Working With Data

Stephen Downes

December 4, 2021

Combining Data

- Data seldom stands alone – data is linked to other data, and can reveal more than was intended (e.g. criminal caught by sister's DNA) (Cohen, et.al., 2018)
- Data on travel may be used for discovering things far beyond mere travel patterns; data on purchases are not solely relevant to purchases; and so on. (Hand, 2019)

ID	zipcode
588	2009
654	2037
527	2000
955	2037
954	2010

zipcode	income
2000	75000
2005	82000
2009	96000
2010	88000
2037	125000

ID	zipcode	income
588	2009	96000
654	2037	125000
527	2000	75000
955	2037	125000
954	2010	88000

https://www.displayr.com/what-is-data-merging/

Unpacking Implications

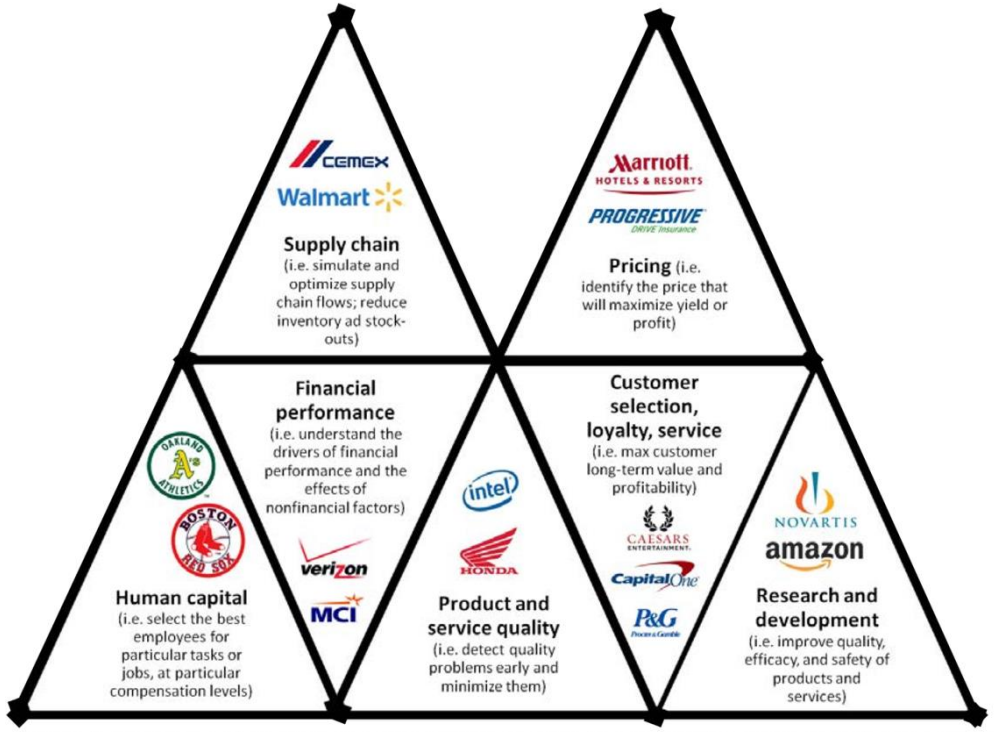


Figure 2. *Examples of analytical competitors' analytical capabilities. Source: author's elaborati*

