Lessons from Archives: Strategies for Collecting Sociocultural Data in ML

Eun Seo Jo, Stanford University

Timnit Gebru, Google

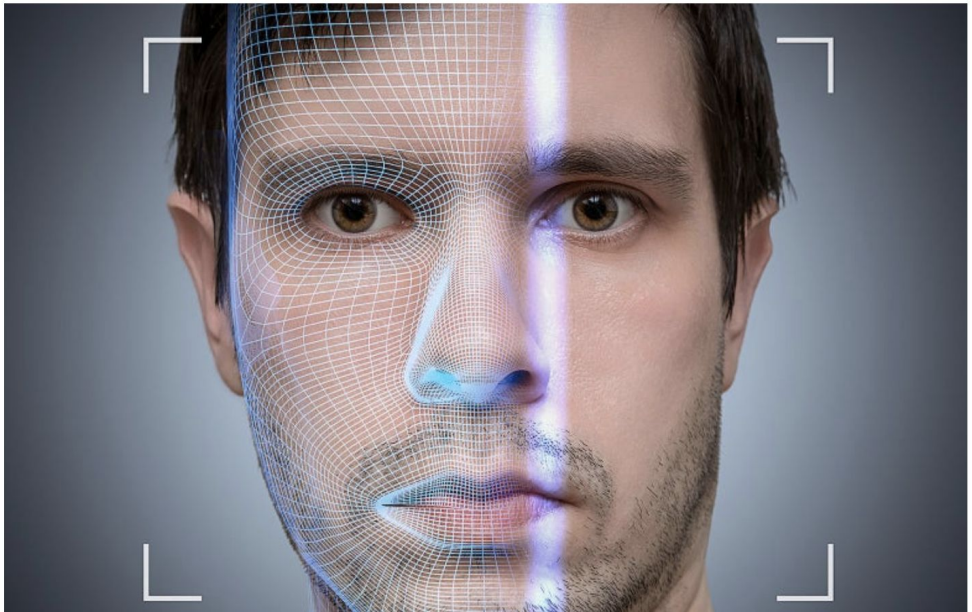
- Many issues in fairness accountability transparency and ethics are rooted in decisions surrounding the data collection, annotation and ownership practices.

The Government Is Using the Most Vulnerable People to Test Facial Recognition Software

Our research shows that any one of us might end up helping the facial recognition industry, perhaps during moments of extraordinary vulnerability.

By OS KEYES, NIKKI STEVENS, and JACQUELINE WERNIMONT

MARCH 17, 2019 • 8:32 PM



FACIAL RECOGNITION • POLITICS • ETHICS • DATA SCIENCE • GENDER • QUEER

Counting the Countless

BY OS KEYES

MARCH 24, 2019

Reconstructed transcript of a talk I gave at Seattle University earlier this year

Good evening everyone! My name is Os, and I'm a PhD student at the University of Washington. According to my website I study gender, data, technology and control; it also says that I'm an inaugural Ada Lovelace Fellow. And I'm here for a variety of reasons, but one of the big ones is that I really enjoy giving talks. Particularly community-oriented talks; remaining grounded in my communities is important to me and for my work to be effective. So I was really pleased when, as a result of my *last* talk here, the Seattle Non-Binary Collective reached out. And they said: "we hear you're a data scientist. Could you do a talk on how trans &/ non-binary people can get involved in data science?"

And I replied: well, to be perfectly honest, I think data science is a profound threat to queer existences. And then for some reason they stopped replying! Who can say why? So when Jodi asked me what I'd like to talk about in this lecture series, I figured I'd do a talk on *that*. Why do I think data science is a profound threat for queer people?

The difficulty of definitions

Error Rate(1-PPV) By Female x Skin Type

	TYPE I	TYPE II	TYPE III	TYPE IV	TYPE V	TYPE VI
Microsoft	1.7%	1.1%	3.3%	0%	23.2%	25.0%
FACE++	11.9%	9.7%	8.2%	13.9%	32.4%	46.5%
IBM	5.1%	7.4%	8.2%	8.3%	33.3%	46.8%

Buolamwini & Gebru FAT* 2018, Slides from Joy Buolamwini

Microsoft improves facial recognition technology to perform well across all skin tones, genders

June 26, 2018 | John Roach

DIGITAL

Amazon Rekognition May Finally Be Audited and Ranked Alongside Other Vendors

A more universal test for facial recognition systems is needed

By Lisa Lacy | February 19, 2019

IBM

IBM Research Blog Topics Labs About

AI

IBM Research Releases 'Diversity in Faces' Dataset to Advance Study of Fairness in Facial Recognition Systems

NEWS MANAGEMENT OVERSIGHT DEFENSE TECH CONTRACTING PAY & B

Senators Are Asking Whether Artificial Intelligence Could Violate U.S. Civil Rights Laws

By Dave Gershgorn | Quartz | September 21, 2018 | 5 Comments

Senators Kamala Harris and Cory Booker both signed letters to federal agencies asking about AI bias. J. Scott Applewhite/AP

RELATED

Shutdown Roundup: NTSB Isn't Investigating Accidents, Warren Warns of Decreased Financial Investigations and More

January 25, 2019 | 1 Comment

Democratic Senator Questions Mulvaney's Hatch Act Compliance

September 21, 2018 | 32 Comments

Why Philadelphia Is on the Federal Government's

Google using dubious tactics to target people with ‘darker skin’ in facial recognition project: sources

By GINGER ADAMS OTIS and NANCY DILLON
NEW YORK DAILY NEWS | OCT 02, 2019 | 6:56 PM

A person takes part in a Google facial recognition project. (Obtained by Daily News)

ARGUMENT

Beijing's Big Brother Tech Needs African Faces

Zimbabwe is signing up for China's surveillance state, but its citizens will pay the price.

