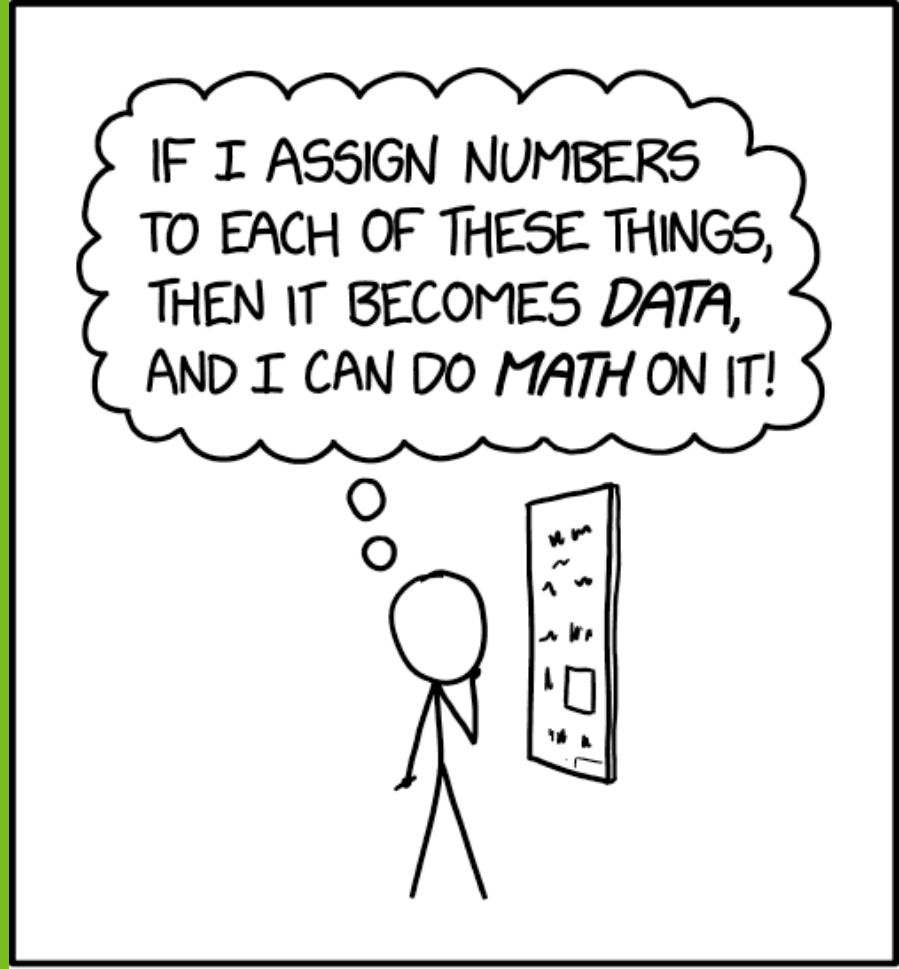
37C00100 Management Information Systems

Challenges with Big Data

Sampsa Suvivuo, Doctoral Researcher
Department of Information and Service Management

26 Mar 2024

Aalto University
School of Business



https://xkcd.com/2610/

Today's topics

General challenges with big data

BREAK

BREAK

Challenges with qualitative big data

What is data?

“Symbols that represent properties of objects and events and their environments” Ackoff (1989, p. 3).

“Simple observations about states of the world” Davenport and Prusak (1997. p 9).

“Numbers, characters or images that designate an attribute of a phenomenon” Royal Society (2012, p. 12).

General assumptions about data (Jones, 2019)

- **Referential**, refers to states of world, objects or entities that exist independently of data
- **Natural**, data has physical existence, and it exists “out there” independently of its use
- **Foundational**, data is the foundation of knowledge
- **Objective**, represents the world without interpretation
- **Equal**, all data can be analysed with same techniques

Davenport, T.H., Prusak, L., 1997. Information Ecology. Oxford University Press, Oxford.
Ackoff, R., 1989. From data to wisdom. J. Appl. Syst. Anal. 16, 3–9.
Royal Society, 2012. Science as an open enterprise. Available at: http://royalsociety.org/policy/projects/science-public-enterprise/report/.
Jones, M. (2019). What we talk about when we talk about (big) data. Journal of Strategic Information Systems 28 (2019) 3–16

Structured data	Semi-structured data	Unstructured data
Clearly defined and highly organized into rows and columns.	Some structural consistency and quantitative elements.	Unorganized data without a clear model, often in textual form. Readable to humans but challenging for programs to understand.
Example: data in relational databases	Example: CSV and JSON files	Example: emails, blogs, books

The amount of data in the world doubles every two years (IDC 2014)

80 – 90 % of this data is qualitative or unstructured (King 2022, Harbert 2022)

IDC. (2014). The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. Retrieved from https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.html
Timothy King, 80 Percent of Your Data Will Be Unstructured in Five Years, Data Management Solutions Review, March 28, 2019. Accessed May 6, 2022.
Tam Harbert, Tapping the power of unstructured data, MIT Sloan School of Management. February 1, 2021. Accessed May 6, 2022

r/Aalto • 10 days ago
soupwanda

Aalto admission process

Hi, applicant here, is there a statistic on how many people pass the preliminary assignments for the bachelor in art & design? do they get strictly selective from the very start?

4 0 Share

https://www.reddit.com/r/Aalto/comments/1bdlut9/aalto_admission_process/

[{'kind': 'Listing', 'data': {'after': None, 'dist': 1, 'modhash': '', 'geo_filter': '', 'children': [{'kind': 't3', 'data': {'approved_at_utc': None, 'subreddit': 'Aalto', 'selftext': 'Hi, applicant here, is there a statistic on how many people pass the preliminary assignments for the bachelor in art & design?\ndo they get strictly selective from the very start? \n', 'user_reports': [], 'saved': False, 'mod_reason_title': None, 'gilded': 0, 'clicked': False, 'title': 'Aalto admission process ', 'link_flair_richtext': [], 'subreddit_name_prefixed': 'r/Aalto', 'hidden': False, 'pwls': 6, 'link_flair_css_class': None, 'downs': 0, 'thumbnail_height': None, 'top_awarded_type': None, 'parent_whitelist_status': 'all_ads', 'hide_score': False, 'name': 't3_1bdlut9', 'quarantine': False, 'link_flair_text_color': 'dark', 'upvote_ratio': 0.81, 'author_flair_background_color': None, 'subreddit_type': 'public', 'ups': 3, 'total_awards_received': 0, 'media_embed': {}, 'thumbnail_width': None, 'author_flair_template_id': None, 'is_original_content': False, 'author_fullname': 't2_pngrnk9k', 'secure_media': None, 'is_reddit_media_domain': False, 'is_meta': False, 'category': None, 'secure_media_embed': {}, 'link_flair_text': None, 'can_mod_post': False, 'score': 4, 'approved_by': None, 'is_created_from_ads_ui': False, 'author_premium': False, 'thumbnail': 'self', 'edited': False, 'author_flair_css_class': None, 'author_flair_richtext': [], 'gildings': {}, 'content_categories': None, 'is_self': True, 'mod_note': None, 'created': 1710316298.0, 'link_flair_type': 'text', 'wls': 6, 'removed_by_category': None, 'banned_by': None, 'author_flair_type': 'text', 'domain': 'self.Aalto', 'allow_live_comments': False, 'selftext_html': '<!-- SC_OFF --><div class="md"><p>Hi, applicant here, is there a statistic on how many people pass the preliminary assignments for the bachelor in art & design?\ndo they get strictly selective from the very start? </p>\n</div><!-- SC_ON -->', 'likes': None, 'suggested_sort': None, 'banned_at_utc': None, 'view_count': None, 'archived': False, 'no_follow': False, 'is_crosspostable': False, 'pinned': False, 'over_18': False, 'all_awardings': [], 'awarders': [], 'media_only': False, 'can_gild': False, 'spoiler': False, 'locked': False, 'author_flair_text': None, 'treatment_tags': [], 'visited': False, 'removed_by': None, 'num_reports': None, 'distinguished': None, 'subreddit_id': 't5_31315', 'author_is_blocked': False, 'mod_reason_by': None, 'removal_reason': None, 'link_flair_background_color': '', 'id': '1bdlut9', 'is_robot_indexable': True, 'num_duplicates': 0, 'report_reasons': None, 'author': 'soupwanda', 'discussion_type': None, 'num_comments': 0, 'send_replies': True, 'media': None, 'contest_mode': False, 'author_patreon_flair': False, 'author_flair_text_color': None, 'permalink': '/r/Aalto/comments/1bdlut9/aalto_admission_process/', 'whitelist_status': 'all_ads', 'stickied': False, 'url': 'https://www.reddit.com/r/Aalto/comments/1bdlut9/aalto_admission_process/', 'subreddit_subscribers': 1859, 'created_utc': 1710316298.0, 'num_crossposts': 0, 'mod_reports': [], 'is_video': False}}], 'before': None}}, {'kind': 'Listing', 'data': {'after': None, 'dist': None, 'modhash': '', 'geo_filter': '', 'children': [], 'before': None}}]

Small data

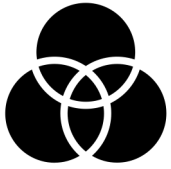
Deliberately produced by the researcher’s actions but at the same time, constrained in size, temporality and flexibility in their generation.

What is Big Data?

VOLUME

Data in enormous quantity

"How much?"



VARIETY

Data contains structured, semi-structured and unstructured data

"In what forms?"

VERACITY

Data is messy, erroneous and noisy

"How reliable?"

VELOCITY

Data is continuously created in real-time

"How fast?"

500 hours of videos uploaded to YouTube per minute

But also…

Value, visibility, variability, versatility, volatility, virtuosity, vitality, visionary, vigour, viability, vibrancy, virility, valueless, vampire-like, venomous, vulgar, violating, very violent, portentous, perverse, personal, productive, partial, practices, predictive, political, provocative, privacy, polyvalent, polymorphous, playful, exhaustivity, fine-grained, relationality, extensionality (Uprichard 2013, Lupton, 2015, Kitchin & McArdle 2016).

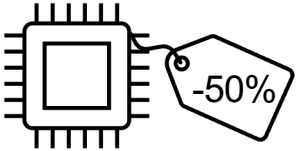
Not to mention the criteria for specialized hardware required for storing and manipulating data or computational difficulties in processing and analyzing the data.

