WASHINGTON STATISTICAL SOCIETY

http://washstat.org/seminars.html#20141103

November 3rd. 2014

Bureau of Labor Statistics Conference Center

On Information Quality (InfoQ) of Official and Establishment Statistics

Ron S. Kenett

ron@kpa-group.com

Joint work with Galit Shmueli and Luciana Dalla Valle

https://sites.google.com/site/datainfoq/presentations

KPA Insights through analytics

UNIVERSITÀ DEGLI STUDI DI TORINO
ALMA UNIVERSITAS TAURINENSIS

Official Statistics and Information Quality

"An issue that can lead to misconception is that many of the concepts used in official statistics often have specific meanings which are based on, but not identical to, their everyday usage meaning.

Official statistics "need to be used to be useful" and utility is one of the overarching concepts in official statistics."

How can official statistics be used to generate information of high quality?

Forbes, S. and Brown, D. (2012) Conceptual thinking in national statistics offices, *Statistical Journal of the IAOS* 28, p 89–98.

Official Statistics and Information Quality

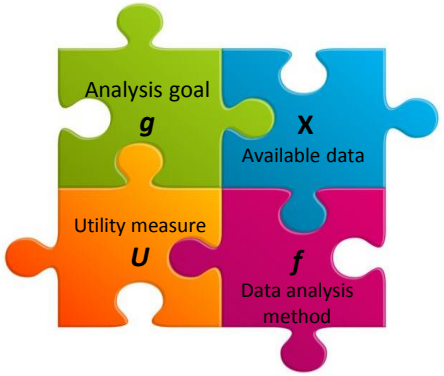
"All staff producing statistics must understand that … their work translate the real world into models that interpret reality and make it measurable for statistical purposes.

The first step … is to define the issue or question(s) that statistical information is needed to inform. That is, to define the objectives for the framework, and then work through those to create its structure and definitions. An important element … is understanding the relationship between the issues and questions to be informed and the definitions themselves."

Forbes, S. and Brown, D. (2012) Conceptual thinking in national statistics offices, *Statistical Journal of the IAOS* 28, p 89–98.

Information Quality

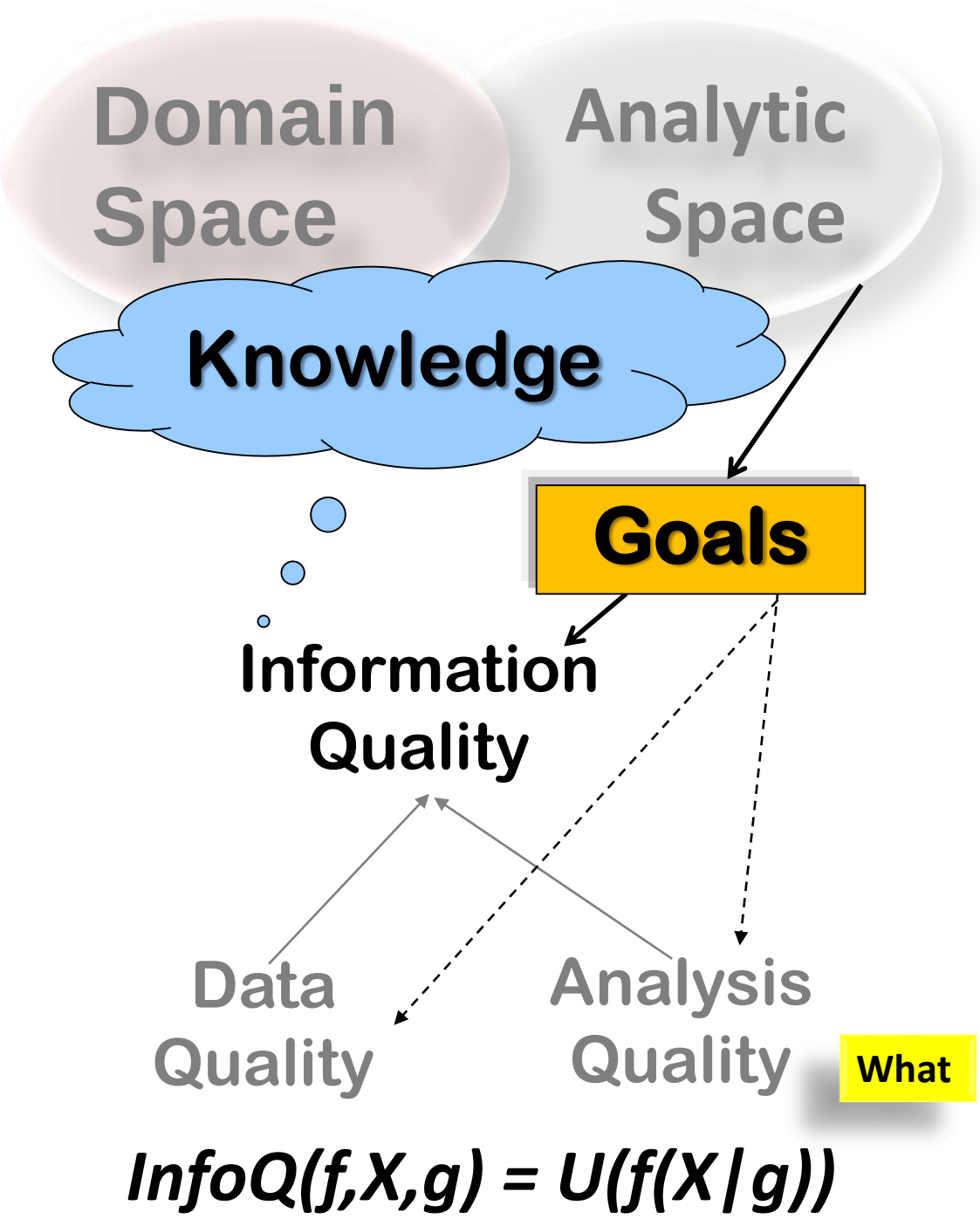
The potential of a particular dataset to achieve a particular goal using a given empirical analysis method



g	**A specific analysis goal**
X	**The available dataset**
f	**An empirical analysis method**
U	**A utility measure**

$$InfoQ(f,X,g) = U(f(X|g))$$

Kenett, R.S. and Shmueli , G. (2014) On Information Quality , *Journal of the Royal Statistical Society, Series A* (with discussion), Vol. 177, No. 1, pp. 3-38. http://ssrn.com/abstract=1464444.

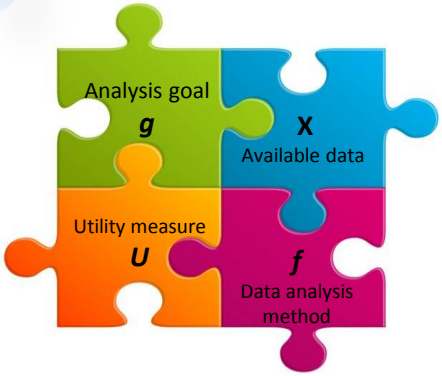


$$InfoQ(f,X,g) = U(f(X|g))$$

InfoQ

1. Data resolution
2. Data structure
3. Data integration
4. Temporal relevance
5. Chronology of data and goal
6. Generalizability
7. Operationalization
8. Communication

How



Assessing Information Quality

Assess components

InfoQ dimensions

1. Data resolution
2. Data structure
3. Data integration
4. Temporal relevance
5. Chronology of data and goal
6. Generalizability
7. Operationalization
8. Communication

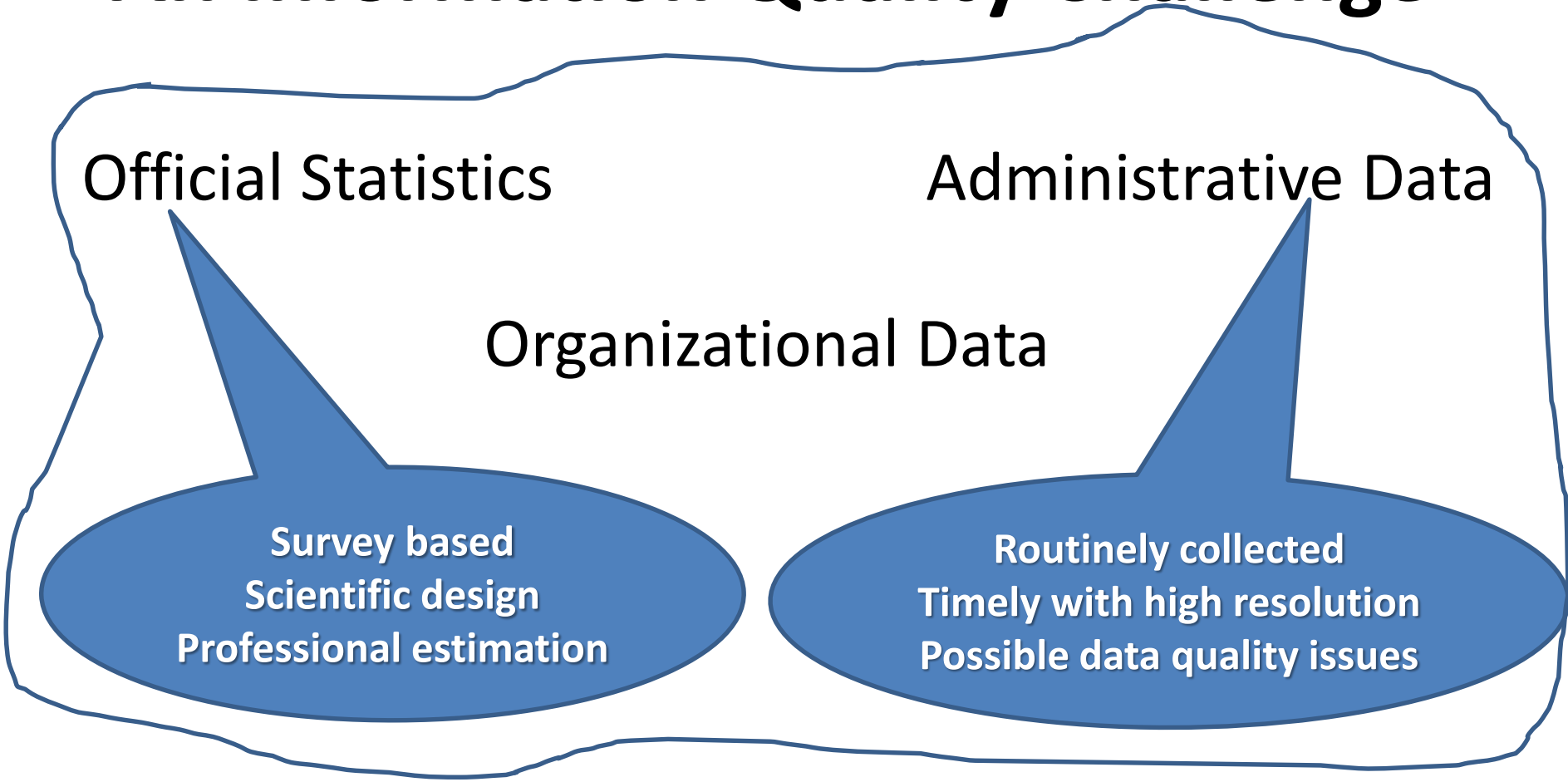
Assess properties

"**Quality of Statistical Data**"
(Eurostat, OECD, NCSES,…)

- Relevance
- Accuracy
- Timeliness and punctuality
- Accessibility
- Interpretability
- Coherence
- Credibility

http://www.nsf.gov/statistics/information-quality.cfm
http://epp.eurostat.ec.europa.eu/portal/page/portal/ver-1/quality/documents/ESQR_FINAL.pdf
http://www.oecd.org/std/qualityframeworkforoecdstatisticalactivities.htm