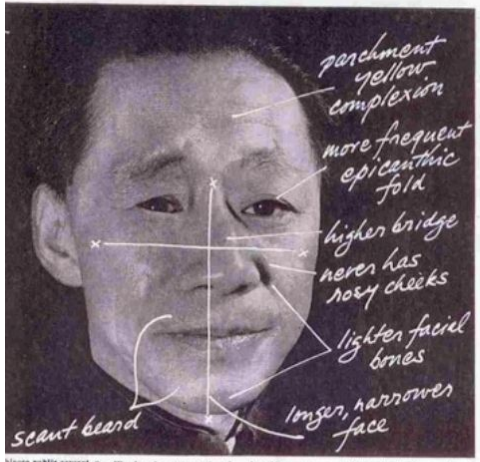
BY AMY HAWKINS

JULY 24, 2018, 10:39 AM

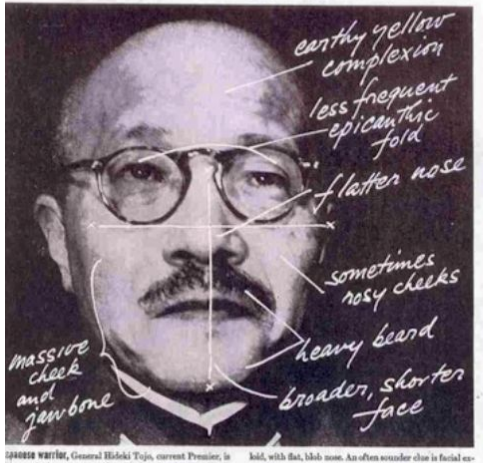
- Sometimes you need to make sure there aren’t disparate error rates across subgroups (e.g. melanoma detection).

- Sometimes the task just should not exist (e.g. automatic gender recognition).

- Sometimes the manner in which the tool is used is very problematic because of who has the power (data) and ability to train powerful models, vs who is subjected to those models.



hinese public servant, Ong Wen-hao, is representative of orth Chinese anthropological group with long, fine-boned es and scant beard. Epicanthic fold of skin above eyelid found in 85% of Chinese. Southern Chinese have round, broad faces, not as massively boned as the Japanese. Except that their skin is darker, this description fits Filipinos who are often mistaken for Japs. Chinese sometimes pass for Europeans; but Japs more often approach Western types.



apanese warlor, General Hideki Tojo, current Premier, is Ye-natural, closer to type of humble Jap than highbred relaives of Imperial Household. Typical are his heavy beard, assive cheek and jaw bones. Peasant Jap is squat Mongoloid, with flat, blob nose. An often sounder clue is facial expression, shaped by cultural, not anthropological, factors. Chinese wear rational calm of tolerant realists. Japs, like General Tojo, show humorless intensity of ruthless mystics.

HOW TO TELL JAPS FROM THE CHINESE

ANGRY CITIZENS VICTIMIZE ALLIES WITH EMOTIONAL OUTBURST AT ENEMY

In the first discharge of emotions touched off by the Japanese assaults on their nation, U. S. citizens have been demonstrating a distressing ignorance on the delicate question of how to tell a Chinese from a Jap. Innocent victims in cities all over the country are many of the 75,000 U. S. Chinese, whose homeland is our stanch ally. So serious were the consequences threatened, that the Chinese consulates last week prepared to tag their nationals with identification buttons. To dispel some of this confusion, LIFE here adduces a rule-of-thumb from the anthropometric conformations that distinguish friendly Chinese from enemy alien Japs.

To physical anthropologists, devoted debunkers of race myths, the difference between Chinese and Japs is measurable in millimeters. Both are related to the Eskimo and North American Indian. The modern Jap is the descendant of Mongoloids who invaded the Japanese archipelago back in the mists of prehistory, and of the native aborigines who possessed the islands before them. Physical anthropology, in consequence, finds Japs and Chinese as closely related as Germans and English. It can, however, set apart the special types of each national group.

The typical Northern Chinese, represented by Ong Wen-hao, Chungking's Minister of Economic Affairs (*left, above*), is relatively tall and slenderly built. His complexion is parchment yellow, his face long and delicately boned, his nose more finely bridged. Representative of the Japanese people as a whole is Premier and General Hideki Tojo (*left, below*), who betrays aboriginal antecedents in a squat, long-torsoed build, a broader, more massively boned head and face, flat, often pug, nose, yellow-ocher skin and heavier beard. From this average type, aristocratic Japs, who claim kinship to the Imperial Household, diverge sharply. They are proud to approximate the patrician lines of the Northern Chinese.

Chinese journalist, Joe Chiang, found it necessary to advertise his nationality to gain admittance to White House press conference. Under Immigration Act of 1924, Japs and Chinese, as members of the "yellow race," are barred from immigration and naturalization.

CONTINUED ON NEXT PAGE

The New York Times

One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority

In a major ethical leap for the tech world, Chinese start-ups have built algorithms that the government uses to track members of a largely Muslim minority group.

US ADULTS INDEXED

130 MILLION

One in two American adults is in a law enforcement face recognition network used in unregulated searches employing algorithms with unaudited accuracy.

The Perpetual Line Up

(Garvie , Bedoya, Frankle 2016)

AMERICA UNDER WATCH

FACE SURVEILLANCE IN THE UNITED STATES

May 16, 2019

Clare Garvie
ClareAngelyn

Laura M. Moy
lauramoy

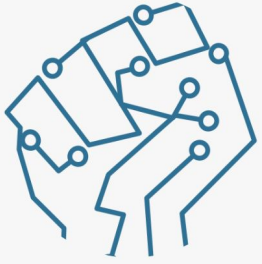
INTRODUCTION

"Real-time video surveillance appears to be a simple question of supply and demand. As the technology improves, we anticipate that real-time face recognition systems will become commonplace." [1]

—*The Perpetual Line-Up*, 2016

Authorities in Guiyang have eyes everywhere. Thanks to a vast, sophisticated camera system blanketing this Southwest Chinese city, police are purportedly able to locate and identify anyone who shows their face in public—in a matter of minutes. They can trace where you have been over the past week. If you are a citizen they can "match your face with your car, match you with your relatives and

Amazon Pushes Facial Recognition to Police. Critics See Surveillance Risk.



Black in AI

Donate

Black in AI (BAI)

Black in AI is a place for sharing ideas, fostering collaborations and discussing initiatives to increase the presence of Black people in the field of Artificial Intelligence. If you are in the field of AI and self-identify as Black, please fill out this Google Form to request to join and we will add you to various platforms that we maintain. We also welcome allies to join our group using the Google form. Allies will be added to our email lists, where we send out group updates and requests for assistance.