Uprichard E (2013) *Big data, little questions*. Discover Society
Lupton D (2015) The thirteen Ps of big data. The Sociological Life,
Kitchin, R. & McArdle, G (2016). *What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets*. Big Data & Society, pp. 1–10

"Big Data is notable not because of its size, but because of its **relationality** to other data. Due to efforts to mine and aggregate data, Big Data is fundamentally **networked**. Its value comes from the **patterns** that can be derived by making **connections** between pieces of data, about an individual, about individuals in relation to others, about groups of people, or simply about the structure of information itself." (boyd & Crawford, 2011).

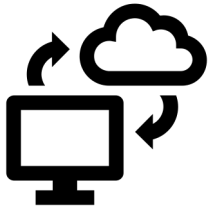
boyd, d., Crawford, K. (2011) Six Provocations for Big Data. Proceedings of Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society"

Drivers of Big Data

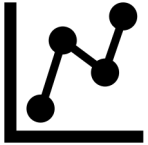


Cheap technology

Smart phones



Cloud computing



Datafication

Pervasive internet

Internet of Things

"…massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves." (Anderson 2008).

Anderson C (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired 16*(7).

Big data hubris

Big data hubris is the belief that that big data are a substitute for, rather than a complement, to traditional data collection and analysis.

Related assumptions:

- more data is always better as quantity equals quality
- big data captures “whole domains” at “full resolution”
- numbers speak for themselves free of human bias or framing
- domain-expertise is no longer needed as results can be interpreted by anyone able to do statistics or data visualization, “correlation trumps causation"
- thus, there is no need for preceding theory, models or hypotheses
- algorithms take the knowledge out of knowledge production
- patterns and relationships in big data are inherently meaningful and truthful

What is happening here?

http://www.simonweckert.com/googlemapshacks.html

Google Flu Trends 2008 - 2015

Posterchild of Big Data and data science success introduced in 2008

- Out of 50 million search terms during influenza seasons between 2003 and 2006, 45 were identified as key terms (fever, aching, flu…)
- The idea was to spot flu trends before Centers for Disease Control (CDC)
- Completely missed massive Swine Flu outbreak in 2009, predictions from 2011 to 2013 were almost completely wrong
- For 108 weeks observation period, the Flu Trends was correct just 8 times
- Tracking outside temperature would have done better job in predicting flu
- At best, 2 weeks ahead of CDC
- Reasons for failure: Most people googling symptoms have some other illness, overfitting the model to training data, e.g. high school basketball → good predictor of a flu

Exhaustivity, N = all?

The size of the big dataset creates the illusion it contains all the relevant information. There is still a distinction between big data and having “everything” or having captured reality.

Even a big dataset is a sample shaped by technology, polices, practices and sampling biases.

Quantity does not absolve from having to consider validity, generalizability, reliability and dependencies in data.

A dataset may have millions entries, but this does not mean it is random or representative. **Without considering the population or dataset’s characteristics, the size of the dataset is meaningless.** Big data has blind spots regarding people who do not register as digital signals (e.g., small children, elderly, populations in hard-to-reach places).

Jones, M (2019). *What we talk about when we talk about (big) data.* Journal of Strategic Information Systems 28 (2019) 3–16

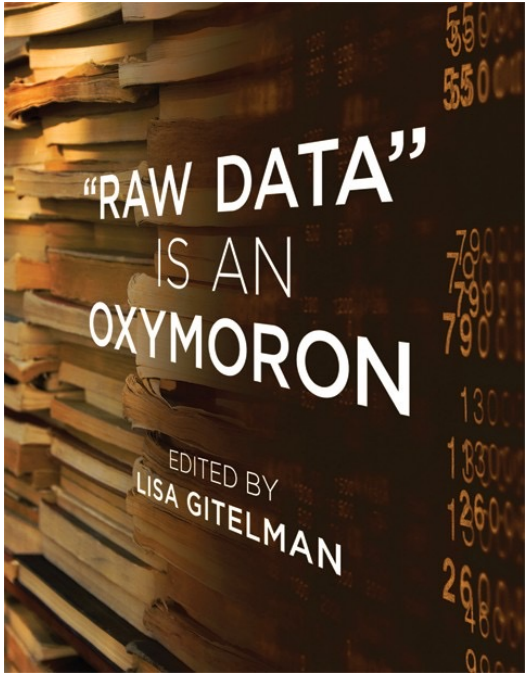
Raw data is an oxymoron

Before we have the results, we have made a myriad of decisions starting from what to look at, how to look at, in what context, what variables to include, how to quantify or operationalize them, what counts as an outlier etc.

Systems are designed to record, process, store and to distribute certain kinds of data → data is always created.

Data cleaning or data preparation, how do we address incomplete or lacking values, how is corrupted or incorrect data identified?

Interpretation begins immediately when a researcher begins to study results and to assign meaning to them → data cannot speak for themselves without human bias or framing.



Crawford, Miltner & Gray (2014). Critiquing Big Data: Politics, Ethics, Epistemology. International Journal of Communication 8 (2014), 1663–1672

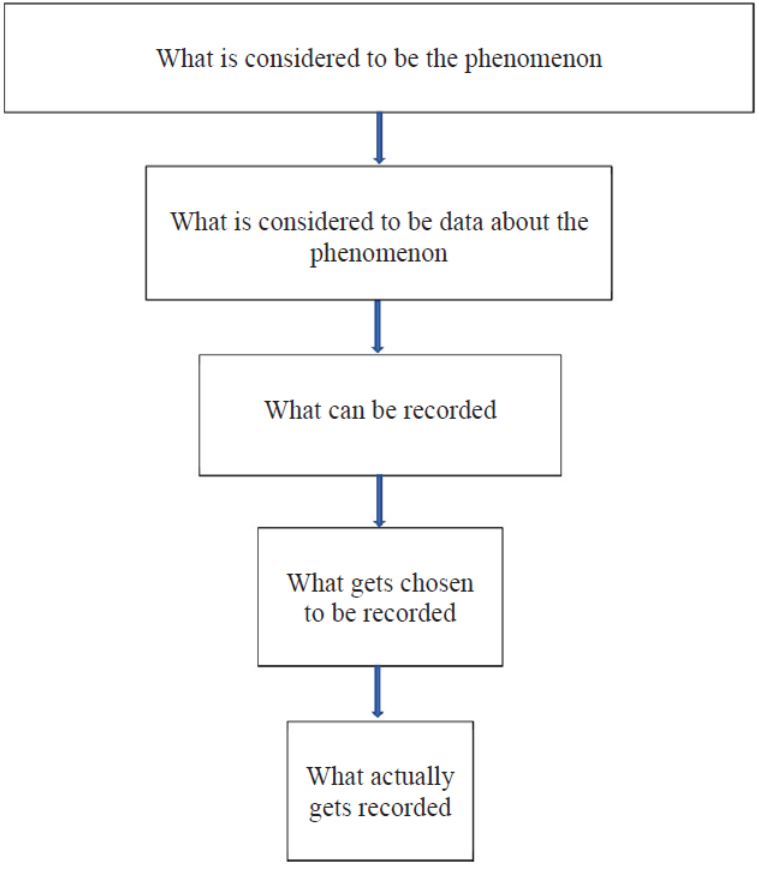


Fig. 2. How data come to be.

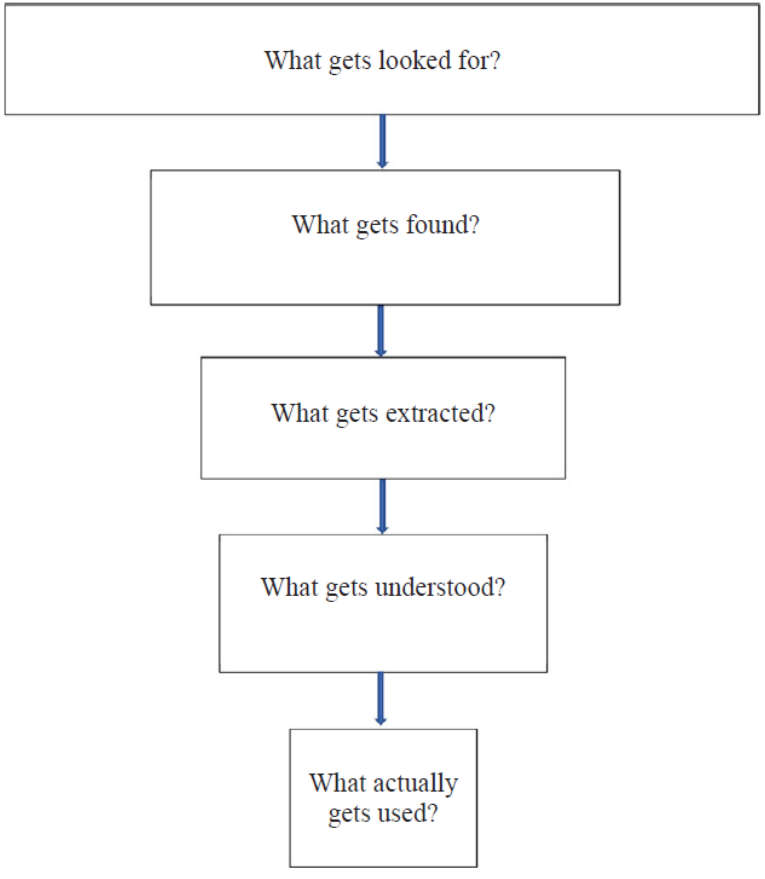


Fig. 3. How data come to be used.

Assumptions about data and big data and potential challenges to them.

Assumption	Description	Potential challenges to the assumption
Data are …		
Referential	Data report a reality that exists independent of themselves	Data may not always report reality accurately or completely
Natural	Data exist independent of their use	Data are constructed
Foundational	Date are the base on which our understanding of the world is built and themselves stand directly on that world	What is taken to be data reflects a particular way of understanding the world
Objective	Data represent the world, without interpretation	What is taken to be data involves subjective judgements
Equal	All data are equivalent	Data vary in their characteristics, provenance and quality
Big data are …		
Revolutionary	Big data are unprecedented and transformational	Big data are not necessarily new and their effects are not determined by their characteristics
Voluminous	Data are accumulating at large scale and at a high rate	The significance of data does not depend on their volume and velocity
Universal	Everything is, or will become, data	Not everything is effectively represented by data. Datafication is selective
Exhaustive	N = all	Data are a sample of the reality they describe. Data that are usable, and used, are a subset of data that are recorded

Jones, M (2019). *What we talk about when we talk about (big) data.* Journal of Strategic Information Systems 28 (2019) 3–16

Unable to answer “hows” or “whys”

Big data research has been criticized for

- focusing on “tactical” issues at the expense of “why” (Grover et al. 2020)
- generating superficial understanding of the phenomenon based on correlations (Hirschheim, 2021)
- producing ostensibly clear presentations hiding the underlying messiness (Walker et al. 2020).

