



The Dirty Side of Data

Diael Thomas
Assessment Specialist, Institutional Effectiveness



Data-Based Decision-Making: Guiding Principle 4

The standards reflect our commitment to data-based decision-making.

Institutions must analyze a range of data, including disaggregated data, to ensure students are appropriately served and institutional mission and goals are met. Institutions should rely upon the data required by the

Commission and additional data used by the institution. Institutions should follow the Commission's evidence expectations that are reflective of a range of data considerations, consonant with higher education expectations, and consistent with the institution's mission. Periodic and systematic evaluation and assessment allow institutions to demonstrate commitment to reflection, and our standards provide the opportunity to evaluate progress toward institutional goals.

Institutions can leverage periodic assessment through each standard, using assessment results for continuous improvement and innovation to ensure levels of quality for constituents.

Background



Hometown:
Houston, TX



B.S. Chemical Engineering



Masters Library and
Information Science



Outline

- What is Data?
- Data as Selection
- Datafication
- Algorithmic Bias
- Towards the Future



What is Data?

“Data means information, more specifically facts, figures, measurements and amounts that we gather for analysis or reference. The term’s meaning also includes descriptive information about things, plants, animals, and people. We collect and store data typically through observation.”

-Market Business News

“Every data set involving people implies subjects and objects, those who collect and those who make up the collected. It is imperative to remember that on both sides we have human beings.”

- Mimi Onuoha (The Point of Collection)

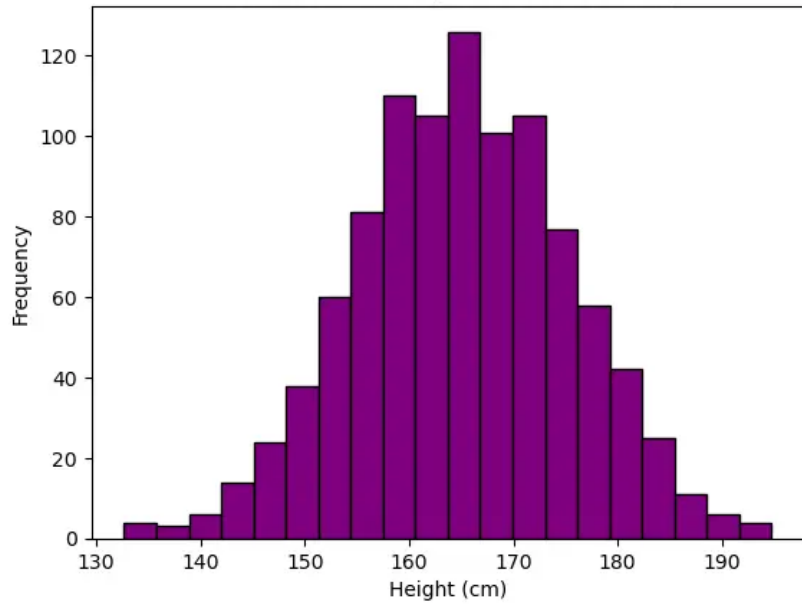
“factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation”

- Merriam-Webster

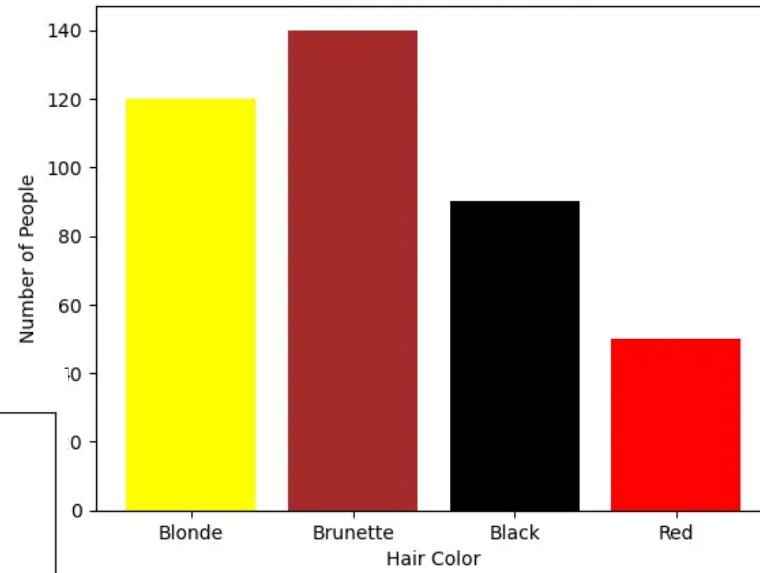
“information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer”

