Organized Chaos: A Framework for Classifying Data Quality Problems

ABSTRACT

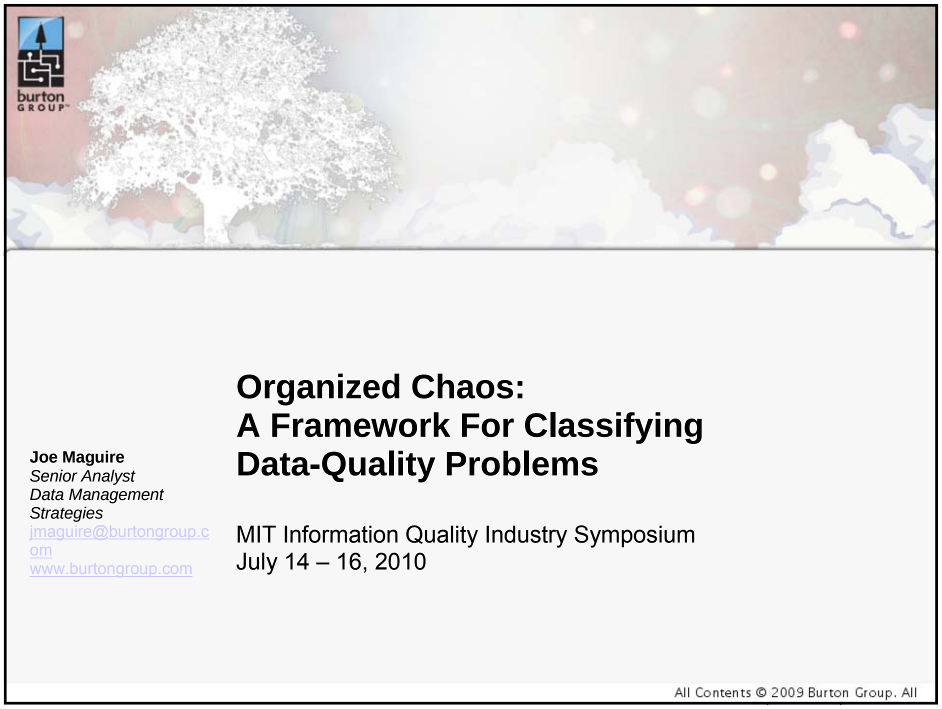
Experts leverage their experience by recognizing patterns. Consciously or unconsciously, experts classify their patterns into mental schemata that formalize their knowledge. Expertise in data quality (DQ) problems is amenable to this approach, but is challenging because DQ problems are about mistakes, and mistakes are a disorderly phenomenon. Nevertheless, a content-neutral framework for classifying DQ problems is possible. The framework presented here helps organizations maximize the value of their experience solving DQ problems in one domain into accelerated solutions in other domains.

Such a framework is especially helpful to practitioners of Master Data Management (MDM), because many approaches to MDM encourage "vertical" thinking, in which businesses contemplate and classify MDM problems according to business topics, such as customers, products, addresses, or other content-specific domains. Such thinking can help businesses organize and prioritize MDM initiatives, but it obscures underlying similarities among MDM problems (and the attendant solutions).

BIOGRAPHY

Joe Maguire
Senior Analyst
Burton Group

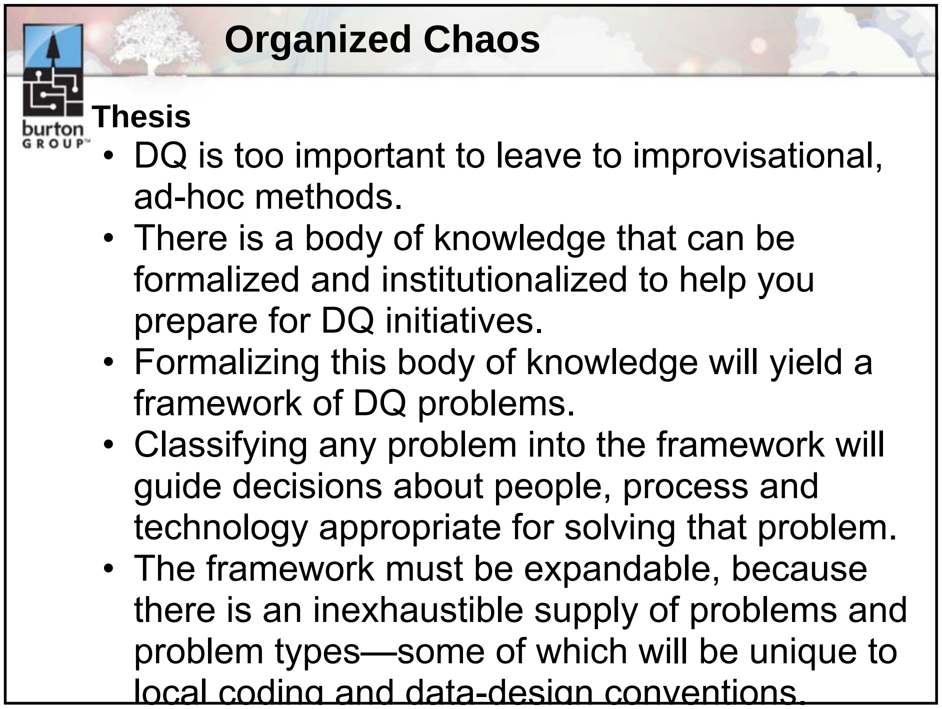
Joe Maguire is Senior Analyst at Burton Group specializing in data-modeling techniques and tools. During his 28 years in the software industry, he has worked in product development (for Digital, Lotus, Microsoft, and Bachman Information Systems) and has consulted for small startups and Fortune-100 companies in a wide range of industries including software R&D, pharmaceuticals, networking and telephony, mass-storage devices, publishing, and environmental engineering. His books—including Mastering Data Modeling: A User-Driven Approach (Addison-Wesley, 2000)—have been reviewed favorably by a wide range of publications including The Mathematica Journal, Science News, The Data Access Newsletter (TDAN.com), The Boston Sunday Globe, and National Public Radio.



Organized Chaos: A Framework For Classifying Data-Quality Problems

Joe Maguire
Senior Analyst
Data Management Strategies
jmaguire@burtongroup.com
www.burtongroup.com

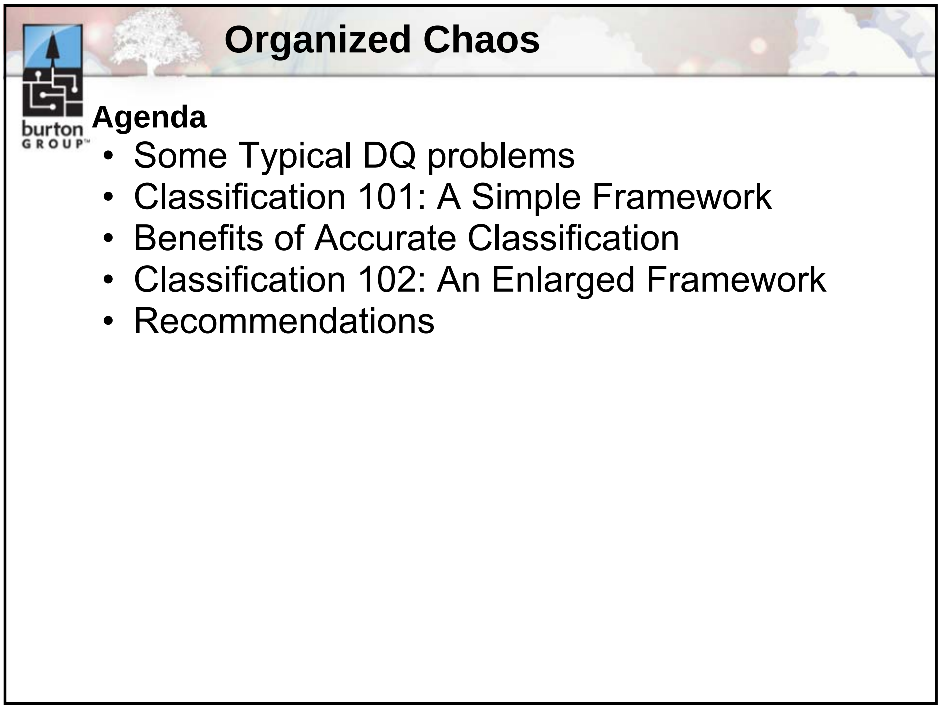
MIT Information Quality Industry Symposium
July 14 – 16, 2010



Organized Chaos

Thesis

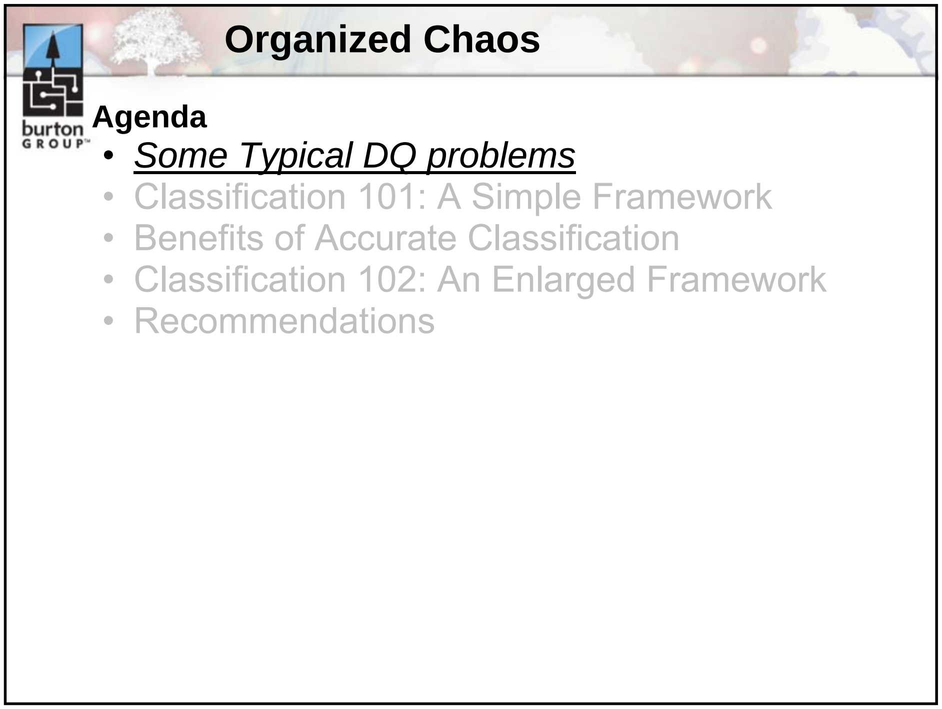
- DQ is too important to leave to improvisational, ad-hoc methods.
- There is a body of knowledge that can be formalized and institutionalized to help you prepare for DQ initiatives.
- Formalizing this body of knowledge will yield a framework of DQ problems.
- Classifying any problem into the framework will guide decisions about people, process and technology appropriate for solving that problem.
- The framework must be expandable, because there is an inexhaustible supply of problems and problem types—some of which will be unique to local coding and data-design conventions.



