

An Introduction to Web Science

RuSSIR, Aug 6-10, 2012

Please interrupt at any point!!

Ingmar Weber

ingmar@yahoo-inc.com

Yahoo! Research Barcelona

Course Outline

- Day 1: Introduction to the Introduction
 - Examples, data sets, presentation of the competition
- Day 2: Web Search and Society
 - Demographics, economy and more
- Day 3: Blogs and Twitter
 - Gender, moods, politics, stock market and more
- **Day 4: Social Networks and Online Dating**
 - **Attractiveness, FB&GPA, FB&Personality and more**
- Day 5: E-commerce and Marketing Studies
 - Brand congruence, Groupon Effect, social ads

What is beautiful is good, even online:
Correlations between photo
attractiveness and text attractiveness
in men's online dating profiles

Rebecca Brand, Abigail Bonatsos, Rebecca
D'Orazio and Hilary DeShong

Computers in Human Behavior 2012

Does photo attractiveness ...



... correlate with text attractiveness

I just finished my studies, in computer science + science education, and I plan to be a teacher for the next school year. ...

I am finished with college, i went to college and majored in criminal justice. ...

I create something new almost every single day, I design, animate or try some kinda crafting. ...

Data

- 50 female psychology students
 - 18-24 years, average 19.0
 - 74% Caucasian, 10% Hispanic
 - Course credit or entry into \$25 drawing
- 100 male profiles and photos from dating site
 - 22-25 years, NYC and Seattle
 - Both recently active and inactive
 - Photos and “About me” were separated

Getting Judgments

- Photos and texts judged by separate people
- For photo: how attractive, from 0 to 4, (i) overall, (ii) for a date, (iii) for sex, and for (iv) for a long-term relationship; also how kind, how confident, how masculine and how symmetrical
- For text: pretty much the same, also how intelligent and how funny
- Each set of 25 profiles rated by 12-13 participants

Evaluating the Results

- Are the four attractiveness questions related?
 - Cronbach's alpha for reliability
 - Summed score for each (photo,judge) X_j
 - Compute variance of the X_j : σ_X^2
 - Compute variance for each question Y_i : $\sigma_{Y_i}^2$
 - $\alpha = \frac{K}{K-1} \left(1 - \frac{\sum_{i=1}^K \sigma_{Y_i}^2}{\sigma_X^2} \right)$, here: alpha = .99
- Related to exam design
 - Different questions should measure “the same thing”
 - The overall exam should tell students apart

Related: Do Different Judges Agree?

Not used by them – but fits in this context

Suppose project proposals are evaluated by A and B

Agreement:

$$(20+15)/50 = 0.70$$

		B	
		Yes	No
A	Yes	20	5
	No	10	15

But how much agreement by chance?

$$\Pr(\text{"yes"} | A) = 50\%, \Pr(\text{"yes"} | B) = 60\%$$

$$\Pr(\text{"both yes"}) + \Pr(\text{"both no"}) = .5 * .6 + .5 * .4 = .5$$

Cohen's kappa

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} = \frac{0.70 - 0.50}{1 - 0.50} = 0.40$$

Back to Study: Correlation Analysis

- Both photo and text attractiveness combined
 - That's what the Cronbach alpha was used for
- Photo attractiveness and text attractiveness
 - $Rho=.24$, $p=.017$
- Photo attractiveness and text confidence
 - $Rho=.25$, $p=.012$
- Photo masculinity and text confidence
 - $Rho=.21$, $p=0.037$

Going Beyond Correlation

- Various things are correlated
- Goal: Isolate a “mediator”

