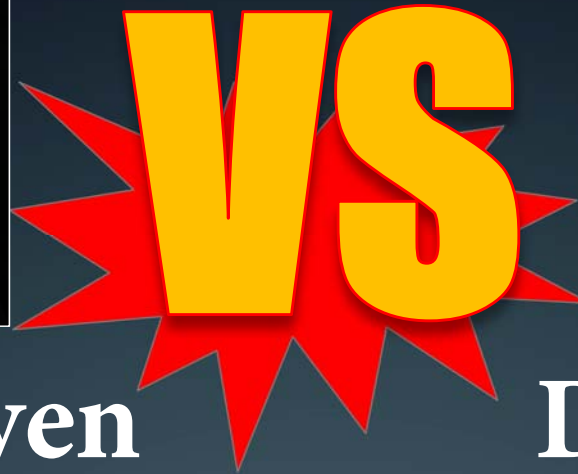


Combining Data- and Theory-Driven Approaches Using Large, Anonymous Datasets of Behavior

Joseph Chancellor, Ph.D.
Post-Doctoral Research Assistant
Department of Psychology



University of Cambridge



VS

Theory-Driven

- Favored in academia.
- Small, but structured data.
- Only what researcher chose to measure.
- Slow and methodical.
- Low predictive accuracy.
- Extremely narrow focus.
- Highly interpretable.

Data-Driven

- Favored in industry.
- Unlimited unstructured data.
- Everything (*except what you actually want*)
- Fast.
- Highest possible accuracy.
- Broader scope.
- Limited interpretability.

facebook

NETFLIX

amazon.com[®]

twitter 

Google[™]

TESCO



Research Interest:
Narcissism
(and Personality)

Data-driven paper: Facebook likes predict county-level crime
(N = 1 million)

Hybrid: Liking narcissistic celebrities on Facebook predicts friend # on Facebook
(N = 1 million)

Theory-driven paper: Narcissism predicts friend # on Facebook
(N = 300)