However, this might be due to the fact that in general, big data is better suited to providing the “what”, the “where” and the “when” but not the “why” or the “how” because of its quantitative heritage.

Grover, V., Lindberg, A., Benbasat, I., & Lyytinen, K. (2020). The Perils and Promises of Big Data Research in Information Systems. *Journal of the Association for Information Systems, 21*, 1-26. doi:https://doi.org/10.17705/1jais.00601

Hirschheim, R. (2021). The attack on understanding: How big data and theory have led us astray: A comment on Gary Smith’s Data Mining Fool’s Gold. Journal of Information Technology, 36(2), 176-183. doi:10.1177/0268396220967677

Walker, M. A., & Boamah, E. F. (2020). The digital life of the #migrantcaravan: Contextualizing Twitter as a spatial technology. Big Data & Society, 7(2), 1-18. doi:10.1177/2053951720978485

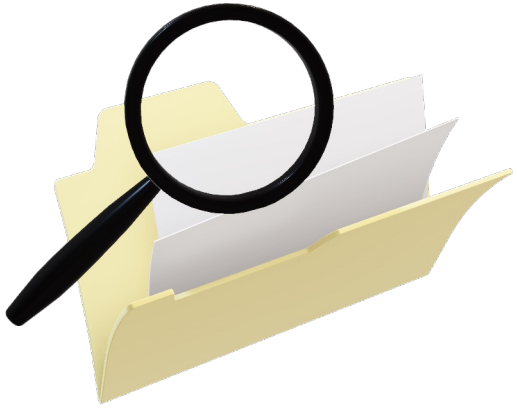
Spurious correlations

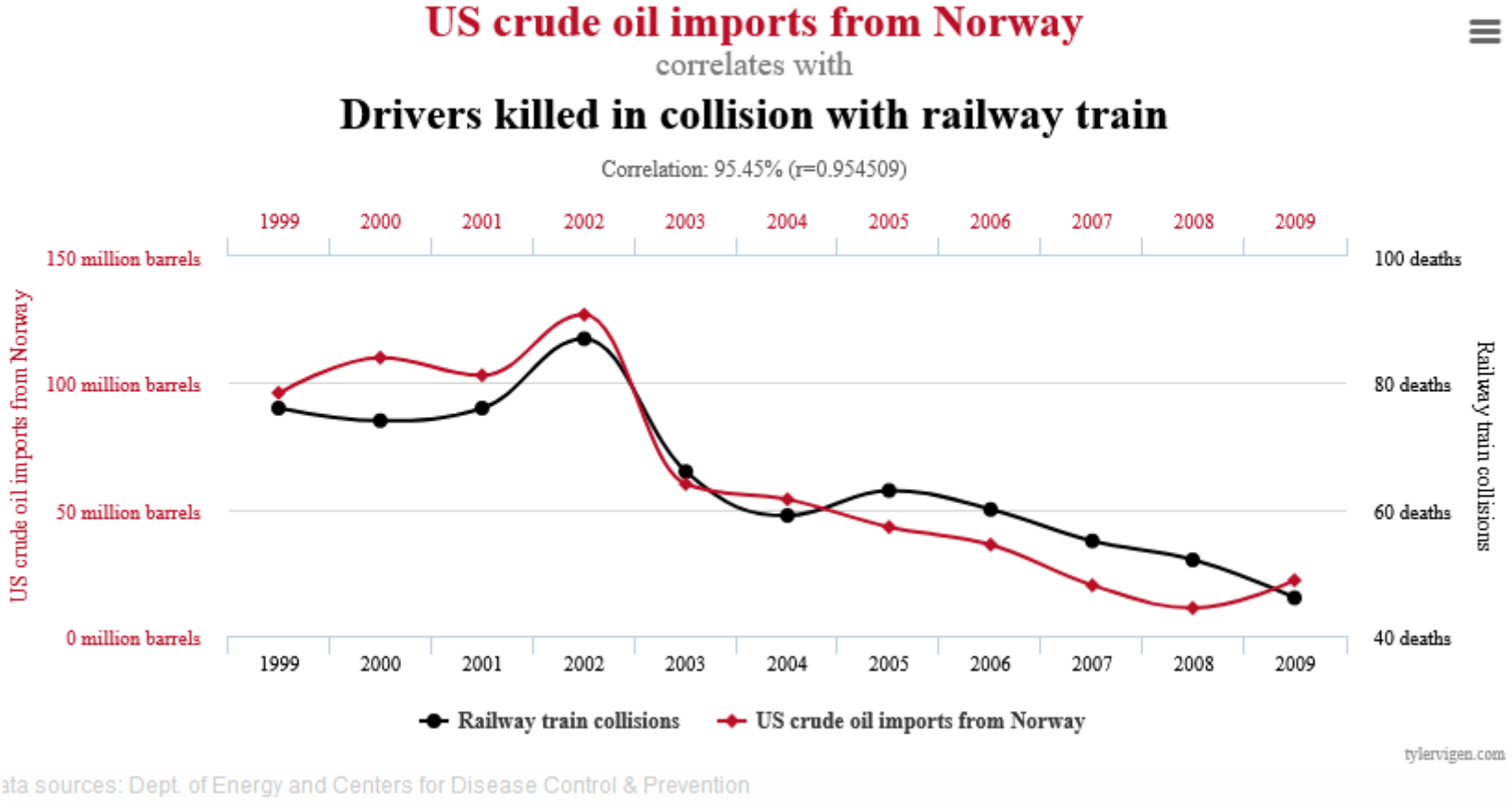
"Deflated p-value" issue: many statistical methods and techniques were developed for smaller samples and **with very large samples almost every relationship appears significant.**

Sizeable datasets have the issue of spurious correlation where variables are correlated but lack causality.

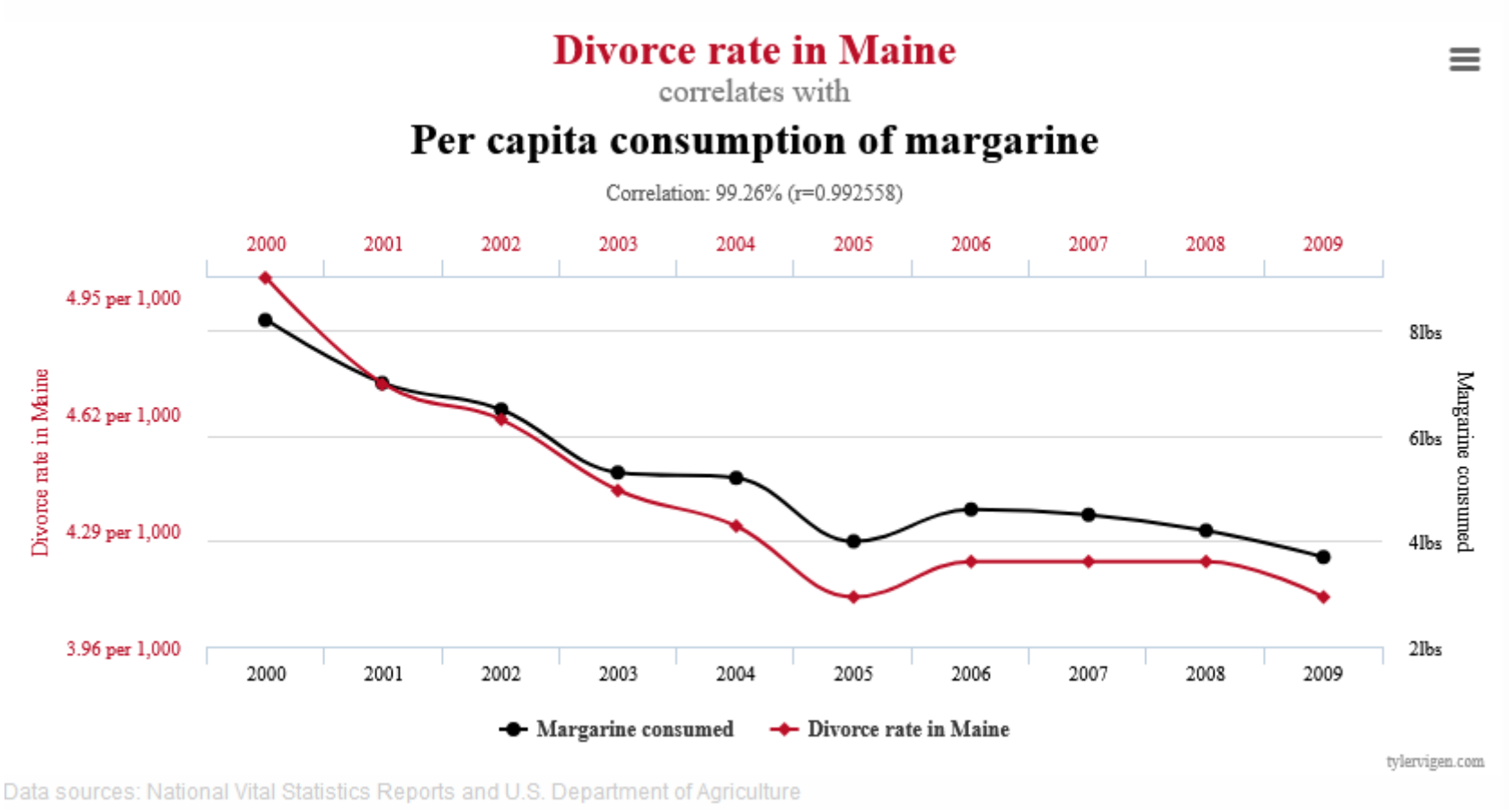
Meaningful relationships don't increase proportionally with the increase in data, you are just making the haystack bigger!

Related concept: Apophenia, seeing patterns and connections where none exists.

