University of BRISTOL

REPHRAIN

National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online

bristol.ac.uk

National Research Centre on
Privacy, Harm Reduction
and Adversarial Influence
Online

REPHRAIN
Protecting citizens online

Psychological Microtargeting in Online Environments

Adam Sutton – NEWS project

REPHRAIN
Protecting citizens online

Acknowledgments

Dr Almog Simchon

Dr Matthew Edwards

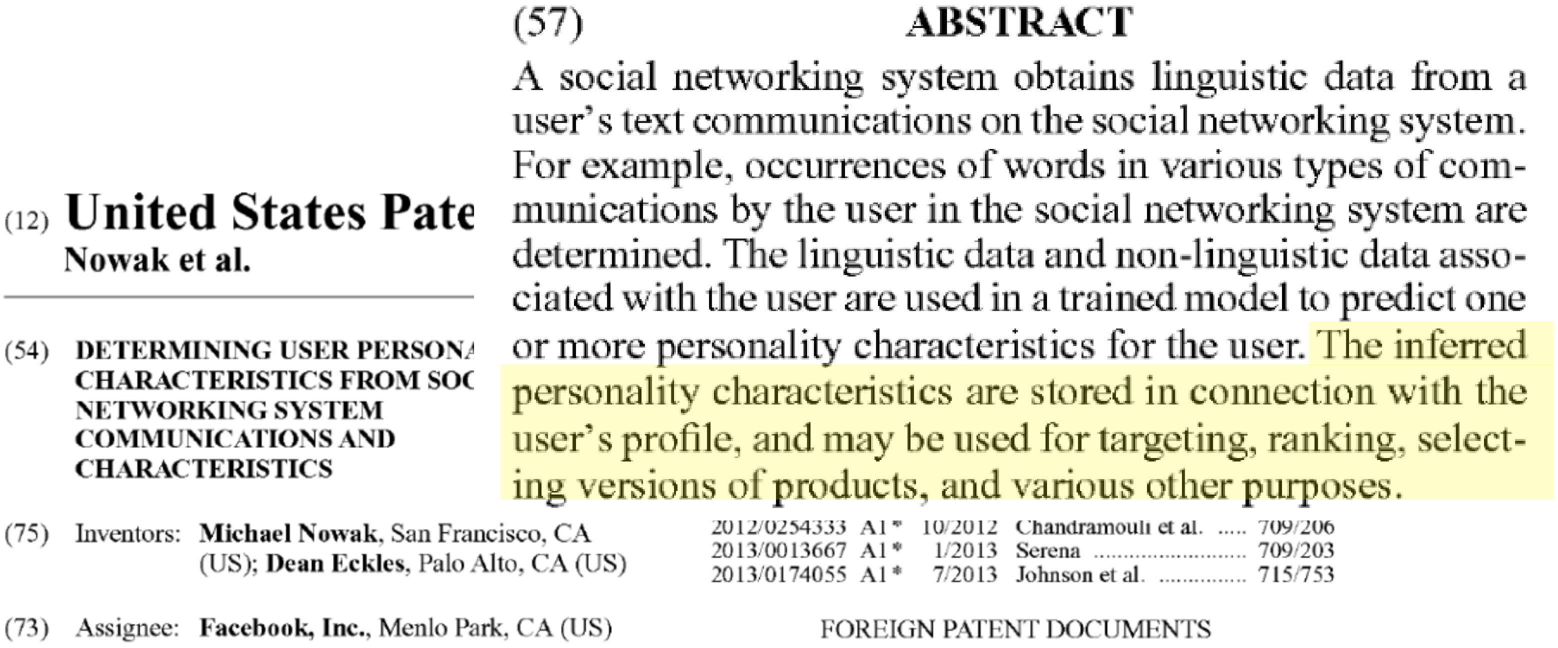
Prof Stephan Lewandowsky

TeDCog

How does Psychological Microtargeting work in political campaigns?

- Target undecided voters
- Derive personality makeup for *each* individual
- Construct a persuasive message to sway individuals for political action or inaction based on ideology AND personality

Who does this?



