Data Management Workgroup

burt.walsh@ast.myflorida.com

michael.avello@ast.myflorida.com

May 17, 2018

Data Management Workgroup

Overview

- Progress on Enterprise Metadata Inventory
- Review of Past Topics
- Data Management Workgroup Goals/Strategy
- More Immediate Goal - Agency Data Sharing
- Record Linkage
- Next Topics

Data Management Workgroup

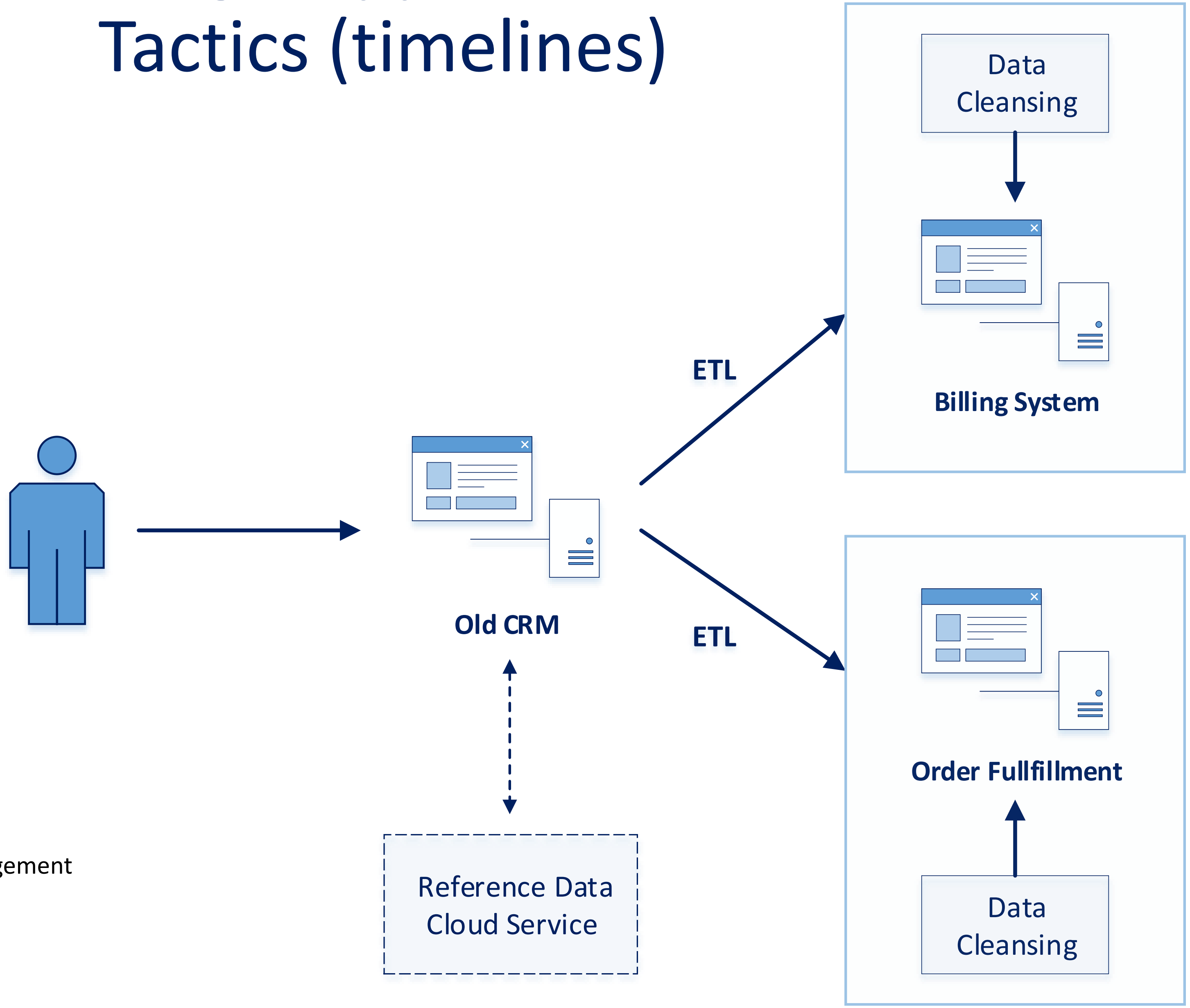
Progress on Enterprise Metadata Inventory

- Aggregation of ingress and egress technologies
- Multiple lines for ingress and egress
- High level overview of dataset/application function use
- Key entities represented in dataset/concepts in reports for instance
- Any questions and/or suggestions?

Review of Past Topics

- Key Dataset Identification/Key Reports
- Data Quality and Metrics
  - Do you have a quality process in place?
  - Is it quantitative or qualitative, or both?
  - Is quality monitored?
  - Is there a remediation strategy?

Strategic Application of Tactics (timelines)



ETL = Extract, Transform, Load
CRM = Customer Relationship Management

Data Management Workgroup

Review of Past Topics

- Variety of Tools (Open Source/Commercial/Cloud Based)
  - Do you have data quality tools?
  - Did you have an opportunity to use any of the demonstrated tools?
  - Are you planning on acquiring tools?
  - How can we help?
- Would a survey (and report) of existing tools used at agencies be valuable?