Like our Facebook Page and follow us on Twitter to learn about our members and various activities!

Tweets by @black_in_ai

Black in AI Retweeted

Thabo Malete
@NerdBw

Highlight of last week: Got the opportunity to present my work on EEG based control at the @black_in_ai conference which ran alongside #NeurIPS18 @biustbw

'Bias deep inside the code': the problem with AI 'ethics' in Silicon Valley

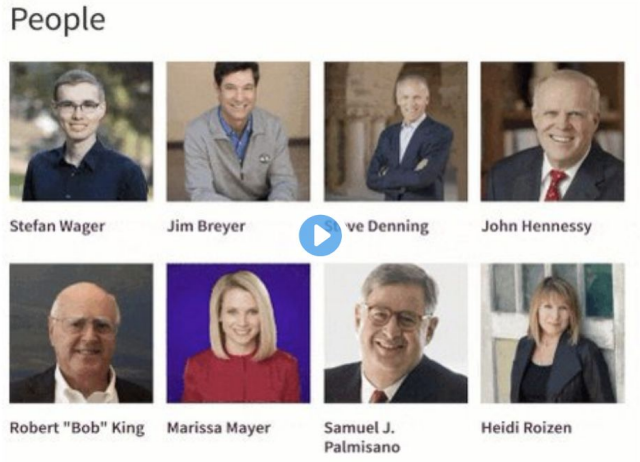
As algorithms play a growing role in criminal justice, education and more, tech advisory boards and academic programs mirror real-world inequality

Chad Loder
@chadloder

Stanford just launched their Institute for Human-Centered Artificial Intelligence (@StanfordHAI) with great fanfare. The mission: "The creators and designers of AI must be broadly representative of humanity."

121 faculty members listed.

Not a single faculty member is Black.



2,464 8:39 PM - Mar 20, 2019

1,501 people are talking about this

…while the fair ML literature has largely focused on "de-biasing" methods and viewed the training data as fixed, most of our interviewees report that their teams consider data collection, rather than model development, as the most important place to intervene

Holstein et al FAT* 2019

- We need a whole specialty in ML just dealing with data. We don’t teach that in classes or treat it as a fundamental part of our scientific publications.

- Issues to do with data are complex. E.g.
    - Consent
    - Power
    - Inclusivity
    - Representation
    - Subgroup classification, annotation
    - Privacy & ethics
    - Transparency

- We need to have an interdisciplinary approach specifically focused on data. We can learn from disciplines such as anthropology and history

Anthropological/Artificial Intelligence & the HAI

26 MARCH 2019

Last week Stanford launched the institute for human-centered artificial intelligence, and to kick things off James Landay posted about the roles AI could play in society, and the importance of exploring smart interfaces.

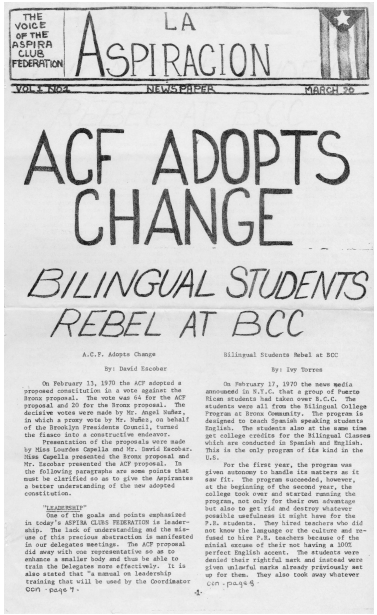
I've followed the HAI's development in passing, and I watched the inaugural event in the background on Monday last week while I was doing other work. I study algorithmic systems that make important decisions about us - which I call "street-level algorithms" in reference to Michael Lipsky's street-level bureaucracies - and some of the work I've done in the past has taken a more careful look at historical parallels between things we see today (like quantified self and piecework) to see if we can learn anything useful either for making sense of phenomena from a sociological perspective, and sometimes for informing the design of systems from an engineering perspective. James is a professor in the Human-Computer Interaction group at Stanford, and I'm a PhD student in that group.

So I was worried to find James leave details out from a series of anecdotes - details that would seriously undermine the point James seemed to be trying to make in his post. I started writing notes to call out how a more cynical perspective might describe the future or remember the past that James writes about; but with the launch of the HAI, the reaction from people around the world, and specifically *the responses from people in the HAI*, it seems like a more serious point that needs to be made.

The voices, opinions, and needs of disempowered stakeholders are being ignored today in favor of stakeholders with power, money, and influence - as they have been historically; our failure to listen promises to doom initiatives like the HAI.

- Archives are the oldest attempt to gather human data

What are Archives?



THE VOICE OF THE ASPIRA CLUB FEDERATION

LA ASPIRACION

VOL. 1 NO. 2 NEWSPAPER MARCH 70

ACF ADOPTS CHANGE

BILINGUAL STUDENTS REBEL AT BCC

A.C.F. Adopts Change

By: David Escobar

On February 13, 1970 the ACF adopted a proposed constitution in a vote against the Bronx proposal. The vote was 64 for the ACF proposal and 20 for the Bronx proposal. The decisive votes were made by Mr. Angel Suñez, in which a proxy vote by Mr. Suñez, on behalf of the Brooklyn Presidents Council, turned the fiasco into a constructive endeavor.

Presentation of the proposals were made by Miss Lourdes Capella and Mr. David Escobar. Miss Capella presented the Bronx proposal and Mr. Escobar presented the ACF proposal. In the following paragraphs are some points that must be clarified so as to give the Aspirantes a better understanding of the new adopted constitution.