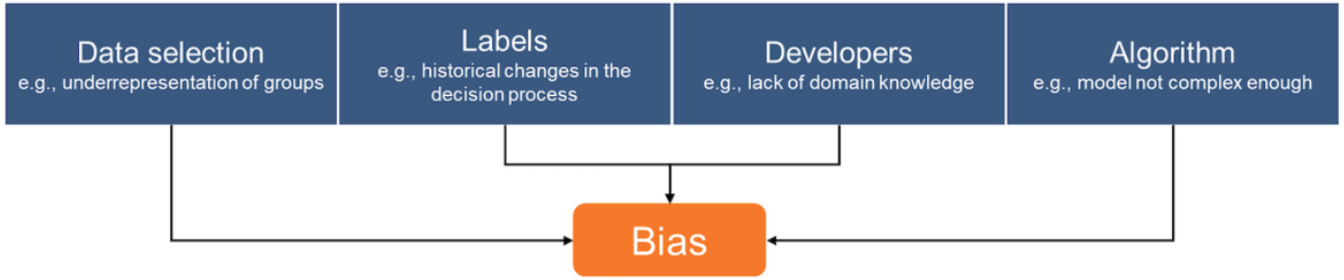
CDS unpacks "the work these data assemblages do in the context of dataveillance, the attrition and loss of privacy, the impact of profiling and social sorting, and the downstream use of data, to mention but a few."

Prinsloo 2019 https://epress.lib.uts.edu.au/journals/index.php/JLA/article/view/6589/7307

https://www.semanticscholar.org/paper/When-big-data-meets-dataveillance%3A-the-hidden-side-Esposti/11aed548023909f4efd4c9b6d7d27d0cf70ae8af

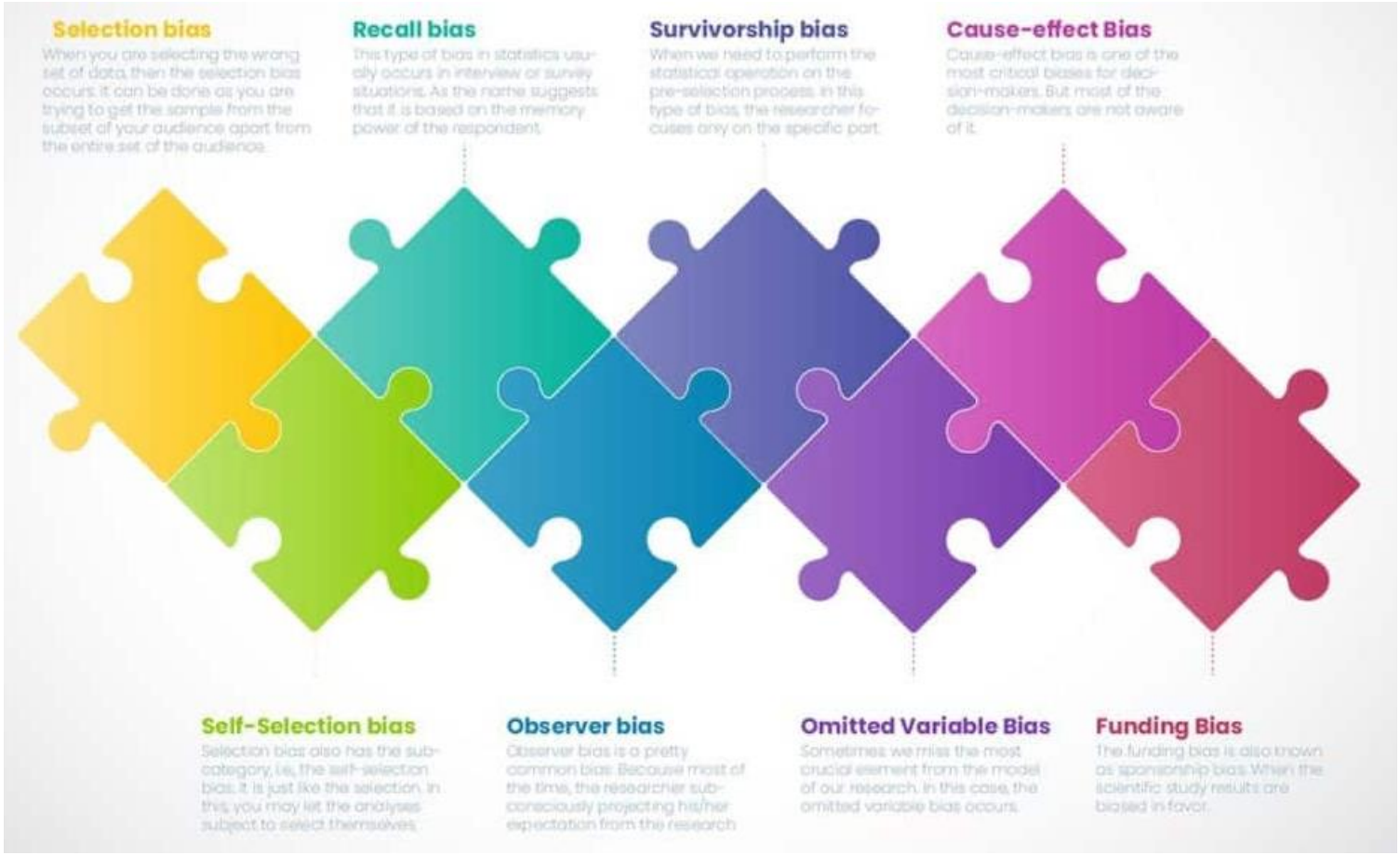
Bias (Causes)

