Original Article

On information quality

Ron S. Kenett[1,2,3], Galit Shmueli[4,*]

Article first published online: 7 FEB 2013

DOI: 10.1111/rssa.12007

RSS Journal Club, June 13, 2013

On Information Quality

Galit Shmueli

SRITNE Chaired Professor of Data Analytics

Indian School of Business

galitshmueli.com

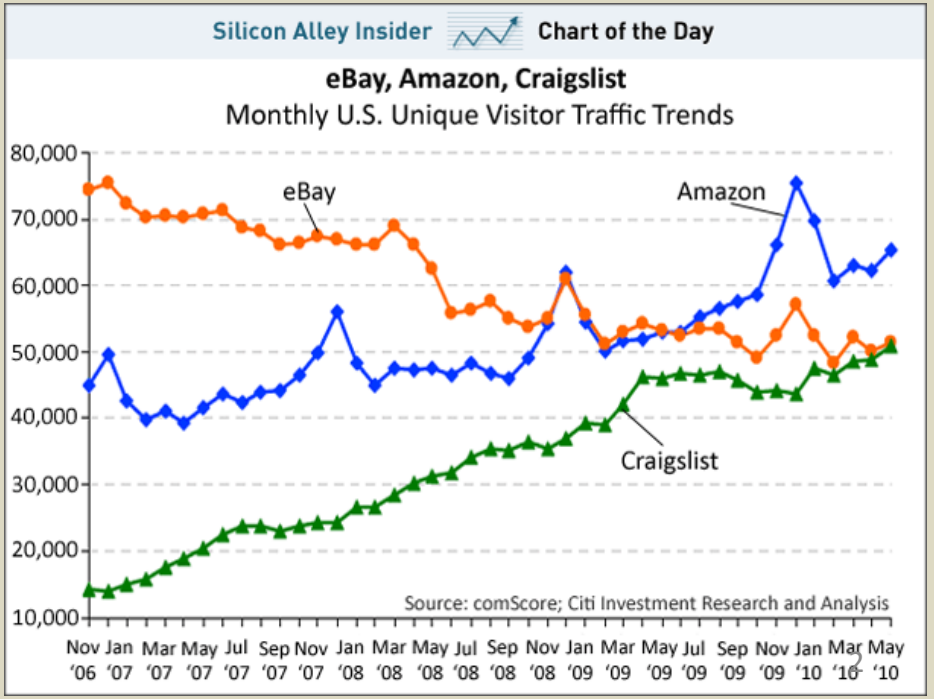
ISB

eBay dataset with millions of auctions

- All camera auctions in 2008
- All camera auctions in 2015
- Auctions for all items in 1/2012
- Sample from all items in 2012

How much will you pay?

Or maybe Craigslist data?