Organized Chaos

Agenda

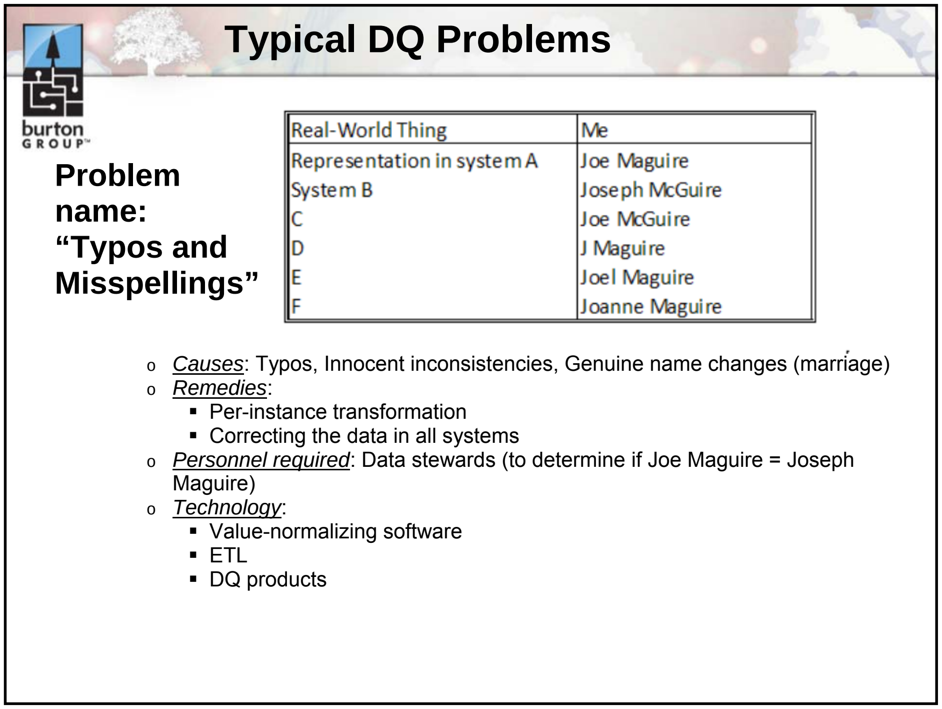
- Some Typical DQ problems
- Classification 101: A Simple Framework
- Benefits of Accurate Classification
- Classification 102: An Enlarged Framework
- Recommendations



Organized Chaos

Agenda

- *Some Typical DQ problems*
- Classification 101: A Simple Framework
- Benefits of Accurate Classification
- Classification 102: An Enlarged Framework
- Recommendations

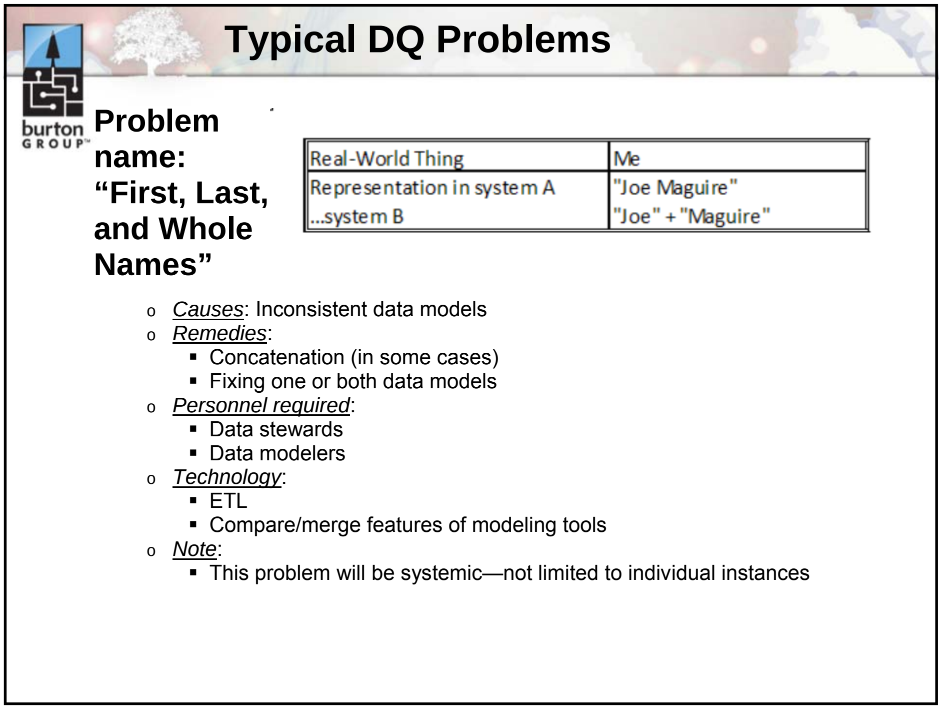


Typical DQ Problems

Problem name: “Typos and Misspellings”

Real-World Thing	Me
Representation in system A	Joe Maguire
System B	Joseph McGuire
C	Joe McGuire
D	J Maguire
E	Joel Maguire
F	Joanne Maguire

- *Causes*: Typos, Innocent inconsistencies, Genuine name changes (marriage)
- *Remedies*:
  - Per-instance transformation
  - Correcting the data in all systems
- *Personnel required*: Data stewards (to determine if Joe Maguire = Joseph Maguire)
- *Technology*:
  - Value-normalizing software
  - ETL
  - DQ products



Typical DQ Problems

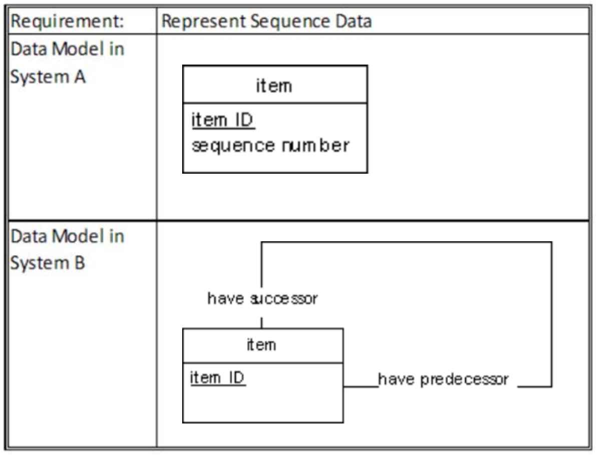
Problem name: “First, Last, and Whole Names”

Real-World Thing	Me
Representation in system A	"Joe Maguire"
...system B	"Joe" + "Maguire"

- *Causes*: Inconsistent data models
- *Remedies*:
  - Concatenation (in some cases)
  - Fixing one or both data models
- *Personnel required*:
  - Data stewards
  - Data modelers
- *Technology*:
  - ETL
  - Compare/merge features of modeling tools
- *Note*:
  - This problem will be systemic—not limited to individual instances

Typical DQ Problems

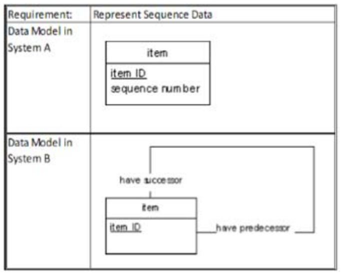
Problem name: “Models for Sequence Data”



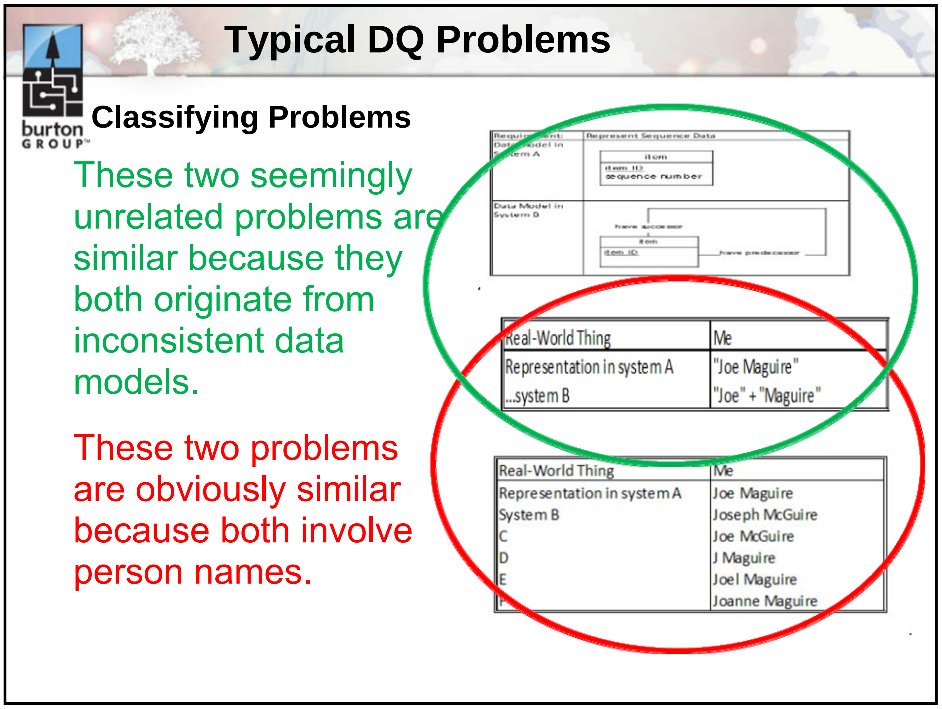
Different systems have subtly different semantics.

Typical DQ Problems

Different systems have subtly different semantics.



- *Causes*: Inconsistent data models
- *Remedies*:
  - Data transformations
  - Fixing one or both data models
- *Personnel required*:
  - Data stewards
  - Data modelers
- *Technology*:
  - ETL
  - Compare/merge features of modeling tools
- *Note*:
  - This problem will be systemic—not limited to individual instances



Typical DQ Problems

Classifying Problems

These two seemingly unrelated problems are similar because they both originate from inconsistent data models.

These two problems are obviously similar because both involve person names.

Requirement:	Represent Sequence Data
Data Model in System A	item item ID sequence number
Data Model in System B	have successor item item ID have predecessor

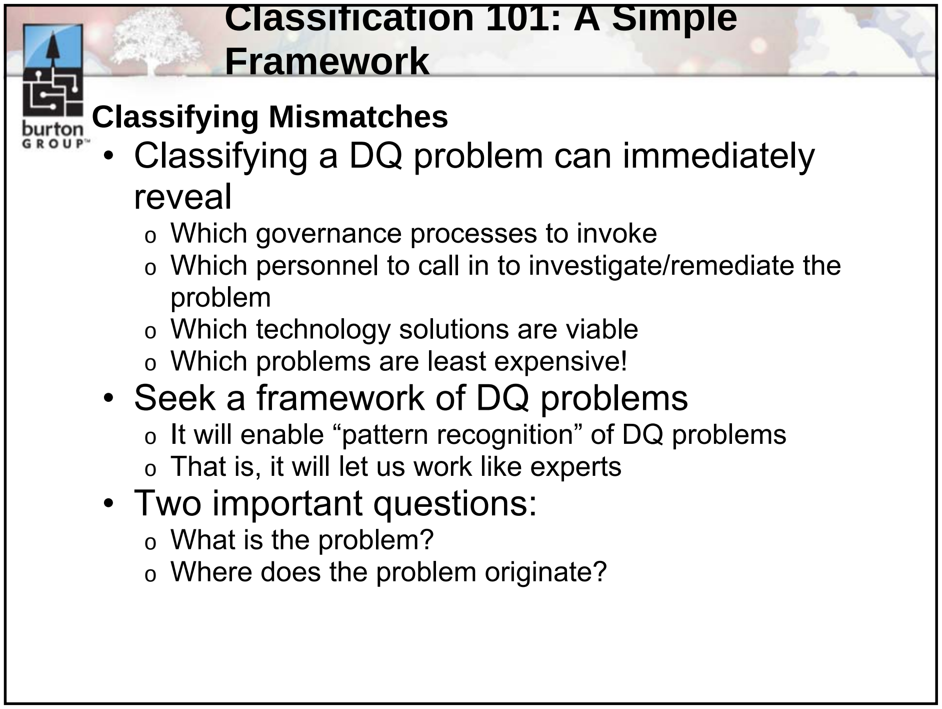
Real-World Thing	Me
Representation in system A	"Joe Maguire"
...system B	"Joe" + "Maguire"

