Data Is People:

Ethical Considerations in Data Collection & Use

Casey Fiesler

Department of Information Science, CU Boulder

Information Science
UNIVERSITY OF COLORADO **BOULDER**

@cfiesler

stuff I do

- online communities & law/governance
- TOS and privacy policies
- technology ethics education
- research ethics
- public perceptions of ethics and of research

things I'm going to talk about

- some* ethical controversies you've heard about
- a study about data scraping & Terms of Service
- a study about the ethics of Twitter research
- my research ethics hill to die on
- my ethics education hill to die on

* every time I give a talk like this there's more.

Cambridge Analytica The Cambridge Analytica Files

Facebook gave data about 57bn friendships to academic

Leaked: Cambridge Analytica's blueprint for Trump victory

Exclusive: Former employee explains how presentation showed technique

Mark Zuckerberg takes out full-page newspaper ads to say 'sorry' for Cambri Analytica scandal

By Nicole Darrah | Fox News

ADAM ROGERS SCIENCE 03.25.18 07:00 AM

THE CAMBRIDGE ANALYTICA DATA APOCALYPSE WAS PREDICTED IN 2007

Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach

Cambridge Analytica active in elections, big data projects for years

KIM HJELMGAARD | USA TODAY
Updated 2:58 p.m. EDT Mar. 22, 2018

The shady data-gathering tactics used by Cambridge Analytica were an open secret to online marketers. I know, because I was one