Text confidence

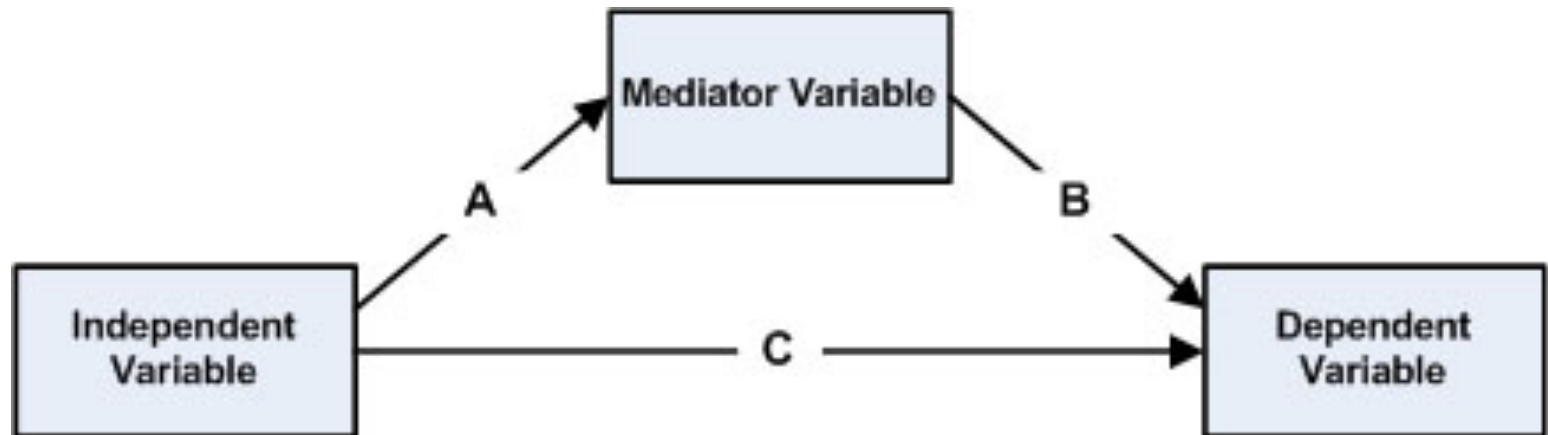


Photo attractiveness

Text attractiveness

Is (Textual) Confidence a Mediator

- First: the independent variable (here, photo attractiveness) must predict the dependent variable (text attractiveness)
- To test this
 - build a linear regression model
$$(\text{text attr.}) = c + \beta_1^* (\text{photo attr.}) + \epsilon$$
 - Then test the hypothesis that $\beta_1 = 0$
Compare errors w/ and w/o β_1

Is (Textual) Confidence a Mediator

- Second: the independent variable must predict the proposed mediator (text confidence)
- Same techniques as before to test this

Is (Textual) Confidence a Mediator

- Third, the mediator must predict the dependent variable when controlling for the independent variable

$$(\text{te. at.}) = c + \beta_1 * (\text{ph. at.}) + \beta_2 * (\text{te. conf.}) + \epsilon$$

β_2 still significant: textual confidence adds inf.

β_1 no longer significant: ph. attr. is redundant

Search for “Sobel Test” for details

Too much face and not enough books:
The relationship between multiple
indices of Facebook use and
academic performance

Reynol Junco

Journal Computers in Human Behavior 2012

Basic Question

- Do your grades suffer if you spend lots of time on Facebook?
- Probably yes, but what if high school GPA is factored in?
- What if other factors are included?
- Does it depend on the type of FB activity?

Getting the Data

- Sent email to 3,866 undergraduate students
- The email contained a link to a survey
 - <http://www.surveymonkey.com/>
- Chance to win 90 \$10 Amazon vouchers
- 1,839 students participated
- Self-reported data
- 61 “outliers” removed, e.g. >10hrs of FB/day

Questions Asked

- How much time do/did you spend on FB
 - Both “on average” and yesterday (FBtime)
- How often do/did you check FB
 - Both “on average” and yesterday (FBcheck)
- Also questions related to type of FB activity
 - Games, status updates, posting links, ...
- Also got their current and past marks
 - High school GPA and college GPA

Basic Results

- Av. 106 min/day (SD 93 min/day) on FB
- No correlation between HSGPA and Fbtime
- Weak corr. between time studying and GPA
 - $r=.22$, $p<.001$
- Weak corr. between time studying and Fbtime
 - $r=-.09$, $p<.001$

Hierarchical regression model exploring how demographics, high school GPA, average minutes/day spent on Facebook, and Facebook activities predict overall GPA ($N = 1771$).

Independent variables	Block 1 demographics β		
Male	-.125***	-	-
African American	-.107*		
Asian American	.013		
Other ethnicity	-.022		
Caucasian	.035		
Less than high school	.009		
High school	.017		
College graduate	.040		
Advanced grad degree	.058*	-	-
		-	-

Interpreting .189

- Increasing FBtime by one SD lowers your GPA by .189 standard deviation

FBtime SD: 93 minutes (av. 106 minutes)

GPA SD: .65 (av. 2.95)

Spending 4 hours/day = 240 min = av + 1.4*SD

Lowers your GPA by .17

The More People I Meet, The More I Like My Dog: A Study of Pet-Oriented Social Networks on the Web

Jennifer Golbeck

First Monday 2011



CONFESSIONS

Ever Been Attacked
By a Dog? I Was
Once, and I'll Never
Forget It

For the love of dog.

dogster

Search!



↑ MAGAZINE

VIDEO

BOOK OF DOG

ANSWERS

GALLERIES

ADOPTION

COMMUNITY

Penny

Labrador Retriever/Basset Hound



Photo Comments

Home: Raleigh, NC

Age: 3 Months Sex: Female Weight: 11-25 lbs

Add This Pup as a Friend



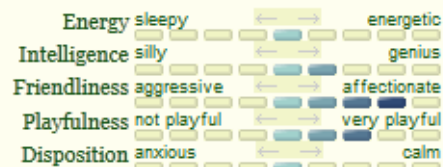
Dogster stats for Penny

Pals: 5 Views: 25

Stars: ★

Leave a bone for Penny

Doggie Dynamics:



Sun Sign:



ARIES

Quick Bio:

-mutt -dog re

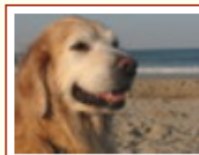
Gotcha Date:
June 3rd 2012

Birthday:
March 30th 2012

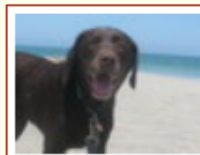
Favorite Toy:
Pink squeaky puppy

Best Tricks:
Sit

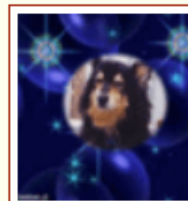
Meet my Pup Pals



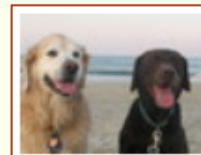
Captain Squirt
of the High
Sea



Captain Chloe
Salty Sea
Dog



Taffy(in
loving
memory



Girls of the
Sea



FREEBIES

Win a Sterling Silver
Kitty Cat Ears Ring
from Welded Heart

Here, Kitty Kitty.

catster

↑ MAGAZINE

VIDEO

BOOK OF CAT

ANSWERS

GALLERIES

ADOPTION

Griswald "Grizzly"

Breed Unknown



Photo Comments

Age: 1 Year Sex: Male

[Send a Feline Friend Request!](#)



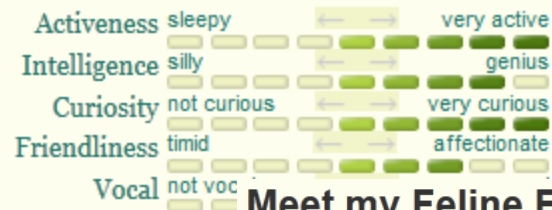
Catster stats for Griswald "Grizzly"

Friends: 20 Views: 63

Stars: ★

[Leave a treat for Griswald "Grizzly"](#)

Kitty Complexion:



Meet my Feline Friends

[See all my Feline Friends](#)

Sun Sign:



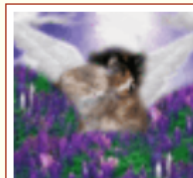
AQUARIUS

Quick Bio:

-cat rescue

Birthday:

February 15th 2011



Angel Pie Joe
in Loving
Memory



Opie
T-bear--In
Loving
Memory



Kittyman



Violet
Loving
Mem

Research Questions

- Are there observable differences in the way people connect in passion-oriented vs. friend-oriented networks?
 - In particular, does the semi-anonymous nature of passion-oriented sites affect the impact of social capital on user behavior?
- Do groups with similar passions utilize passion-oriented networking websites in similar or significantly different ways?
 - In particular, how do dog and cat owners utilize their sites?

A Bit of Context

Table 1. Results of the University of Maryland study on cardiac patient survival rates.

	Pets	No Pets	Total
Living	50	28	78
Dead	3	11	14
Total	53	39	92

Survival rates one year after heart attack. (1977-79)

Other studies have reconfirmed health benefits of owning a pet.

Oddities of Pet-Networks

- Connections mostly virtual
 - Between strangers with same “passion”
- The owner is hidden
 - All actions performed by the pets
- Break-down by population density
 - Dogs are more frequent in rural households


Urban vs. Rural

- Gilbert, Karahalios, and Sandvig, 2008: “The network in the garden: an empirical analysis of social media in rural life” (using Myspace):
- Urban users join earlier, have more friends and the friends live farther away than for rural users
- Usually explained using “social capital”: “social networks have value”
- What happens when user’s identity is hidden?

