International Association of Survey Statisticians (IASS)

Webinar 23 – November 30, 2022

Bridging BigData and Sampling Methodology: What is "big" and where is the "bridge"

Fulvia Mecatti

University of Milano-Bicocca

WHY . . . bridging

Sample survey theory and method have been a reliable "gold standard" framework for over 100 years to create information and new knowledge from data

… and now a paradigm shift

- **in data management and statistical analysis**
- **dragged by digitization and the explosion of the Internet**

SIGNS from the past

- Declining response rates
- Increasing costs to fill in accelerated information needs
- Penetration of digital devices and the Internet

↑

Web Surveys as a first natural response

THE BigData ERA

- Traffic on the WWW, e.g. Social networks, commenting & blog posting
- Software & sensors monitoring business transactions, shipments, suppliers & customers
- digitized health records
- videos & images & audios, e.g. surveillance cameras, Instagram, YouTube
- click streams & log-files, e.g. on e-commerce platforms & IoT
- search queries, e.g. all our Google-ing
- satellite imagery & geo-localized data, e.g. from mobile phones
- . . .

THE BigData ERA

- a tempting abundance of datasets collected passively, self-creating & self-updating
- at great speed and mostly free/lower cost
- for a large portion ready-for-access from our own computer and cell phone

Why should we not want to use it, to harness their power to answer questions, networking, connect … *and* to deal with issues challenging survey sampling theory and practice

IS there a THREATEN?

- Is the impact of the BigData Era threatening our science? Our own competencies?
- A menace to statisticians and scientist at large?

IS there a THREATEN?

WIRED

“The End of Theory: The Data Deluge Makes the Scientific Method Obsolete” (2008)

https://www.wired.com/2008/06/pb-theory/

… 60 years ago, digital computers made information readable. 20 years ago, the Internet made it reachable. 10 years ago, the first search engine crawlers made it a single database […measured in PetaByte=1 million GigaByte and stored in cloud].

… faced with massive data, [the] approach to science - hypotheses, model, test - is becoming obsolete…

IS there a THREATEN?

S. Lohr 2016
Distinguished Lecture

JOINT PROGRAM IN SURVEY METHODOLOGY

"Reaching for a random sample
in the age of big data is like
clutching at a horse whip
in the era of the motor car."

Mayer-Schönberger and Cukier (2014), *Big Data: A Revolution That Will Transform How We Live, Work, Think*

prof. of Internet Governance and Regulation @Oxford Univ formerly @Harvard
data editor of the Economist

IS there a THREATEN?

- Is BigData underplaying survey sampling theory & practice ?
- Supplanting it as a low-cost futuristic option?

IS there a THREATEN?

Survey Methodology, June 2020

Are probability surveys bound to disappear for the production of official statistics?

Jean-François Beaumont[1]

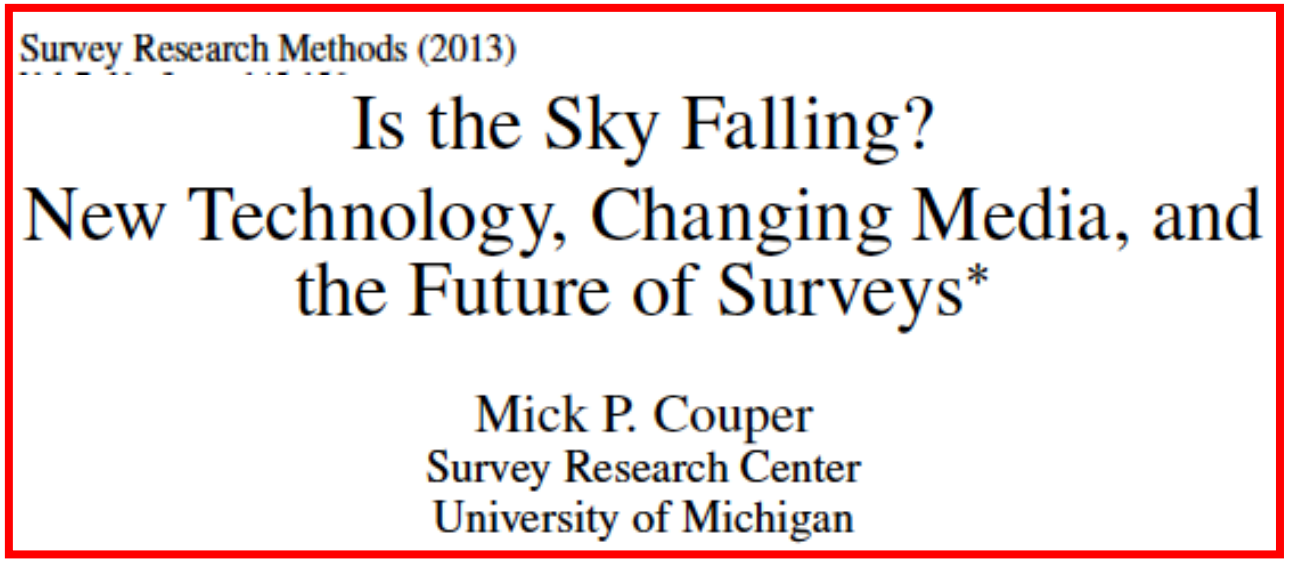
Survey Research Methods (2013)

Is the Sky Falling?
New Technology, Changing Media, and the Future of Surveys*

Mick P. Couper
Survey Research Center
University of Michigan

IS there a THREATEN?

- **Limits of BigData, e.g.** *Opportunity for mischief*
- *Big data are here to stay*
- *we should welcome – rather than fear or oppose – these new developments*



Survey Research Methods (2013)

Is the Sky Falling?
New Technology, Changing Media, and the Future of Surveys*

Mick P. Couper
Survey Research Center
University of Michigan

IS there a THREATEN?

Survey Methodology, June 2020

Are probability surveys bound to disappear for the production of official statistics?

Jean-François Beaumont[1]

1. *the decline in response rates in probability surveys*
2. *the high cost of data collection*
3. *the increased burden on respondents*
4. *the desire for access to "real-time" statistics* ←
5. *the proliferation of non-probability data sources (as do is BigData)* ←

The 5 key factors that make the question relevant . . .

IS there a THREATEN?

Survey Methodology, June 2020

Are probability surveys bound to disappear for the production of official statistics?

Jean-François Beaumont[1]

. . . that have brought to

a wind of change has been blowing over national statistical agencies, and other data sources [*different from standard (probability) surveys*] *are being increasingly explored. "*

Ri-POSITIONING

Survey Methodology, December 2017

Sample survey theory and methods: Past, present, and future directions

J.N.K. Rao and Wayne A. Fuller[1]

"[…] improved data collection devices, and availability of auxiliary data, some of which will come from "Big Data". Survey taking will be impacted by changing cultural behavior and by a changing physical-technical environment."

