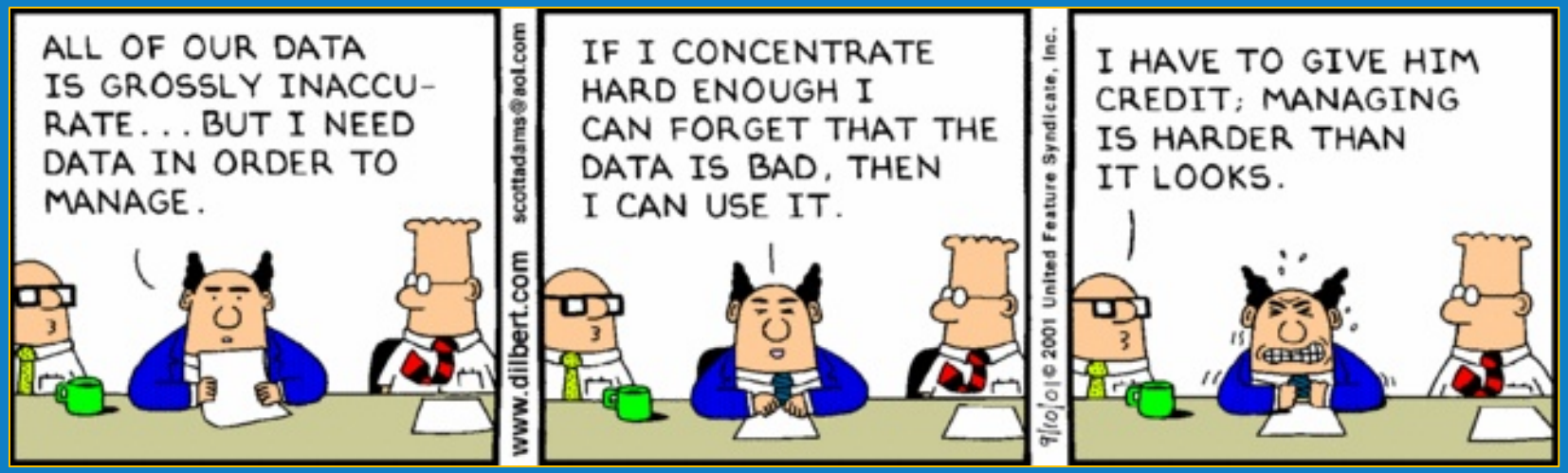
OFFICE OF INSTITUTIONAL RESEARCH & ACADEMIC PLANNING [IRAP]