Market researchers have used these tricks for years

By Alexandra Samuel | Mar 25, 2018, 1:19pm EDT

Forbes / Washington

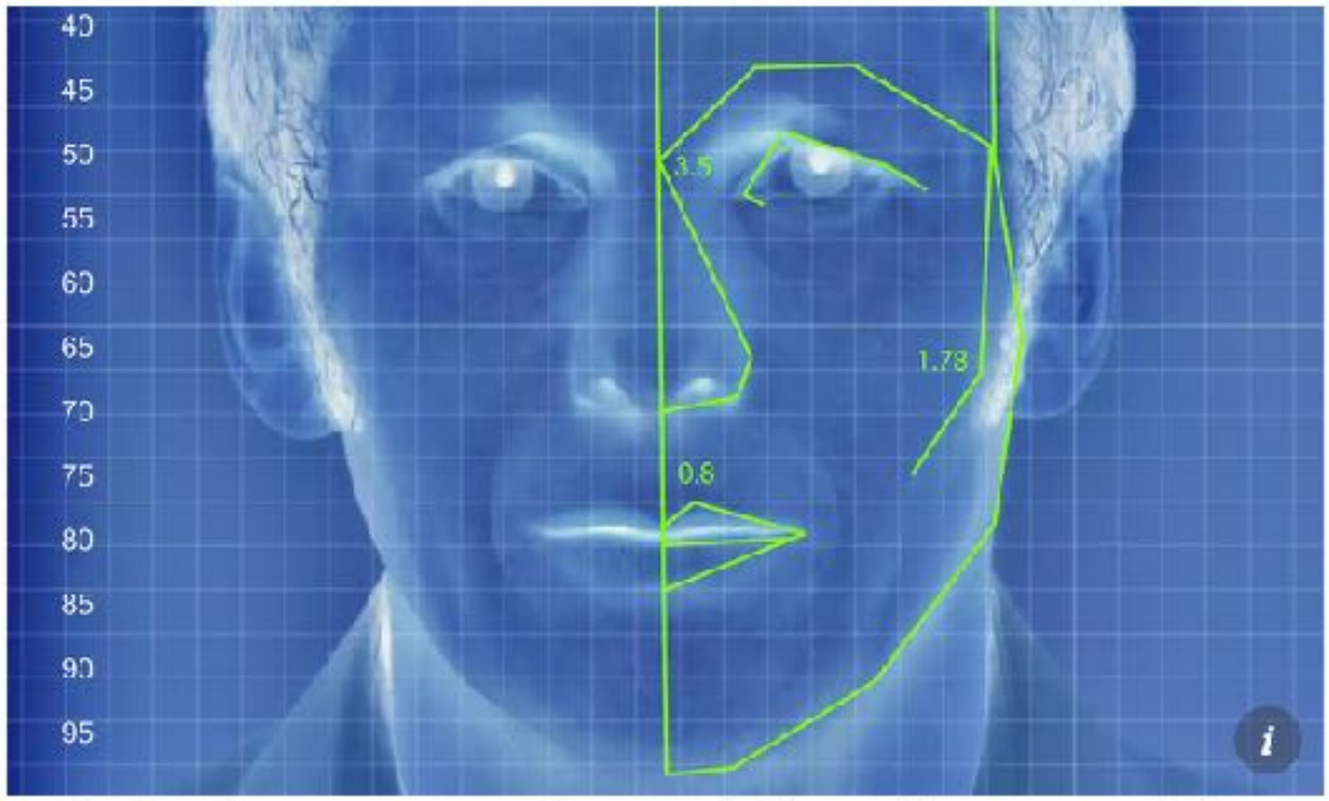
The Little Black Book of

JUN 28, 2014 @ 01:10 PM 79,834 VIEWS

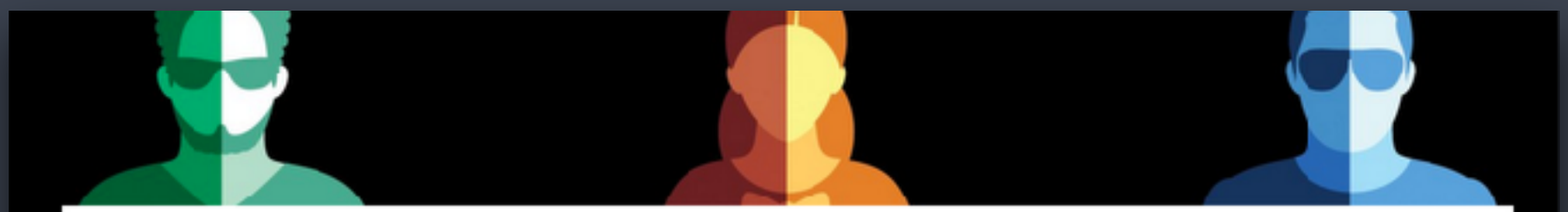
Facebook Manipulated User News Feeds To Create Emotional Responses

New AI can guess whether you're gay or straight from a photograph

An algorithm deduced the sexuality of people on a dating site with up to 91% accuracy, raising tricky ethical questions



Artificial intelligence can accurately guess whether people are gay or straight based on photos of their faces, according to new research that suggests machines can have significantly better "gaydar" than humans.



70,000 OkCupid Users Just Had Their Data Published

WRITTEN BY JOSEPH COX

May 12, 2016 // 01:44 PM EST

A student and a co-researcher have publicly released a dataset on nearly 70,000 users of the dating site OkCupid, including their sexual turn-ons, orientation, usernames and more. And critics say it may be possible to work out users' real identities from the published data.

The situation is raising questions about what type of data researchers should be allowed to collect en masse, repackage and perhaps distribute.

- What defines “public” data?
- Is consent sometimes appropriate for use of public data?
- Is it acceptable for researchers to violate TOS? If research breaks TOS, are there additional oversights needed to mitigate risk?
- How do we account for user expectations even if we find them unreasonable?
- How do we address the use of stolen, leaked, or deleted data?
- What are proper anonymization practices for online data?
- Should we have shared expectations of ethical oversight?
- What precautions should we take for studying vulnerable populations online?
- What are our obligations to the communities we study?

U.S. § 46.102

(f) **Human subject** means a living individual about whom an investigator (whether professional or student) conducting research obtains

(1) Data through intervention or interaction with the individual, or

(2) Identifiable private information.

How do you decide whether it is okay to collect data?

How do you decide whether it is okay to collect data?

- "Do the Terms of Service or other rules prohibit data collection?"
- "Is the data *public*?"

ABOUT BLOG FAQS

Unauthorized Activities

When using the Services, you agree not to:

Create a handle for the purpose of preventing others from using that handle.