Ri-POSITIONING

Positioning Household Surveys for the Next Decade

Haoyi Chen (UNSD Inter-Secretariat Working Group on Household Surveys)

2021 https://unece.org/sites/default/files/2021-09/DC2021_S4_Chen_UNSD_A.pdf

"Household surveys are facing funding challenges and skepticism on their continued utility within the changing data landscape […including] *How can we establish sustainable household survey programs that are resilient and versatile to future shocks like COVID-19?"*

IASS Webinar 11 - Nov 2021

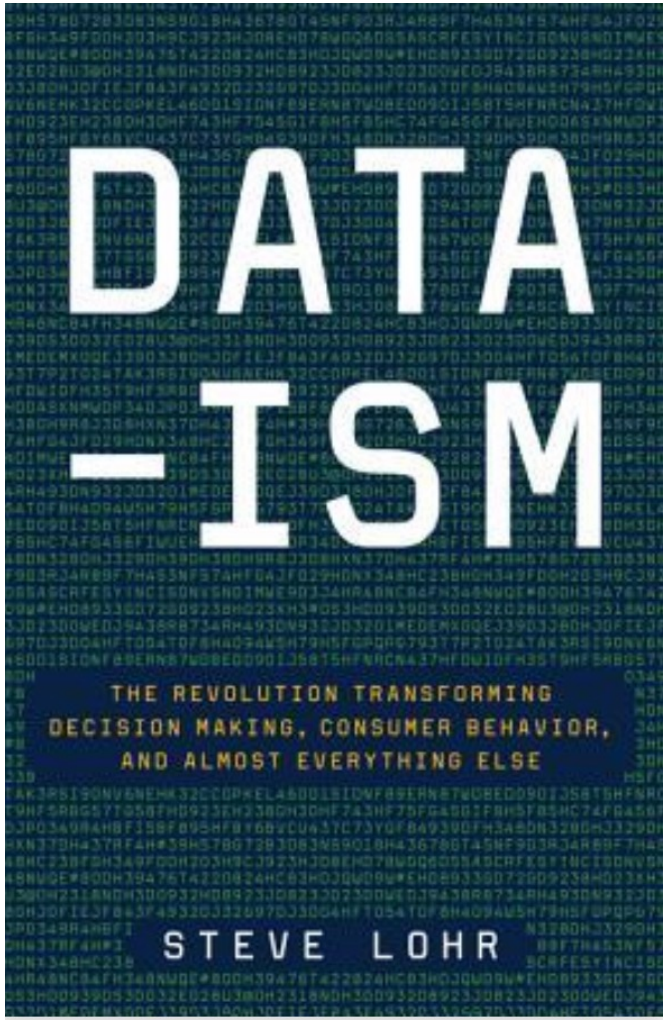
G. Carletto co-director ISWGHS

http://isi-iass.org/home/wpcontent/uploads/IASS_HouseholdSurveys_Nov23_webinar.pdf

A compelling CALL-FOR-ACTION

S-Index

Statistics
Sampling
Survey …



S. Lohr 2016

Distinguished Lecture

JOINT PROGRAM IN SURVEY METHODOLOGY

WHAT IS BIG … data

- trending buzzword
- blanket term (*Wikipedia)*
- *Organic data* R.Groves 2011 PubOpQ; *Found data* AAPOR 2015
- *"a collective term for the increasingly diverse range of data sources available through 'web of everything' "* S-M.Tam&F. Clarke 2015 IRS
- https://ischoolonline.berkeley.edu/blog/what-is-big-data/ 2019

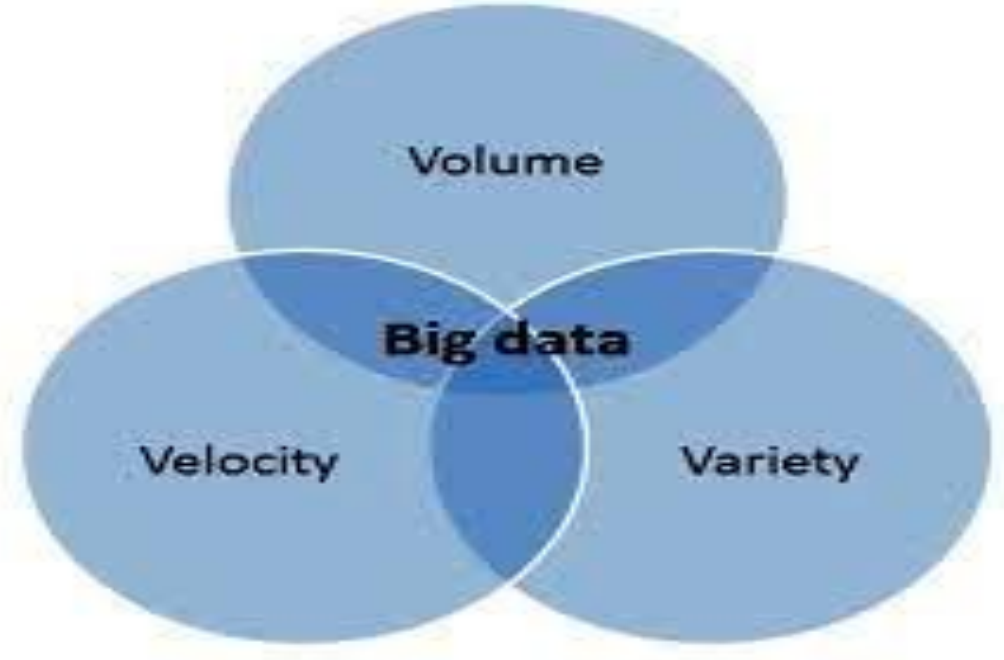
  "Today, the concept of big data is not only less compelling, but it's also potentially misleading"

. . . A blurry definition

. . . not simply a matter of SIZE

WHAT IS BIG: the Vs definition(s)

1. Volume
2. Velocity
3. Variety



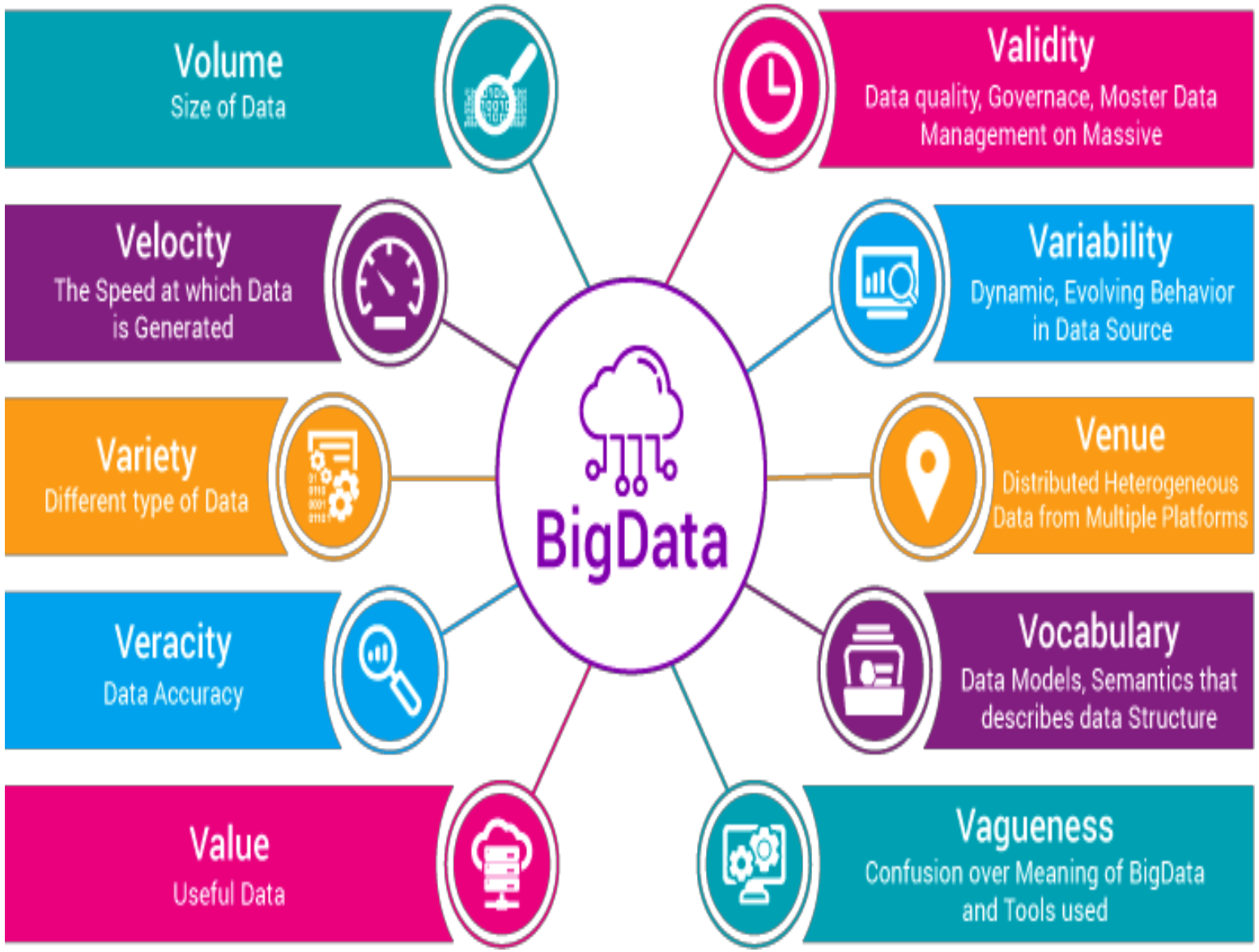
industry analyst Doung Laney
2001

WHAT IS BIG: the Vs definition(s)

