

Turning the page on paper-based assessments:

Three techniques and one technology to transform patient-reported assessments

Chris Gibbons

Director of Health Assessment and Innovation
The Psychometrics Centre

NIHR Research Fellow
Cambridge Centre for Health Services Research
University of Cambridge

TODAY'S TALK

Describe

 Three cutting-edge techniques and one open-source technology to improve patient reported outcome measures

Demonstrate

 How modern psychometrics can transform patient outcome and experience assessment

Discuss

 Where we might be heading in this world of big data and computational social science?

THE TERMS

• Patient-reported outcome and experience measures

Psychometrics

- Item response theory
- Item banks
- eHealth
- Machine learning
- Computer adaptive testing
- Computational social science

THE TERMS

- Item response theory
- Item banks
- eHealth
- Machine learning
- Computer adaptive testing
- Computational social science

THE TERMS

ePsychometrics

- Item response theory
- Item banks
- eHealth
- Machine learning
- Computer adaptive testing
- Computational social science

PROMS AND PREMS

- Thousands, suitable for use in diverse conditions
- Many applications (trials, clinical practice, epidemiology, audit, quality, improvement)
- Patient-centred
- Ad hoc use (motivated clinicians)
- Take a long time to administer, score and interpret
- Scores can be difficult to understand
- Not much incentive



PROMS EVIDENCE

- Systematic reviews (e.g., Valderas et al., 2008)
- Effect of PROM feedback to clinicians on processes and outcomes of care.
- 2/3rd of studies show impact on processes
- Half show impact on outcomes
- Heterogeneity (,of course)



WHAT SEEMS TO WORK?

- Specific PROM training/supported implementation
- Timely feedback
- Well aligned with practice
- Information integrated into available systems
- Information that is linked to specific clinical action
- Interventions with a formal, structured, feedback process perform better (Krageloh, 2014)





RESEARCH

Patients' and doctors' views on depression severity questionnaires incentivised in UK quality and outcomes framework: qualitative study

Christopher Dowrick, professor of primary medical care, Geraldine M Leydon, senior research fellow, Anita McBride, research fellow, Amanda Howe, professor of primary care, Hana Burgess, academic foundation trainee, Pamela Clarke, research assistant, Sue Maisey, research associate, Tony Kendrick, professor of primary medical care.

WHAT DO PATIENTS THINK?

- PROMS for Depression in primary care in the UK
- Patients favoured the PROMS
- They saw them as an efficient and structured supplement to medical judgement
- They saw them as evidence that clinicians were taking their problems seriously through full assessment



UPDATED REVIEW PLANNED 2016



Routine provision of information on patient-reported outcome measures to healthcare providers and patients in clinical practice (Protocol)

Gonçalves Bradley DC, Gibbons C, Ricci-Cabello I, Bobrovitz NJH, Gibbons EJ, Kotzeva A, Alonso J, Fitzpatrick R, Bower P, van der Wees PJ, Rajmil L, Roberts NW, Taylor RS, Greenhalgh J, Porter I, Valderas JM



PREM PROBLEMS



How good is your GP at taking your problems seriously?

Very poor	Poor	Neither poor nor good	Good	Very good
9.1%		21.2%	69.7%	



PREM PROBLEMS



How good is your GP at taking your problems seriously?

Very poor	Poor	Neither poor nor good	Good	Very good
9.1%		21.2%	69.7%	



PREM PROBLEMS

People who said "Good" or "Very good"

"Taking your problems seriously? Well no he didn't."

"No, he didn't listen to me."

"Well no, he didn't really ask about symptoms."



IMPROVING MEASUREMENT IN CLINICAL PRACTICE

- Tackle the barriers!
- Make measurement more efficient
- Engage patients and clinicians with instant feedback
- Align feedback more closely with care
- Address response biases

IMPROVING MEASUREMENT IN CLINICAL PRACTICE

- Tackle the barriers!
- Make measurement more efficient
- Engage patients and clinicians with instant feedback
- Align feedback more closely with care
- Address response biases

Hypothesis:
All possible using ePsychometrics

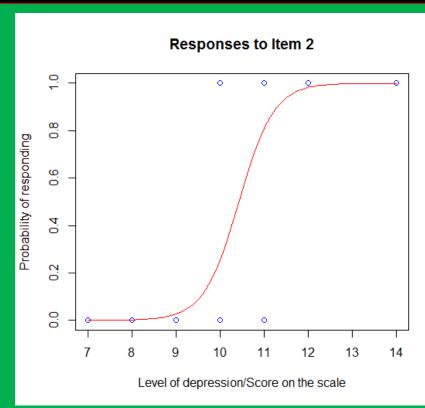
IMPROVING MEASUREMENT IN CLINICAL RESEARCH

ePsychometrics can also transform...

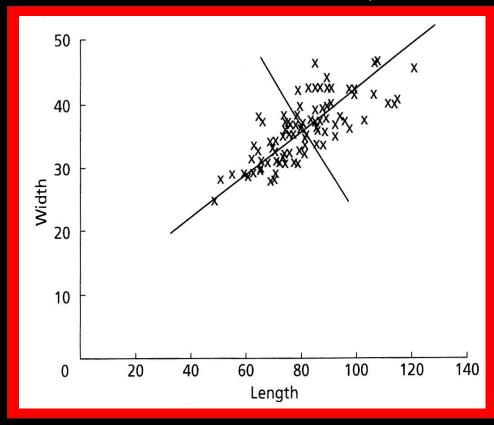
- Clinical trials (recruitment, reliability, attrition, comparability)
- Comparative research
- Patient-centred Big Data analytics

WHAT IS 'PSYCHOMETRICS'?

