Microsoft Research

Microsoft

The Human Manifold

On the Predictability of Human Online Behaviour and its Consequences

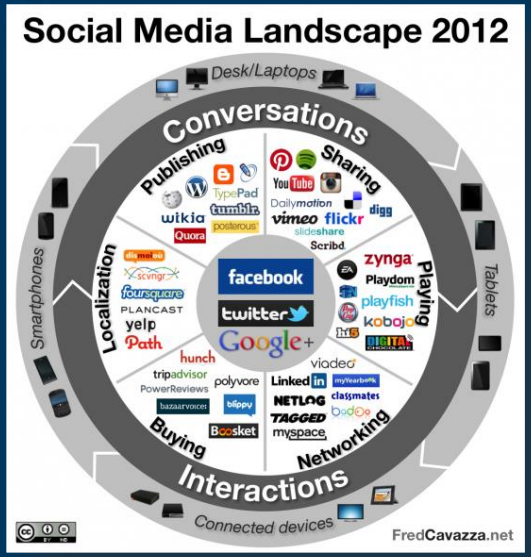
Thore Graepel

MSR Cambridge: Yoram Bachrach, Pushmeet Kohli, and Milad Shokouhi
Cambridge University: Michal Kosinski and David Stillwell

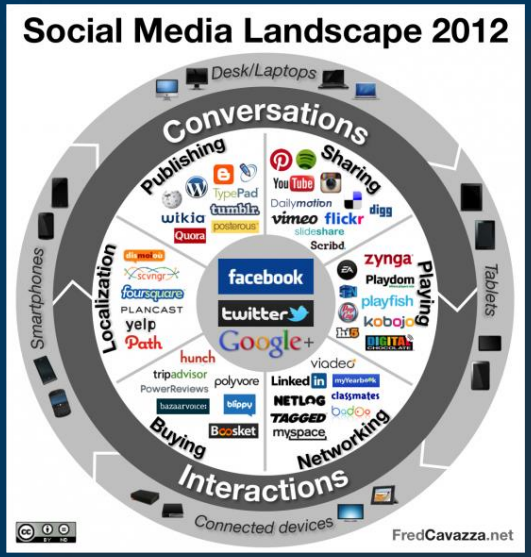
LSOLDM Workshop, 24th September 2013, Cumberland Lodge

The Digital Traces: Online Data

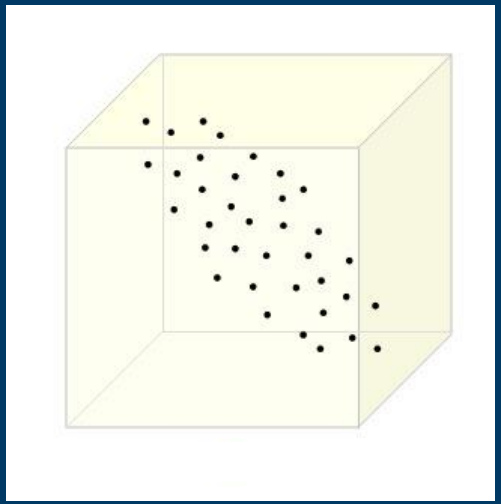
The Digital Traces: Online Data



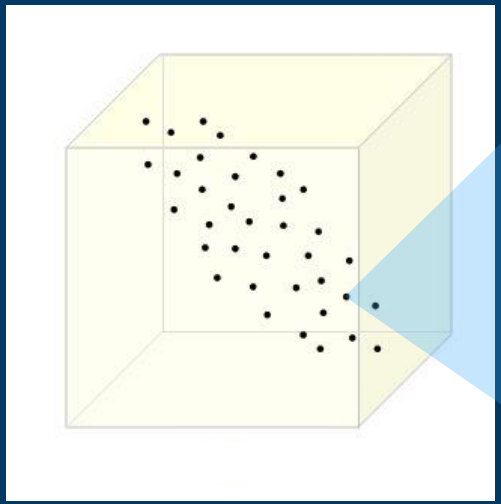
The Digital Traces: Online Data

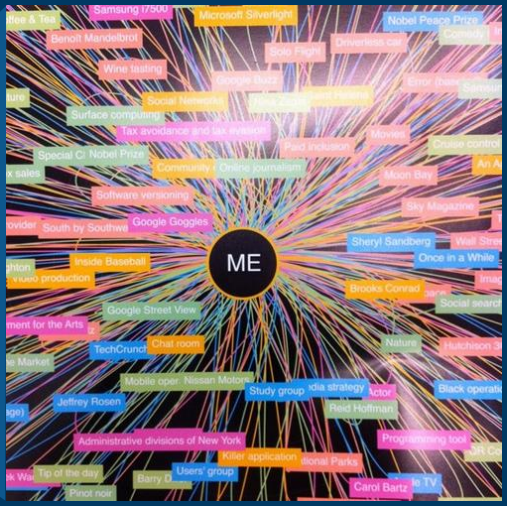


High-Dimensional Observation Space



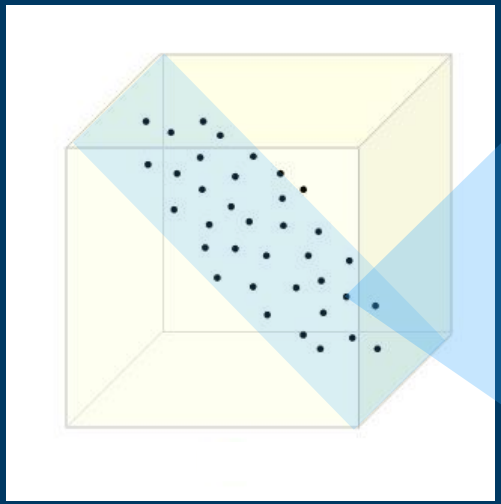
High-Dimensional Observation Space

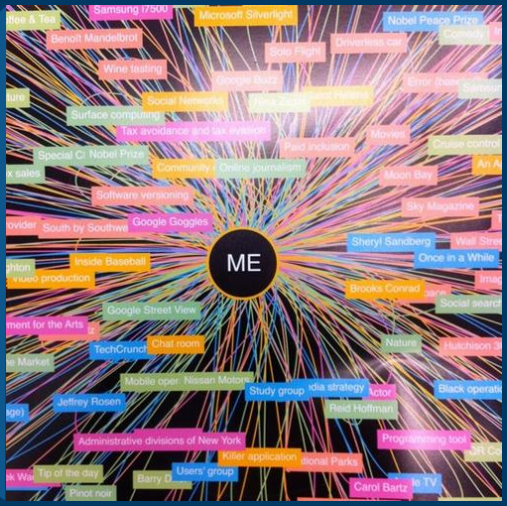




www.marketersstudio.com

High-Dimensional Observation Space





www.marketersstudio.com

Mapping the Human Manifold

-
  -
  -
  -
-
  -
  -
  -
-
  -

Mapping the Human Manifold

- Scientific potential - Understanding:
  - Better understand human behaviour
  - Understand commonality and individual differences among people
  - Obtain psychometric measurements at an unprecedented scale
-
  -
  -
  -
-
  -