(57) **ABSTRACT**

A social networking system obtains linguistic data from a user's text communications on the social networking system. For example, occurrences of words in various types of communications by the user in the social networking system are determined. The linguistic data and non-linguistic data associated with the user are used in a trained model to predict one or more personality characteristics for the user. The inferred personality characteristics are stored in connection with the user's profile, and may be used for targeting, ranking, selecting versions of products, and various other purposes.

(12) **United States Pate**
Nowak et al.

(54) **DETERMINING USER PERSON
CHARACTERISTICS FROM SOC
NETWORKING SYSTEM
COMMUNICATIONS AND
CHARACTERISTICS**

(75) Inventors: **Michael Nowak**, San Francisco, CA (US); **Dean Eckles**, Palo Alto, CA (US)

(73) Assignee: **Facebook, Inc.**, Menlo Park, CA (US)

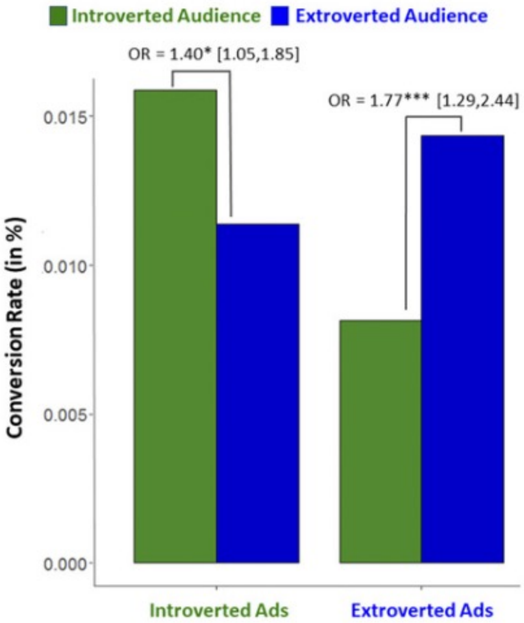
2012/0254333	A1*	10/2012	Chandramouli et al.	709/206
2013/0013667	A1*	1/2013	Serena	709/203
2013/0174055	A1*	7/2013	Johnson et al.	715/753

FOREIGN PATENT DOCUMENTS

Personality-based Microtargeting

- Does it work?
  - Evidence suggests it does (Matz et al., 2017; *PNAS*)
- Is it a problem?
  - Depend on who you ask: unaccepted in Germany but passes in the US
  - Personalization for political campaigning is unaccepted across the board

  (Kozyreva et al, 2021; *Humanit. Soc. Sci. Commun*)



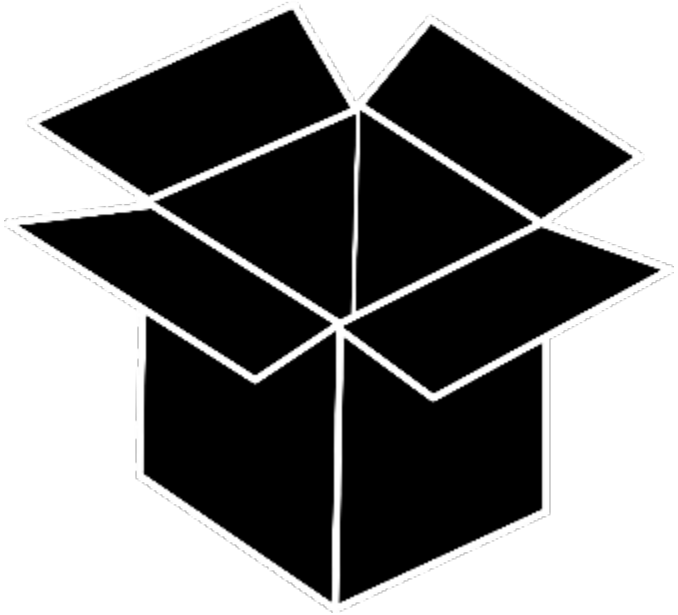
Solution

- Boosting: empowering individuals to make **informed** decisions
  - Letting people know psychological microtargeting exists + information about their personality leads to accurate identification of such attempts (Lorenz-Spreen et al., 2021; *Sci Rep*)

- How can we boost individuals in online environments?
- How can we know when people are being microtargeted?