An Information Quality Challenge



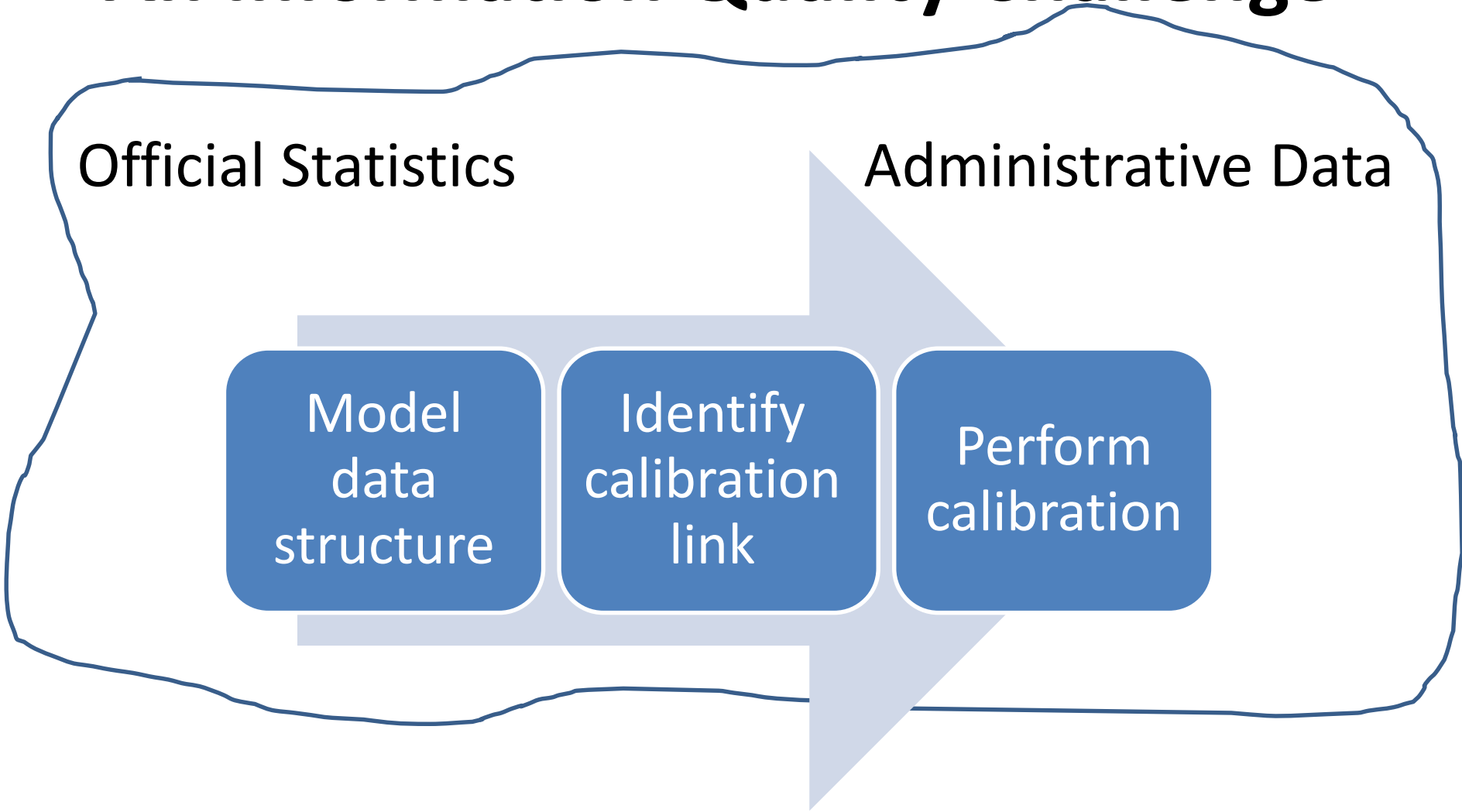
Small area estimation with external benchmarks

Pfefferman, D (2013) New Important Developments in Small Area Estimation, *Statistical Science*, Vol. 28, No. 1, 40–68

An Information Quality Challenge

- Model data structure
  - Conduct multivariate data analysis with graphical models of official statistics and administrative data
- Identify calibration link
  - Identify links with content commonality
- Perform calibration
  - Condition official statistics to reflect strength of calibration links in administrative data

An Information Quality Challenge



This calibration affects all InfoQ dimensions: 1) Data resolution, 2) Data structure, 3) Data integration, 4) Temporal relevance, 5) Chronology of data and goal, 6) Generalizability, 7) Operationalization, 8) Communication

An Italian Case Study

$$InfoQ(f,X,g) = U(f(X|g))$$

f: Use of **Vines** and **Bayesian networks** to model the dependence structure of the variables in the data set and to calculate the conditional rank correlations

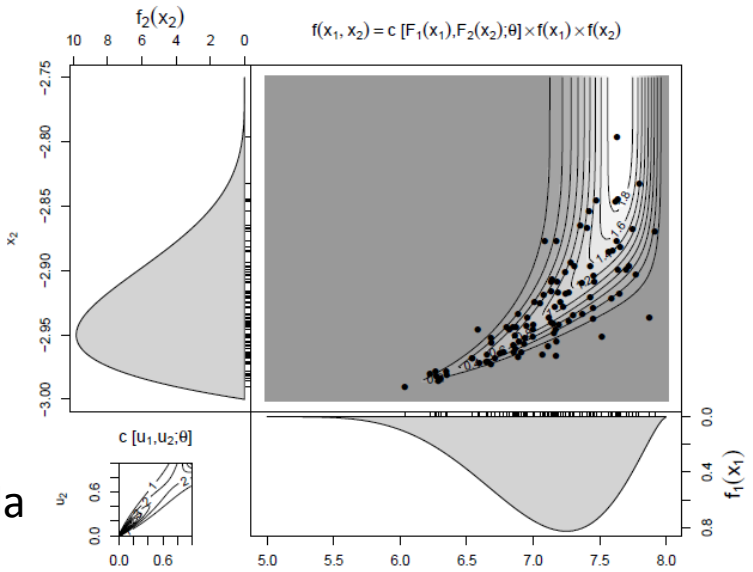
g: Understand the influence on sales of several variables, such as number of employees, to make predictions and derive diagnostics.

X: combined survey data and individual company performance with data reported to the stock exchange.

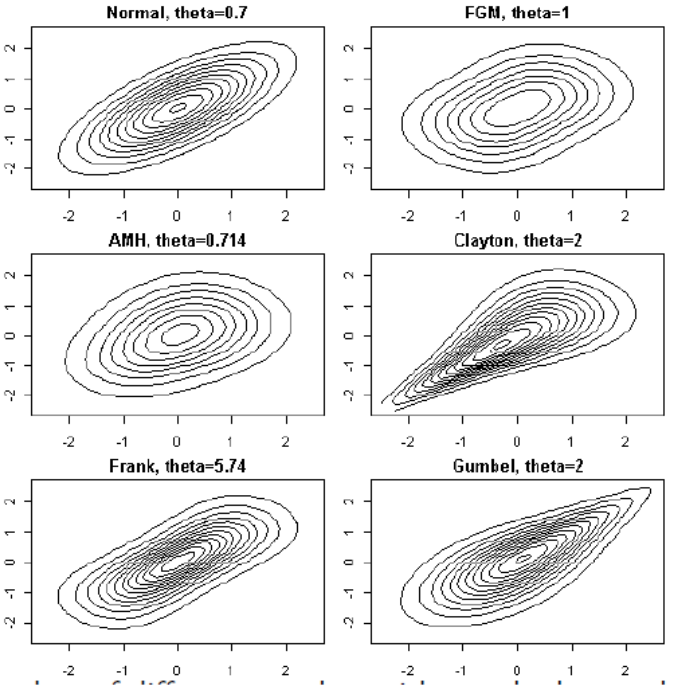
U: Sales prediction error in employment policy economic programs

Kenett, Ron S., Applications of Bayesian Networks (2012). Available at SSRN: http://ssrn.com/abstract=2172713 or http://dx.doi.org/10.2139/ssrn.2172713

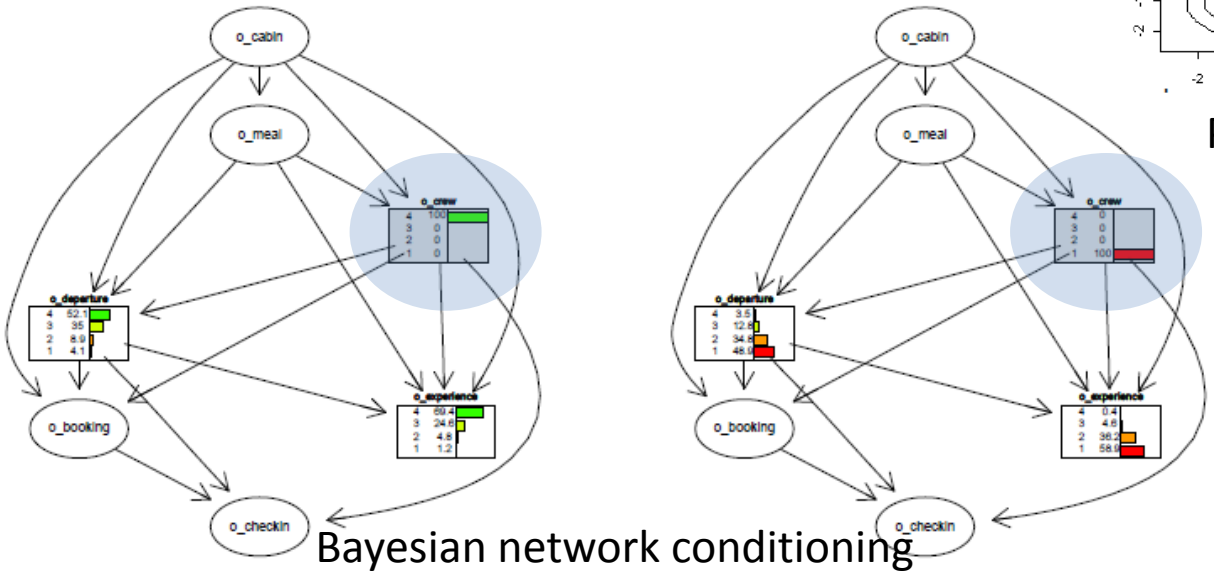
Vines (PCCs) and Bayesian networks



Copula



Paired Copula Constructions



Bayesian network conditioning

Non parametric Bayesian belief nets (NPBBNs): directed graphs, that use pair copulas to model the dependencies, and allow for diagnosis and prediction via conditionalization.(Kurowicka and Cooke, 2006)

I: The Assolombarda Data

- Assolombarda is an Italian association of about 5,000 firms located in the province of Milan and in other provinces of the north of Italy, and represents manufacturing and service companies.
- The associated firms employ about 300,000 workers locally and several hundred thousands in the whole country.
- Assolombarda periodically collects data through questionnaires sent to the associated firms, in order to gather information about the economic climate, firms' activity and production, and the number and types of employees.
- The data analyzed contains information collected through one of the association surveys in 2007, and it is about 167 firms located in the provinces of Milan and Lodi.