Mapping the Human Manifold

- Scientific potential - Understanding:
  - Better understand human behaviour
  - Understand commonality and individual differences among people
  - Obtain psychometric measurements at an unprecedented scale
- Business potential - Predicting:
  - Develop fine-grained and predictive psycho-demographic user profiles
  - Increase user satisfaction by deep personalization for products and services
  - Increase revenue by providing more engaging ads and recommendations
-

Mapping the Human Manifold

- Scientific potential - Understanding:
  - Better understand human behaviour
  - Understand commonality and individual differences among people
  - Obtain psychometric measurements at an unprecedented scale
- Business potential - Predicting:
  - Develop fine-grained and predictive psycho-demographic user profiles
  - Increase user satisfaction by deep personalization for products and services
  - Increase revenue by providing more engaging ads and recommendations
- How: Find mapping to interpretable dimensions
  - Personality, Intelligence, Happiness, etc.

Big Five Personality traits

Openness

- **Appreciation of art, emotion, adventure, and variety of experience**

Conscientiousness

- **Self-discipline, act dutifully, and aim for achievement**

Extraversion

- **Energy, positive emotions, seek social stimulation**

Agreeableness

- **Compassionate and cooperative rather than suspicious**

Neuroticism

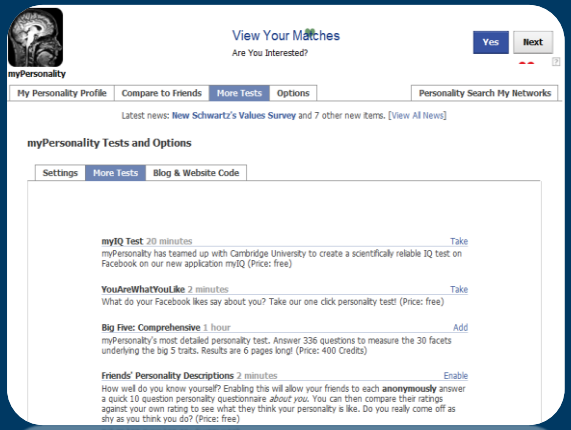
- **Experience unpleasant emotions easily, such as anger and anxiety**

A Treasure Chest of Data: MyPersonality

-
-
  -
  -
  -
-
  -
  -
  -

A Treasure Chest of Data: MyPersonality

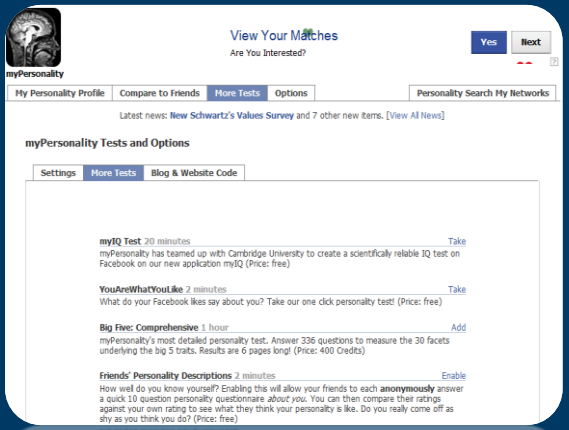
- Facebook App since 2008
-
  - .
  - .
  - .
-
  - .
  - .
  - .



Data: www.myPersonality.org
David Stillwell, Michal Kosinski
Cambridge Psychometrics Centre

A Treasure Chest of Data: MyPersonality

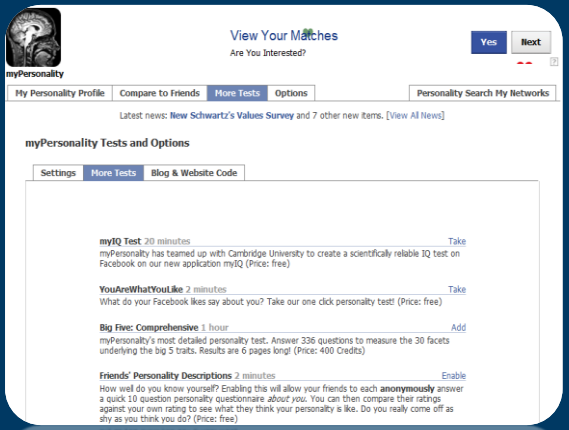
- Facebook App since 2008
- Over 8 Million psychometric test results
  - Personality
  - Intelligence
  - Happiness
-
  - .
  - .
  - .



Data: www.myPersonality.org
David Stillwell, Michal Kosinski
Cambridge Psychometrics Centre

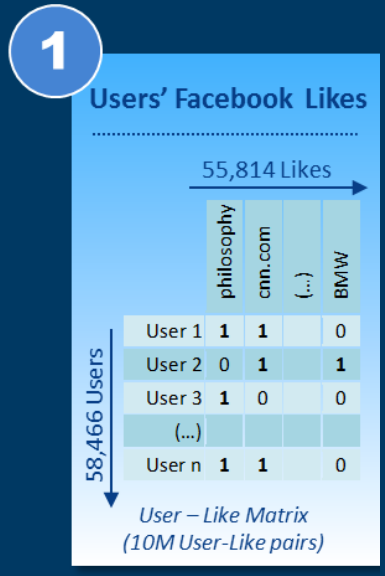
A Treasure Chest of Data: MyPersonality

- Facebook App since 2008
- Over 8 Million psychometric test results
  - Personality
  - Intelligence
  - Happiness
- Volunteered user profiles
  - Relationship status, age, gender
  - Facebook Likes
  - Friendship network