1. Volume
2. Velocity
3. Variety
4. Value
5. Veracity

WHAT IS BIG: the Vs definition(s)

2014,
Data Science Central,
Kirk Born



WHAT IS BIG: the Vs definition(s)

11.	Volatility	Duration of Usefulness
12.	Visualization	Data Act/ Data Process
13.	Virality	Spreading Speed rate at which the data is broadcast
14.	Viscosity	Lag of Event time difference

International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 09 | Sep-2017 www.irjet.net p-ISSN: 2395-0072

The 17 V's Of Big Data

Arockia Panimalar.S [1], Varnekha Shree.S[2], Veneshia Kathrine.A[3]

[1] Assistant Professor, Department of BCA & M.Sc SS, Sri Krishna Arts and Science College, Tamilnadu, India
[2,3] III BCA, Department of BCA & M.Sc SS, Sri Krishna Arts and Science College, Tamilnadu, India

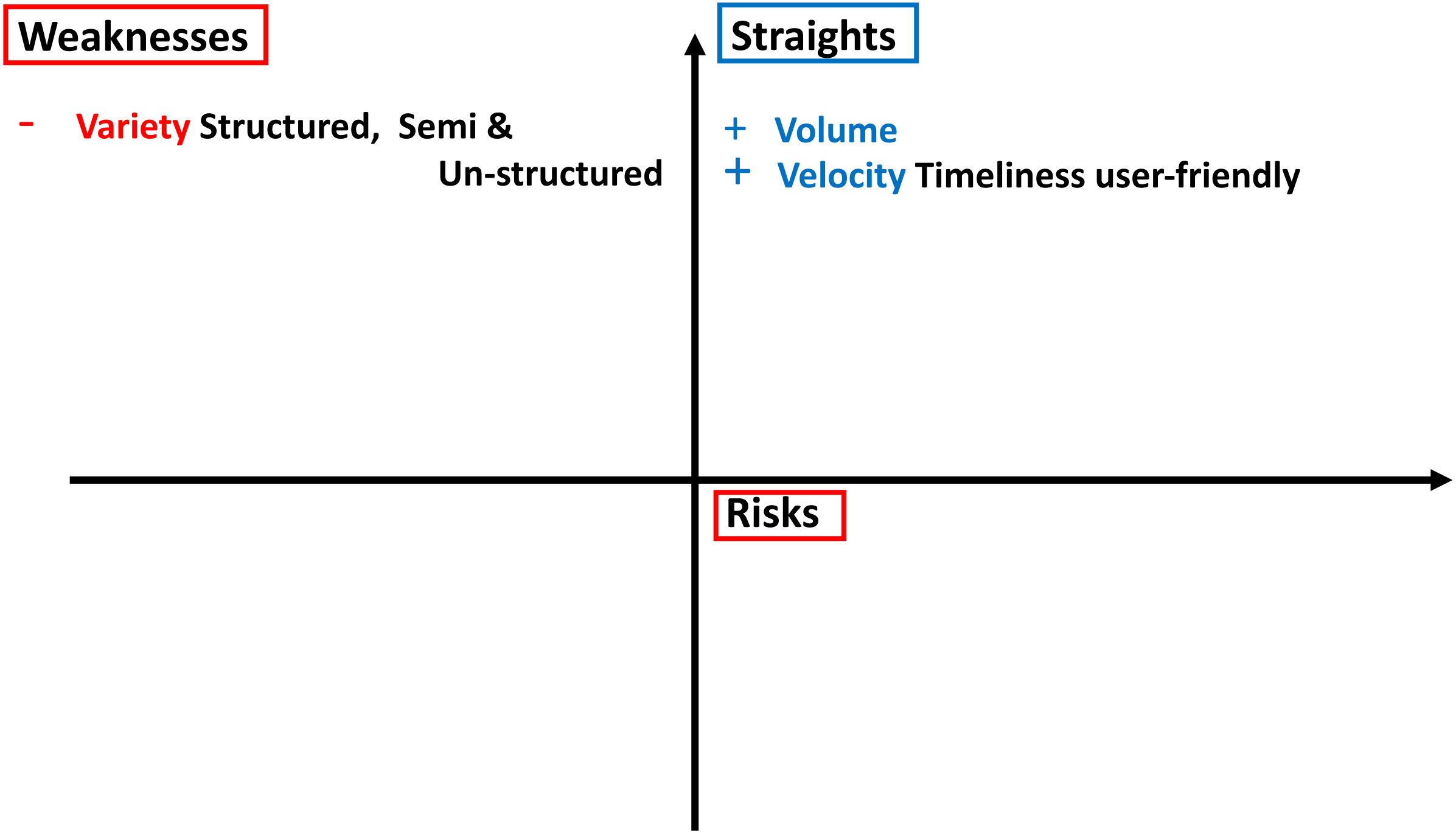
---

A. Verbosity — **Redundancy** of sources **waste** of storage space and time

B. Voluntariness — **Easy** and **full availability** to a range of users' context

C. Versatility — **Flexibility** to be used differently in different context

WHAT IS BIG: a personal classification


Weaknesses
- Variety Structured, Semi & Un-structured
Straights
+ Volume
+ Velocity Timeliness user-friendly
Risks

Weaknesses

- Variety
- Value low signal to noise ratio

Big Data – Big Noise UNIDO 2015

https://unstats.un.org/unsd/accsub/2015docs-26th/Presentation-UNIDO.pdf

Straights

+ Volume + Variety
+ Velocity
+ Value BigData-BigImpact WEF 2012

https://www.weforum.org/reports/big-data-big-impact-new-possibilities-international-development

