Demystifying Data Quality Tools

2016 CAS RPM Seminar

Verisk Insurance Solutions

SERVE | ADD VALUE | INNOVATE

Antitrust Notice

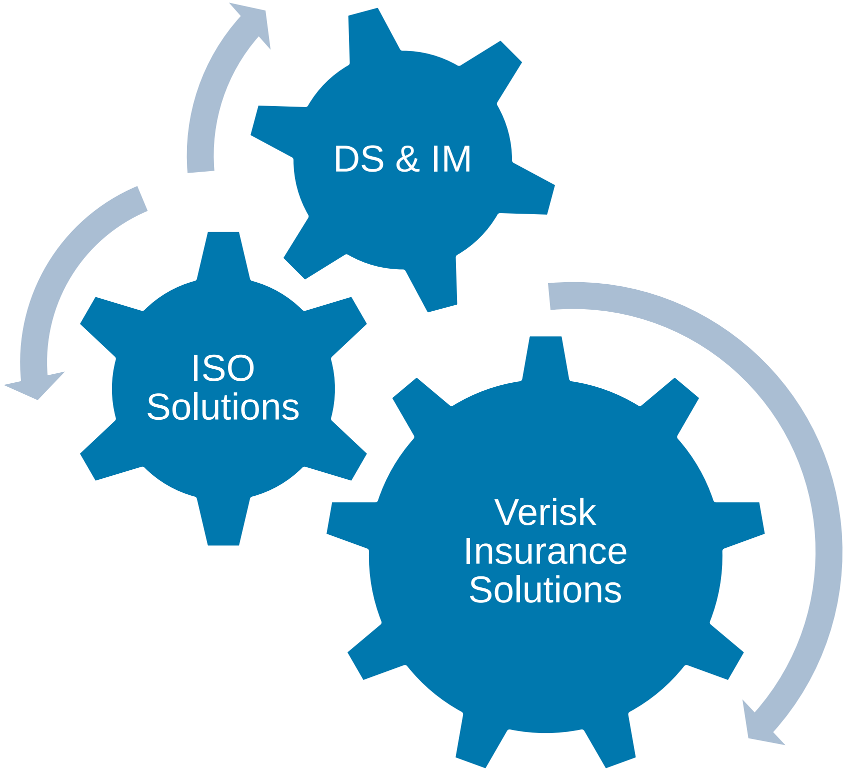
- The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.
- Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.
- It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

Speakers

Joe Izzo
Senior Vice President
Data Strategy & Information Management (DS & IM)

Tracy Spadola
Vice President
Strategic Data Operations

Hernan L. Medina
Director
Analytical Data Management



Data Flowing to ISO Solutions

Almost 3 billion records processed each year

Over 1,800 insurers provide data

Roughly 18 billion records in commercial and personal lines

Introduction

Data quality — primary concern

Actuarial Standard of Practice 23

Best practices / commonly used techniques

References

Learning Objectives

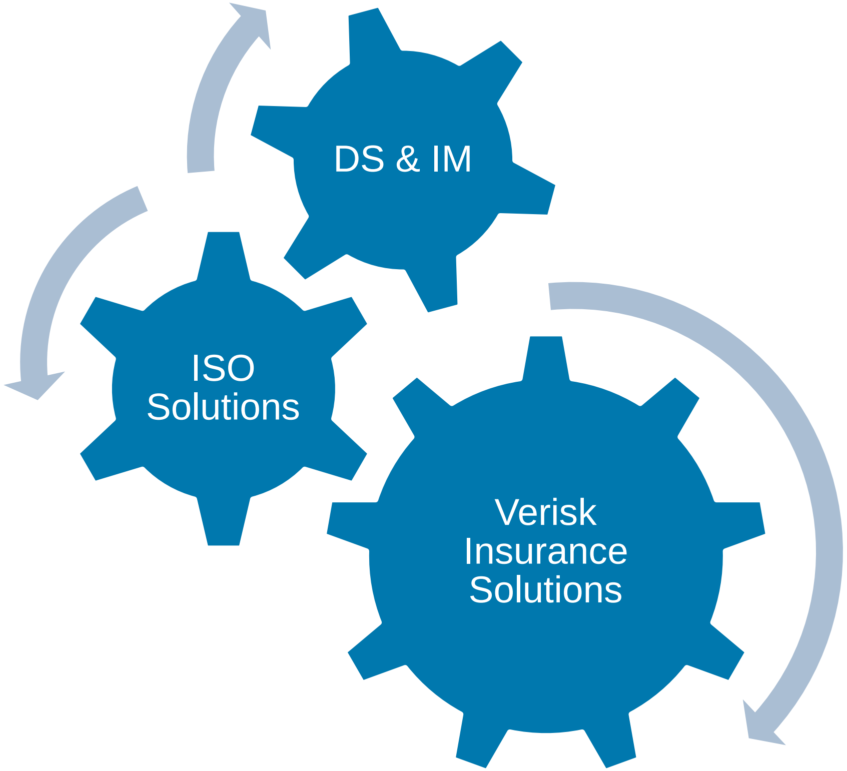
Achieve an understanding of key data quality concepts.