Real-World Thing	Me
Representation in system A	Joe Maguire
System B	Joseph McGuire
C	Joe McGuire
D	J Maguire
E	Joel Maguire
F	Joanne Maguire

Organized Chaos

Agenda

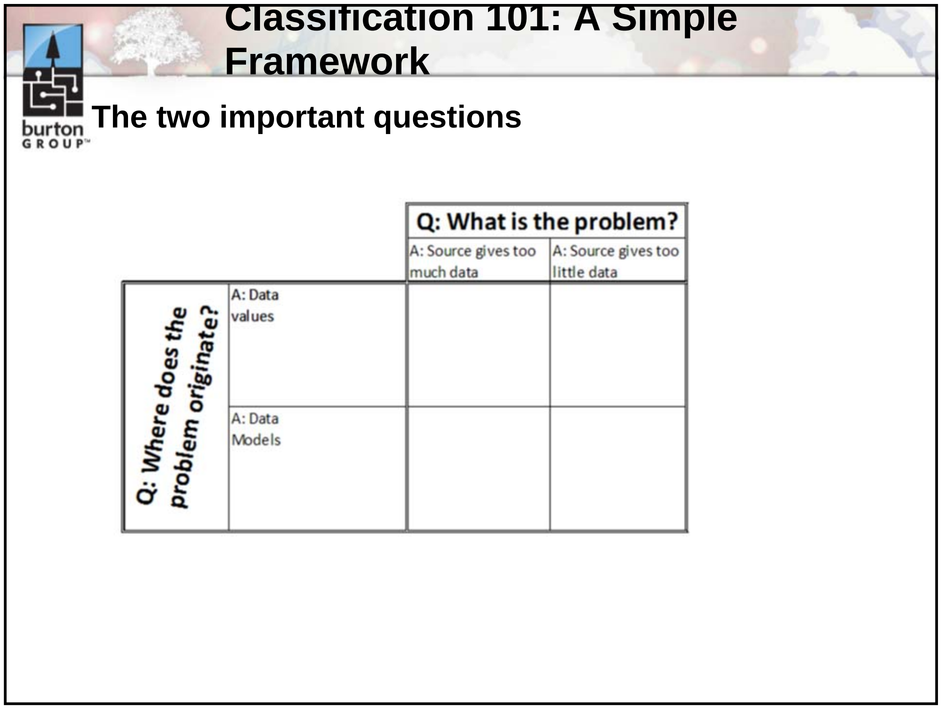
- Some Typical DQ problems
- *Classification 101: A Simple Framework*
- Benefits of Accurate Classification
- Classification 102: An Enlarged Framework
- Recommendations



Classification 101: A Simple Framework

Classifying Mismatches

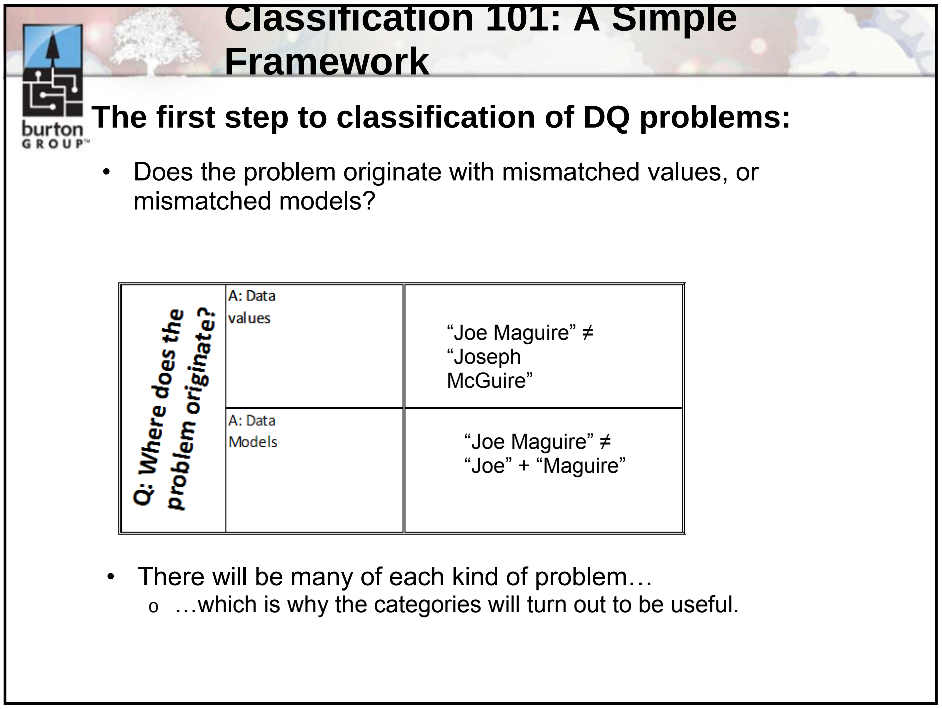
- Classifying a DQ problem can immediately reveal
  - Which governance processes to invoke
  - Which personnel to call in to investigate/remediate the problem
  - Which technology solutions are viable
  - Which problems are least expensive!
- Seek a framework of DQ problems
  - It will enable "pattern recognition" of DQ problems
  - That is, it will let us work like experts
- Two important questions:
  - What is the problem?
  - Where does the problem originate?



Classification 101: A Simple Framework

The two important questions

		Q: What is the problem?	
		A: Source gives too much data	A: Source gives too little data
Q: Where does the problem originate?	A: Data values		
	A: Data Models		



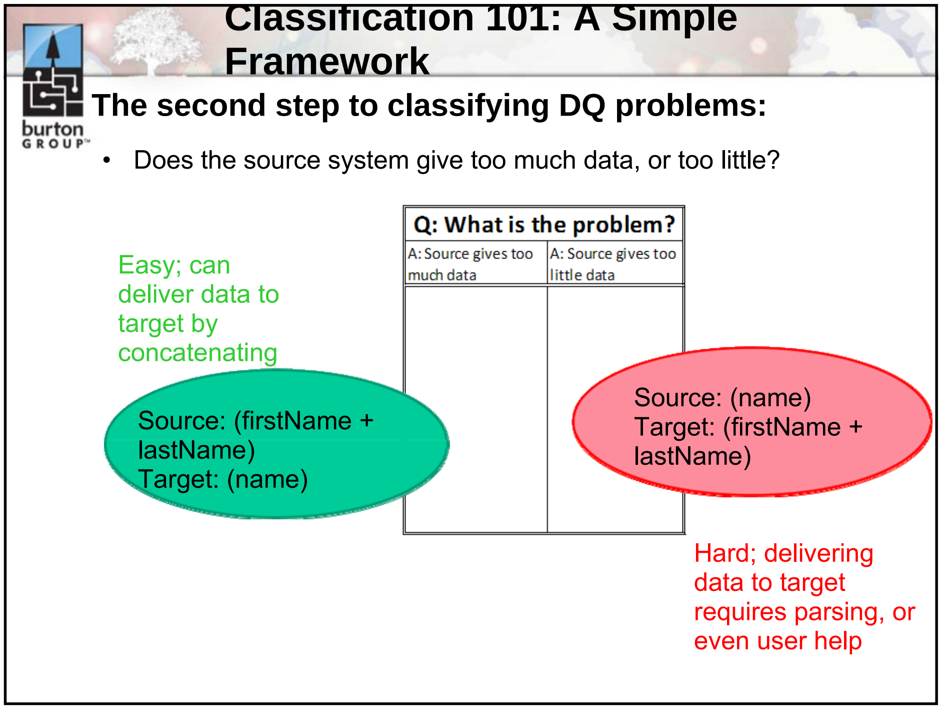
Classification 101: A Simple Framework

The first step to classification of DQ problems:

- Does the problem originate with mismatched values, or mismatched models?

Q: Where does the problem originate?		
	A: Data values	“Joe Maguire” ≠ “Joseph McGuire”
	A: Data Models	“Joe Maguire” ≠ “Joe” + “Maguire”

- There will be many of each kind of problem…
  - …which is why the categories will turn out to be useful.



Classification 101: A Simple Framework

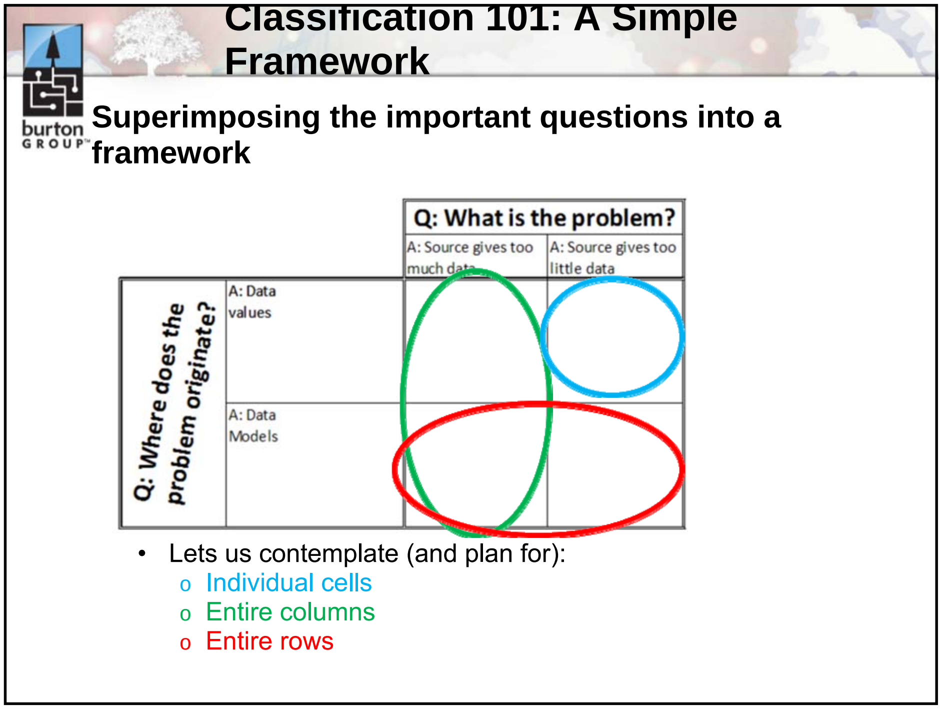
The second step to classifying DQ problems:

- Does the source system give too much data, or too little?