CAIR 2016 Conference

OLA POPOOLA – Director: Reporting & Analytics

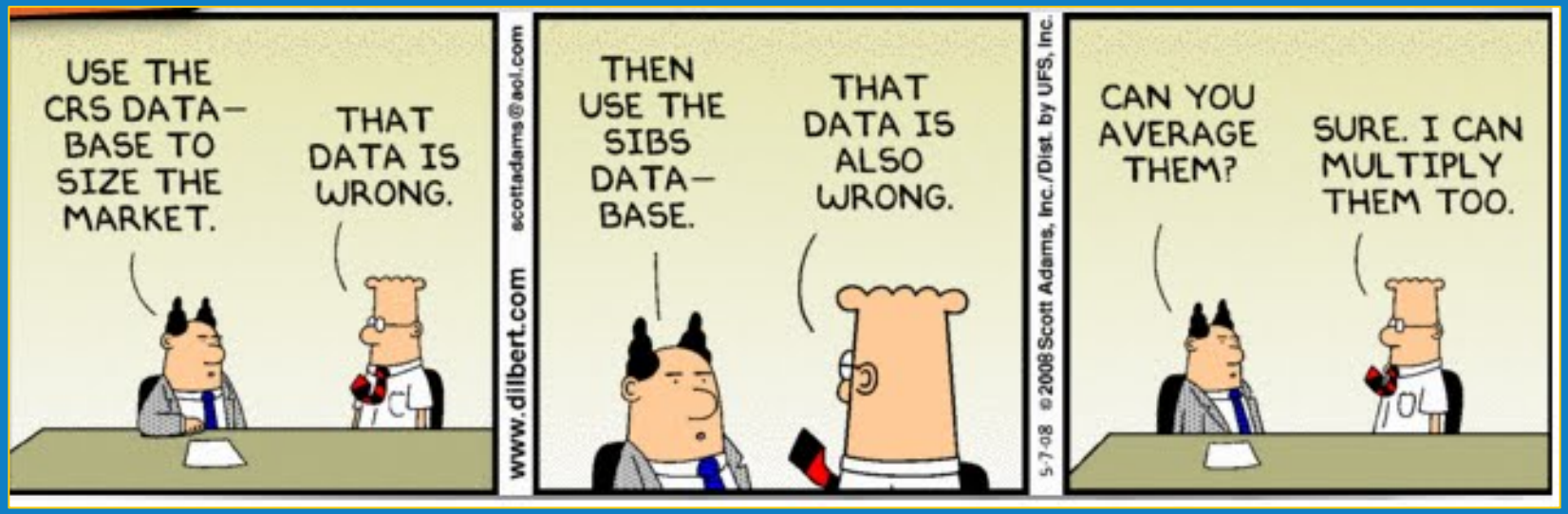
October 17, 2016

Dealing with bad data?



1

2



Agenda

- Desired Presentation Outcomes
- Data Quality
  - Attributes of Data Quality
  - Causes & Costs of Poor Quality Data
- The UC Data Validation Framework
- Creating Your Own Data Quality Management Program
- Final Thoughts

Presentation Outcomes

Desired Presentation Outcomes

- A better understanding of data quality from an IR perspective
- Exposure to data quality principles, methods and techniques that enable continuous improvement in data quality
- How to conduct simple data quality audits by implementing a successful Data Quality Program (DQP)


data
Quality

What is Data Quality?

Definition 1

The state of completeness, validity, consistency, timeliness and accuracy that makes data appropriate for a *specific use*. (Government of British Columbia)

Definition 2

The quality of a particular dataset or record is to describe the *fitness* of that dataset or record for a particular use that one may have in mind for the data. (Chrisman, 1991)