Data Management Workgroup

Review of Past Topics

- Data Lineage
  - Do you know the source of the data?
  - How many steps back?
- Memorandum on Understanding (MOU) / Data Use License (DUL)

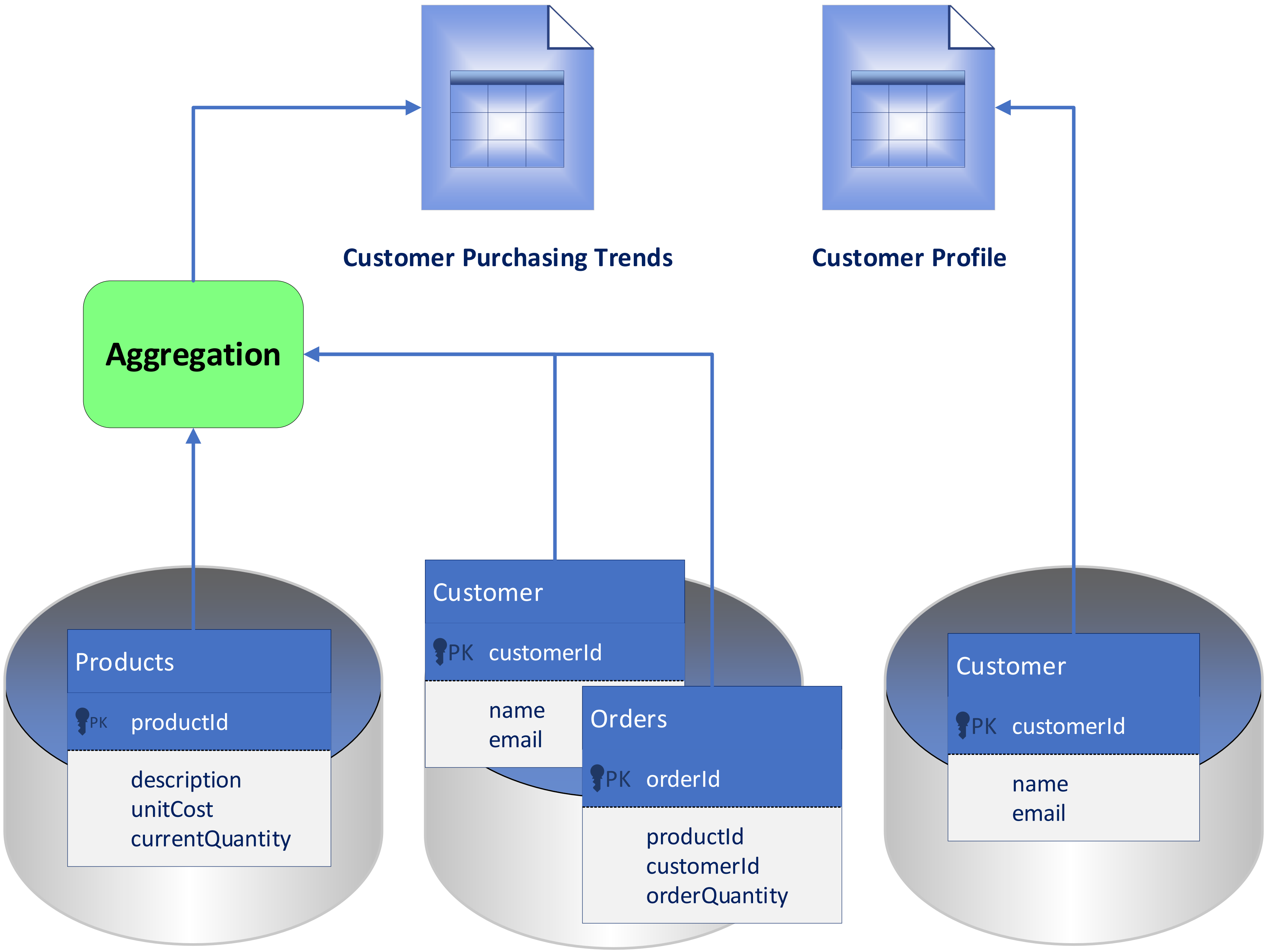
Data Management Workgroup

Review of Past Topics

Performance Measure/Risk Management

- How do we know we are doing well?
  - (Key Performance Indicators)
- What things should be on our problem radar?
  - (Key Risk Indicators)
- Is this data conveyed to management mostly by reports?

Data Lineage for Reporting



Data Management Workgroup

Data Lineage Mapping

Report	Field (Concept)	Data Source Type	Source Field	Description
Customer Retention	Years Retained	Aggregate (SQL)	DB.Table.field	Number of years person retained as a customer
Customer Retention	Expenditures by Year	Aggregate	DB Stored Procedure Name	Expenditures (USD) per year for each year the person is retained as a customer
Customer Retention	Geographic Region Map	GIS	CRM	Geospatial data store

Review of Past Topics

Report Data Source Mapping

- Is report data quality important?
- Can you guarantee quality if asked?
- Do you have control of the quality? Where is the true source?
- Is it important to map sources to report entities?
- We can help with any mapping!