REPHRAIN
Protecting citizens online

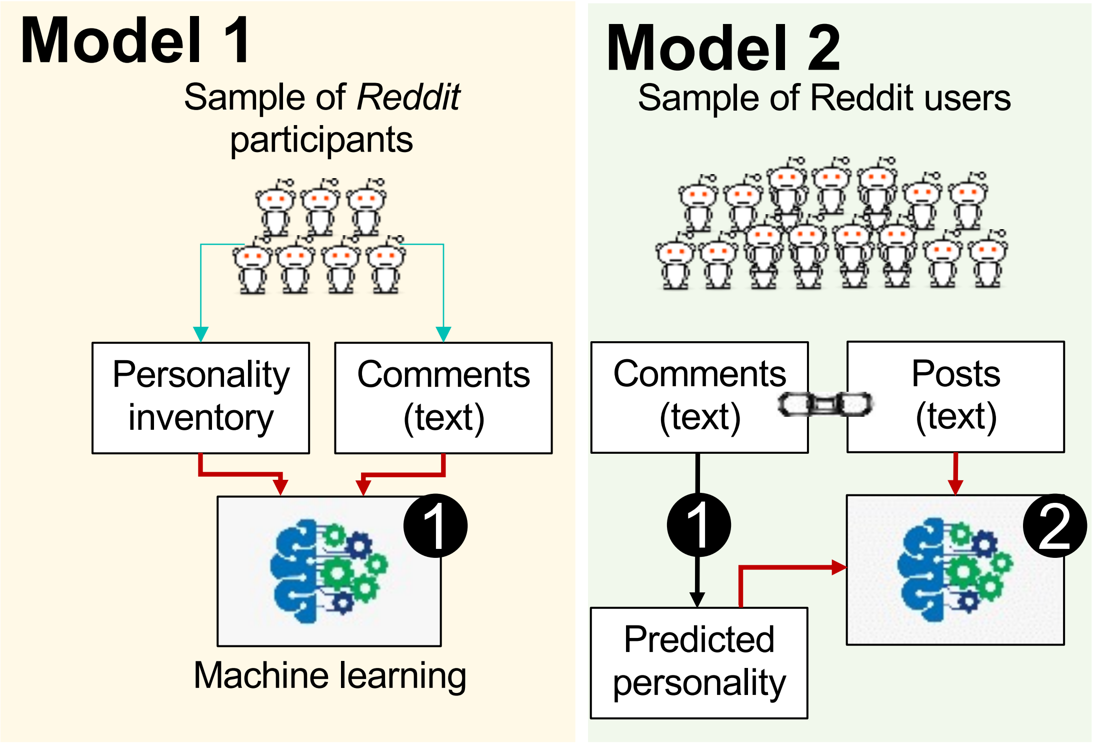
Uncover the algorithms in action



The Current Project

- Population: Reddit users of fiction-writing communities
- Text-based models:
  - Model 1: predict stable psychological characteristics based on the text people produce
  - Model 2: predict stable psychological characteristics based on the text people consume
- Find if indeed psychologically concordant messages are more persuasive
- Apply in the real world

Model 1 & Model 2



Model 1

Sample of *Reddit* participants

Personality inventory

Comments (text)

1

Machine learning

Model 2

Sample of Reddit users

Comments (text)

Posts (text)