-Cambridge

Distribution of Heights in a Community



Distribution of Hair Colors



Philadelphia, Pennsylvania
Updated a few minutes ago

82 °F / °C 82° / 69°
Precipitation: 30%
Wind: 8 MPH
Humidity: 36%

Mostly sunny · Tue 27, 5:00 PM

1 AM 4 AM 7 AM 10 AM 1 PM 4 PM 7 PM 10 PM

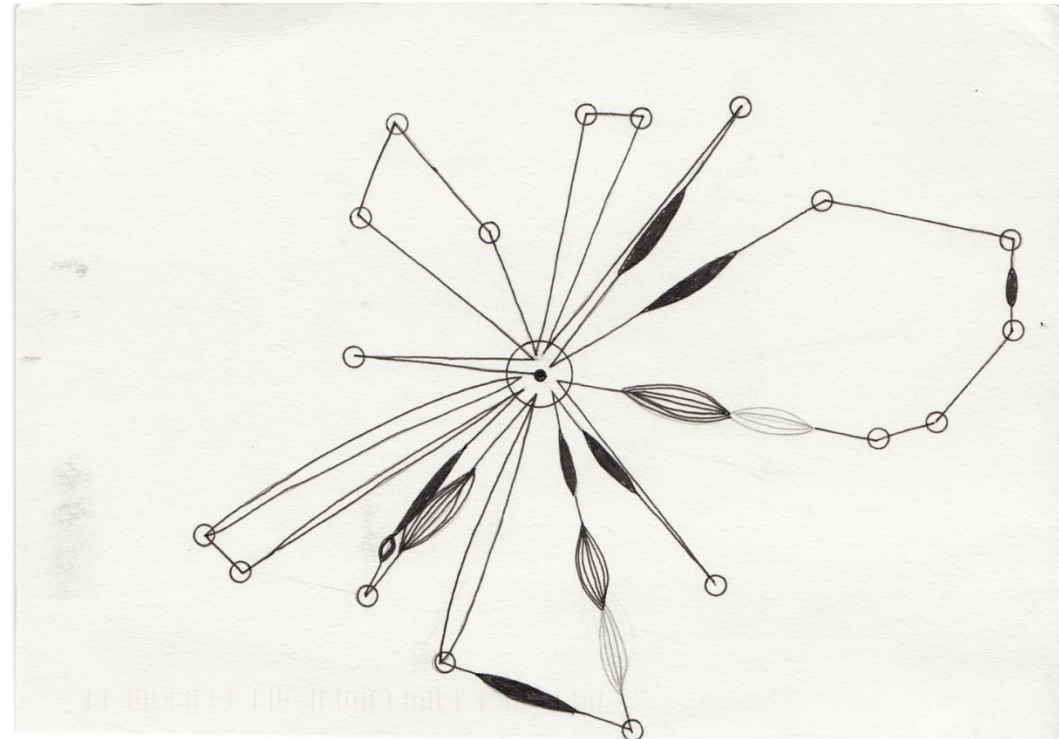
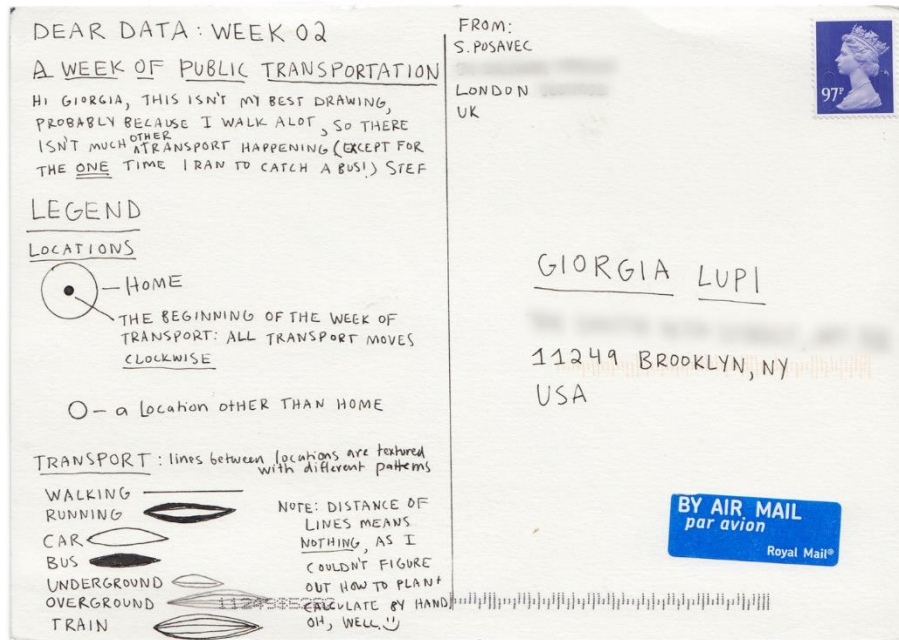
Date	Icon	High	Low
Mon 21		83°	65°
Tue 22		76°	59°
Fri 23		81°	62°
Sat 24		84°	66°
Sun 25		87°	69°
Mon 26		88°	69°
Tue 27		82°	69°
Wed 28		88°	72°
Thu 29		89°	73°