Attributes of Data Quality

Accurate	Complete	Flexible
Timely	Consistent	Available

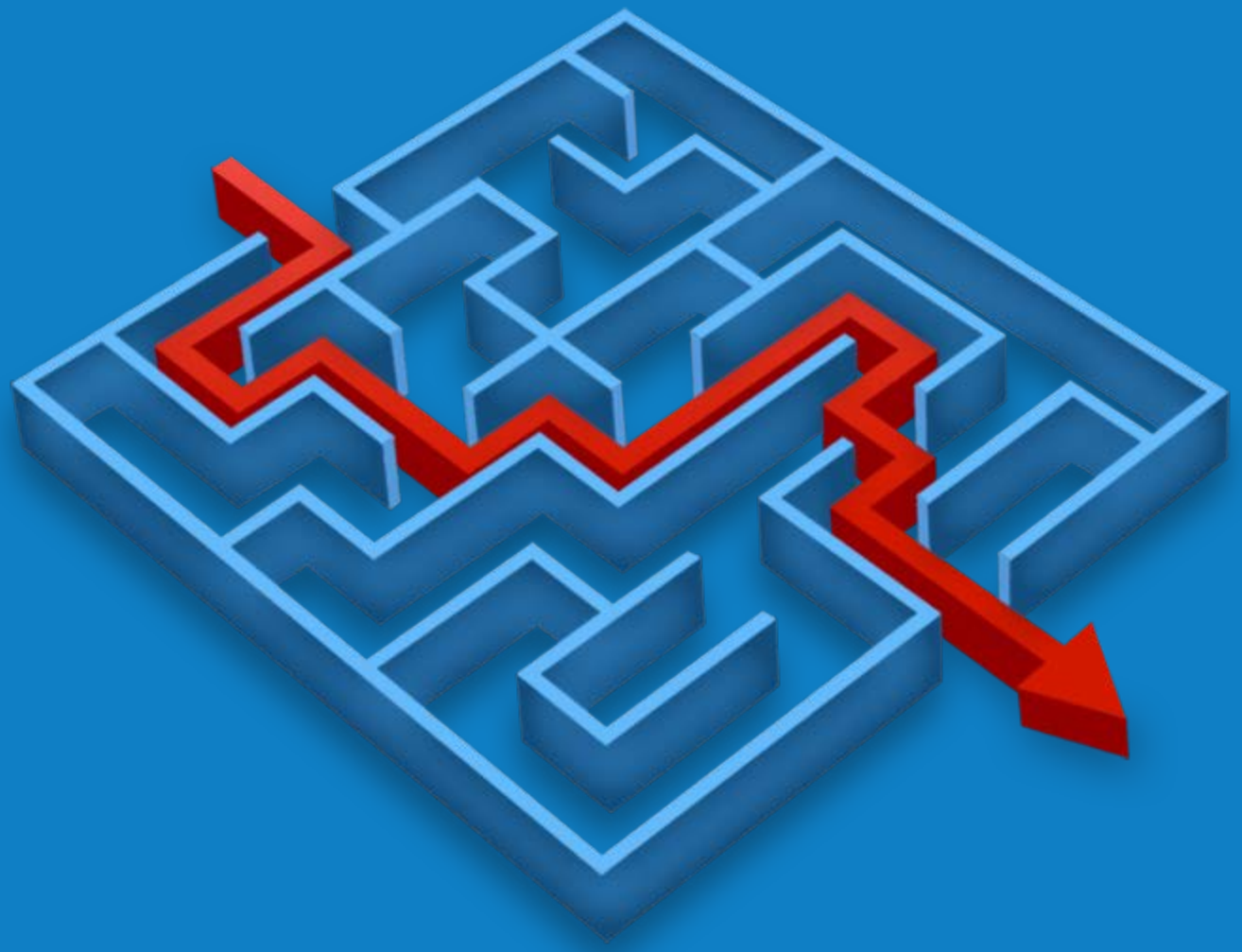
Causes of Poor Quality Data

- Lack of data governance
- User errors – manual data entry
- Lack of identified “authoritative” data sources
- Complex IT infrastructure
- Bad business processes
- Silo-driven solutions
- Multiple disconnected processes
- Tactical initiatives to “re-solve” data accuracy rather than understanding and addressing root cause

The Cost of Poor Data Quality

- Wasted revenue - $3.1 trillion in US alone (2016)
- Mistrust
- Bad or delayed decisions
- Impacted funding
- Constant rework
- Missed opportunities

