Practical Guide to Ethical Data Science

What are we learning?

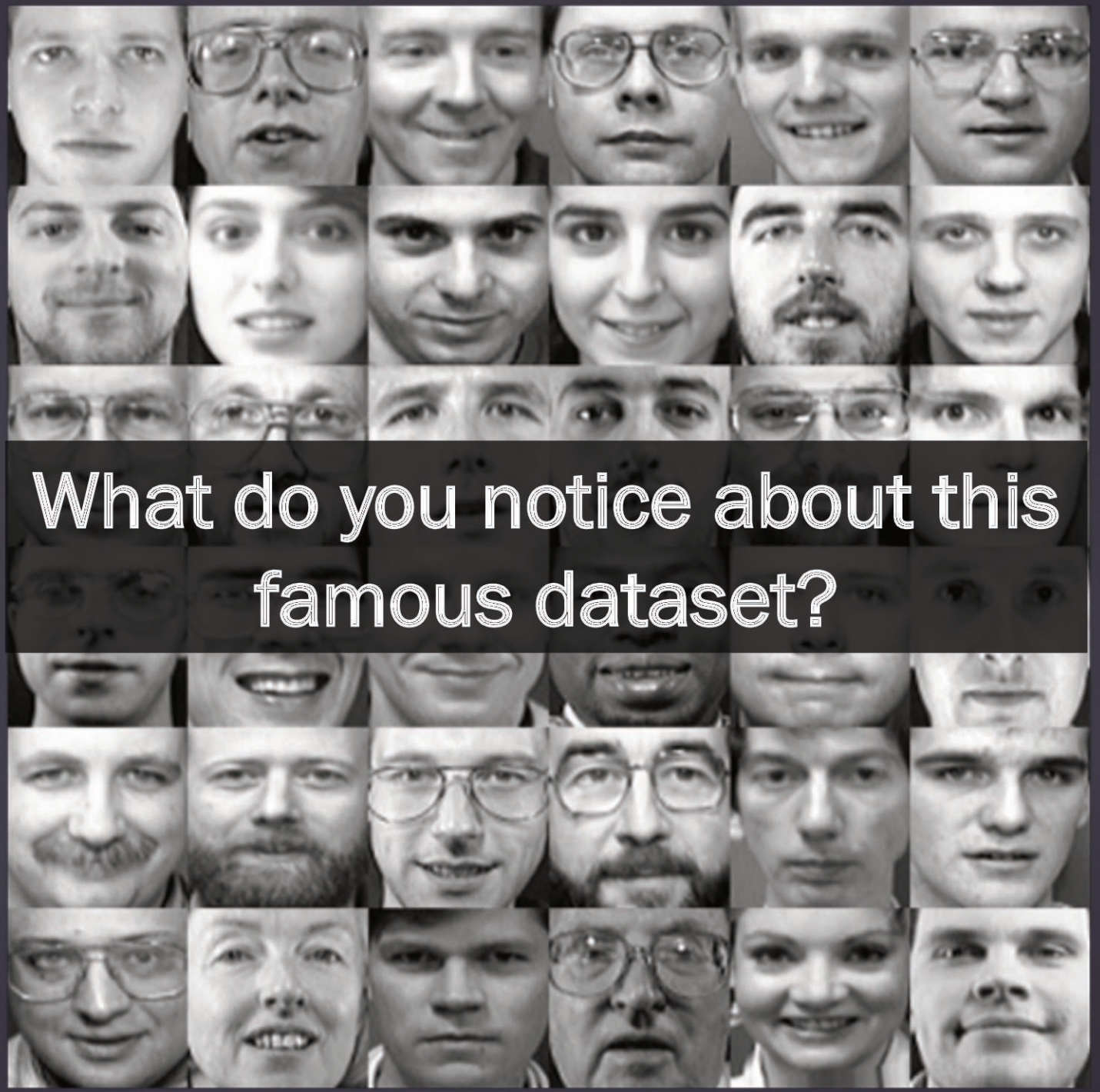
Not a rulebook

- Questions to ask
- Frameworks
- Case studies

Why?

- Data Science has a history of harm
- Ethics come before bottom line
- Field of Data Science newly prioritizing ethics


DATASETS


What do you notice about this famous dataset?

Dataset bias

- Recall from our discussion of unbalanced classes that relative class frequency in our datasets can cause problems
- Many models use class frequency to:
  - Determine prior probability of that class: Infrequent classes in training are assumed to be less likely in test
  - Determine the “emphasis” the model places on each class: This why we sometimes use class weights or sampling to emphasize minority classes

Dataset bias

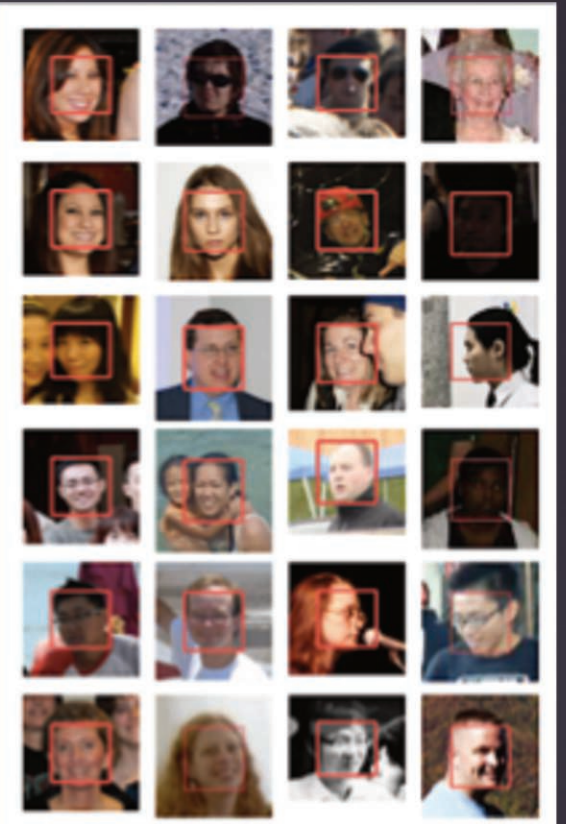
- Many existing face datasets are built using existing face recognition models
- Megaface: largest publicly available set of facial images (1 Million)
  - Began with the 100M photo Flickr dataset
  - Applied state-of-the-art Head-Hunter model to find images with faces