71° 70° 72° 74° 80° 81° 78° 75°
3% 15% 24% 5% 6%

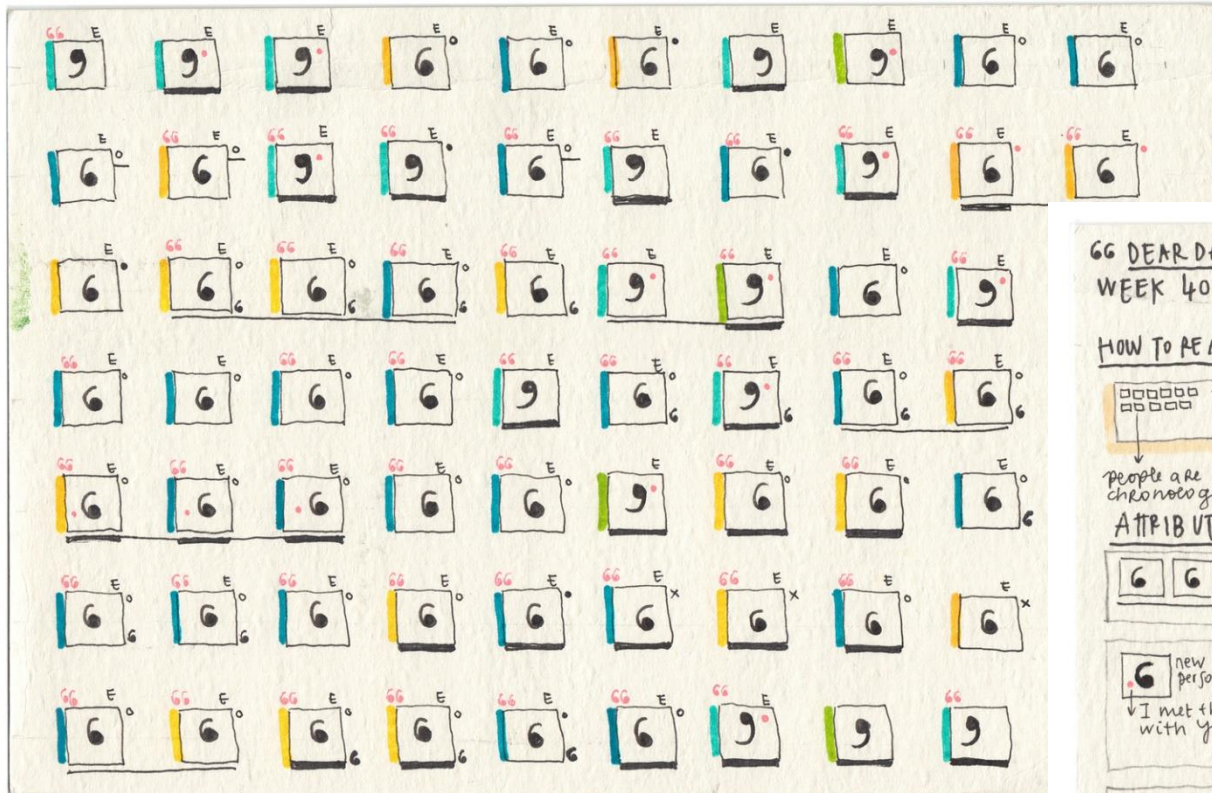
1 AM 4 AM 7 AM 10 AM 1 PM 4 PM 7 PM 10 PM

Dear Data Project

“Instead of using data just to become more efficient, we argue we can use data to become more humane and to connect with ourselves and others at a deeper level.”



Dear Data Project



66 DEAR DATA
WEEK 40: NEW PEOPLE!

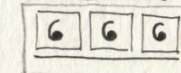
HOW TO READ IT:



Every symbol is a new person I talked to during this week (no waitresses/sales people included)

people are arranged in chronological order

ATTRIBUTES:



line below indicates I met them together/they were together

(A)

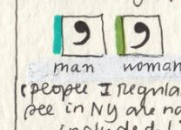
new person I never met before:



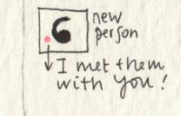
man woman (for at least I didn't remember of ☺)

(B)

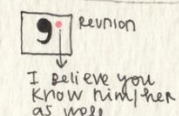
reunion = person I haven't met in 1 year:



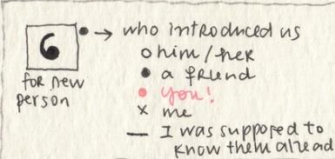
man woman (people I regularly see in NY are not included!)



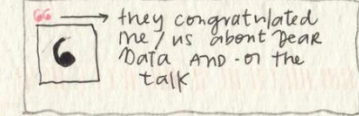
new person I met them with you!