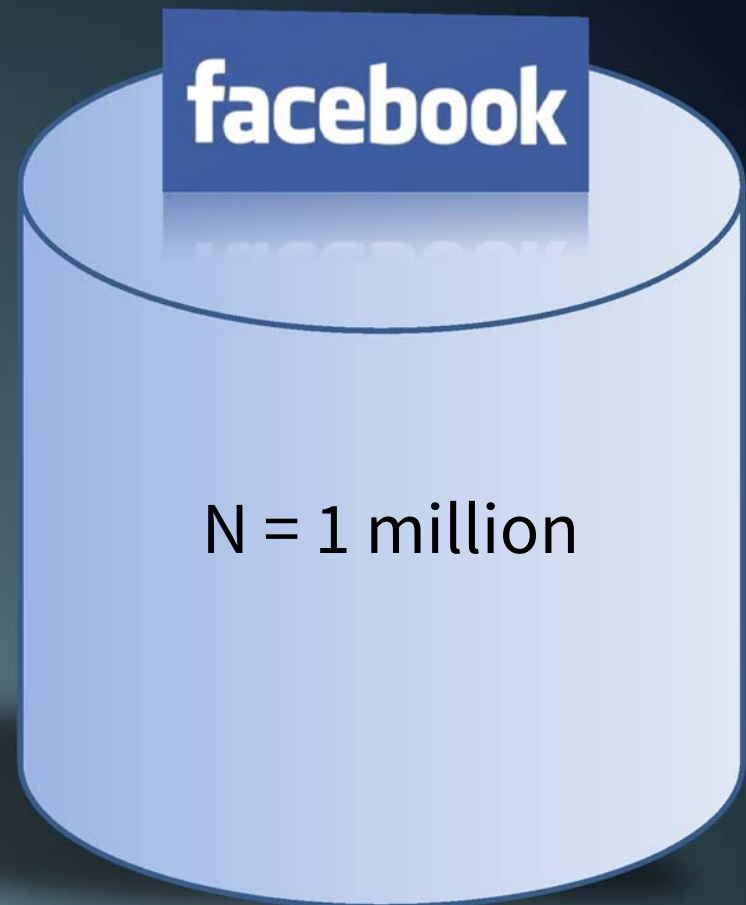


Survey Battery

Narcissism, Personality

N = 500-5,000

Linked to Big Data



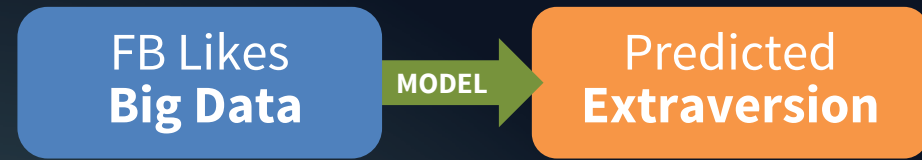
facebook

N = 1 million

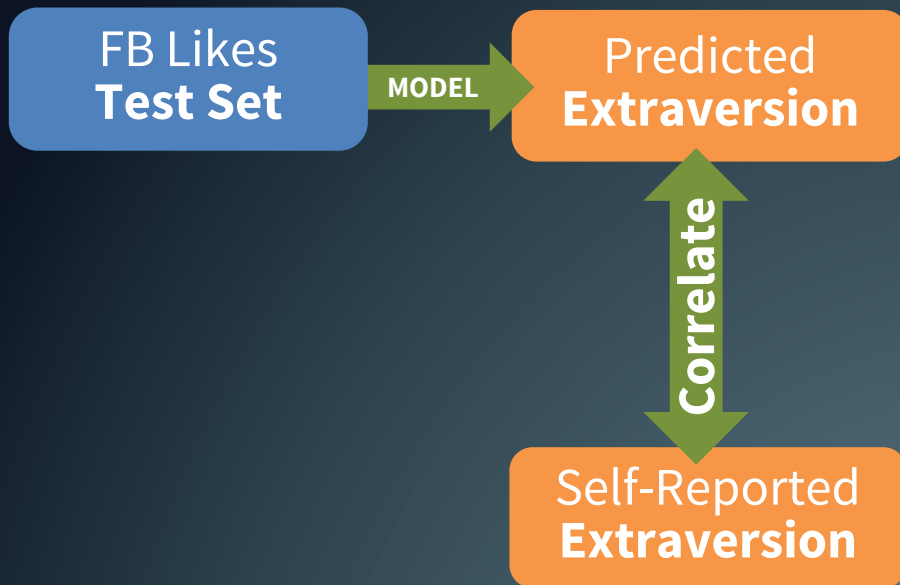
Train Model



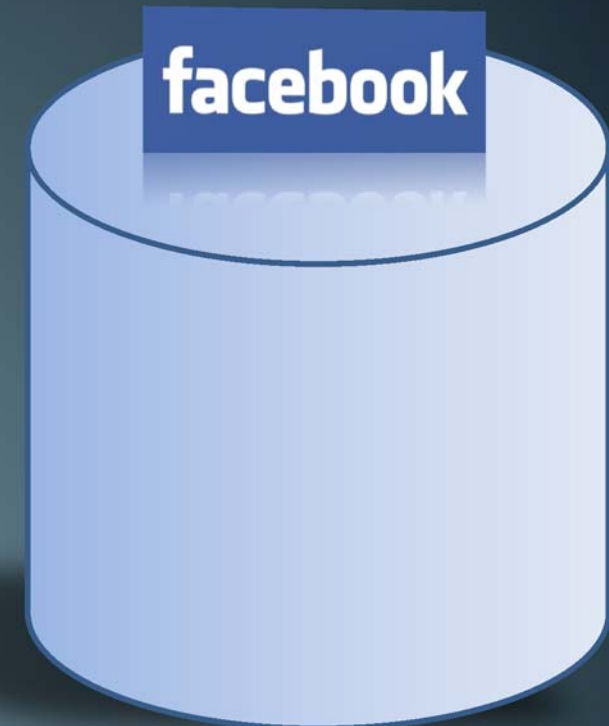
Mass Prediction



Cross-Validate Model



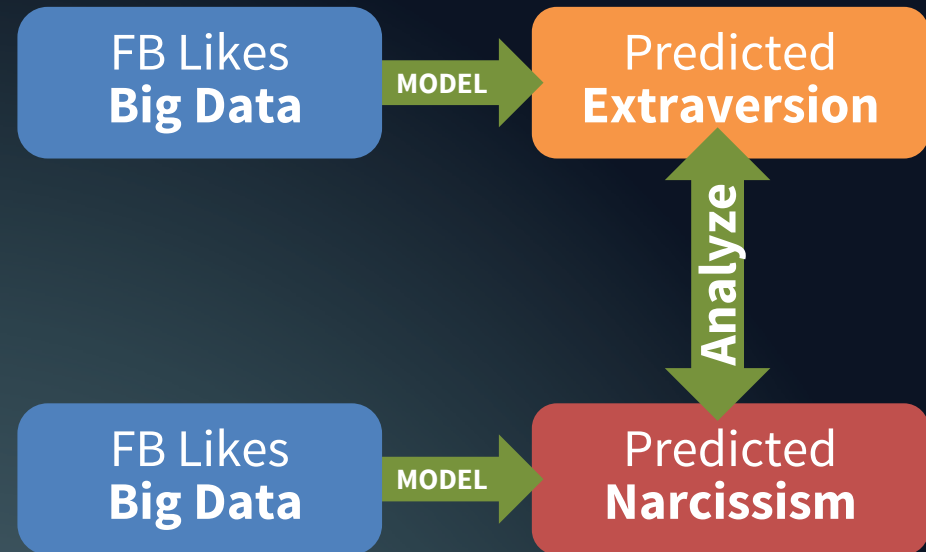
Seeder Sample 1



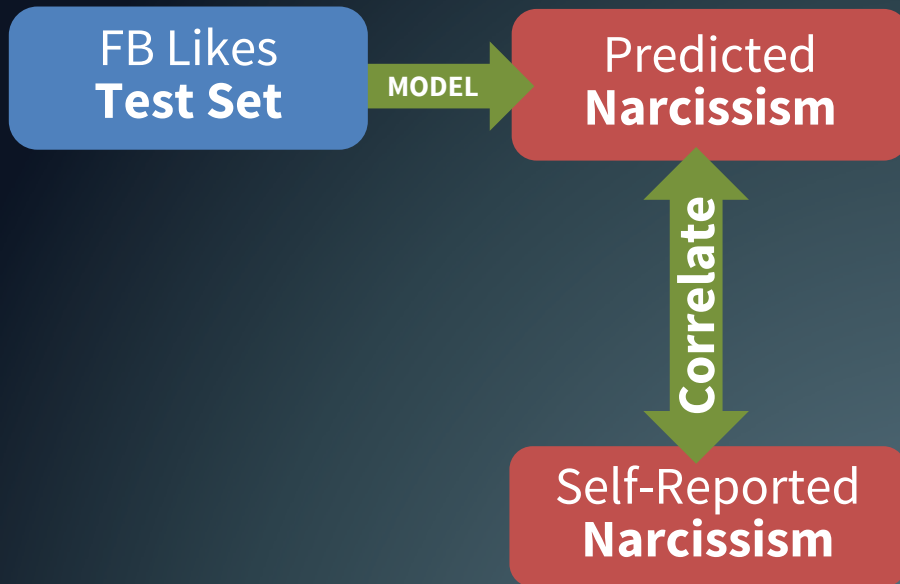
Train Model