Dataset bias

The MegaFace Benchmark: 1 Million Faces for Recognition at Scale

Ira Kemelmacher-Shlizerman Steve Seitz Daniel Miller Evan Brossard
Dept. of Computer Science and Engineering
University of Washington

Face processing. We downloaded the highest resolution available per photo. The faces are detected using the Head-Hunter[2] algorithm by Mathias et al. [20], which reported state of the art results in face detection, and is especially robust to a wide range of head poses including profiles. We crop detected faces such that the face spans 50% of the photo height, thus including the full head (Fig. 3). We further estimate 49 fiducial points and yaw and pitch angles, as computed by the IntraFace[3] landmark model [34].



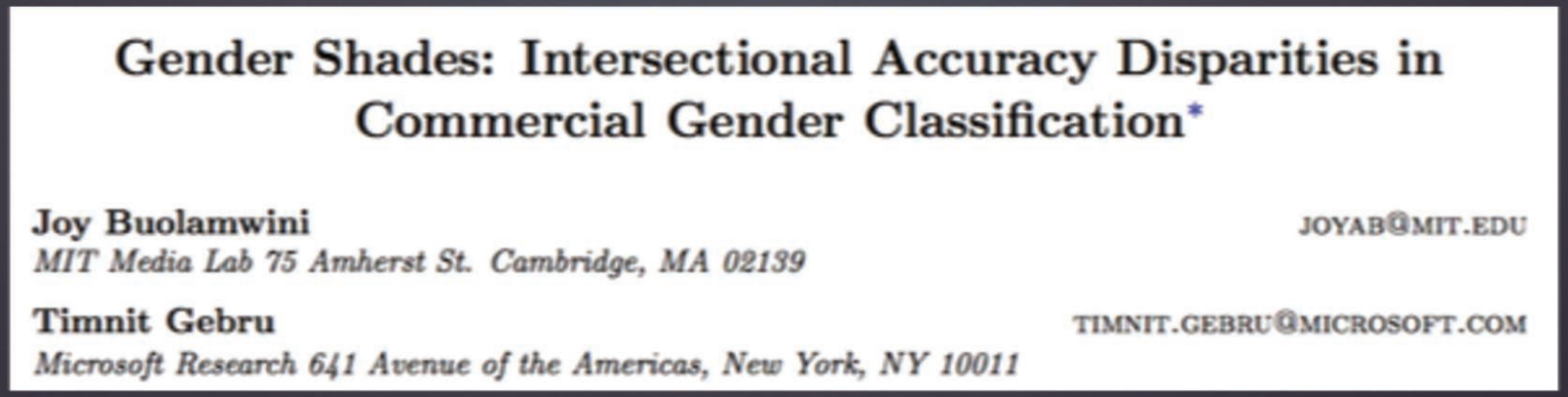
Random sample of MegaFace photos (with provided detection boxes in red)

Dataset bias

- Many existing face datasets are built using existing face recognition models
- Megaface: largest publicly available set of facial images (1 Million)
  - Began with the 100M photo Flickr dataset
  - Applied state-of-the-art Head-Hunter model to find images with faces
- Does this raise any concerns?

Assessment of dataset/model bias

- Annotate existing datasets for gender and skin color
- Use annotations to build carefully balanced dataset
- Balanced dataset can
  1. Train models with fewer problems of bias
  2. Evaluate existing models for bias



Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini JOYAB@MIT.EDU
MIT Media Lab 75 Amherst St. Cambridge, MA 02139

Timnit Gebru TIMNIT.GEBRU@MICROSOFT.COM
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