Q: What is the problem?	
A: Source gives too much data	A: Source gives too little data
Source: (firstName + lastName) Target: (name)	Source: (name) Target: (firstName + lastName)

Easy; can deliver data to target by concatenating

Hard; delivering data to target requires parsing, or even user help

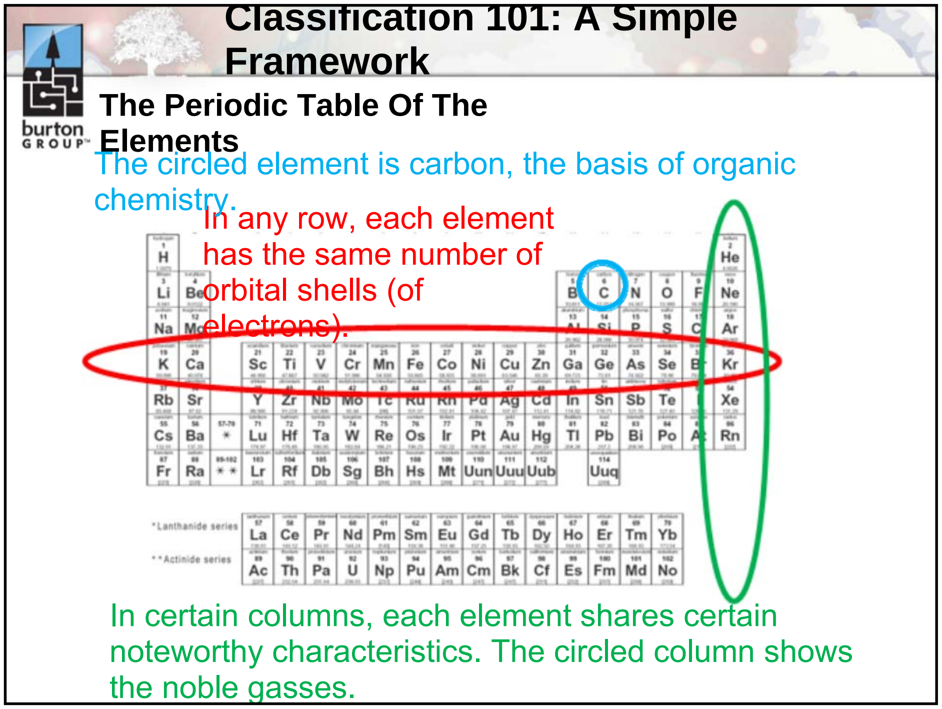


Classification 101: A Simple Framework

Superimposing the important questions into a framework

Q: Where does the problem originate?		Q: What is the problem?	
		A: Source gives too much data	A: Source gives too little data
	A: Data values		
	A: Data Models		

- Lets us contemplate (and plan for):
  - Individual cells
  - Entire columns
  - Entire rows



Classification 101: A Simple Framework

The Periodic Table Of The Elements

The circled element is carbon, the basis of organic chemistry.

In any row, each element has the same number of orbital shells (of electrons).

H																		He
Li	Be												B	C	N	O	F	Ne
Na	Mg												Al	Si	P	S	Cl	Ar
K	Ca		Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Sr		Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
Cs	Ba	57-70 *	Lu	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
Fr	Ra	89-102 * *	Lr	Rf	Db	Sg	Bh	Hs	Mt	Uun	Uuu	Uub		Uuq				

*Lanthanide series	La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb
* *Actinide series	Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No

In certain columns, each element shares certain noteworthy characteristics. The circled column shows the noble gasses.

Organized Chaos

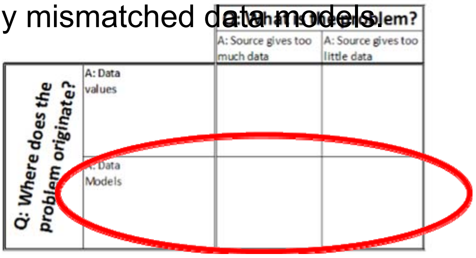
Agenda

- Some Typical DQ problems
- Classification 101: A Simple Framework
- ***<u>Benefits of Accurate Classification</u>***
- Classification 102: An Enlarged Framework
- Recommendations

Benefits of Accurate Classification

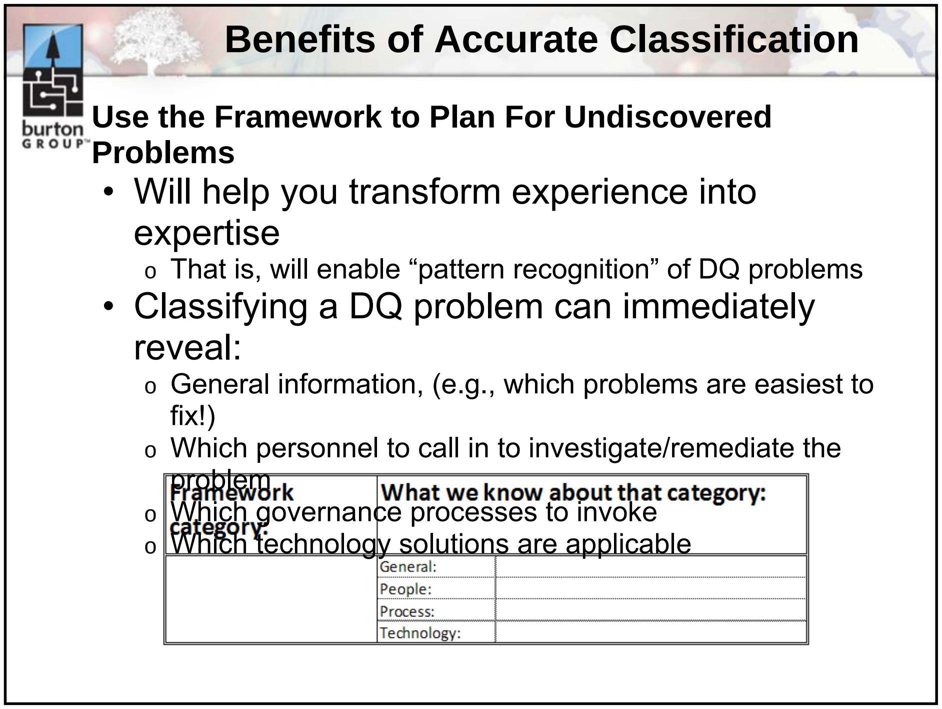
Example of the benefits for problems originating in data models

- There should be a certain set of repeatable behaviors that we invoke every time we encounter a problem causes by mismatched data models.



		Q: What is the problem?	
		A: Source gives too much data	A: Source gives too little data
Q: Where does the problem originate?	A: Data values		
	A: Data Models		

- We don't have to anticipate every possible data-model mismatch, we can benefit merely by including a category in the framework for data-model mismatches.
- In other words, we can plan for undiscovered problems.

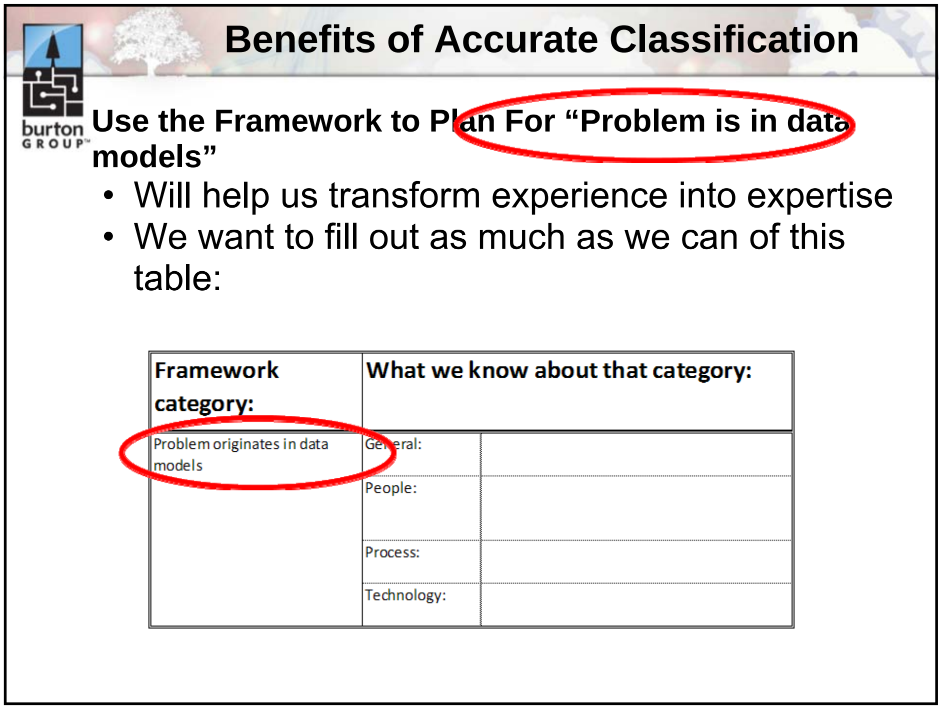


Benefits of Accurate Classification

Use the Framework to Plan For Undiscovered Problems

- Will help you transform experience into expertise
  - That is, will enable "pattern recognition" of DQ problems
- Classifying a DQ problem can immediately reveal:
  - General information, (e.g., which problems are easiest to fix!)
  - Which personnel to call in to investigate/remediate the problem
  - Which governance processes to invoke
  - Which technology solutions are applicable

Framework category:	**What we know about that category:**	
	General:	
	People:	
	Process:	
	Technology:	

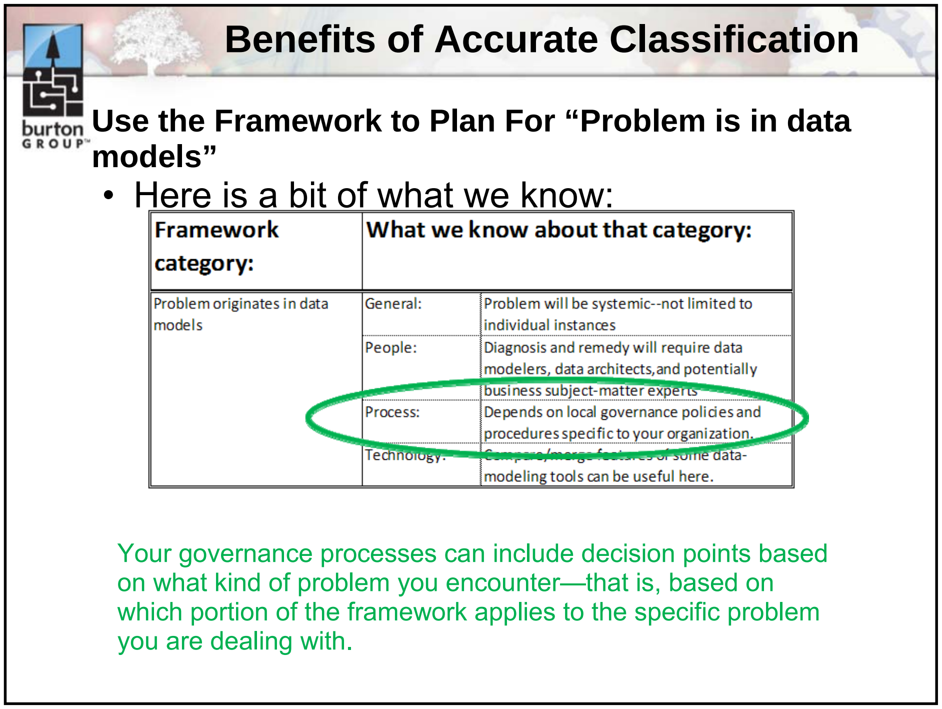


Benefits of Accurate Classification

Use the Framework to Plan For "Problem is in data models"

- Will help us transform experience into expertise
- We want to fill out as much as we can of this table:

Framework category:	**What we know about that category:**	
Problem originates in data models	General:	
	People:	
	Process:	
	Technology:	



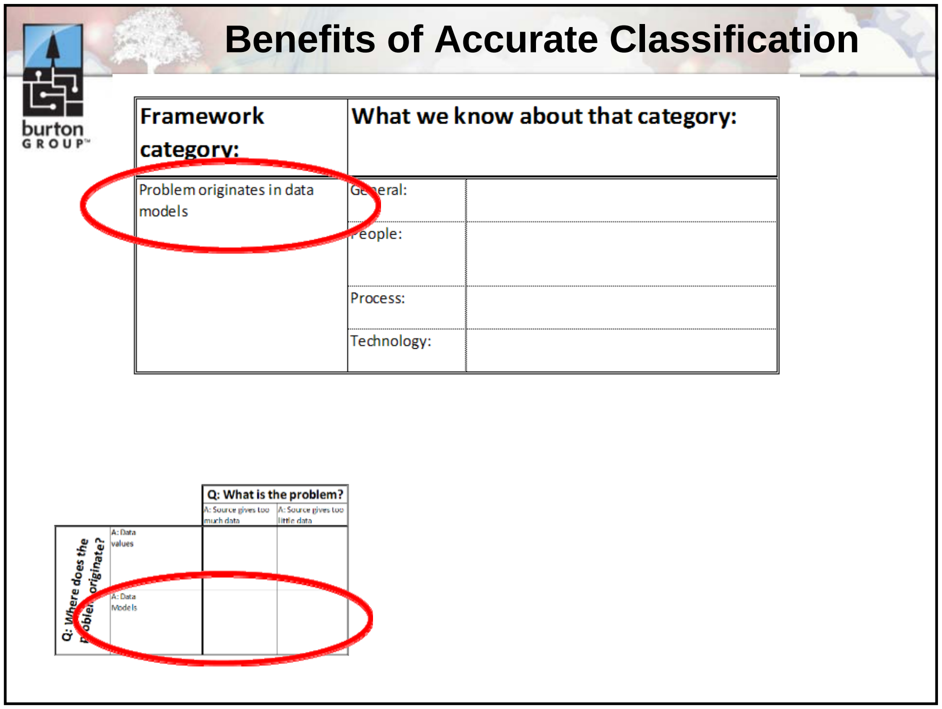
Benefits of Accurate Classification

Use the Framework to Plan For “Problem is in data models”

- Here is a bit of what we know:

Framework category:	What we know about that category:	
Problem originates in data models	General:	Problem will be systemic--not limited to individual instances
	People:	Diagnosis and remedy will require data modelers, data architects, and potentially business subject-matter experts
	Process:	Depends on local governance policies and procedures specific to your organization.
	Technology:	Compare/merge features of some data-modeling tools can be useful here.