Sampling from a Network

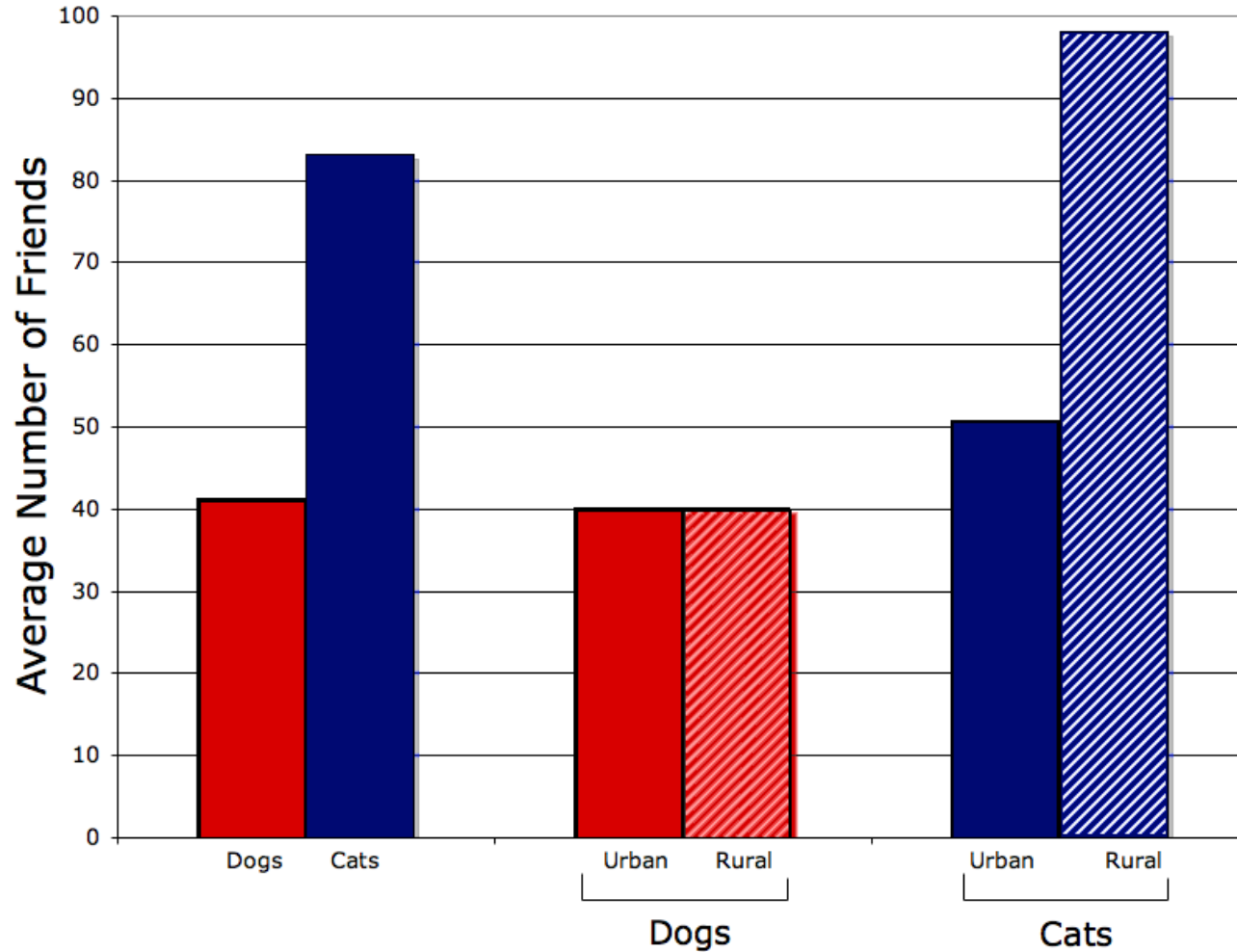
- How to sample uniformly at random
 - In this case easy:
<http://www.dogster.com/dogs/1253721>
 - Consecutive user ID at the end
- To sample urban vs. rural
 - Same Methodology as other study
 - Classify ZIP codes into range from urban to rural
 - Select 2,000 locations from either extreme end
 - Try to find a profile for that location

Comparing non-Normal Distributions

- t-tests are standard to compare means
 - Assumes a normal distribution  power law
- Non-parametric tests exist: Mann-Whitney U
 - Count how often an X comes before a Y
 - THHHHTTTT
 - For T: $U_T = 0 + 4 + 4 + 5 + 5 = 18$
 - For H: $U_H = 1 + 1 + 1 + 1 + 3 = 7$
 - $U = \min(U_T, U_H)$ --- but: $U_T + U_H = n_T * n_U$