Identify concepts or processes commonly associated with data quality tools.

Become aware of the functionality available in data quality management tools.

Data Quality Concepts

Tracy Spadola
Vice President
Strategic Data Operations



Data Quality Defined

"Data are of high quality if they are fit for their intended uses in operations, decision making and planning. Data are fit for use if they are free of defects and possess desired features"

- Joseph Juran

Fit for Use

Free of Defects	Possess Desired Features
Accessible	Relevant
Accurate	Comprehensive
Complete	Easy to Read
Reliability	Easy to Interpret
Timely	Proper Level of Detail

Data Quality: The Field Guide T. Redman, Ph.D

Data Quality Concepts

Completeness

Reliability

Accuracy

Consistency

Relevance

Completeness

- Degree to which data values are present
  - for any required attribute
  - which required records are present

- Is all of the needed data available?
- Are all of the data fields populated (if they should be)?

Reliability

- Sufficiently complete, free of errors
- Fit for purpose and context

- Impacted by Source and other data quality measures

Accuracy

- Degree to which the data value represents it's source

- Beyond Validity
- Representation of "real life"

Consistency

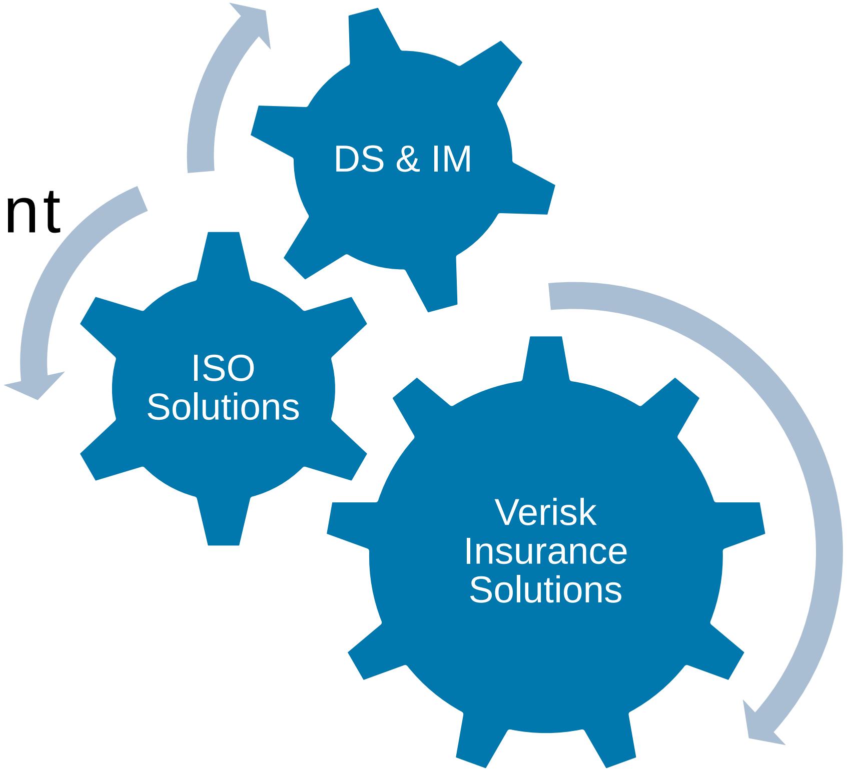
- Degree to which the data satisfies business rules
- Looks at relationships between data elements or sources
  - State Code = NY
  - Zip Code = 60606
    - While both data elements are valid, They are inconsistent