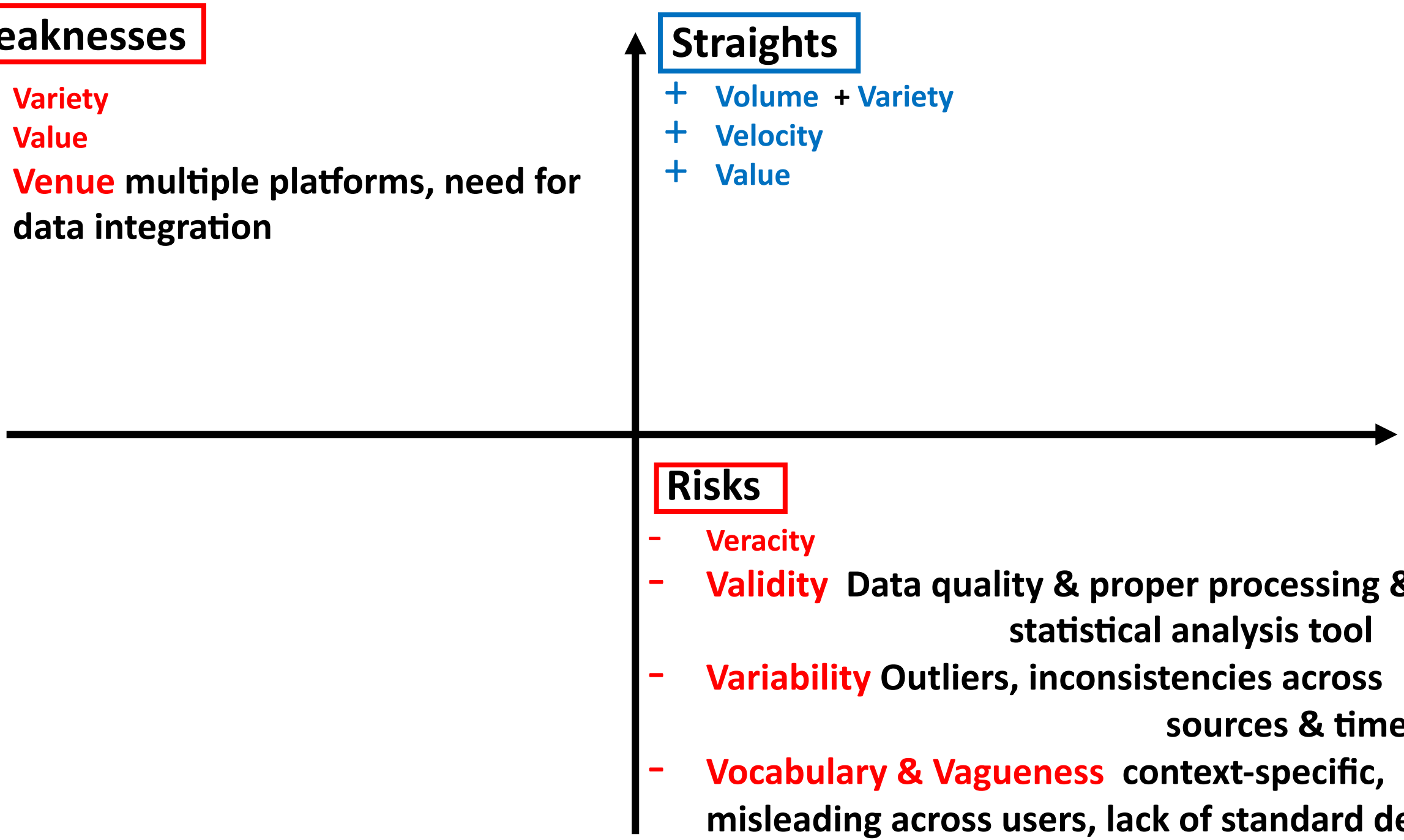
"declared data a new class of economic asset, like currency or gold"