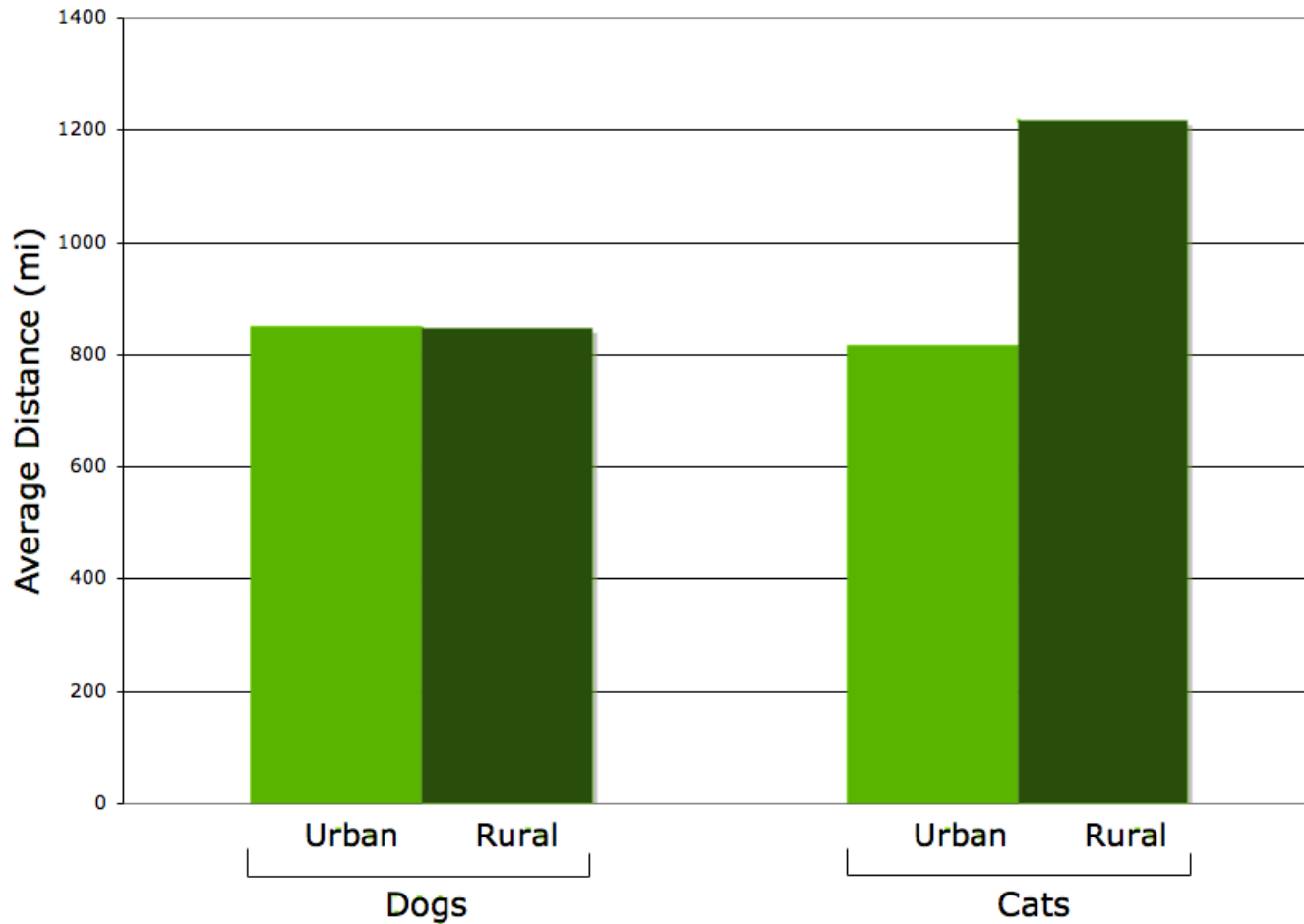
$$z = \frac{U - m_U}{\sigma_U}, \quad m_U = \frac{n_1 n_2}{2}, \quad \sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}.$$

Findings



Less pet-related interactions in the real world appear to be “compensated”.

Findings



A factor ~10 larger than for MySpace study

Findings

	Rural	Urban	z	p-value
Cats				
N	1740	2000		
Member ID	342090	195192	-25.2	< .001
Number of Friends	98.1	50.6	2.6	< .01
Distance between Friends	815.6	1218.0		< .001
Dogs				
N	1669	2000		
Member ID	290903	204788	-15.5	< .001
Number of Friends	39.9	39.9	-0.1	> .920
Distance between Friends	849.0	845.6		> .5

Urban area first, despite lacking “social capital”

Inferring social ties from geographic coincidences

David Crandall, Lars Backstrom, Dan Cosley,
Siddharth Suri, Daniel Huttenlocher and Jon
Kleinberg

PNAS 2010

Suppose that ...

- ... we frequently go to the same pubs
- ... we're going to the same university
- ... we're going to the same super market
- Does that mean we know each other?

Research Questions

- Given that two people have been in approximately the same geographic locale at approximately the same time, on multiple occasions, how likely are they to know each other?
- Furthermore, how does this likelihood depend on the spatial and temporal proximity of the co-occurrences?