Relevance

- Relationship of the data to a task or decision
- Data is relevant if
  - It contributes to the completion of the task
  - It contributes to the making of a decision

Data Quality Tools

Hernan L. Medina
Director
Analytical Data Management

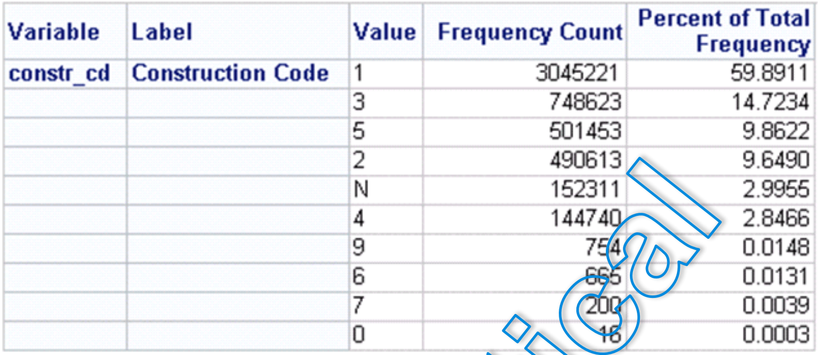


Sample Data Quality Tool Features

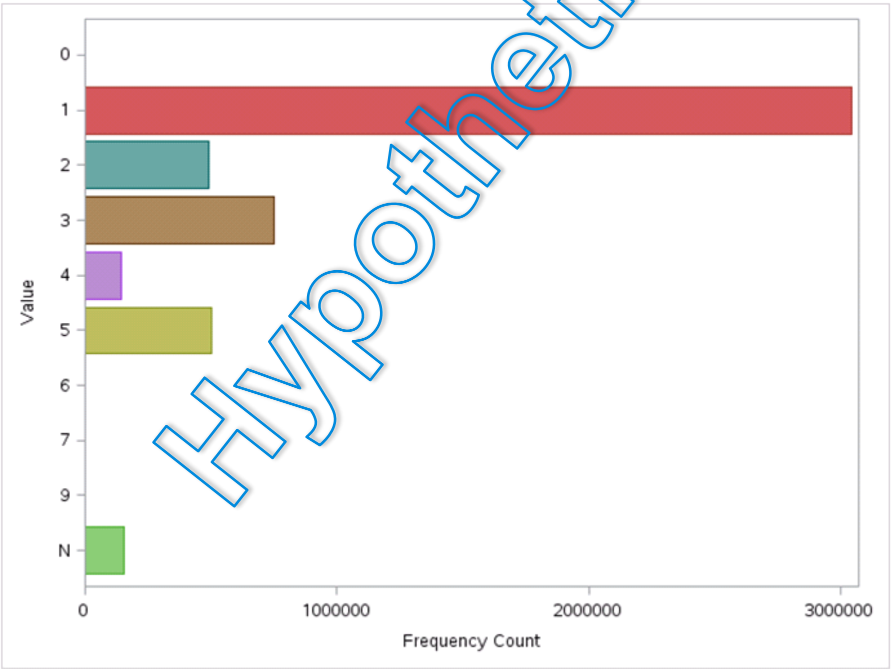
- Data profiling, data quality assessment
- Normalizing, parsing
- Deduplication, entity resolution
- Metadata management

Data Profiling, Quality Assessment

- Examining and analyzing data to measure or improve its quality. (Strube and Russell)
- Examining a data source to produce metadata about its attributes and the relationships between them. (Maydanchik)



Variable	Label	Value	Frequency Count	Percent of Total Frequency
constr_cd	Construction Code	1	3045221	59.8911
		3	748623	14.7234
		5	501453	9.8622
		2	490613	9.6490
		N	152311	2.9955
		4	144740	2.8466
		9	754	0.0148
		6	665	0.0131
		7	200	0.0039
		0	16	0.0003



Data Profiling, Quality Assessment

Basic statistics: min, max, mean, etc.