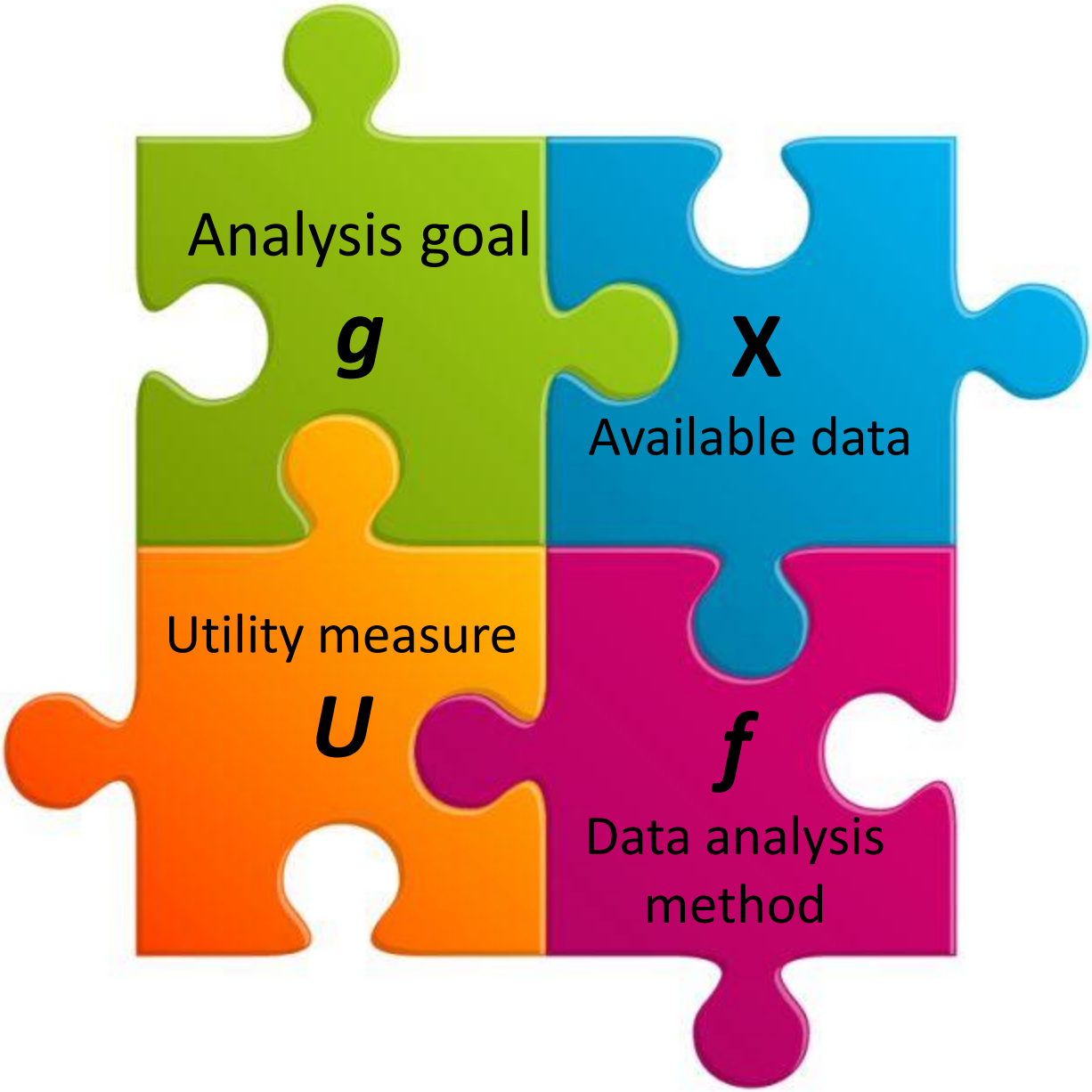


What is the potential of a dataset to generate knowledge?

"statisticians working in a research environment… may well have to explain that the data are **inadequate** to answer a particular question."

Statistics: A Very Short Introduction (Hand 2008)

Pre-data, post-data, post-analysis


Analysis goal
g
X
Available data
Utility measure
U
f
Data analysis method

Domain Goal

What, why, when, where, how

→ Analysis Goal

Explain, predict, describe

Enumerative, analytic

Exploratory, confirmatory

Quality of Goal Specification

- "error of the third kind" - giving the right answer to the wrong question – Kimball
- "Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise" - Tukey

Data Source

- Primary, secondary
- Observational, experiment
- Single, multiple sources
- Collection instrument, protocol

Data Size and Dimension

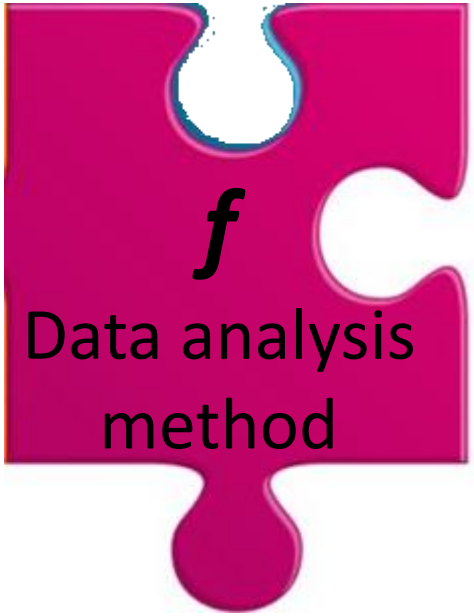
- # observations
- # variables

Data Type

- Continuous, categorical, mix
- Structured, un-, semi-structured
- Cross-sectional, time series, panel, network, geographical

Data Quality *U(X|g)*

- "Zeroth Problem - How do the data relate to the problem, and what other data might be relevant?" - Mallows
- MIS/Database - usefulness of queried data to person querying it.
- *Quality of Statistical Data* (IMF, OECD) - usefulness of summary statistics for a particular goal (7 dimensions)