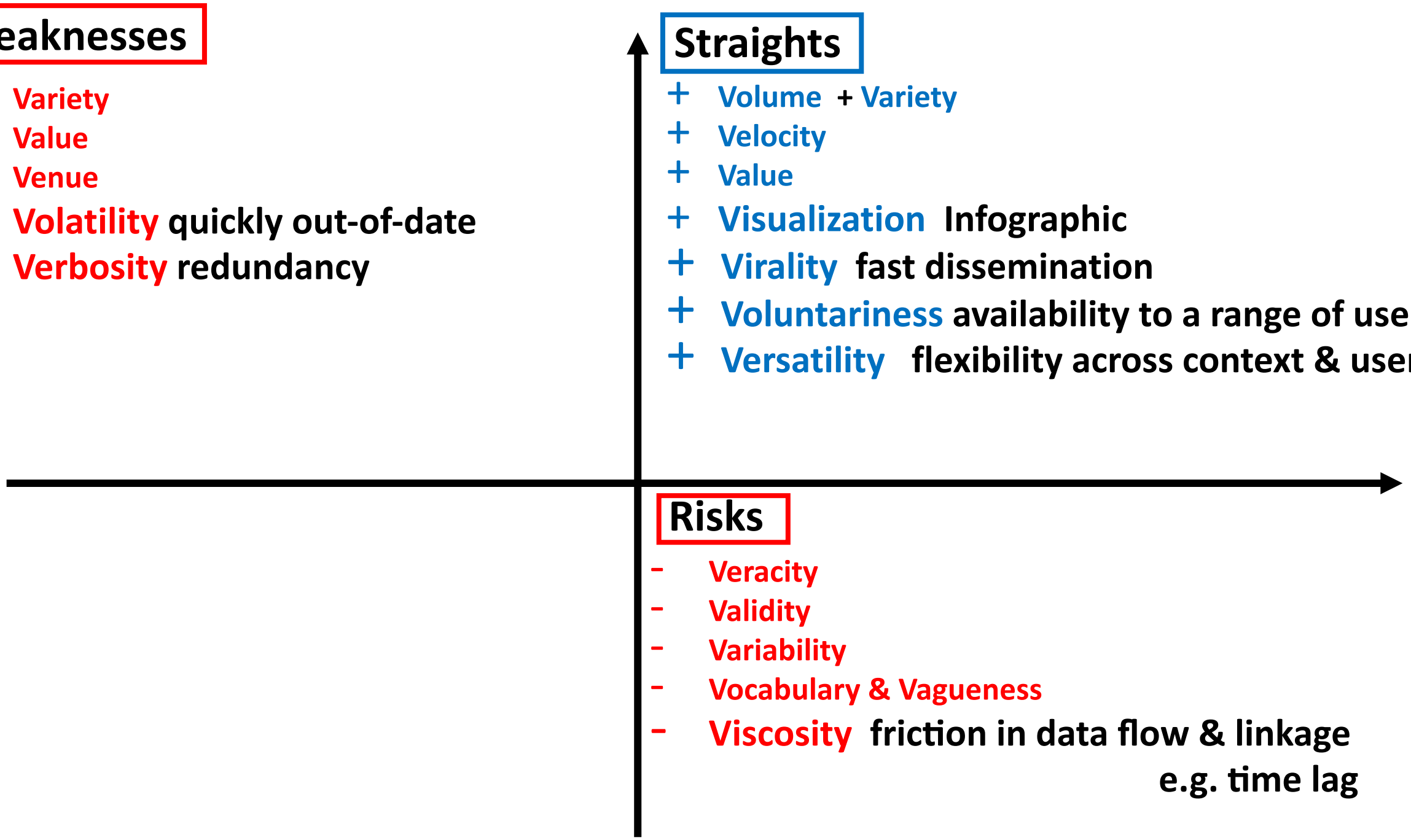
Risks

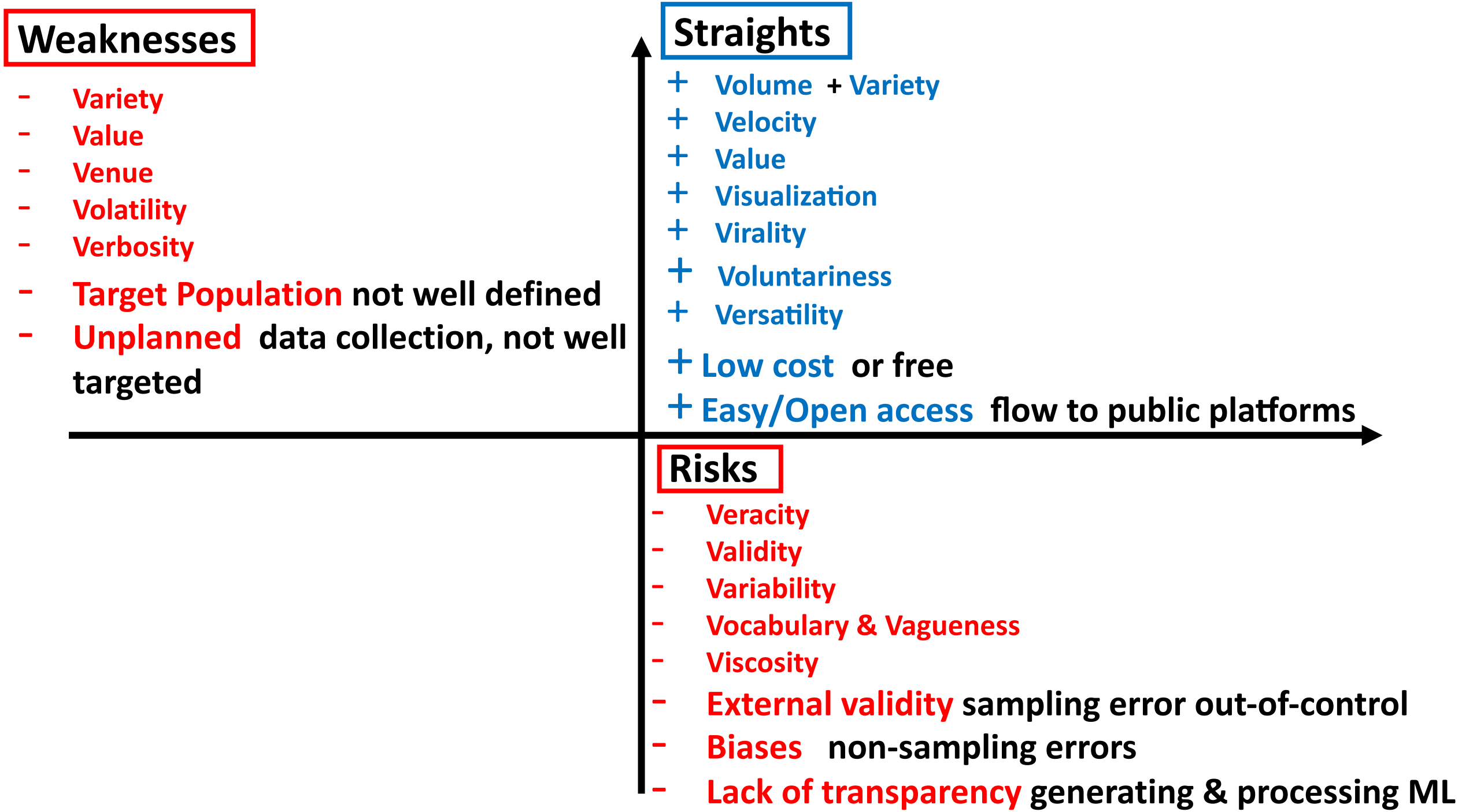
- Veracity accuracy & truthfulness of data & trustfulness of data-source

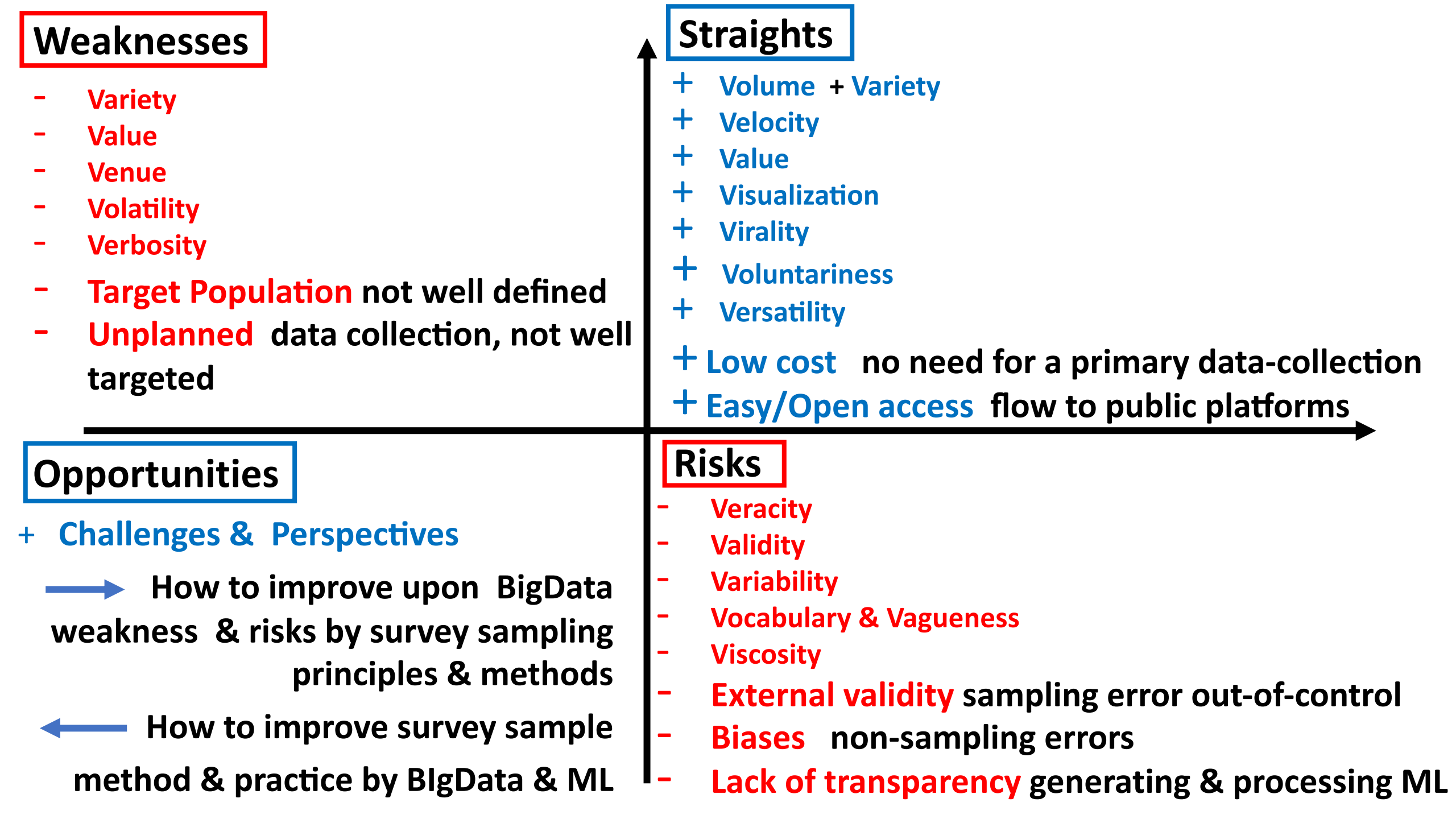
"Like good wine, the provenance of the data we analyze is important, as it is quality"