Data Management Workgroup

Data Management Workgroup Goal

Establish and promote data management standards to ensure the responsible use and full power of the state's data so it can be used to serve the citizens of Florida.

Data Management Workgroup

Data Governance

*"Data Governance is a quality control discipline for managing, using, **improving and protecting organizational information**. Effective data governance enhances the **quality, availability, integrity, and protection** of a company's data by fostering **cross-organizational collaboration and structured policy-making**."*

- IBM

DAMA-DMBOK2 Data Management Framework

Data Management Workgroup

Data Management Workgroup Strategy

- Discuss data management topics and generic best practices
- Assess the current maturity level
- Understand each agency's particular needs
- Develop improvement plans
- Serve agencies in the plan execution/governance
- Memorialize shared consensus on best practices as standards

Data Management Workgroup

What is Currently Ahead of Us

- Enterprise Data Inventory (living document)
- Convenience of dataset discovery
- Convenience of data sharing between state agencies
- Context for data and improved quality
- Tools, tools, tools
- Serving the citizens of Florida!

Agencies Leveraging State Data

AGENCY FOR STATE TECHNOLOGY

Data Catalog

Searchable Metadata

(for agencies only)

Data Quality Record Linkage

& Privacy-Protected Record Linkage

MOUs

Data Use Agreements

(Data Use Licenses)

Techniques & Tools

python

jupyter

R

Data Quality

If data has value in a certain context, then it may be an indication that more rigorous data quality rules are required.

Data Quality Problems

- Data entry errors
- Missing integrity constraints (e.g., age = 345)
- Multiple conventions for data input
- Multiple entries for the same object
- Data tables not normalized
- Systematic conflicts between Database Management Systems (DBMS)

Implications of Data Quality Problems

- Data integration efforts compromised
- Data migration efforts stalled
- Disjointed & fragmented reporting sources
- Inefficient / ineffective business processes
- Impact to delivery of services and benefits
- Missed information ➝ missed opportunities

Source: Askham, N., Cook, D., Doyle, M., Fereday, H., Gibson, M., Landbeck, U., ... & Schwarzenbach, J. (2013). The six primary dimensions for data quality assessment. DAMA UK Working Group, 432-435.

Deduplication/Record Linkage

- What are the key entities/concepts that drive business?
- What constitutes a duplicate (key fields)?
- Are the fields the same as the measured data quality fields?
- Do you meet the same expectations of quality over multiple data sets?
- Are certain fields more important?
- Does the entity (concept) span tables or data stores?
- How do we “measure” similarity?

Data Quality

Dimensions & Measurements for Data Quality Assessment

Data Quality Dimensions

Describe a feature (characteristic, attribute or facet) of data that can be measured or assessed against defined standards in order to determine the quality of data.

Completeness
- Are all data sets and data items recorded?

Consistency
- Can we match the data set across data stores?

Uniqueness
- Is there a single view of the data set?

Validity
- Does the data match the rules?

Accuracy
- Does the data reflect the data set?

Source: Askham, N., Cook, D., Doyle, M., Fereday, H., Gibson, M., Landbeck, U., ... & Schwarzenbach, J. (2013). The six primary dimensions for data quality assessment. DAMA UK Working Group, 432-435.

Data Quality Assessment Dimensions & Measurements

Title	Uniqueness
Definition	No thing will be recorded more than once based upon how that thing is identified.
Reference	Data item measured against itself or its counterpart in another data set or database.
Measure	Analysis of the number of things as assessed in the 'real world' compared to the number of records of things in the data set. The real world number of things could be either determined from a different and perhaps more reliable data set or a relevant external comparator.
Scope	Measured against all records within a single data set
Unit of Measure	Percentage
Type of Measure: • Assessment • Continuous • Discrete	Discrete
Related dimension	Consistency
Optionality	Dependent on circumstances
Example(s)	A school has 120 current students and 380 former students (i.e. 500 in total) however; the Student database shows 520 different student records. This could include Fred Smith and Freddy Smith as separate records, despite there only being one student at the school named Fred Smith. This indicates a uniqueness of 500/520 x 100 = 96.2%
Pseudo code	(Number of things in real world)/(Number of records describing different things)
External Validation	IAM Asset Information Quality Handbook Principles of Data Management, Keith Gordon

Source: Askham, N., Cook, D., Doyle, M., Fereday, H., Gibson, M., Landbeck, U., ... & Schwarzenbach, J. (2013). The six primary dimensions for data quality assessment. DAMA UK Working Group, 432-435.