Most frequent values

Shape of distribution

Data type, length, format

Frequency of null values

Data Profiling, Quality Assessment

Business rule management

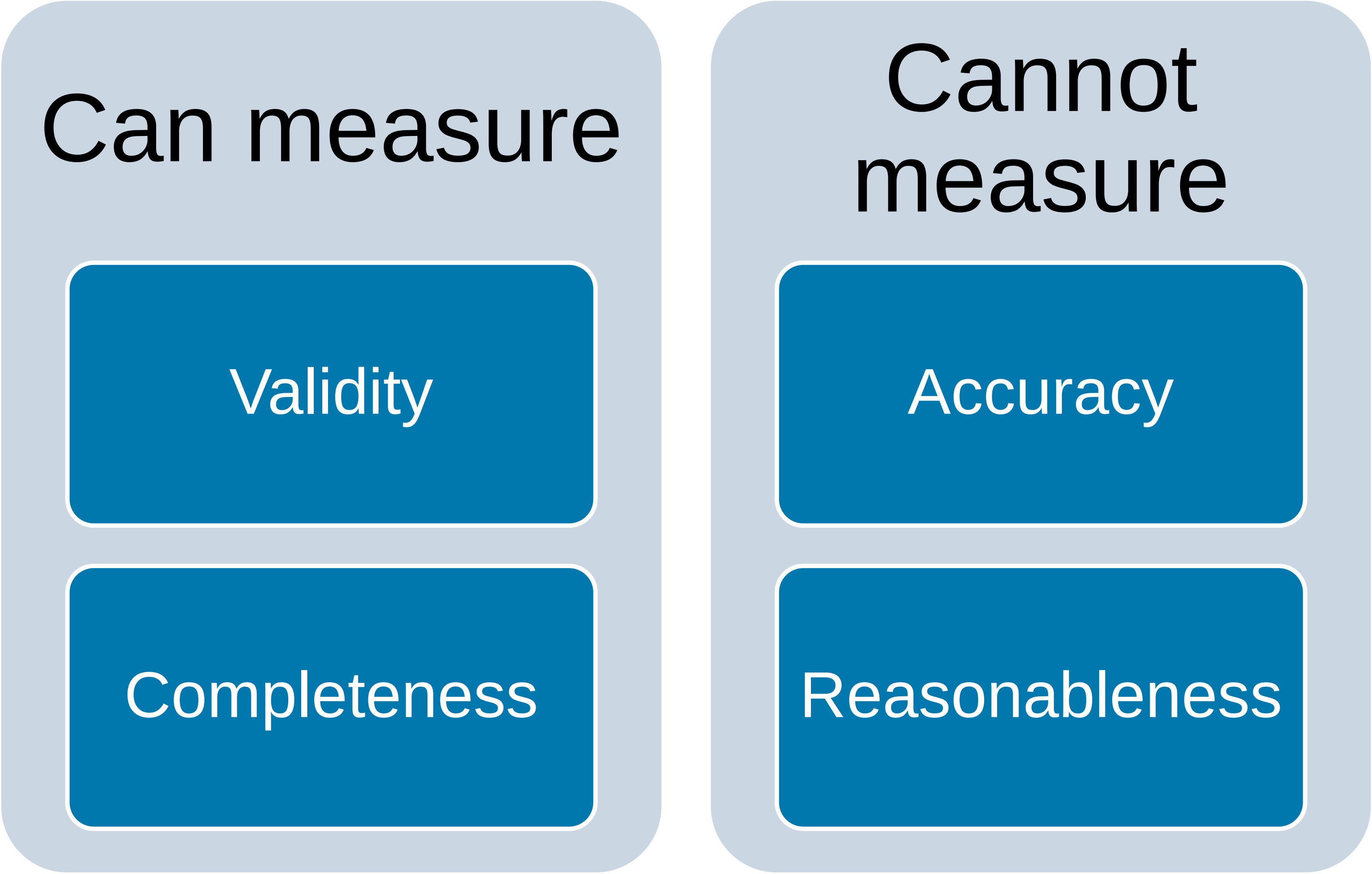
Business rule discovery