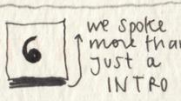
reunion I believe you know him/her as well



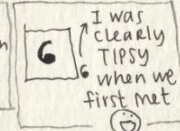
who introduced us for new person
 • him/her
 • a friend
 • you!
 x me
 — I was supposed to know them already



they congratulated me/us about Dear Data AND - on the talk



we spoke more than just a INTRO



I was clearly TIPSY when we first met

MILANO 15.06.11

FROM: Postale (L)



SEND TO:

STEFANIE POSAVEC

LONDON

- UK -

ENGLAND



Data as Selection

Are photographs objective?

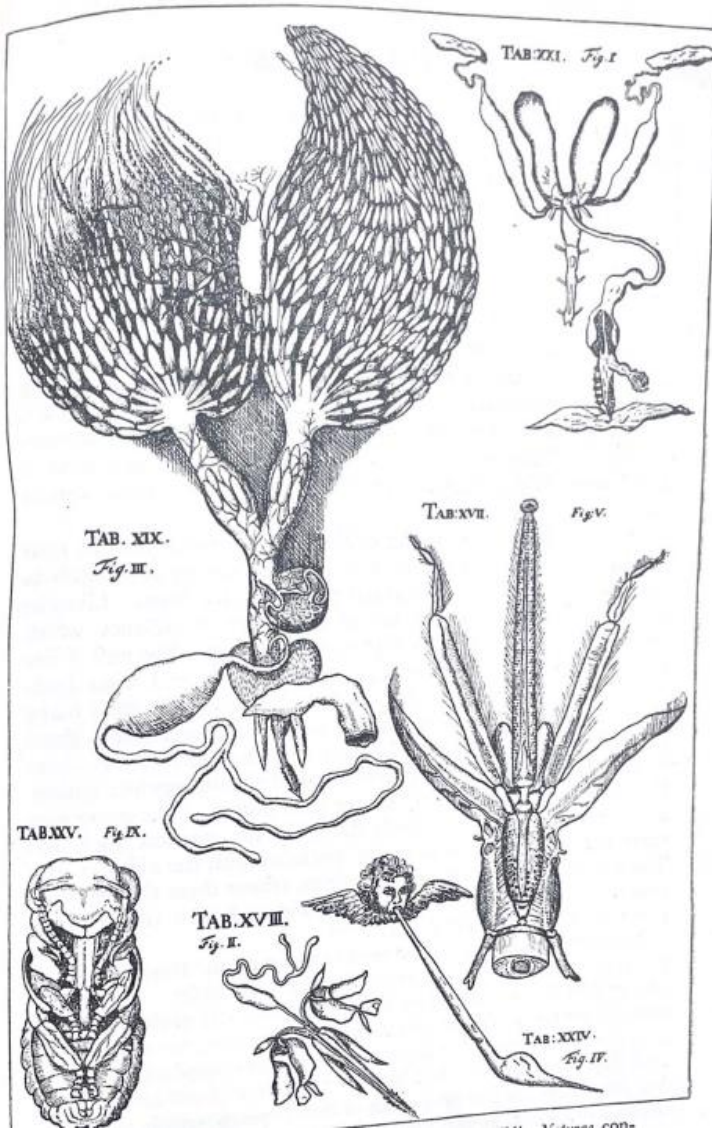
“In deciding how a picture should look, in preferring one exposure to another, photographers are always imposing standards on their subjects. Although there is a sense in which the camera does indeed capture reality, not just interpret it, **photographs are as much an interpretation of the world as paintings and drawings are.**”

(Sontag 17)