Sell or buy handles.

Impersonate another person in a manner that is intended to or does mislead, confuse or deceive

Post or share another individual's private information without their express authorization and p

Defame, abuse, bully, harass, stalk, threaten, or otherwise violate the legal rights of others.

Use racially or ethnically offensive language.

Discuss or incite illegal activity.

Post or share Submissions that contain pornography or graphic violence.

Post or share anything that exploits children or minors or that depicts cruelty to animals.

Post or share Submissions that violate any third party right, including any copyright, trademark other intellectual property or proprietary right.

Disseminate any unsolicited or unauthorized advertising, promotional materials, 'junk mail', 'sp solicitation.

Use any robot, spider, crawler, scraper or other automated means to access the Services.

Take any action that imposes an unreasonable or disproportionately large load on our infrastruc

Use or develop any third party applications that interact with the Services or Submissions with

Alter the opinions or comments posted by others on the Services.

Data Collection Provisions in Terms of Service

You will not collect users' content or information, or otherwise access Facebook, using automated means (such as harvesting bots, robots, spiders, or scrapers) without our prior permission.

We prohibit crawling, scraping, caching or otherwise accessing any content on the Service via automated means, including but not limited to, user profiles and photos (except as may be the result of standard search engine protocols or technologies used by a search engine with Instagram's express consent).

Scrape or copy profiles and information of others through any means (including crawlers, browser plugins and add-ons, and any other technology or manual work);

Use FetLife to do any academic or corporate research without the expressed written consent of BitLove

Most data collection provisions are context-agnostic.

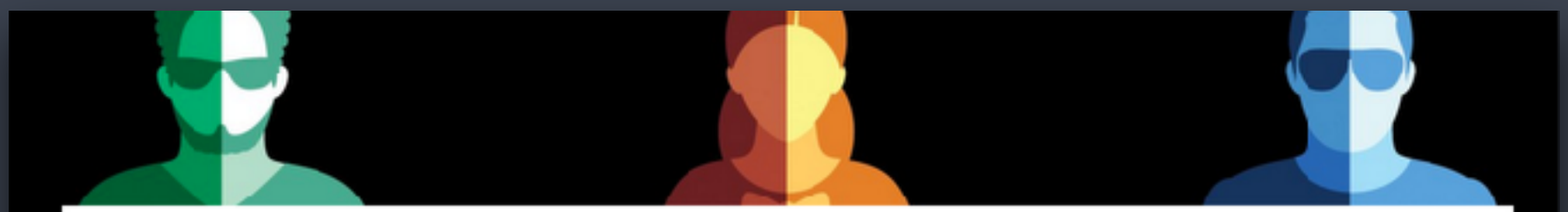
- What kind of data you're collecting
- What you're going to do with it
- Who you are
- What the expectations and norms of the users are

TOS-based decisions assume:

- Violating TOS is inherently unethical
- Violating TOS is the *only* thing that could make data collection unethical

How do you decide whether it is okay to collect data?

- "Do the Terms of Service or other rules prohibit data collection?"
- "Is the data *public*?"



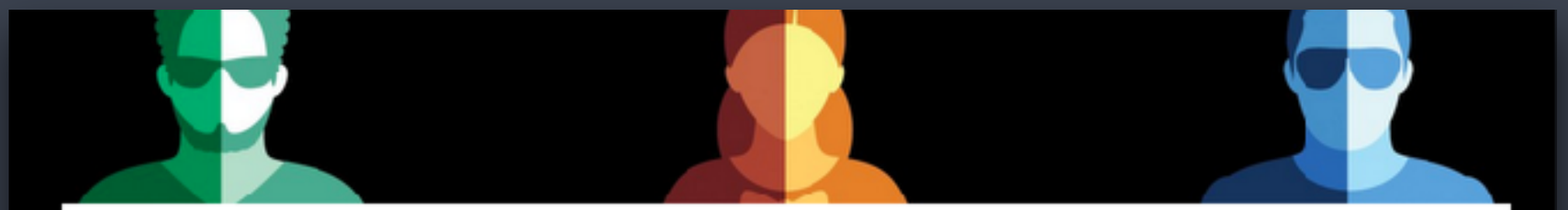
70,000 OkCupid Users Just Had Their Data Published

WRITTEN BY JOSEPH COX

May 12, 2016 // 01:44 PM EST

A student and a co-researcher have publicly released a dataset on nearly 70,000 users of the dating site OkCupid, including their sexual turn-ons, orientation, usernames and more. And critics say it may be possible to work out users' real identities from the published data.