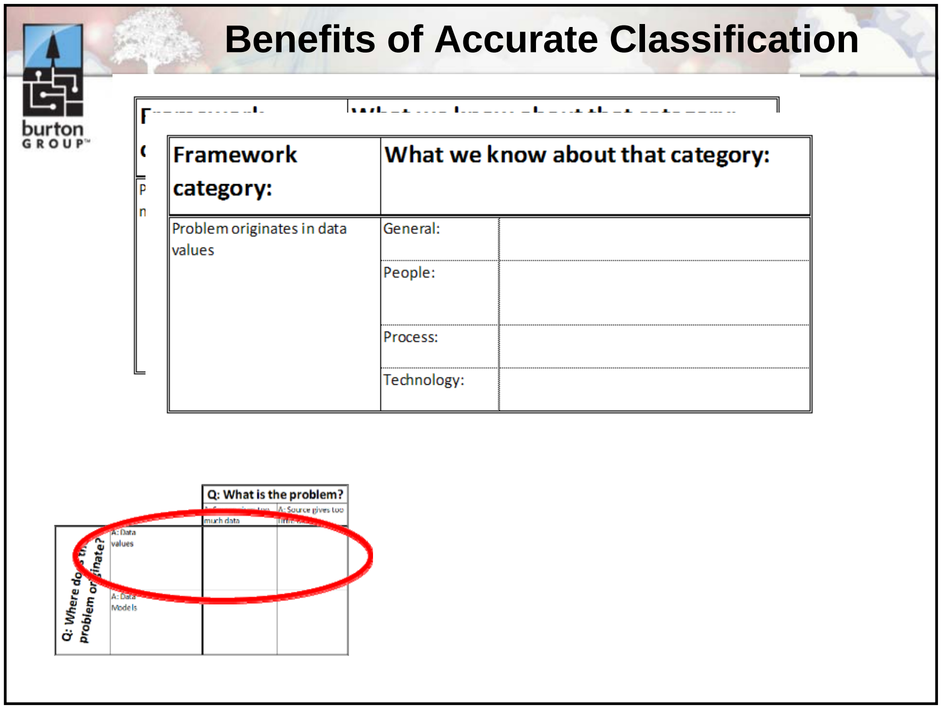
Your governance processes can include decision points based on what kind of problem you encounter—that is, based on which portion of the framework applies to the specific problem you are dealing with.



Benefits of Accurate Classification

Framework category:	What we know about that category:	
Problem originates in data models	General:	
	People:	
	Process:	
	Technology:	

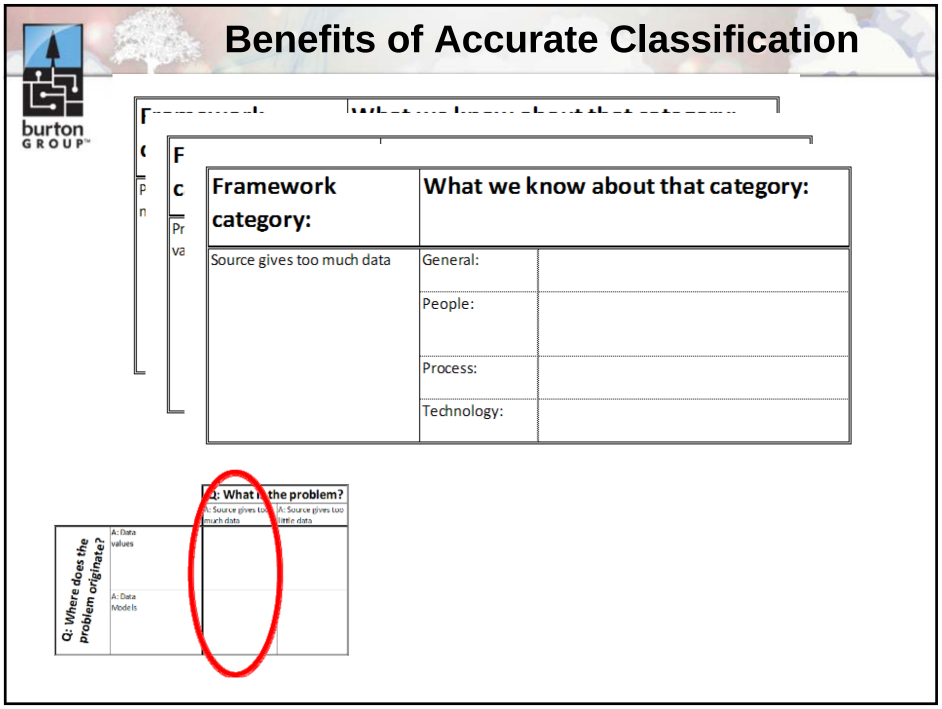
		Q: What is the problem?	
		A: Source gives too much data	A: Source gives too little data
Q: Where does the problem originate?	A: Data values		
	A: Data Models		



Benefits of Accurate Classification

Framework category:	What we know about that category:	
Problem originates in data values	General:	
	People:	
	Process:	
	Technology:	

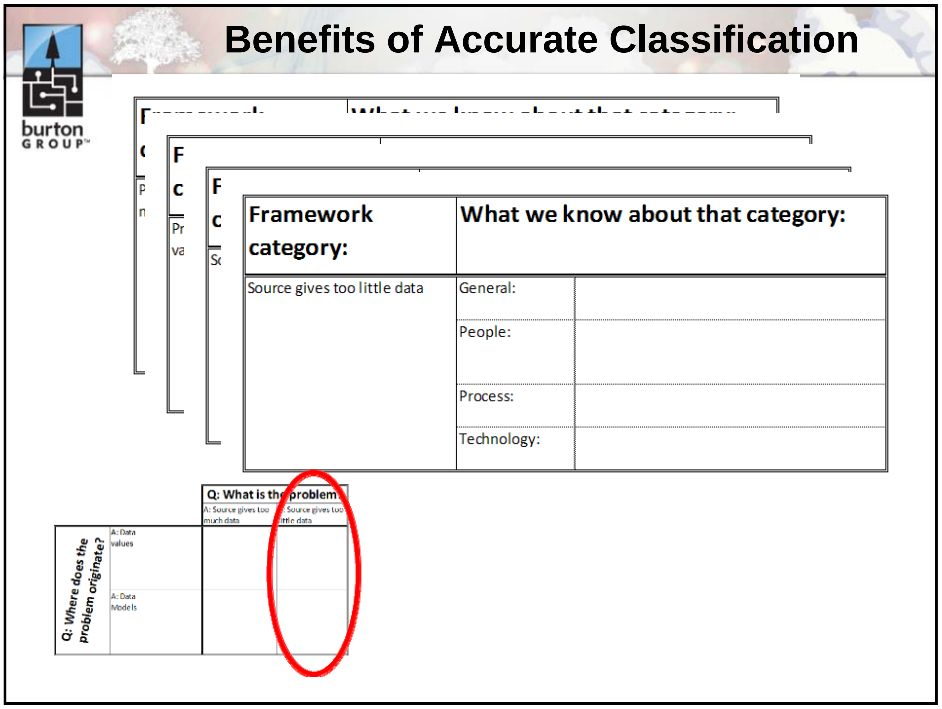
		Q: What is the problem?	
		A: Source gives too much data	A: Source gives too little data
Q: Where does the problem originate?	A: Data values		
	A: Data Models		



Benefits of Accurate Classification

Framework category:	What we know about that category:	
Source gives too much data	General:	
	People:	
	Process:	
	Technology:	

		Q: What is the problem?	
		A: Source gives too much data	A: Source gives too little data
Q: Where does the problem originate?	A: Data values		
	A: Data Models		

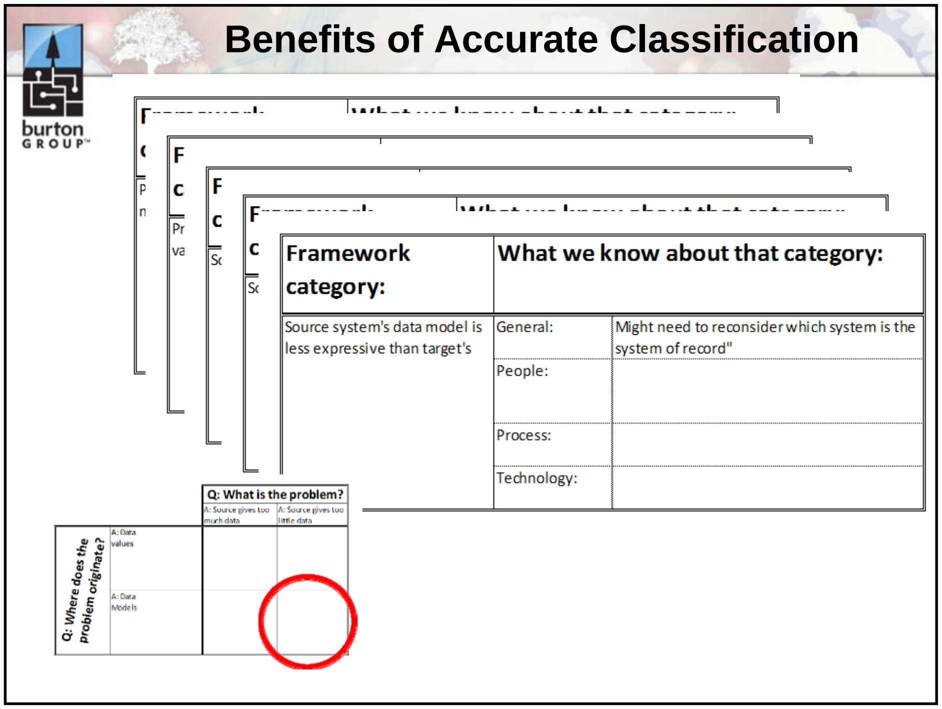


Benefits of Accurate Classification

burton GROUP

Framework category:	What we know about that category:	
Source gives too little data	General:	
	People:	
	Process:	
	Technology:	