Data Quality Assessment Dimensions & Measurements (example dataset)

Techniques & Tools

Age Entry Errors

rowID	firstName	lastName	age	street	apt	city	state	zip	ssn
R01	John	Willams	38	643 Gulf Ln		Half Moon Bay	CA	94013	A2B-4C-678D
R02	Richard	Alpert	151	15 Black Rock St		Cannon Beach	OR	97110	A3B-5C-78D1
R03	Ana Lucia	Cortez	39	48 Ocean Park Ave		Santa Monica	CA	90405	A4B-6C-891D
R04	Joh	Williams	42	642 Gulf Lane		Halfmoon Bay	CA	94013	A2B-4C678D
R05	Daniel	Faraday	48	23 Martin St		Essex	MA	01929	A5B-7C-9123
R06	Jonathan	Williams	42*	643 Gulf Lane		Half Moon Bay	CA	94013	A2B4C-678D
R07	Penny	Widmore	43	1623 Hawthorne Road		Palos Verdes	CA	90275	A6B-8C-123D
R08	Austen	Kate	38	1516 Ontario Street		Ames	IA	50014	A78-9B-234D
R09	John	William	47	403 Stadium Dr	B-005	Tallahassee	FL	32304	A2B4C678D
R10	Benjamin	Linus	63	815 Oceanic Ave		Portland	OR	97205	A8B-C2-345D

AGENCY FOR STATE TECHNOLOGY

Techniques & Tools

Data Preprocessing & Cleaning with R

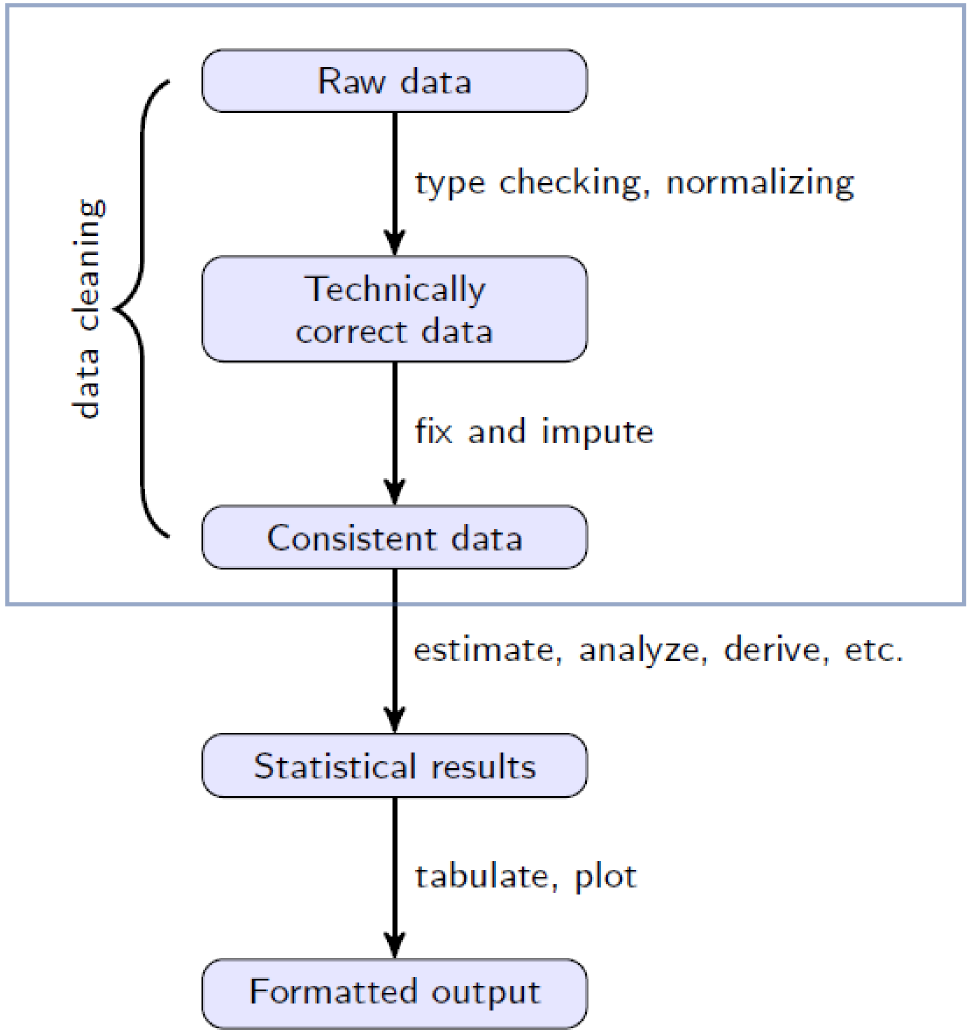


Figure 1: Statistical analysis value chain

Source: De Jonge, E., & van der Loo, M. (2013). An introduction to data cleaning with R. Heerlen: Statistics Netherlands. p. 7.

Data Quality Assessment Dimensions & Measurements

(data cleaning with R)

Techniques & Tools

(1) From Raw Data to "Technically Correct" Data

Table 1: Steps to take when converting lines in a raw text file to a `data.frame` *with correctly typed columns.*

	Step	result
1	Read the data with `readLines`	`character`
2	Select lines containing data	`character`
3	Split lines into separate fields	`list` of `character` vectors
4	Standardize rows	`list` of equivalent vectors
5	Transform to `data.frame`	`data.frame`
6	Normalize and coerce to correct type	`data.frame`

```
(M <- matrix(
  unlist(standardFields)
  , nrow=length(standardFields)
  , byrow=TRUE))
##      [,1]    [,2]   [,3]
## [1,] "Gratt" "1861" "1892"
## [2,] "Bob"   NA     "1892"
## [3,] "Emmet" "1871" "1937"
colnames(M) <- c("name","birth","death")
(daltons <- as.data.frame(M, stringsAsFactors=FALSE))
##    name birth death
```