Dalla Valle, L. (2014) Official Statistics Data Integration Using Copulas, *Quality Technology & Quantitative Management,* Vol. 11, No. 1, pp. 111-131, http://www.cc.nctu.edu.tw/~qtqm/upcomingpapers/2014V11N1/2014V11N1_F8.pdf

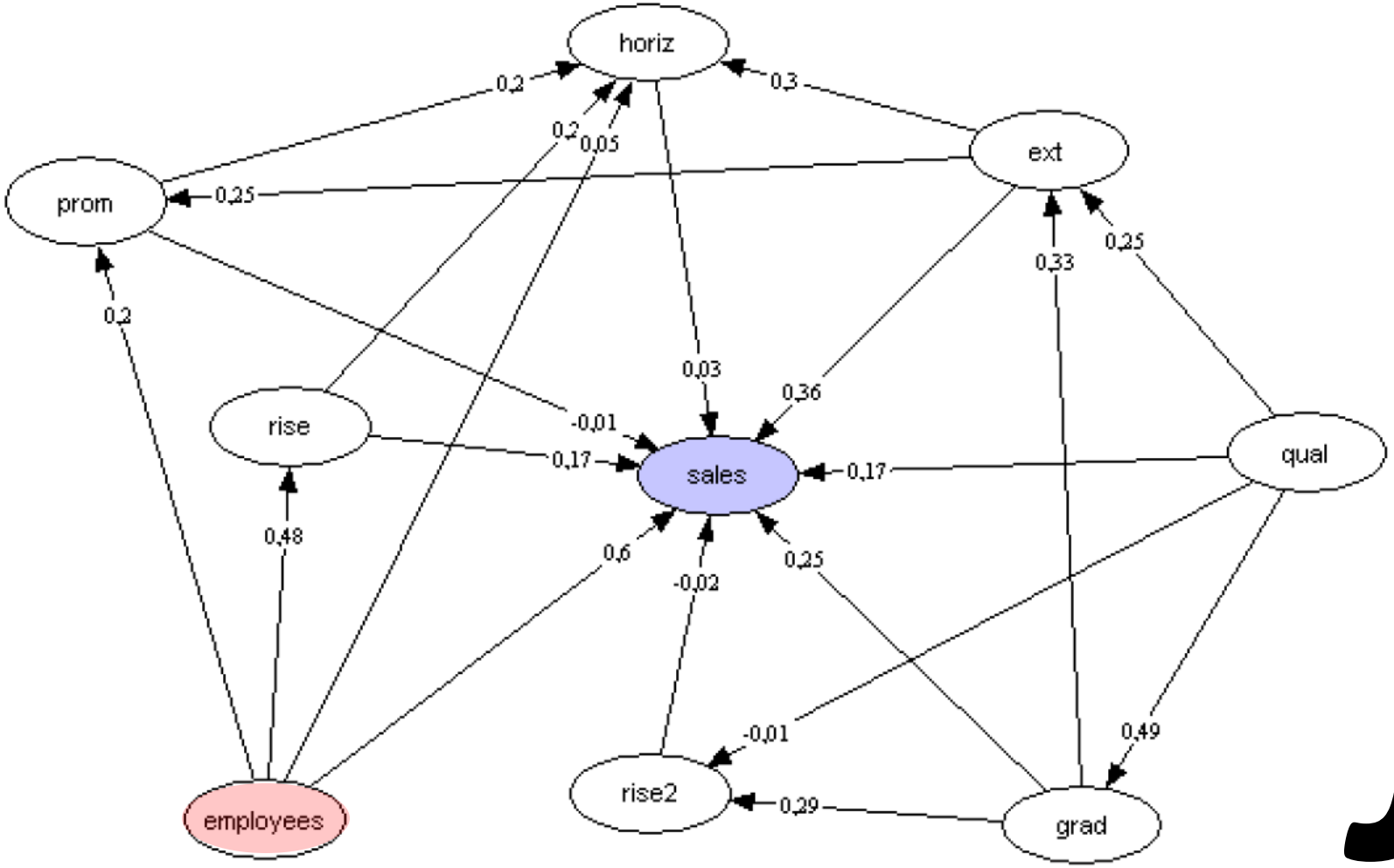
I: The Assolombarda Data

The variables in the dataset are:

- *sales*: firm annual turnover;
- *emp*: average number of employees;
- *rise*: number of managers receiving wage rise;
- *rise2*: number of managers that will receive wage rise in the following year;
- *prom*: number of employees gaining a promotion;
- *horiz* : number of employees involved in horizontal movements;
- *ext*: number of people employed in the external market;
- *grad*: number of newly-graduated employees;
- *qual*: number of newly-qualified employees.

Dalla Valle, L. (2014) Official Statistics Data Integration Using Copulas, *Quality Technology & Quantitative Management,* Vol. 11, No. 1, pp. 111-131, http://www.cc.nctu.edu.tw/~qtqm/upcomingpapers/2014V11N1/2014V11N1_F8.pdf

I: The Assolombarda Data

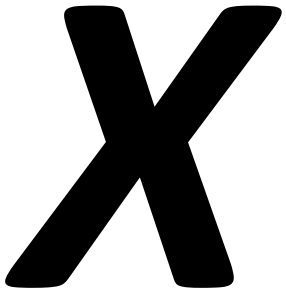


Dalla Valle, L. (2014) Official Statistics Data Integration Using Copulas, *Quality Technology & Quantitative Management,* Vol. 11, No. 1, pp. 111-131, http://www.cc.nctu.edu.tw/~qtqm/upcomingpapers/2014V11N1/2014V11N1_F8.pdf

II: The FTSE-MTB Data

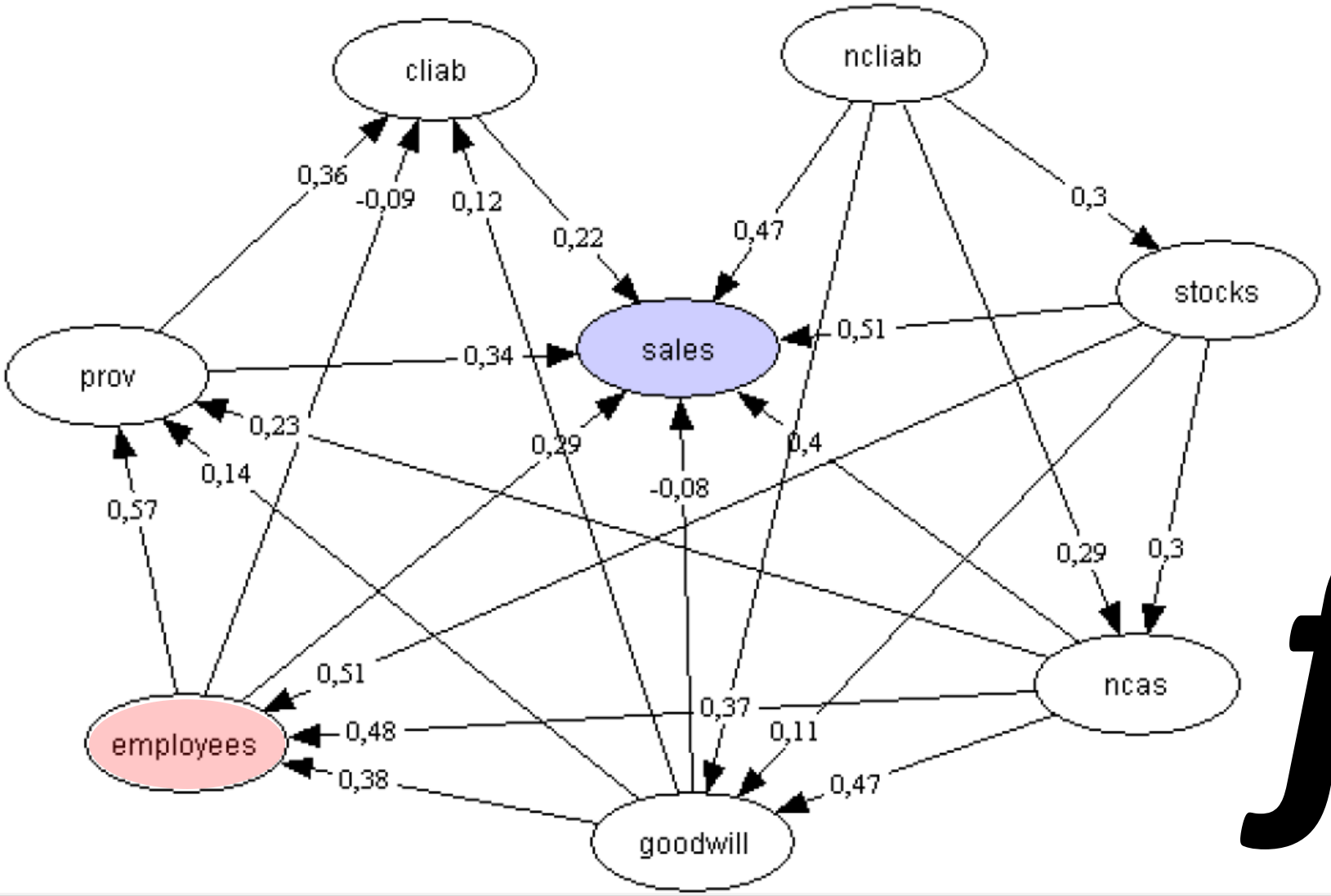
The FTSE-MIB is the benchmark stock market index for the Italian national stock exchange and consists of the 40 most-traded stock classes on the exchange. The dataset analyzed here contains information from the balance sheets of the 40 largest Italian firms belonging to the Italian stock market. The variables used in the analysis are:

- *sales*: firm annual turnover;
- *emp*: average number of employees;
- *goodwill*: difference between the balance sheet assets and the sum of intangible assets and equipment at market value;
- *ncas*: non-current financial assets;
- *stocks*: stocks and work in progress;
- *prov*: provisions for liabilities and non-recurring expenses;
- *ncliab*: non-current liabilities;
- *cliab*: current liabilities.



Dalla Valle, L. (2014) Official Statistics Data Integration Using Copulas, *Quality Technology & Quantitative Management,* Vol. 11, No. 1, pp. 111-131, http://www.cc.nctu.edu.tw/~qtqm/upcomingpapers/2014V11N1/2014V11N1_F8.pdf

II: The FTSE-MTB Data



Dalla Valle, L. (2014) Official Statistics Data Integration Using Copulas, *Quality Technology & Quantitative Management,* Vol. 11, No. 1, pp. 111-131, http://www.cc.nctu.edu.tw/~qtqm/upcomingpapers/2014V11N1/2014V11N1_F8.pdf

#1 Data Resolution

Data collected at the company level.

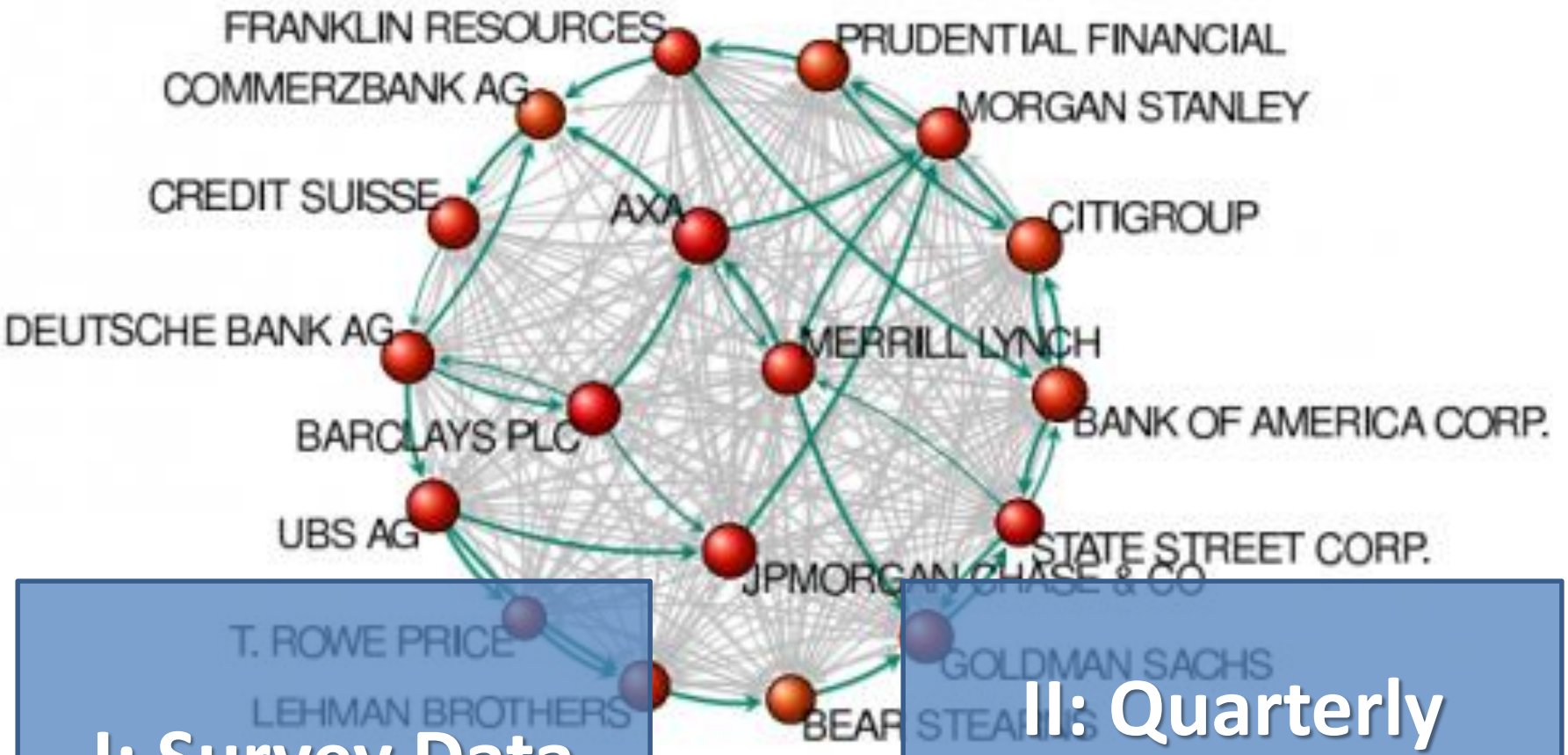
I: Periodic survey waves of self reports