Statistical models and methods

- Parametric, semi-, non-parametric
- Classic, Bayesian

Data mining algorithms

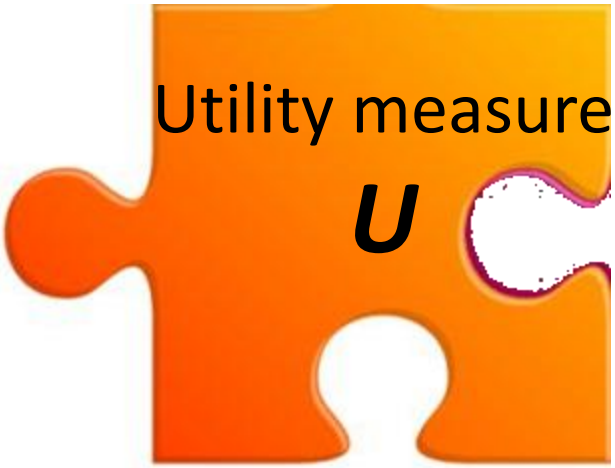
Graphical methods

Operations research methods

Analysis Quality

- "poor models and poor analysis techniques, or even analyzing the data in a totally incorrect way." - Godfrey
- Analyst expertise
- Software availability
- The focus of statistics education

Domain goal → Analysis goal



- Predictive accuracy, lift
- Goodness-of-fit
- Statistical power, statistical significance
- Strength-of-fit
- Expected costs, gains
- Bias reduction, bias-variance tradeoff

Analysis utility → Domain utility

Quality of Utility Measure

- Adequate metric from analysis standpoint (R^2, holdout data)
- Adequate metric from domain standpoint

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

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

Depends on quality of *g, X, f, U* and relationship between them

Statistical Approaches for Increasing InfoQ

Study Design (Pre-Data)

- DOE
- Clinical trials
- Survey sampling
- Computer experiments

Randomization, Stratification, Blinding, Placebo, Blocking, Replication, Sampling frame, Link data collection protocol with appropriate design

Post-Data-Collection

- Data cleaning and preprocessing
- Re-weighting, bias adjustment
- Meta analysis

Recovering "real data" vs. "cleaning for the goal" Handling missing values, outlier detection, re-weighting, combining results

Assessing InfoQ

InfoQ dimensions

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

3 V's of *Big Data*

- Volume
- Variety
- Velocity

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

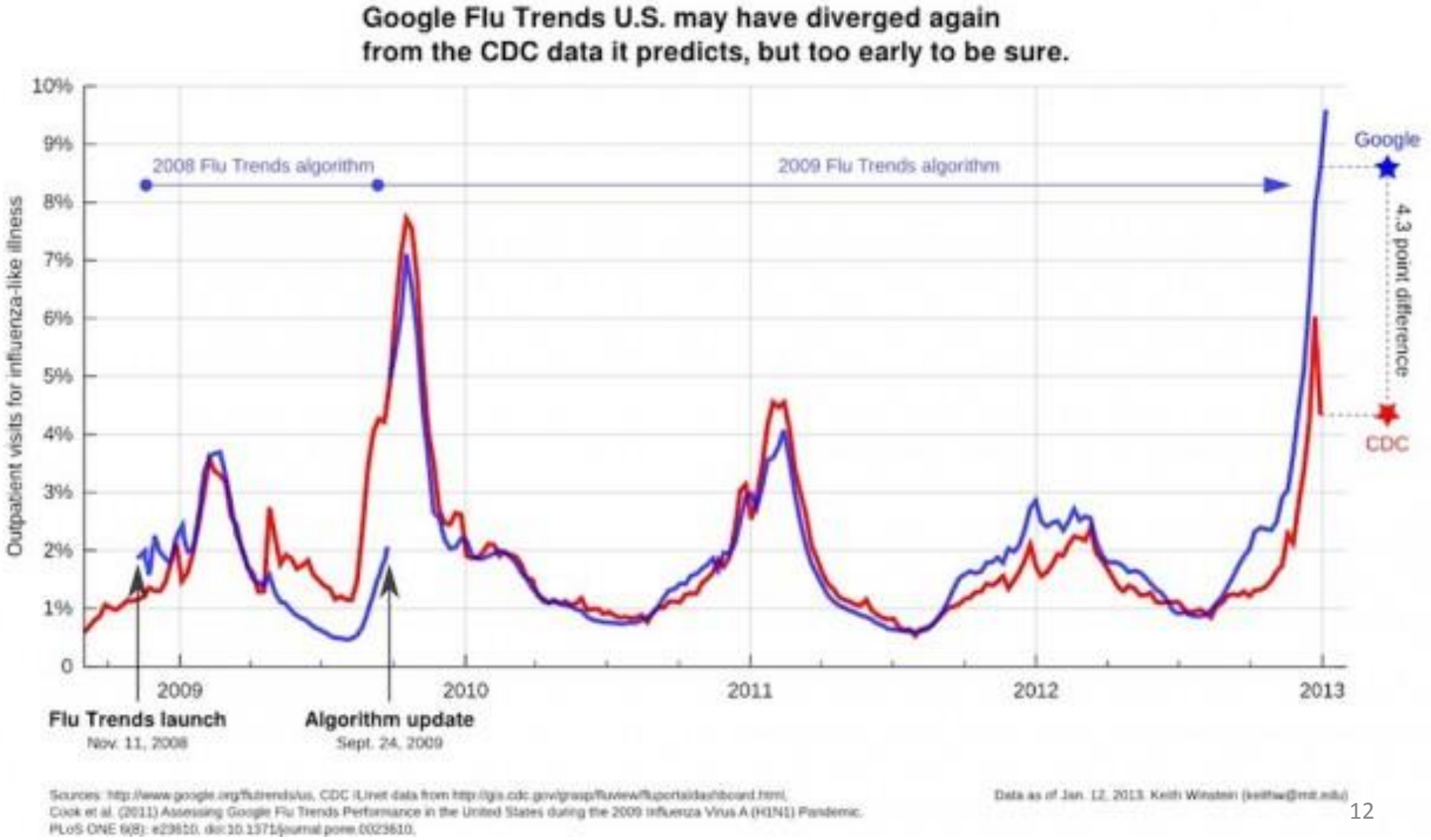
- Relevance
- Accuracy
- Timeliness and punctuality
- Accessibility
- Interpretability
- Coherence
- Credibility