(2) Consistent Data

Missing values, special values, (obvious) errors, and outliers are either removed, corrected, or imputed (replaced with substitute values).

```
# numerical rules
age >= 0
height > 0
age <= 150
age > yearsmarried

# categorical rules
status %in% c("married","single","widowed")
agegroup %in% c("child","adult","elderly")
if ( status == "married" ) agegroup %in% c("adult","elderly")

# mixed rules
if ( status %in% c("married","widowed")) age - yearsmarried >= 17
if ( age < 18 ) agegroup == "child"
if ( age >= 18 && age <65 ) agegroup == "adult"
if ( age >= 65 ) agegroup == "elderly"
```

Data Quality Assessment Dimensions & Measurements

(data cleaning with R)

Techniques & Tools

Correct Age Errors

Normalize SSN Values (remove dashes)

rowID	firstName	lastName	age	street	apt	city	state	zip	ssn
R01	John	Willams	38	643 Gulf Ln		Half Moon Bay	CA	94013	A2B-4C-678D
R02	Richard	Alpert	151	15 Black Rock St		Cannon Beach	OR	97110	A3B-5C-78D1
R03	Ana Lucia	Cortez	39	48 Ocean Park Ave		Santa Monica	CA	90405	A4B-6C-891D
R04	Joh	Williams	42	642 Gulf Lane		Halfmoon Bay	CA	94013	A2B-4C678D
R05	Daniel	Faraday	48	23 Martin St		Essex	MA	01929	A5B-7C-9123
R06	Jonathan	Williams	42*	643 Gulf Lane		Half Moon Bay	CA	94013	A2B4C-678D
R07	Penny	Widmore	43	1623 Hawthorne Road		Palos Verdes	CA	90275	A6B-8C-123D
R08	Austen	Kate	38	1516 Ontario Street		Ames	IA	50014	A78-9B-234D
R09	John	William	47	403 Stadium Dr	B-005	Tallahassee	FL	32304	A2B4C678D
R10	Benjamin	Linus	63	815 Oceanic Ave		Portland	OR	97205	A8B-C2-345D

Data Quality Assessment Dimensions & Measurements

(data deduplication with R)

Techniques & Tools

Soundex codes and algorithms can be used to detect similarity.

```
soundex('Ann') == soundex('Anne') == 'A500'
```

rowID	firstName	lastName	age	street	apt	city	state	zip	ssn
R01	John	Willams	38	643 Gulf Ln		Half Moon Bay	CA	94013	A2B-4C-678D
R02	Richard	Alpert	151	15 Black Rock St		Cannon Beach	OR	97110	A3B-5C-78D1
R03	Ana Lucia	Cortez	39	48 Ocean Park Ave		Santa Monica	CA	90405	A4B-6C-891D
R04	Joh	Williams	42	642 Gulf Lane		Halfmoon Bay	CA	94013	A2B-4C678D
R05	Daniel	Faraday	48	23 Martin St		Essex	MA	01929	A5B-7C-9123
R06	Jonathan	Williams	42*	643 Gulf Lane		Half Moon Bay	CA	94013	A2B4C-678D
R07	Penny	Widmore	43	1623 Hawthorne Road		Palos Verdes	CA	90275	A6B-8C-123D
R08	Austen	Kate	38	1516 Ontario Street		Ames	IA	50014	A78-9B-234D
R09	John	William	47	403 Stadium Dr	B-005	Tallahassee	FL	32304	A2B4C678D
R10	Benjamin	Linus	63	815 Oceanic Ave		Portland	OR	97205	A8B-C2-345D

The Soundex code for “John” is J500 ... for “Jonathan,” J535... for “Joh,” it’s J000... for “ohn,” it’s O500*.

* The first letter of the code is simply the first letter of the word; otherwise, the Soundex algorithm will disregard vowels.

Data Quality Assessment Dimensions & Measurements

(data deduplication with R)

Techniques & Tools

Soundex Codes and Algorithms

Driver's License Analyzer: Florida Results

Input

Florida Driver's License Number: A

Output

First Name: M.
Middle Initial: D.
Last Name: Starts with a A, followed by a labial (B, F, P, or V), followed by a long liquid (L). Perhaps it sounds like "Aebele"
Last Name Guesses: Avila, Abel, Abell, Apple, Appel, Able, Apel, Avilla, Appell, Auvil, Abila, Abele, Abella, Avella, Apfel, Apolo, Avolio, Appello, Avello, Auble, Apollo, Abilay, Abuhl, Ablao, Avola
Gender: male
Date of Birth:
Wild Guess: M. D. Avila, male, born

What is a Soundex code? - High Programmer
www.highprogrammer.com › Alan De Smet › Unique ID
Soundex is a hashing system for english words. ... You might be interested in a history of various different versions of the **Soundex** coding system. ... So Pfizer becomes Pizer, Sack becomes Sac, Czar becomes Car, Collins becomes Colins, and Mroczak becomes Mrocak.