Methodology

- Geo-tagged photos on Flickr
- Contact list on Flickr

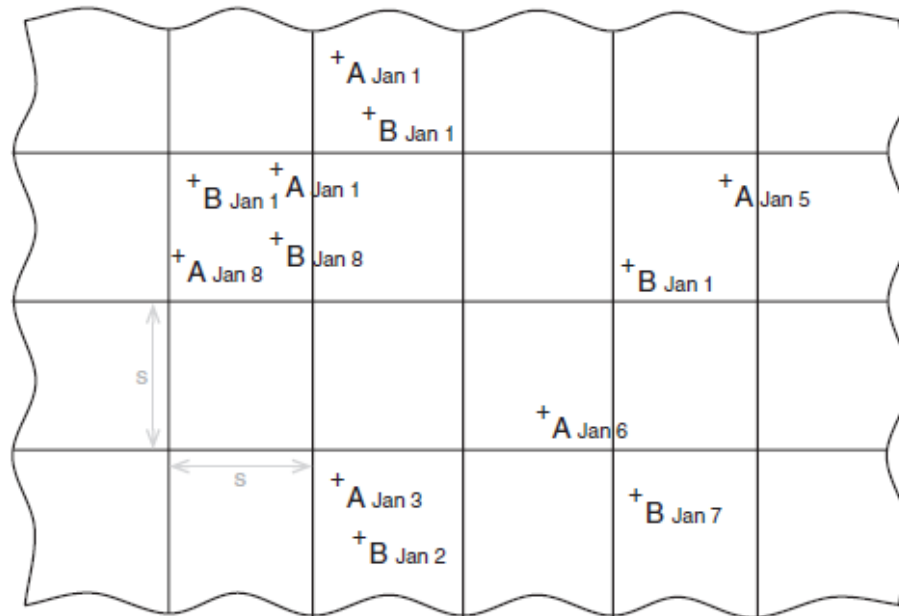
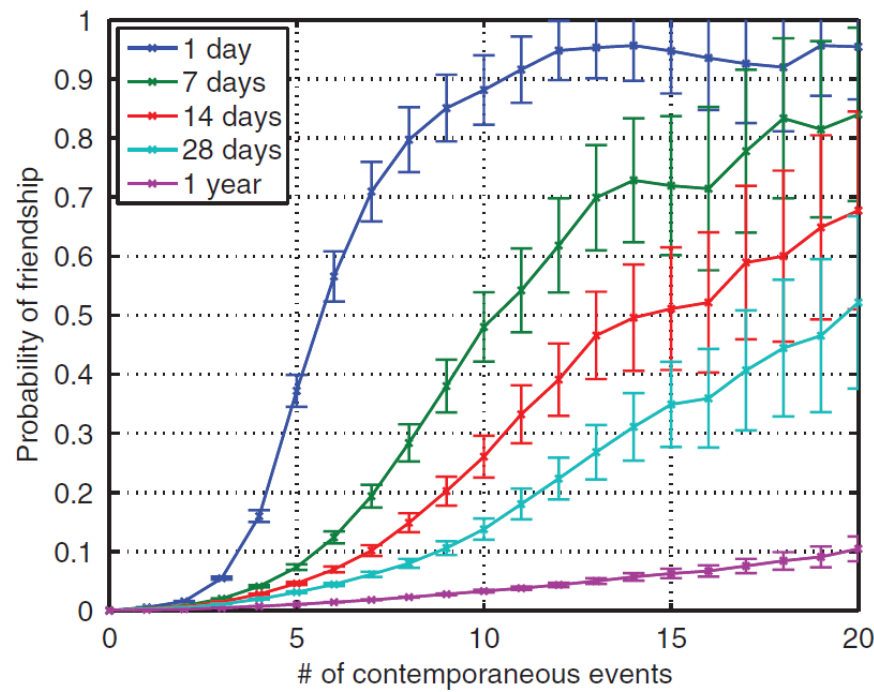


Fig. 1. Illustration of how spatio-temporal co-occurrences are counted, for some sample time-stamped observations of individuals *A* and *B*. The world is divided into discrete cells of size $s \times s$, and we count the number of cells k in which the two individuals have been observed within a time threshold of t days—in this case, $k = 3$ when t is 2.

Details

- Social graph snapshot in April 2008
- Use spatio-temporal occurrences appearing *after* this date
- 38 million geo-tagged photos



$s = 0.1^{\circ}$

A Simple Model

- N geogr. cells, M people, each with one contact
- M/2 pairs of friends, each day chooses a location
- Jointly with prob. β , independently with $1-\beta$
- Prob. being friend visiting same cell on k consecutive days:

$$P(F|C_k) = \frac{P(F)P(C_k|F)}{P(C_k)}$$

$$P(F) = 1/(M-1)$$

$$P(C_k|F) = p_1^k \quad \text{where} \quad p_1 = \beta + (1-\beta)/N$$

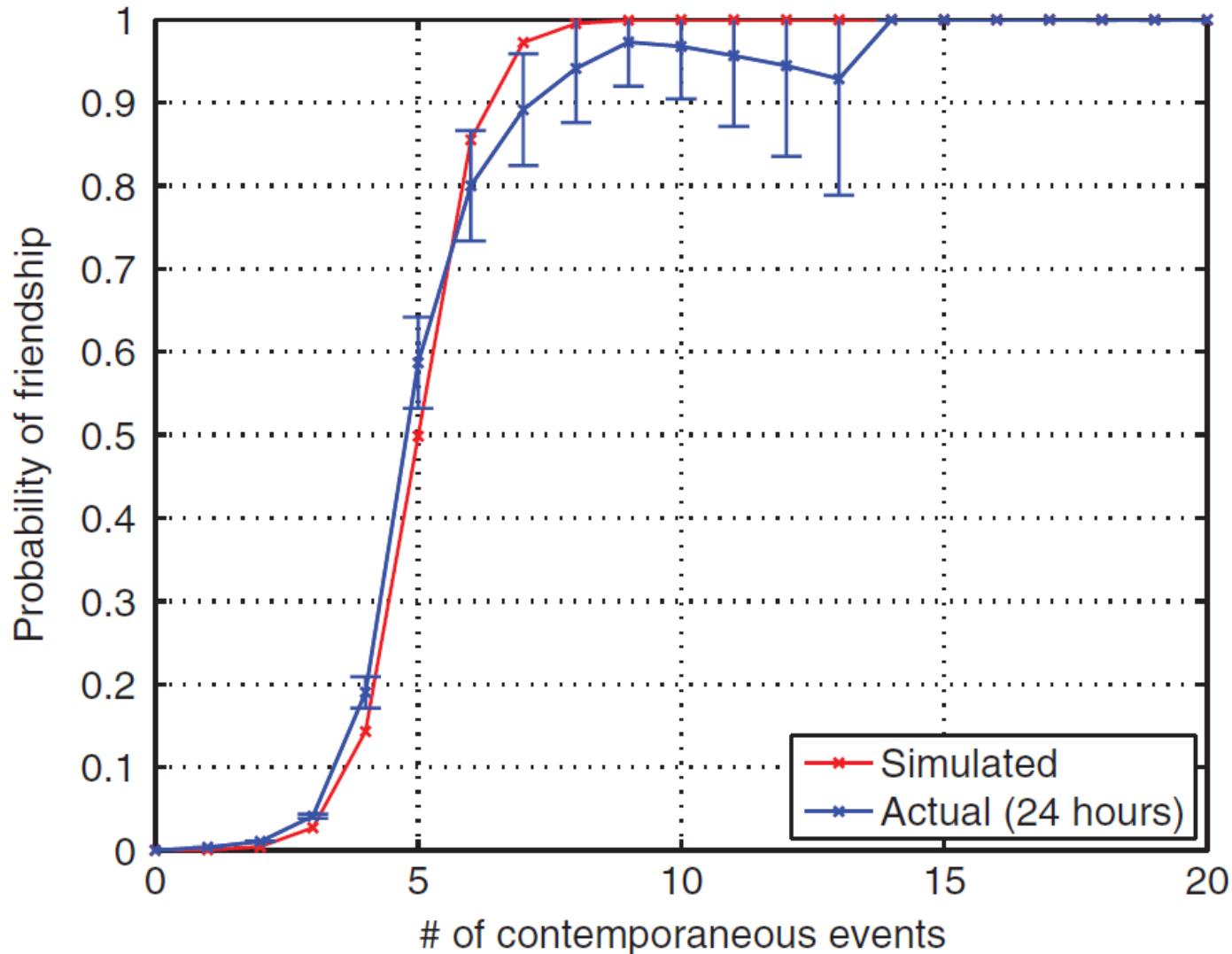
Putting It All Together

$$\begin{aligned}P(C_k) &= P(C_k|F)P(F) + P(C_k|\bar{F})P(\bar{F}) \\&= p_1^k \cdot \frac{1}{M-1} + p_2^k \cdot \frac{M-2}{M-1}\end{aligned}$$

$$p_1 = \beta + \frac{1-\beta}{N} \quad p_2 = \frac{1}{N}$$

$$P(F|C_k) = \frac{p_1^k}{p_1^k + p_2^k(M-2)}$$

A Pretty Good Fit



$$M = 7,500, N = 100, \beta = 0.05$$

$$t = 1, s = 1$$

Refining the Model

- Each user pair is assigned a home cell according to Flickr's empirical distribution
- On a given day each person goes out (and takes a photo) with probability α
- Go to same location with probability β
- Sample power law distribution centered “home” and with exponent γ

Model-Fitting and Comparing Distributions

$$M = 7,500, \quad N = 64,800, \quad \alpha = 0.29, \quad \beta = 0.12, \quad \gamma = 1.8$$

- Kolmogorov-Smirnov statistics
- Test for equality of two distributions
- Gives a distance measure

- Kolmogorov–Smirnov statistic

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{X_i \leq x} \qquad D_{n,n'} = \sup_x |F_{1,n}(x) - F_{2,n'}(x)|,$$

- Reject null hypothesis (same distribution) if:

$$\sqrt{\frac{nn'}{n+n'}} D_{n,n'} > K_\alpha. \quad \Pr(K \leq K_\alpha) = 1 - \alpha.$$

Personality and Patterns of Facebook Usage

Yoram Bachrach, Michal Kosinski, Thore
Graepel, Pushmeet Kohli and David Stillwell

WebSci'12

The Big Five Personality Model

- Openness
 - inventive/curious vs. consistent/cautious
- Conscientiousness
 - efficient/organized vs. easy-going/careless
- Extraversion
 - outgoing/energetic vs. solitary/reserved
- Agreeableness
 - friendly/compassionate vs. cold/unkind
- Neuroticism
 - sensitive/nervous vs. secure/confident

Openness to Experience

- inventive/curious vs. consistent/cautious

Appreciation for art, emotion, adventure, unusual ideas, curiosity, and variety of experience. Openness reflects the degree of intellectual curiosity, creativity and a preference for novelty and variety. Some disagreement remains about how to interpret the openness factor, which is sometimes called "intellect" rather than openness to experience.

Conscientiousness

- Conscientiousness – (efficient/organized vs. easy-going/careless)

A tendency to show self-discipline, act dutifully, and aim for achievement; planned rather than spontaneous behavior; organized, and dependable.