Item response theory



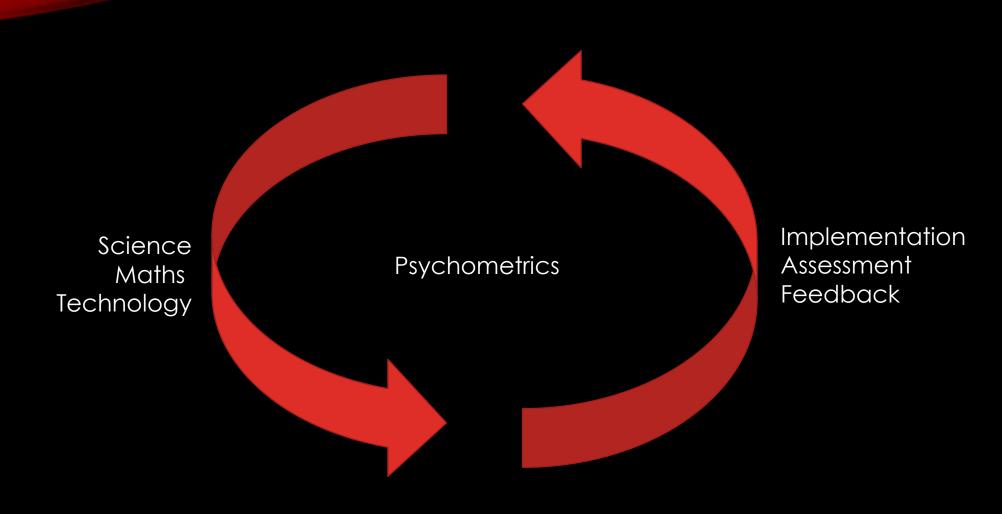
Classical test theory



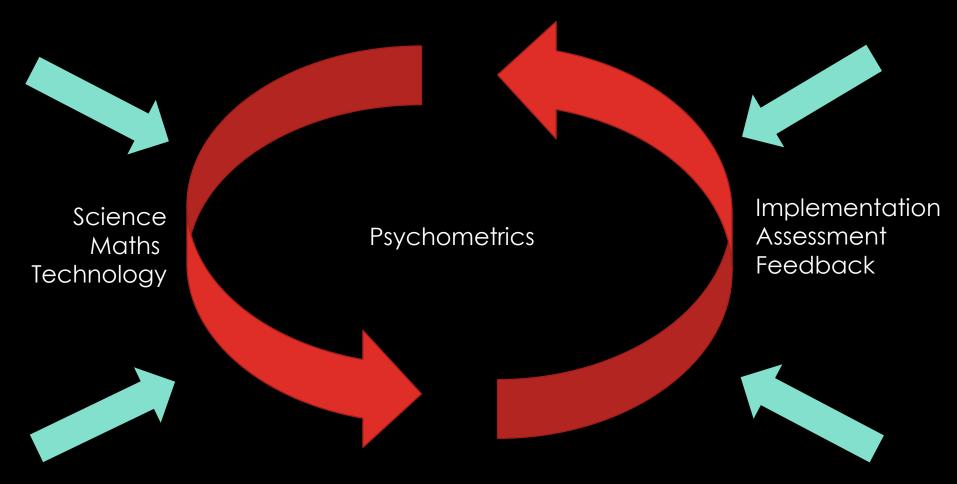
The Psychometrics Centre



Psychometrics



The Psychometrics Centre



Research, leadership, innovation, support

THREE TECHNIQUES & ONE TECHNOLOGY

Computer adaptive testing

(Efficiency and precision)

Prediction and feedback

(Validation and engagement)

Machine learning

(New forms of data)



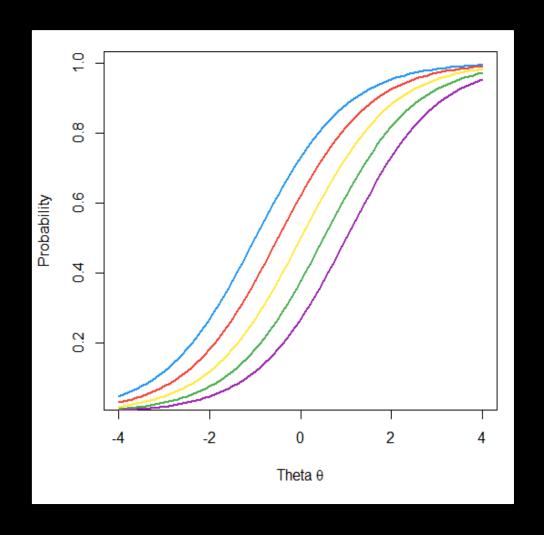
CONCERTO

DEVELOPING A PREDICTIVE COMPUTER ADAPTIVE TEST FOR THE WHOQOL QUALITY OF LIFE MEASURE



ITEM RESPONSE THEORY

- Probabilistic psychometric theory
- The more you have the more you agree...
- Sample free and test free
- Individual reliability
- Permits computer adaptive testing
- Allows predictions



- "Computer system which iteratively 'learns' about the test taker and selects the best item from a large 'bank' of items"
- Interfaces with item response theory
- Could be efficient, more reliable and better targeted
- Increasing use
- Used in international high stakes educational assessments (USA, UK and Australia)

Item Bank









Computer adaptive test



Participant



Item Bank



Computer adaptive test



Participant



Item Bank

2

3 4

 $\left(\begin{array}{c}6\end{array}\right)\left(\begin{array}{c}5\end{array}\right)$

 $\left(\begin{array}{c}7\end{array}\right)\left(\begin{array}{c}8\end{array}\right)$