		Q: What is the problem?	
		A: Source gives too much data	A: Source gives too little data
Q: Where does the problem originate?	A: Data values		
	A: Data Models		

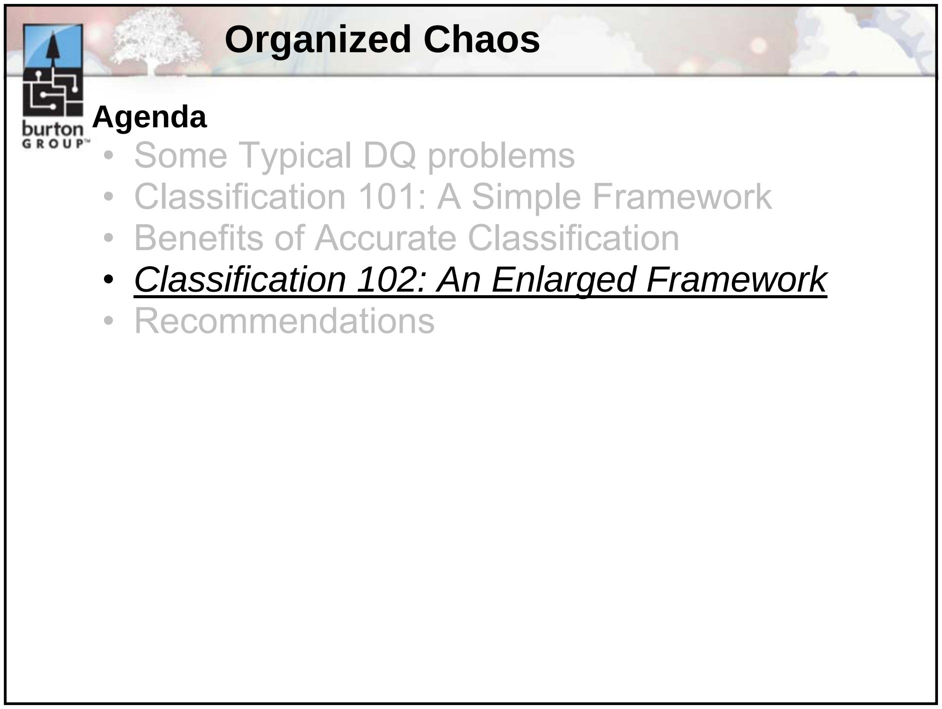


Benefits of Accurate Classification

burton GROUP

Framework category:	What we know about that category:	
Source system's data model is less expressive than target's	General:	Might need to reconsider which system is the system of record"
	People:	
	Process:	
	Technology:	

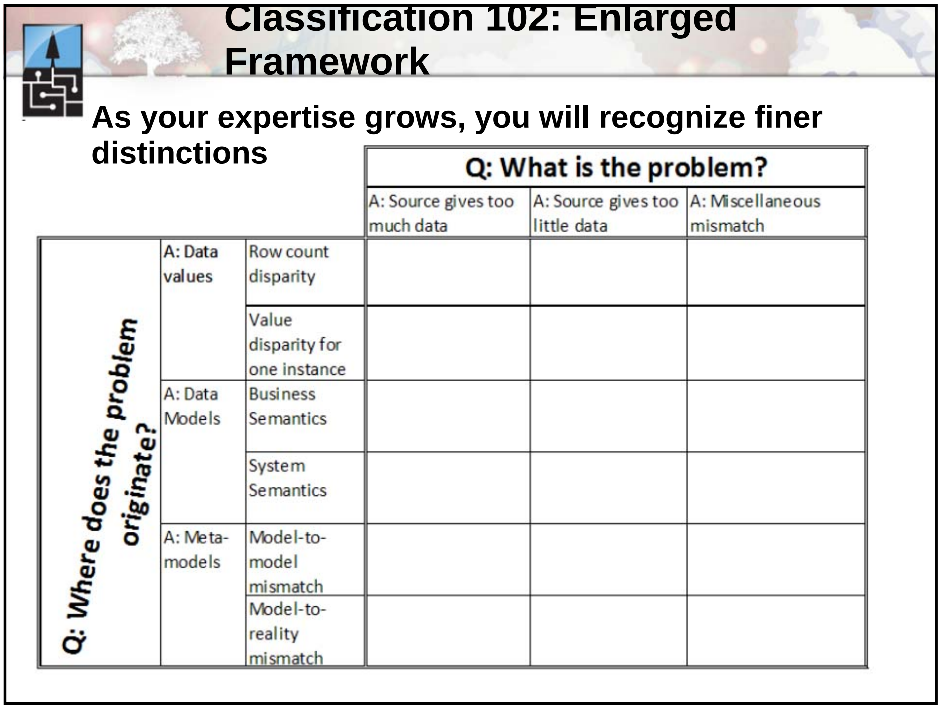
		Q: What is the problem?	
		A: Source gives too much data	A: Source gives too little data
Q: Where does the problem originate?	A: Data values		
	A: Data Models		



Organized Chaos

Agenda

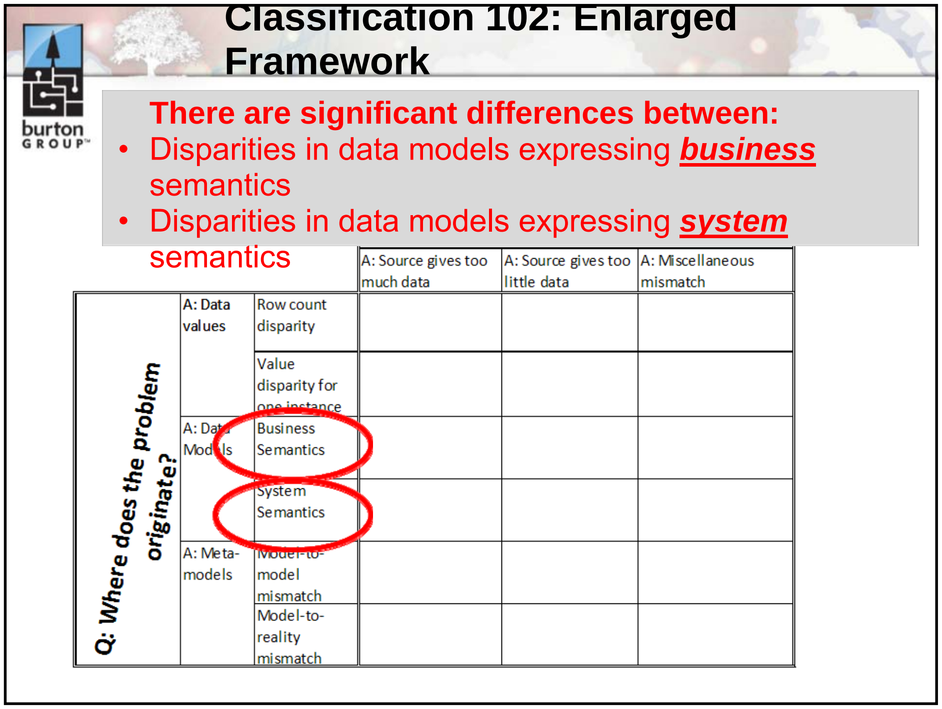
- Some Typical DQ problems
- Classification 101: A Simple Framework
- Benefits of Accurate Classification
- *Classification 102: An Enlarged Framework*
- Recommendations



Classification 102: Enlarged Framework

As your expertise grows, you will recognize finer distinctions

			Q: What is the problem?		
			A: Source gives too much data	A: Source gives too little data	A: Miscellaneous mismatch
Q: Where does the problem originate?	A: Data values	Row count disparity			
		Value disparity for one instance			
	A: Data Models	Business Semantics			
		System Semantics			
	A: Meta-models	Model-to-model mismatch			
		Model-to-reality mismatch			



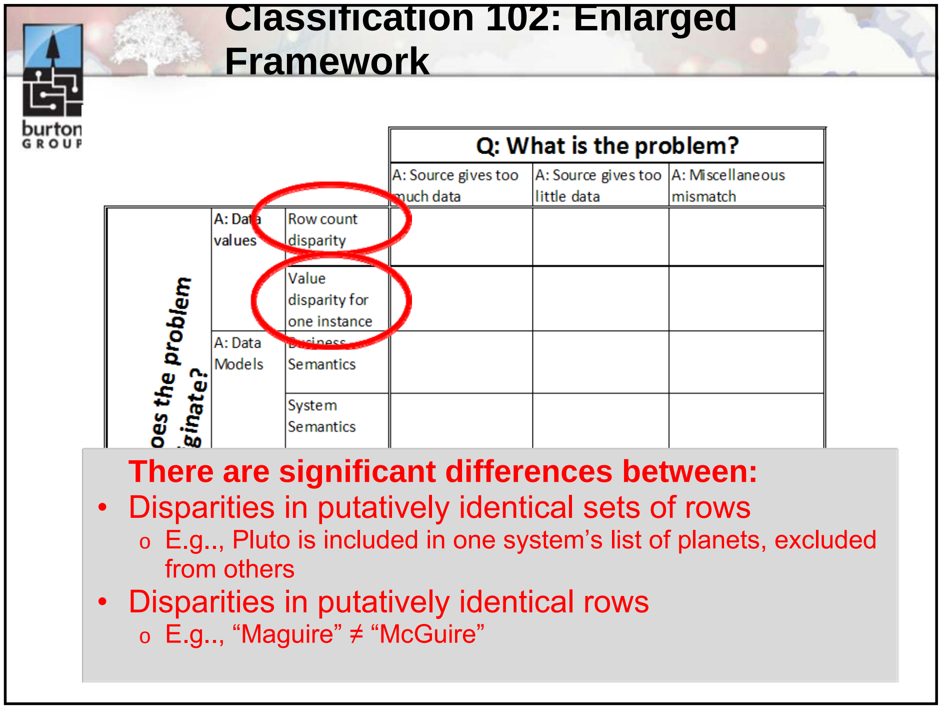
Classification 102: Enlarged Framework

burton GROUP

There are significant differences between:

- Disparities in data models expressing ***business*** semantics
- Disparities in data models expressing ***system*** semantics

Q: Where does the problem originate?			A: Source gives too much data	A: Source gives too little data	A: Miscellaneous mismatch
	A: Data values	Row count disparity			
		Value disparity for one instance			
	A: Data Models	Business Semantics			
		System Semantics			
	A: Meta-models	Model-to-model mismatch			
		Model-to-reality mismatch			



Classification 102: Enlarged Framework

burton GROUP

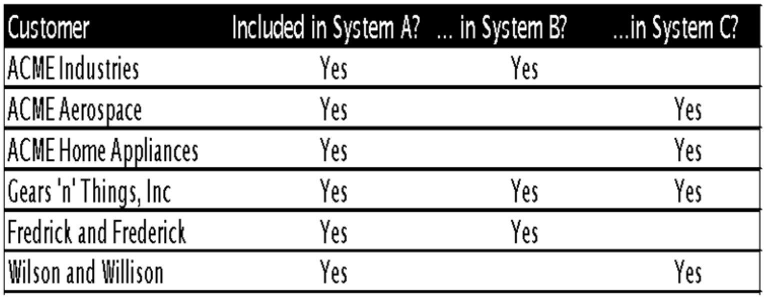
			Q: What is the problem?		
			A: Source gives too much data	A: Source gives too little data	A: Miscellaneous mismatch
oes the problem ginate?	A: Data values	Row count disparity			
		Value disparity for one instance			
	A: Data Models	Business Semantics			
		System Semantics			

There are significant differences between:

- Disparities in putatively identical sets of rows
  - E.g.., Pluto is included in one system's list of planets, excluded from others
- Disparities in putatively identical rows
  - E.g.., "Maguire" ≠ "McGuire"

Classification 102: Enlarged Framework

Row count disparity: Customers



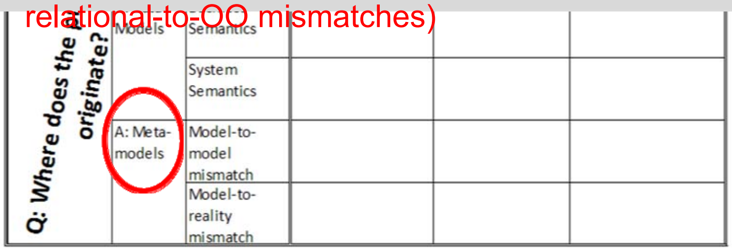
Customer	Included in System A?	... in System B?	...in System C?
ACME Industries	Yes	Yes	
ACME Aerospace	Yes		Yes
ACME Home Appliances	Yes		Yes
Gears 'n' Things, Inc	Yes	Yes	Yes
Fredrick and Frederick	Yes	Yes	
Wilson and Willison	Yes		Yes

- Which system is correct?
- Is any one of the three systems correct?
- The problem is not one or more inaccurate rows—the problem is *omitted* or *extraneous* rows.