The situation is raising questions about what type of data researchers should be allowed to collect en masse, repackage and perhaps distribute.



70,000 OkCupid Users Just Had Their Data Published

A student and a co-r
of the dating site Ok
and more. And critic
published data.

The situation is raisi
allowed to collect en

Emil OW Kirkegaard @KirkegaardEmil · May 8

The OKCupid paper has now been submitted. This means that the dataset is now public! Enjoy! :) openpsych.net/forum/showthre…

26 38

Ethan Jewett @esjewett · May 11

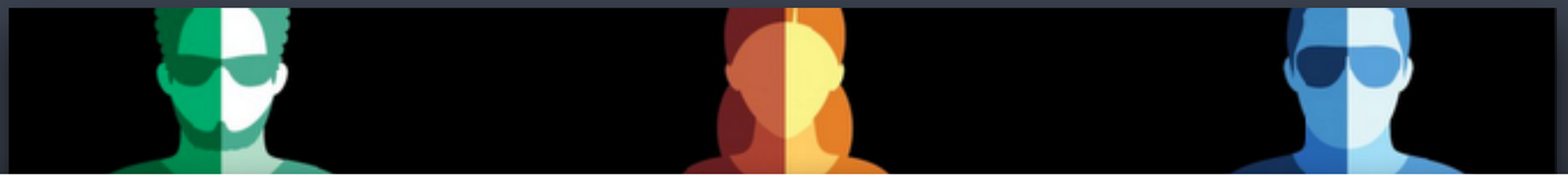
@KirkegaardEmil This data set is highly re-identifiable. Even includes usernames? Was any work at all done to anonymize it?

3 9

Emil OW Kirkegaard
@KirkegaardEmil

Follow

@esjewett No. Data is already public.



Scott B. Weingart
@scott_bot

Follow

With these details, I roughly estimate I could ~90% accurately connect sexual preferences & histories to real names of >10,000 OkC users.

RETWEETS	LIKES
36	18

2:23 PM - 11 May 2016

t the dataset is

cludes

published data.

The situation is raisi

allowed to collect en

Emil OW Kirkegaard
@KirkegaardEmil

Follow

@esjewett No. Data is already public.

RQ: How do Twitter users feel about the use of their tweets in research?

Fiesler, C. & Proferes, N. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society* 4(1).

methods

- survey (demographics, Twitter use, Likert scale for comfort level, open-ended questions)
- Twitter users via Mechanical Turk
- Final N: 268

61% had no idea that researchers use tweets in research

61% had no idea that researchers use tweets in research

47% think researchers are not permitted to without permission

61% had no idea that researchers use tweets in research

47% think researchers are not permitted to without permission

61% of these think it is breaking ethical rules for researchers

61% had no idea that researchers use tweets in research

47% think researchers are not permitted to without permission

- 61% of these think it is breaking ethical rules for researchers
- 23% of these think it is a violation of Twitter's Terms of Service

61% had no idea that researchers use tweets in research

47% think researchers are not permitted to without permission

- 61% of these think it is breaking ethical rules for researchers
- 23% of these think it is a violation of Twitter's Terms of Service

[O]ur public user profile information and **public Tweets** are immediately delivered via SMS and our APIs to our partners and other third parties, including search engines, developers, and publishers that integrate Twitter content into their services, and **institutions such as universities** and public health agencies that analyze the information for trends and insights.When you share information or content like photos, videos, and links via the Services, **you should think carefully about what you are making public**.

Regardless of whether you want them to use your tweets specifically, do you think that researchers **should** be able to use tweets in research without permission?