Scientific Illustrations

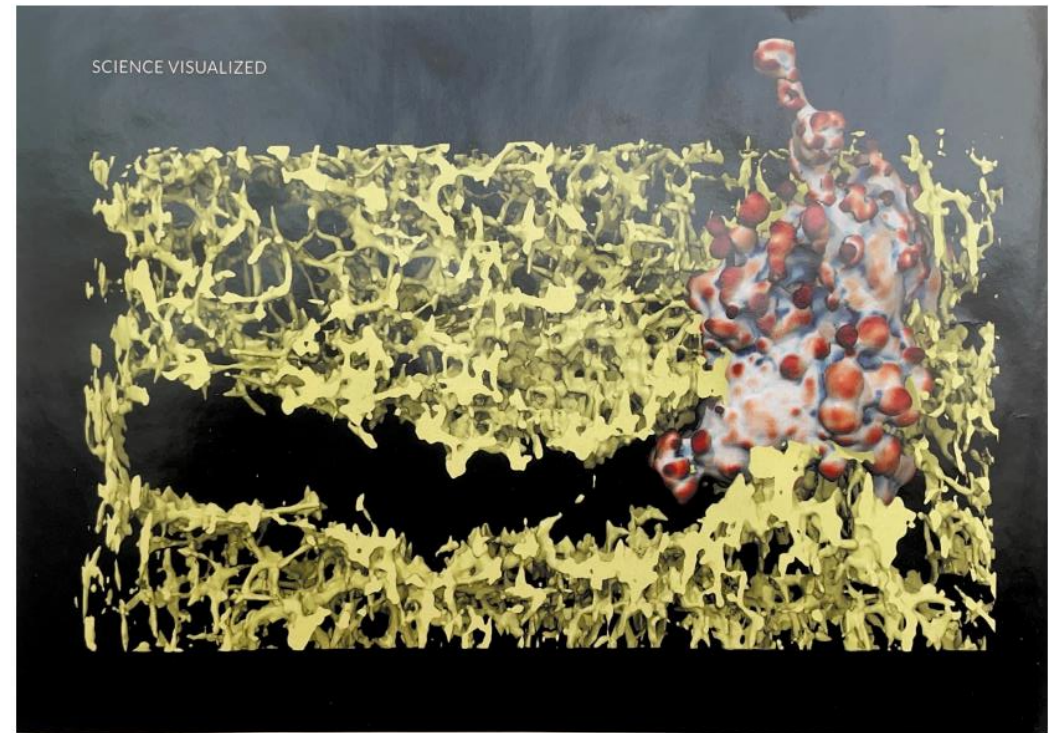
“Science journals viewed photography as a way of seeing that enabled ‘mechanical objectivity’. There was **greater trust in the power of ground lenses and silver halides to capture the world in a way that the eye cannot.** As with all visual technologies, however, it needed selection, organization and interpretation for data to be rendered into a comprehensible image.”
(Belknap 27)



Drawings from *Biblia Naturae* on the honeybee. The lower right illustration shows “the use of dyed liquids blown into the trachea as an aid for their preparation.”

Scientific Photographs


“Scientists soon realized that photographic devices introduce their own distortions into the images that they produce, and that no eye provides an unmediated view onto nature. From the perspective of scientific visualization, **the idea that machines allow us to see true has long been outmoded**...there is a continuing tendency to characterize the objective as that which speaks for itself without the interference of human perception, interpretation, judgement, and so on.”
(Feinberg 54)




Digital reconstruction of a melanoma cell using fleshy membrane protrusions (red) to tunnel through tissue (yellow)

“Data will always be a subjective interpretation of someone’s reality, a specific presentation of the goals and perspectives we choose to prioritize in this moment. That’s a power held by those of us responsible for sourcing, selecting, and designing this data and developing the models that interpret the information.”

(Raji)




Data that is observed, has an observer
Data that is collected, has a collector
Data is about people





Rapid “Datafication”

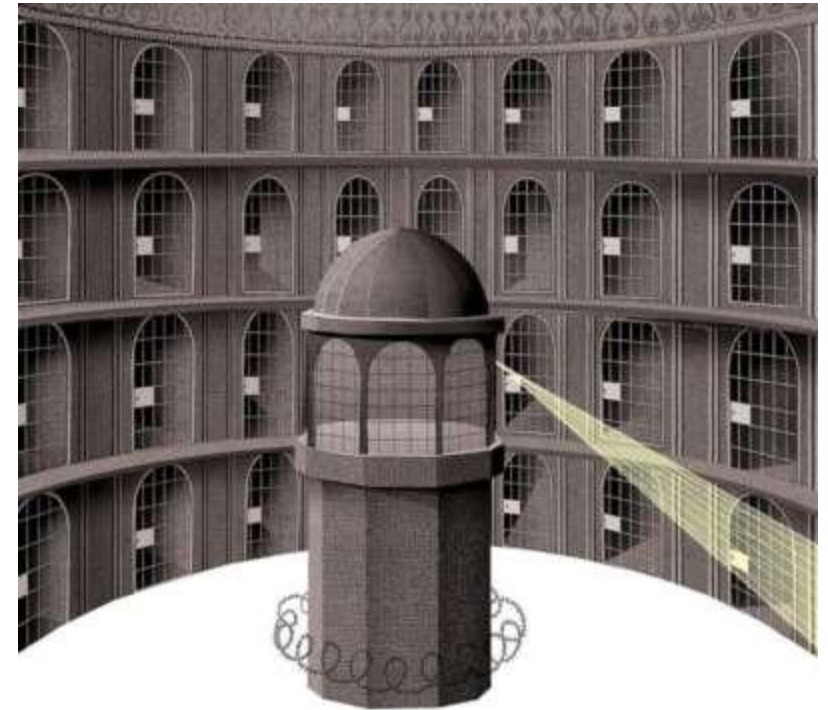
Data is speedy, accessible, revealing, panoramic,
prophetic, and smart



“The answers we need, the solutions to some problem we might not even know that we had, are often thought to be somewhere in the data. It would seem that if only we had more data and more analytics we would know more and waste less.” (Beer 14)

Effects on our “subjects”

- Survey fatigue
- Frustration and anger
 - How are these results being used?
What is the action?
- Surveillance state
 - Panoptic Society
 - The right to be forgotten





Algorithmic Bias





Pasco's sheriff created a futuristic program to stop crime before it happens.