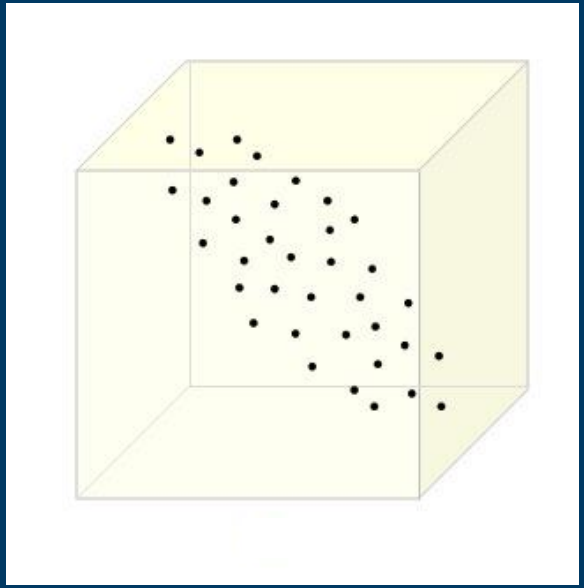


Data: www.myPersonality.org
David Stillwell, Michal Kosinski
Cambridge Psychometrics Centre

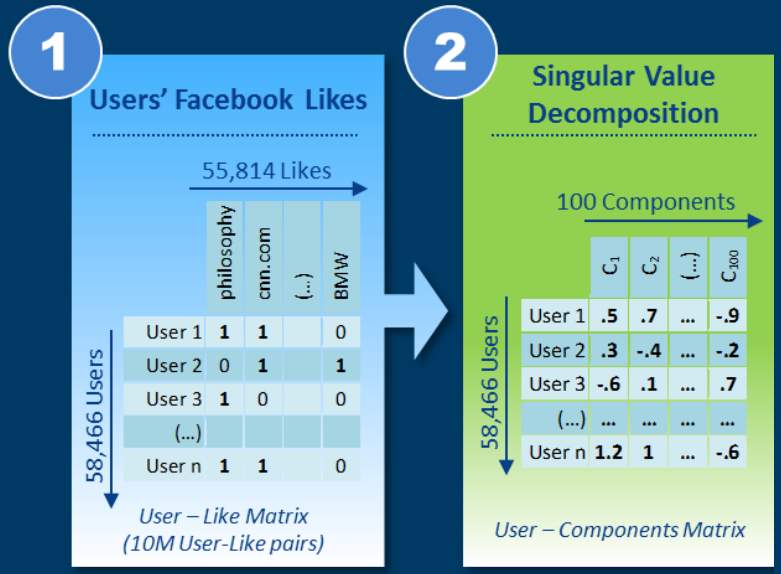
The Magic of Machine Learning



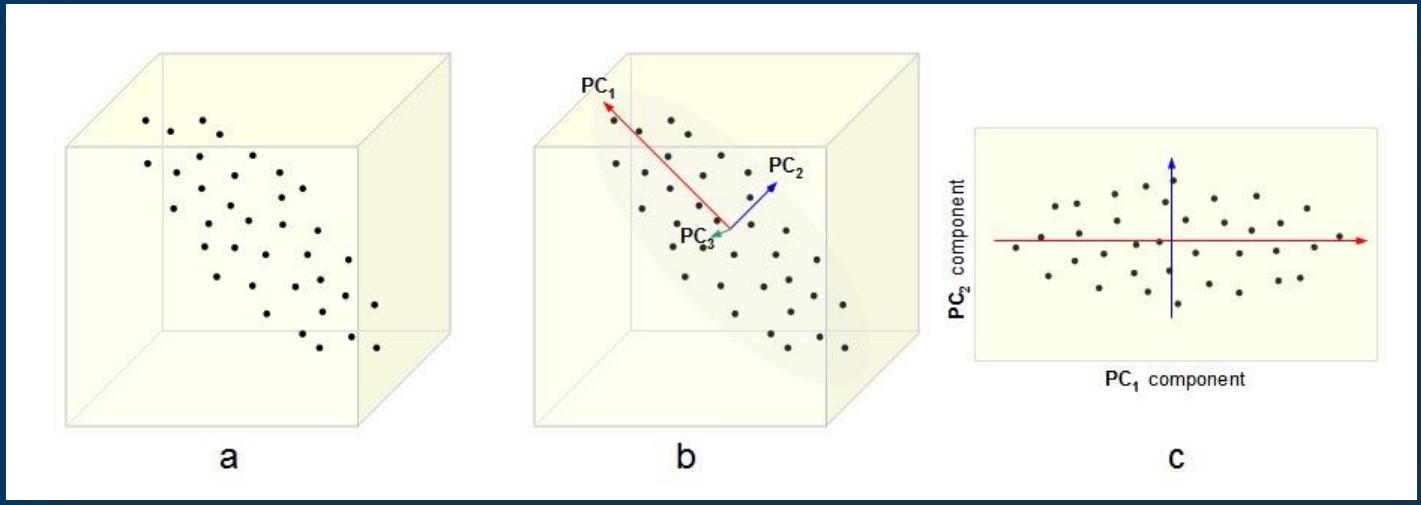