Data Quality Assessment Dimensions & Measurements

(data deduplication with R)

Techniques & Tools

python jupyter R

There are a number of other metrics that can be used to measure data closeness or similarity.

e.g., the Levenshtein Distance Algorithm (1965)

```
> LevDistance <- function (str1, str2)
+ {
+   if (typeof(str1) != "character" && class(str1) != "factor")
+     stop(sprintf("Illegal data type: %s", typeof(str1)))
+   if (class(str1) == "factor")
+     str = as.character(str1)
+   if (typeof(str2) != "character" && class(str2) != "factor")
+     stop(sprintf("Illegal data type: %s", typeof(str2)))
+   if (class(str2) == "factor")
+     str = as.character(str2)
+   if ((is.array(str1) || is.array(str2)) && dim(str1) != dim(str2))
+     stop("non-conformable arrays")
+   if (length(str1) == 0 || length(str2) == 0)
+     return(integer(0))
+   l1 <- length(str1)
+   l2 <- length(str2)
+   out <- .C("levenshtein", as.character(str1), as.character(str2),
+             l1, l2, ans = integer(max(l1, l2)), PACKAGE = "RecordLinkage")
+   if (any(is.na(str1), is.na(str2)))
+     out$ans[is.na(str1) | is.na(str2)] = NA
+   if (is.array(str1))
+     return(array(out$ans, dim(str1)))
+   if (is.array(str2))
+     return(array(out$ans, dim(str2)))
+   return(out$ans)
+ }
>
> LevDistance("John", "Jon")
[1] 1
```

Data Quality Assessment Dimensions & Measurements

(data deduplication with Python)

Techniques & Tools

Temporal / Transient Data
(addresses can change over time)

rowID	firstName	lastName	age	street	apt	city	state	zip	ssn
R01	John	Willams	38	643 Gulf Ln		Half Moon Bay	CA	94013	A2B-4C-678D
R02	Richard	Alpert	151	15 Black Rock St		Cannon Beach	OR	97110	A3B-5C-78D1
R03	Ana Lucia	Cortez	39	48 Ocean Park Ave		Santa Monica	CA	90405	A4B-6C-891D
R04	Joh	Williams	42	642 Gulf Lane		Halfmoon Bay	CA	94013	A2B-4C678D
R05	Daniel	Faraday	48	23 Martin St		Essex	MA	01929	A5B-7C-9123
R06	Jonathan	Williams	42*	643 Gulf Lane		Half Moon Bay	CA	94013	A2B4C-678D
R07	Penny	Widmore	43	1623 Hawthorne Road		Palos Verdes	CA	90275	A6B-8C-123D
R08	Austen	Kate	38	1516 Ontario Street		Ames	IA	50014	A78-9B-234D
R09	John	William	47	403 Stadium Dr	B-005	Tallahassee	FL	32304	A2B4C678D
R10	Benjamin	Linus	63	815 Oceanic Ave		Portland	OR	97205	A8B-C2-345D

Data Quality Assessment Dimensions & Measurements (data deduplication with Python)

Techniques & Tools

Emphasis of Business Rules

rowID	firstName	lastName	age	street	apt	city	state	zip	ssn
R01	John	Willams	38	643 Gulf Ln		Half Moon Bay	CA	94013	A2B-4C-678D
R02	Richard	Alpert	151	15 Black Rock St		Cannon Beach	OR	97110	A3B-5C-78D1
R03	Ana Lucia	Cortez	39	48 Ocean Park Ave		Santa Monica	CA	90405	A4B-6C-891D
R04	Joh	Williams	42	642 Gulf Lane		Halfmoon Bay	CA	94013	A2B-4C678D
R05	Daniel	Faraday	48	23 Martin St		Essex	MA	01929	A5B-7C-9123
R06	Jonathan	Williams	42*	643 Gulf Lane		Half Moon Bay	CA	94013	A2B4C-678D
R07	Penny	Widmore	43	1623 Hawthorne Road		Palos Verdes	CA	90275	A6B-8C-123D
R08	Austen	Kate	38	1516 Ontario Street		Ames	IA	50014	A78-9B-234D
R09	John	William	47	403 Stadium Dr	B-005	Tallahassee	FL	32304	A2B4C678D
R10	Benjamin	Linus	63	815 Oceanic Ave		Portland	OR	97205	A8B-C2-345D

Business rules tell us that SSN represents the strongest indicator of similarity (duplication).

Data Quality Assessment Dimensions & Measurements

(data deduplication with Python)

Techniques & Tools

Soundex functions

Temporal Data

Business Rules Emphasis

rowID	firstName	lastName	age	street	apt	city	state	zip	ssn
R01	John	Willams	38	643 Gulf Ln		Half Moon Bay	CA	94013	A2B-4C-678D
R02	Richard	Alpert	151	15 Black Rock St		Cannon Beach	OR	97110	A3B-5C-78D1
R03	Ana Lucia	Cortez	39	48 Ocean Park Ave		Santa Monica	CA	90405	A4B-6C-891D
R04	Joh	Williams	42	642 Gulf Lane		Halfmoon Bay	CA	94013	A2B-4C678D
R05	Daniel	Faraday	48	23 Martin St		Essex	MA	01929	A5B-7C-9123
R06	Jonathan	Williams	42*	643 Gulf Lane		Half Moon Bay	CA	94013	A2B4C-678D
R07	Penny	Widmore	43	1623 Hawthorne Road		Palos Verdes	CA	90275	A6B-8C-123D
R08	Austen	Kate	38	1516 Ontario Street		Ames	IA	50014	A78-9B-234D
R09	John	William	47	403 Stadium Dr	B-005	Tallahassee	FL	32304	A2B4C678D
R10	Benjamin	Linus	63	815 Oceanic Ave		Portland	OR	97205	A8B-C2-345D

Putting it all together—with an understanding of the primary business rules—increases the confidence level for finding duplicate records in the database.