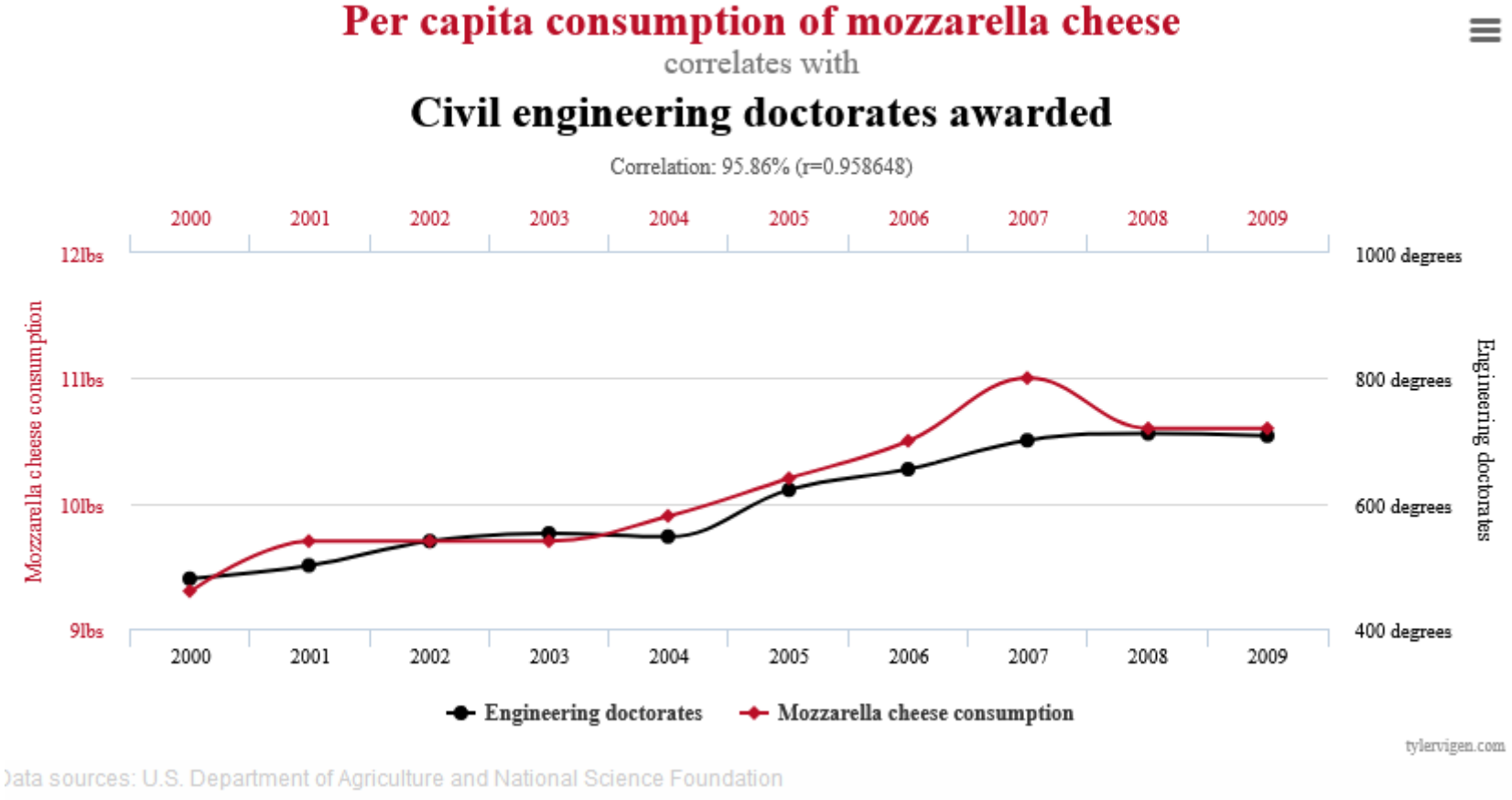




https://www.tylervigen.com/spurious-correlations



https://www.tylervigen.com/spurious-correlations



https://www.tylervigen.com/spurious-correlations

Effects of context loss

"If it [data] *is collected and processed at a great distance, it becomes both easier to mythologize and harder to verify",* (Dalton et al. 2016).

The distance the big data puts between the researcher and the participants, the *"view from 30 000 feet",* decontextualizes and sterilizes the data by focusing on behavior rather than its meaning (Fielding 2020).

Using a dataset collected by someone else might lead to a blurring of the context and circumstances the data was created in i.e., how the data came to be → being far-removed from the data creation makes it difficult to detect evident errors.

Willfully ignoring context results in reductionist analysis ignoring effects of culture, politics, governance etc.

Dalton, C. M., Taylor, L., & Thatcher, J. (2016). Critical Data Studies: A dialog on data and space. *Big Data and Society*, *3*(1), p 5. https://doi.org/10.1177/2053951716648346

Frey, W. R., Patton, D. U., Gaskell, M. B., & McGregor, K. A. (2020). Artificial Intelligence and Inclusion: Formerly Gang-Involved Youth as Domain Experts for Analyzing Unstructured Twitter Data. *Social Science Computer Review*, *38*(1), 42–56. https://doi.org/10.1177/0894439318788314

Fielding, N. G. (2020). Critical Qualitative Research and Impact in the Public Sphere. *Qualitative Inquiry*, *26*(2), 142–152. https://doi.org/10.1177/1077800419857746

Convenience sample

According to Lazer and Radford (2017) in social sciences big data is almost always a “convenience sample”

- Variables are selected for their traceability rather than based on the understanding of the phenomenon.
- Big data is collected not because it is the only viable way to study the phenomenon, but because it is easier, faster and cheaper to collect.

Dalton et al. (2016) talk about the danger of big data studies becoming a form of “land-rover research” where the researcher is reluctant to leave the comfort of one’s office to go to the field

Lazer, D., & Radford, J. (2017). Data ex Machina: Introduction to Big Data. *Annual Review of Sociology*, *43*(1), 19–39. https://doi.org/10.1146/annurev-soc-060116-053457

Dalton, C. M., Taylor, L., & Thatcher, J. (2016). Critical Data Studies: A dialog on data and space. *Big Data and Society*, *3*(1), p 5. https://doi.org/10.1177/2053951716648346

Critical Data Studies

As a counter to big data hype, a Critical Data Studies (CDS) stream of research was born in the early 2010s.

CDS study cultural, ethical, legal, social and critical challenges pertaining to big data and takes the stance that big data's connection to the world goes beyond realm of traditional data science (Iliadis & Russo 2016).

Data, along with related infrastructures, are informed by specific histories, biases, worldviews, ideologies, and philosophies of the people who created and used them, but these tend to remain hidden.

Data as a form of power, Big Data is not neutral phenomena, data is never raw but always "cooked."

Iliadis, A. & Russo, F. (2016). Critical data studies: An introduction. Big Data & Society July–December 2016: 1–7

Opportunistic behaviour

Researchers tend to work under the “Ideal User Assumption” expecting all the subject users to operate in good faith and not trying to manipulate the system or engage in opportunistic behaviour.

For example, a person might not be who they purport to be or a human at all, or they might create alternative accounts to fabricate support for themselves, writing false reviews etc.

The better tools we have to monitor open-source information, the more tempting it becomes to try to manipulate those signals for economic or political gain.

'Obama Is Dead' Tweet Makes For Flash Crash

Apr. 23, 2013 2:44 PM ET | SPY, USO, XLF | 32 Comments

Paulo Santos

Marketplace

Follow

Perhaps you're looking around for what made the market crash 1% in seconds. Look no further. What happened is that Fox's twitter feed was hacked. The hacker, using his command of Fox News tweet feed, then issued the following tweet:

So there you have it. Obama was assassinated with two shots. The market reaction was massive, one still wonders if it were machines doing the selling, or flesh and blood humans. Among other assets, there was an impact on S&P futures, Crude, USD/JPY, you name it. The moves were real and tradable. I reproduce a few of these moves below.

Socio-technical issues

Issues with accuracy and completeness:

- errors and failures in observing, differences in what is observed and what is recorded, deliberate or accidental changes to record, failure or corruption of data storage, malfunctioning sensors, loose cables, inconsistent time zone management, server outages, incomplete or inconsistent event logging

Change in instrumentation vs change in use

Technology and its use changes over time

How system should be used vs how it is used in practice

How system records and affects behaviour

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. Journal of the Association for Information Systems, 12(12), 767-797.
Jones, M (2019). *What we talk about when we talk about (big) data.* Journal of Strategic Information Systems 28 (2019) 3–16

Big data divides and data fumes

“Big Data Divide”, the data-haves and the data-have-nots, the data-rich and the data-poor

- Asymmetric relationship between those who collect, store, and mine large quantities of data, and those whom data collection targets.
- Also, significant investment to create commercial value compared to solving social and environmental challenges.
- Between companies, those who have more data have the advantage.
    - Example: email spam filters, Elisa vs Google

“Data fumes” describes a situation where corporations generating big data are selective in who they give access to leading to researchers and others working with fumes.

Ethics and privacy

Key ethical questions: informed consent, anonymization, minimization of harm and searchability.

Framing Big Data as an “asset”, “ability” or “technique” downplays ethical aspects of data.

Some big data studies are built upon “dataveillance”, “uberveillance”, profiling and lack of privacy.

Big data researchers often make the claim that since the data is accessible and public, ethical considerations don’t apply. Thinking being that publicly available data cannot cause harm to an individual → being in public vs being public.

There are also tensions between academic, commercial and regulatory practices and contradictions in ethical behavior and academic rigor. For example, Twitter allows the tweets to be used for other purposes but forbids any manipulation of the messages i.e., anonymization or paraphrasing creating tension with academic practices.

Duality of Big Data

Big data has potential for both good and bad. For example:

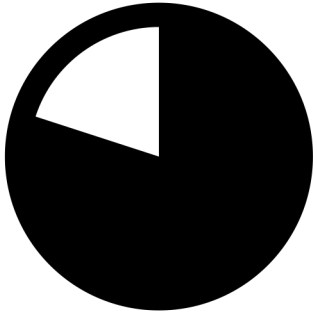
- We can improve healthcare, or we can discriminate in healthcare and related insurance
- We can improve store layout and offer better recommendation engines, or we can engage in price discrimination and targeted adds
- Researchers like big data for its unobtrusiveness and candour as interviews and surveys are artificial situations. However, it is also surveillance without subjects being aware of it

Big data absorbs the biases and prejudices in the data meaning that it will enforce those biases and prejudices.