"LEADERSHIP"

One of the goals and points emphasized in today's ASPIRA CLUBS FEDERATION is leadership. The lack of understanding and the misuse of this precious abstraction is manifested in our delegates meetings. The ACF proposal did away with one representative so as to enhance a smaller body and thus be able to train the delegates more effectively. It is also stated that "a manual on leadership training that will be used by the Coordinator

con. page 6

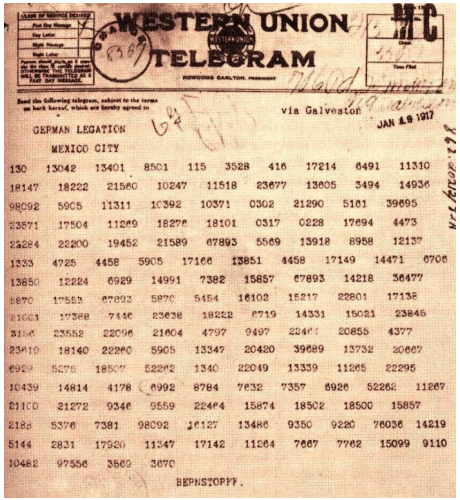
Bilingual Students Rebel at BCC

By: Ivy Torres

On February 17, 1970 the news media announced in N.Y.C. that a group of Puerto Rican students had taken over B.C.C. The students were all from the Bilingual College Program at Bronx Community. The program is designed to teach Spanish speaking students English. The students also at the same time get college credits for the Bilingual Classes which are conducted in Spanish and English. This is the only program of its kind in the U.S.

For the first year, the program was given autonomy to handle its matters as it saw fit. The program succeeded, however, at the beginning of the second year, the college took over and started running the program, not only for their own advantage but also to get rid and destroy whatever possible usefulness it might have for the P.R. students. They hired teachers who did not know the language or the culture and refused to hire P.R. teachers because of the minial excuse of their not having a 100% perfect English accent. The students were denied their rightful mark and instead were given unlawful marks already previously set up for them. They also took away whatever

con. page 6



WESTERN UNION TELEGRAM

via Galveston

JAN 19 1917

GERMAN LEGATION

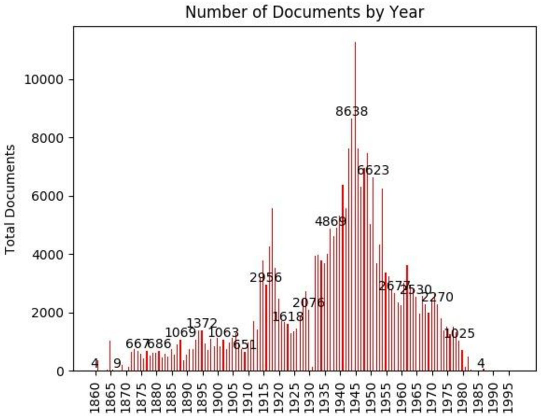
MEXICO CITY

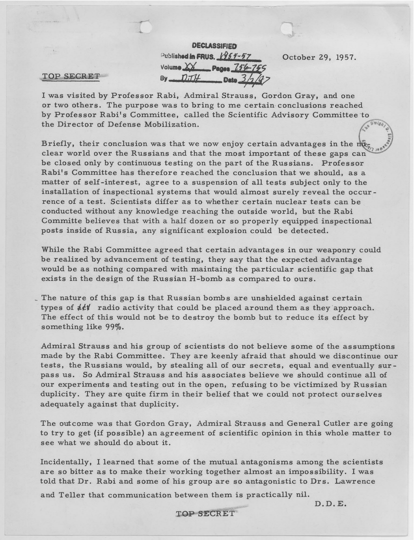
130 13042 13401 8501 115 3528 416 17214 6491 11310
18147 18222 21560 10247 11518 23677 13805 3494 14936
98092 5905 11311 10392 10371 0302 21290 5161 39695
23571 17504 11269 18276 18101 0317 0228 17694 4473
24284 22200 19452 21589 07893 5569 13918 8958 12137
1333 4725 4458 5905 17166 13851 4458 17149 14471 6706
13850 12224 6929 14991 7382 15857 67893 14218 36477
5870 17553 67893 5870 5454 16102 15217 22801 17138
21001 17388 7446 23638 18222 6719 14331 15021 23845
3146 23552 22096 21604 4797 9497 2240[illegible] 20855 4377
23610 18140 22260 5905 13347 20420 39689 13732 20667
6929 5575 18507 52262 1340 22049 13339 11265 22295
10439 14814 4178 6992 8784 7632 7357 6926 52262 11267
21100 21272 9346 9559 22464 15874 18502 18500 15857
2188 5376 7381 98092 16127 13486 9350 9220 76036 14219
5144 2831 17926 11347 17142 11264 7667 7762 15099 9110
10482 97556 3569 3670

BERNSTORFF.

- Institutional Missions /Agenda
- “Balanced and Representative Record”
- Private (Rockefeller Archives) & Public (UN Archives)

Case Study 1:
U.S. State Department





DECLASSIFIED
Published in FRUS. 1955-57
Volume XX Pages 756-755
By DJH Date 3/7/97

October 29, 1957.

TOP SECRET

I was visited by Professor Rabi, Admiral Strauss, Gordon Gray, and one or two others. The purpose was to bring to me certain conclusions reached by Professor Rabi's Committee, called the Scientific Advisory Committee to the Director of Defense Mobilization.

Briefly, their conclusion was that we now enjoy certain advantages in the nuclear world over the Russians and that the most important of these gaps can be closed only by continuous testing on the part of the Russians. Professor Rabi's Committee has therefore reached the conclusion that we should, as a matter of self-interest, agree to a suspension of all tests subject only to the installation of inspectional systems that would almost surely reveal the occurrence of a test. Scientists differ as to whether certain nuclear tests can be conducted without any knowledge reaching the outside world, but the Rabi Committe believes that with a half dozen or so properly equipped inspectional posts inside of Russia, any significant explosion could be detected.

While the Rabi Committee agreed that certain advantages in our weaponry could be realized by advancement of testing, they say that the expected advantage would be as nothing compared with maintaing the particular scientific gap that exists in the design of the Russian H-bomb as compared to ours.

The nature of this gap is that Russian bombs are unshielded against certain types of ~~act~~ radio activity that could be placed around them as they approach. The effect of this would not be to destroy the bomb but to reduce its effect by something like 99%.