II: Quarterly stock exchange reports

Goal: Predict sales using # employees in the context of a regional development plan

Dalla Valle, L. (2014) Official Statistics Data Integration Using Copulas, *Quality Technology & Quantitative Management,* Vol. 11, No. 1, pp. 111-131, http://www.cc.nctu.edu.tw/~qtqm/upcomingpapers/2014V11N1/2014V11N1_F8.pdf

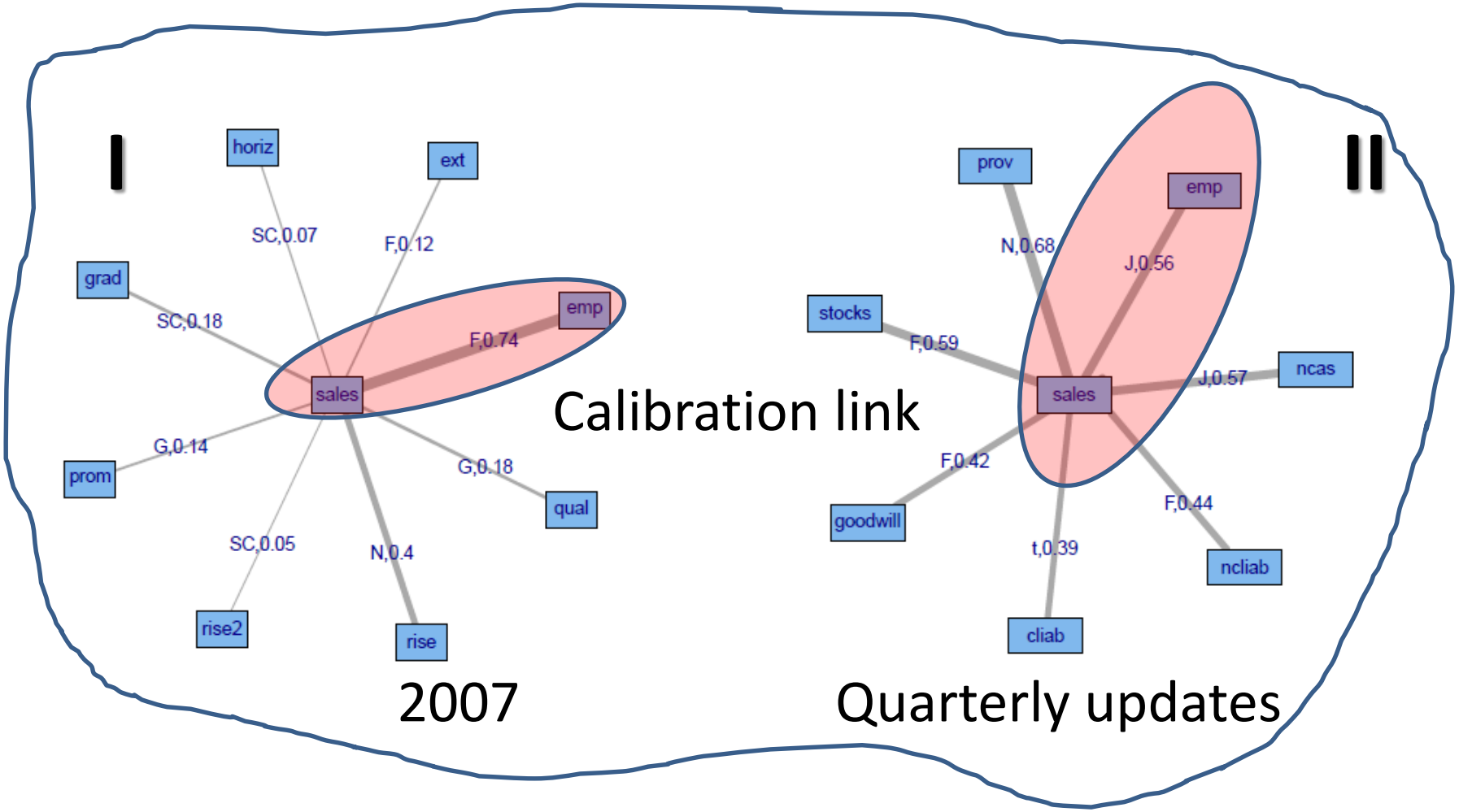
#2 Data Structure



I: Survey Data

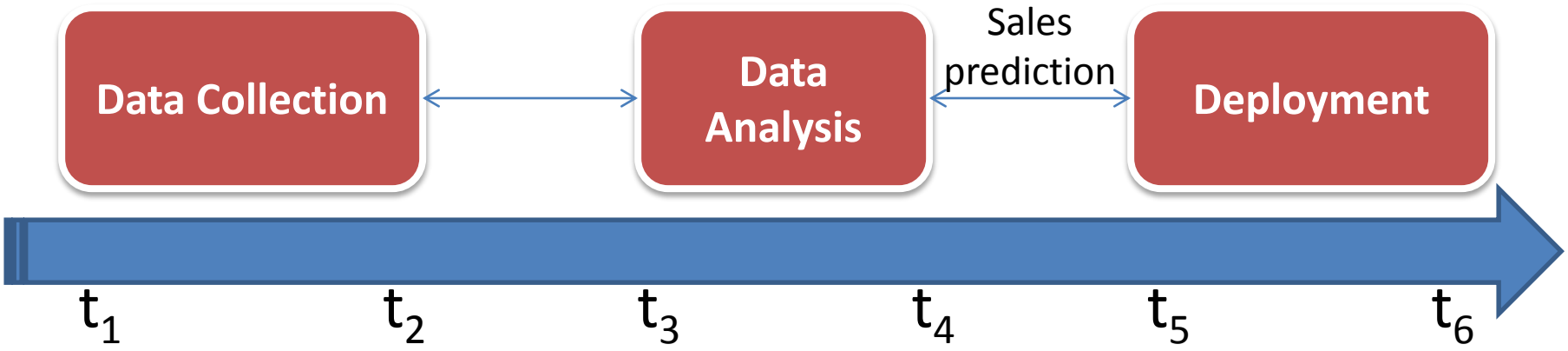
II: Quarterly Reports

#3 Data Integration



Dalla Valle, L. (2014) Official Statistics Data Integration Using Copulas, *Quality Technology & Quantitative Management,* Vol. 11, No. 1, pp. 111-131, http://www.cc.nctu.edu.tw/~qtqm/upcomingpapers/2014V11N1/2014V11N1_F8.pdf

#4 Temporal Relevance



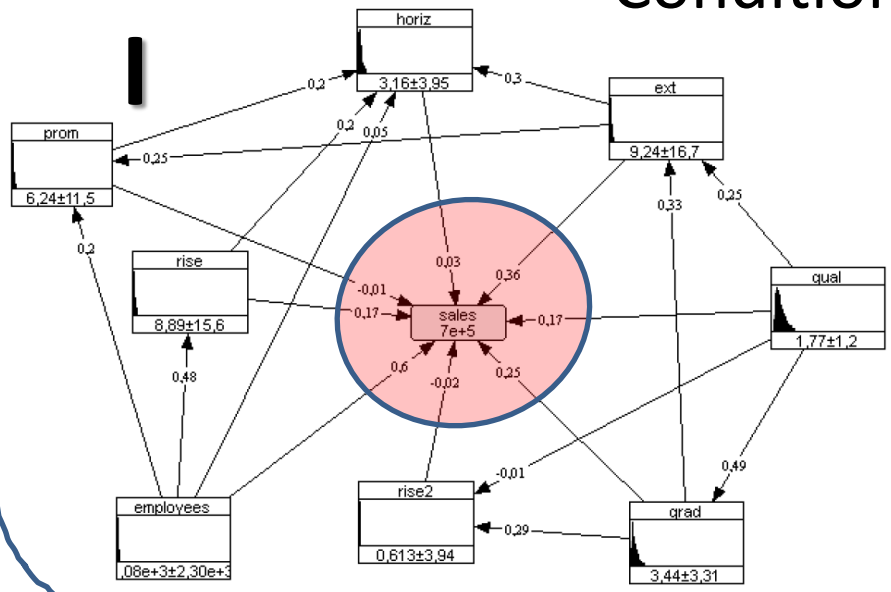
Quarterly updates

2007 2014 2015?

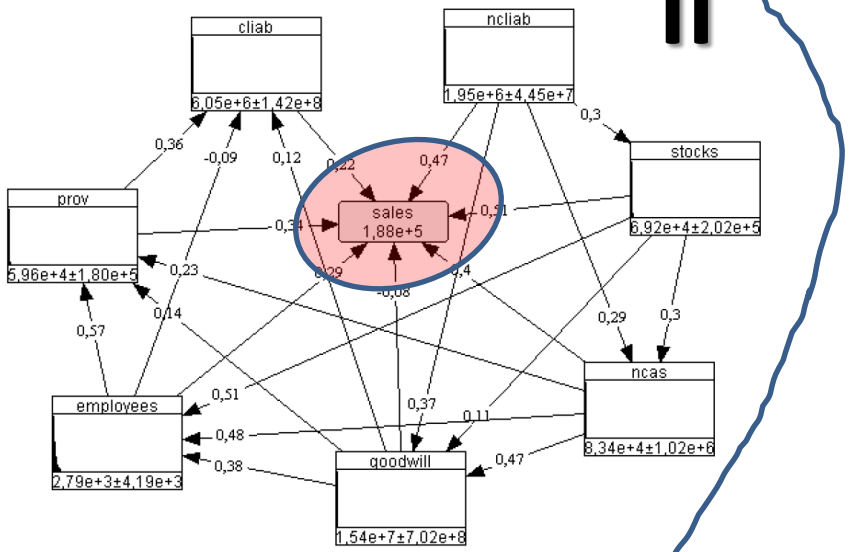
#5 Chronology of Data & Goal

For a specific locality
(Small Area Estimation)