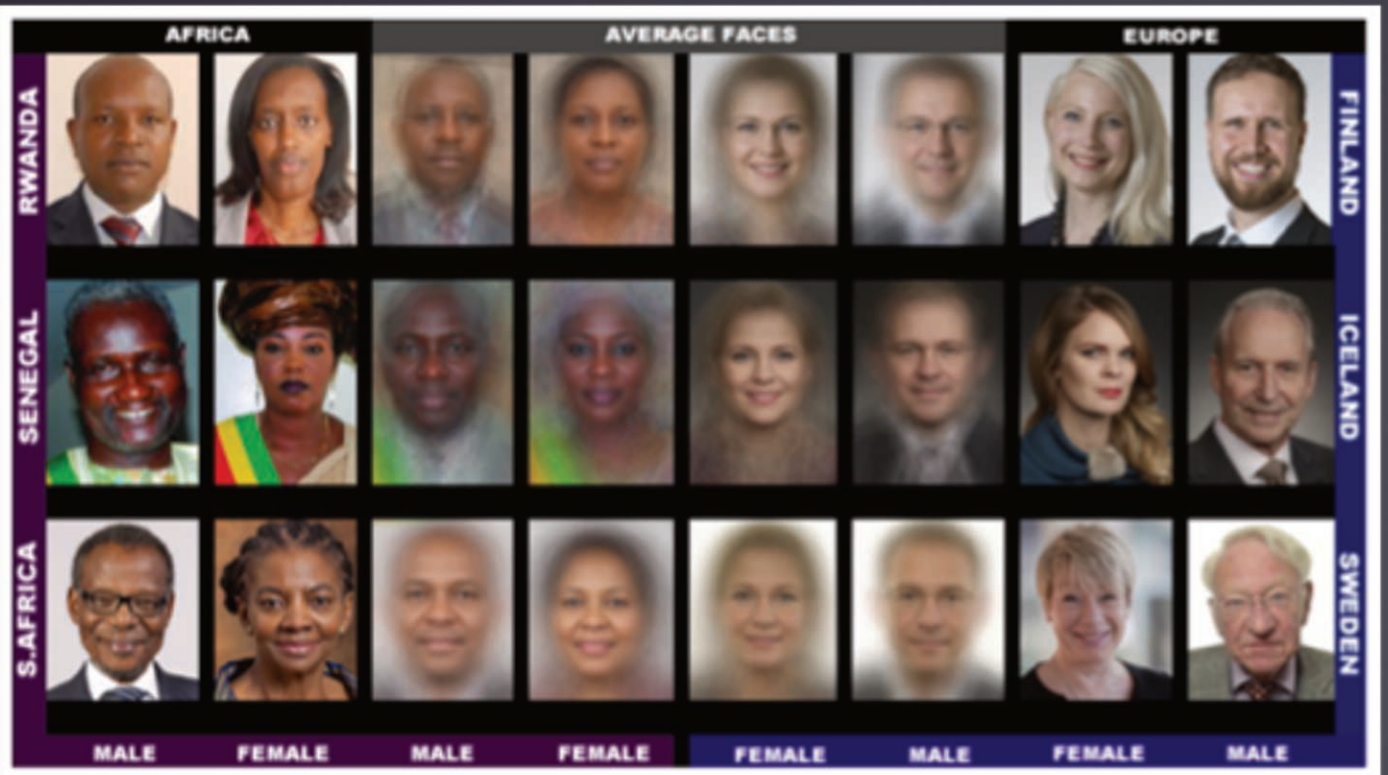


Figure 1: Example images and average faces from the new Pilot Parliaments Benchmark (PPB). As the examples show, the images are constrained with relatively little variation in pose. The subjects are composed of male and female parliamentarians from 6 countries. On average, Senegalese subjects are the darkest skinned while those from Finland and Iceland are the lightest skinned.

Figure 3: The percentage of darker female, lighter female, darker male, and lighter male subjects in PPB, IJB-A and Adience. Only 4.4% of subjects in Adience are darker-skinned and female in comparison to 21.3% in PPB.

Assessment of model bias

Existing off-the-shelf facial recognition models

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

Table 4: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-PPV), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the PPB dataset. All classifiers have the highest error rates for darker-skinned females (ranging from 20.8% for Microsoft to 34.7% for IBM).

Assessment of model bias

women

men

Women with dark skin

Men with light skin

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	100
	Error Rate(%)	**6.3**	**10.7**	**2.6**	**12.9**	**0.7**	**20.8**	**6.0**	**1.7**	**0.0**
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	100	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	16.3	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	99.3	94.0	99.2
	Error Rate(%)	**10.0**	**21.3**	**0.7**	**16.5**	**4.7**	**34.5**	**0.7**	**6.0**	**0.8**
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	98.9	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	23.4	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	99.7
	Error Rate(%)	**12.1**	**20.3**	**5.6**	**22.4**	**3.2**	**34.7**	**12.0**	**7.1**	**0.3**
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	99.6	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	25.2	17.7	5.20	0.4

Table 4: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-PPV), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the PPB dataset. All classifiers have the highest error rates for darker-skinned females (ranging from 20.8% for Microsoft to 34.7% for IBM).

Assessment of bias

- First example of bias inheritance
  - Biased datasets (early face datasets) *used to train* →
  - Biased models (previous state of the art) *used to build* →
  - Biased datasets (current face datasets) *used to train* →
  - Biased models (current state of the art and off the shelf)
- Balancing datasets is hard work
- A well balanced dataset can be used to
  - Train better models
  - Evaluate existing models


Consent

Consent

- Consent: meaningful approval by those affected
- Important for
  - Gathering data
  - Those using a model

Consent: Questions

- Do I need to ask for consent?
- How should I ask for consent?
  - Weakest: Hide ”we’re gathering your data” in the EULA no one reads
  - Weak: Setting in app that defaults to “on” and controls whether data is gathered. No user prompt
  - Strong: Periodic checks that user understands what data is being gathered and approves
- What do I do if a user doesn’t or revokes consent?
  - Weakest: Users who don’t consent may not use the app and all existing data is retained
  - Right to be forgotten: Policy that allows users to revoke consent to use their data, *after the data has been gathered*

Consent: Answering questions

What's the burden of consent? How important is it to make sure those affected are consenting?

1. Will users be negatively affected?
   - How significant is the effect?
   - How likely is it to cause harm?
2. Do users know what they're getting into?
   - Is there any chance that a user will be surprised to learn of what you're doing?
3. Can users withdraw consent at any time?

Consent: Answering questions

A project uses a well-known and publicly maintained dataset. The product will be an article informing readers on a topic. What's the burden of consent?