Admiral Strauss and his group of scientists do not believe some of the assumptions made by the Rabi Committee. They are keenly afraid that should we discontinue our tests, the Russians would, by stealing all of our secrets, equal and eventually surpass us. So Admiral Strauss and his associates believe we should continue all of our experiments and testing out in the open, refusing to be victimized by Russian duplicity. They are quite firm in their belief that we could not protect ourselves adequately against that duplicity.

The outcome was that Gordon Gray, Admiral Strauss and General Cutler are going to try to get (if possible) an agreement of scientific opinion in this whole matter to see what we should do about it.

Incidentally, I learned that some of the mutual antagonisms among the scientists are so bitter as to make their working together almost an impossibility. I was told that Dr. Rabi and some of his group are so antagonistic to Drs. Lawrence and Teller that communication between them is practically nil.

D.D.E.

TOP SECRET

- Historians at the State Department
- Culled from State/Presidential/CIA etc. Archives
- Declassified Documents, Hand Selected
- Procedure determining relevance + scope

Case Study 2: Internet Archive

Internet Archive is a non-profit library of millions of free books, movies, software, music, websites, and more.

349B 20M 4.8M 5.0M 1.8M 351K 3.3M 200K 402K

INTERNET ARCHIVE

- Crawling Webpages
- 30+ petabytes/330 billion webpages
- 20+ years of internet
- No institutional mission (bigger → better)

Data Collection Spectrum

Data Collection Spectrum

“If it is available to us, we ingest it.”
- NLP ML Engineer
(Holstein et al., 2019)

Data Collection Spectrum

“If it is available to us, we ingest it.”
- NLP ML Engineer (Holstein et al., 2019)

“Driven by collection initiatives and agendas”
-University Archivist

Data Collection Spectrum

Poor Supervision of Data Collection and Handling

Supervision

Professional and Selective Curation

Data Collection Spectrum

Poor Supervision of Data Collection and Handling

Commercial Products/Research Output, Accuracy (minimal public outcry, consumer complaints, Holstein et al., 2019)

Supervision

Objectives

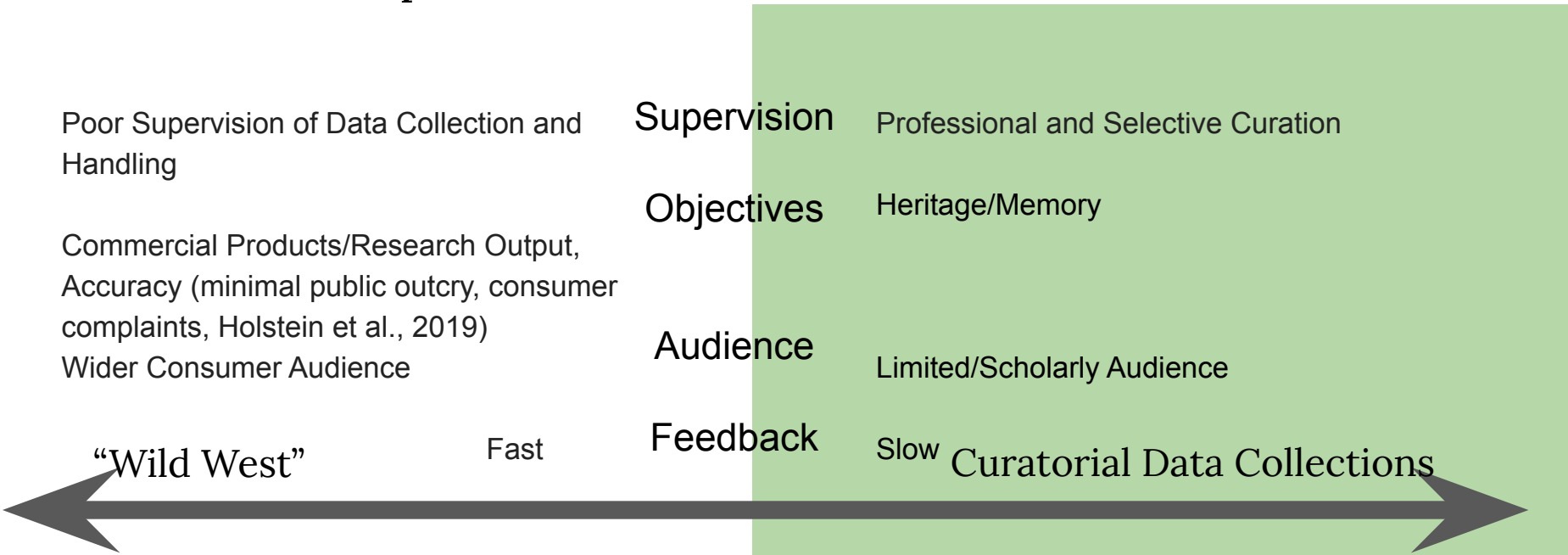
Professional and Selective Curation

Heritage/Memory

Data Collection Spectrum

Poor Supervision of Data Collection and Handling	Supervision	Professional and Selective Curation
Commercial Products/Research Output, Accuracy (minimal public outcry, consumer complaints, Holstein et al., 2019)	Objectives	Heritage/Memory
Wider Consumer Audience	Audience	Limited/Scholarly Audience