- An incomplete or skewed training dataset
- Labels used for training: training data is labeled in order to teach the model how to behave, and humans create these labels…
- Features and modeling techniques - the measurements used as inputs for machine-learning models, or the actual model training itself



Josh Feast. (2019). 4 Ways to Address Gender Bias in AI. Harvard Business Review Online. November 20, 2019. https://hbr.org/2019/11/4-ways-to-address-gender-bias-in-ai

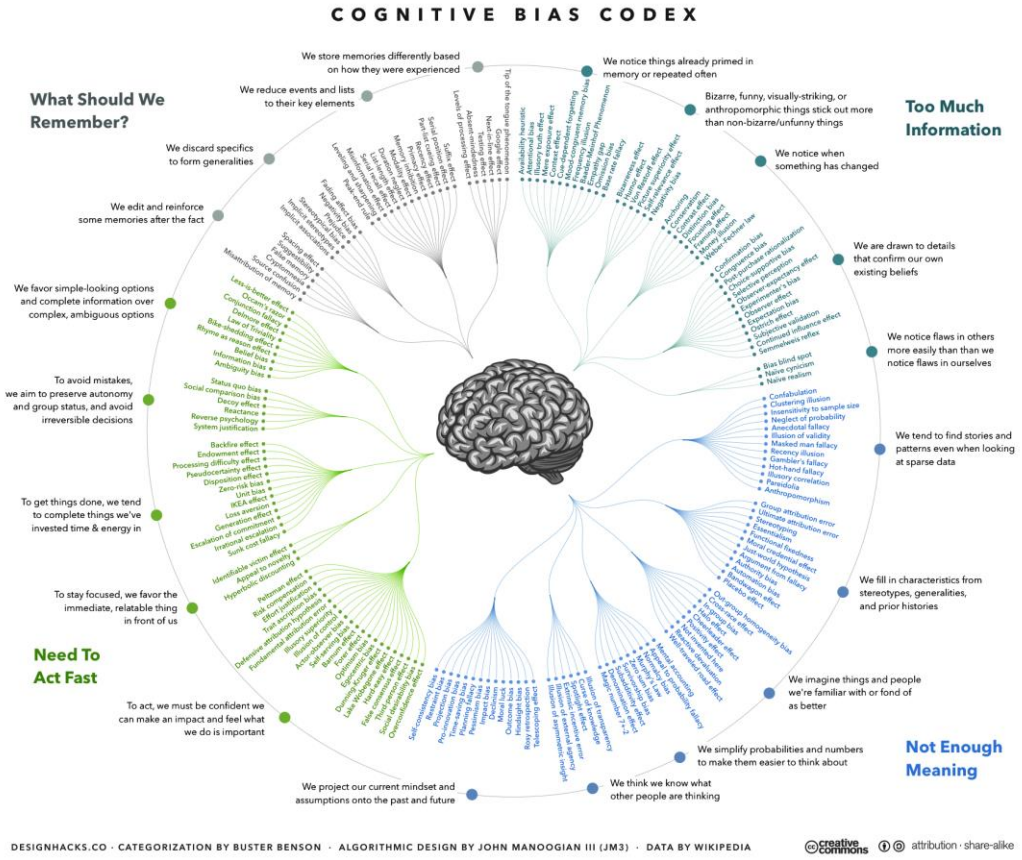
Bias (Types)