Computer adaptive test

Participant



Item



1

Item bank

Computer

adaptive test



Response

Participant



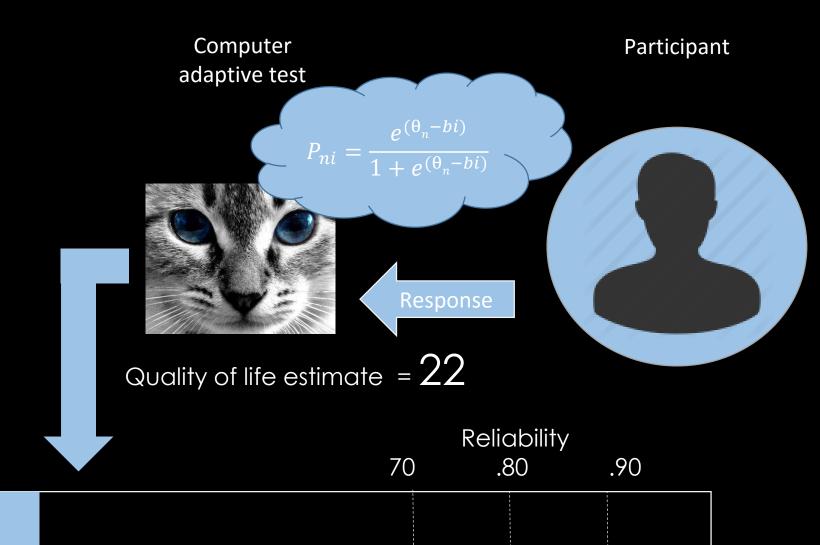
Item bank

2

3 4

6 5

7



Item Bank

2

3 4

6 5

7 8

Computer adaptive test

Best item

Quality of life estimate =22

Participant



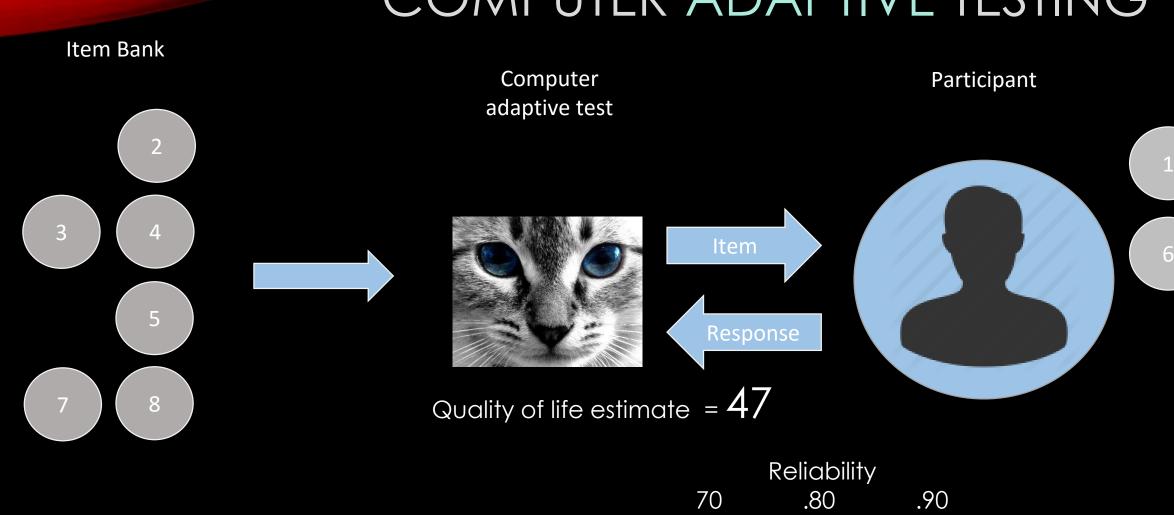
Reliability

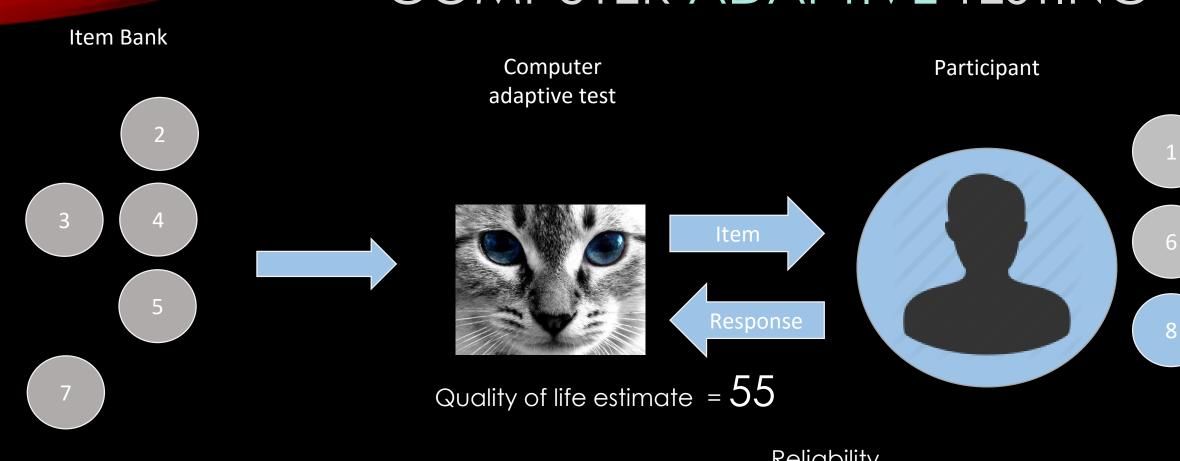
08.

70

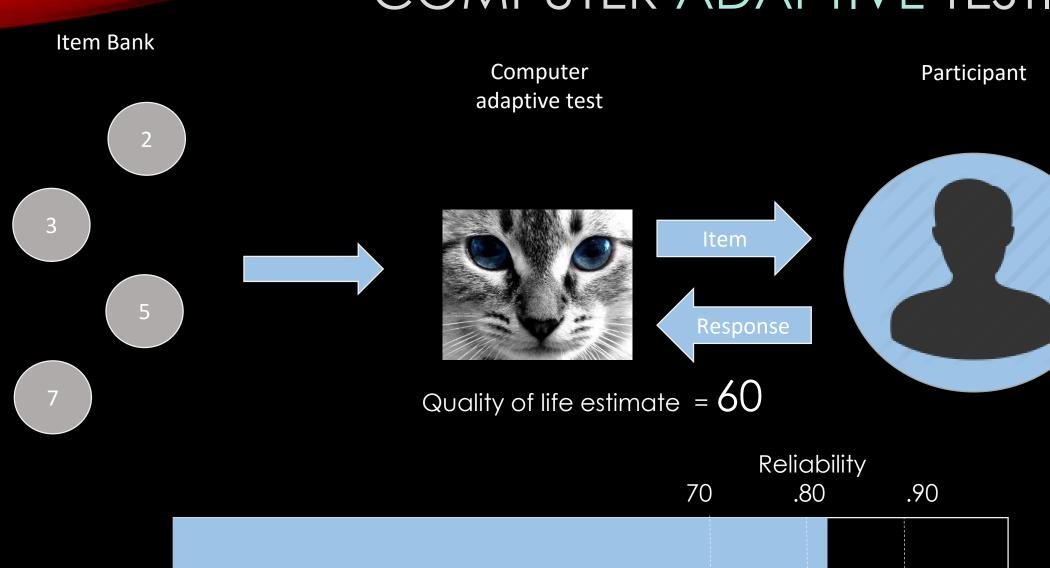
.90

1





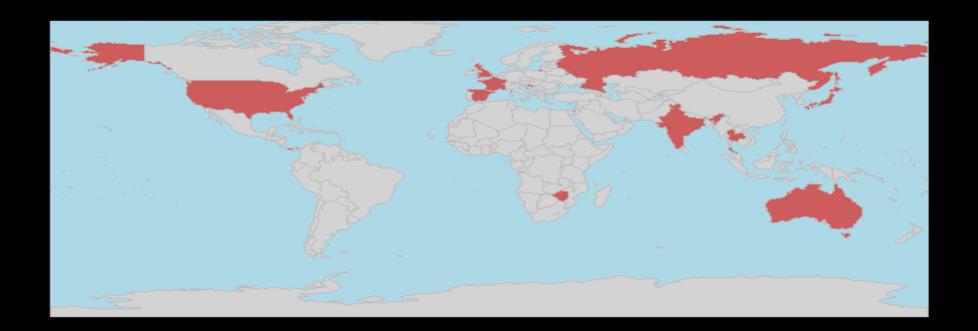






The WHOQOL

- Multi-dimensional measure of global quality of life
- Physical, psychological, social and environmental domains
- Fifteen field centers



Item bank development

- IRT analysis of the WHOQOL-100 on UK sample (n = 320)
- Four domains fit the Partial Credit Model (p>0.5)
- 52 items removed
- Mean 11 items per bank
- Banks were suitable for patients with long-term conditions

Computer adaptive test

- WHOQOL can be 82% shorter *and* more reliable
- (Gibbons et al., In Press. J Med Int Res)

• WHOQOL-CAT assessments are accurate and comparable between countries (Gibbons *et al.*, 2015 .Qual Life Res)





Computer adaptive test



Personalised Quality of Life (QoL) feedback Unique Identifier - S5QHJ

Scale	Score
Physical QoL	60
Psychological QoL	71
Social QoL	53
Environmental QoL	76

The scores above are worked out from the answers that you gave to the questions you have just completed. Your results are given on the chart above. A higher score means that you have a higher quality of life. The table below gives you a little bit more information about what each of the scores mean.

Physical Quality of Life

Your physical quality of life includes things like how well you are able to move around, how much energy you have or how much you are in pain.

Your score of 60 on this scale indicates that your physical quality of life is normal. The majority of people in the United Kingdom report a similar quality of life to you.

If you are worried about your physical quality of life then you should contact your doctor. You can <u>click here to find a local doctor</u> or <u>click here to access local emergency services</u>

PREDICATE STUDY

- <u>Predictions in a Computer Adaptive Testing Environment</u>
- Assess validity in the real-world
- CAT is completed until stopping rule is met
- An algorithm predicts what test-takers would answer to unseen questions
- Test-takers rate the predictions (Correct, close, wrong)
- PREDICTMYQOL.com



PREDICATE module



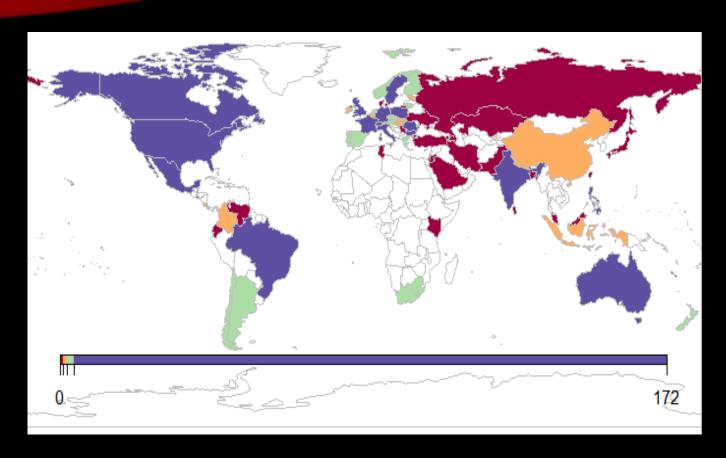
Here are the four questions and the answers we predict that you would have given, based on your previous responses.

Question 1	Do you have enough energy for everyday life?	
Answer	Moderately	

We are 69% sure that we're right, what do you think?