Data Collection Spectrum

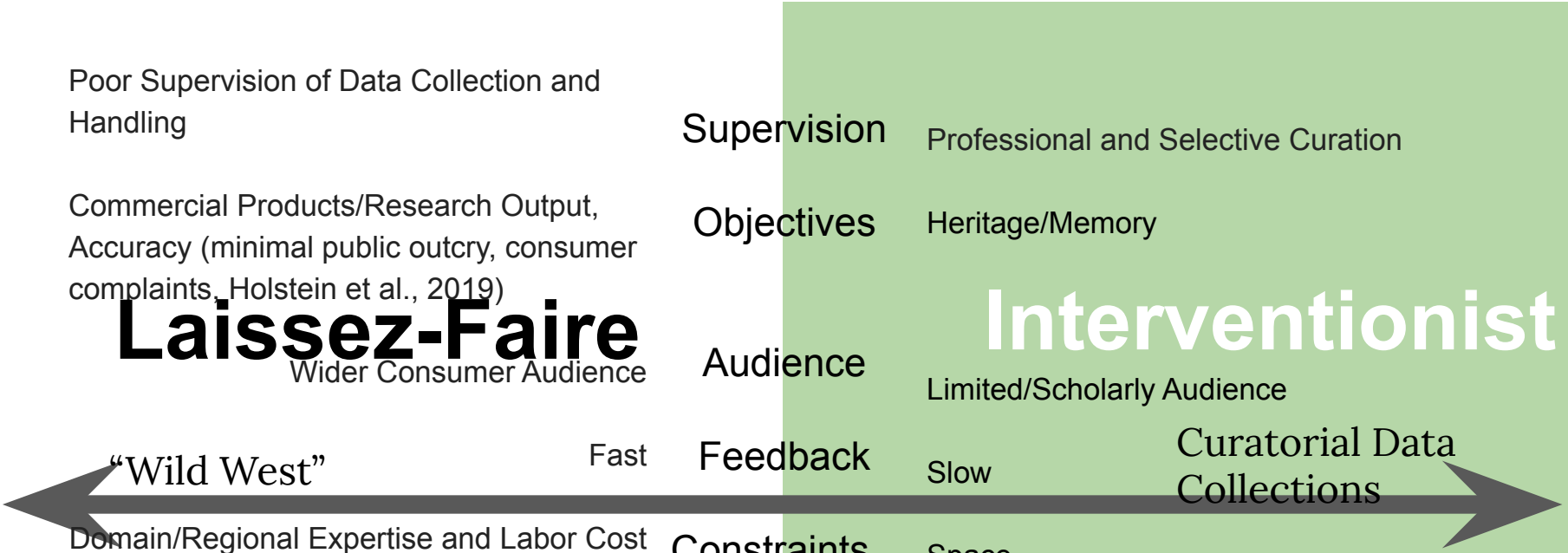


Data Collection Spectrum

Poor Supervision of Data Collection and Handling	Supervision	Professional and Selective Curation
Commercial Products/Research Output, Accuracy (minimal public outcry, consumer complaints, Holstein et al., 2019)	Objectives	Heritage/Memory
Wider Consumer Audience	Audience	Limited/Scholarly Audience
Fast	Feedback	Slow

Concerned with Size and Inclusivity	Concerned with Authenticity, Privacy, Value
Curates 'natural' data BUT replicates social biases	**Catered and Carefully Considered BUT Politically biased by the Curators' Agenda/Initiatives**

Data Collection Spectrum



From MACRO to MICRO

Mission Statement

Collection Development Policy

Appraisal

Processing

From MACRO to MICRO

Mission Statement

Collection Development Policy

Appraisal

Processing

From MACRO to MICRO

Mission Statement

Collection Development Policy

Appraisal

Processing

From MACRO to MICRO

Mission Statement

Collection Development Policy

Appraisal

Processing

From MACRO to MICRO

Mission Statement

Collection Development Policy

Appraisal

Processing

1. Mission Statement

- Highest level of agenda formulation determining topic/concepts of concern.

2. Collection Development Policy

- A more specific policy drawn from the mission statement about what is collected, what is not, where to search for sources.

3. Appraisal

- Evaluation based on criteria of whether a given selection of sources is worth collecting.
- Asking whether this collection fits the outlines of the mission statement, evaluating the rarity and authenticity of the provenance, and its value for future generations.

4. Processing/Indexing (Micro-Appraisal)

- Processing the sources individually or at the folder/document level, including indexing them and updating the finding aid.
- Sources may be discarded on the grounds of privacy concerns/irrelevance.

Shared Concerns

Digitization is seen as democratization of the archive. But we must think critically about what we are digitizing. Digitization programs could reinforce cultural stereotypes and canonicities.What are the criteria used to select manuscripts?

- Andrew Prescott, “Why do we Digitize? The Case for Slow Digitization”

Shared Concerns

“It seems clear that the archive is primarily the product of a judgment, the result of the exercise of a specific power and authority, which involves placing certain documents in an archive at the same time as others are discarded. The archive, therefore, is fundamentally a matter of discrimination and of selection, which, in the end, results in the granting of a privileged status to certain written documents, and the refusal of that same status to others, thereby judged ‘unarchivable.' The archive is, therefore, not a piece of data, but a status."

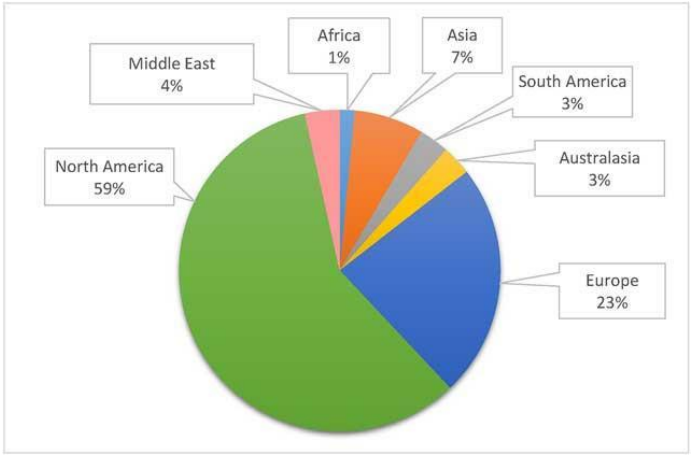
-Achille Mbembe, ``The Power of the Archive and its Limits"

Model 1: Consortium Models

“Ideally, we would have area specialists for each region”

- Internet Archive Archivist

- 1980s onwards → Consortium Agreements
- HathiTrust, Ivy Plus, OhioLINK
- International Coalition of Library Consortia (ICOLC)
- Human Genome Project
- Minimize overlap in collection efforts
- Open access to wider users
- Share resources such as human expertise