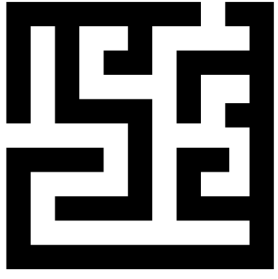
To recap

- Raw data is an oxymoron
- Reductionism
- Convenience sample
- Without context, data loses meaning
- Distance to data makes it hard to verify and easy to mystify
- Without domain-expertise, interpretations become anemic
- Sampling bias, representativeness
- Cannot reach marginal or offline phenomenon
- Spurious correlations
- Ill-suited for “why” and “how” questions
- Grey area between those who create the data and those represented by it
- Mostly generated by corporations, “data fumes”
- Dataveillance
- Contradicting criteria for ethical behavior and academic integrity
- Ethical issues with informed consent, collecting, sharing combining, access to data, searchability, and privacy

Challenges With Qualitative Big Data



80 % of data is estimated to be qualitative



Complicated analysis

Qualitative methods requiring manual persual struggle with the width and breath of big data

Methodological review

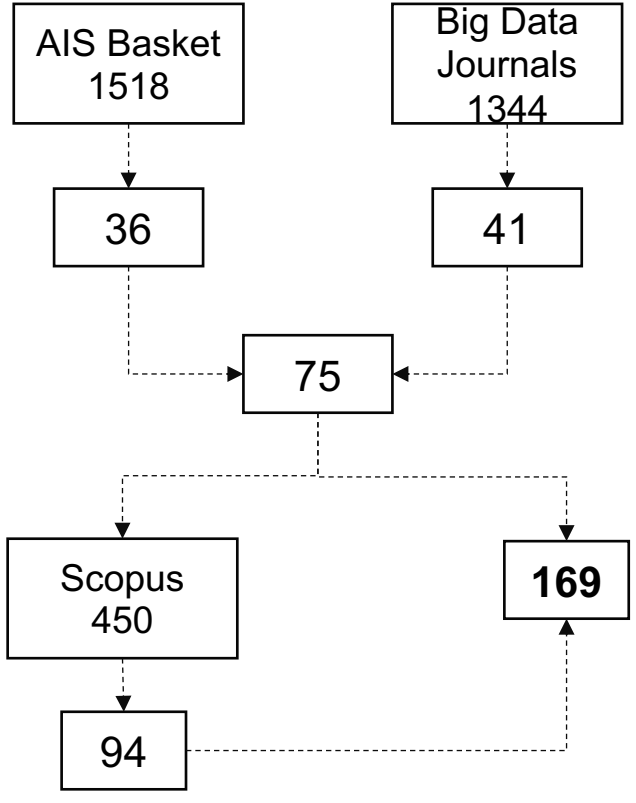
MOSTLY PEER-REVIEWED

No gap-spotting, the aim is to describe and synthesize challenges encountered by researchers with qualitative big data

Journal paper focus

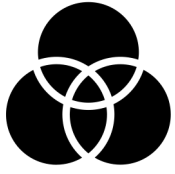
Criteria:

- Empirical
- Qualitative big data
- 30 000 observations or more



VOLUME

Locating relevant data

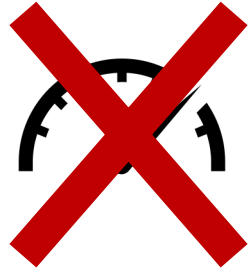


VARIETY

Preserving richness in data

VERACITY

Addressing noise in data



VELOCITY

Types of machine learning*

ML: type of an artificial intelligence that tries to learn and make predictions from the data. For example, speech and image recognition, automated stock trading and recommendation engines

Supervised

- Data is split into training and test data
- The algorithm is given examples of what to look for
- Used for example, classification and regression
- Examples: Support Vector Machines, Naïve Bayes, Decision trees

Unsupervised

- No training data
- The algorithm is applied directly to data and tries to learn on its own
- Used for example, clustering, dimensionality reduction
- Examples: Latent Dirichlet Allocation, Latent Semantic Analysis

*Also, reinforced and deep learning

Good discussion on different modes of machine learning
https://www.ibm.com/cloud/learn/machine-learning

Locating relevant data

For 10 gigabytes of relevant data there could be *“non-trivial amount of irrelevant data”* (Yue et al. 2019)

Dictionary/lexicon approach

- General purpose dictionaries
- Tailored dictionaries and topic modelling

Concentrated approach

- Concentrating search around events or external data

Yue, W. T., Wang, Q.-H., & Hui, K.-L. (2019). See No Evil, Hear No Evil? Dissecting the Impact of Online Hacker Forums. *MIS Quarterly, 43*(1), 73-95. doi:https://doi.org/10.25300/MISQ/2019/13042

General dictionaries

General purpose dictionaries capture certain emotions or have been tailored for a certain purpose, but they are not specifically tailored for the researcher's data. For example:

- Financial sentiment dictionary with 2 329 negative and 297 positive words
- A list of abusive words
- Domain-specific dictionaries e.g., medical

Example: Word-Emotion Association (a.k.a. NRC Emotion Lexicon)

- Eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust)
- two sentiments (negative and positive)
- 14 000 words

https://nrc.canada.ca/en/research-development/products-services/technical-advisory-services/sentiment-emotion-lexicons

Tailored dictionaries

General-purpose dictionaries might perform badly in capturing nuances/characteristics, or a suitable dictionary might simply be unavailable.

Noting that researchers defining words associated with disease outbreaks might miss colloquial and informal terms, a probabilistic Naïve Bayes classifier was used to create a tailored dictionary based on Yelp reviews. This way, words such as “pungency”, “barely edible” and “wiping nose” that authors might not have realized to include in the dictionary and which are unlikely to be included in any other preconceived dictionary were included. (Mejia et al. 2019).

Building a lexicon of 326 complaint n-grams and 354 compliment n-grams out of 2 000 tweets (Gunarathne et al. 2018).

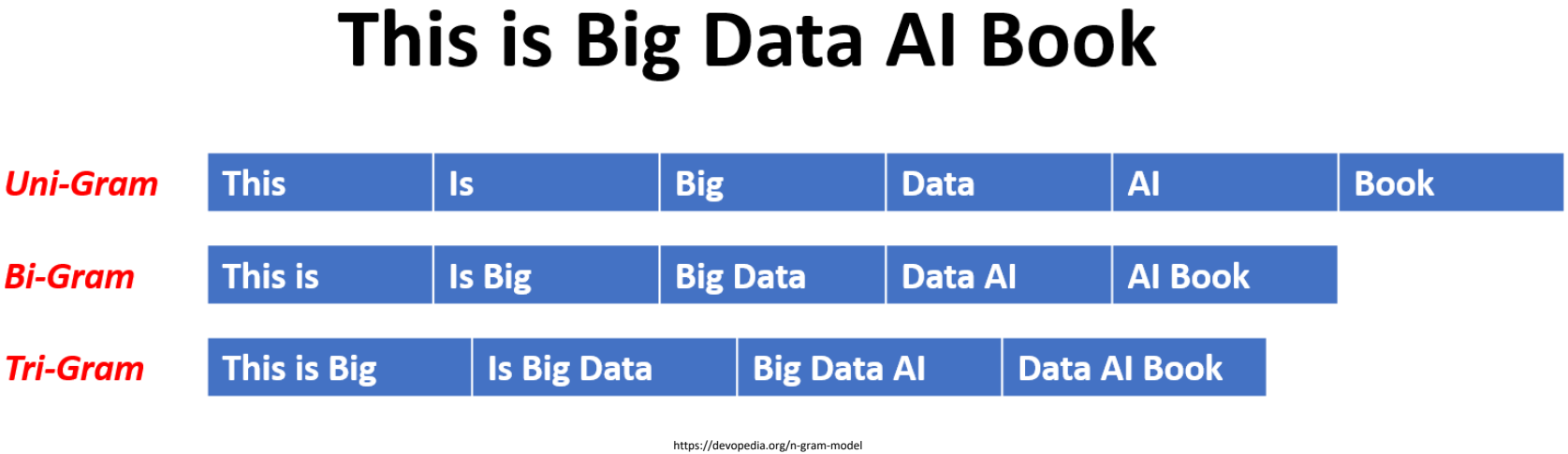
Manually coding 3 086 forum replies to create a training data and using a support vector machine to code the remaining data. (Chen et al. 2019).

Gunarathne, P., Rui, H., & Seidmann, A. (2018). When Social Media Delivers Customer Service: Differential Customer Treatment in the Airline Industry. MIS Quarterly 42(2), 489-520. doi:https://doi.org/10.25300/MISQ/2018/14290

Mejia, J., Mankad, S., & Gopal, A. (2019). A for Effort? Using the Crowd to Identify Moral Hazard in New York City Restaurant Hygiene Inspections. Information Systems Research 30(4), 1363-1386 doi:https://doi.org/10.1287/isre.2019.0866

Chen L., Baird, A., & Straub, D. (2019). Fostering Participant Health Knowledge and Attitudes: An Econometric Study of a Chronic Disease-Focused Online Health Community. *Journal of Management Information Systems, 36*(1), 194-229. doi:https://doi.org/10.1080/07421222.2018.1550547

N-grams



https://devopedia.org/n-gram-model

Unigrams
social: 134
data: 118
media: 88
information: 87
our: 81
analysis: 81
online: 79
study: 74
have: 67
learning: 63
twitter: 63
sentiment: 58
model: 55
their: 55
using: 53
content: 53
customer: 49
more: 49
tweets: 46
which: 44