High-Dimensional Observation Space



The Magic of Machine Learning

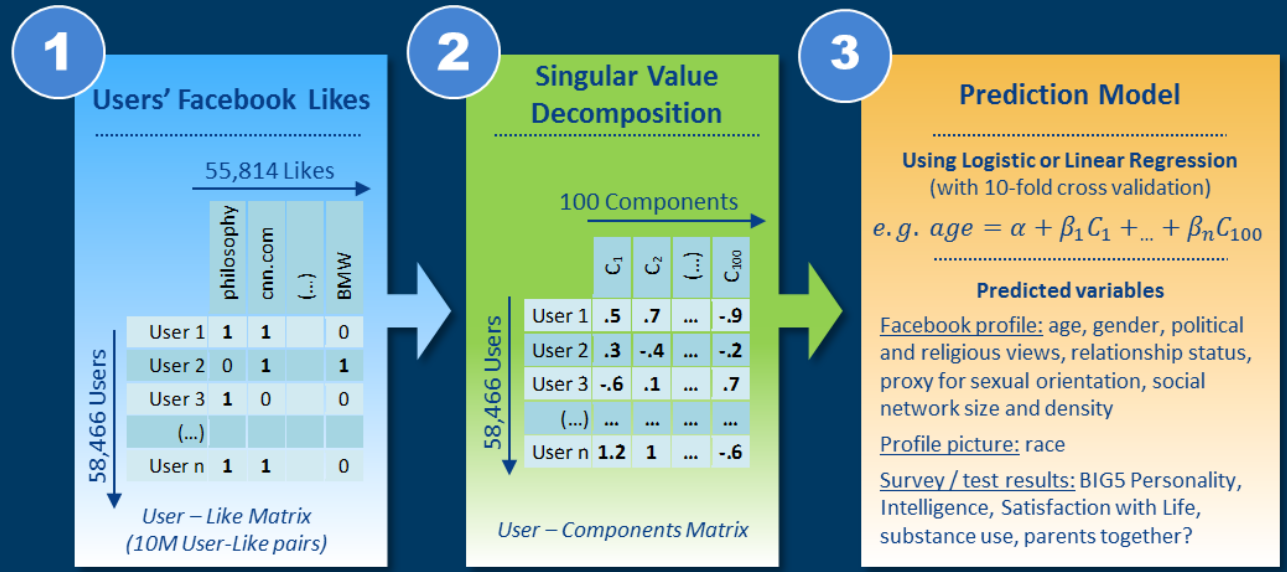


Mapping the Manifold

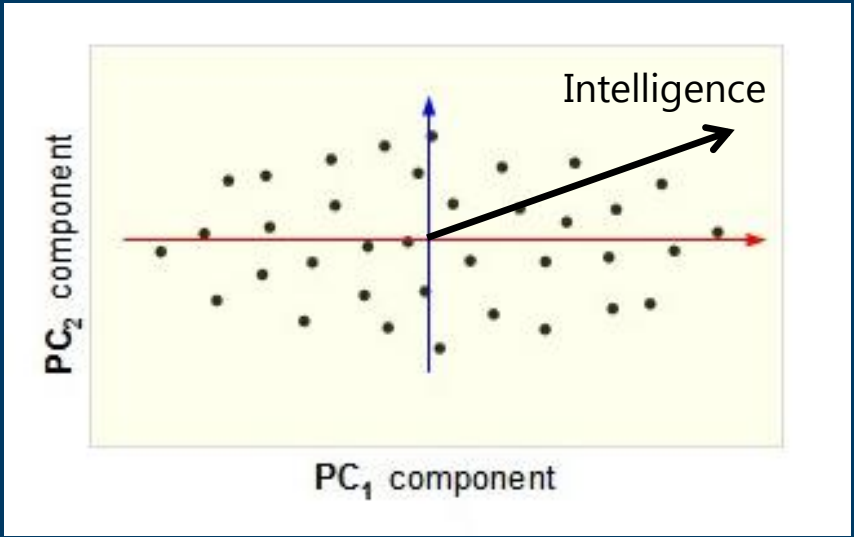


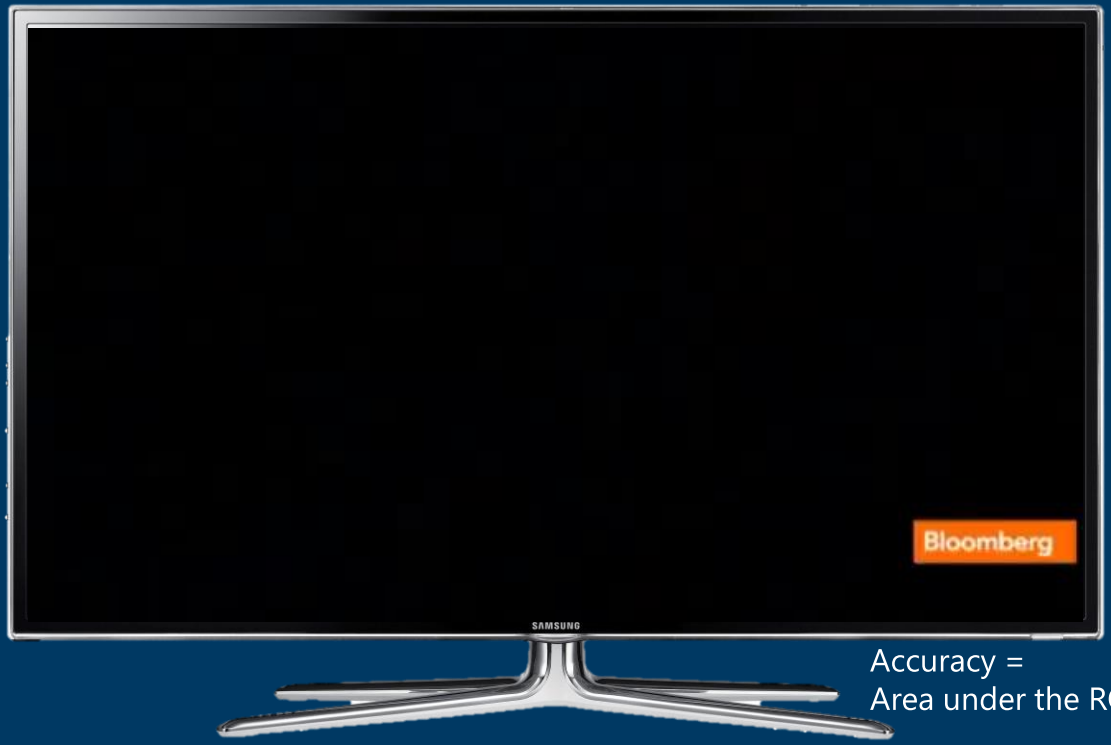
100 out of 55,814 dimensions explain 28% of variance