Marketing Research

- Recency
- Accuracy
- Availability
- Relevance

#1 Data Resolution

Measurement scale and aggregation level



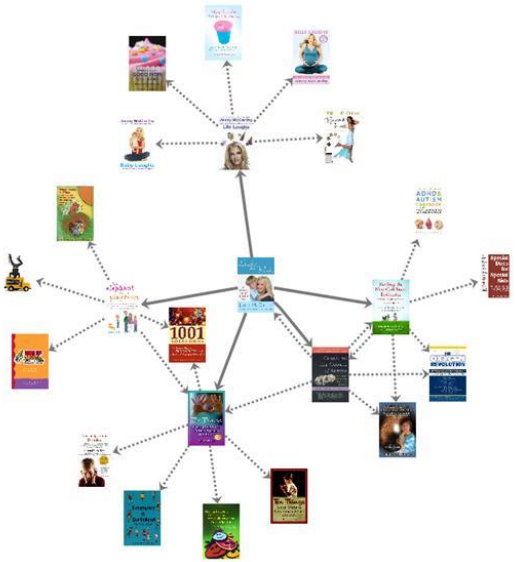
#2 Data Structure

Data Types

- Time series, cross-sectional, panel
- Geographic, spatial, network
- Text, audio, video, semantic
- Structured, semi-, non-structured
- Discrete, continuous