It monitors and harasses families across the county.

Tech

Self-driving cars more likely to drive into black people, study claims

New study suggests autonomous vehicles might be racist

Anthony Cuthbertson • Wednesday 06 March 2019 13:58 GMT • [Comments](#)



Automatic soap dispenser sparks 'racism' outrage after footage shows it doesn't work for dark-skinned people



A VIDEO OF THE AUTOMATIC DISPENSER WENT VIRAL ONLINE

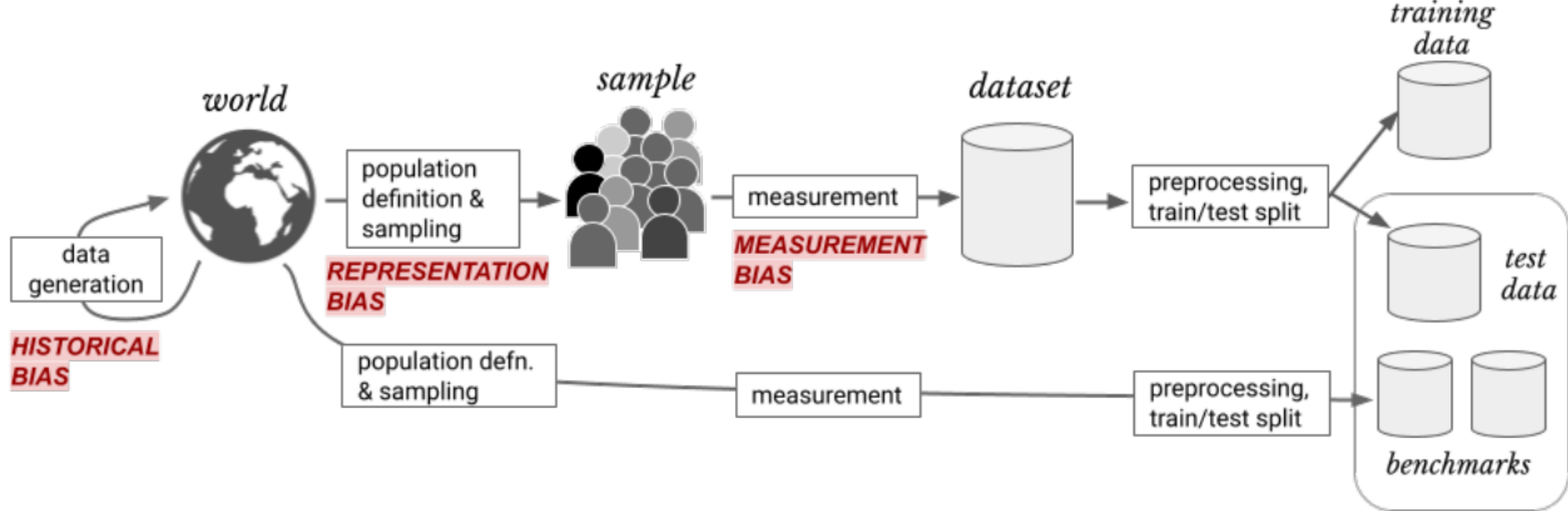
HARRIET PAVEY
18 AUGUST 2017

Woman In China Says Colleague's Face Was Able To Unlock Her iPhone X

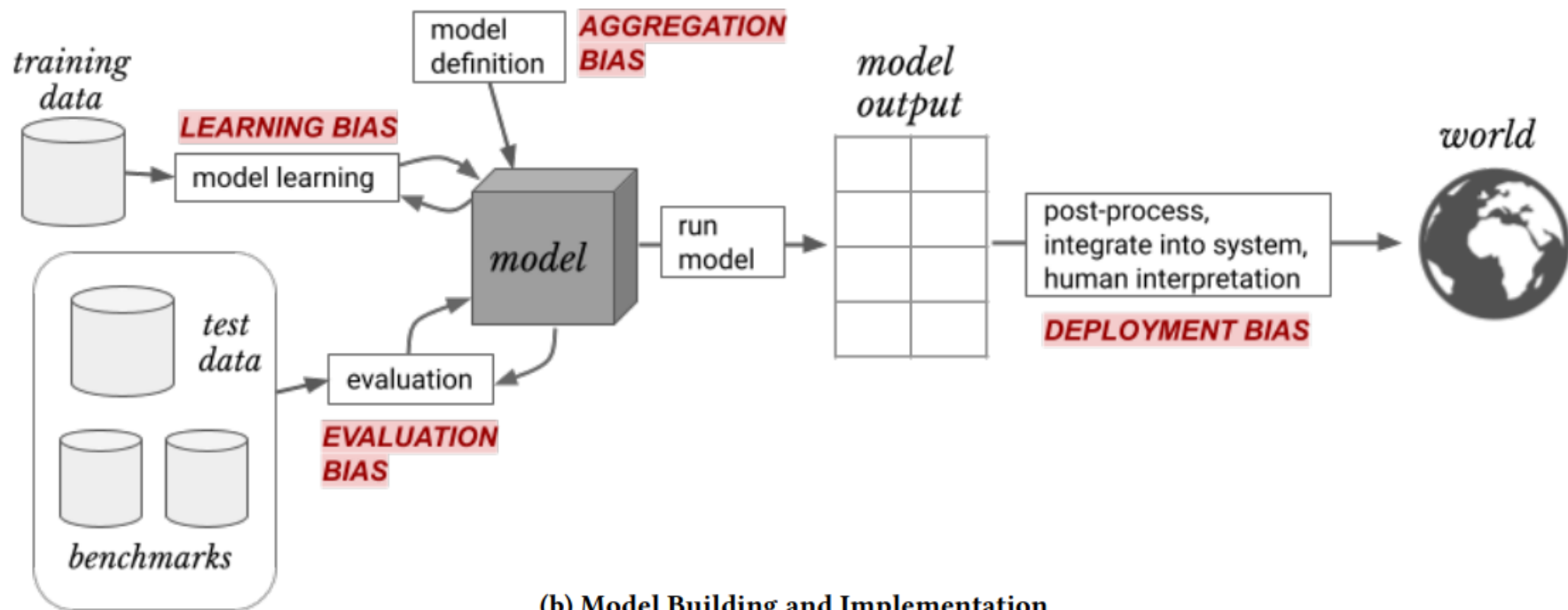
It could also be cheeky passcode training, an Apple spokesman says.

By Mary Papenfuss

Dec 14, 2017, 10:55 PM EST | **Updated** Dec 15, 2017



(a) Data Generation



(b) Model Building and Implementation

Types of Biases

- Historical bias – even if data is measured perfectly, it can capture biases in the world *as it is*
- Representation Bias
 - The target population should reflect the use population
 - The model contains under-represented groups
- Measurement Bias – a proxy is used to approximate a construct
 - In college admissions, predictor of student success \neq GPA



Towards the Future

Learning Analytics

- “Learning analytics has emerged as a solution to address prevalent challenges in education, such as student retention, widening access and personalized support for a massive student cohort” (Tsai 556)
- “provide personalized feedback at scale and to identify variables and behaviours that promote student success and address the need for quality assurance of educational services” (556)

Recommendations

- Prioritize collection of data on gender, race, ethnicity, and national origin
- Collect data on disability status, dialect, socioeconomic status, urbanicity, native language, second-language learner status, national region, parental education background, military–connectedness, and migrant status
- Involve members of communities potentially impacted throughout the entire development and use process

Datasheets for Datasets

- For creators, the objective is to encourage reflection on creating, distributing, and maintaining a dataset including underlying assumptions
- For consumers, the objective is to provide information to make informed decisions

Datasheets for Datasets