The Magic of Machine Learning



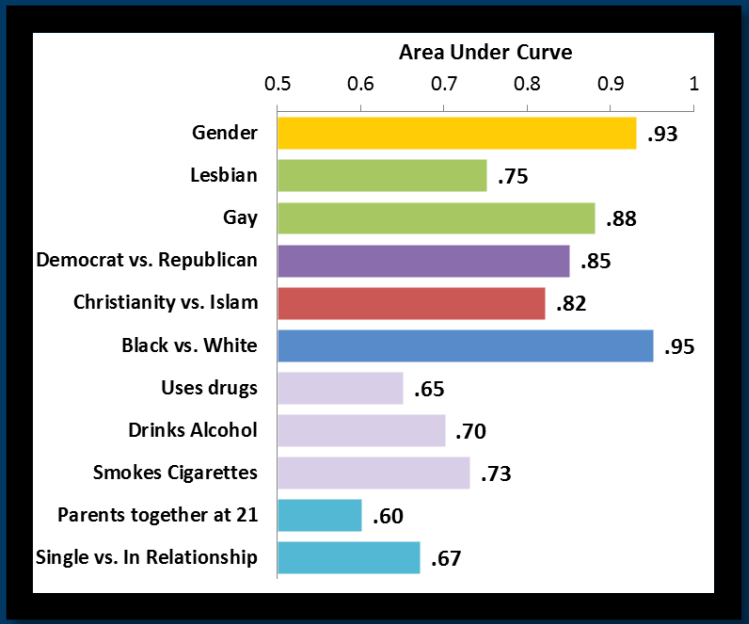
Making Predictions



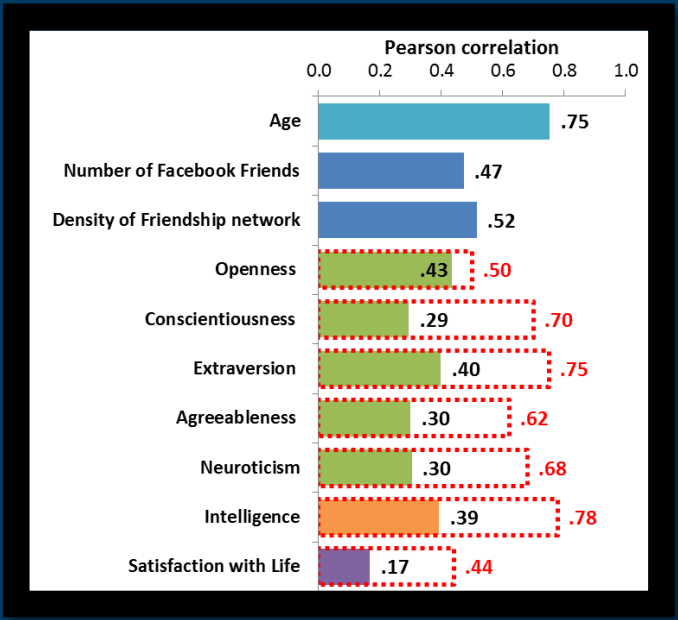


Accuracy =
Area under the ROC curve

Prediction Accuracy: Binary variables



Prediction Accuracy: Numeric Variables



SAMSUNG

What is there to like? Intelligence

IQ		
High	The Godfather	Jason Aldean
	Mozart	Tyler Perry
	Thunderstorms	Sephora
	The Colbert Report	Chiq
	Morgan Freemans Voice	Bret Michaels
	The Daily Show	Clark Griswold
	Lord Of The Rings	Bebe
	To Kill A Mockingbird	I Love Being A Mom
	Science	Harley Davidson
	Curly Fries	Lady Antebellum
		Low

Which Likes? Happiness

Satisfaction With Life		
Satisfied	Sarah Palin	Hawthorne Heights
	Glenn Beck	Kickass
	Proud To Be Christian	Atreyu (Metal Band)
	Indiana Jones	Lamb Of God
	Swimming	Gorillaz
	Jesus Christ	Science
	Bible	Quote Portal
	Jesus	Stewie Griffin
	Being Conservative	Killswitch Engage
	Pride And Prejudice	Ipod
		Dissatisfied

Which Likes? Extraversion

Extraversion		
Outgoing & Active	Beerpong	RPGs
	Michael Jordan	Fanfiction.Net
	Dancing	Programming
	Socializing	Anime
	Chris Tucker	Manga
	I Feel Better Tan	Video Games
	Modeling	Role Playing Games
	Cheerleading	Minecraft
	Theatre	Voltaire
	Flip Cup	Terry Pratchet
		Shy & Reserved

Toronto Star

ON ONO

NEWS

You are what you 'like,' Facebook study finds

(And if you like Colbert and curly fries, you're one of the smart ones)

LESLEY CIARULA TAYLOR
STAFF REPORTER

Liking curly fries is a smart thing to do.

So smart that a computer crunching millions of Facebook "likes," independent of the information in users' profiles, can figure out all kinds of personal and often private details about them, including their IQ.

A study by Michal Kosinski and David Stillwell of the University of Cambridge's Psychometrics Centre found that "likes" could predict, with varying degrees of success, whether someone used drugs, smoked, had divorced parents and leaned liberal or conservative.

"The best predictors of high intelligence include 'thunderstorms,' *'The Colbert Report,'* 'science' and 'curly fries,' " the study determined.

"To be honest, we were mindblown when we saw the results for the first time," Kosinski told the Star.

"We could predict so many things with so much accuracy. To be able to predict someone's IQ or personality or whether their parents were divorced was very surprising, even to me, and I predict personality traits all the time."

The study took the likes (registered when a Facebook user clicks "like" on something, and stored on the site) of 60,000 volunteers in the United States. It looked at "photos, friends' status updates, pages of products, sports, musicians, books, restaurants or popular websites."

"Likes represent a very generic class of digital records, similar to web search queries, web browsing histories and credit card purchases," the study said.

Isn't this dangerous?

"This is one of the main points of the paper," Kosinski said. ". . . What used to be expensive and labour-intensive to find out can now be done very quickly and potentially more accurately simply by asking someone to access his Facebook profile.

"Our results suggest these results can be done without your consent behind the scenes."

The researchers found the secrets of more selective Facebook users were actually easier to crack open than the people "who spend half their life online."

The volunteers had from one to 700 likes each, with the average being 170. Highly active Facebook users had a higher level of online noise that could obscure an accurate reading on their private lives, Kosinski said.

Even more troubling, he said, was the computer program's ability to predict the sexual orientation of someone who hadn't declared one on Facebook. "In Canada, it's not a big issue. In Iran, people might have an issue with it," he said.