Consent: Answering questions

A project uses a well-known and publicly maintained dataset. The product will be an article informing readers on a topic. What's the burden of consent?

- Very low burden of consent

Consent: Answering questions

Consider a project that scrapes one of the largest data sites for personal answers to questions about dating. The product will be a (somewhat anonymized) dataset packaging these responses for other data scientists to use.

- What's the burden of consent here?

Case study: OKCupid dataset

3.1 *Which data was collected*

We did not gather all the possible data about the users. Specifically, we gathered the following datapoints:

- Profile information: username, age, gender(s), location, religion-related opinions, astrology-related opinions, interested in, number of photos, etc. (36 variables)
- OKCupid personality scales (50 variables)
  - These are personality dimensions that OKCupid calculates automatically. No information is given about how they are calculated as far as we know.
- Answers given to the top 2600 questions on the site.

Data we did not gather include:

- The profile text.
- The profile photos.
- Explanations given to chosen answers.
- Profile height.

Gathering the photos would have taken up a lot of hard drive space but could be done in a future scraping. Be advised that scraping and releasing users' photos may be illegal due to copyright or privacy laws. The other data were not collected because we forgot to include them in the scraper.

After collecting the data, they were processed in R to create one large datafile. Either during the data collection or the processing, some mistakes were made that left some variables corrupted. This left of total of 2541 questions, 50 personality scales and 29 variables related to each profile with uncorrupted data. The corrupted variables were the variables that had the lowest response rates (N<100), so they were nearly useless for analysis anyway.

Due to privacy concerns (Hackett, 2016), the username and city variables were removed from the published version of the dataset.

Weapons of Math Destruction

What is a Weapon of Math Destruction?

- Opacity: Can a model be inspected?
  - Black boxes have more potential for harm than those we can inspect and interpret
- Checks: Who makes sure a model doesn't harm others?
  - Can be internal or external. Models without checks have more potential for harm
  - Institutional Review Board: common in some academic fields
- Scale: Who does your model affect?
  - Models that affect many have more potential for harm. Especially dangerous are models that are used outside of the original intended purpose

What is a Weapon of Math Destruction?

- Proxies are a common source of problems
- Many of the things we want to affect are difficult or impossible to measure:
  - Customer satisfaction
  - Political beliefs
  - …
- When we can't measure something directly, one solution is to use a proxy
  - Customers who return soon are likely to be satisfied and we can measure that.

Case Study: Loans

I work at a bank where loan officers currently decide if each loan is approved or not. We want to build an algorithm to determine if someone is a good candidate for a loan. I have a table with our past loan applicants, the outcome of the application, and other features.

We don't have a ground truth for who should get a loan. How should I select (or engineer) a proxy for that?

Case Study: Crime

We want to predict areas where crimes are most likely to occur so that our police force can be used efficiently. Are arrest locations a good proxy for estimating crime?

Case Study: Video

I run a video site whose revenue comes from ads. So we're more profitable the more time users spend watching videos.

I'm building a recommender but I don't have a ground truth for what users want to watch next. Which features should I select (or engineer)?

Case Study: Video

Which feature do I select/engineer for our recommender proxy?

- Likes: Predict if a user will like the next video
- Next viewtime: Predict the amount of time the user will spend on next video
- Session viewtime: Predict the amount of time the user will spend on this session (across multiple videos)

YouTube doesn't tell us what proxies they use for recommendation, but it seems that something like to session viewtime is highly weighted

- Possible problem: What if videos that keep users watching are bad for users?

Case Study: Video


THE WALL STREET JOURNAL.

Subscribe Now | Sign

U.S. Edition | November 8, 2018 | Today's Paper | Video

World U.S. Politics Economy Business Tech Markets Opinion Life & Arts Real Estate WSJ. Magazine

TECH

How YouTube Drives People to the Internet's Darkest Corners

Google's video site often recommends divisive or misleading material, despite recent changes designed to fix the problem


WIRED

Long Reads

Children's YouTube is still churning out blood, suicide and cannibalism

Children's search terms on YouTube are still awash with bizarre and sometimes disturbing bootleg content. Can anything be done to stem the tide?