Exactly right, just what I was thinking!	
Close, but it's not quite right	
The prediciton is wrong, I was thinking of a very differe	nt answer

PREDICATE Study



- 699 completions (52 countries)
- 173 participants from the UK
- Average age = 45 ± 10
- 56% male
- 50 patients from UK report long-term conditions (diabetes, depression, arthritis, chronic obstructive pulmonary disease, coronary heart disease)

PREDICATE results

• For the UK sample, results were, on average, 94% "right" or "close"

- 'Exactly right' 69% of the time (predicted accuracy was 78%)
- "Close, but not quite right" 25% of the time
- "Wrong" 6% of the time [non-UK 10% wrong]

- Prediction performance was better in the UK (where item bank is validated) than in the rest of the world

ATLanTiC conclusions

- The item bank is brief and can be administered in 120 seconds
- As much as a 82% increase in efficiency compared to paper-based versions
- Simulations translated well to 'real world' CAT
- Prediction study supports the validity of the measurement model and 'backfilling' unanswered item bank items
- Feedback well received and instantaneous
- People like the predictions

USING MACHINE LEARNING TO MAKE SENSE OF GP PERFORMANCE DATA



OPEN TEXT FEEDBACK

- Many questionnaires include open-text elements to add further information
- May contain important information missed by questionnaires
- Typically underused / ignored
- Time-consuming to use human analysts

PATIENT EXPERIENCE









We can make sense of this data using machine learning and natural language processing



GMC Colleague Questionnaire

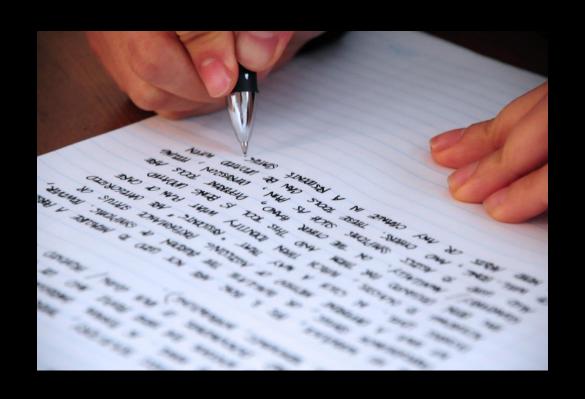


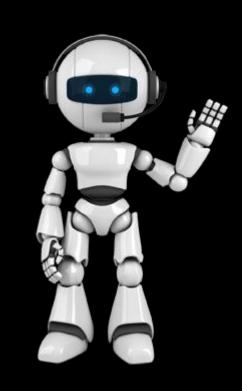
1636 Comments

> 548 Doctors

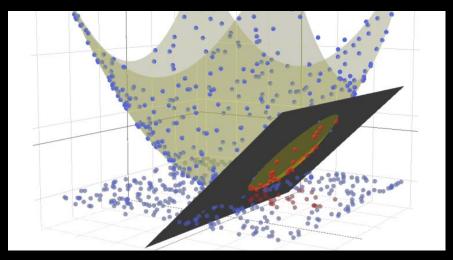
20-item questionnaire

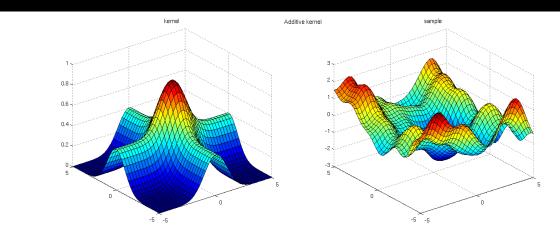
Free-text information





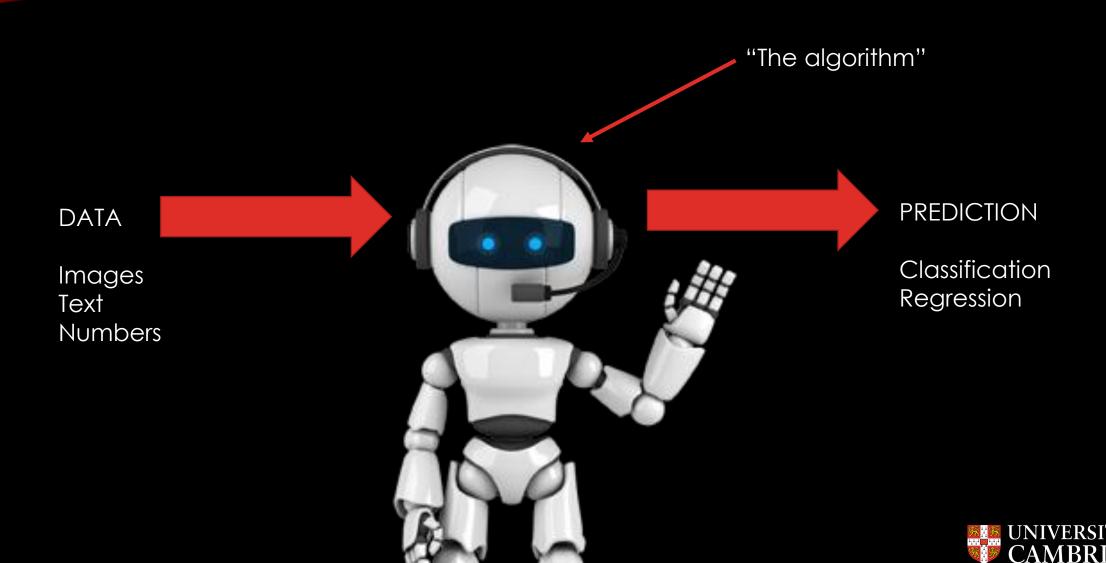


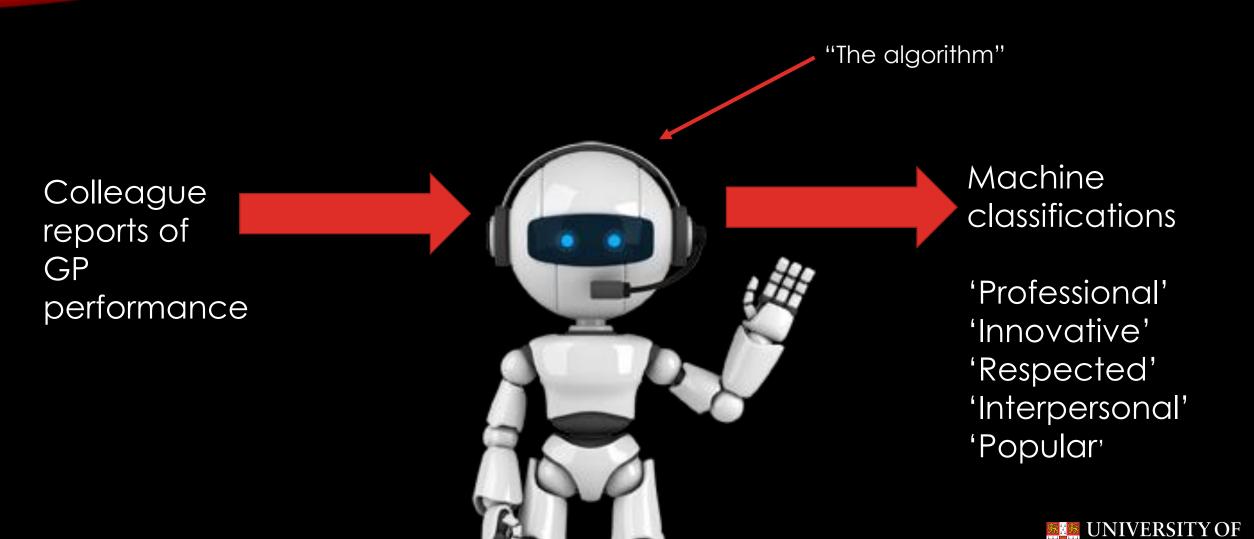










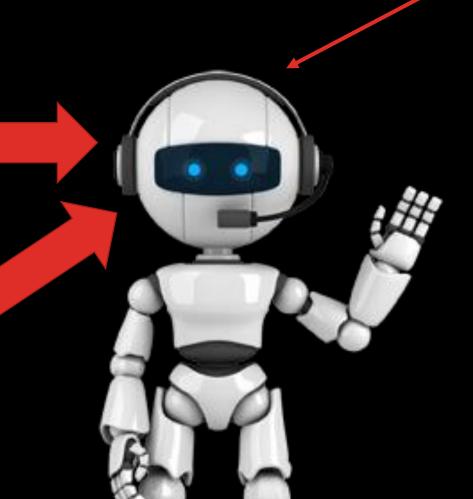


TRAINING

"The algorithm"