To quantify what a like might mean, the object had to have at least 20 likes, Kosinski said. That meant eliminating 11 million unique likes among the volunteers.

The team was able to predict with 95-per-cent certainty a person's race, 93 per cent their sex, 88 per cent their sexual orientation, 82 per cent their religion, 73 per cent whether they smoked and 60 per cent whether they had divorced parents.

"Individuals with parents who separated have a higher probability of liking statements preoccupied with relationships, such as 'If I'm with you, then I'm with you. I don't want anybody else,' " the study said.

The study, published Monday in the Proceedings of the National Academy of Sciences, was less accurate on personality, Kosinski said, reflecting the more arbitrary, changing nature of mood. Their accuracy on intelligence was highest of the psychological traits, at 78 per cent.

Next step, he said, is to figure out what all this means.

"There is no obvious correlation between curly fries and intelligence," the study noted.

Toronto Star
The Times
THE TIMES | Tuesday March 12 2013
News
How Facebook 'likes' reveal all of your character secrets in a click
Murad Ahmed
Technology Reporter
Do you feel a peculiar thrill during a thunderstorm; are you a fan of Pride and Prejudice who also enjoys creating scrapbooks? Then you're probably very smart, content with life and in a relationship.
Or do you prefer the sprinter Usain Bolt, have a tattoo and use an iPod? Then you're more likely to be single, a keen drinker and unhappy with your circumstances.
Those are some of the conclusions from a new study published yesterday which claims that the way we use Facebook reveals a lot — perhaps too much — about our personal characteristics and most intimate details.
Psychologists and computer scientists at the University of Cambridge have analysed tens of thousands of Facebook users, tracking the pages on which they clicked the "Like" button — blue thumbs-up sign familiar to the social network's billion users.
These likes, of anything from a political party to an amusing photograph, are seen by a user's friends, but can often be viewed by anyone else on the Internet. The researchers were able to use this Facebook activity to predict details accurately such as a person's
What choices say about you
Pages "liked" by social network users that reveal personality traits
Satisfied with life
Indiana Jones
Pride and Prejudice
Swimming
In a relationship
Weightwatchers (fronted by Patsy Kensit)
Huggies
Liberal and artistic
Sylvia Plath
Plato
Oscar Wilde
High IQ
Lord of the Rings
The Godfather
Mozart
'We were surprised by what you can find out through songs, movies and quotes'
what appear to be incongruous or random links between a person's online preferences and real-world behaviour. The research suggests that those who like Jane Austen's novel Pride and Prejudice tend to be satisfied with their life, while those who endorsed the page "I like Lyrics That Actually Mean Something" are likelier to be drug users.
Mr Kosinski said his techniques may help psychologists to make mental health diagnoses. But he worried that governments and companies could abuse his methods. For example, he suggested that illiberal regimes may be tempted to analyse their citizens' online activities to persecute political opponents or gay people.
"I'm a big fan of these technology companies like Google, Facebook, Microsoft," he added. "I don't believe they store this data for evil reasons. They just do it because it's easy."
Facebook declined to comment.
BOODLES
BRITISH EXCELLENCE SINCE 1798
A4 | TORONTO STAR
ON ONO

The Telegraph

Features

Clicks that tell the world all it needs to know about you

Psychological profiles are an eerily accurate reflection of our activity online, says **Harry Wallop**

Toronto Star

News

ll of ick

ON ON0

A4 | TORONTO STAR

NE

You

(And if

and cur

one of

LESLEY

STAFF REP

Liking

do.

So s

ing

dep

ers'

of

ab

ternet. The researchers were able to

this Facebook activity to predict

found that

with varying degrees

whether someone used drug

smoked, had divorced parents and

produc

THE TIMES | Tuesday March 12 20

How F

your cl

Murad Ahmed

Technology

Reporter

Do you feel a pecu-

liar thrill during a

thunderstorm; are you a

fan of *Pride and Prejudice* who

also enjoys creating scrapbooks?

Then you're probably very smart, con-

tent with life and in a relationship.

Or do you prefer the sprinter Usain

Bolt, have a tattoo and use an iPod?

Then you're more likely to be single,

a keen drinker and unhappy with

your circumstances.

Those are some of the conclu-

sions from a new study published

yesterday which claims that the

way we use Facebook reveals a lot

— perhaps too much — about our

personal characteristics and most in-

timate details.

Psychologists and computer scien-

tists at the University of Cambridge

have analysed tens of thousands of

Facebook users, tracking the pages on

which they clicked the "Like" button —

the blue thumbs-up sign familiar to the

social network's billion users.

These likes, of anything from a polit-

ical party to an amusing photograph,

can be seen by a user's friends, but can

also be viewed by anyone else on the

Satisf

with li

Indiana Jo

Pride and

Prejudice

Swimming

times

5

cent

50 per

vorced

ts who

obability

occupied

as 'If I'm

you. I don't

study said.

Monday in

he National

was less accu-

Kosinski said.

bitrary, chang-

Their accuracy

highest of the

, at 78 per cent.

d, is to figure out

is.

vious correlation

fries and intelli-

y noted.

ear to be incongruous or ran-

between a person's online

es and real-world behaviour.

search suggests that those who

usten's novel *Pride and Prej-*

to be satisfied with their life,

who endorsed the page *"I*

That Actually Mean Some-

keliher to be drug users.

ski said his techniques may

ologists to make mental

loses. But he worried that

and companies could

ethods. For example, he

t illiberal regimes may be

analyse their citizens'

es to persecute political

gay people.

an of these technology

e Google, Facebook.

added. "I don't believe

data for evil reasons.

ecause it's easy."

fined to comment.

LES

Financial Times

13

Facebook Likes show big data brings big responsibility

Michal Kosinski

Big data is subject to much hype. The ability to manipulate vast swaths of information at warp seed is transforming businesses across the world. But as with all technologies there are risks. All users of technology should take note: with big data comes big responsibility.

Toronto Star

News

Daily Mail

Facebook threat to users' privacy

By **Andrew Levy**

FACEBOOK users are at risk of unwittingly revealing personal details simply by 'liking' pages on the site dedicated to anything from celebrities to charities, researchers warn.