Data Cleaning & Deduplication

RecordLinkage Python Package

Techniques & Tools

```python
# pandas is used for data manipulation.
# jellyfiish is needed for approximate string comparison and string encoding.
# numexpr is used to speed up numeric comparisons.
import pandas
import numpy
import scipy
import sklearn
import jellyfish
import numexpr
```

```python
# Load the dataset
import recordlinkage
from recordlinkage.datasets import load_febrl4
```

```python
dfA, dfB = load_febrl4()
dfA
```

```python
indexer = recordlinkage.FullIndex()
pairs = indexer.index(dfA, dfB)
```

```python
print (len(dfA), len(dfB), len(pairs))
```

Data Cleaning & Deduplication

Techniques & Tools

RecordLinkage Python Package

```python
# full code

import recordlinkage
from recordlinkage.datasets import load_febrl4

dfA, dfB = load_febrl4()

# Indexation step
indexer = recordlinkage.BlockIndex(on='given_name')
pairs = indexer.index(dfA, dfB)

# Comparison step
compare_cl = recordlinkage.Compare()

compare_cl.exact('given_name', 'given_name', label='given_name')
compare_cl.string('surname', 'surname', method='jarowinkler', threshold=0.85, label='surname')
compare_cl.exact('date_of_birth', 'date_of_birth', label='date_of_birth')
compare_cl.exact('suburb', 'suburb', label='suburb')
compare_cl.exact('state', 'state', label='state')
compare_cl.string('address_1', 'address_1', threshold=0.85, label='address_1')

features = compare_cl.compute(pairs, dfA, dfB)

# Classification step
matches = features[features.sum(axis=1) > 3]
print(len(matches))
```

Data Cleaning & Deduplication

RecordLinkage Python Package

Techniques & Tools

```
"""
The comparing of record pairs starts when the compute method is called. All
attribute comparisons are stored in a DataFrame with horizontally the features
and vertically the record pairs. The first 10 comparison vectors are:

"""
features
```

rec_id	rec_id	given_name	surname	date_of_birth	suburb	state	address_1
rec-1070-org	rec-3024-dup-0	1	0.0	0	0	1	0.0
	rec-2371-dup-0	1	0.0	0	0	0	0.0
	rec-4652-dup-0	1	0.0	0	0	0	0.0
	rec-4795-dup-0	1	0.0	0	0	1	0.0
	rec-1314-dup-0	1	0.0	0	0	1	0.0
rec-2371-org	rec-3024-dup-0	1	0.0	0	0	0	0.0
	rec-2371-dup-0	1	1.0	1	1	1	1.0
	rec-4652-dup-0	1	0.0	0	0	1	0.0
	rec-4795-dup-0	1	0.0	0	0	0	0.0
	rec-1314-dup-0	1	0.0	0	0	0	0.0