Colleague reports of GP performance (50% of comments)

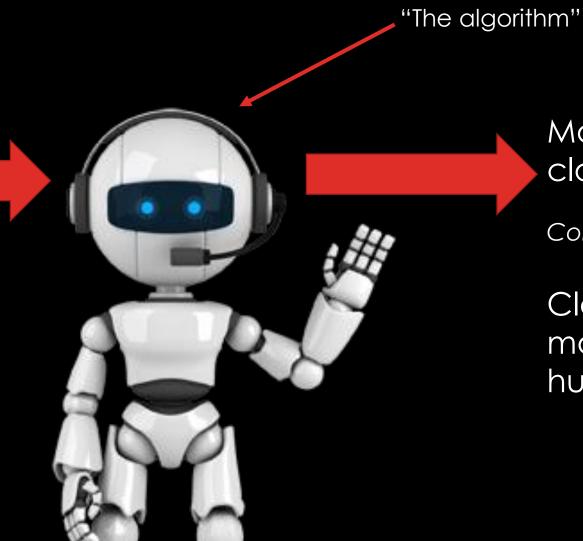
Classifications made by human raters





VALIDATING

Colleague reports of GP performance



Machine classifications

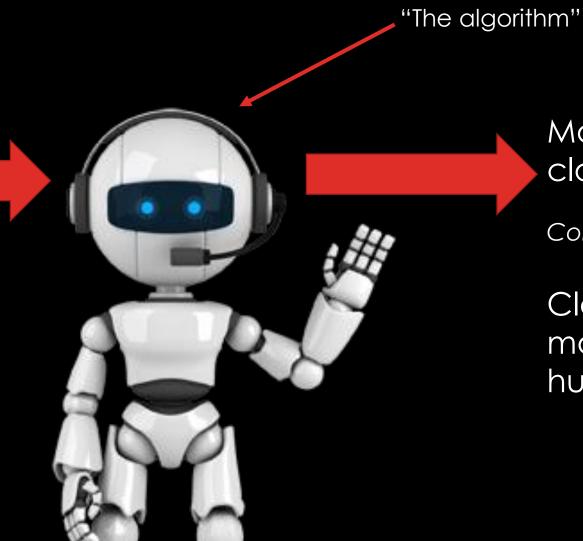
Compared with

Classifications made by human raters



VALIDATING

Colleague reports of GP performance



Machine classifications

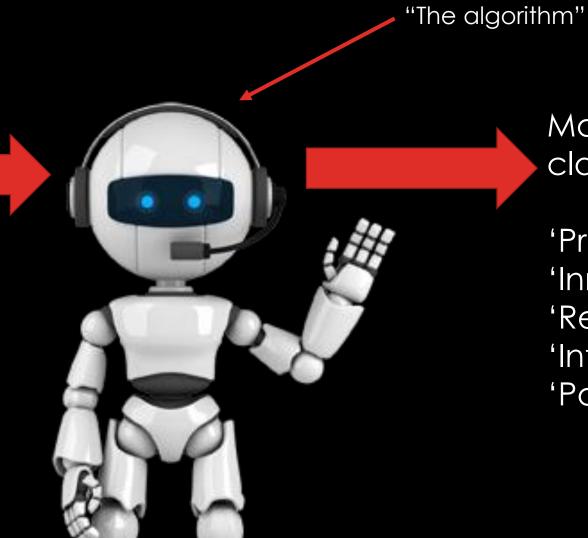
Compared with

Classifications made by human raters



DEPLOYMENT

Colleague reports of GP performance



Machine classifications

'Professional'
'Innovative'
'Respected'
'Interpersonal'
'Popular'



COMMENTS

Theme	Comment	
Innovator	"It is clear from the advice he gives that he is aware of [the] current good practice, is highly motivated, very practical and very much a team player. His advice, when working with consultant colleagues was respected, and he recognised where practice/primary care limitations were and yet looked for opportunities for change and improvement."	
	"She has an admirable level of commitment and enthusiasm for her patients and her work. She has been instrumental in promoting change and improvement in her department. She is a great asset to the department and the hospital."	
Interpersonal	"She is a very good, committed colleague always keen to improve, very liked by her patients and highly valued by all who work with her."	
	"Very approachable and professional."	
Popular	"Excellent well liked and easy working colleague."	
Popular	"Very popular doctor. Works to high standards."	
Duefeesienel	"I find this doctor to be very efficient, caring, honest and very professional."	
Professional	"I find that he very easy and helpful to work with, he always has time for patients and staff."	
Respected	"A first class colleague."	
Respected	"Pleasant and valued colleague."	

INTER-RATER AGREEMENT

<u>Code</u>	Agreement (Kappa)	
Innovator	.98	
Interpersonal	.80	
Popular	.97	
Professional	.82	
Respected	.87	



<u>Performance</u>		
<u>N</u>	<u>T-test</u>	<u>Sig.</u>
48	.00	.99
435	1.98	.04
107	88	.38
,	2.51	.01
10	_	<.001
·		.001
	48	N T-test 48 .00 435 1.98 107 88 643 2.51 264 3.75



<u> </u>	<u>Code</u>	Performance		
		N	<u>T-test</u>	<u>Sig.</u>
	Innovator	48	.00	.99
In	terpersonal	435	1.98	.04
	Popular	107	88	.38
F	Professional	643	2.51	.01
	Respected	264	3.75	<.001
Ai	ny category	1069	•77	.001



<u>Code</u>	<u>Performance</u>		
	<u>N</u>	<u>T-test</u>	Sig.
Innovator	48	.00	.99
Interpersonal	435	1.98	.04
Popular		88	.38
Professional	643	2.51	.01
		2.51	.01
Respected	264	3.75	<.001
Any category	1069	·77	.001



<u>Code</u>	Performance		
	<u>N</u>	<u>T-test</u>	Sig.
Innovator	48	.00	.99
Interpersonal	435	1.98	.04
Popular	107	88	.38
Professional	643	2.51	.01
Respected	-	3.75	<.001
Any category		·77	.001



 Machine learning algorithms can classify open-text reports of doctors' performance with human-level accuracy

 Machine-made classifications can signal significant differences in doctors' performance



SUMMARY SO FAR

 Computer adaptive testing is an acceptable way to make assessments shorter and more reliable

 Machine learning may be used to make sense of open-text data in questionnaires



The problem



Concerto

Open source tool for the development of online assessments utilising computer adaptive testing and tailored feedback





The solution



Research, tools, training, and support

People



Prof John Rust Director Psychometrics Centre



Przemyslaw Lis Software Engineer University of Cambridge



Dr David Stillwell Deputy Director Psychometrics Centre



Dr Tomoya Okubo Psychometrician NCUEE Tokyo



Dr Michal Kosinski Computational Psychology Stanford University



Dr Chris Gibbons Director of eHealth Assessment Psychometrics Centre

Vision

- Adaptive and electronic measurement should be widely available to non-experts
- It should always be free and open-source
- Assessment experience should be made as enjoyable as possible
- Data security protocol should be dictated by the test-developer (your servers, not ours)
- Where expertise is needed, it be should be readily available as consultancy, training, or free guides online





open source_

















open source_

Collaborators

Patient and stakeholder feedback

Case-specific adaptation

C

Deployment of open-source software

Better products More widely used

More patient benefit!