By K.G ORPHANIDES

23 Mar 2018

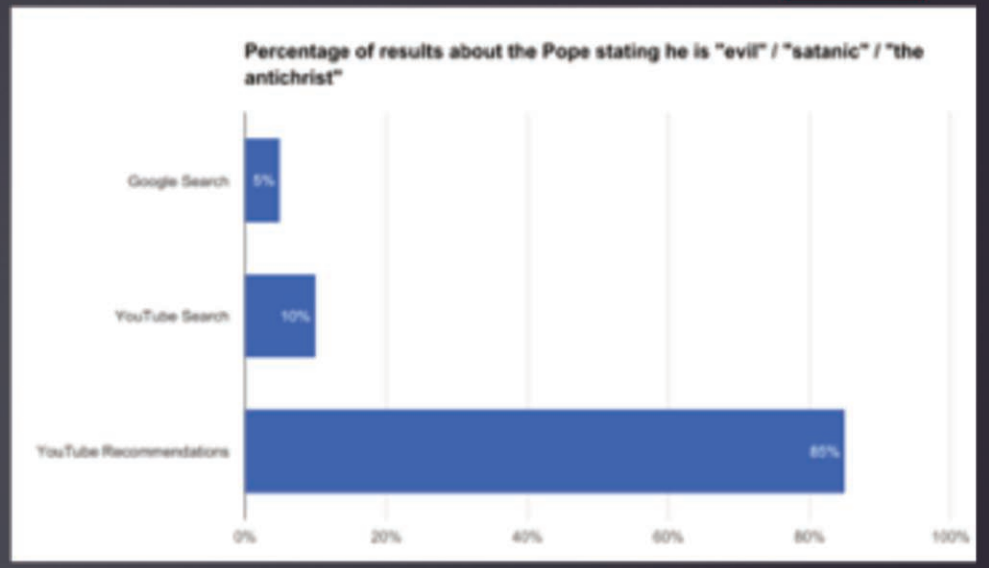

Bloomberg Businessweek

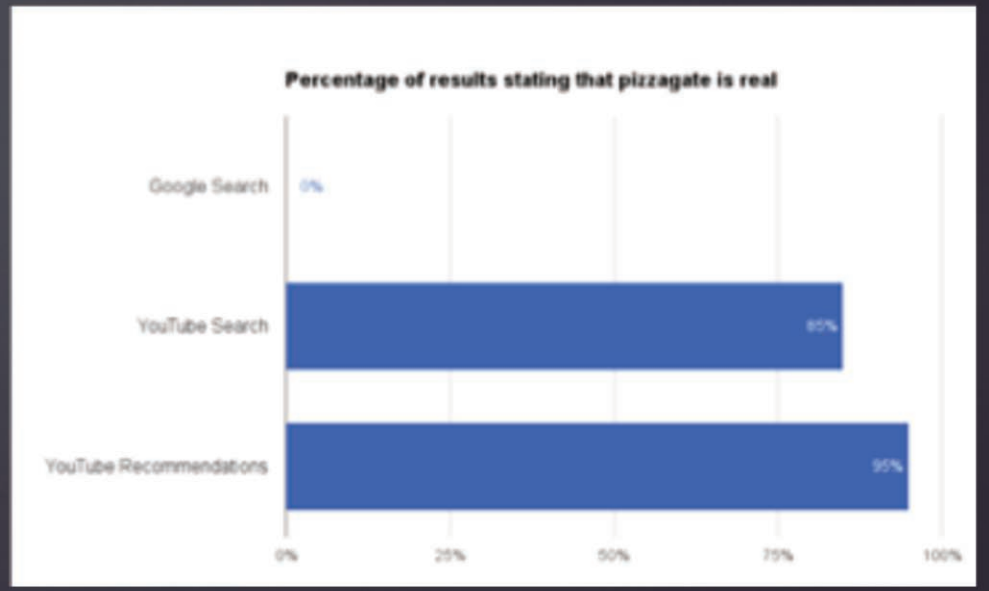
YouTube's Plan to Clean Up the Mess That Made It Rich

Extremist propaganda, dangerous hoaxes, videos of tasered rats —the company is having its worst year ever. Except financially.