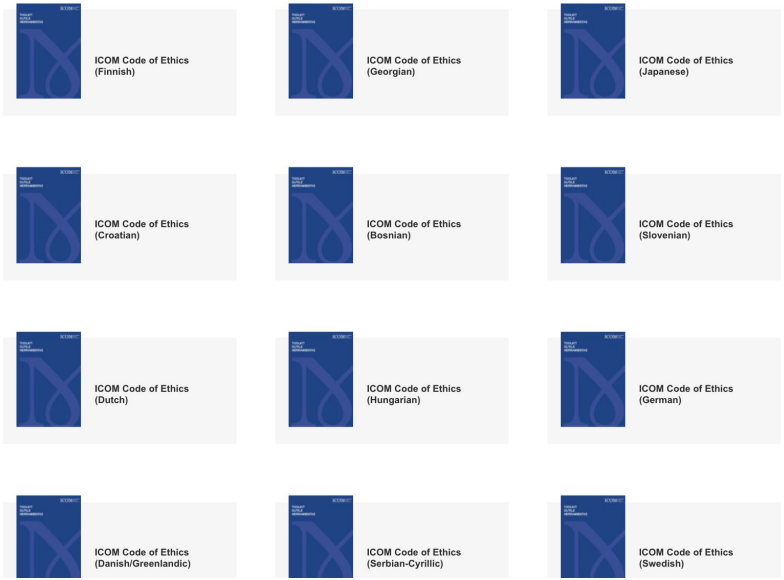
Regional distribution of ICOLC member consortia (Feather 2015)

Model 2:
Participatory Archives/ Community Archives

- Allowing locals/grassroots to define their own collections
- Oral Histories
- Local Infrastructure
- Preservation Outreach (eg. Community History Toolkit)
- Mukurtu: Tool to Provide Access to Native Languages/Cultures - “using their own protocols”, “manage and share their digital cultural heritage in their own way”

Model 3:
Codes of Conduct/Acquisition Policy

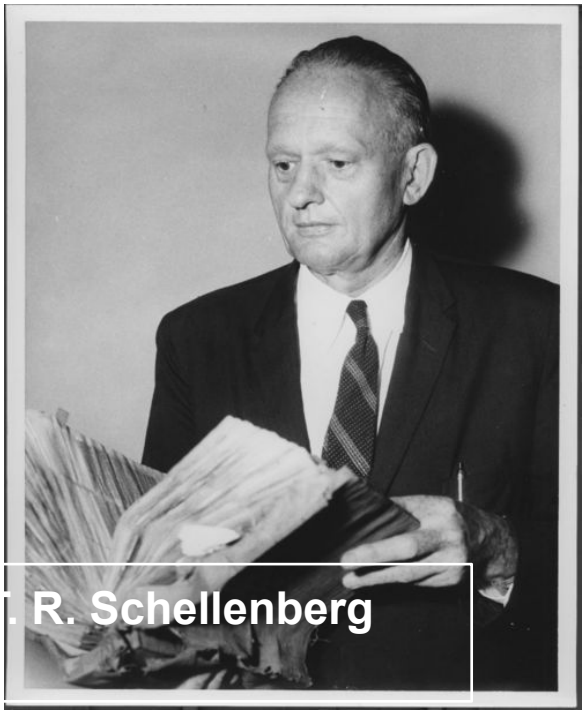
- International Council of Museum (ICOM): Code of Ethics for Museums
- “...minimum standards of professional practice and performance...”
- International Council on Archives (ICA): Code of Ethics
- Ethics, Privacy, Preservation, Access Standards
- “Right to be Forgotten”: European Union General Data Protection Regulation 2016
- Society for American Archivists (SAA)



Model 4: Appraisal

- 5 Major Concepts & Methodologies

1. "Unbiased" Record Creator
2. Organizational History
3. Patterns of history/Speculation about Future Usage
4. Professional Archivists
5. Archival Darwinism: Random or Natural Selection



F. Gerald Ham, *Selecting and Appraising Archives and Manuscripts*

Appraisal Report Example

Background

Description

Evaluation

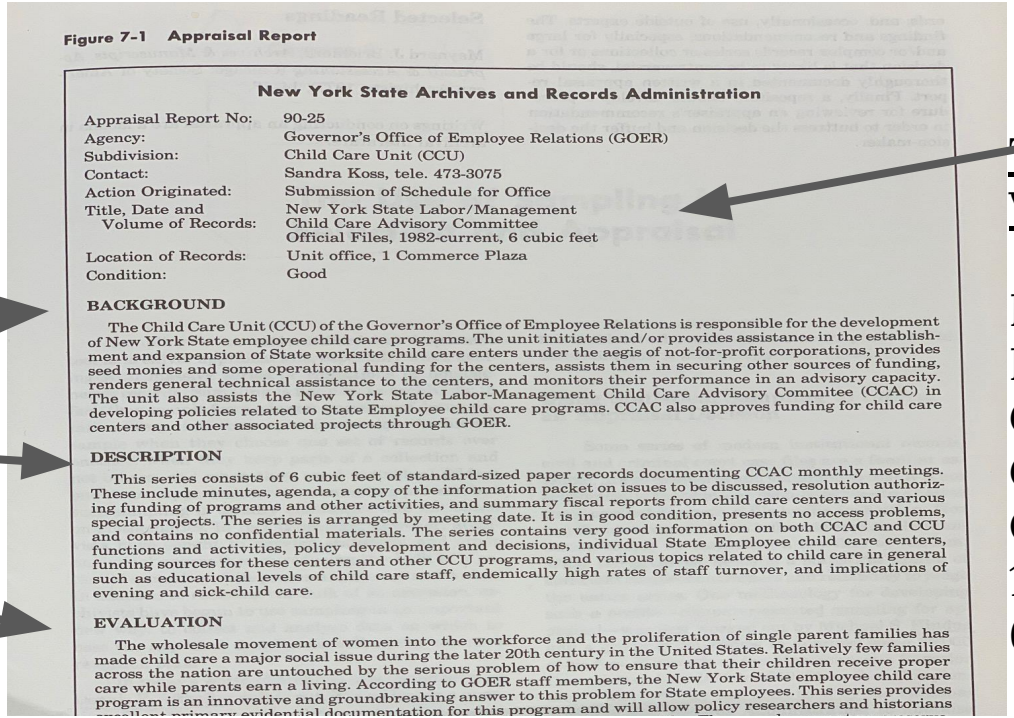


Figure 7-1 Appraisal Report

New York State Archives and Records Administration

Appraisal Report No:	90-25
Agency:	Governor's Office of Employee Relations (GOER)
Subdivision:	Child Care Unit (CCU)
Contact:	Sandra Koss, tele. 473-3075
Action Originated:	Submission of Schedule for Office
Title, Date and Volume of Records:	New York State Labor/Management Child Care Advisory Committee Official Files, 1982-current, 6 cubic feet
Location of Records:	Unit office, 1 Commerce Plaza
Condition:	Good