The UC Implementation



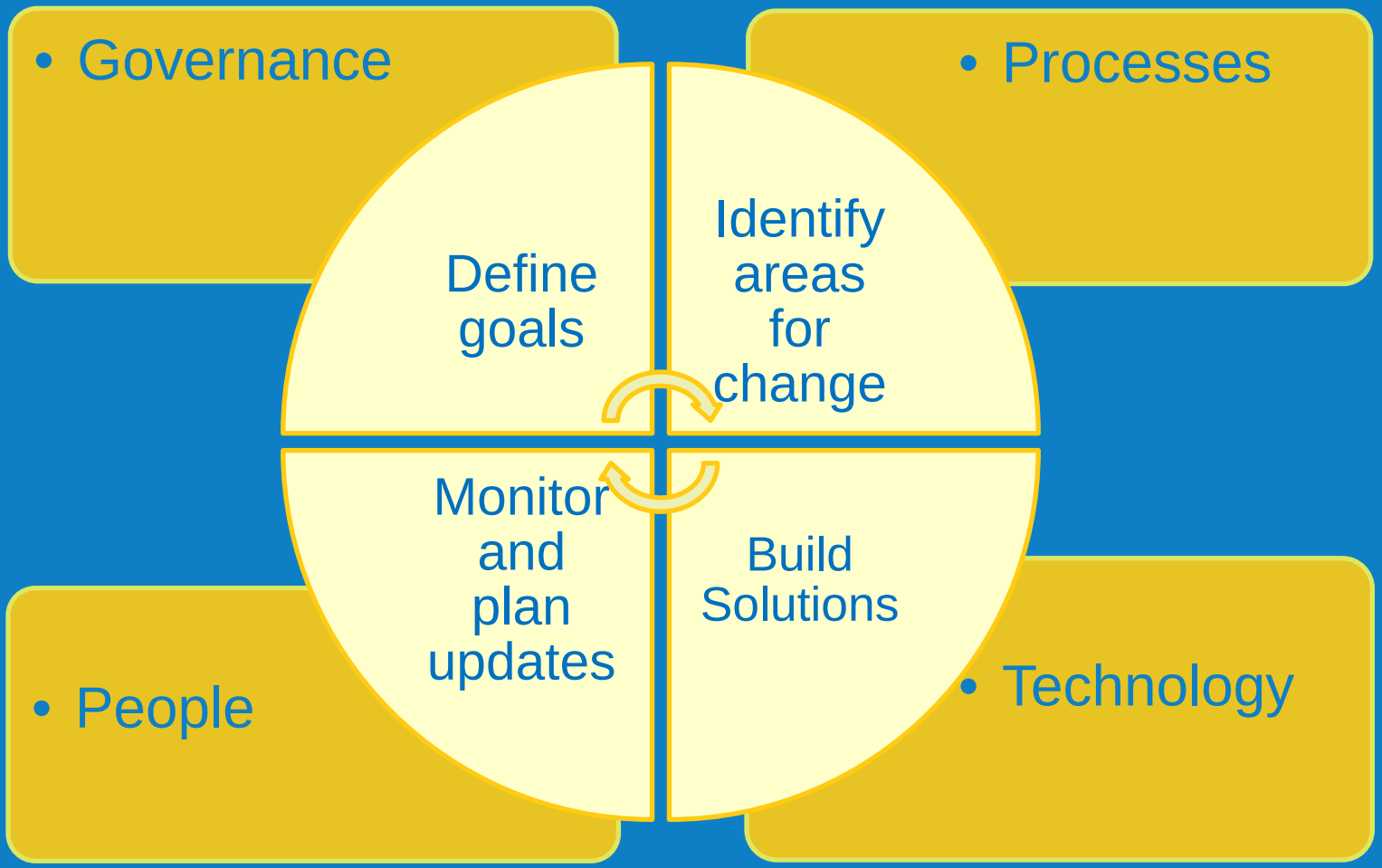
The UC Story

- Challenge associated with data submission from 10 different campus locations, central office, three laboratories and ANR with diverse transactional systems
- Implementation of a new data warehouse called for an extensive review of data quality processes
- Selection of a data quality methodology that involved business practice review and change.

UC Quality Program Guidelines

- UC Applicable
  - For UC business; based on user needs
- Flexible
  - Adaptable to evolving data content areas
- Scalable
  - Could be expanded or reduced in scale
  - Could be deployed across multiple UC locations
- Prudent
  - Minimal implementation costs
- Complementary
  - Compatible with UC Standards