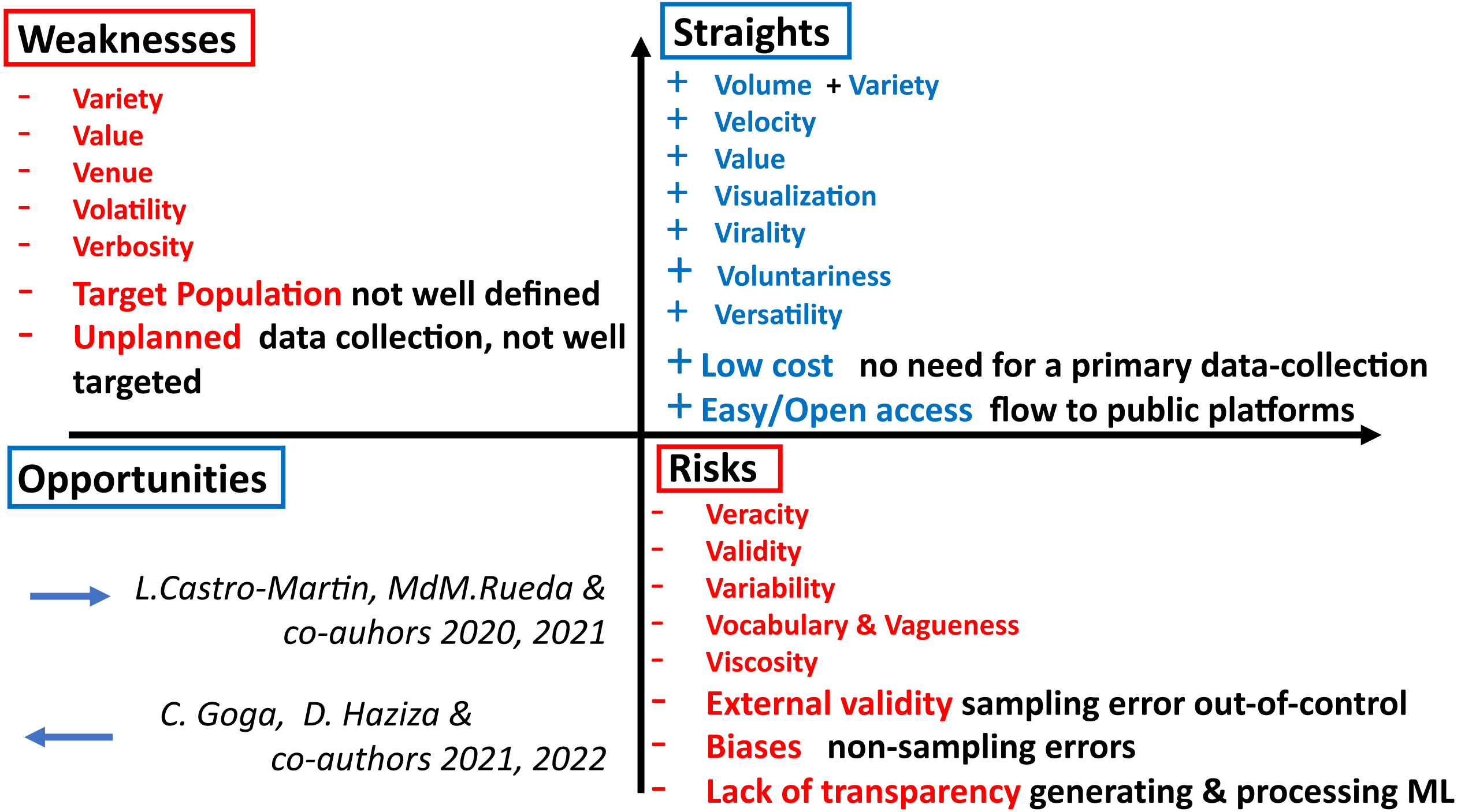
M.P.Couper 2013 SurResMeth


Weaknesses
- Variety
- Value
- Venue multiple platforms, need for data integration
Straights
+ Volume + Variety
+ Velocity
+ Value
Risks
- Veracity
- Validity Data quality & proper processing & statistical analysis tool
- Variability Outliers, inconsistencies across sources & time
- Vocabulary & Vagueness context-specific, misleading across users, lack of standard def


Weaknesses
- Variety
- Value
- Venue
- Volatility quickly out-of-date
- Verbosity redundancy
Straights
+ Volume + Variety
+ Velocity
+ Value
+ Visualization Infographic
+ Virality fast dissemination
+ Voluntariness availability to a range of users
+ Versatility flexibility across context & users
Risks
- Veracity
- Validity
- Variability
- Vocabulary & Vagueness
- Viscosity friction in data flow & linkage
e.g. time lag


Weaknesses
- Variety
- Value
- Venue
- Volatility
- Verbosity
- Target Population not well defined
- Unplanned data collection, not well targeted
Straights
+ Volume + Variety
+ Velocity
+ Value
+ Visualization
+ Virality
+ Voluntariness
+ Versatility
+ Low cost or free
+ Easy/Open access flow to public platforms
Risks
- Veracity
- Validity
- Variability
- Vocabulary & Vagueness
- Viscosity
- External validity sampling error out-of-control
- Biases non-sampling errors
- Lack of transparency generating & processing ML


Weaknesses
- Variety
- Value
- Venue
- Volatility
- Verbosity
- Target Population not well defined
- Unplanned data collection, not well targeted
Straights
+ Volume + Variety
+ Velocity
+ Value
+ Visualization
+ Virality
+ Voluntariness
+ Versatility
+ Low cost no need for a primary data-collection
+ Easy/Open access flow to public platforms
Opportunities
+ Challenges & Perspectives
How to improve upon BigData weakness & risks by survey sampling principles & methods
How to improve survey sample method & practice by BIgData & ML
Risks
- Veracity
- Validity
- Variability
- Vocabulary & Vagueness
- Viscosity
- External validity sampling error out-of-control
- Biases non-sampling errors
- Lack of transparency generating & processing ML


Weaknesses
- Variety
- Value
- Venue
- Volatility
- Verbosity
- Target Population not well defined
- Unplanned data collection, not well targeted
Straights
+ Volume + Variety
+ Velocity
+ Value
+ Visualization
+ Virality
+ Voluntariness
+ Versatility
+ Low cost no need for a primary data-collection
+ Easy/Open access flow to public platforms
Opportunities
L.Castro-Martin, MdM.Rueda & co-auhors 2020, 2021
C. Goga, D. Haziza & co-authors 2021, 2022
Risks
- Veracity
- Validity
- Variability
- Vocabulary & Vagueness
- Viscosity
- External validity sampling error out-of-control
- Biases non-sampling errors
- Lack of transparency generating & processing ML

WHERE IS THE BRIDGE

WHERE IS THE BRIDGE

- **NON-PROBABILITY SAMPLING**

  unknown, unplanned & inherently biased mechanism generating the BigData

*Non-probability surveys have been around for a long time, but the recent attention that has been paid to such methods can be partly attributed to the **rise of BIgData***

AAPOR Report 2013
Task Force on Non-probability sampling

WHERE IS THE BRIDGE

- **NON-PROBABILITY SAMPLING**

J-F.Beaumont 2022 IASS Webinar 13

J-F. Beaumont & JNK.Rao 2021 Survey Statistician

PL.Conti 2022 SIS

- **Bias can not be corrected upon sample data itself**
- **Extra-info required:**
  1. **available probability reference sample**
  2. **non-testable assumptions**
- **Several methods based on Propensity Score adjustment**

WHERE IS THE BRIDGE

- **NON-PROBABILITY SAMPLING**

L.Chen P.Li, JNK.Rao & C.Wu 2022 CanJStat

Pseudo empirical likelihood inference for non-probability survey samples

WHERE IS THE BRIDGE

- **EL nonparametric quasi-likelihood method for *iid* inference**
  Owen 1988
- **EL nonparametric Likelihood-ratio type Confidence Intervals with no variance estimation needed and range-respecting (unlike customary large-sample normal theory CI)**
- **Not directly applying to complex sample designs**
- **Pseudo-EL does based on a discrete probability measure on sample data & inclusion probabilities considered in PE(log)L function**

C.Wu&JNK.Rao (2006) CanJStat
C.Wu&M.E.Thompson 2020 Springer Ch. 8

Pseudo-EL APPROACH

L.Chen P.Li, JNK.Rao & C.Wu 2022 CanJStat

Pseudo empirical likelihood inference for non-probability survey samples

- **Data:** **NP** $\{(x_i, y_i), i \in s_A\}$

- **PS or participating probability:**

strong-ignorability condition

Rosenbaum & Rubin (1983) Biometrika

Pseudo-EL APPROACH

L.Chen P.Li, JNK.Rao & C.Wu 2022 CanJStat

Pseudo empirical likelihood inference for non-probability survey samples

- **Data: NP** $\{(x_i, y_i), i \in s_A\}$ **& reference** $\{(x_i, \pi_i^B), i \in s_B\}$ **with** $d_i^B = 1/\pi_i^B$ **known**
- **PS or participating probability:** $\pi_i^A = P\{i \in s_A | (x_i, y_i)\} = P\{i \in s_A | x_i\}$
- **Estimated PS** $\hat{\pi}_i^A$ **by using** x_i **from** s_B **leading to** $\hat{d}_i^A = 1/\hat{\pi}_i^A$