1

2

Predicted personality

REPHRAIN
Protecting citizens online

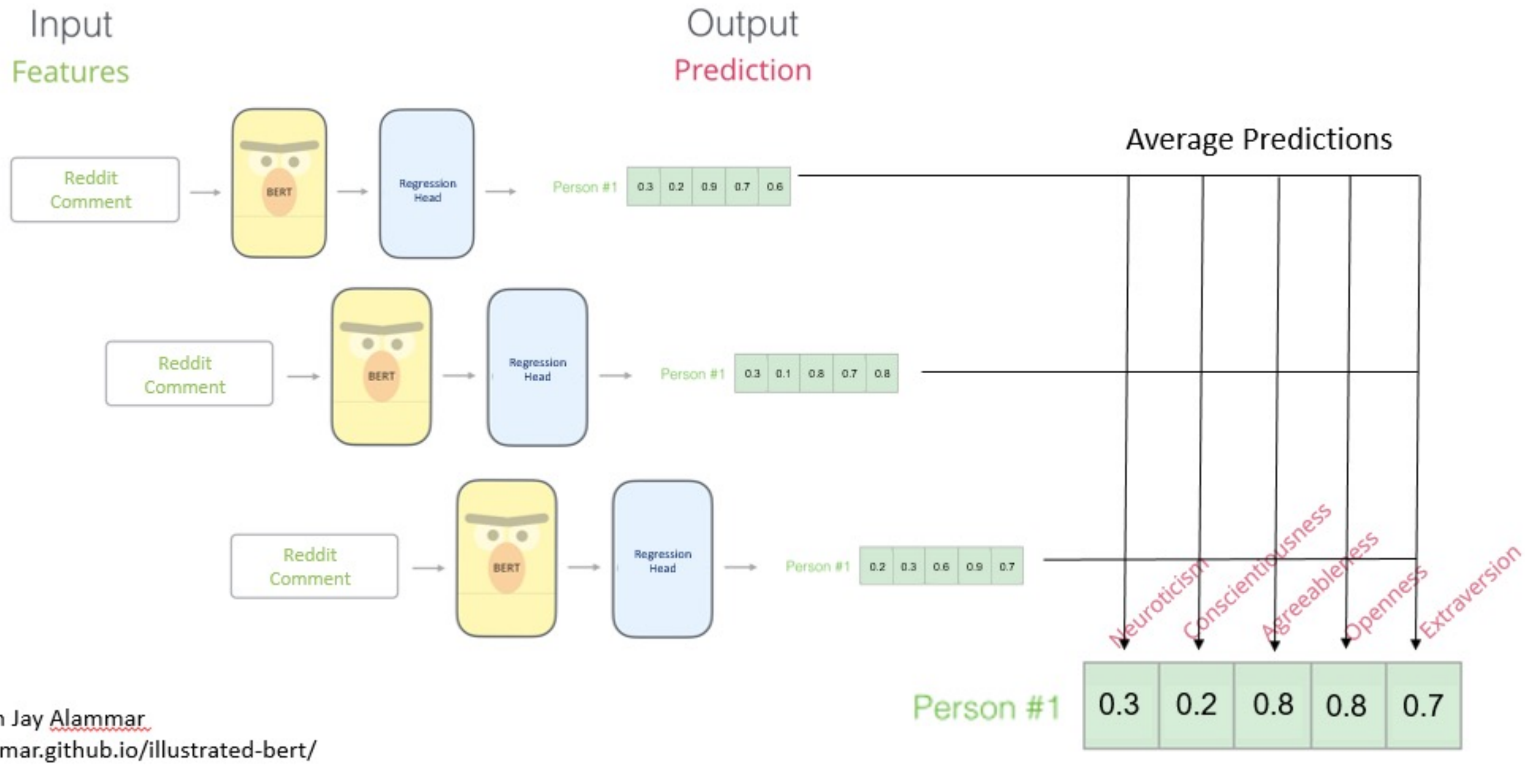
Model 1: Method

- NEWS - collection
  - Communities of Political News
  - 18,293 Potential Participants
  - 1,063 sent PMs
  - 123 participants
  - 290,000 comments
- Measures
  - BFI-2 (Soto & John, 2017; *JPSP*)
  - SVS-PVQ (Schwartz, 1992; 2012)

- VW Fiction collection
  - Communities of Fiction writing
  - 32,344 Potential Participants
  - 9,244 sent PMs
  - 1,100 participants
  - 650,000 comments
- Measures
  - BFI-2 (Soto & John, 2017; *JPSP*)
  - SVS-PVQ (Schwartz, 1992; 2012)