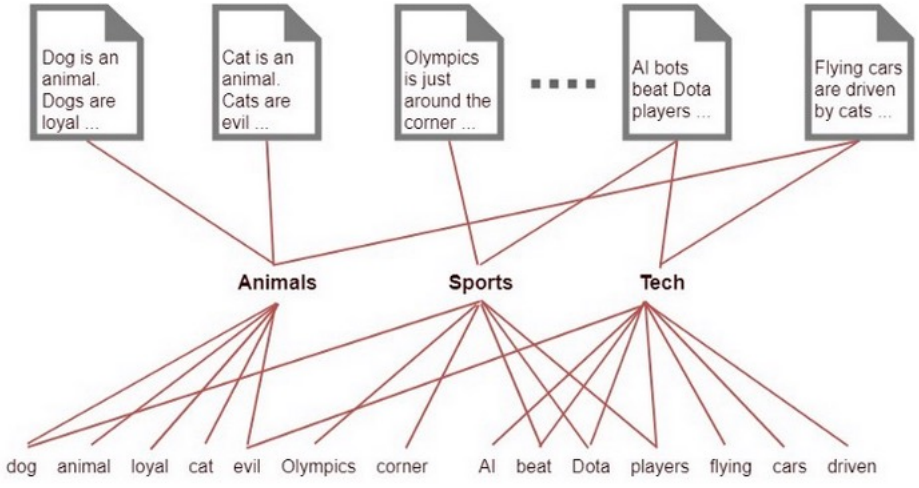
Bigrams
social media: 83
machine learning: 30
sentiment analysis: 28
big data: 26
topic modeling: 16
deep learning: 13
online reviews: 12
neural network: 12
truth discovery: 10
customer care: 10
text mining: 9
convolutional neural: 9
have been: 9
keyword ambiguity: 9
risk factors: 8
but also: 8
facial images: 8
our study: 8
has been: 8
relationship between: 8

Trigrams
convolutional neural network: 7
natural language processing: 7
digital customer care: 6
triggers risk factors: 5
use social media: 5
social media data: 5
social media platforms: 5
safety management systems: 4
long short-term memory: 4
social network analysis: 4
little known about: 4
intra-organizational blogging platforms: 4
customer support community: 4
firm equity value: 4
focal human brand: 4
dynamic topic modeling: 4
2013 presidential election: 4
semantic soft factors: 4
amazon ec2 cloud: 4
identification risk factors: 3

Topic modelling

An unsupervised machine learning approach to finding clusters of associated words and phrases i.e. topics in documents

Two common “bag-of-words” models: Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA).



LDA creates topics based on words’ probabilistic co-occurrence. Ignores word order and word’s grammatical role.
LSA creates a term-document matrix which is then reduced in size leaving only the most important terms.
Captures polysemy partially but ignores word order.

LDA

Documents' probabilities of belonging to a topic

```
[[0.0173246  0.0173246  0.01732459 ... 0.01732459 0.0173246  0.75745563]
 [0.01732128 0.01732125 0.01732127 ... 0.01732125 0.01732125 0.33057636]
 [0.02325244 0.02325244 0.02325244 ... 0.02325244 0.02325244 0.43793031]
 ...
```

Generated topics

```
['upwork', 'portfolio', 'clients', 'time', 'client', 'tracker', 'app', 'jobs', 'use', 'outside']
['upwork', 'proposals', 'proposal', 'talent', 'rising', 'badge', 'approved', 'client', 'tips', 'job']
['upwork', 'client', 'data', 'starting', 'freelancer', 'hire', 'report', 'started', 'good', 'scammers']
['upwork', 'client', 'support', 'thoughts', 'tos', 'marketing', 'contact', 'invitations', 'customer', 'job']
['user', 'deleted', 'wrong', 'newbie', 'upwork', 'charge', 'unresponsive', 'red', 'client', 'dispute']
['feedback', 'client', 'upwork', 'give', 'potential', 'profile', 'banned', 'clients', 'account', 'long']
['help', 'need', 'advice', 'please', 'assistant', 'upwork', 'profile', 'rate', 'verification', 'services']
['scam', 'upwork', 'review', 'account', 'suspended', 'new', 'profile', 'client', 'contract', 'close']
['upwork', 'boosting', 'job', 'finding', 'update', 'client', 'work', 'type', 'difficult', 'fiverr']
['first', 'job', 'upwork', 'get', 'legit', 'client', 'got', 'beginner', 'taxes', 'land']
['payment', 'upwork', 'client', 'refund', 'score', 'work', 'success', 'start', 'take', 'job']
['connects', 'cover', 'letter', 'upwork', 'client', 'private', 'tax', 'job', 'writing', 'lol']
['top', 'rated', 'get', 'upwork', 'getting', 'plus', 'invites', 'jss', 'clients', 'views']
['client', 'price', 'fixed', 'upwork', 'project', 'hourly', 'contract', 'job', 'problem', 'allowed']
['question', 'upwork', 'contract', 'contracts', 'end', 'test', 'right', 'fees', 'direct', 'hey']
```

Topic modelling examples

When key terms do not occur simultaneously, arranging terms under topics can help to find relevant documents. Rather than trying to come up with keywords to locate texts on taste or mouthfeel of coffee, these can be turned into topics (“finishing”, “mouthfeel”, “aftertaste”), allowing the algorithm to identify relevant texts. (Ngoc et al. 2019).

To identify tweets related to inflation among 11.1 million tweets, Angelico et al. (2022) applied LDA to create 50 topics from which they chose two inflation-related topics with the top 10 words (e.g., “inflation”, “wages”, “deflation”, “euro”, and “price”) from the associated tweets for further examination. Then, they created N-grams from the 1,534,743 remaining tweets with words such as “price,” “expensive,” and “inflation” and manually coded whether the N-gram referred to increasing or decreasing inflation.

Ngoc, T. N. T., Thu, H. N. T., & Nguyen, V. A. (2019). Mining aspects of customer’s review on the social network. Journal of Big Data, 6(22), 1-21.

Angelico, C., Marcucci, J., Miccoli, M., & Quarta, F. (2022). Can we measure inflation expectations using twitter? Journal of Econometrics, 228(2), 259-277.

Top2Vec

248 topics found

```
[1560 1106  897  748  714  605  560  491  470  4
  351  338  320  317  316  315  308  305  301  2
  274  266  253  253  240  235  232  231  230  2
  210  206  206  206  205  203  203  203  198  1
  184  183  183  180  179  175  174  173  170  1
  155  153  151  151  151  151  147  146  145  1
  142  141  141  140  140  138  137  136  135  1
  122  120  120  120  119  119  118  117  113  1
  107  107  106  106  105  104  104  103  103  1
   95   95   95   94   94   94   93   93   91
   88   87   87   86   85   85   85   85   85
   81   81   80   78   78   77   77   76   76
   72   72   72   71   69   69   69   68   68
   64   61   61   60   60   60   59   59   59
   56   55   55   55   55   55   55   54   54
   52   52   52   52   52   51   49   49   49
   47   47   46   45   45   44   43   43   40
   36   35   35   35   34   33   28   23   23
```

Topics closest to a given keyword “alternative” by topic score
[200 201 56 139 190]
[0.67116353 0.4964827 0.39442092 0.29924601 0.2

Topic: 200
Word: ['alternative' 'upwork' 'suggestions' 'opt
 'working' 'jobs' 'opportunity' 'work' 'better'
 'trying' 'else' 'method' 'other' 'hiring' 'move
 'proposals' 'changing' 'future' 'change' 'hires
 'freelancers' 'tips' 'try' 'offers' 'plus' 'car
 'recommend' 'best' 'changes' 'freelance' 'tried
 'agency' 'freelancer' 'talent' 'fix' 'again’ ‘a

Topic: 190
Word: ['question' 'questions' 'asking' 'answer'
 'answers' 'issue' 'issues' 'explain' 'asked' 'h
 'survey' 'about' 'response' 'point' 'advice' 'r
 'regarding' 'request' 'understand' 'know' 'conf
 'whata' 'probably' 'id' 'interesting' 'seems' '
 'try' 'lol' 'ok' 'call' 'interview' 'check' 'hi
 'why' 'interviews' 'trouble' 'think']

Concentrated search

Search around events

Rather than trying to peruse and absorb the complete data, McKenna (2019) centered the search around times when game patches were introduced.

Connect to external data

A study on the online-hacker forum discussions' impact on the extent of distributed denial of service attacks had to locate relevant content among 2 960 893 posts in 355 222 threads. Relevant posts were found by looking for posts with mentions of the port numbers listed in threat databases (Yue et al. 2019).

- Port numbers in computers are used to differentiate transactions over a network (web service, mail service, file transfer etc.) and in this case, could be used to see what service or vulnerability the threat was targeting helping the authors to zero in on relevant posts.

McKenna, B. (2019). Creating convivial affordances: A study of virtual world social movements. *Information Systems Journal, 30*(1), 185-214. doi:https://doi.org/10.1111/isj.12256

Yue, W. T., Wang, Q.-H., & Hui, K.-L. (2019). See No Evil, Hear No Evil? Dissecting the Impact of Online Hacker Forums. *MIS Quarterly, 43*(1), 73-95. doi:https://doi.org/10.25300/MISQ/2019/13042

Locating relevant data

Challenge	Solution
Manually perusing and absorbing voluminous data is unfeasible.	Manually coding a small sample and using it to train a classifier. Crowdsourcing can increase and/or verify the annotated sample. Using a readymade or a tailored dictionary or a topic model.
Relevant search terms do not occur simultaneously in documents.	Arranging terms under topics and using the topics to find relevant documents.
Identifying posts of interest among millions of posts.	Connecting data to other events or sources.
Consumers' search intents vary while the used keywords might be the same	Focusing on the search goal and using a dictionary relevant to it

Addressing noise (kohina)

We can distinguish between data in principle and data in practice (Jones 2019).

Relevant data might contain incorrect, misclassified and sometimes fraudulent data or some other form of uncertainty.

Together the irrelevant data and data of uncertain veracity obscure relevant data and hamper algorithms' performance by making them learn from unrelated or faulty data.

Humans can address many of the issues almost subconsciously, but a machine must be explicitly told how to address each type of noise.

Jones, M (2019). *What we talk about when we talk about (big) data*. Journal of Strategic Information Systems 28 (2019) 3–16

Irrelevant data

Words correctly related to the topic in principle could in practice be used in unrelated contexts.

For example:

- Study on asthma risk factors collecting tweets like *“I will call you asthma, because you take my breath away”.*
- Tweets on police being metaphorically blind, when studying rumours about a police officer being blinded by a firecracker during 2017 G20 summit demonstrations.
- Trying to identify sales and promotion of wildlife products on Twitter. Talk of banning the sales of ivory rather than selling it, discussing wildlife preservation and related law and policy or even mentions of animal cartoon characters contain relevant keywords but are false positives.
- “Apple orchard” vs Apple (company)

Out-of-vocabulary words

Informal language, abbreviations, misspellings, punctuation errors, non-dictionary slang, wordplay, comparative sentences, negations, transferred negations, double negations, sarcasm, unwanted languages, spam and emoticons all constitute noise for algorithms if not properly addressed.