35% "Yes"

65% "No"

"I would not want my tweets to be used in a study unless I was informed."

"I would like to know if the research would serve a greater purpose."

"If my tweets were being used in a large scale study, I really wouldn't care. If anything was being personally picked out about me in a small study, I would care."

How would you feel if a tweet of yours was used in a research study and…

	Very uncomfortable	Somewhat uncomfortable	Neither uncomfortable or comfortable	Somewhat comfortable	Very comfortable
... you were not informed at all?	35.1%	31.7%	16.4%	13.4%	3.4%
... you were informed about the use after the fact?	21.3%	29.1%	20.5%	22.0%	7.1%
… it was analyzed along with millions of other tweets?	2.6%	18.7%	25.5%	30.0%	23.2%
… it was analyzed along with only a few dozen tweets?	16.5%	30.3%	24.0%	20.2%	9.0%
… it was from your "protected" account?	54.9%	20.5%	13.8%	6.0%	4.9%
… it was a public tweet you had later deleted?	31.3%	32.5%	20.5%	10.4%	5.2%
… no human researchers read it, but it was analyzed by a computer program?	2.6%	14.3%	30.5%	32.3%	20.3%
… the human researchers read your tweet in order to analyze it?	9.7%	27.6%	25.0%	25.4%	12.3%
... the researchers also analyzed your public profile information, such as location and username?	32.2%	23.2%	21.0%	13.9%	9.7%
… the researchers didn't have any of your additional profile information?	4.9%	15.4%	25.1%	34.1%	20.6%
... your tweet was quoted in a published research paper, attributed to your Twitter handle?	34.3%	21.6%	21.6%	13.1%	9.3%
... your tweet was quoted in a published research paper, attributed anonymously?	9.0%	16.8%	26.5%	28.4%	19.4%

Contextual Factor Examples

study purpose

If it was for a conservative cause I would be more forgiving. But that is unlikely with academic researchers, who are inherently biased, i.e., liberal slanted.

tweet content

If it's personal, has identifying information, or embarrassing/ offensive/private, I don't want my tweets used.

anonymity

As long as my name wasn't tied to it I wouldn't care.

dissemination

I honestly wouldn't mind [if researchers used my tweets] as long as I was told up front and I had the option to read the findings. I wouldn't want my name or handle associated or given credit anywhere though.

“publicness” is not the only context that matters.*

* my research ethics hill to die on

potential best practices

- if it is reasonable to ask (non-IRB) permission, consider doing so
- consider informing *after*, including dissemination
- anonymize identifying information
- when using certain methods that might cause more discomfort (e.g., small N, quotes), think more carefully about study context (e.g., sensitive topics)

design ideas

- opt-out option on platforms
- pre-emptive clear explanations to users
- “inform bots”
- BUT NOT making data more difficult to collect

design ideas

- opt-out option on platforms
- pre-emptive clear explanations to users
- “inform bots”
- BUT NOT making data more difficult to collect

“Research is a noble pursuit.” “if it’s for science!”

other things we might want to know from users

- attitudes cross-platform
- reactions to research
- what influences expectations
- how “public” is perceived
- … and other stuff

CNN International
January 14, 2016 ·

Like Page

British actor Alan Rickman was well-known for roles like Snape in the Harry Potter films:

Actor Alan Rickman dies at 69

EDITION.CNN.COM

Like Comment Share

14K Top Comments

30,837 shares 1.7K Comments

Write a comment...

[redacted] What sad start of year for famous people and celebrities in the World, Lemmy, Bowie and now Alan Rickman.

Like · Reply · 59 · January 14, 2016 at 6:09am

[redacted] Ya I was just talking to someone who said exactly that. Seems like everyone is going out of the picture.