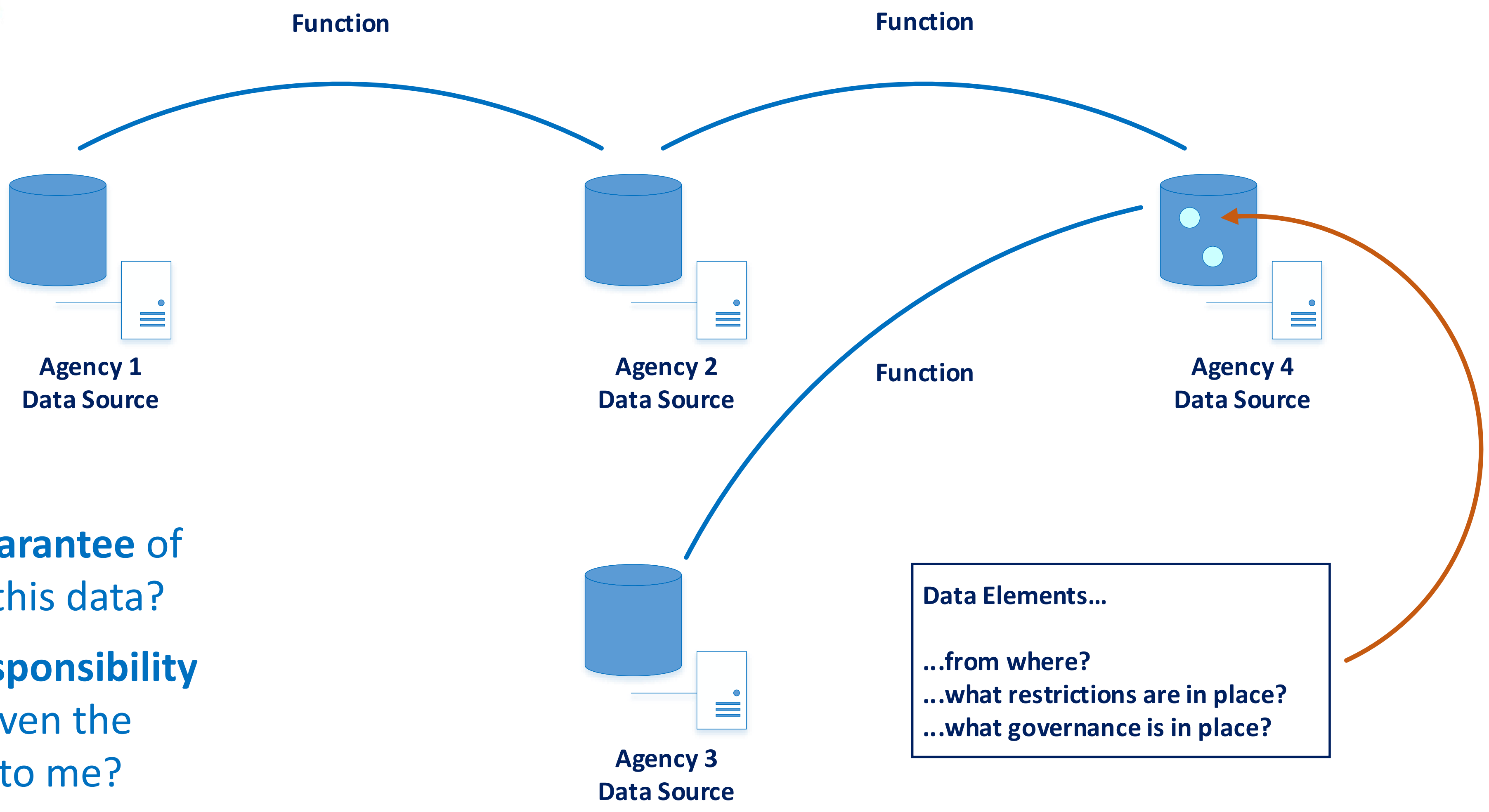
Data Management Workgroup

Data Sharing Concerns (start for next meeting)

- Do you know the lineage of your data?
- Do you know the quality of every key data element? Even data from other sources?
- Should there be an **expectation of quality** for the data being received?
  - Is there a definition of quality?
  - Is the data fit for purpose?
  - Does the data include metadata for context?
- Who should be responsible for data misuse?

Data Management Workgroup

Why Share Data



- What is my **guarantee** of the quality of this data?
- What is my **responsibility** for this data given the license issued to me?

Data Sharing Agreements

- Should the process to create these be formalized?
- What are you licensing?
  - Behavior
  - Responsibility (in both directions)
- Are you making an assessment of the data sharing party?
  - Is this context driven?
  - Are there certain licensed attributes of the party?

Data Management Workgroup

Data Sharing Agreements

- Purpose for using the data
- Justification (rationale) for issuance of license
- Licensing entity / certifications
- Audit and disclosure requirements
- Indemnification
- Limitation of liability
- Requirements for data provider (e.g., data quality, data formats, timeliness)

Data Management Workgroup

Next Topics

- Inter-agency approaches to data sharing
- Sunshine law Chapter 119, Florida Statutes
- Enabling Act: Florida Statutes, Sections 163.61-.64 (2017) permitting governmental agencies within Florida to participate in “collaborative client information systems” like the IDS
- Privacy Rules and Regulations (HIPPA, FERPA, DPPA, GDPR)
- Privacy Preserving Record Linkage
- What should the structure of an MOU/DUA/DUL be and can we streamline the process?

References

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. IEEE Transactions on knowledge and data engineering, 19(1), 1-16. Retrieved April 27, 2018 from the Purdue University, Department of Computer Science website, https://www.cs.purdue.edu/homes/ake/pub/TKDE-0240-0605-1.pdf

Askham, N., Cook, D., Doyle, M., Fereday, H., Gibson, M., Landbeck, U., ... & Schwarzenbach, J. (2013). The six primary dimensions for data quality assessment. DAMA UK Working Group, 432-435. Retrieved April 23, 2018 from EM360Tech website, https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf

De Jonge, E., & van der Loo, M. (2013). An introduction to data cleaning with R. Heerlen: Statistics Netherlands. Retrieved April 16, 2018 from The Comprehensive R Archive Network (CRAN) website, https://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf

Florida Department of State. (2012). Address Confidentiality Exemption Request Form Revised 08-2012 (2), Public Records Exemption Request to the Florida Department of State. Retrieved May 2, 2018 from the Florida Department of State website, http://dos.myflorida.com/media/696331/dos119-public-records-exemption-form.pdf

Questions?

Burt Walsh, Chief Data Officer
burt.walsh@ast.myflorida.com

Michael Avello
michael.avello@ast.myflorida.com

Agency For State Technology
4050 Esplanade Way, Suite 115
Tallahassee, Florida 32399