Noise like this is often addressed in pre-processing the data by removing messages' features, such as misspellings, slang or emoticons, for the algorithm's sake.

At the same time, some of the data's richness is lost. For example, uppercase and extended words (e.g. "FUNNYYY", "looool") alongside emoticons and URLs can be used to decipher sentiment and intent instead of just being removed as noise.

- Verifying words against a dictionary of 1 million distinct words. The words not found in the dictionary are either corrected to comply with grammar, ignored or subjected to further analysis for their sentiment (Cury 2019)

Cury, R. M. (2019). Oscillation of tweet sentiments in the election of João Doria Jr . for Mayor. *Journal of Big Data*, 1-15. doi:https://doi.org/10.1186/s40537-019-0208-1

Several discussions in a thread

Several different conversations might be taking place within a single discussion thread.

This could mean that some recorded sentiments are directed to something a specific commentator said rather than the nominal topic or original post, for example.

Cases like these can be difficult for algorithms to identify and address

- Having to address multiple conversations within discussion forum threads, Abbasi et al. (2019) converted 5 million posts into 26 million sentences with more focus and consistency.
- A similar approach was adopted in Zhang et al. (2017) where documents were analysed at a paragraph level to better identify relevant parts.
- Moving between sentence level and document or domain level features can help in taking into account the effect of context.

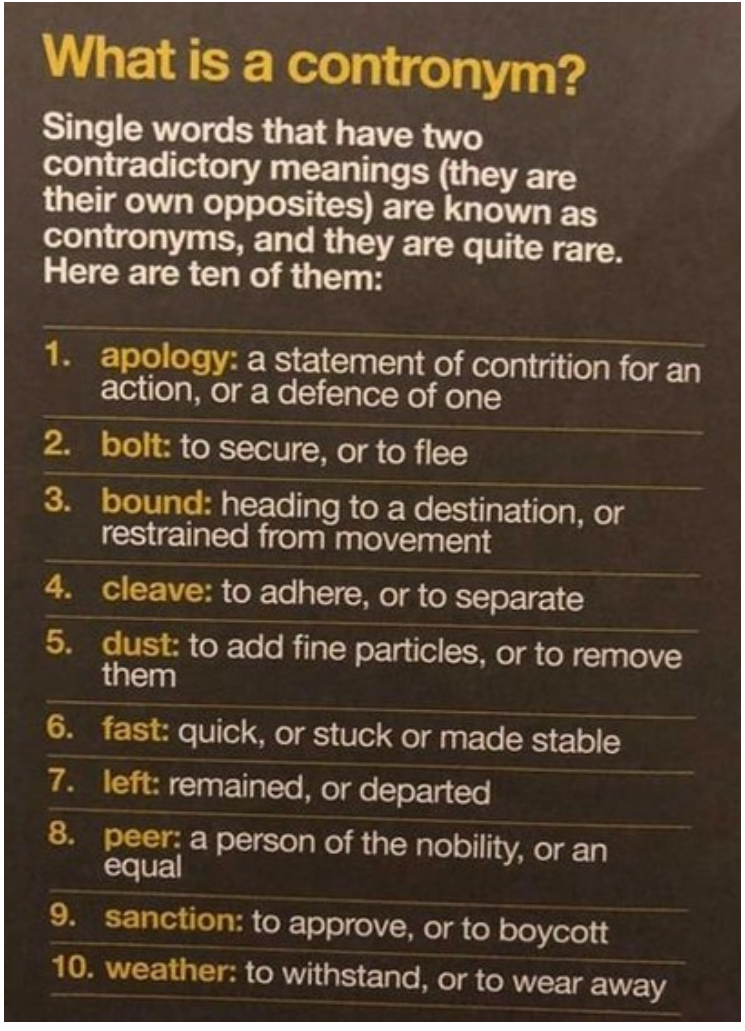
Abbasi, A., Li, J., Adjeroh, D., Abate, M., & Zheng, W. (2019). Don't Mention It? Analyzing User-Generated Content Signals for Early Adverse Event Warnings. *Information Systems Research, 30*(3), 1007-1028. doi:https://doi.org/10.1287/isre.2019.0847

Zhang, J.-B., Sun, Y.-X., & Zhan, D.-C. (2017). Multiple-instance learning for text categorization based on semantic representation. *Big Data & Information Analytics, 2*(1), 69–75. doi:https://doi.org/10.3934/bdia.2017009

Context's effect

Polysemy: words might have different meanings depending on the context.

- "This toy is not bad at all, my two-year old plays with it all the time."
- "I'd kill for a cookie right now."
- In Arabic, popular names "Saeed" and "Amal" also mean "happy" and "hope", respectively, and their use depends on the context.



Accuracy-coverage trade-off

If the accuracy is very good, the dataset might not have adequate coverage.

If there is too much variety in the data i.e., it covers too much, accuracy might suffer.

For their study of locating relevant data from all English-speaking forums indexed by Google's discussion forum search Geva, et al. (2017) opted to use just brand names for cars without any additional refinements when balancing between accuracy and coverage. "Chevrolet Malibu" or "Chevrolet Spark" are in principle more detailed queries than "Chevrolet", but introduce noise to the sample in the form of irrelevant results such as the city of Malibu in California.

In training of a facial recognition algorithm, images' backgrounds add noise. They do not contribute to facial recognition but add unnecessary details the algorithm does not know to ignore, decreasing the algorithm's accuracy. To cancel this effect, backgrounds can simply be cropped out.

Geva, T., Oestreicher-Singer, G., Efron, N., & Shimshoni, Y. (2017). USING FORUM AND SEARCH DATA FOR SALES PREDICTION OF HIGH-INVOLVEMENT PROJECTS. *MIS Quarterly*, *41*(1), 65-82.

Addressing noise

Challenge	Solution
Informal language, abbreviations, misspellings, punctuation errors, non-dictionary slang, wordplay, emoticons, URLs.	Natural language processing, machine learning, domain adaptation, data pre-processing. Using a dictionary or the subword feature to check for out-of-vocabulary words.
There could be several discussions within one thread or document.	Breaking posts into sentences and documents into paragraphs.
Issues relating to a given phenomenon may be presented in various forms as it is highly likely that different terms are used to refer to the same topic.	Constructing a dictionary directly from the text with Naïve Bayes classifiers and applying SVD to reduce the number of keywords.
Depending on the context, the words' sentiment may change.	Combining sentence-level features with domain sensitive features.

Preserving richness

Hallmark of qualitative data is its rich descriptions enabling answering “why” and “how” questions.

Quantification of qualitative data reduces this richness for example by:

- disregarding grammar and word order
- ignoring certain forms of language that could help to determine the sentiment and intent of the author
- hiding messiness of data under seemingly clean charts
- lack of context

“Fixing bias in democracy by giving more votes to those who pay more taxes”

https://www.kansalaisaloite.fi/fi/aloite/432

An algorithm has hard time deciphering that the premise of the petition is the issue behind low support.

Popularity vs representability

Many systems have mechanisms to elevate popular content meaning that the result is not representative of the whole content but a picture of what is popular → categorizing content according to its similarities and then sampling it ensuring that less popular topics are also picked for reading. (Guo et al. 2017).

Supervised machine learning algorithms, are limited by the requirement of predefined topics (=training data), which is why interesting topics might remain hidden if too small portion of data is coded or coding is too coarse.

Unsupervised approaches have similar issue. For example, a topic could emerge when an algorithm is told to create eight topics but otherwise the topic would not emerge → triangulating the number of topics, newer algorithms.

Guo, X., Wei, Q., Chen, G., Zhang, J., & Qiao, D. (2017). Extracting representative information on intra-organizational blogging platforms. MIS Quarterly, 41(4), 1105-1127.

Analysing sentiment

Sentiment analysis is used to classify text into positive, neutral and negative or some other predetermined emotions such as trust, surprise, joy and disgust.

For example, Liu et al. (2020) studied crowdsourcing communities for open innovation and analyzed 43 550 product ideas submitted to electronics manufacturer's new product development community by categorizing the feedback valences into positive and negative.

However, approaches like this do not enable studying on a scale what the users were feeling positive or negative towards to, decreasing the richness in the data.

- This can be alleviated (to a degree) by using aspect-based sentiment analysis to connect the sentiment to a particular aspect. For example, beyond classifying tweets about an airline as positive and negative, tweets could connect these sentiments to staff, luggage, comfort, airport, punctuality, cabin crew behaviour, food quality and loss of baggage making for more fine-grained data.

"This earbud has good sound quality but poor battery life"

Liu, Q., Du, Q., Hong, Y., Fan, W., & Wu, S. (2020). User idea implementation in open innovation communities: Evidence from a new product development crowdsourcing community. *Information Systems Journal, 30*(5), 899-927. doi:10.1111/isj.12286

Ho, S. Y., Choi, K. W. S., & Yang, F. F. (2019). Harnessing aspect-based sentiment analysis: How are tweets associated with forecast accuracy? *Journal of the Association for Information Systems 20*(8), 1174-1209.