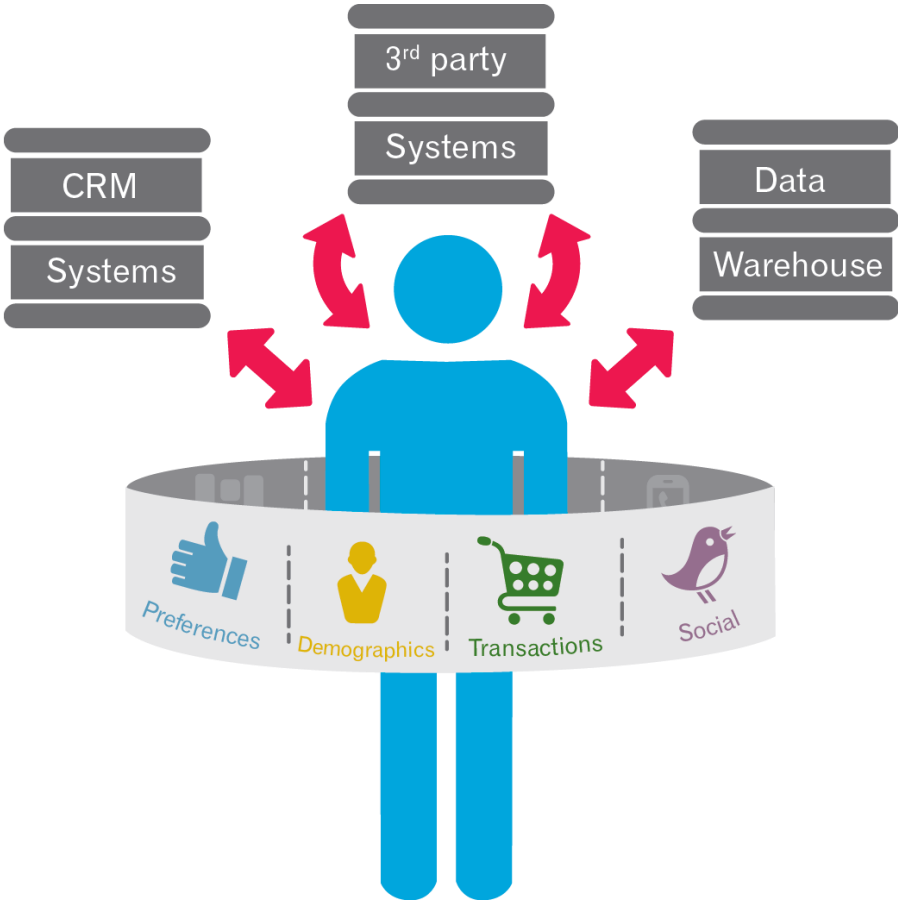
Data Characteristics

Corrupted and missing values due to study design or data collection mechanism



#3 Data Integration

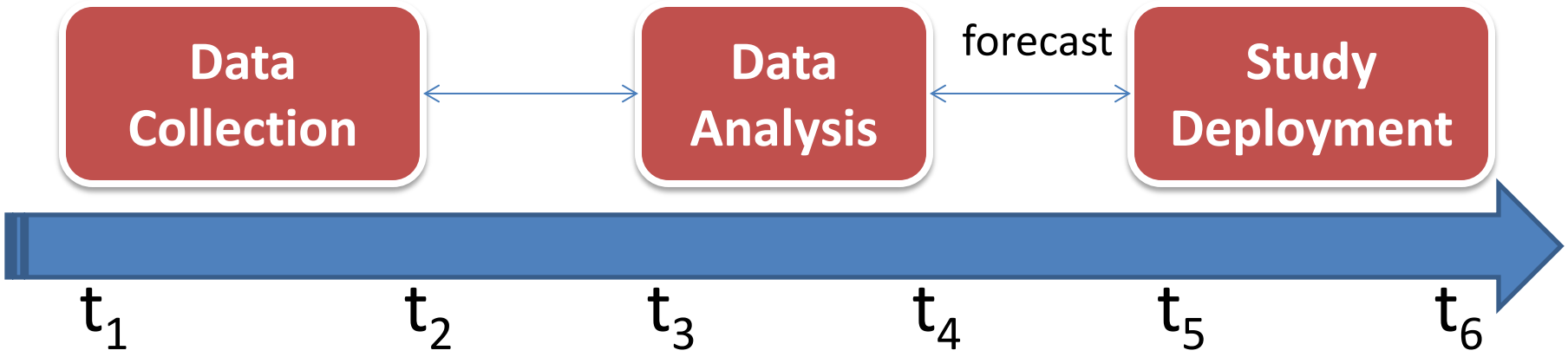
Linkage, privacy-preserving methods: Increase or decrease InfoQ?