National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online

Model 1: Text Modeling



Adapted from Jay Alammar
https://jalammar.github.io/illustrated-bert/

Model 1: Results for NEWS

- Training set: 100% - Fiction
- Test set: News Set (N = 123)

  Average performance:

  Pearson's *r* = 0.24

  Correlation between the model predictions and ground truth
- Small test set limited to difficulties with reddit sampling

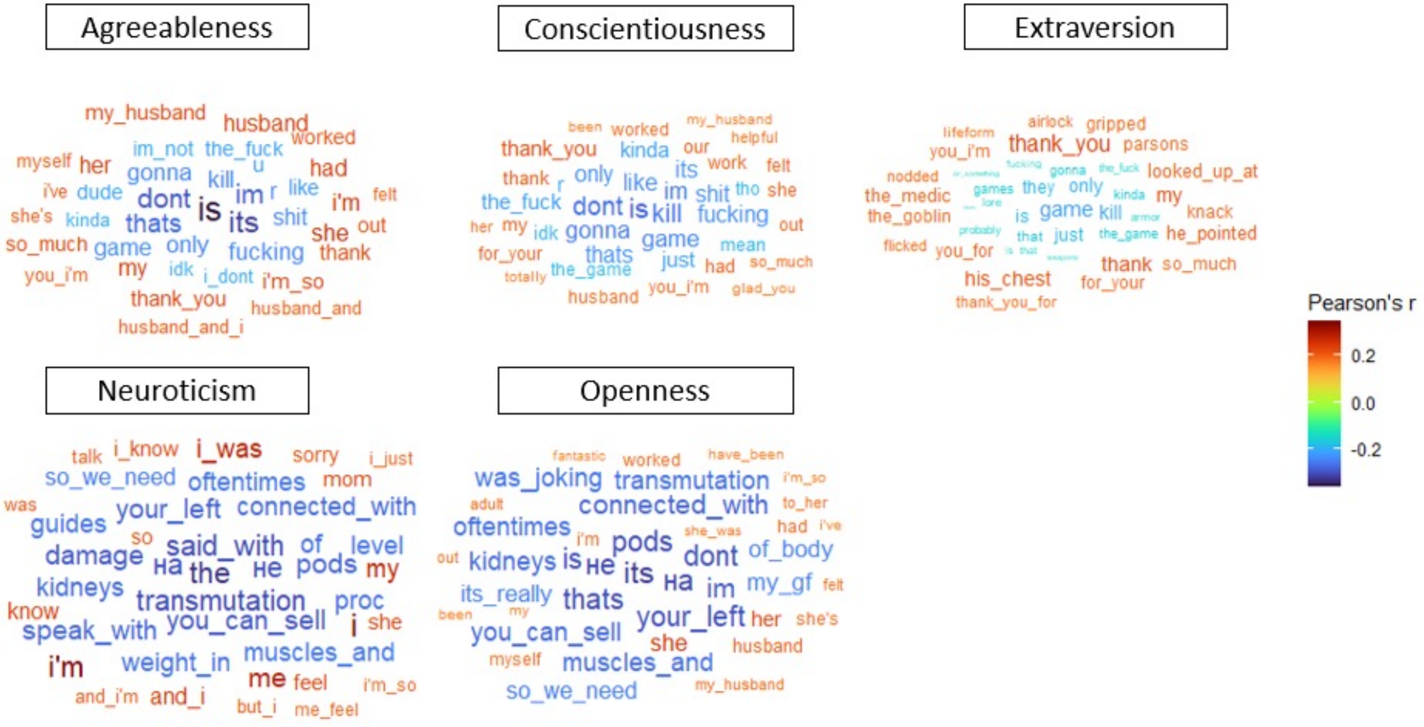
Personality Dimension	Pearson's *r*
Extraversion	0.22
Agreeableness	0.23
Conscientiousness	0.25
Neuroticism	0.27
Openness to Experience	0.23