Sexuality, drug use, political views and religious beliefs can be accurately predicted by monitoring users' activity on the social networking website, they said.

The team from Cambridge University focused their research on Facebook's system of liking pages – the seemingly innocuous act of clicking a button illustrated with a thumbs up.

Worryingly, the researchers found that liking even apparently unrelated information still can be used to accurately predict personal details.

For example, the researchers found drug use is suggested by 'liking' milkshakes and swimming, while high IQs are indicated by showing a taste for curly fries, and the Godfather movies.

The study was carried out by Cambridge's Psychometrics Centre and based on the Facebook profiles of 58,000 people in the US.

Their 'likes' were fed into a computer algorithm which was used to predict a range of personality traits. Researchers predicted male sexuality with 88 per cent accuracy. They also had an 85 per cent success rate with political leanings and 82 per cent with religion.

Dr Gus Hosein, of campaigners Privacy International, said: 'It's a nightmare scenario that Facebook is entirely responsible for setting up. This information can be used to pre-categorise people.

'Banks could use it to decide who gets a loan. It also creates the perfect surveillance state for governments.'

Facebook declined to comment yesterday.

The Telegraph

Toronto Star

A4 | TORONTO STAR

THE TIMES | Tuesday March 12 20

How F

your cl

Murad Ahmed
Technology
Reporter

LESLEY
STAFF REP

Questions for User Privacy

Facebook threat to users' privacy

FACEBOOK users are at risk of unwit- By Andrew Levy

Questions for User Privacy

Facebook threat to users' privacy

FACEBOOK users are at risk of unwit- By Andrew Levy

- Are users aware to what degree we can infer their personal traits and attributes from their digital traces?

Questions for User Privacy

Facebook threat to users' privacy

FACEBOOK users are at risk of unwit- By Andrew Levy

- Are users aware to what degree we can infer their personal traits and attributes from their digital traces?
- If they were, would they make that data public?

Questions for User Privacy

Facebook threat to users' privacy

FACEBOOK users are at risk of unwit- By Andrew Levy

- Are users aware to what degree we can infer their personal traits and attributes from their digital traces?
- If they were, would they make that data public?
- Do these predictions just summarize what users are signalling to their friends using Facebook Likes?

Questions for User Privacy

Facebook threat to users' privacy

FACEBOOK users are at risk of unwit- By Andrew Levy

- Are users aware to what degree we can infer their personal traits and attributes from their digital traces?
- If they were, would they make that data public?
- Do these predictions just summarize what users are signalling to their friends using Facebook Likes?
- Should companies/services be allowed to use this information for commercial purposes?

Questions for User Privacy

Facebook threat to users' privacy

FACEBOOK users are at risk of unwit- By Andrew Levy

- Are users aware to what degree we can infer their personal traits and attributes from their digital traces?
- If they were, would they make that data public?
- Do these predictions just summarize what users are signalling to their friends using Facebook Likes?
- Should companies/services be allowed to use this information for commercial purposes?
- Can technology help to mitigate these challenges?

Conclusions

-
  -
  -
-
-
  -
  -
  -

Conclusions

- Mapping the manifold of human online behaviour
  - Science: Understanding human behaviour and large-scale psychometrics
  - User benefit: Deep personalization and adaptation to user preferences
-
-
  -
  -
  -

Conclusions

- Mapping the manifold of human online behaviour
  - Science: Understanding human behaviour and large-scale psychometrics
  - User benefit: Deep personalization and adaptation to user preferences
- Balance need for privacy with quality of service
-
  .
  .
  .

Conclusions

- Mapping the manifold of human online behaviour
  - Science: Understanding human behaviour and large-scale psychometrics
  - User benefit: Deep personalization and adaptation to user preferences
- Balance need for privacy with quality of service
- Challenges and future work
  - Define a new predictive and interpretable framework for human behaviour
  - Which “Likes” do people get exposed to and how? Is there a filter bubble?
  - Discover causal relationships from observational data

Private traits and attributes are predictable from digital records of human behavior

Michal Kosinski[a,1], David Stillwell[a], and Thore Graepel[b]

[a]The School Lane, The Psychometrics Centre, University of Cambridge, Cambridge CB2 3RQ United Kingdom; and [b]Microsoft Research, Cambridge CB1 2FB, United Kingdom

Edited by Kenneth Wachter, University of California, Berkeley, CA, and approved February 12, 2013 (received for review October 29, 2012)

We show that easily accessible digital records of behavior, Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender. The analysis presented is based on a dataset of over 58,000 volunteers who provided their Facebook Likes, detailed demographic profiles, and the results of several psychometric tests. The proposed model uses dimensionality reduction for preprocessing the Likes data, which are then entered into logistic/linear regression to predict individual psychodemographic profiles from Likes. The model correctly discriminates between homosexual and heterosexual men in 88% of cases, African Americans and Caucasian Americans in 95% of cases, and between Democrat and Republican in 85% of cases. For the personality trait "Openness," prediction accuracy is close to the test–retest accuracy of a standard personality test. We give examples of associations between attributes and Likes and discuss implications for online personalization and privacy.

social networks | computational social science | machine learning | big data | data mining | psychological assessment

A growing proportion of human activities, such as social interactions, entertainment, shopping, and gathering information, are now mediated by digital services and devices. Such digitally mediated behaviors can easily be recorded and analyzed, fueling the emergence of computational social science (1) and new services such as personalized search engines, recommender systems (2), and targeted online marketing (3). However, the widespread availability of extensive records of individual behavior, together with the desire to learn more about customers and citizens, presents serious challenges related to privacy and data ownership (4, 5).

We distinguish between data that are actually recorded and information that can be statistically predicted from such records. People may choose not to reveal certain pieces of information about their lives, such as their sexual orientation or age, and yet this information might be predicted in a statistical sense from other aspects of their lives that they do reveal. For example, a major US retail network used customer shopping records to predict pregnancies of its female customers and send them well-timed and well-targeted offers (6). In some contexts, an unexpected flood of vouchers for prenatal vitamins and maternity clothing may be welcome, but it could also lead to a tragic outcome, e.g., by revealing (or incorrectly suggesting) a pregnancy of an unmarried woman to her family in a culture where this is unacceptable (7). As this example shows, predicting personal information to improve products, services, and targeting can also lead to dangerous invasions of privacy.