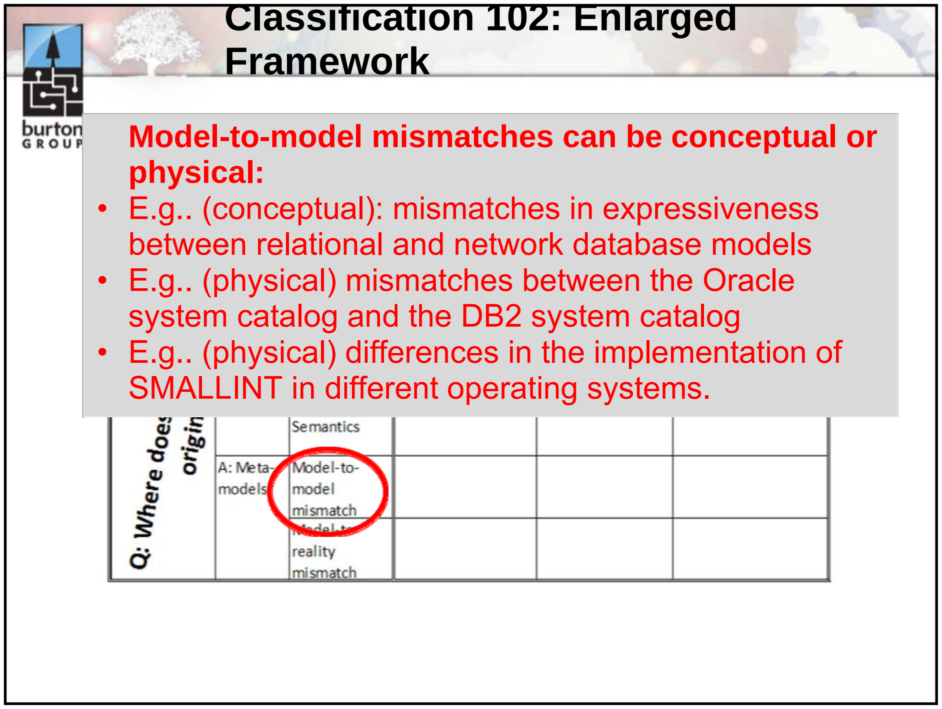
Classification 102: Enlarged Framework

One source of problems is mismatched metamodels:

- E.g.., mismatches in expressiveness between relational and network database models
- Not all of these mismatches are DQ problems, some exist between layers of the application stack (e.g., relational-to-OO mismatches)



Q: Where does the p originate?	Models	Semantics			
		System Semantics			
	A: Meta-models	Model-to-model mismatch			
		Model-to-reality mismatch			

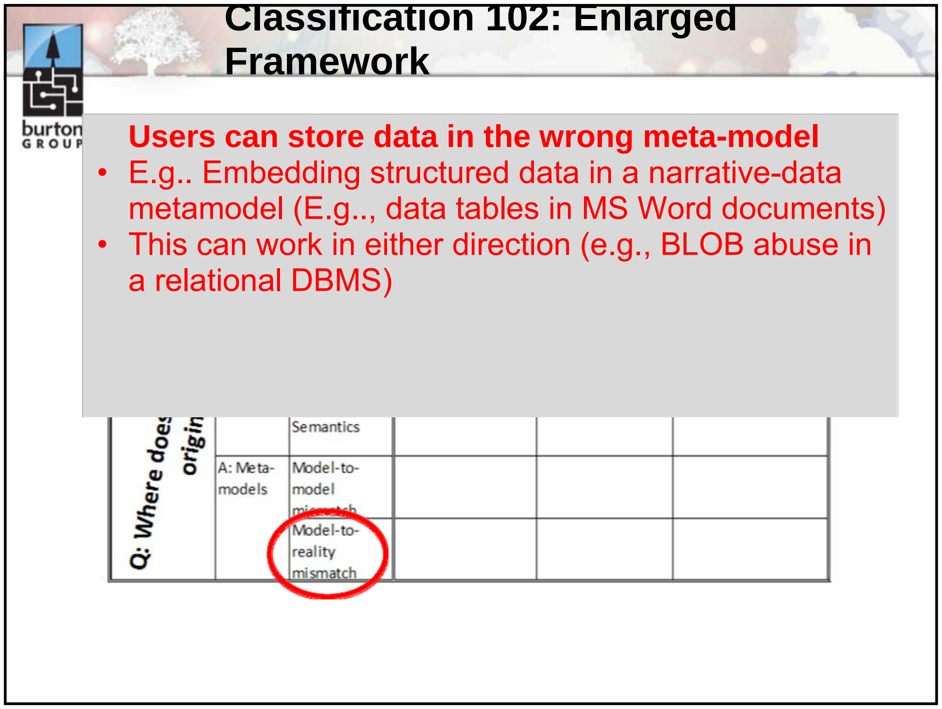


Classification 102: Enlarged Framework

Model-to-model mismatches can be conceptual or physical:

- E.g.. (conceptual): mismatches in expressiveness between relational and network database models
- E.g.. (physical) mismatches between the Oracle system catalog and the DB2 system catalog
- E.g.. (physical) differences in the implementation of SMALLINT in different operating systems.

Q: Where does origin		Semantics			
	A: Meta-models	Model-to-model mismatch			
		Model-to-reality mismatch			

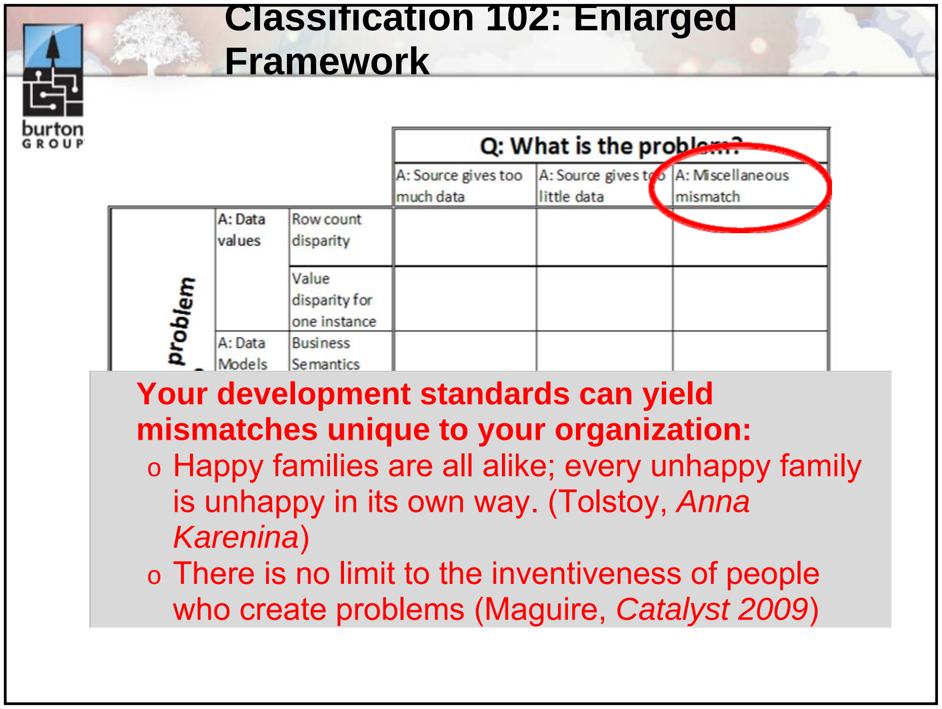


Classification 102: Enlarged Framework

Users can store data in the wrong meta-model

- E.g.. Embedding structured data in a narrative-data metamodel (E.g.., data tables in MS Word documents)
- This can work in either direction (e.g., BLOB abuse in a relational DBMS)

Q: Where does origin		Semantics			
	A: Meta-models	Model-to-model mismatch			
		Model-to-reality mismatch			

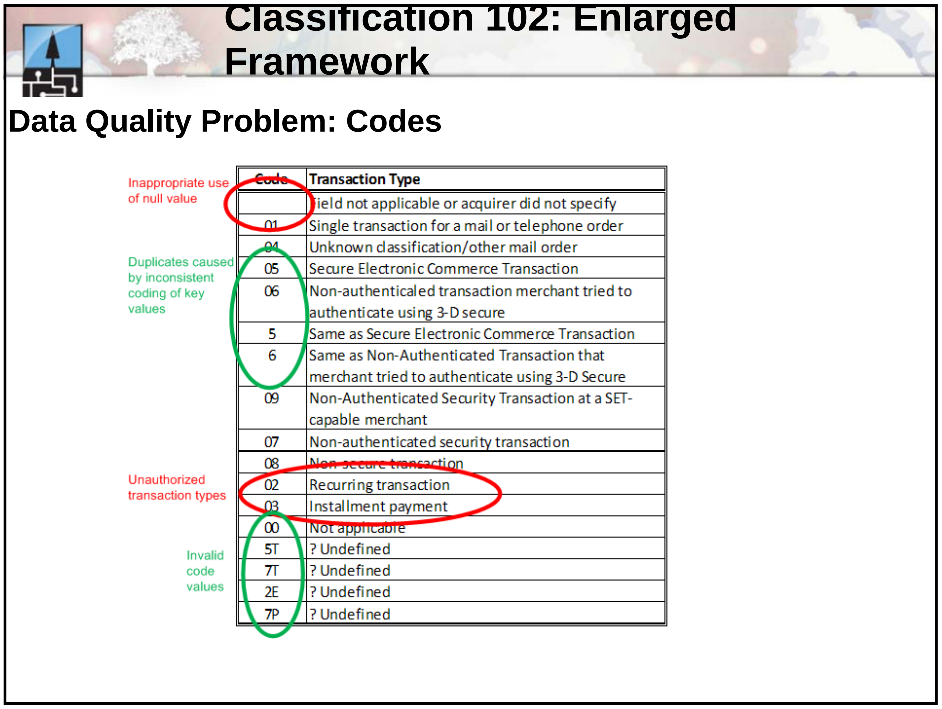


Classification 102: Enlarged Framework

			Q: What is the problem?		
			A: Source gives too much data	A: Source gives too little data	A: Miscellaneous mismatch
problem	A: Data values	Row count disparity			
		Value disparity for one instance			
	A: Data Models	Business Semantics			

Your development standards can yield mismatches unique to your organization:

- Happy families are all alike; every unhappy family is unhappy in its own way. (Tolstoy, *Anna Karenina*)
- There is no limit to the inventiveness of people who create problems (Maguire, *Catalyst 2009*)