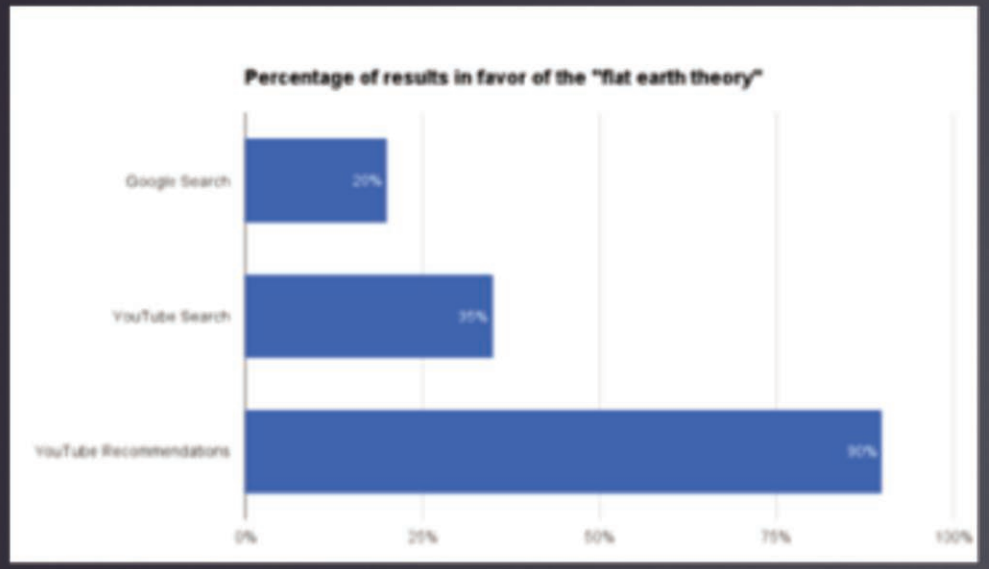
YouTube recommendations

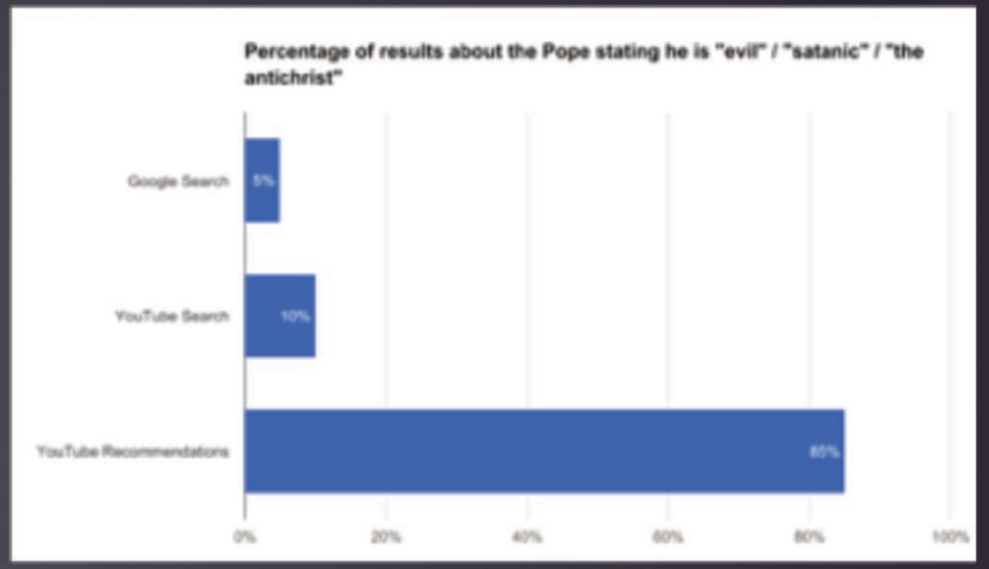
- This is research from Guillaume Chaslot, a former Google engineer who built a scraper to measure Youtube recommendations
- The researchrs entered queries like “Pope” and annotated the fraction of results that were conspiracy theories