Predicting individual traits and attributes based on various cues, such as samples of written text (8), answers to a psychometric test (9), or the appearance of spaces people inhabit (10), has a long history. Human migration to digital environment renders it possible to base such predictions on digital records of human behavior. It has been shown that age, gender, occupation, education level, and even personality can be predicted from people's Web site browsing logs (11–15). Similarly, it has been shown that personality can be predicted based on the contents of personal Web sites (16), music collections (17), properties of Facebook or Twitter profiles such as the number of friends or the density of friendship networks (18–21), or language used by their users (22). Furthermore, location within a friendship network at Facebook was shown to be predictive of sexual orientation (23).

This study demonstrates the degree to which relatively basic digital records of human behavior can be used to automatically and accurately estimate a wide range of personal attributes that people would typically assume to be private. The study is based on Facebook Likes, a mechanism used by Facebook users to express their positive association with (or "Like") online content, such as photos, friends' status updates, Facebook pages of products, sports, musicians, books, restaurants, or popular Web sites. Likes represent a very generic class of digital records, similar to Web search queries, Web browsing histories, and credit card purchases. For example, observing users' Likes related to music provides similar information to observing records of songs listened to online, songs and artists searched for using a Web search engine, or subscriptions to related Twitter channels. In contrast to these other sources of information, Facebook Likes are unusual in that they are currently publicly available by default. However, those other digital records are still available to numerous parties (e.g., governments, developers of Web browsers, search engines, or Facebook applications) and, hence, similar predictions are unlikely to be limited to the Facebook environment.

The design of the study is presented in Fig. 1. We selected traits and attributes that reveal how accurate and potentially intrusive such a predictive analysis can be, including "sexual orientation," "ethnic origin," "political views," "religion," "personality," "intelligence," "satisfaction with life" (SWL), substance use ("alcohol," "drugs," "cigarettes"), "whether an individual's parents stayed together until the individual was 21 y old," and basic demographic attributes such as "age," "gender," "relationship status," and "size and density of the friendship network." Five Factor Model (9) personality scores ($n = 54,373$) were established using the International Personality Item Pool (IPIP) questionnaire with 20 items (25). Intelligence ($n = 1,350$) was estimated using Raven's Standard Progressive Matrices (SPM) (26), and SWL ($n = 2,340$) was measured using the SWL Scale (27). Age ($n = 52,700$; average, $\mu = 25.6$; SD = 10), gender ($n = 57,505$; 62% female), relationship status ("single"/"in relationship"; $n = 46,027$; 49% single), political views ("Liberal"/"Conservative"; $n = 9,752$;

Author contributions: M.K. and T.G. designed research; M.K. and D.S. performed research; M.K. and T.G. analyzed data; and M.K., D.S., and T.G. wrote the paper.

Conflict of interest statement: D.S. received revenue as owner of the myPersonality Facebook application.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the myPersonality Project database (www.mypersonality.org/wiki/).

[1]To whom correspondence should be addressed. E-mail: mk583@cam.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1218772110/-/DCSupplemental.

SOCIAL SCIENCES

NAS Embargo

PNAS

Inferring the Demographics of Search Users

When Social Data Met Search Queries

Bin Bi*
UCLA
United States
bbi@cs.ucla.edu

Milad Shokouhi
Microsoft Research
United Kingdom
milads@microsoft.com

Michal Kosinski
University of Cambridge
United Kingdom
michal@michalkosinski.com

Thore Graepel
Microsoft Research
United Kingdom
thoreg@microsoft.com

ABSTRACT

Knowing users' views and demographic traits offers a great potential for personalizing web search results or related services such as query suggestion and query completion. Such signals however are often only available for a small fraction of search users, namely those who log in with their social network account and allow its use for personalization of search results. In this paper, we offer a solution to this problem by showing how user demographic traits such as age and gender, and even political and religious views can be efficiently and accurately inferred based on their search query histories. This is accomplished in two steps; we first train predictive models based on the publically available myPersonality dataset containing users' Facebook Likes and their demographic information. We then match Facebook Likes with search queries using Open Directory Project categories. Finally, we apply the model trained on Facebook Likes to large-scale query logs of a commercial search engine while explicitly taking into account the difference between the traits distribution in both datasets. We find that the accuracy of classifying age and gender, expressed by the area under the ROC curve (AUC), are 77% and 84% respectively for predictions based on Facebook Likes, and only degrade to 74% and 80% when based on search queries. On a US state-by-state basis we find a Pearson correlation of 0.72 for political views between the predicted scores and Gallup data, and 0.54 for affiliation with Judaism between predicted scores and data from the US Religious Landscape Survey. We conclude that it is indeed feasible to infer important demographic data of users from their query history based on labelled Likes data and believe that this approach could provide valuable information for personalization and monetization even in the absence of demographic data.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]:

General Terms

Algorithms, Human Factors

Keywords

User demographics, Personalized search, Social networks

1. INTRODUCTION

In recent years, we have been witnessing the rapid emergence of social networks and an increasing amounts of user generated data. Meanwhile, it became apparent that the relevance of search results can be improved by personalization, i.e., by taking into account additional information about the user, such as interests, demographic and psychological traits, social background, or the context of the search. As a consequence, search engines have been evolving into *social-aware* platforms, Google's social layer (Google+), and Bing's *social purse* being perhaps the two most noteworthy examples.

While leveraging the background information about the users in ranking models has shown significant promise in enhancing users' search experience both in academic [Carmel et al., 2009] and industrial[1] studies, obtaining such features for all users can be difficult. For instance, a recent study suggests that only about 22% of Bing users are logged into Facebook accounts while searching[2], and even them may have not given the search engine access to their profile information. It would therefore be useful to be able to infer characteristics of users relevant to their search experience from information more readily available in the contexts of a search engine, such as the search query histories.

This paper addresses the question of how demographic traits and users' views can be inferred based on the query histories. The main challenge, however, lies in the fact that only a very limited amount of data is available to allow training models for predicting such traits based on the search

*The work was done during Bin's internship at Microsoft Research Cambridge.

[1]Google blog, http://bit.ly/YaJv8m1.
[2]Search Engine Land: http://wp.me/p1RJWS-1qN

WWW 2013