BACKGROUND

The Child Care Unit (CCU) of the Governor's Office of Employee Relations is responsible for the development of New York State employee child care programs. The unit initiates and/or provides assistance in the establishment and expansion of State worksite child care enters under the aegis of not-for-profit corporations, provides seed monies and some operational funding for the centers, assists them in securing other sources of funding, renders general technical assistance to the centers, and monitors their performance in an advisory capacity. The unit also assists the New York State Labor-Management Child Care Advisory Commitee (CCAC) in developing policies related to State Employee child care programs. CCAC also approves funding for child care centers and other associated projects through GOER.

DESCRIPTION

This series consists of 6 cubic feet of standard-sized paper records documenting CCAC monthly meetings. These include minutes, agenda, a copy of the information packet on issues to be discussed, resolution authorizing funding of programs and other activities, and summary fiscal reports from child care centers and various special projects. The series is arranged by meeting date. It is in good condition, presents no access problems, and contains no confidential materials. The series contains very good information on both CCAC and CCU functions and activities, policy development and decisions, individual State Employee child care centers, funding sources for these centers and other CCU programs, and various topics related to child care in general such as educational levels of child care staff, endemically high rates of staff turnover, and implications of evening and sick-child care.

EVALUATION

The wholesale movement of women into the workforce and the proliferation of single parent families has made child care a major social issue during the later 20th century in the United States. Relatively few families across the nation are untouched by the serious problem of how to ensure that their children receive proper care while parents earn a living. According to GOER staff members, the New York State employee child care program is an innovative and groundbreaking answer to this problem for State employees. This series provides excellent primary evidential documentation for this program and will allow policy researchers and historians

<u>Title, Date, and Volume of Records:</u>

New York State Labor/Management Child Care Advisory Committee Official Files, 1982-current, 6 cubic feet

Model 5: Cultural/Demographic Surveys & Guidelines

"We don't know what we don't know"

- Internet Archive Archivist

- Uneven Digital Access
- eg. Native American groups running their own radio stations
- Cultural Ban/Taboo
- Critical Languages
- Organizations eg. Lighthouse for the Blind

Table 1

	N	%	Cumulative downwards %	Cumulative upwards %
more than 100 million	8	0.13		99.9
10–99.9 million	72	1.2	1.3	99.8
1–9.9 million	239	3.9	5.2	98.6
100,000–999,999	795	13.1	18.3	94.7
10,000–99,999	1,605	26.5	44.8	81.6
1,000–9,999	1,782	29.4	74.2	55.1
100–999	1,075	17.7	91.9	25.7
10–99	302	5.0	96.9	8.0
1–9	181	3.0	99.9	

David Crystal,
Language Death (2000)

Transparency and Inclusivity:

Mission Statements, Collection Policies, Community Archives

- Setting data collection agenda based on digital availability inevitably produces misrepresentation of data
- Left without active management of data composition, these methods can lead to poor or limited scope in demographics
    - Eg. what kind of users use Reddit?
    - Resulting models are heavily reliant on the specific source (ie. all of this was trained on Reddit data thus reflecting all users of Reddit).

- Archives have faced similar critiques of exclusivity. Traditional archives had focused on state and government documents, to preserve the documents of the governing and social elites

- To further inclusivity and data collection from the long tail, archives have expanded mediums for data input through

  - active collection policies

  - prioritizing data collection from minority groups

  - democratizing the data collection process via community archives.

- Rather than starting with data sets by availability, data collection in archives starts with a statement of concept or topics that often signal a commitment to collecting information about underrepresented groups.

- Community archives are projects of data or document collection in the ownership of the group that is being represented.
    - Projects such as historypin provide platforms for local communities to define and contribute their own cultural and heritage collections.
    - Widen the channel of user input in data collection.
- While many of the initiatives have grassroots beginnings, foundations and institutions have been actively funding and promoting archiving in the periphery.

- This model of decentralization also enables minority groups to consent to and define their own categorization.

- Some cultures demand non-Western systems of representation.

- Mukurtu is an example of a content management system built to allow indigenous communities to house their own materials “using their own protocols.”

Power: Data Consortia

- To increase parity in data ownership, archival and library sciences have developed a consortial model.

- Consortia have several mutual benefits for participating groups the main being the ability to gain economies of scale.

- Groups of libraries can make expensive purchases such as subscriptions to academic journals, pool resources for large scale projects

Ethics & Privacy: Codes of Ethics and Conduct

- Many organizations (e.g. Partnership in AI, IEEE Global Initiative on Ethics of Autonomous & Intelligent Systems)
- But these measures have yet to address the ethics associated with data collection let alone provide enforcement mechanisms
- Handling human information exposes archivists to various ethical dilemmas:
  - selecting which documents to toss or keep
  - granting access to sensitive content
  - dealing with intellectual property
- Overlapping organizations across archives, libraries, and museums independently have codes of ethics and conduct.

Ethics & Privacy: Codes of Ethics and Conduct

- Enforcing ethical regulation in data collection in ML faces challenges because of the lack of an incentive mechanism for researchers and practitioners

- Most archivists are full-time professional data collectors.

- **Archives work by a membership system whereby breaching the code of conduct could result in losing professional membership (ARA 2018).**

- Many suborganizations of archival and records collectors have ethics panels or committees that evaluate each alleged violation case by case (ARA 2018; ICRM 2016).

Conclusion

Societal biases enter when we

- Formulate what problems to work on
- Collect training and evaluation data
- Architect our models and loss functions
- Analyze how our models are used in society

Conclusion

Our papers, classes and incentive structures only focus on

- **Architect our models and loss functions**

Conclusion

We need to learn from other disciplines who have looked at data much more rigorously & work with them to draw lessons for ML.