Was the “raw” data saved in addition to the preprocessed/cleaned data? (e.g., to support unanticipated future uses)

The raw unprocessed data (consisting of images of faces and names of the corresponding people in the images) is saved.

Is the preprocessing software available?

While a script running a sequence of commands is not available, all software used to process the data is open source and has been specified above.

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?

There are some potential limitations in the dataset which might bias the data towards a particular demographic, pose, image characteristics etc.

- The Viola-Jones detector can have systematic errors by race, gender, age or other categories
- Due to the Viola-Jones detector, there are only a small number of side views of faces, and only a few views from either above or below
- The dataset does not contain many images that occur under extreme (or very low) lighting conditions
- The original images were collected from news paper articles. These articles could cover subjects in limited geographical locations, specific genders, age, race, etc. The dataset does not provide information on the types of garments worn by the individuals, whether they have glasses on, etc.
- The majority of the dataset consists of White males
- There are very few images of people who are under 20 years old
- The proposed train/test protocol allows reuse of data between View 1 and View 2 in the dataset. This could potentially introduce very small biases into the results

Dataset Distribution

How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)

The dataset can be downloaded from <http://vis-www.cs.umass.edu/lfw/index.html#download>. The images can be downloaded as a gzipped tar file.

What license (if any) is it distributed under? Are there any copyrights on the data?

The crawled data copyright belongs to the news papers that the data originally appeared in. There is no license, but there is a request to cite the corresponding paper if the dataset is used: Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.

Are there any fees or access/export restrictions?

There are no fees or restrictions.

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset?

The dataset is hosted at the University of Massachusetts and all and comments can be sent to: Gary Huang - gbyhuang@cs.umass.edu.

Will the dataset be updated? If so, how often and by whom?