Assessment against business rules

Charts showing data quality state

Control charts showing data quality trends

Data Profiling, Quality Assessment



Can measure

Validity

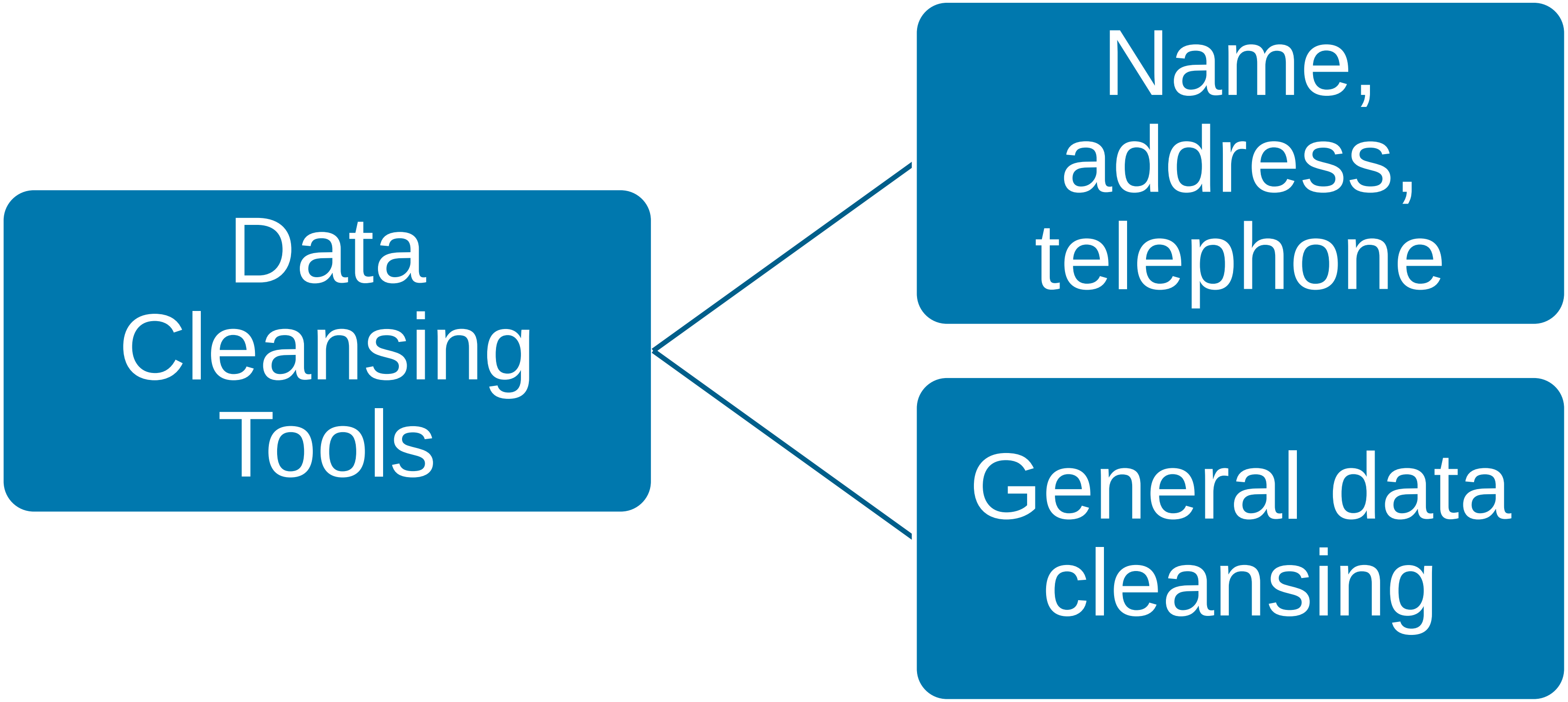