https://statanalytica.com/blog/bias-in-statistics/

Cognitive Bias Codex

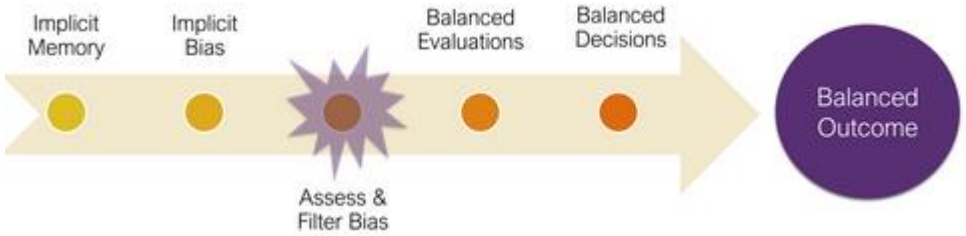
https://greatminds.consulting/insight/cognitive-bias-a-mental-shortcut-that-causes-a-mistake-in-reasoning



https://greatminds.consulting/wp-content/uploads/2019/08/The_Cognitive_Bias_Codex_-_180_biases_designed_by_John_Manoogian_III_jm3.png

Bias (Remedies)

- Ensure diversity in the training samples
- Ensure that humans labeling come from diverse backgrounds.
- Measure accuracy levels separately for different demographic categories to identify when one category is being treated unfavorably.
- Solve for unfairness by collecting more training data associated with sensitive groups.



Josh Feast. (2019). 4 Ways to Address Gender Bias in AI. Harvard Business Review Online. November 20, 2019. https://hbr.org/2019/11/4-ways-to-address-gender-bias-in-ai

Data and Objectivity

- Data should never be accepted on faith
- There is no such thing as context-free data; data cannot manifest the kind of perfect objectivity that is sometimes imagined.
- Lisa Gitelman and Virginia Jackson point out in the introduction to the aptly titled “Raw Data Is an Oxymoron“ (2013)