#4 Temporal Relevance

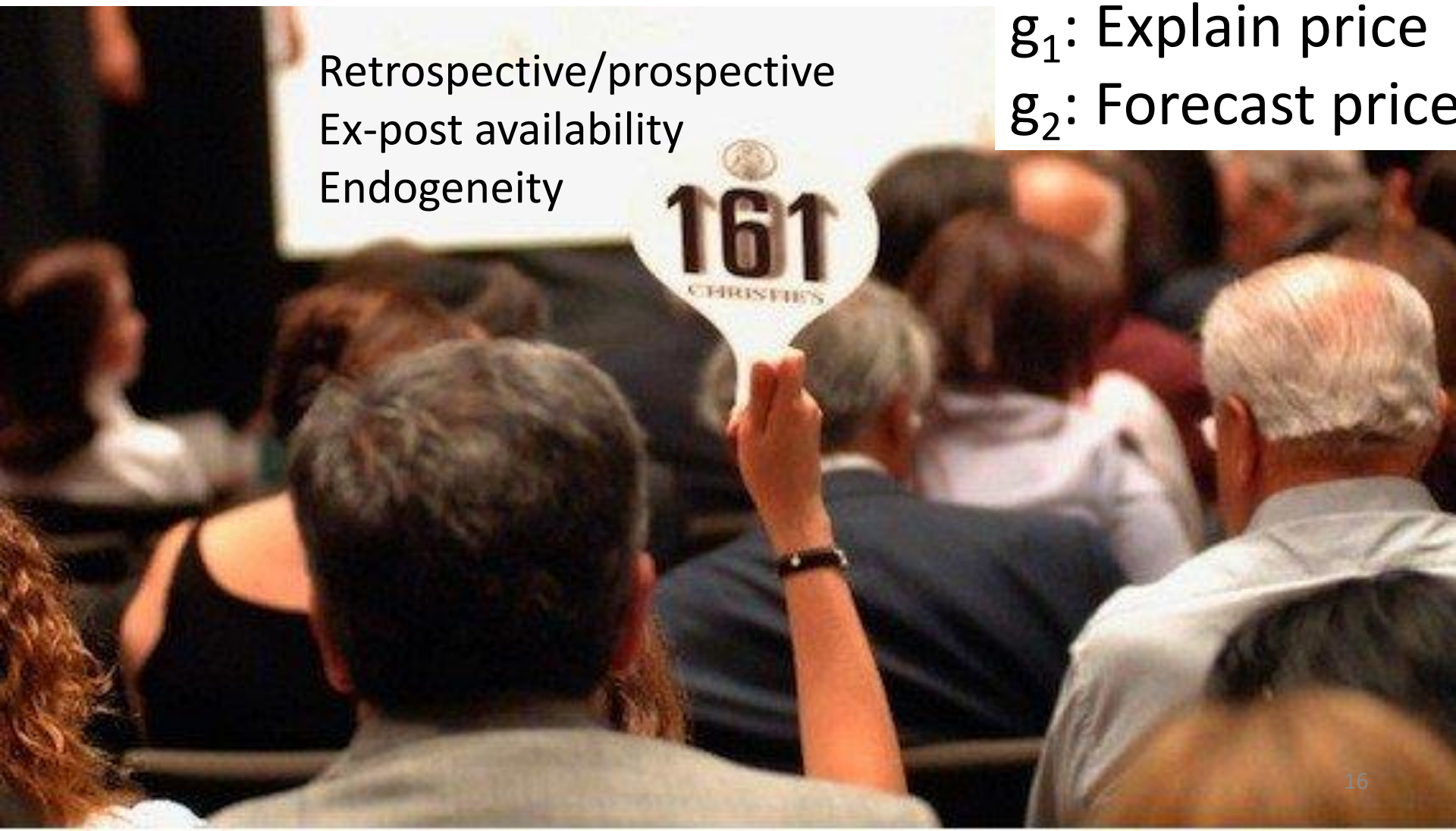
Collection Timeliness
(relevance to *g*)

Analysis Timeliness
(solving the right
problem too late)



g: Prospective vs. retrospective; longitudinal vs. snapshot
Nature of X, complexity of *f*