Elements of UC DQM



UC Data Infrastructure

Input Files

Data Processing

Reporting & Analytics

ETL

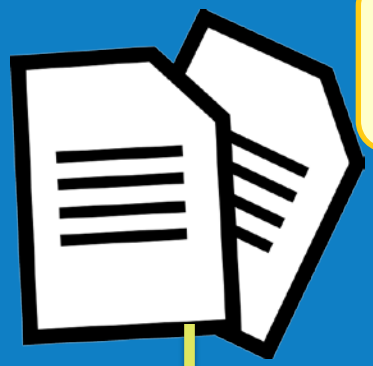
Staging Layer

ETL

Reporting Layer

Data Marts

UC Data Validation Framework



Standard input files

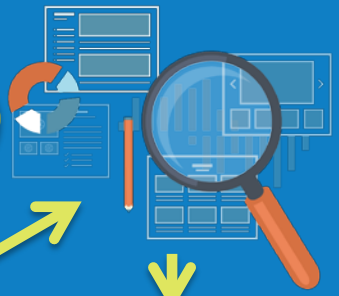
N1

N6

Review data audit reports

Review data validation reports

N3



N5

Reporting Layer

N2

N4

Data Staging Layer

Data Collection & File Specs…

- File specifications: data collection instrument
- Proper data collection instrumental to integrity of research
- A good file specification:
  - Clarifies how you expect all institutions to submit their data
  - Clarifies the length, format, error levels and valid values
  - Has an accompanying overview and file characteristics that contains:
    - File submission schedule
    - File physical characteristics
    - Any special conventions
  - Has an accompanying code book

Error Groups

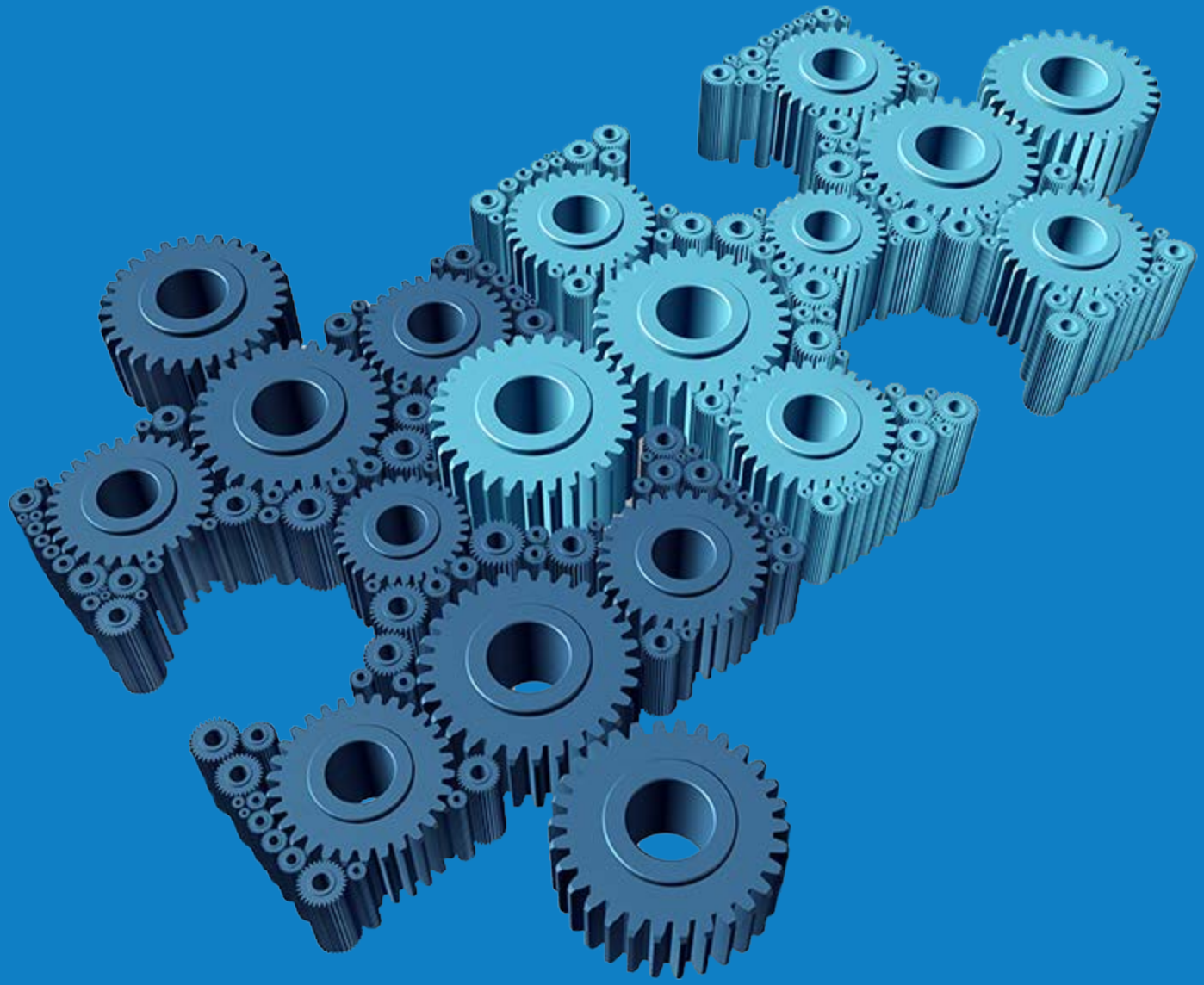
- Error Framework
  - Database Tables
- Rejected Files (R)
  - Header Record Type
- Severe Errors (S)
  - Invalid Campus Code
  - Invalid Student ID
- Element Errors (E)
  - Invalid Sex Code
- Group Errors (G)
  - Campus-College-Major combinations