Brand recognition
Academic collaboration
Consultancy
Skills and learning
Paid support



In practice













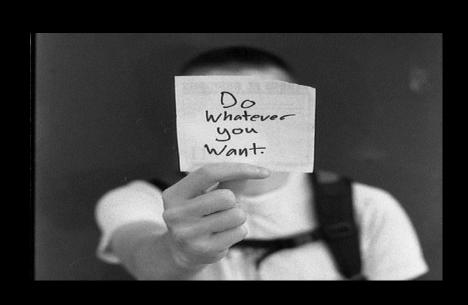






Flexibility

- Item type (text, pictures, sounds, movies)
- Test type (Computer adaptive test, short-form questionnaire)
- Scoring (IRT scoring or sum scores)
- Feedback (patient/clinician/graphs/text/links/e-mail)
- Longitudinal assessment
- 'Log on' details
- Facebook connection
- Data security
- Data storage



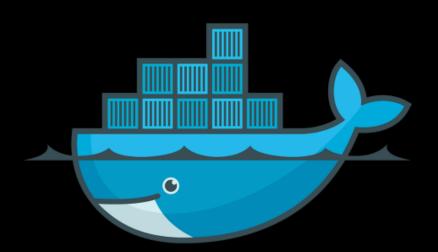


Concerto



<u>Installable on different platforms</u>

- Cloud servers (free trial on Amazon)
- Local server
- (at your University/hospital)
- Docker
- Windows 10 tablet
- Linux Installation
- User-defined data security
- Flexible speed and data-storage for scalable tests (up to millions of users)







BONUS STUDY

PREDICTING PERSONALITY FROM DIGITAL FOOTPRINT

MYPERSONALITY

- Example of Big Data analytics in social science, using personality questionnaires
- Big Data widely used by Big Business
- Google, Facebook, Apple, Amazon, Barclays...
- "Inevitable application in health care..." (Murdoch & Detsky, 2013, JAMA)
- What might big data analytics using questionnaires or PROMs look like?



MYPERSONALITY

Psychometric tests (BIG-5) hosted on Facebook

Developed and managed by psychometrics centre members

More than <u>3,100,000</u> people have completed the Big-5 questionnaire

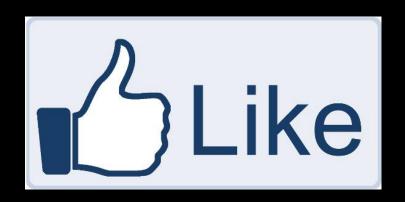
Large proportion of users shared their Facebook 'Like' information



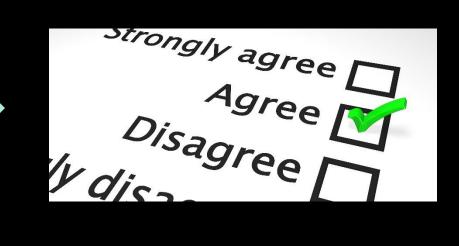
MYPERSONALITY

Question:

Can you build a predictive model of a psychometric test (BIG-5) using Facebook 'likes'









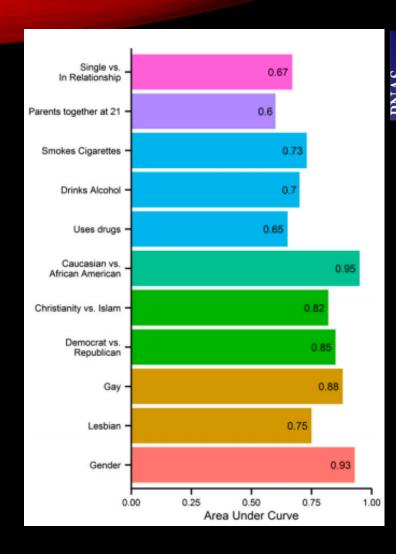
LIKES AND ITEMS



Do you like University of Birmingham? Yes | No



MY PERSONALITY



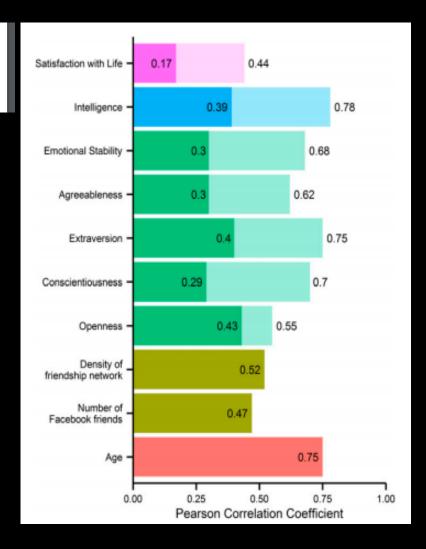
Private traits and attributes are predictable from digital records of human behavior

Michal Kosinski^{a,1}, David Stillwell^a, and Thore Graepel^b

*Free School Lane, The Psychometrics Centre, University of Cambridge, Cambridge CB2 3RQ United Kingdom; and *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 browsing logs (11–15). Similarly, it has been shown that personality can be predicted based on the contents of personal Web sites (16),





MY PERSONALITY

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

Michal Kosinski^{a,1}, David Stillwell^a, and Thore Graepel^b

*Free School Lane, The Psychometrics Centre, University of Cambridge, Cambridge CB2 3RQ United Kingdom; and hMicrosoft Research, Cambridge CB1 2FB,

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 browsing logs (11–15). Similarly, it has been shown that personality Likes, can be used to automatically and accurately predict a range can be predicted based on the contents of personal Web sites (16),

Trait		Selected most predictive Likes			
ÒI	High	The Godfather Mozart Thunderstorms The Colbert Report Morgan Freemans Voice The Daily Show Lord Of The Rings To Kill A Mockingbird Science Curly Fries	Jason Aldean Tyler Perry Sephora Chiq Bret Michaels Clark Griswold Bebe I Love Being A Mom Harley Davidson Lady Antebellum	Low	
Satisfaction With Life	Satisfied	Sarah Palin Glenn Beck Proud To Be Christian Indiana Jones Swimming Jesus Christ Bible Jesus Being Conservative Pride And Prejudice	Hawthorne Heights Kickass Atreyu (Metal Band) Lamb Of God Gorillaz Science Quote Portal Stewie Griffin Killswitch Engage Ipod	Dissatisfied	



MY PERSONALITY

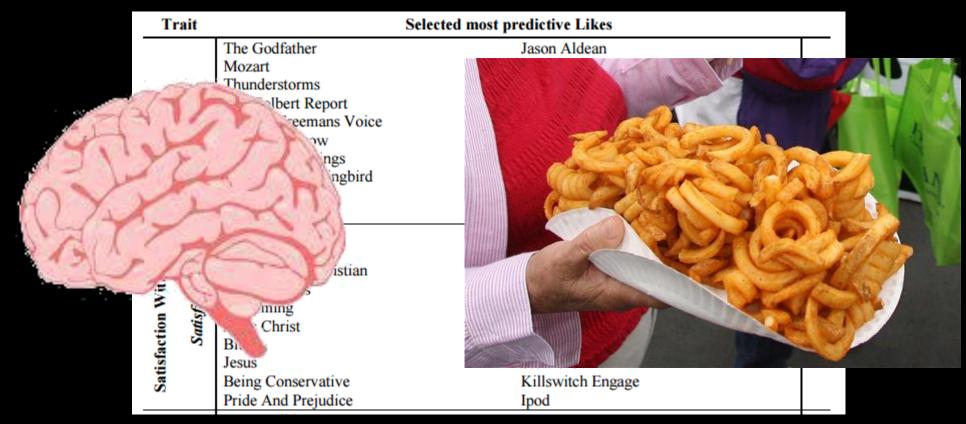
Private traits and attributes are predictable from digital records of human behavior