Like · Reply · January 14, 2016 at 2:15pm

4.2.1 Incommensurate with Relationship

Facebook was flooded with reactions to Alan Rickman's death, many expressing strong emotional reactions. Just as Rickman's character in the fifth *Harry Potter* film called for Harry to "control [his] emotions," grief policers in this context found overt displays of sadness distasteful. They expressed a norm that grief is specific to personal, rather than parasocial relationships. One user went so far as to call for a general silence of expressions of grief about the actor:

> *Okay people!!! I understand that Alan Rickman was a pretty good actor (even if he usually played grumpy types)....but PLEASE- stop with the whole "Boo Hoo, I'm sooo sad!" shit. You never met him, you didn't know him.... move on to being sad about something that matters[...] (Comment #204, Rickman)*

"She said yes!" – Liminality and Engagement Announcements on Twitter

Munmun De Choudhury[†1], Michael Massimi[‡]
[†]Georgia Institute of Technology, Atlanta GA
[‡]Microsoft Research, Cambridge, UK

Abstract
Social media sites enable people to share
how and why they are used in the con
explore the behavior of individuals under
that they are engaged to be married. I
behavior manifested in Twitter that cha
of Twitter postings of engaged individual
in behavior that can be gleaned from so
for the statistical analysis. Our findings
of a major milestone in life, and bear im
Keywords Engagement, liminality, Twit

Relationship bonding: continuance, roles
My fiancee made red wine risotto with crimini mushrooms and shrimp for dinner. jealous? You should be.
I love my fiance!! He's already planned theme weeknights when we merge into a family.
#thingsthatmakemesmile cards from him. He writes the sweetest things and leaves them in the most random places
Togetherness rituals: occurrence, repetitiveness, attendance
There's only1thing better than a morning run; a morning run w/<user>
Watching tv with my beautiful pregnant fiancee! (at home sweet home) <url>
Out to dinner with my lovely girl.. <url>
Enjoyable activities (couple-centric)
We just got back from a camping trip. Definitely had nothing to be afraid of at night.. my fiance is the scariest thing in the woods.
Exhausted. And my alarm is set for 4am (less than three hours!) so I can make sure my fiance is up for a fishing trip. #TooEarly
Some video footage of when me and my girlfriend went around Chester and visited the museum etc.... <url>
Wedding event update
Wedding picture. union station. 12/3/11 ;) ;) <url>
We were just given the most incredibly gorgeous chinese meditation bell as a nye/wedding gift. wow. #dharma #hsil <url>
Headed to the wedding with my bro @<user> <url>
Wedding planning
Almost just died florist had the wrong flowers written down for my #wedding
Picking out wedding ring bands today #yayyy
My wedding dress came in! :)))))))))). can't waiiiiit! to see it, try it on! hope it's not too small...

Table 3: Topic categories in Twitter posts characterizing users' relationship investment and progression during the phase succeeding engagement.

"She said yes!" – Liminality and Engagement Announcements on Twitter

Munmun De Choudhury[†1], Michael Massimi[‡]
[†]Georgia Institute of Technology, Atlanta GA
[‡]Microsoft Research, Cambridge, UK

Abstract
Social media sites enable people to share
how and why they are used in the con
explore the behavior of individuals under
that they are engaged to be married. I
behavior manifested in Twitter that cha
of Twitter postings of engaged individual
in behavior that can be gleaned from so
for the statistical analysis. Our findings
of a major milestone in life, and bear im
Keywords Engagement, liminality, Twit

Relationship bonding: continuance, roles
My fiancee made red wine risotto with crimini mushrooms and shrimp for dinner. jealous? You should be.
I love my fiance!! He's already planned theme weeknights when we merge into a family.
#thingsthatmakemesmile cards from him. He writes the sweetest things and leaves them in the most random places
Togetherness rituals: occurrence, repetitiveness, attendance
There's only1thing better than a morning run; a morning run w/<user>
Watching tv with my beautiful pregnant fiancee! (at home sweet home) <url>
Out to dinner with my lovely girl.. <url>
Enjoyable activities (couple-centric)
We just got back from a camping trip. Definitely had nothing to be afraid of at night.. my fiance is the scariest thing in the woods.
Exhausted. And my alarm is set for 4am (less than three hours!) so I can make sure my fiance is up for a fishing trip. #TooEarly
Some video footage of when me and my girlfriend went around Chester and visited the museum etc.... <url>
Wedding event update
Wedding picture. union station. 12/3/11 ;) ;) <url>
We were just given the most incredibly gorgeous chinese meditation bell as a nye/wedding gift. wow. #dharma #hsil <url>
Headed to the wedding with my bro @<user> <url>
Wedding planning
Almost just died florist had the wrong flowers written down for my #wedding
Picking out wedding ring bands today #yayyy
My wedding dress came in! :)))))))))). can't waiiiiit! to see it, try it on! hope it's not too small...