Doubly robust inference with non-probability survey samples

Y.Chen, P.Li& C.Wu 2020 JASA

Pseudo-EL APPROACH

L.Chen P.Li, JNK.Rao & C.Wu 2022 CanJStat

Pseudo empirical likelihood inference for non-probability survey samples

1. **Discrete probability measure** $\{p_i,\ i = 1 \ldots n_A\}$

2. **log Pseudo-EL function** $l_{PEL}(p_i,\ i = 1 \ldots n_A) = n_A \sum_{i \in s_A} \hat{d}_i^A \ \log(p_i)$

3. **Normalization constraint** $\sum_{i \in s_A} p_i = 1$

4. **Parameter constraint**
   e.g. for estimating pop mean $\sum_{i \in s_A} p_i y_i = \mu_y$

Pseudo-EL APPROACH

L.Chen P.Li, JNK.Rao & C.Wu 2022 CanJStat

Pseudo empirical likelihood inference for non-probability survey samples

$$\max_{\substack{p_i \\ \sum_{i\in s_A} p_i = 1}} l_{PEL}(p_i,\ i = 1 \ldots n_A) \longrightarrow \hat{p}_i = \hat{d}_i^A \longrightarrow \hat{\mu}_{PEL} = \frac{\sum_{i\in s_A} y_i \hat{d}_i^A}{\sum_{i\in s_A} \hat{d}_i^A} = \frac{\hat{\mu}_{y\,HT}}{\hat{N}_{HT}^A}$$

- **Likelihood ratio-type CI** *C.Wu & JNK.Rao 2006 CanJStats*
- **(simplified) bootstrap calibration for variance estimation**
  C.Wu & JNK.Rao 2010 StatsProbLetters
- **Calibration constraint(s) can be accommodated as additional estimating function (moment constraint)**

WHERE IS THE BRIDGE

- **DATA INTEGRATION & DATA COMBINATION**

JNK.Rao 2021 $Sankhyā$
SL.Lohr 2021 SurvMeth

Multiple-frame surveys for a multiple-data-source world

the use of MF is suggested as an organizing principle to combine data from multiple sources

WHERE IS THE BRIDGE

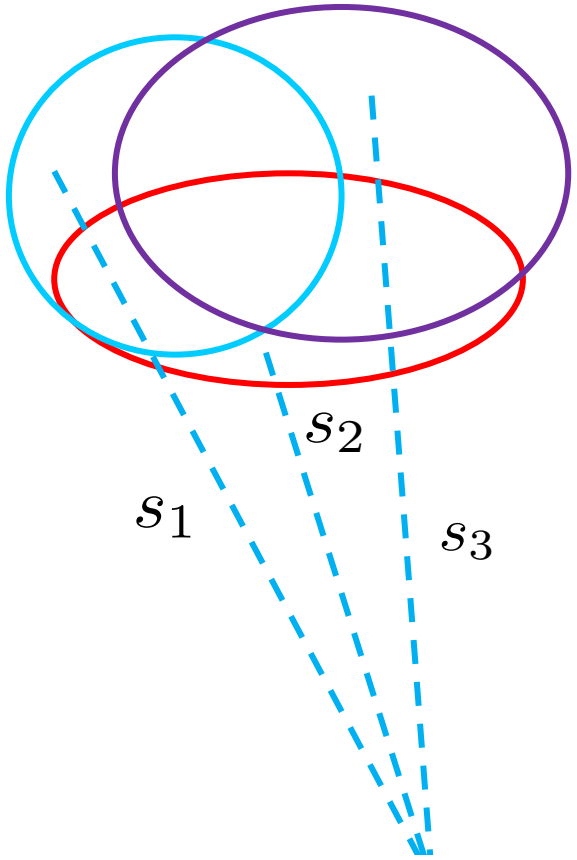
- **DATA INTEGRATION & DATA COMBINATION**

JNK.Rao 2021 $Sankhyā$
SL.Lohr 2021 SurvMeth

Multiple-frame surveys for a multiple-data-source world

"Although the types of data (and the speed with which some types of data can be collected) have changed in recent years, the basic structure of the problem for combining data sources is unchanged from the earliest dual-frame surveys"

DUAL & MULTIPLE FRAME SURVEY



Data & Estimates

- Overlap among frames leads to mixed & multiple opportunities of selecting the same unit in the final sample
- Increased inclusion probability for units appearing in more than one frame (whether or not multiple–listed units are in fact selected)
- Chances for overlapping frame-samples (whether or not duplications actually occur in the final sample)
- Overlap complexity quickly increases with the number of frames used

WHERE IS THE BRIDGE

- **DATA INTEGRATION & DATA COMBINATION**
- **MF surveys are essentially a matter of combining data from different frame-samples**

  AC.Singh & F.Mecatti (2011) JOS

  F.Mecatti & AC.Singh (2014) JFrenchStatSoc
- **A fresh view & research perspectives under the simplified, unified & principled MF multiplicity approach**
- **Expanding the "traditional" structure of overlapping frames (Area and List frames) to include BigData frames**

  administrative records & business transactions, Sensor & satellite data, Internet & Social media, non-probability

WHERE IS THE BRIDGE

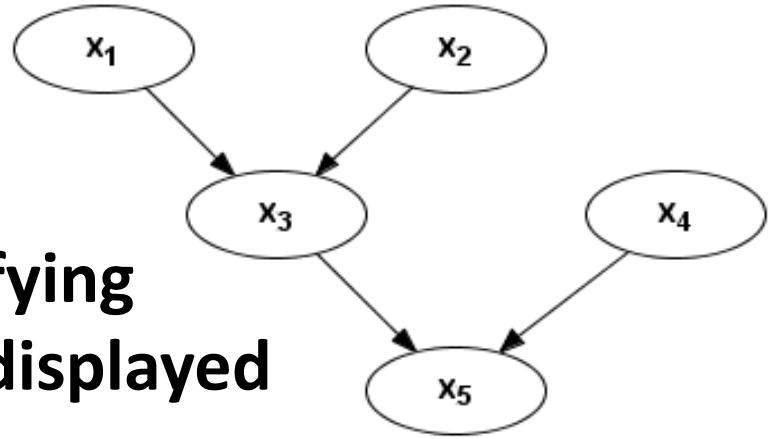
- **EMERGING NEEDS & APPLICATION AREAS involving BigData associated to ML & AI**
  - **Environmental data science**
  - **Computational Social Science**
  - **Causal inference from observational data & causal statistical learning** *L.Giammei 2022*

https://web.uniroma1.it/memotef/sites/default/files/Giammei_onlinefirst_Phd_2022.pdf

Improved new-generation ML methodologies transparent & statistically inferential

WHERE IS THE BRIDGE

- **EMERGING NEEDS & APPLICATION AREAS**

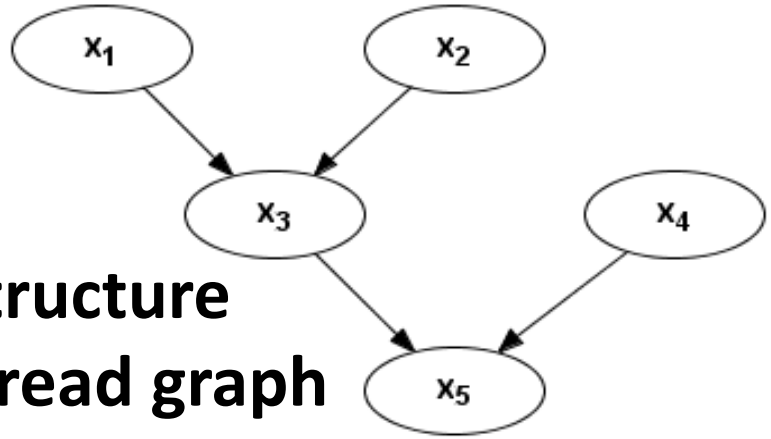


Bayesian Networks

- **Graphical Multivariate Statistical Models satisfying Sets of conditional independence statements displayed in a DAG**
- **Estimated via structural learning, e.g. PC constrain-based algorithm, fed with large dataset, starting on a complete undirected graph, via a sequence of conditional independence tests.**