Model 1: Results for Fiction

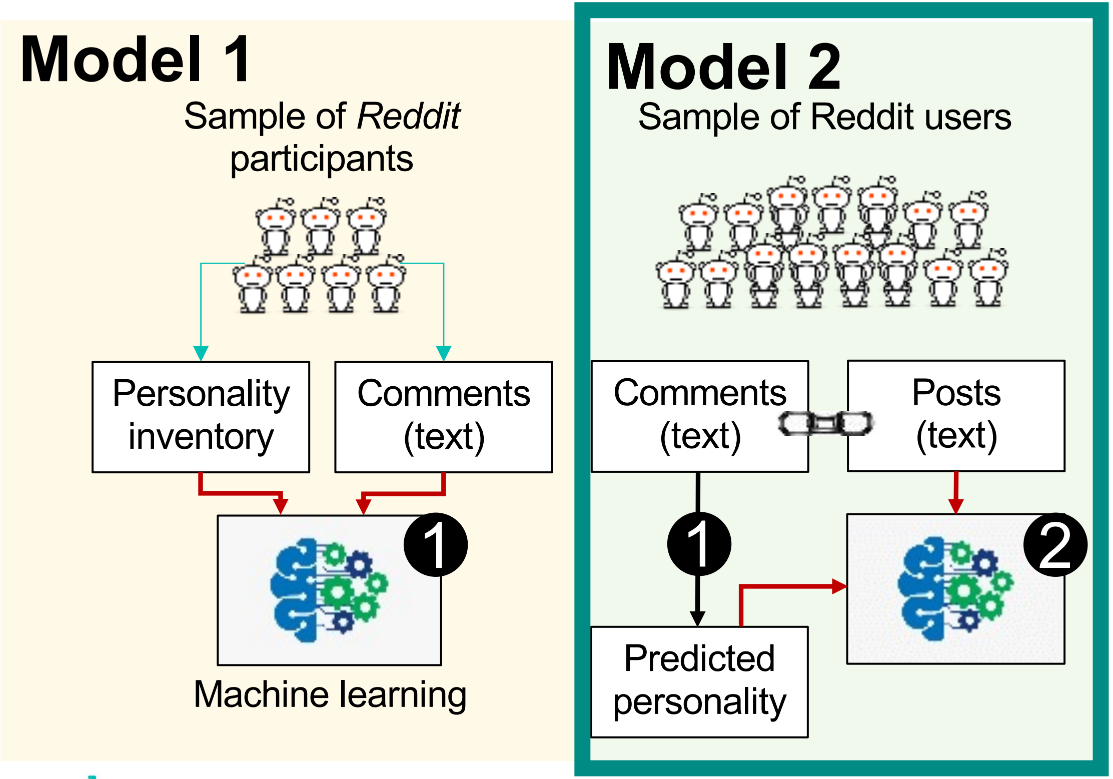
- Training set: 80% Fiction
- Test set: 20% Fiction Set (N = 215)
  - Average performance:
  - Pearson's *r* = 0.33
  - Correlation between the model predictions and ground truth
- Performance within meta-analytic estimates (SOTA; Eichstaedt et al., 2021; *Psych Methods*)

Personality Dimension	Pearson’s *r* [95% CI]
Extraversion	0.26 [0.13, 0.38]
Agreeableness	0.35 [0.23, 0.46]
Conscientiousness	0.37 [0.25, 0.28]
Neuroticism	0.28 [0.15, 0.40]
Openness to Experience	0.39 [0.28, 0.50]

What linguistic features best predict personality?



Model 2 – Currently Fiction Only



Model 1

Sample of *Reddit* participants

Personality inventory

Comments (text)

1

Machine learning

Model 2

Sample of Reddit users

Comments (text)

Posts (text)