Conditioning of Output



2007

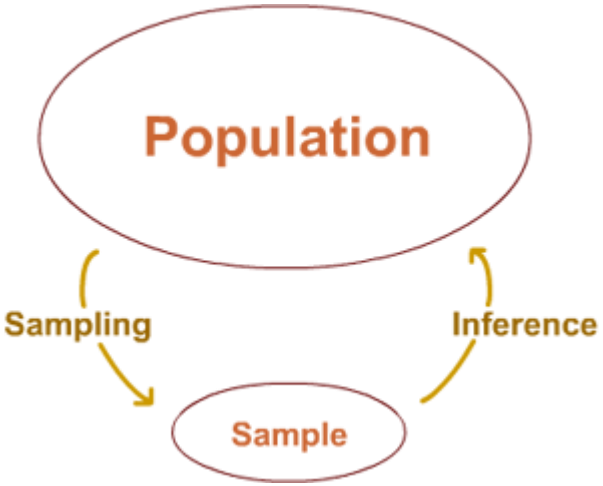


Quarterly updates

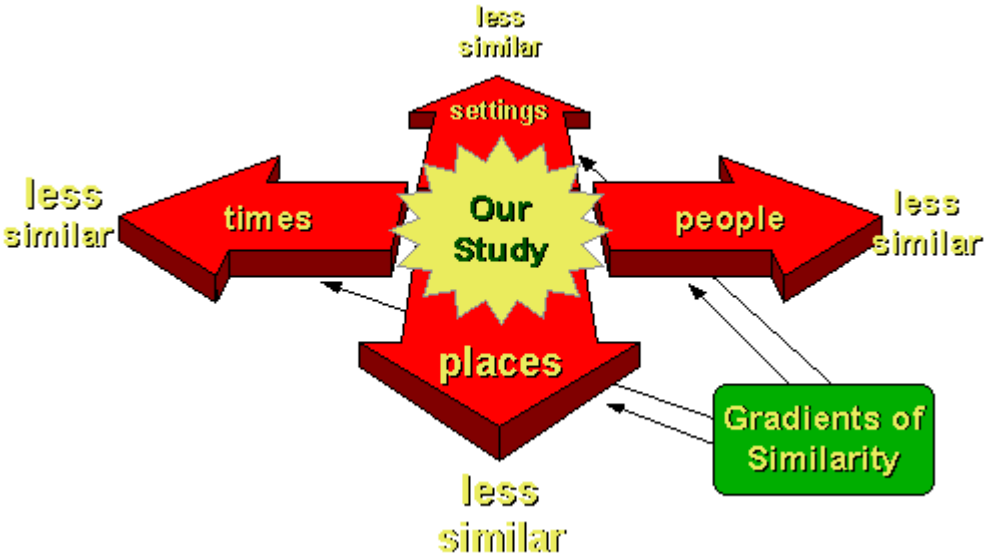
Dalla Valle, L. (2014) Official Statistics Data Integration Using Copulas, *Quality Technology & Quantitative Management,* Vol. 11, No. 1, pp. 111-131, http://www.cc.nctu.edu.tw/~qtqm/upcomingpapers/2014V11N1/2014V11N1_F8.pdf

#6 Generalizability

Statistical generalizability

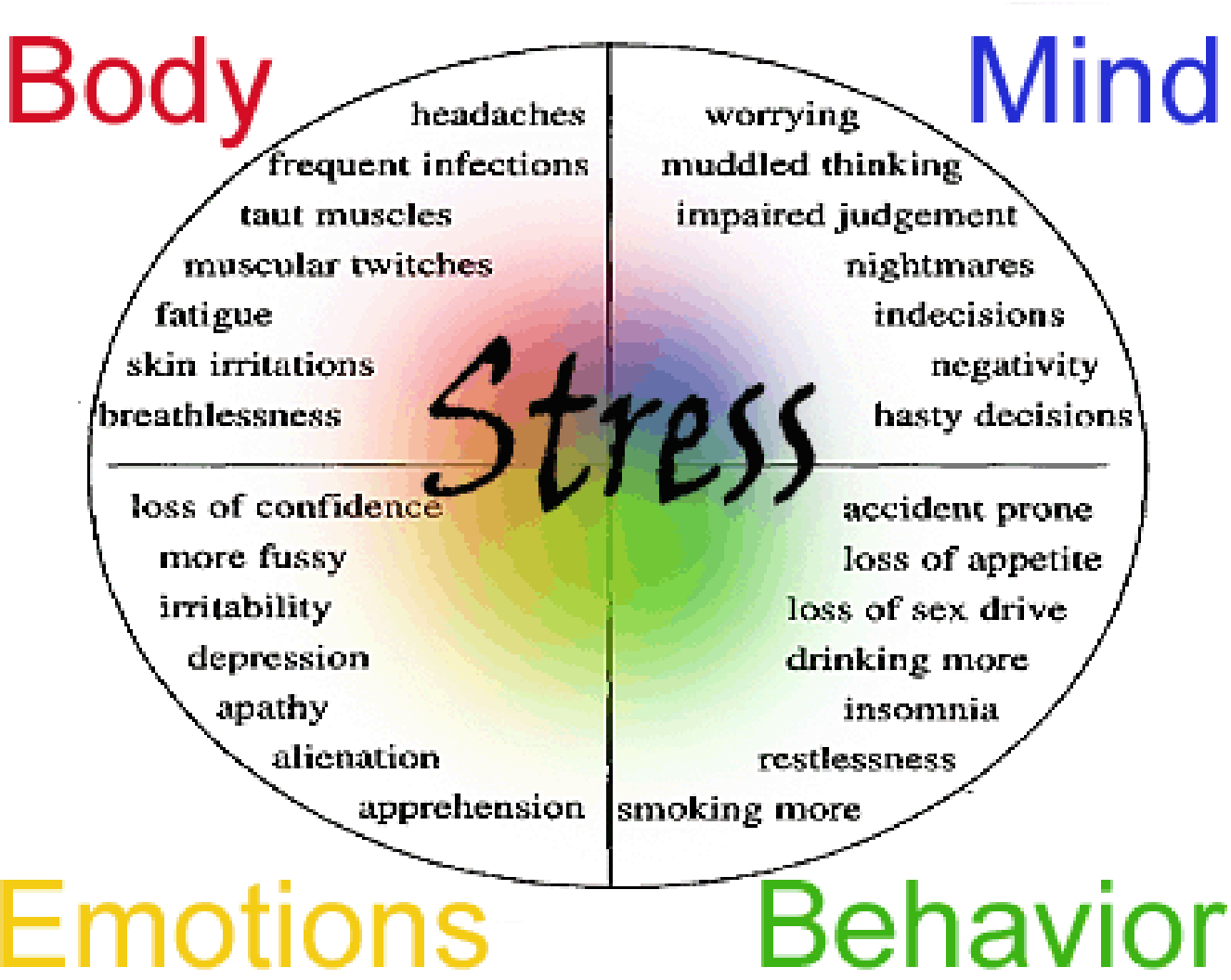


Scientific generalizability



Lombardi? The North of Italy? Italy?

#7 (Construct) Operationalization



#7 (Action) Operationalization

In the pre-publication drafts of *Quality, Productivity, and Competitive Position* Dr. Deming wrote:

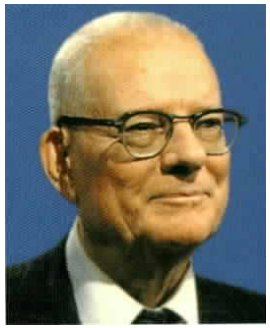
> "An operational definition consists of (1) a criterion to be applied to an object or a group of objects, (2) a test of compliance for the object or group, and (3) a decision rule for interpreting the test results as to whether the object or group is, or is not, in compliance."

In Dr. Deming's own conversations, when individuals would start telling him about what they or their organization were planning to do, he would invariably have one of two responses for them: "By what method?" or "How will you know?" Either one of these questions would generally end the conversation since the individual would have no answer. After discerning this pattern to Dr. Deming's responses, it finally occurred to me that these two questions corresponded to the last two parts of an operational definition. This realization, in turn, resulted in a generalization of an operational definition to become:

(1) What do you want to accomplish?

(2) By what method will you accomplish it?

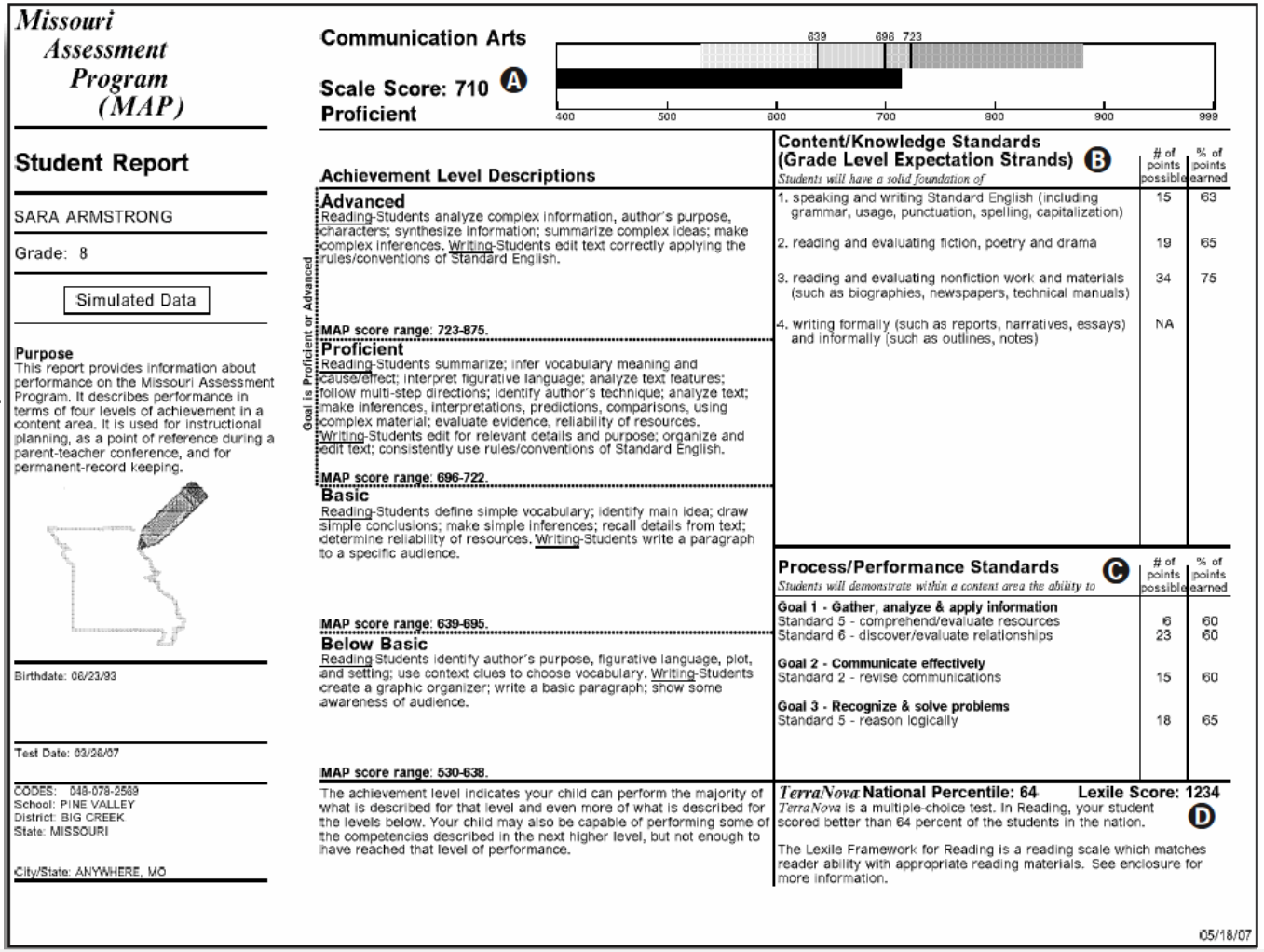
(3) How will you know when you have accomplished it?



#7 (Action) Operationalization

National Education Goals Panel (NEGP) recommended that states answer four questions on their student reports:

1. How did my child do?
2. What types of skills or knowledge does his or her performance reflect?
3. How did my child perform in comparison to other students in the school,
district, state, and, if available, the nation?
4. What can I do to help my child improve?



Missouri Assessment Program (MAP)

Student Report

SARA ARMSTRONG

Grade: 8

Simulated Data

Purpose
This report provides information about performance on the Missouri Assessment Program. It describes performance in terms of four levels of achievement in a content area. It is used for instructional planning, as a point of reference during a parent-teacher conference, and for permanent-record keeping.

Birthdate: 08/23/93

Test Date: 03/26/07

CODES: 048-078-2589
School: PINE VALLEY
District: BIG CREEK
State: MISSOURI

City/State: ANYWHERE, MO

Communication Arts

Scale Score: 710 A
Proficient

639 696 723
400 500 600 700 800 900 999

Goal is Proficient or Advanced

Achievement Level Descriptions

Advanced
Reading-Students analyze complex information, author's purpose, characters; synthesize information; summarize complex ideas; make complex inferences. Writing-Students edit text correctly applying the rules/conventions of Standard English.