WHERE IS THE BRIDGE

- **EMERGING NEEDS & APPLICATION AREAS**

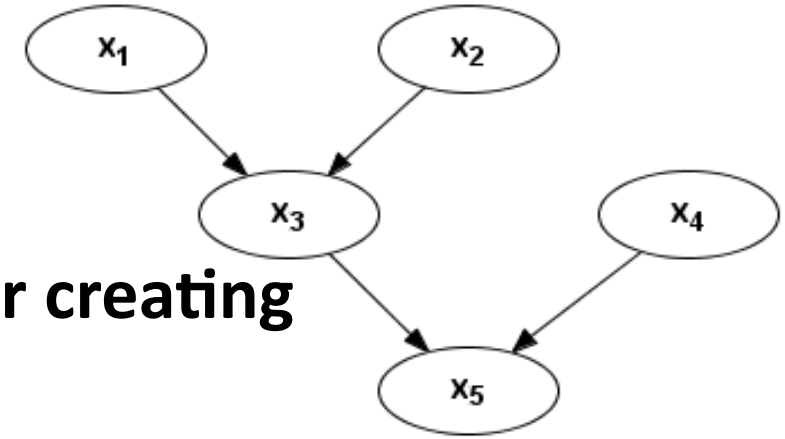


Bayesian Networks (BNs)

- **The estimated BN is a map of the association structure between all variables displayed in the easy-to-read graph**
- **The graphical interface has an analytic counterpart, in the form of estimated probability distributions (joint, marginal and conditional)**

WHERE IS THE BRIDGE

- **EMERGING NEEDS & APPLICATION AREAS**



Bayesian Networks (BNs)

- **The estimated BN is also a predictive engine for creating scenarios and for what-if analysis**
- **Any supplied new piece of information, propagates throughout the map and updates occur across the whole estimated BN**

WHERE IS THE BRIDGE

- **EMERGING NEEDS & APPLICATION AREAS**

  Enhancing Treatment Effect Evaluation in Observational Study: a Propensity Score Method based on Bayesian Networks

  F.Cugnata, P. M.V. Rancoita, PL.Conti, A.Briganti, C. Di Serio, F. Mecatti & Paola Vicard SIS 2022

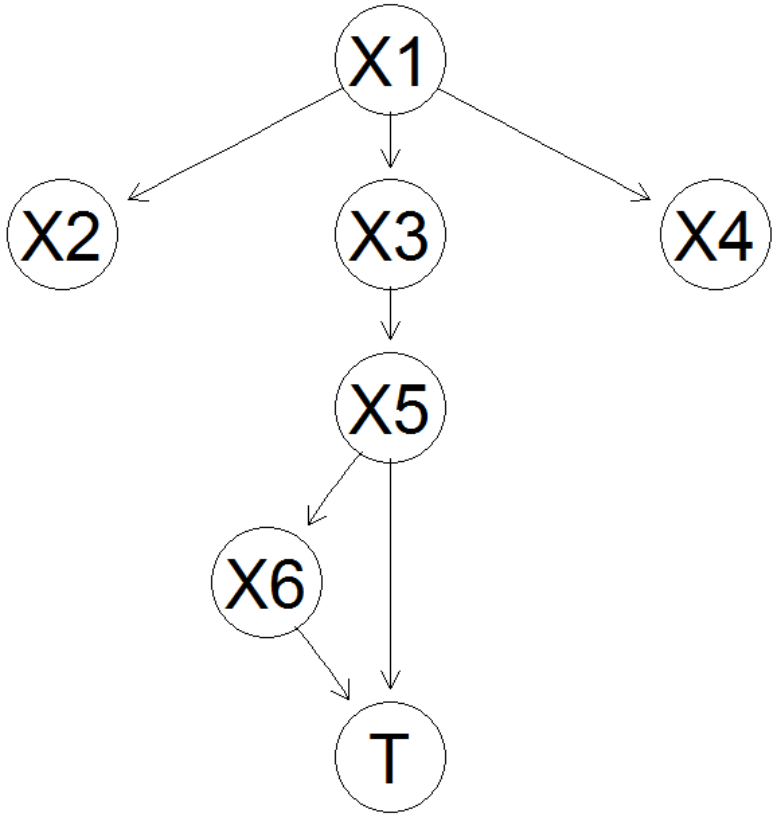
- **Observational data-sets are often larger-scale and easier to attain than experimental data , i.e. gold standard for causal inference**
- **Current practice based on Potential Outcomes approach with a logistic model for the Propensity Score** *Imbens&Rubin 2015 CambridgeUPress*

WHERE IS THE BRIDGE

- **EMERGING NEEDS & APPLICATION AREAS**

  Bayesian Networks (BNs)

- **to estimate PS, upon discrete covariates** x
- **then estimate the Average Treatment Effect**

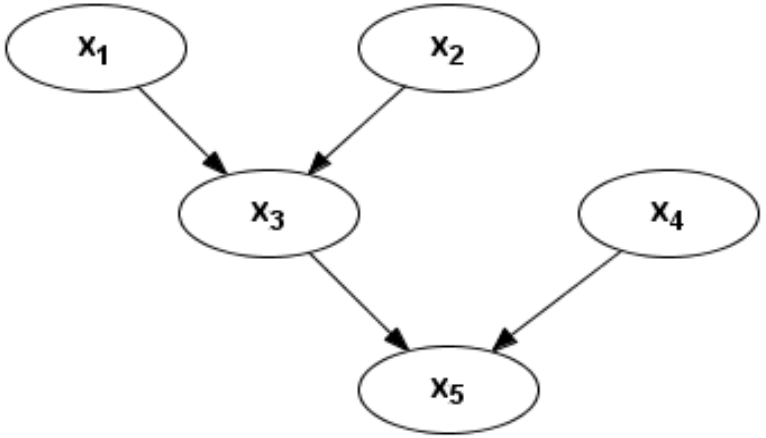


$$\hat{ATE}_{HT} = \frac{1}{n}\sum_{i=1}^{n} I_{(Y_i=1)} I_{(T_i=k)} \hat{p}_k(\boldsymbol{x}_i)^{-1};\ \ k = 0, 1$$

$$\hat{ATE}_{H} = \frac{1}{\sum_{i=1}^{n} I_{(T_i=k)} \hat{p}_k(\boldsymbol{x}_i)^{-1}} \sum_{i=1}^{n} I_{(Y_i=1)} I_{(T_i=k)} \hat{p}_k(\boldsymbol{x}_i)^{-1};\ \ k = 0, 1.$$

WHERE IS THE BRIDGE

- **EMERGING NEEDS & APPLICATION AREAS**



Bayesian Networks (BNs)

Offers attractive statistical properties:

1. **un-sensitive to miss-specified PS logistic model (or any parametric model)**
2. **Estimators are genuine MLE asymp. unbiased & efficient**
3. **Potential to go beyond ATE estimation, e.g. Test**
4. **Potential to integrate BN approach with other approaches, e.g. calibration**

WHERE IS THE BRIDGE

- **EMERGING NEEDS & APPLICATION AREAS**

Bayesian Networks (BNs)

F.Mecatti, P.Vicard, F.Musella, L.Giammei
Significance Sept 2022
https://rss.onlinelibrary.wiley.com/doi/epdf/10.1111/1740-9713.01684

How BNs can help improving measurement, monitoring and prediction of gender equality

Thanks for your attention

Grazie

Дякую

谢谢你

Σας ευχαριστώ

Gracias

Tack

Obrigado

Danke

Merci

Спасибо