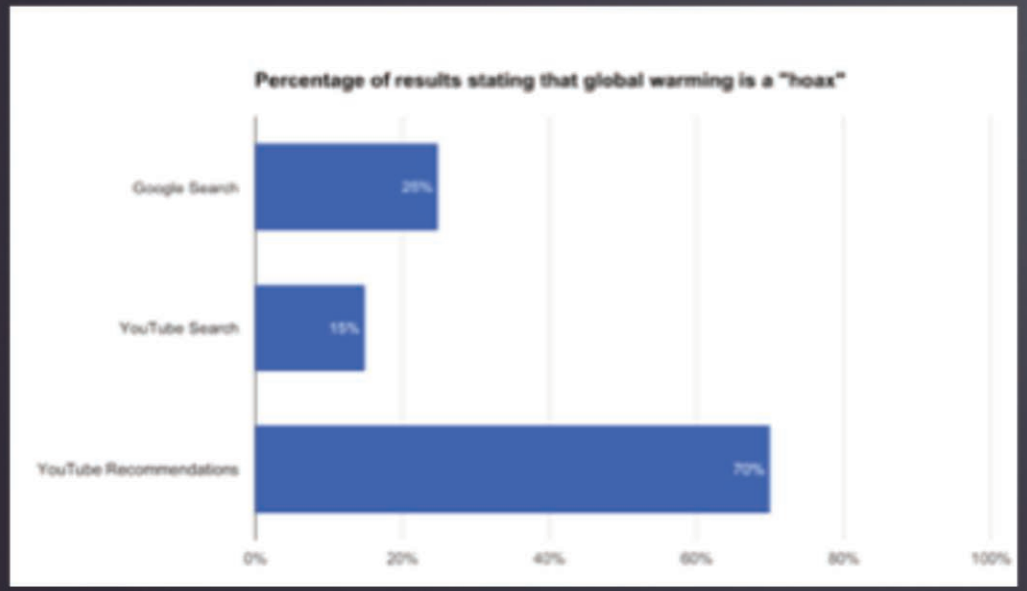


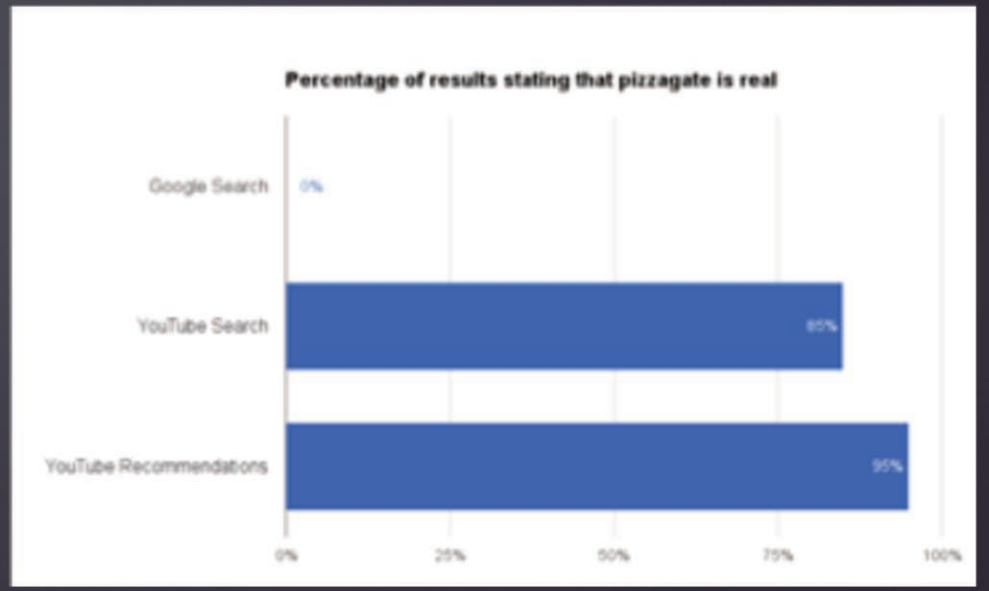


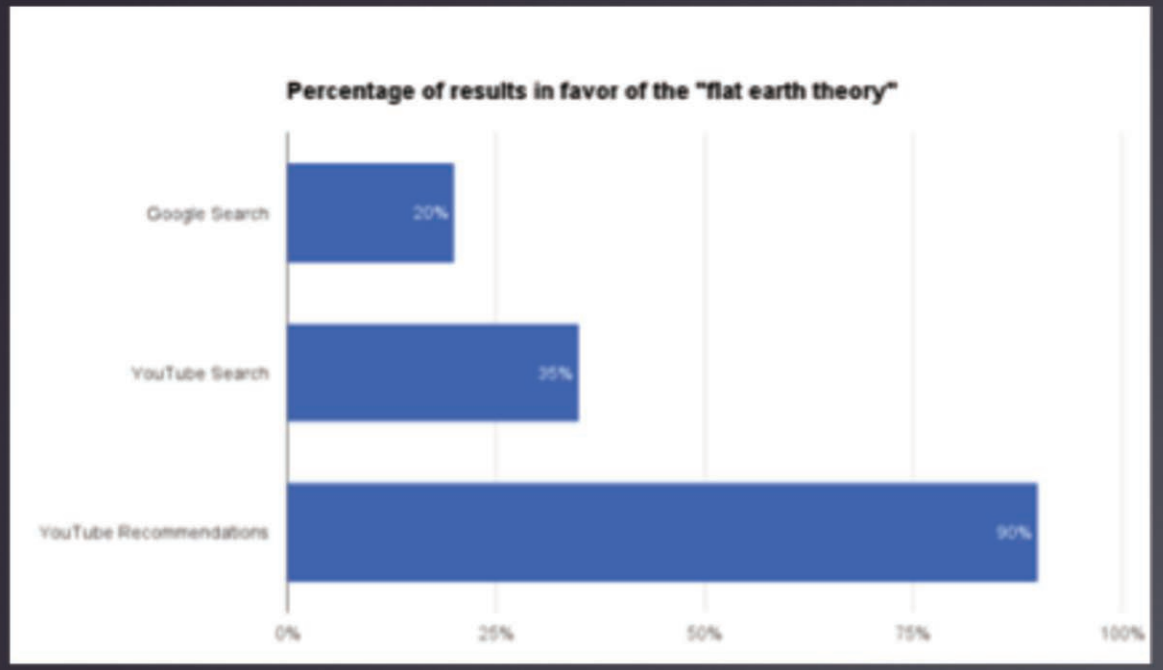
YouTube recommendations

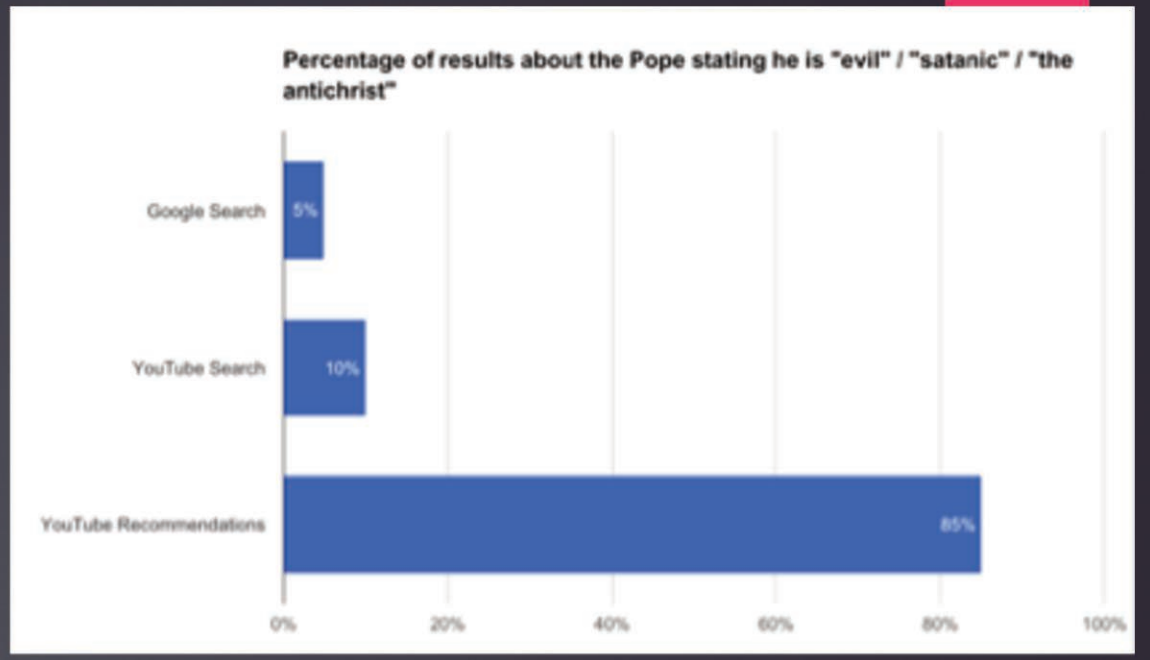


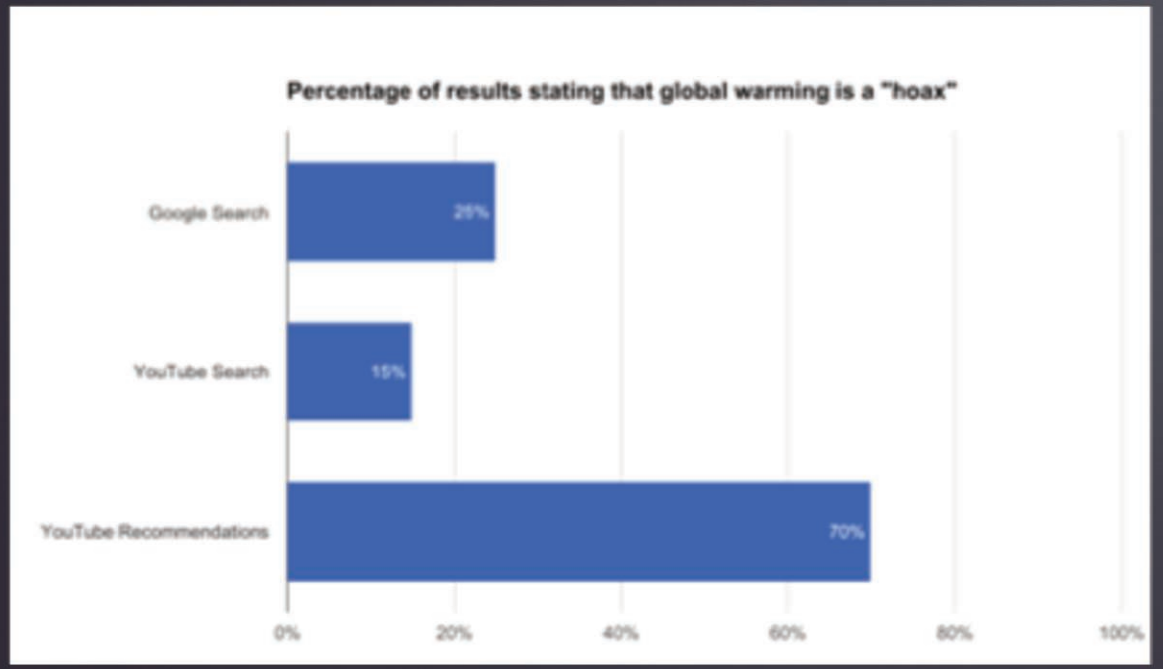


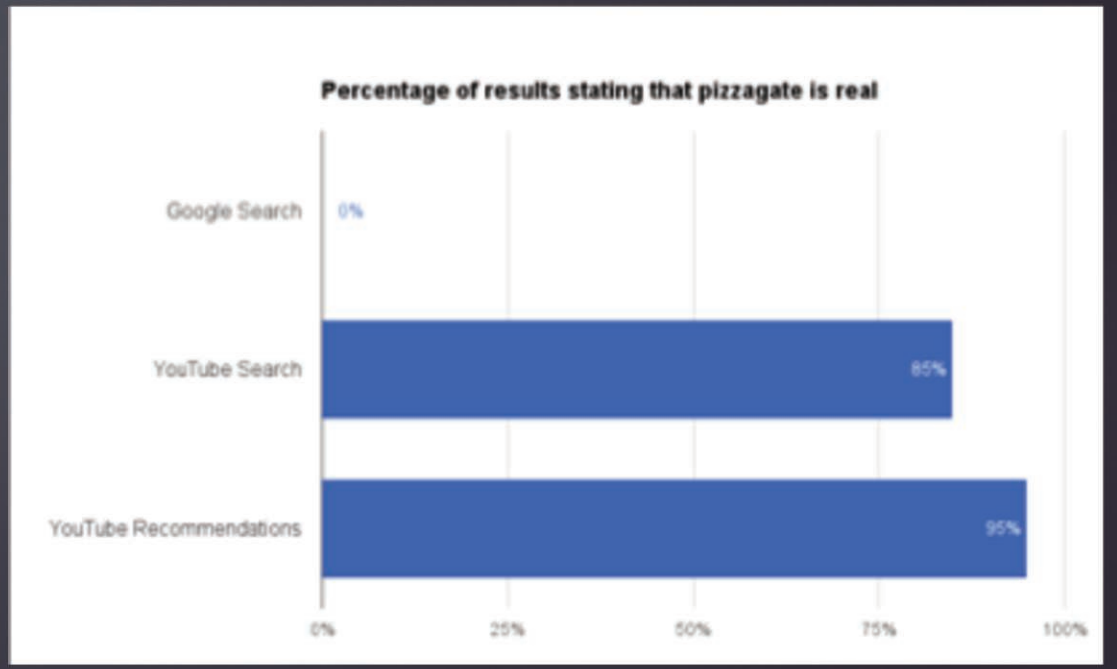





Percentage of results in favor of the "flat earth theory"
Google Search
20%
YouTube Search
35%
YouTube Recommendations
90%
0%
25%
50%
75%
100%


Percentage of results about the Pope stating he is "evil" / "satanic" / "the antichrist"
Google Search
5%
YouTube Search
10%
YouTube Recommendations
85%
0%
20%
40%
60%
80%
100%


Percentage of results stating that global warming is a "hoax"
Google Search
25%
YouTube Search
15%
YouTube Recommendations
70%
0%
20%
40%
60%
80%
100%


Percentage of results stating that pizzagate is real
Google Search
0%
YouTube Search
85%
YouTube Recommendations
95%
0%
25%
50%
75%
100%

Advice

- Regularly audit your assumptions and the effects of your models
- Customer interviews
- Avoid bias inheritance
  - Using past biased data or models to build new ones
- Give users options instead of forcing the use of your product
- Be extremely skeptical

Where to look

- Algorithmic Violence – Mimi Onuoha
- Gender Shades - Joy Buolamwini
- Weapons of Math Destruction – Cathy O'Neil
- AlgoTransparency - Guillaume Chaslot

François Chollet
@fchollet

Replying to @fchollet

Our challenge is thus to design information filtering and distillation technology that allow us to retain our agency, that empower us with greater control over our lives rather than manipulating us. More like Wikipedia and Google Search, and less like Facebook.

7:29 PM - Jan 20, 2018

297 88 people are talking about this