MAP score range: 723-875.

Proficient
Reading-Students summarize; infer vocabulary meaning and cause/effect; interpret figurative language; analyze text features; follow multi-step directions; identify author's technique; analyze text; make inferences, interpretations, predictions, comparisons, using complex material; evaluate evidence, reliability of resources. Writing-Students edit for relevant details and purpose; organize and edit text; consistently use rules/conventions of Standard English.

MAP score range: 696-722.

Basic
Reading-Students define simple vocabulary; identify main idea; draw simple conclusions; make simple inferences; recall details from text; determine reliability of resources. Writing-Students write a paragraph to a specific audience.

MAP score range: 639-695.

Below Basic
Reading-Students identify author's purpose, figurative language, plot, and setting; use context clues to choose vocabulary. Writing-Students create a graphic organizer; write a basic paragraph; show some awareness of audience.

MAP score range: 530-638.

The achievement level indicates your child can perform the majority of what is described for that level and even more of what is described for the levels below. Your child may also be capable of performing some of the competencies described in the next higher level, but not enough to have reached that level of performance.

Content/Knowledge Standards (Grade Level Expectation Strands) B
Students will have a solid foundation of

	# of points possible	% of points earned
1. speaking and writing Standard English (including grammar, usage, punctuation, spelling, capitalization)	15	63
2. reading and evaluating fiction, poetry and drama	19	65
3. reading and evaluating nonfiction work and materials (such as biographies, newspapers, technical manuals)	34	75
4. writing formally (such as reports, narratives, essays) and informally (such as outlines, notes)	NA	

Process/Performance Standards C
Students will demonstrate within a content area the ability to

	# of points possible	% of points earned
Goal 1 - Gather, analyze & apply information		
Standard 5 - comprehend/evaluate resources	6	60
Standard 6 - discover/evaluate relationships	23	60
Goal 2 - Communicate effectively		
Standard 2 - revise communications	15	60
Goal 3 - Recognize & solve problems		
Standard 5 - reason logically	18	65

TerraNova **National Percentile: 64** **Lexile Score: 1234** D

TerraNova is a multiple-choice test. In Reading, your student scored better than 64 percent of the students in the nation.

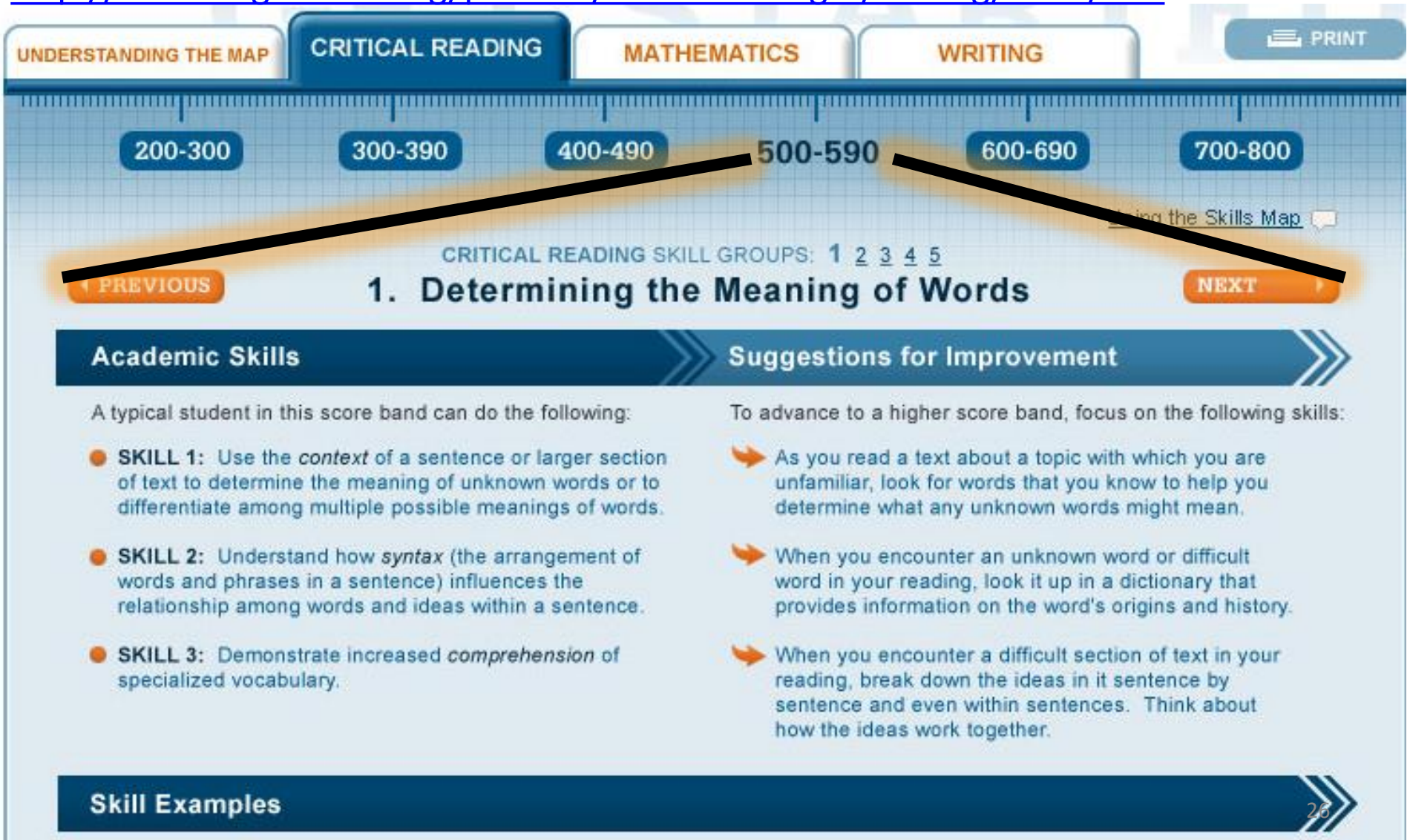
The Lexile Framework for Reading is a reading scale which matches reader ability with appropriate reading materials. See enclosure for more information.

05/18/07

Goodman , D. and. Hambleton , R.(2004). Student Test Score Reports and Interpretive Guides: Review of Current Practices and Suggestions for Future Research, *Applied Measurement in Education*, 17:2, 145-220

#7 (Action) Operationalization

http://sat.collegeboard.org/practice/sat-skills-insight/writing/band/200



#7 (Action) Operationalization

TABLE 1 National Overall Average Mathematics Proficiency and Achievement Levels, Grades 4, 8, and 12

16

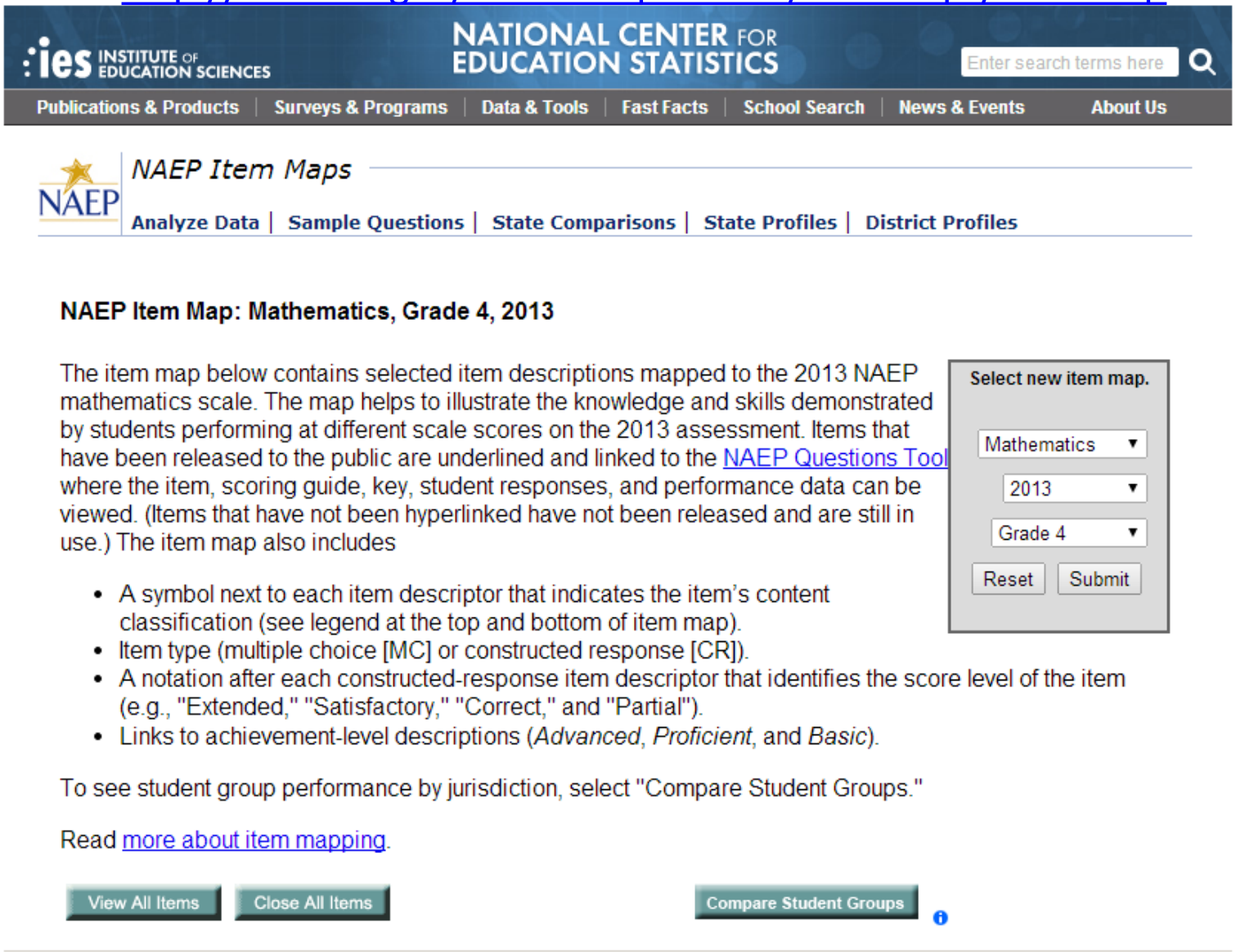
			Percentage of Students At or Above			
Grades	Assessment Years	Average Proficiency	Advanced	Proficient	Basic	Percentage Below Basic
4	1992	218(0.7)>	2(0.3)	18(1.0)>	61(1.0)>	39(1.0)<
	1990	213(0.9)	1(0.4)	13(1.1)	54(1.4)	46(1.4)
8	1992	268(0.9)>	4(0.4)	25(1.0)>	63(1.1)>	37(1.1)<
	1990	263(1.3)	2(0.4)	20(1.1)	58(1.4)	42(1.4)
12	1992	299(0.9)>	2(0.3)	16(0.9)	64(1.2)>	36(1.2)<
	1990	294(1.1)	2(0.3)	13(1.0)	59(1.5)	41(1.5)

When asked what the 18% in line 1 meant, 53% of the policy makers responded incorrectly Implying low InfoQ of the report