Michal Kosinski^{a,1}, David Stillwell^a, and Thore Graepel^b

*Free School Lane, The Psychometrics Centre, University of Cambridge, Cambridge CB2 3RQ United Kingdom; and *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 can be predicted based on the contents of personal Web sites (16),

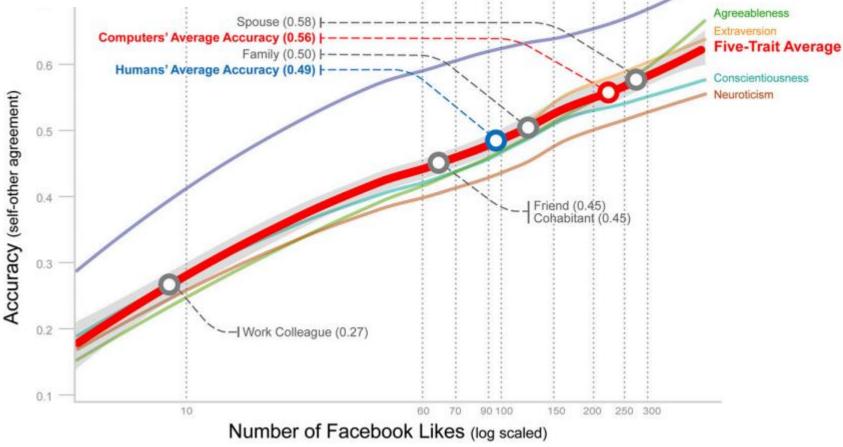




Wu Youyou^{a,1,2}, Michal Kosinski^{b,1}, and David Stillwell^a

^aDepartment of Psychology, University of Cambridge, Cambridge CB2 3EB, United Kingdom; and ^bDepartment of Computer Science, Stanford University, Stanford, CA 94305

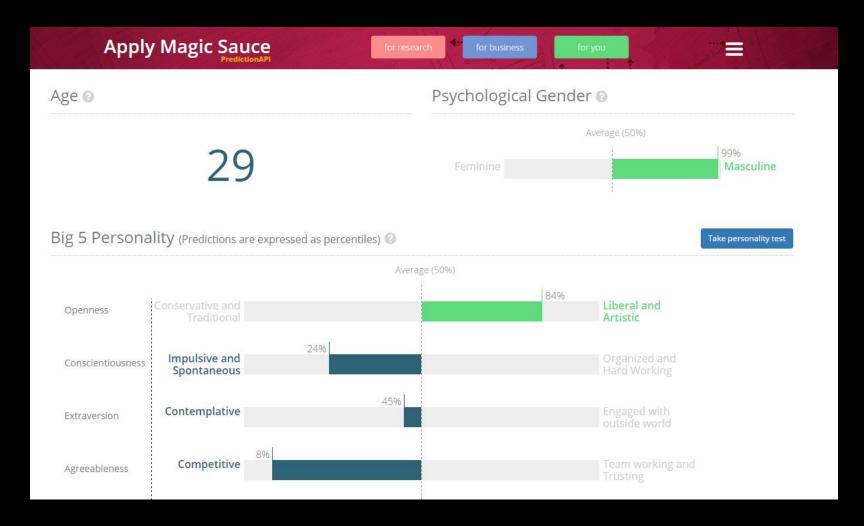
Edited by David Funder, University of California, Riverside, CA, and accepted by the Editorial Board December 2, 2014 (received for review September 28, 2014)





APPLYMAGICSAUCE.COM

- API which allows companies to make personality predictions at scale
- Hilton, Wrigleys, Barclays
- Using psychological information to tailor information improves conversion rate and satisfaction (Matz, 2015)





LEARNING POINTS

On feedback

- 3,100,000 people completed a personality test that had up to 336 items!
- No monetary incentive just feedback
- People want to learn about themselves!
- Research planned how does feedback affect response rates, reliability etc

On data sharing and collaborative / open science

- All these data (apart from Likes) are available for free to anyone (since 2012)
- No regrets!



WHY MIGHT THIS 'INFERENTIAL' APPROACH BE USEFUL FOR HEALTH SCIENCES

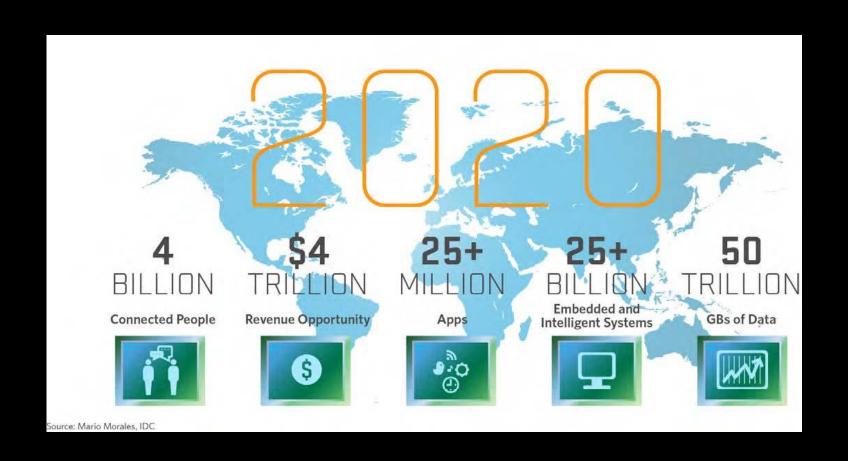
- Data exist on a massive scale (and increasingly so)
- 'Learning data cities' are being planned (cf. Buchan)
- May be less biased than questionnaire items
- Real-time monitoring/assessment/feedback
- No measurement latency (can be done anywhere)



WHAT MIGHT WE USE FOR 'INFERENTIAL PSYCHOMETRICS'?

- Facebook likes
- Twitter 'follows'
- Twitter updates
- Images (image recognition e.g., Facebook profile photos)
- Instagram
- 'Wearables'
- Geo-location data
- Movement data
- Phone activity
- Interactions with friends (proximity of two devices)

INTERNET OF THINGS





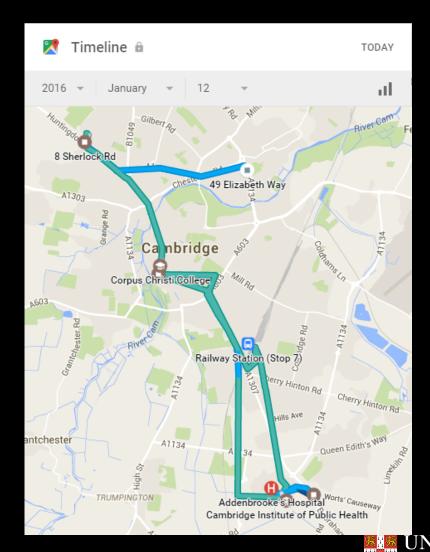
SF-36 & GOOGLE DATA?

The following questions are about activities you might do during a typical day. In the past 1-week does your health limit you in these activities? If so, how much?