Unknown

How will updates be communicated? (e.g., mailing list, GitHub)

All changes to the dataset will be announced through the LFW mailing list. Those who would like to sign up should send an email to lfw-subscribe@cs.umass.edu.

Is there an erratum?

Errata are listed under the “Errata” section of <http://vis-www.cs.umass.edu/lfw/index.html>

If the dataset becomes obsolete how will this be communicated?

All changes to the dataset will be announced through the LFW mailing list.

Is there a repository to link to any/all papers/systems that use this dataset?

Papers using this dataset and the specified training/evaluation protocols are listed under “Methods” section of <http://vis-www.cs.umass.edu/lfw/results.html>

If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

Unknown

Table 1: Lessons from Archives: summaries of approaches in archival and library sciences to some of the most important topics in data collection, and how they can be applied in the machine learning setting.

Consent	(1) Institute data gathering outreach programs to actively collect underrepresented data (2) Adopt crowdsourcing models that collect open-ended responses from participants and give them options to denote sensitivity and access
Inclusivity	(1) Complement datasets with “Mission Statements” that signal commitment to stated concepts/topics/groups (2) “Open” data sets to promote ongoing collection following mission statements
Power	(1) Form data consortia where data centers of various sizes can share resources and the cost burdens of data collection and management
Transparency	(1) Keep process records of materials added to or selected out of dataset. (2) Adopt a multi-layer, multi-person data supervision system.
Ethics & Privacy	(1) Promote data collection as a full-time, professional career. (2) Form or integrate existing global/national organizations in instituting standardized codes of ethics/conduct and procedures to review violations

Let's talk about
equity and data!



**COMMUNITY
COLLEGE OF
PHILADELPHIA**

**Institutional
Effectiveness**

Sources

- Baker, Ryan S., and Aaron Hawn. "Algorithmic Bias in Education." *International Journal of Artificial Intelligence in Education*, vol. 32, no. 4, 2022, pp. 1052-1092. ProQuest, <https://www.proquest.com/scholarly-journals/algorithmic-bias-education/docview/2932309122/se-2>, doi:<https://doi.org/10.1007/s40593-021-00285-9>.
- Beer, David. *The Data Gaze: Capitalism, Power and Perception*. SAGE Publications Ltd, 2019.
- Belknap, Geoffrey. "The Evolution of Scientific Illustration." *Nature*, vol. 575, no. 7781, Nov. 2019, pp. 25–28. ProQuest, <https://doi.org/10.1038/d41586-019-03306-9>.
- Blanchette, Jean-François, and Deborah G. Johnson. "Data Retention and the Panoptic Society: The Social Benefits of Forgetfulness." *Information Society*, vol. 18, no. 1, Jan. 2002, pp. 33–45. EBSCOhost, <https://doi-org.ezproxy.ccp.edu/10.1080/01972240252818216>.
- Feinberg, Melanie. *Everyday Adventures with Unruly Data*. The MIT Press, 2022.
- Francis, Peter, et al. "Thinking Critically about Learning Analytics, Student Outcomes, and Equity of Attainment." *Assessment & Evaluation in Higher Education*, vol. 45, no. 6, Sept. 2020, pp. 811–21. EBSCOhost, <https://doi.org/10.1080/02602938.2019.1691975>.
- Gebru, Timnit, et al. *Datasheets for Datasets*. arXiv:1803.09010, arXiv, 1 Dec. 2021. *arXiv.org*, <http://arxiv.org/abs/1803.09010>.
- Jo, Eun Seo, and Timnit Gebru. "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ACM, 2020, pp. 306–16. DOI.org (Crossref), <https://doi.org/10.1145/3351095.3372829>.
- Lupi, Giorgia, and Stefanie Posavec. "The Project." *Dear Data*, www.dear-data.com/theproject. Accessed 21 Aug. 2024.
- Noble, Safiya Umoja. *Algorithms of Oppression*. New York University Press, 2018.
- Raji, Deborah. "How Our Data Encodes Systematic Racism." *MIT Technology Review*, MIT Technology Review, 21 Jan. 2021, www.technologyreview.com/2020/12/10/1013617/racism-data-science-artificial-intelligence-ai-opinion/.
- Sontag, Susan. *On Photography*. Penguin Classics, 2008.
- Suresh, Harini, and John Guttag. "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle." *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, Association for Computing Machinery, 2021, pp. 1–9. *ACM Digital Library*, <https://doi.org/10.1145/3465416.3483305>.
- Tsai, Yi-Shan, et al. "Empowering Learners with Personalised Learning Approaches? Agency, Equity and Transparency in the Context of Learning Analytics." *Assessment & Evaluation in Higher Education*, vol. 45, Aug. 2020, pp. 554–67. EBSCOhost, <https://doi.org/10.1080/02602938.2019.1676396>.

Thank You



[E: dsthomas@ccp.edu](mailto:dsthomas@ccp.edu)

P: 8148

O: A7-118