Our Toolset

UC Data Quality Toolset

- Atlassian JIRA
- IBM DB2 Database
- IBM DataStage
- IBM Cognos
- Microsoft Excel

Creating Your Data Quality Program



Key Requirements for a DQP

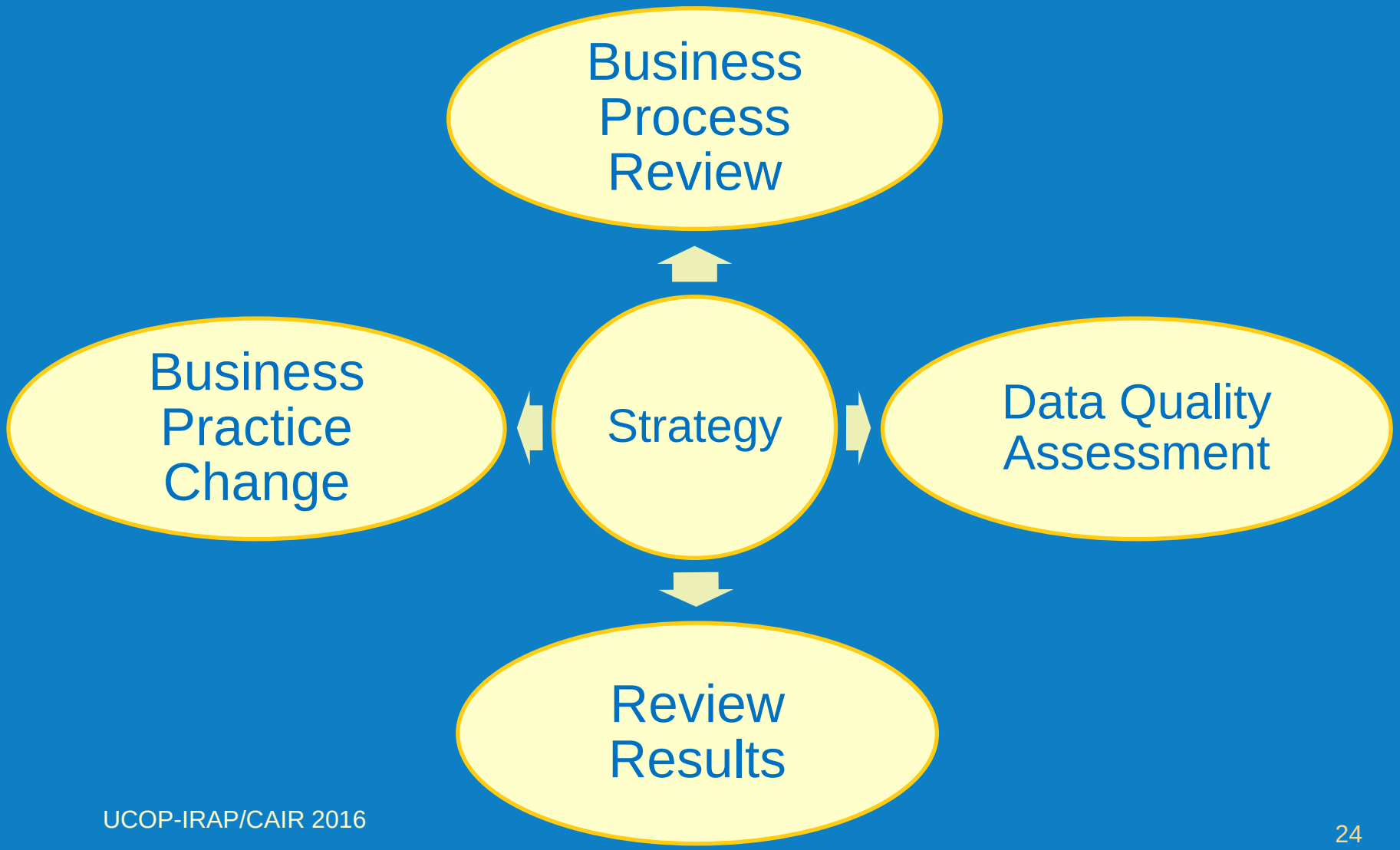
A Data Quality Vision

A Data Quality Strategy

Develop A Data Quality Vision

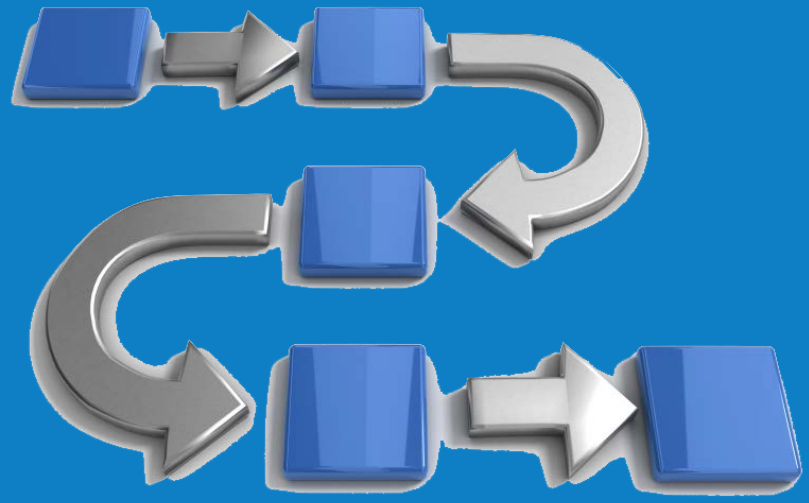
- Every organization needs:
  - A vision with respect to having good quality data
  - An accompanying policy to implement that vision
  - A strategy for implementation
- Every organization should look:
  - For efficiencies in data collection and quality control processes
  - Beyond immediate use and examine user requirements
  - For ways to build networks and partnerships