Table 3: Topic categories in Twitter posts characterizing users' relationship investment and progression during the phase succeeding engagement.

lisa
@abrideabudget

Follow

Exhausted. And my alarm is set for 4am (less than three hours!) so I can make sure my fiance is up for a fishing trip. #TooEarly

11:13 PM - 15 Jul 2013

In a phone interview this week, Ms. Sokolowski, now 34, described the alteration of her social media usage as a logical progression.

"You don't want to tweet too much about your boyfriend because it could make you look clingy," she said. But once he had proposed, "I started to tweet about him more. We were going to be entwined for the rest of our lives."

Now, Ms. Sokolowski's fishing trip tweet has been cited in a new research effort by two computer scientists who examined how the soon-to-be-married used Twitter. (The study quoted a number of Twitter posts "anonymously," but I easily located and contacted Ms. Sokolowski. I informed her that her Tweet had been quoted in the study and obtained permission from her to write about it.)

s on Twitter

, roles
crimini mushrooms and shrimp for dinner. jealous? You

theme weeknights when we merge into a family.
im. He writes the sweetest things and leaves them in the

repetitiveness, attendance
ng run; a morning run w/<user>
t fiancee! (at home sweet home) <url>
l>
e)
Definitely had nothing to be afraid of at night.. my fiance is

n (less than three hours!) so I can make sure my fiance is

~~y girlfriend went around Chester and visited the museum~~

1 ;) ;) <url>
y gorgeous chinese meditation bell as a nye/wedding gift.

<user> <url>

lowers written down for my #wedding
#yayyy
an't waiiiiit! to see it, try it on! hope it's not too small...

haracterizing users' relationship investment and progression

during the phase succeeding engagement.

72% of PubMed articles using Twitter data quoted at least one tweet, leading to a Twitter user 84% of the time.

Ayers, John W., Theodore L. Caputi, Camille Nebeker & Mark Dredze. 2018. Don't quote me: Reverse identification of research participants in social media studies. *npj Digital Medicine* 1(1), p. 30.

“publicness” is not the only context that matters.*

* my research ethics hill to die on

Cambridge Analytica The Cambridge Analytica Files

Facebook gave data about 57bn friendships to academic

Leaked: Cambridge Analytica's blueprint for Trump victory

Exclusive: Former employee explains how presentation showed technique

Mark Zuckerberg takes out full-page newspaper ads to say 'sorry' for Cambri Analytica scandal