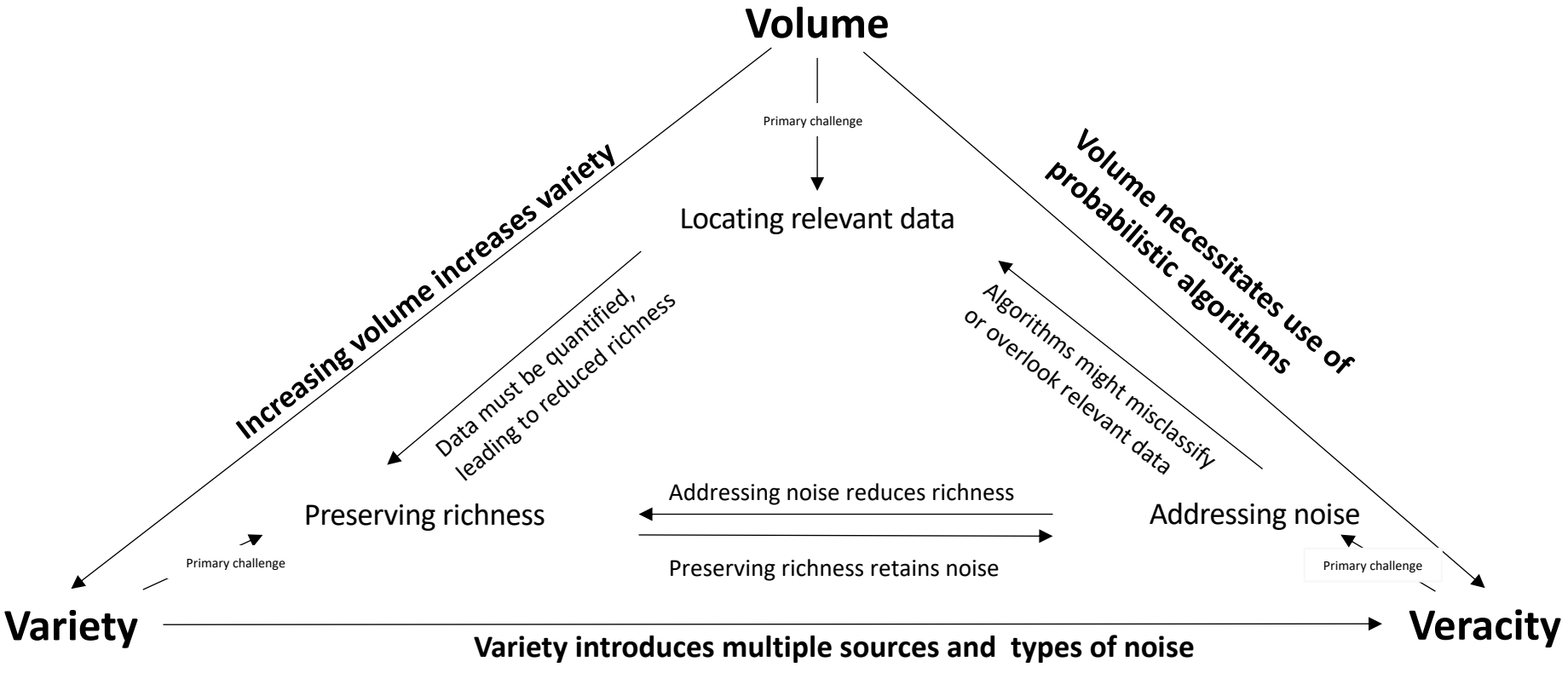
Kumar, S., & Zymbler, M. (2019). A machine learning approach to analyze customer satisfaction from airline tweets. Journal of Big Data, 6(62), 1-16.

Oh, Y. K., & Yi, J. (2021). Asymmetric effect of feature level sentiment on product rating: An application of bigram natural language processing (nlp) analysis. Internet Research, 32(3), 1023-1040.

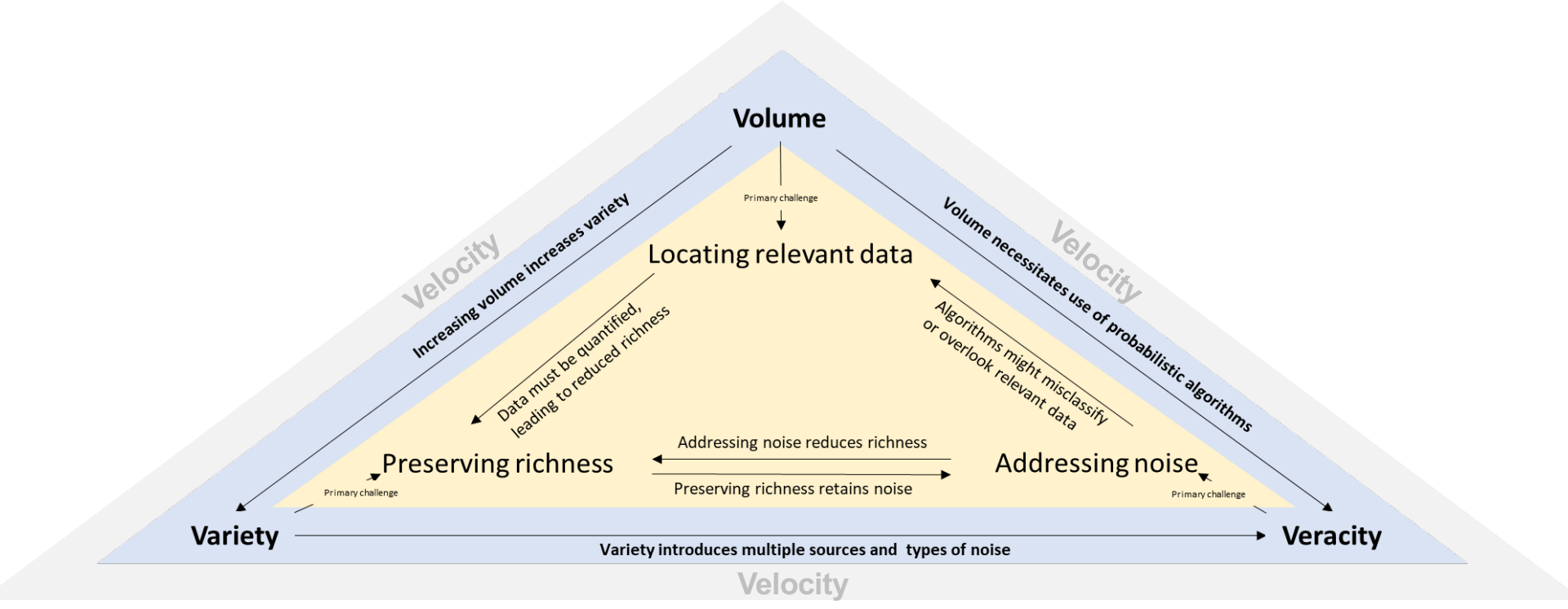
Preserving richness

Challenges	Solution
A voluminous qualitative corpus must be turned into a quantitative data	Use crowdsourcing and mixed-method approach as a stopgap solution before the development of new theories and techniques.
When content is classified (e.g. into sentiments) some of the “rich description” is inevitably lost	Looking beyond the usual tools, generating dictionaries from the data to supplement readymade dictionaries, using aspect-based sentiment analysis or establishing association rules between sentiments and issues.
Extracting a representative sample, not what is popular	Categorizing content based on their similarities between each other and then taking a sample.

Intertwined challenges



Intertwined challenges

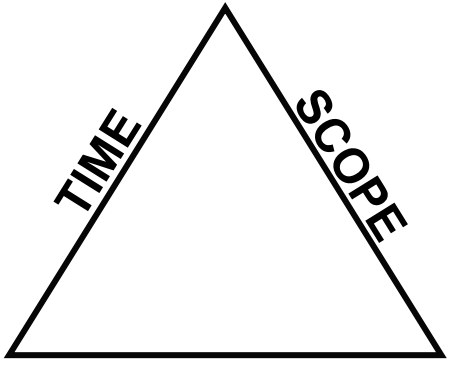


Intertwined challenges

The more voluminous the dataset, the more difficult it becomes to form an understanding of the data or to ensure that the data is relevant and properly coded as these tasks are given primarily for a probabilistic algorithm.

As volume increases, the variety of data types, how data is structured as well as the different motives and writing styles, also increase. This variety represents the richness qualitative data is known for. However, the more there is variety in data, the more there are various types and sources of noise.

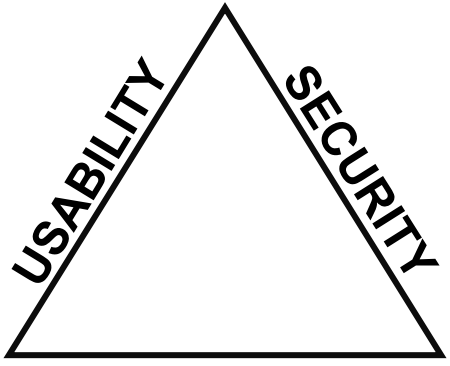
Low veracity of (noisy) data means that significant treatment of data is needed reducing either variety, volume or both.



BUDGET

Project management triangle

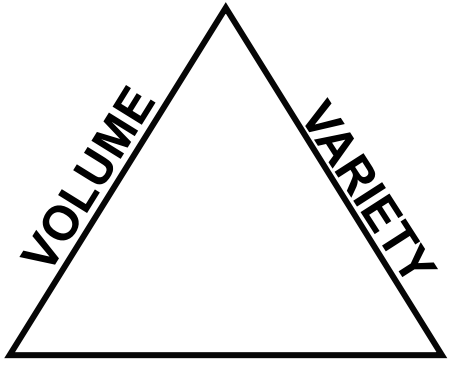
"Good, fast, cheap. Pick two."



PERFORMANCE

Secure systems design

“Useful and connected system cannot be completely secure.”



VERACITY

Qualitative big data analysis

”Quantity, reliability, richness. Which will be prioritized?”

Findings 1/2

While preserving richness might be the biggest challenges regarding qualitative big data, most of the identified solutions targeted locating relevant data or addressing noise in the data.

It is up to researchers to decide how to best manage the systemic effect of the three challenges and how much are they willing or able to give of one to gain more of another. The way a researcher solves a particular challenges might create additional or increased problems in other challenges but a challenges may also be addressed in a manner where less must be compensated for elsewhere.

Examples

Instead of removing noisy features, sometimes they can be accommodated: recognize and recode, shortening words with more than two consecutive same letters, breaking out-of-vocabulary words into subwords retaining more of the data's richness.

Choice of algorithms: aspect-based sentiment analysis vs regular sentiment analysis would preserve more of the information enabling richer analysis.

Considering context's effect: examining sentence level and document or domain level features together. For example, if a sentence appears to contain a negative expression in otherwise positive document, is it possible, that in the given context, the expression is not negative?

Findings 2/2

Often the solutions combine both human and machine pattern recognition zooming on a subset of the data deemed relevant in one way or another, or by scaling-up a manually coded subset of data with the help of machine learning.

However, in the reviewed articles human intelligence is often involved only at the beginning preparing the data and at the end interpreting results meaning that the rich and nuanced understanding underlining qualitative data is lost.

To counter this, extant research recommends adopting mixed-methods approach combining qualitative small data studies with big data studies while iterating between small and big data parts of the study.

Thank you

sampsa.suvivuo@aalto.fi

Aalto University
School of Business