1

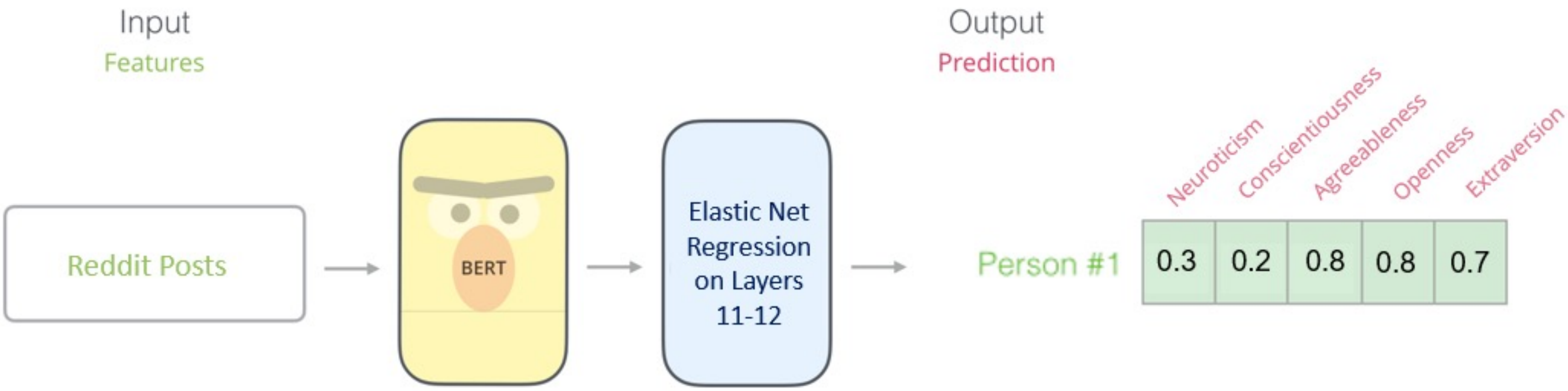
2

Predicted personality

Model 2: Method for Fiction

- Reddit collection
  - N = 4,466 participants
  - 1,756,819 comments – Used for Model 1 predictions.
  - 4,466 unique pieces of fiction – Used for Model 2 training.

Model 2: Consumed Fiction Predictions



Model 2: Results

- 5-fold Cross Validation

  Average Performance:

  Pearson's *r* = 0.106

- Ground Truth
  – N = 689

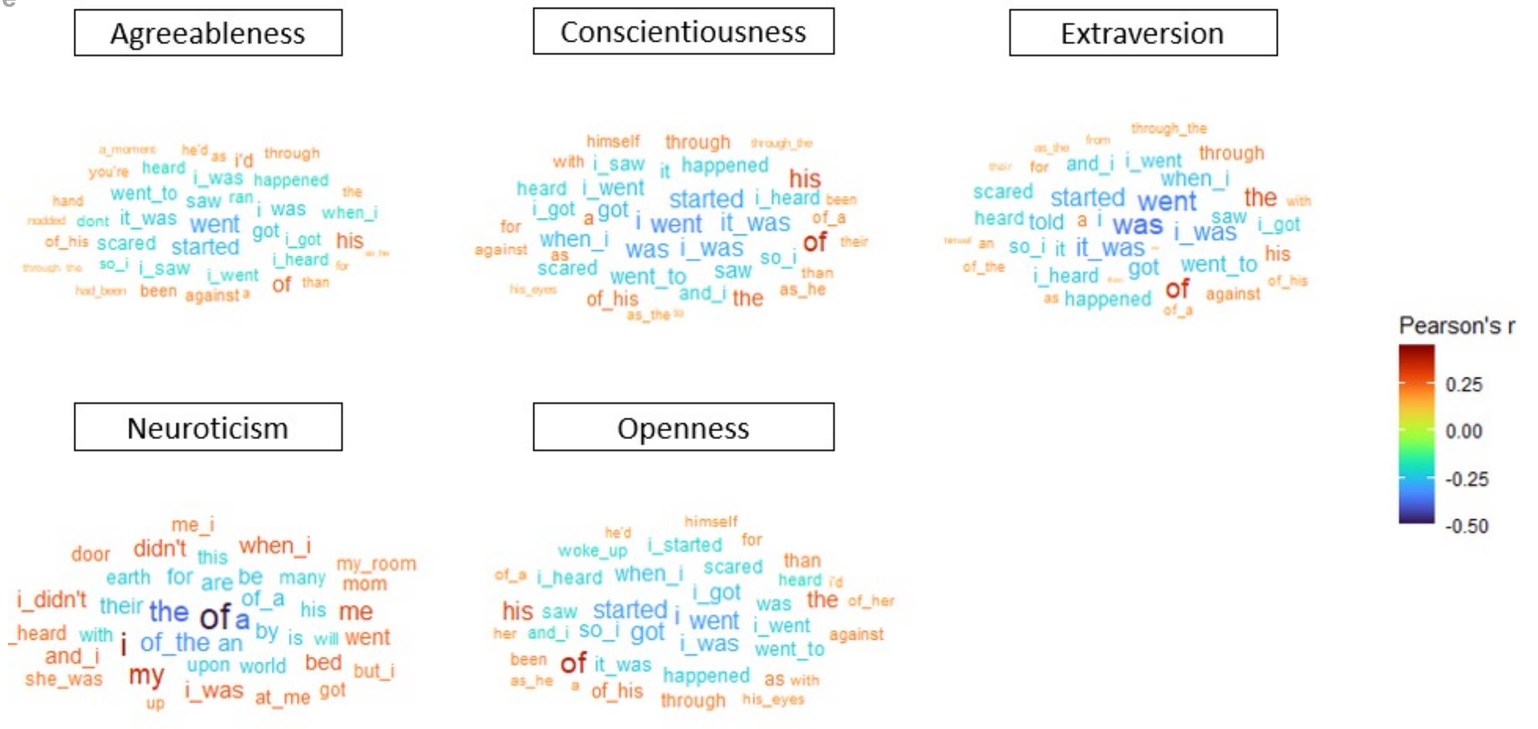
Personality Dimension	Pearson's *r*
Extraversion	0.07
Agreeableness	0.11
Conscientiousness	0.11
Neuroticism	0.13
Openness to Experience	0.11