#5 Chronology of Data & Goal

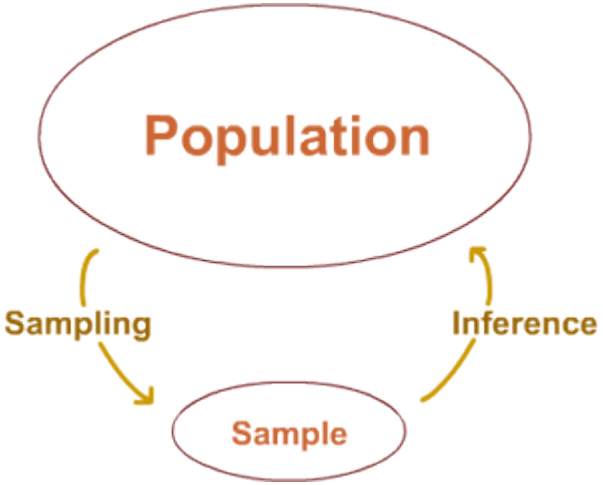


g_1: Explain price

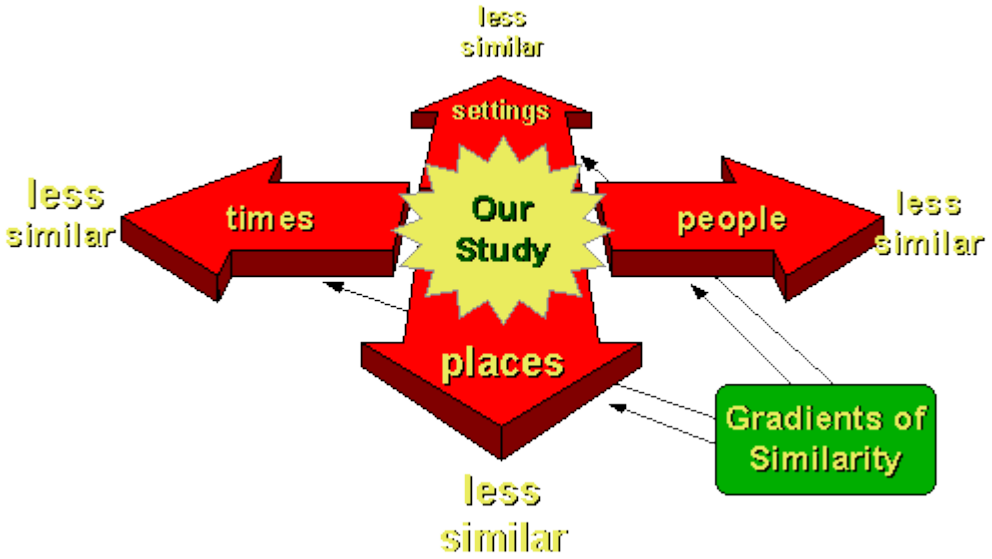
g_2: Forecast price

#6 Generalizability

Statistical generalizability



Scientific generalizability

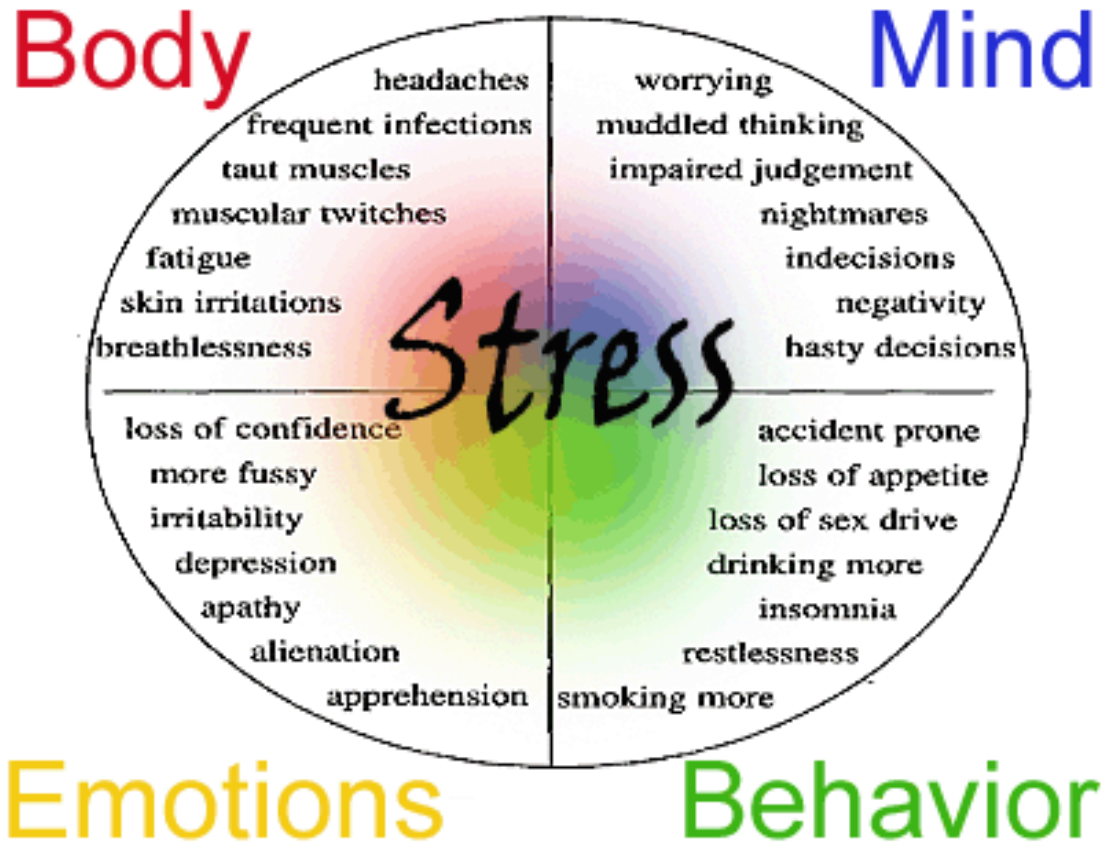


Definition of *g*
Choice of X, *f, U*

#7 Construct Operationalization

χ construct

$X = \theta(\chi)$ operationalization (measurable)



- Causal explanation vs. prediction, description
- Theory vs. data
- Data: Questionnaire, physio measurement

#8 Communication

Visual, written, and verbal presentations and reports

Knowledge must reach the right person at the right time

- Mentoring
- Manuscript reviewing
- Data made available to others
- EDA and shared visualization dashboards

"In the last three years, there has been a concerted effort by those in Washington to reduce government spending and reign in the national debt.

One reason for the budget cuts?

Research by two Harvard economists, Ken Rogoff and Carmen Reinhart. The pair found that **when a country owes more than 90 percent of their GDP, it slides into recession**."