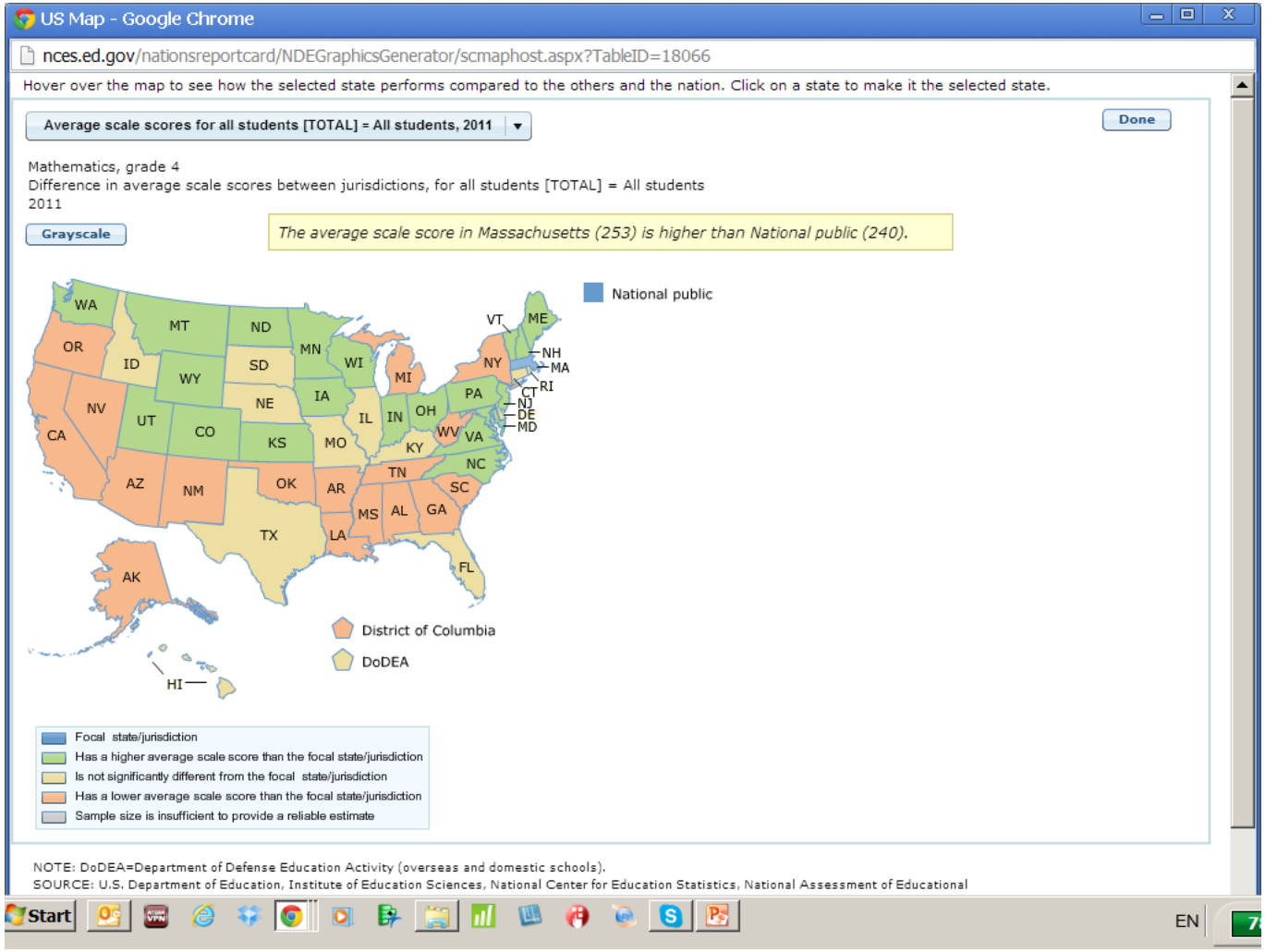
NAEP Executive Summary Report

#8 Communication

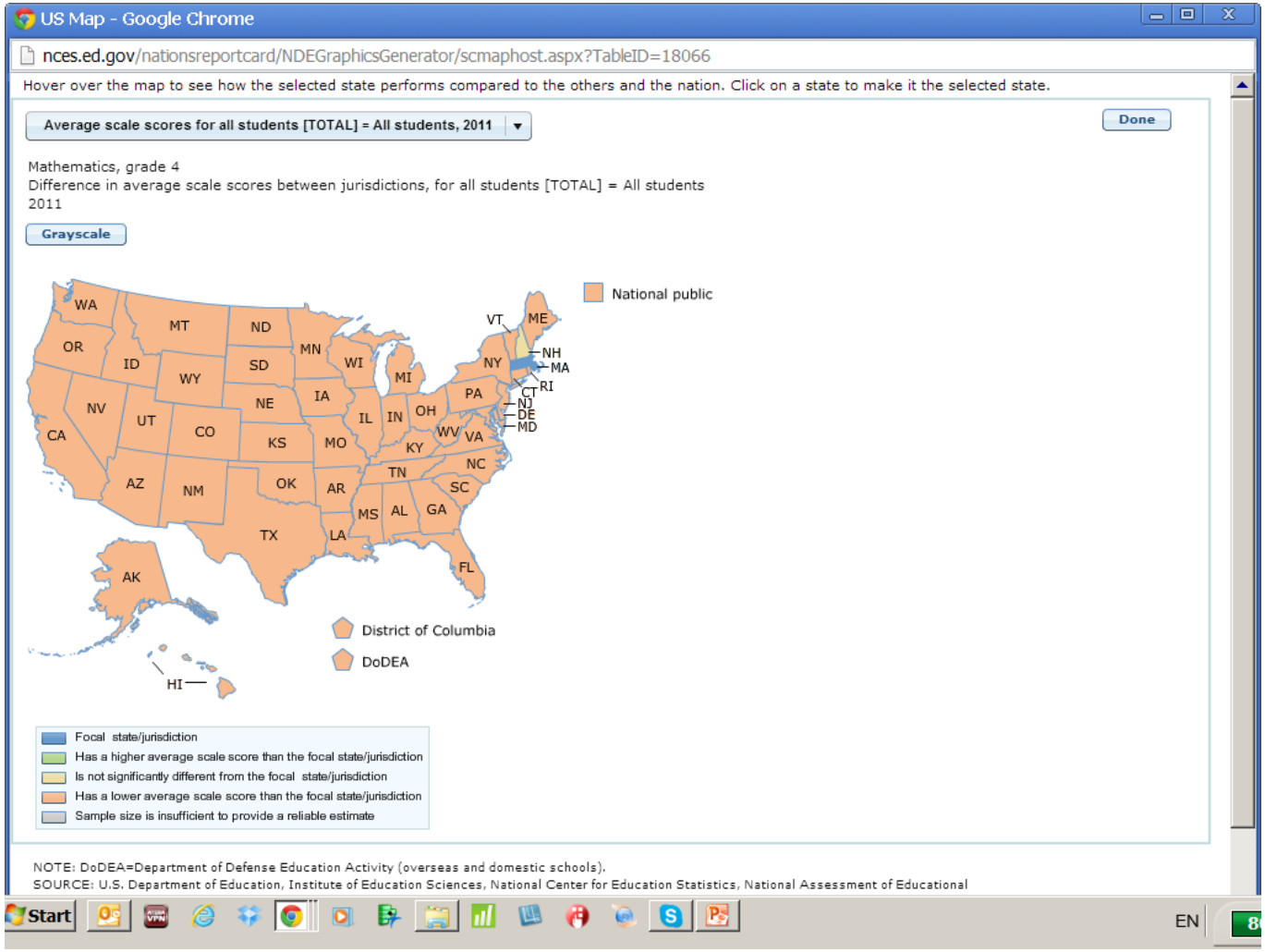
http://nces.ed.gov/nationsreportcard/itemmaps/index.asp