All from Radan, 2019: 16

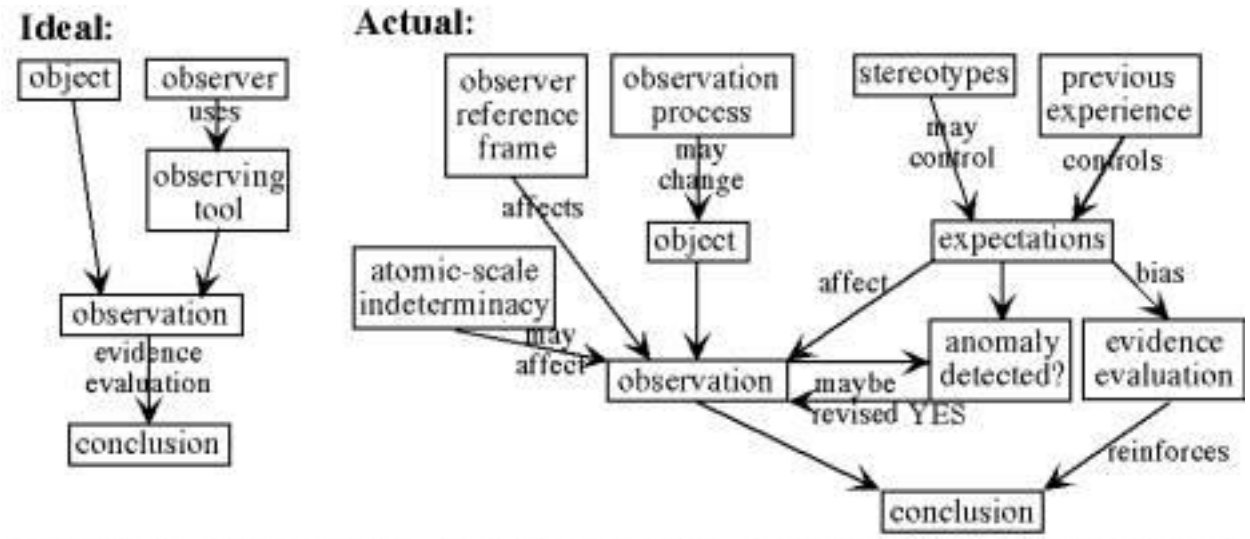


Figure 22. Flowcharts of ideal and actual interactions concerning experimental observations.

Eugenia Lee https://slideplayer.com/slide/14962871/

https://emotionalcompetency.com/sci/sm6.htm

Data Risks

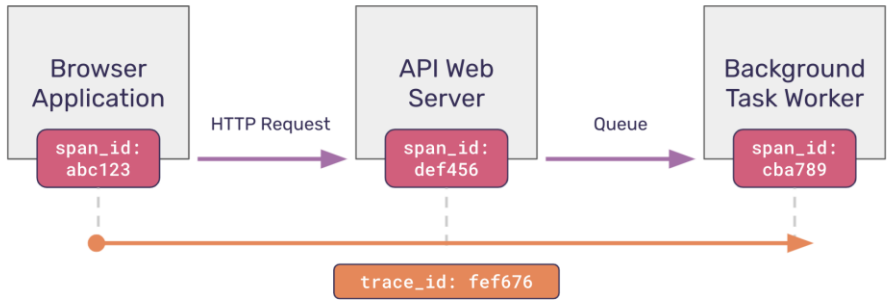
- The danger of stale and outdated data (Cohen, et.al., 2014)
- Sensitive files subject to regulations like GDPR, HIPAA and PCI (Varonis, 2019)
- Lack of space, time, and social context limitation on scope of data (Hand, 2018)
- Use for unexpected purposes and to reveal unexpected information
- Risk of exceptional intrusiveness since
- Potential for misuse, privacy breach, blackmail, and other crimes.
- Ghost users and accounts

We need to be clear about who benefits and who incurs the risk

https://www.varonis.com/blog/data-risk-report-highlights-2019/

Data Tracing

- Using markers or signs to track the flow of data through a network
- “We call this new verification method ‘radioactive’ data because it is analogous to the use of radioactive markers in medicine.”
- Examples: watermarks, embedded metadata, hash addressing



Alexandre Sablayrolles, Matthijs Douze, Hervé Jégou. (2020). Using ‘radioactive data’ to detect if a data set was used for training. Facebook Artificial Intelligence. February 05, 2020. https://ai.facebook.com/blog/using-radioactive-data-to-detect-if-a-data-set-was-used-for-training/

Data Ownership

- Data ownership refers to both the possession of and responsibility for information. Ownership implies power as well as control.
  - Who owns the data?
  - Who owns the information the data is about?

https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/dotopic.html

Loshin (2002) identifies a list of parties laying a potential claim to data:

- Creator – The party that creates or generate data
- Consumer – The party that uses the data owns the data
- Compiler - This is the entity that selects and compiles information from different information sources
- Enterprise - All data that enters the enterprise or is created within the enterprise is completely owned by the enterprise
- Funder - the user that commissions the data creation claims ownership
- Decoder - In environments where information is “locked” inside particular encoded formats, the party that can unlock the information becomes an owner of that information
- Packager - the party that collects information for a particular use and adds value through formatting the information for a particular market or set of consumers
- Reader as owner - the value of any data that can be read is subsumed by the reader and, therefore, the reader gains value through adding that information to an information repository
- Subject as owner - the subject of the data claims ownership of that data, mostly in reaction to another party claiming ownership of the same data
- Purchaser/Licenser as Owner – the individual or organization that buys or licenses data may stake a claim to ownership