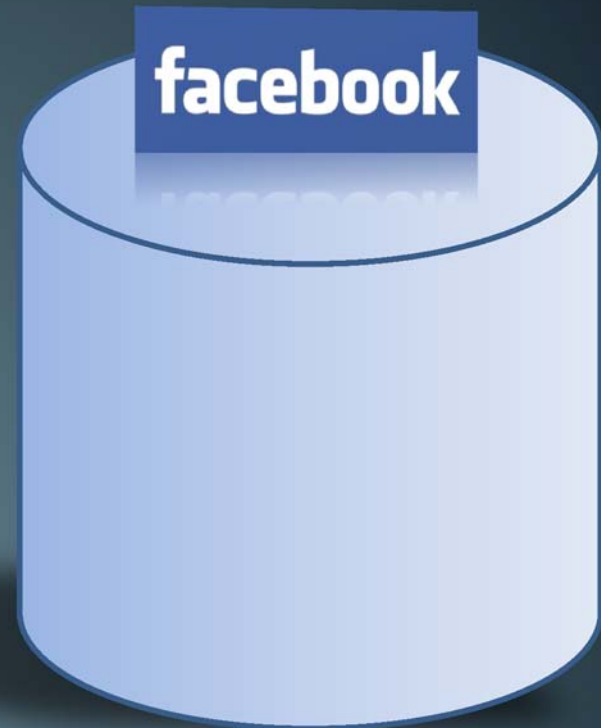
Mass Prediction



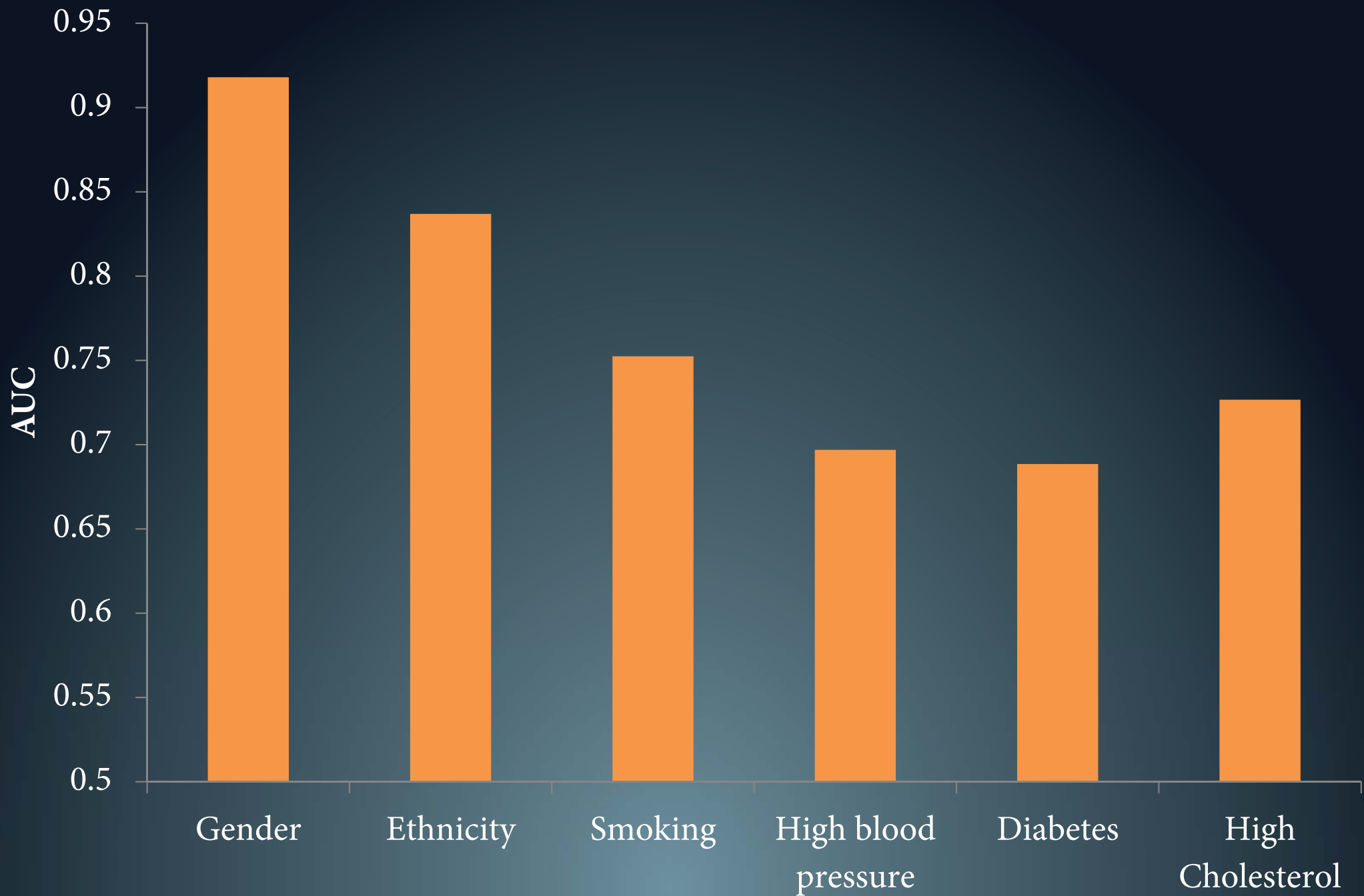
Cross-Validate Model



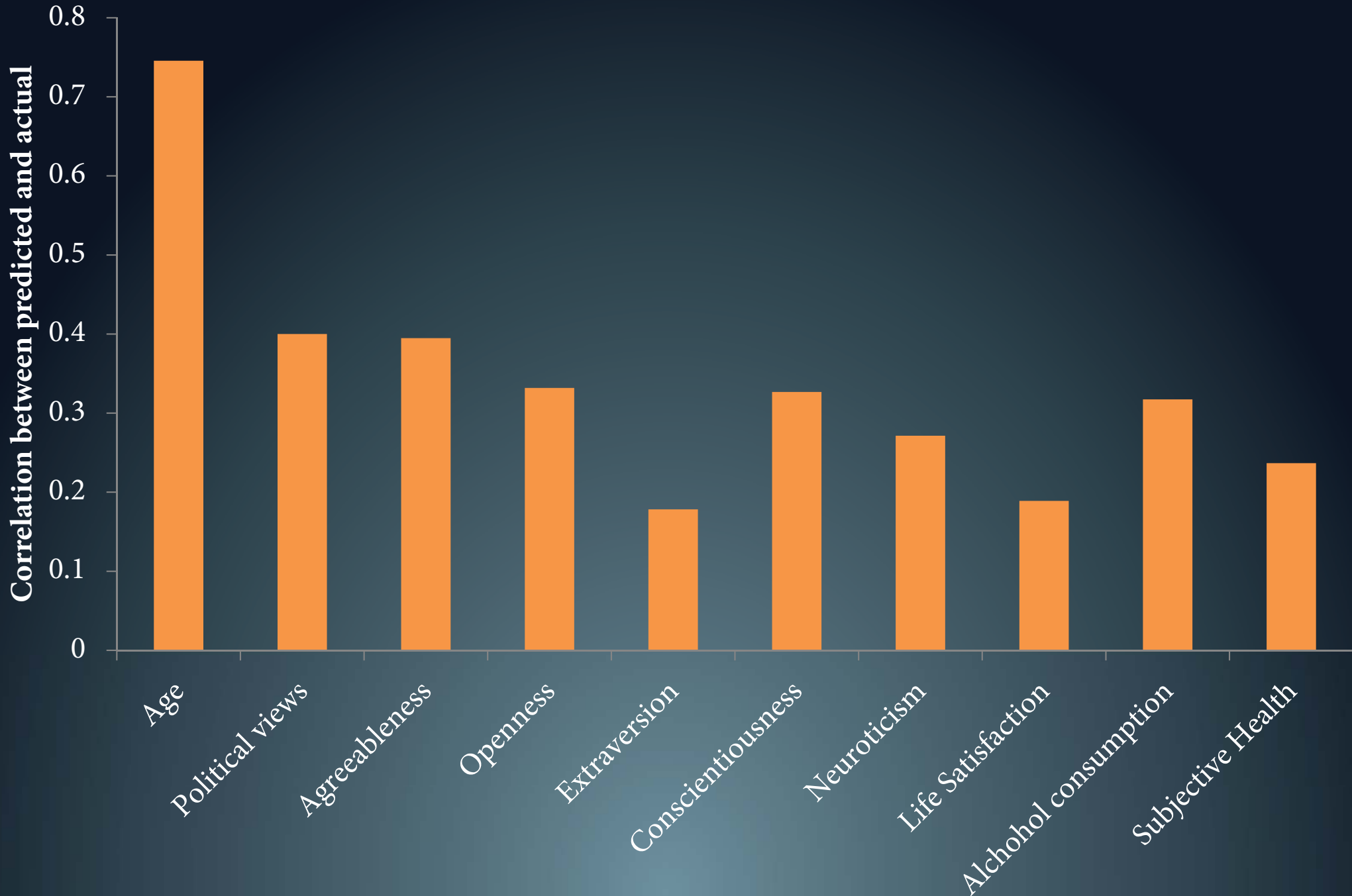
Seeder Sample 2



Model Accuracy: Dichotomous



Model Accuracy: Continuous



Similarity in insights gained from actual and predicted scores

Actual

	<i>h</i>	<i>e</i>	<i>n</i>	<i>c</i>	<i>o</i>
<i>h</i>					
<i>e</i>	.41				
<i>n</i>	-.33	-.46			
<i>c</i>	.21	.37	-.53		
<i>o</i>	.01	.32	-.09	.23	
<i>a</i>	.01	.10	-.31	.60	.33

Predicted

	<i>h</i>	<i>e</i>	<i>n</i>	<i>c</i>	<i>o</i>
<i>h</i>					
<i>e</i>	.39				
<i>n</i>	-.21	-.40			
<i>c</i>	.29	.18	-.65		
<i>o</i>	-.30	-.10	.35	-.30	
<i>a</i>	.09	-.20	-.33	.78	-.08

Challenges

Learning curve.

Sparsity.

Size of N for surveys?

Not all variables model well.

Convincing people it's real.

Real Example

Nested Social Contexts Moderate Link between Agreeableness and Well-Being

Self-Reported Data
(N = 35,695)

State levels of agreeableness moderate agreeableness-LS link, $\beta=.021$.

Computer-Predicted Data
(N = 3,982,299)

State levels of agreeableness moderate agreeableness-LS link, $\beta=.078$.

City levels of agreeableness moderate agreeableness-LS link, $\beta=.101$.

Social network levels of agreeableness moderate agreeableness-LS link, $\beta=.185/.263$.