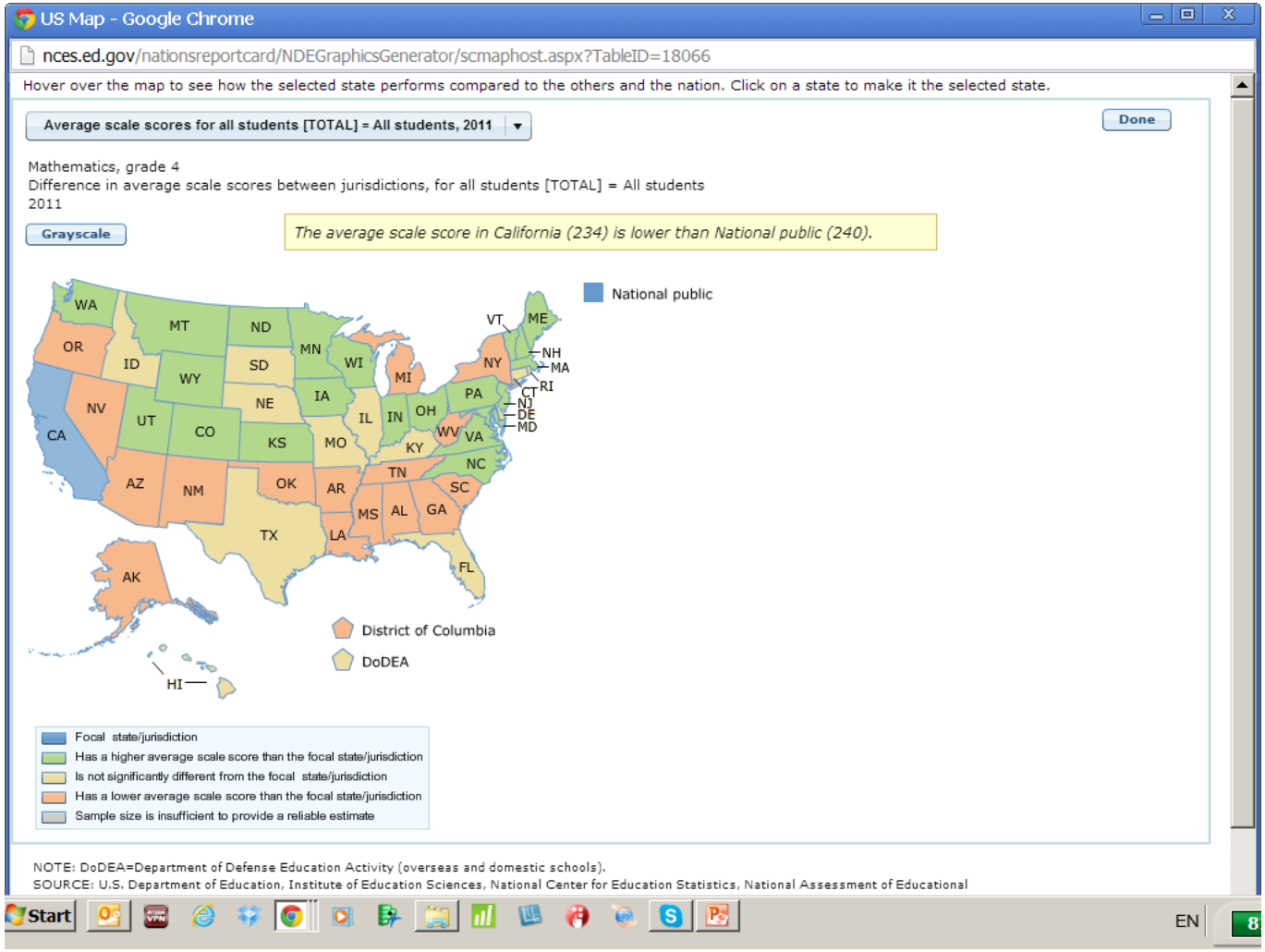
#8 Communication



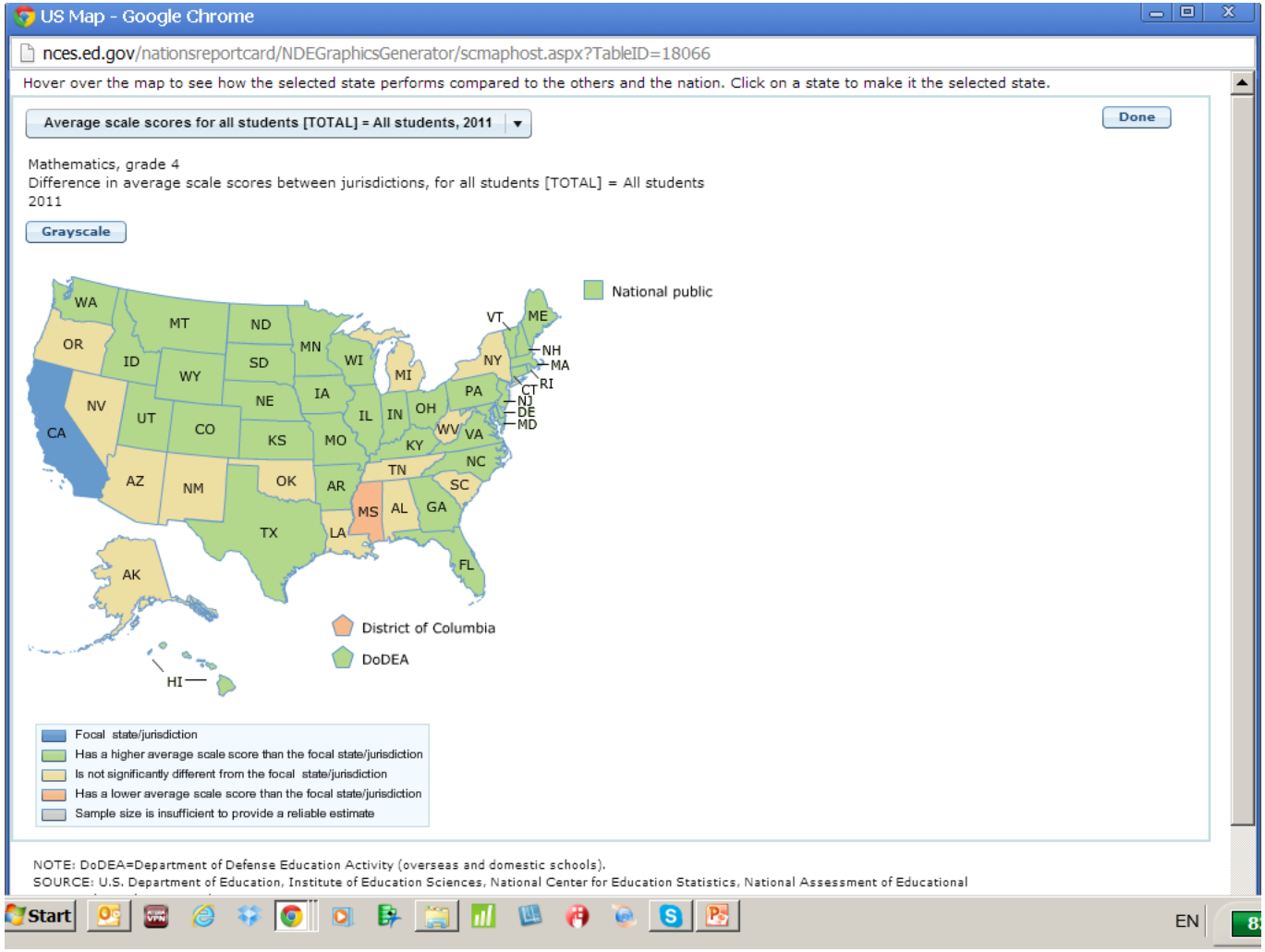
#8 Communication



#8 Communication



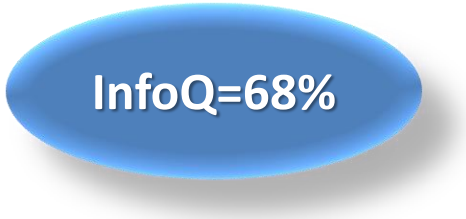
#8 Communication



Assessing InfoQ in Practice

Rating-based assessment

1-5 scale on each dimension:

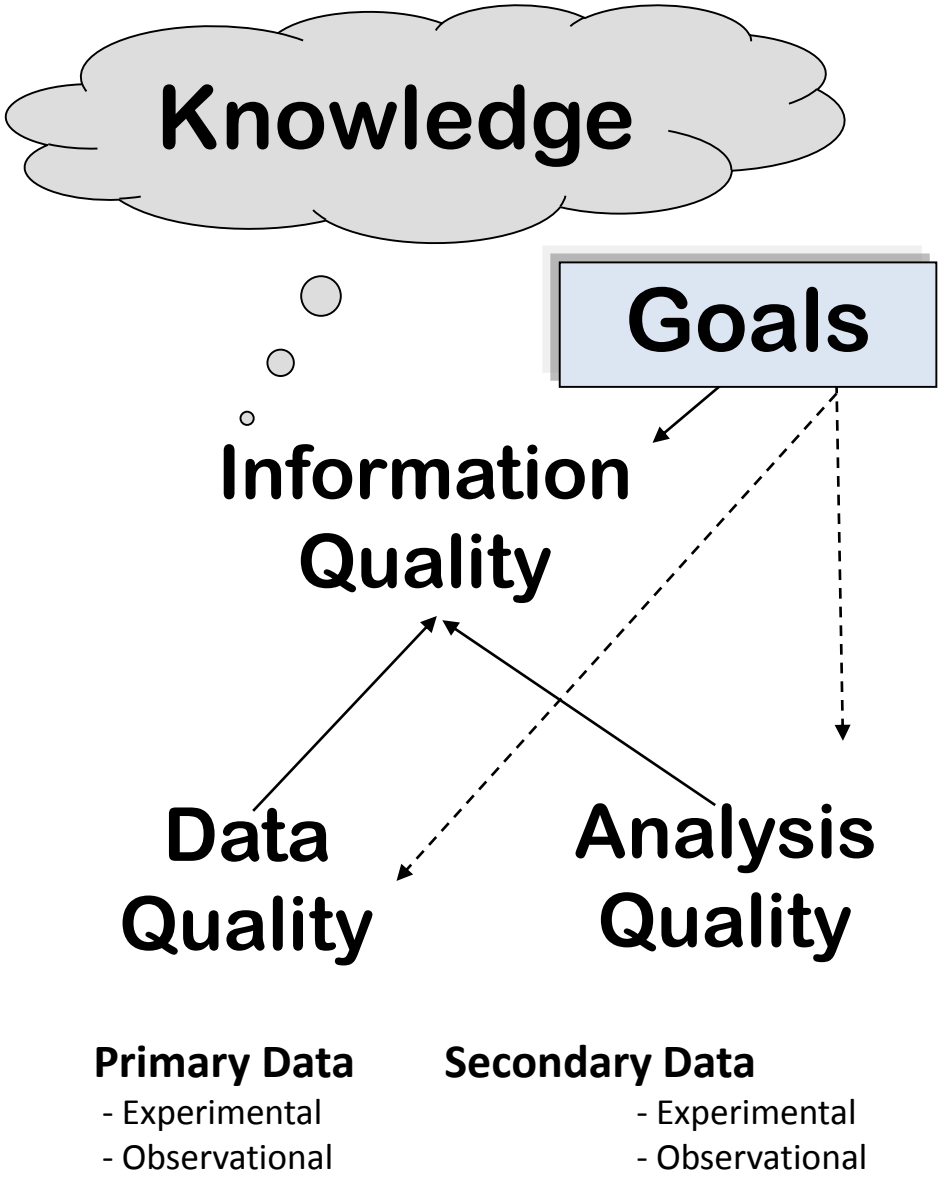


#	Dimension	Note	Value	Index
1	Data resolution		5	1.0000
2	Data structure		4	0.7500
3	Data integration		5	1.0000
4	Temporal relevance		5	1.0000
5	Generalizability		3	0.5000
6	Chronology of data and goal		5	1.0000
7	Concept operationalization		2	0.2500
8	Communication		3	0.5000
	InfoQ Score =	0.68		

$$InfoQ\ Score = [d_1(Y_1)\ \ d_2(Y_2)\ \ \ldots\ \ d_8(Y_8)]^{1/8}$$

Experience from two research methods courses

- Preparing a PhD research proposal (U Ljubljana, 50 students, goo.gl/f6bIA)
- Post-hoc evaluation of five completed studies (CMU, 16 students, goo.gl/erNPF)



Information Quality (InfoQ)

$$InfoQ(f,X,g) = U(f(X|g))$$

g	A specific analysis goal	
X	The available dataset	
f	An empirical analysis method	
U	A utility measure	**What**

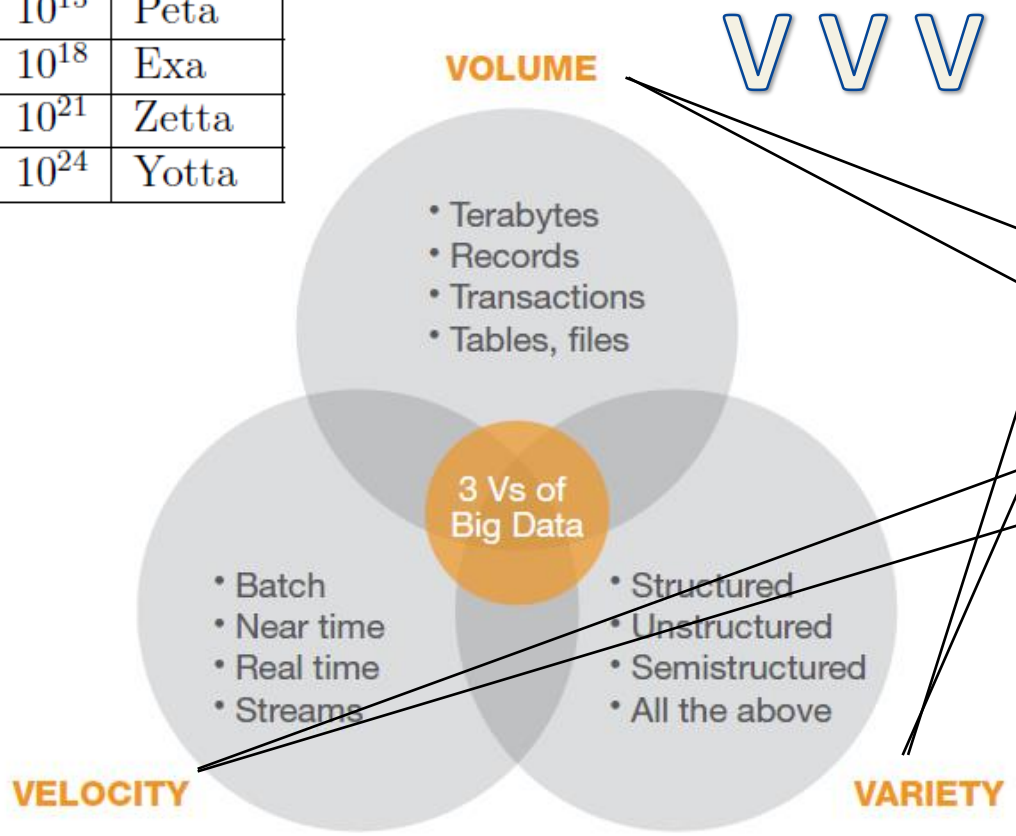
How

1. Data resolution
2. Data structure
3. Data integration
4. Temporal relevance
5. Chronology of data and goal
6. Generalizability
7. Operationalization
8. Communication

Big Data

Power	Prefix
10^{9}	Giga
10^{12}	Tera
10^{15}	Peta
10^{18}	Exa
10^{21}	Zetta
10^{24}	Yotta

Big Data Analytics



V V V

InfoQ

1. Data resolution
2. Data structure
3. Data integration
4. Temporal relevance
5. Chronology of data and goal
6. Generalizability
7. Operationalization
8. Communication

Russom, P., Big Data Analytics, TDWI Best Practices Report, Q4 2011

Thank you for your attention

- Hambleton, R.K. (2002). How can we make NAEP and state test score reporting scales and reports more understandable? In R. W. Lissitz & W. D. Schafer (Eds.), *Assessment in educational reform* (pp. 192–205). Boston, MA: Allyn & Bacon.
- Kenett, R.S. (2008), From Data to Information to Knowledge, *Six Sigma Forum Magazine*, pp. 32-33.
- Kenett, R.S. and Salini, S. (2012). *Modern Analysis of Customer Satisfaction Surveys: with applications using R*, John Wiley and Sons.
- Kenett, R.S. and Shmueli, G. (2014). On Information Quality, *Journal of the Royal Statistical Society, Series A* (with discussion), Vol. 177, No. 1, pp. 3-38. http://ssrn.com/abstract=1464444.
- Dalla Valle, L. (2014) Official Statistics Data Integration Using Copulas, *Quality Technology and Quantitative Management*, Vol. 11, No. 1, pp. 111-131.
- Kenett, R.S. and Shmueli, G. (2014). From Quality to Information Quality in Official Statistics, *Journal of Official Statistics, submitted*