Personality Dimension	Pearson's *r* [95% CI]
Extraversion	0.06 [-0.02, 0.13]
Agreeableness	0.12 [0.05, 0.20]
Conscientiousness	0.09 [0.02, 0.16]
Neuroticism	0.08 [0.01, 0.16]
Openness to Experience	0.08 [0.01, 0.15]

Size Matters

- Isn't an effect size of *r* = 0.1 negligible?
  - Yes, for a particular event, not at the aggregate or at scale
- Real world examples:
  - The effect of antihistamines on runny nose and sneezing: *r* = 0.11
  - The effect of ibuprofen on pain relief: *r* = .14
  - The correlation between extraversion and spent on holiday shopping: *r* = .09

What linguistic features best predict personality?



REPHRAIN
Protecting citizens online

What can we say about Model 2?

- Content vs. Style
  - Seems to be more sensitive to “how” in contrast to “what”
  - Which can be consistent with “linguistic markers”
- The linguistic features seem to overlap
  - Evidence for the General Factor of Personality? (Musek, 2007; *J Res Pers*)
    - Five personality dimensions can be reduced to two dimensions (Neuroticism and Agr/Con/Ext/Opn)

The current project

- Population: Reddit users of fiction-writing and political news communities
- Text-based models:
  - Model 1: predict stable psychological characteristics based on the text people produce
  - Model 2: predict stable psychological characteristics based on the text people consume <- We are here
- Find if indeed psychologically concordant messages are more persuasive
- Apply in the real world

Next steps and challenges

- Behavioral studies based on the applicability of Model 2
  - Are personality-congruent political ads rated as more persuasive? (different sample; experimental setting)
- Real-world application
  - How can we harness the science of boosting in developing interventions “in the wild”?

National Research Centre on
Privacy, Harm Reduction
and Adversarial Influence
Online

REPHRAIN
Protecting citizens online

To learn more about REPHRAIN, our future plans and how to get involved:

www.rephrain.ac.uk

@REPHRAIN1

rephrain-centre@bristol.ac.uk

We would love to hear from you. Thank you!

TeDCog group:
sks.to/tedcog

bristol.ac.uk

TeDCog