… Fixing this Excel error transforms high-debt countries from recession to growth

AMERICAN PUBLIC MEDIA

Marketplace Economy

LATEST STORIES • SONGS • PODCASTS • BEYOND PAYDAY • INTERVIEW: DONALD RUMSFELD

ECONOMY

Like 305 Tweet 40 Share 4 +1 7 Share 106

The Excel mistake heard round the world

	B	C	I	J	K	L	M
2			Real GDP growth				
3			Debt/GDP				
4	Country	Coverage	30 or less	30 to 60	60 to 90	90 or above	30 or less
26			3.7	3.0	3.5	1.7	5.5
27	Minimum		1.6	0.3	1.3	-1.8	0.8
28	Maximum		5.4	4.9	10.2	3.6	13.3
29							
30	US	1946-2009	n.a.	3.4	3.3	-2.0	n.a.
31	UK	1946-2009	n.a.	2.4	2.5	2.4	n.a.
32	Sweden	1946-2009	3.6	2.9	2.7	n.a.	6.3
33	Spain	1946-2009	1.5	3.4	4.2	n.a.	9.9
34	Portugal	1952-2009	4.8	2.5	0.3	n.a.	7.9
35	New Zealand	1948-2009	2.5	2.9	3.9	-7.9	2.6
36	Netherlands	1956-2009	4.1	2.7	1.1	n.a.	6.4
37	Norway	1947-2009	3.4	5.1	n.a.	n.a.	5.4
38	Japan	1946-2009	7.0	4.0	1.0	0.7	7.0
39	Italy	1951-2009	5.4	2.1	1.8	1.0	5.6
40	Ireland	1948-2009	4.4	4.5	4.0	2.4	2.9
41	Greece	1970-2009	4.0	0.3	2.7	2.9	13.3
42	Germany	1946-2009	3.9	0.9	n.a.	n.a.	3.2
43	France	1949-2009	4.9	2.7	3.0	n.a.	5.2
44	Finland	1946-2009	3.8	2.4	5.5	n.a.	7.0
45	Denmark	1950-2009	3.5	1.7	2.4	n.a.	5.6
46	Canada	1951-2009	1.9	3.6	4.1	n.a.	2.2
47	Belgium	1947-2009	n.a.	4.2	3.1	2.6	n.a.
48	Austria	1948-2009	5.2	3.3	-3.8	n.a.	5.7
49	Australia	1951-2009	3.2	4.9	4.0	n.a.	5.9
50							
51			4.1	2.8	2.8	=AVERAGE(L30:L44)	

www.marketplace.org/topics/economy/excel-mistake-heard-round-world

Assessing InfoQ in Practice

Rating-based assessment

1-5 scale on each dimension:

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

Experience from two research methods courses

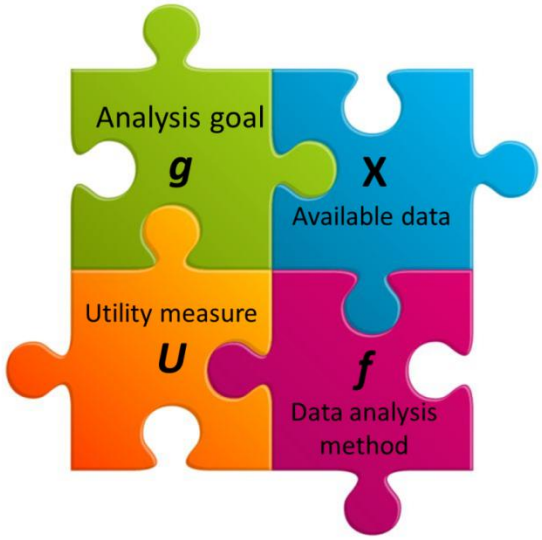
- Preparing a PhD research proposal (U Ljubljana, 50 students, goo.gl/f6bIA)
- Post-hoc evaluation of five completed studies (CMU, 16 students, goo.gl/erNPF)

Shmueli and Kenett (2013), "An Information Quality (InfoQ) Framework for Ex-Ante and Ex-Post Evaluation of Empirical Studies", *Proceedings of IDAM 2013*, Springer-Verlag.

InfoQ: Strengths and Challenges

InfoQ approach streamlines questioning of data value

- "Why should we invest in data?" – management
- Compare value of potential datasets, analyses
- Prioritize/rank projects
- Strengthen functional – analytical relationship



Multiple goals:

- Goals can change during study: Reevaluate InfoQ
- Multiple goals: Prioritize.
  - clinical trials: effect of new drug, adverse effects

To Do:

- Improve InfoQ assessment
- Alternative InfoQ assessment approaches (pilot study, EDA, other)
- Further dimensions (data privacy, human subject compliance and risk)
- Effect of technological advances on InfoQ