Define a Data Quality Strategy



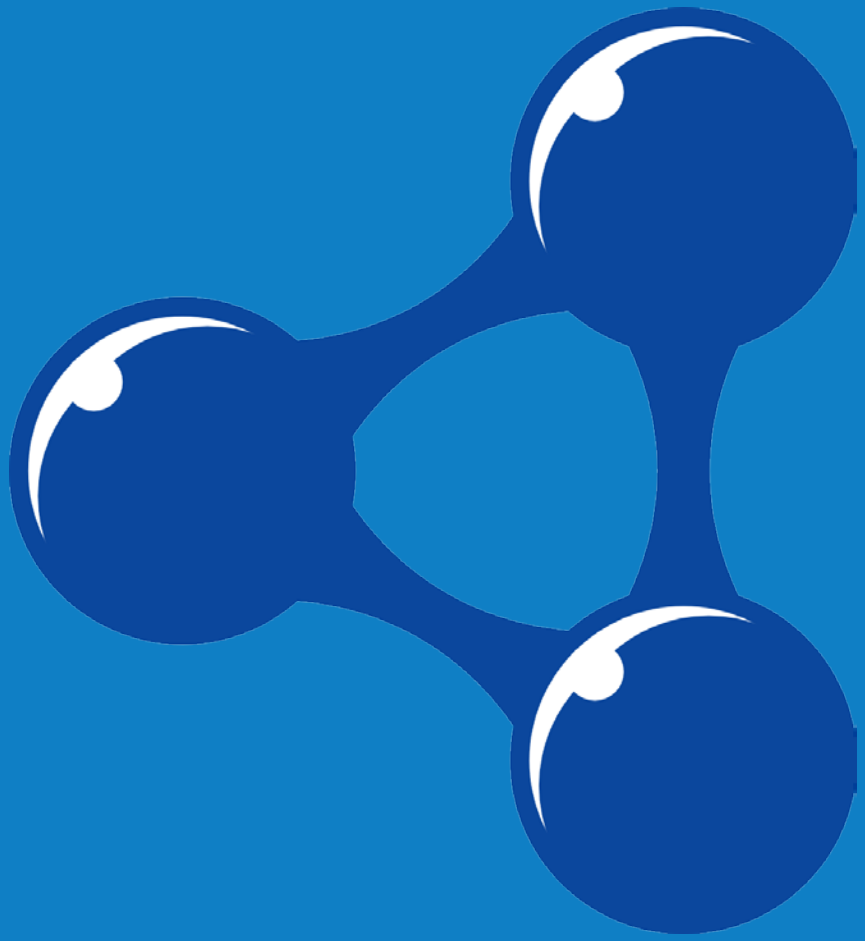
Review Your Business Processes

- How and when is data collected?
- Where is data stored?
- Is the same data stored in more than one system?
- Who creates the data?
- Who uses the data?
- What kind of quality checks already exist?



Do a Data Quality Assessment

- What are the quality criteria?
- What are the acceptable range of values?
- What kind of thresholds should be in place?
- What are your business rules?



Review Your Results

- Develop a systematic approach to reviewing results
- Develop a process for data cleaning or correction
- Identify source of data problems
- Communicate!

Implement Necessary Changes

- Implement changes to improve data quality
  - Centralize reference data codes
  - Consolidate data collection and storage
- Adopt ongoing data quality review process
  - Review data regularly
  - Communicate quality improvements

Where Are We Now?

Where are we now?

- Standardizing input file specifications across all content areas
- Implementation of a requirement managements tool
- Documenting business related quality rules
- Promoting data governance through the creation of a data operations website
- Improving communication between the data creators and IRAP
- Improving relationship between IRAP and IT

Final Thoughts

- Quality data is achievable if you are willing to:
  - Take a critical look at your existing data
  - Implement changes to how you collect and manage data
  - Invest the time to educate and communicate with data creators and users
  - Make data quality improvements an ongoing process

It's not the things you don't know that matter, it's the things you know that aren't so.

Will Rogers, Famous Okie GI specialist

Fast is fine but accuracy is final.

Wyatt Earp - Officer of the law, gambler and saloon keeper in the Wild West

Good data are the data you already have.

Dr. Edgar Horwood - Founder of the Urban and Regional Information Systems

UNIVERSITY
OF
CALIFORNIA