Completeness

Cannot measure

Accuracy

Reasonableness

Normalizing, Parsing

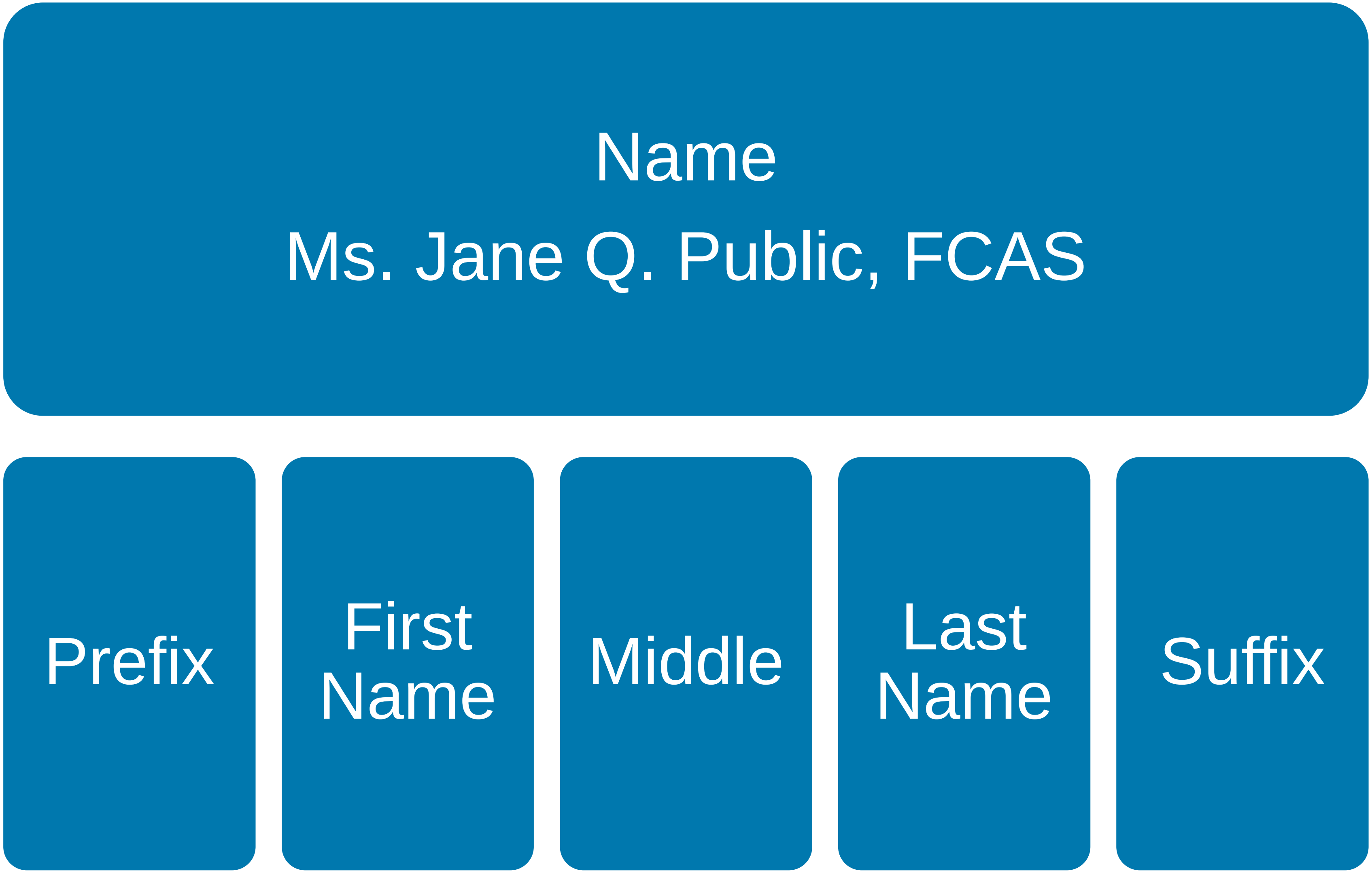


Data Cleansing Tools

Name, address, telephone

General data cleansing

Normalizing, Parsing