(Please circle one number on each line)

ACTIVITIES			Yes Limited A little	No, Not Limited At All
3a:	Vigorous activities, such as running, lifting heavy Objects, participating in strenuous sports	1	2	3
3b:	Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	1	2	3
3c:	Lifting or carrying groceries	1	2	3
3d:	Climbing several flights of stairs	1	2	3
3e:	Climbing one flight of stairs	1	2	3
3f:	Bending, kneeling, or stooping	1	2	3
3g:	Walking more than one kilometre	1	2	3
3h:	Walking half a kilometre	1	2	3
3i:	Walking 100 metres	1	2	3
3g ww:	Wheeling more than one kilometre	1	2	3
3h ww:	Wheeling half a kilometre	1	2	3
3i ww:	Wheeling 100 metres	1	2	3
3j:	Bathing or dressing yourself	1	2	3

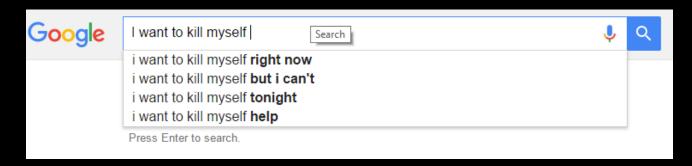
^a Modified from SF-36¹: Items 3 (a to j) are the original SF-36 questions, while 3g ww to 3i ww (shaded area) comprise the supplementary SF-36ww modification.



MENTAL HEALTH AND GOOGLE DATA?

 Do you have thoughts of suicide?

Have you been experiencing delusions?







GOOGLE AND THOMAS INSEL

Former head of NIMH in the USA

Moved to Google in late 2015.

"Technology can have greater impact on mental healthcare than on the care for heart disease, diabetes, cancer or other diseases...

It could transform this area in the next five years." (Insel, 2015)





IS USING DATA LIKE THIS A BIT CREEPY*?

*OFF-PUTTING TO PATIENTS



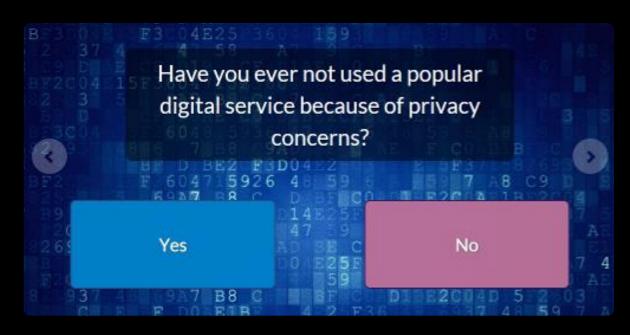
Assessment

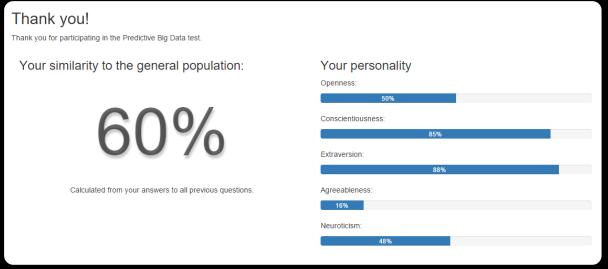
- 27 Yes/No Questions on aspects of Big Data
- Measuring attitudes towards Personalisation, Ethics, Wearable Tech, Internet of Things, Finance, Policy-Making, Cloud Storage, etc.
- 20-item BIG5 Personality test at the end

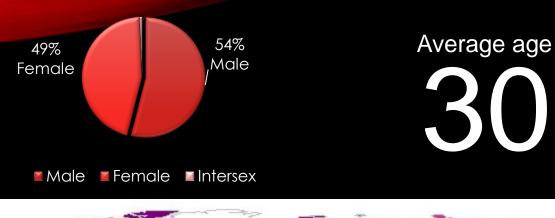
Feedback

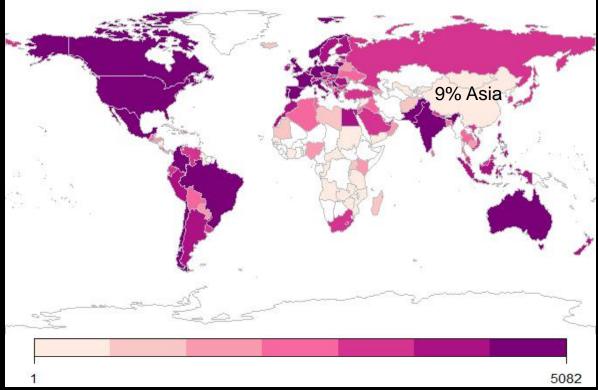
 Participants given real-time feedback on the similarity of their answers to all previous participants

PredictiveDataProject.com









Sample (by Oct '15)

No. of test sessions in total (Sept)

33,937

Answered At Least One Question

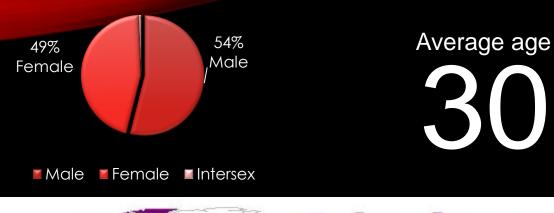
19,126

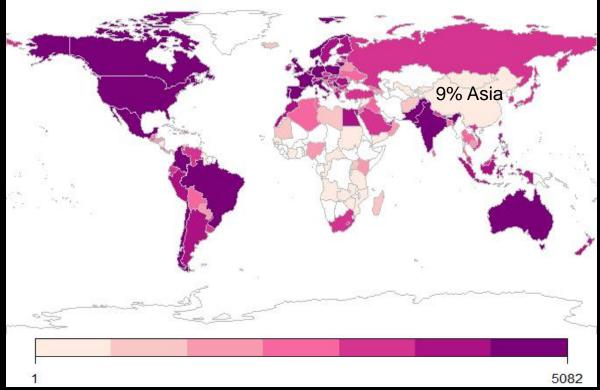
All Big Data Questions

10,411

All Personality Questions

8,871





Sample (by Oct '15)

No. of test sessions in total (Sept)

33,937

Answered At Least One Question

19,126

All Big Data Questions

10,411

All Personality Questions

8,871

Results: Audit of Global Public Opinion

Are those with access to your personal data able to accurately predict your future behaviour? (n=19,100)

62% NO 38% YES

Should predictive technologies be used to improve the quality of healthcare, for example by helping doctors to recommend personalised nutrition and exercise plans? (n=14,187)

16% NO 84% YES

Should predictive technologies be used assess your eligibility for a mortgage? (n=13,725)

62% NO 38% YES

Do you think your organisation ought to invest in predictive technologies? (n=2,830 CMOs)

62% NO 73% YES

Is it important for you to understand the psychological attributes of your customers? (n=2,489 CMOs)

8% NO 92% YES

PATIENTS AND SOCIAL MEDIA

Downloaded from http://qualitysafety.bmj.com/ on February 9, 2016 - Published by group.bmj.com

BMJ Quality & Safety Online First, published on 13 October 2015 as 10.1136/bmigs-2015-004489

ORIGINAL RESEARCH

Linking social media and medical record data: a study of adults presenting to an academic, urban emergency department

Kevin A Padrez, 1,2 Lyle Ungar, 2,3,4 Hansen Andrew Schwartz, 2,3,4 Robert J Smith, 2,4,5 Shawndra Hill, 2,6 Tadas Antanavicius, 2,4 Dana M Brown, 2,5 Patrick Crutchley, 2,4 David A Asch, 2,7 Raina M Merchant 2,5

71% of patients
were happy to
share social media
information with
their doctor



CONCLUSION

- Modern psychometrics offers many opportunities for improved research and clinical practice
- Computer adaptive tests work well in the 'real world'
- Predictions are engaging
- myPersonality showed what will/could be possible for 'inferential psychometrics' in health research and the power of feedback
- The future is exceptionally exciting in this field!



THANK YOU!

cg598@cam.ac.uk

concertoplatform.com drchrisgibbons.com

@DrCGibbons

@concerto_cam

@campsych