By Nicole Darrah | Fox News

ADAM ROGERS SCIENCE 03.25.18 07:00 AM

THE CAMBRIDGE ANALYTICA DATA APOCALYPSE WAS PREDICTED IN 2007

Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach

Cambridge Analytica active in elections, big data projects for years

KIM HJELMGAARD | USA TODAY
Updated 2:58 p.m. EDT Mar. 22, 2018

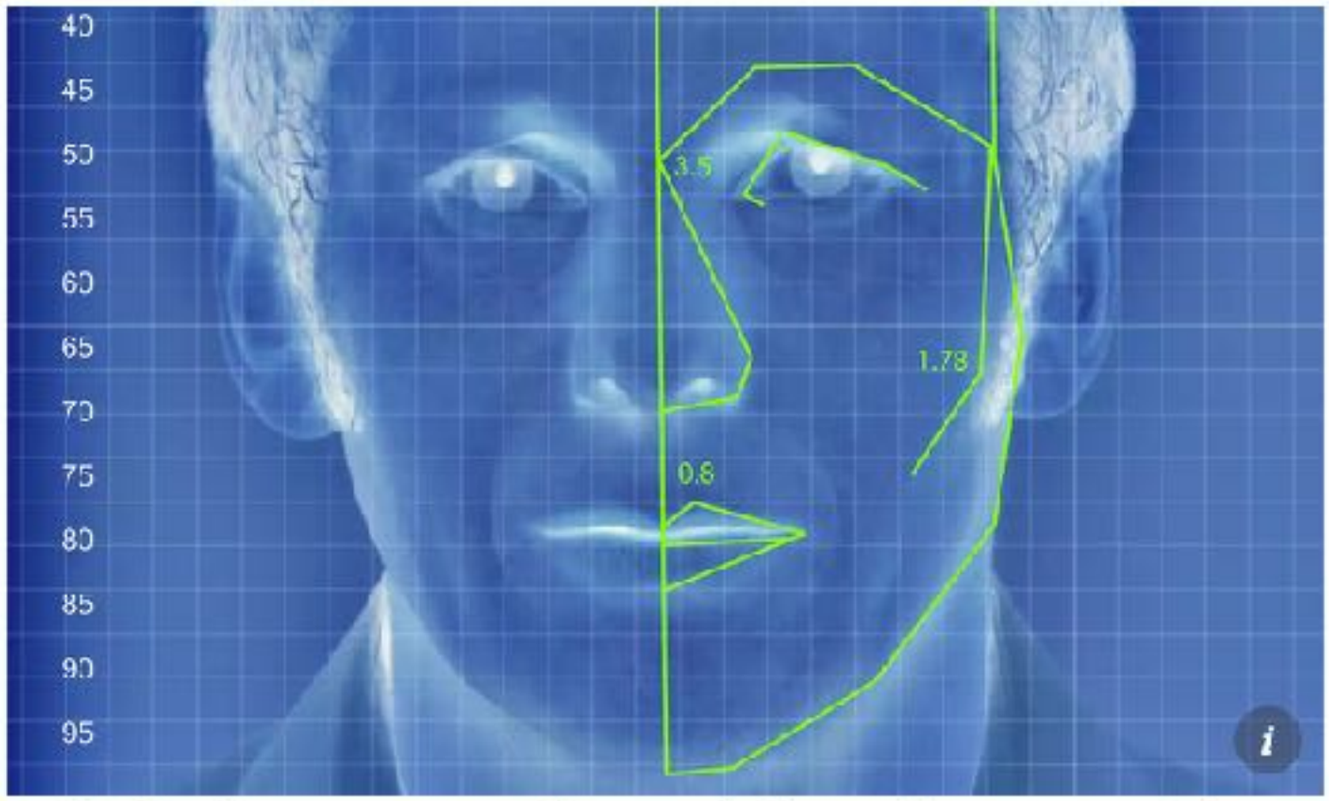
The shady data-gathering tactics used by Cambridge Analytica were an open secret to online marketers. I know, because I was one

Market researchers have used these tricks for years

By Alexandra Samuel | Mar 25, 2018, 1:19pm EDT

New AI can guess whether you're gay or straight from a photograph

An algorithm deduced the sexuality of people on a dating site with up to 91% accuracy, raising tricky ethical questions



Artificial intelligence can accurately guess whether people are gay or straight based on photos of their faces, according to new research that suggests machines can have significantly better "gaydar" than humans.

Algorithm Predicts If Twitter Users Are Becoming Mentally Ill

Can Facebook's Machine-Learning Algorithms Accurately Predict Suicide?

Multimodal Classification of Moderated Online Pro-Eating Disorder Content

Instagram photos reveal predictive markers of depression

Artificial intelligence could identify gang crimes—and ignite an ethical firestorm

By **Matthew Hutson** | Feb. 28, 2018 , 8:00 AM

"I'm just an engineer," he said.

ethics is not a specialization.*

* my ethics education hill to die on

PERVADE

PERVASIVE DATA ETHICS for computational research

- ethics of and user attitudes toward data sharing
- representation & reactions to research in the media
- ethical practices & norms among researchers
- role and experiences of IRBs
- data ethics in industry
- quantification and metrics for ethical risks

Let's talk more!

casey.fiesler@colorado.edu