Name

Ms. Jane Q. Public, FCAS

Prefix

First Name

Middle

Last Name

Suffix

Normalizing, Parsing

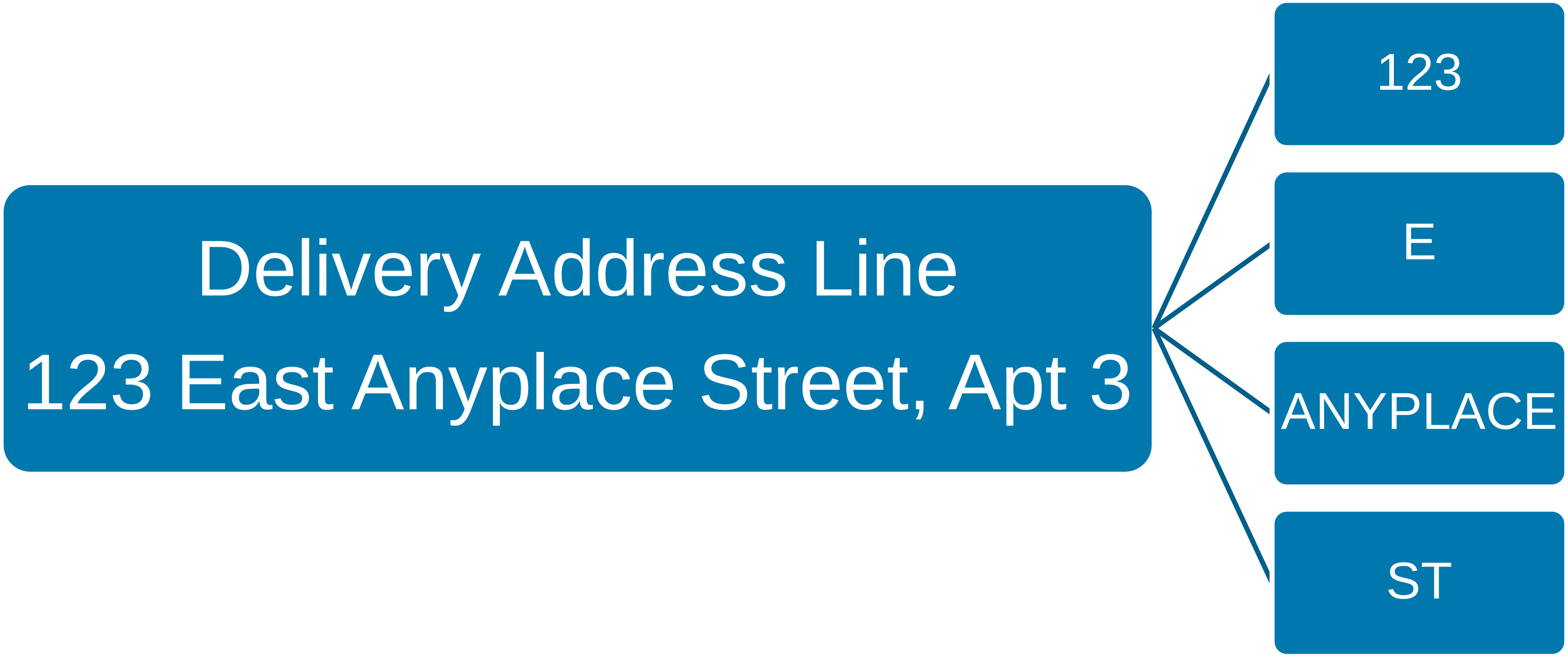
Ms. Jane Q. Public
123 East Anyplace Street
Anytown, NY 12345-6789