Classification 102: Enlarged Framework

Data Quality Problem: Codes

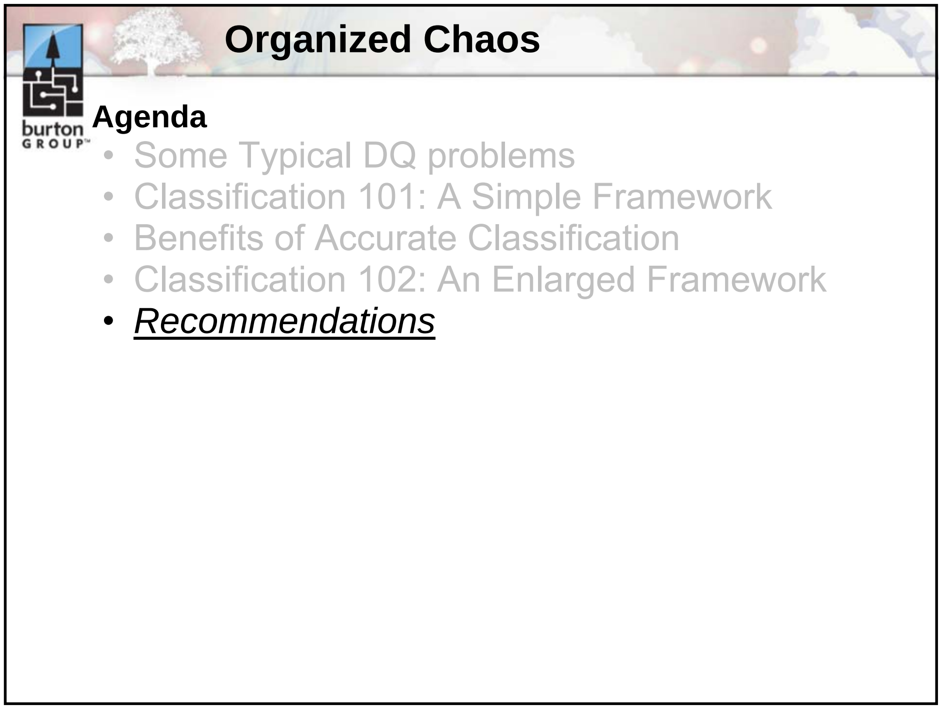
Code	Transaction Type
	Field not applicable or acquirer did not specify
01	Single transaction for a mail or telephone order
04	Unknown classification/other mail order
05	Secure Electronic Commerce Transaction
06	Non-authenticaled transaction merchant tried to authenticate using 3-D secure
5	Same as Secure Electronic Commerce Transaction
6	Same as Non-Authenticated Transaction that merchant tried to authenticate using 3-D Secure
09	Non-Authenticated Security Transaction at a SET-capable merchant
07	Non-authenticated security transaction
08	Non-secure transaction
02	Recurring transaction
03	Installment payment
00	Not applicable
5T	? Undefined
7T	? Undefined
2E	? Undefined
7P	? Undefined

Inappropriate use of null value

Duplicates caused by inconsistent coding of key values

Unauthorized transaction types

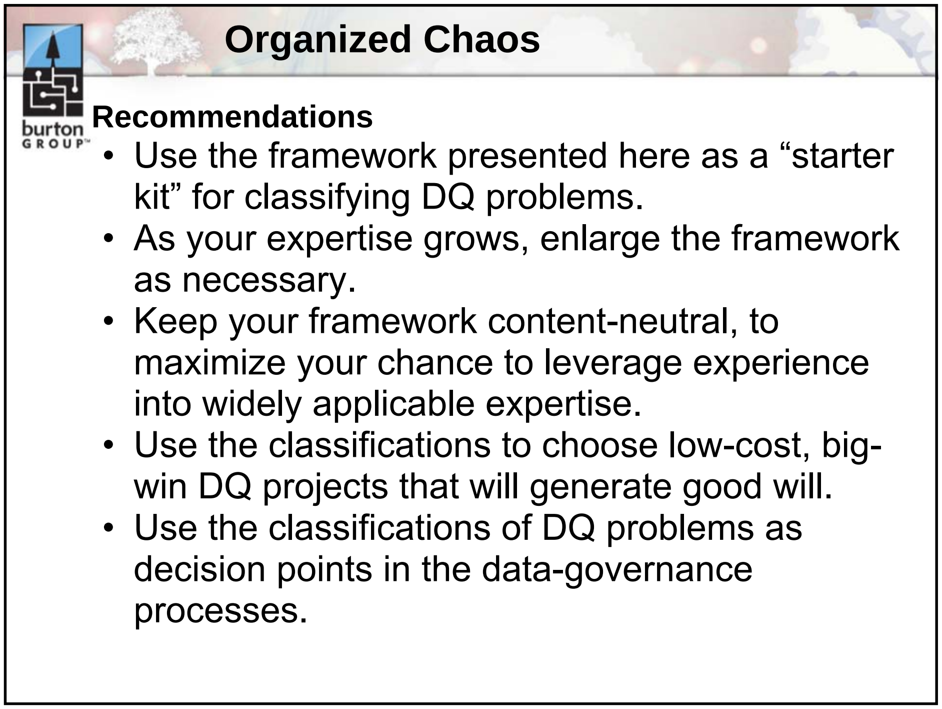
Invalid code values



Organized Chaos

Agenda

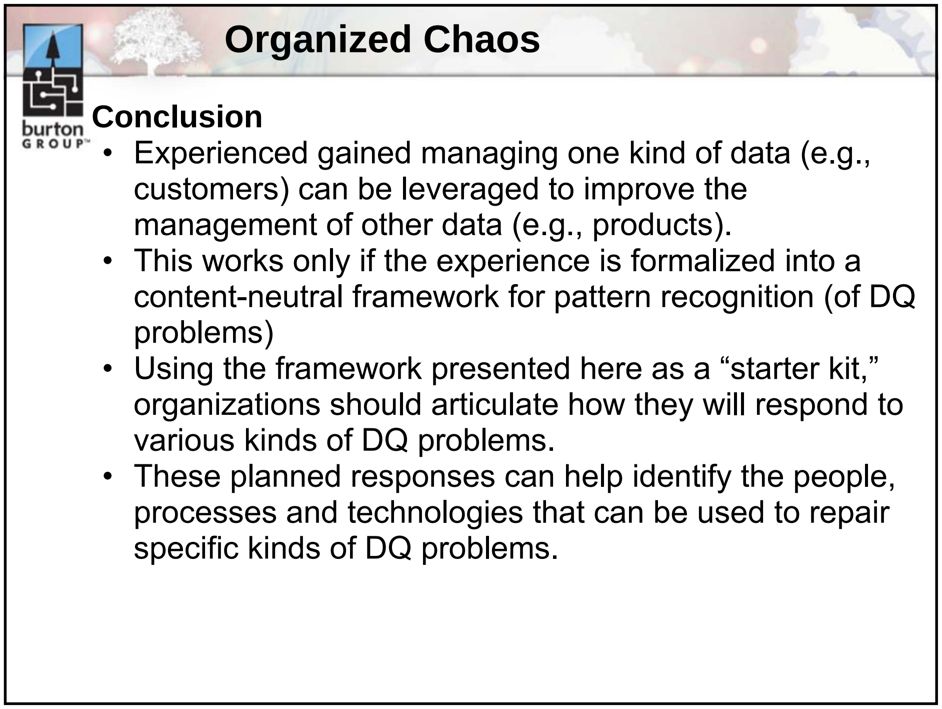
- Some Typical DQ problems
- Classification 101: A Simple Framework
- Benefits of Accurate Classification
- Classification 102: An Enlarged Framework
- *Recommendations*



Organized Chaos

Recommendations

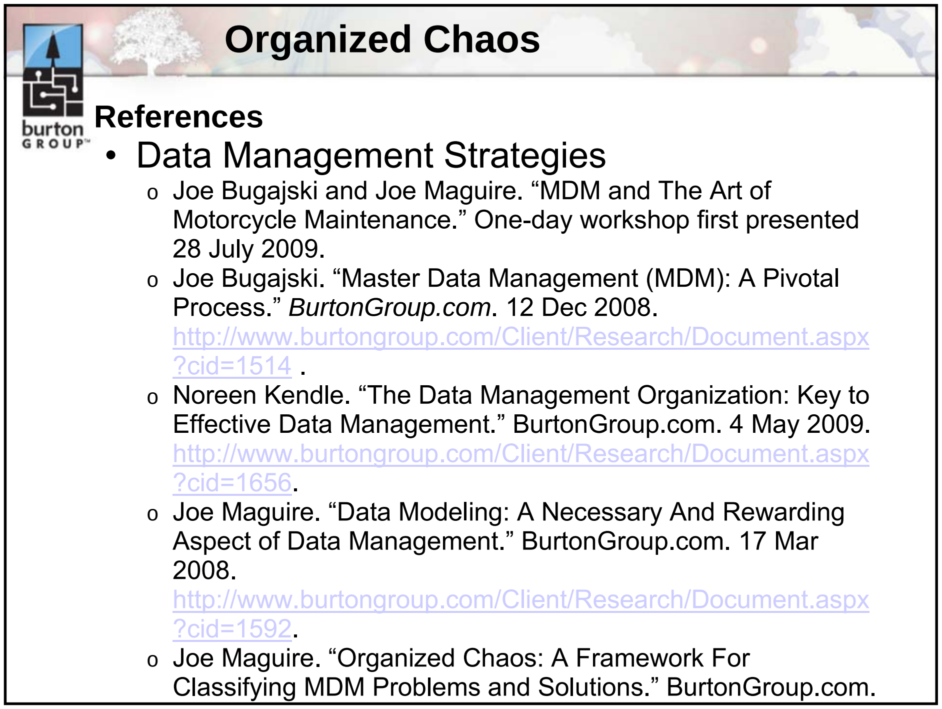
- Use the framework presented here as a "starter kit" for classifying DQ problems.
- As your expertise grows, enlarge the framework as necessary.
- Keep your framework content-neutral, to maximize your chance to leverage experience into widely applicable expertise.
- Use the classifications to choose low-cost, big-win DQ projects that will generate good will.
- Use the classifications of DQ problems as decision points in the data-governance processes.



Organized Chaos

Conclusion

- Experienced gained managing one kind of data (e.g., customers) can be leveraged to improve the management of other data (e.g., products).
- This works only if the experience is formalized into a content-neutral framework for pattern recognition (of DQ problems)
- Using the framework presented here as a "starter kit," organizations should articulate how they will respond to various kinds of DQ problems.
- These planned responses can help identify the people, processes and technologies that can be used to repair specific kinds of DQ problems.



Organized Chaos

References

- Data Management Strategies

  - Joe Bugajski and Joe Maguire. "MDM and The Art of Motorcycle Maintenance." One-day workshop first presented 28 July 2009.
  - Joe Bugajski. "Master Data Management (MDM): A Pivotal Process." *BurtonGroup.com*. 12 Dec 2008. http://www.burtongroup.com/Client/Research/Document.aspx?cid=1514 .
  - Noreen Kendle. "The Data Management Organization: Key to Effective Data Management." BurtonGroup.com. 4 May 2009. http://www.burtongroup.com/Client/Research/Document.aspx?cid=1656.
  - Joe Maguire. "Data Modeling: A Necessary And Rewarding Aspect of Data Management." BurtonGroup.com. 17 Mar 2008. http://www.burtongroup.com/Client/Research/Document.aspx?cid=1592.
  - Joe Maguire. "Organized Chaos: A Framework For Classifying MDM Problems and Solutions." BurtonGroup.com.