USPS standards

Missing elements

Normalization codes

Normalizing, Parsing



Delivery Address Line

123 East Anyplace Street, Apt 3

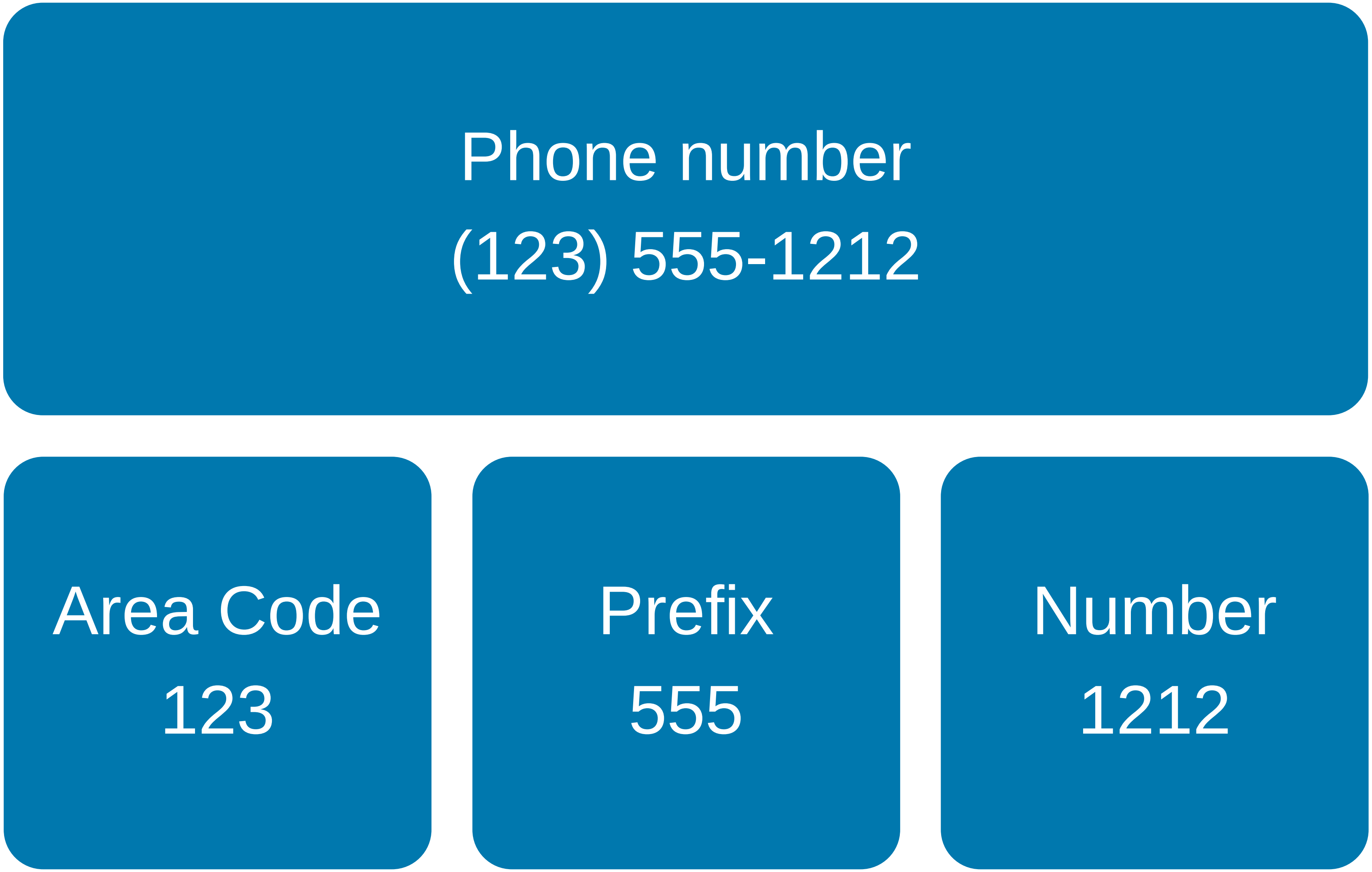
123

E

ANYPLACE

ST

Normalizing, Parsing



Phone number

(123) 555-1212

Area Code

123

Prefix

555

Number

1212

Deduplication, Entity Resolution

Name	Address	Phone
J. Q. Public	123 Anyplace St Anytown, NY 12345-6789	555-1212
Jane Public	123 E Anyplace Street Anytown, NY 12345	(123) 555-1212
J. Public		(123) 555-1212

JANE	Q	PUBLIC	123	E	ANYPLACE	ST	ANYTOWN	NY	12345	6789	123	555	1212

Metadata Management

- "Management and quality control of data definition and information architecture development." (English)

- Data dictionary
  - Claims
  - Incurred Loss

References

- CAS Data Management Educational Materials Working Party, Actuarial I. Q. (Information Quality), CAS eForum, Winter 2008
- English, Larry P., Improving Data Warehouse and Business Information Quality, Wiley Computer Publishing, 1999
- Fisher, Anthony (Tony), The Data Asset: Govern Your Data For Business Success, Wiley & SAS Business Series, 2009.
- Maydanchik, Arkady, Data Quality Assessment, Technics Publications, 2007.
- Redman, Thomas C., Data Driven: Profiting From Your Most Important Business Asset, Harvard Business Press, 2008.
- Sebastian-Coleman, Laura, Measuring Data Quality for Ongoing Improvement, Morgan Kaufmann Business Intelligence Series, 2013
- Strube, J. and Russell, B., Actuarial Data Management In A High-Volume Transactional Processing Environment, CAS eForum, Winter 2005

Contact Information

Joseph M. Izzo
Senior Vice President
Data Strategy and Information Management
ISO Solutions
Joseph.Izzo@verisk.com
(201) 469-2308

Tracy Spadola
Vice President
Strategic Data Operations
ISO Solutions
Tracy.Spadola@verisk.com
(201) 469-2213

Hernan L. Medina
Director
Analytical Data Management
ISO Solutions
Hernan.Medina@verisk.com
(201) 469-3829

This material was used exclusively as an exhibit to an oral presentation. It may not be, nor should it be relied upon as reflecting, a complete record of the discussion.

Verisk
Insurance Solutions