Extraversion

- outgoing/energetic vs. solitary/reserved

Energy, positive emotions, surgency, assertiveness, sociability and the tendency to seek stimulation in the company of others, and talkativeness.

Agreeableness

- friendly/compassionate vs. cold/unkind

A tendency to be compassionate and cooperative rather than suspicious and antagonistic towards others.

Neuroticism

- sensitive/nervous vs. secure/confident

The tendency to experience unpleasant emotions easily, such as anger, anxiety, depression, or vulnerability. Neuroticism also refers to the degree of emotional stability and impulse control, and is sometimes referred to by its low pole – "emotional stability".

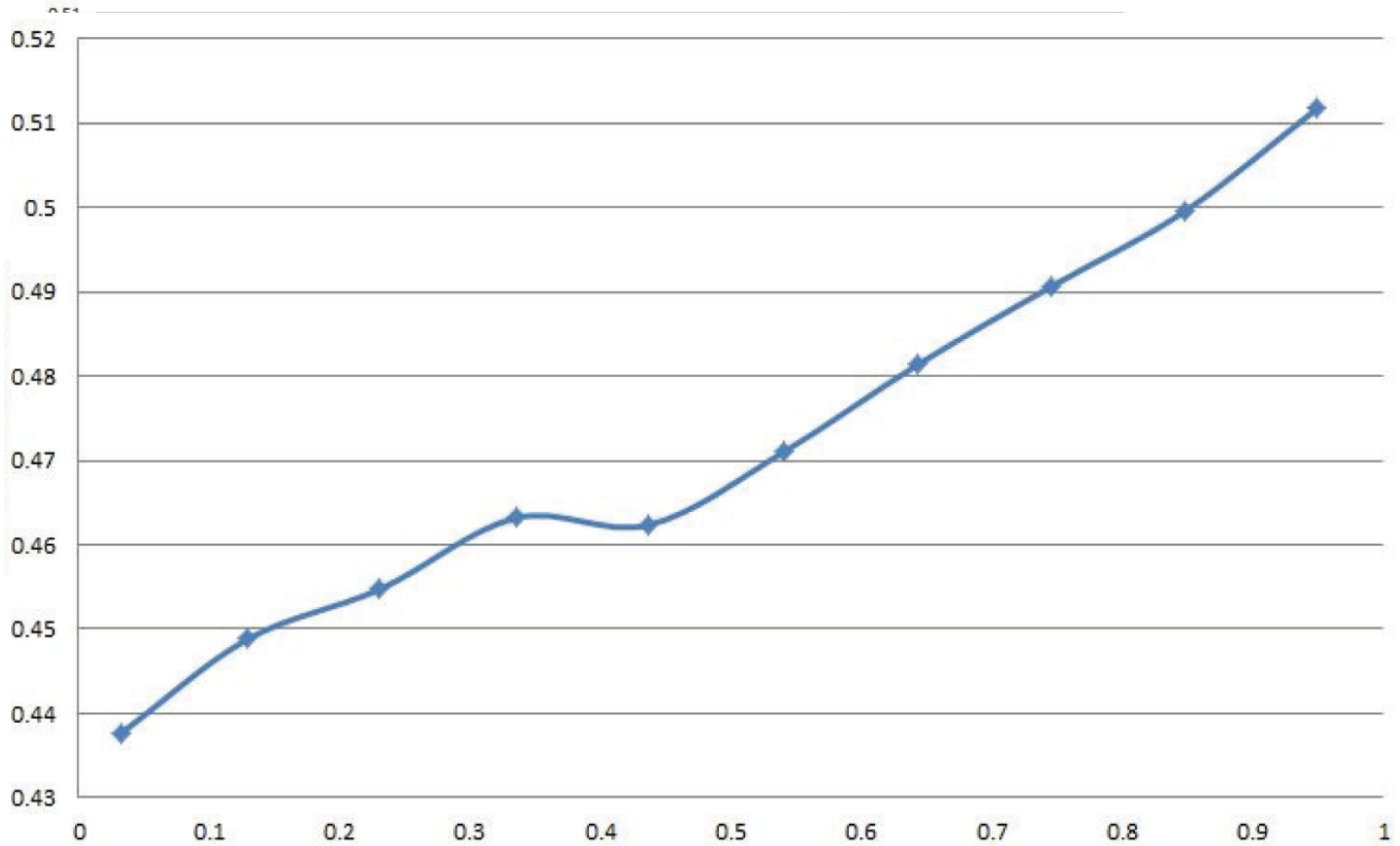
Data set

- 180,000 users profiles from “myPersonality”
 - <http://www.psychometrics.cam.ac.uk/page/255/mypersonality.htm>
- Not always full profile access
 - But at least 15,000 for each feature
- Average age 24 years, 58% female

Feature	Details
Friends	number of Facebook friends
Groups	number of associations with groups
Likes	number of Facebook “likes”
Photos	number of photos uploaded by user
Statuses	number of status updates by user
Tags	number of times others “tagged” user in photos

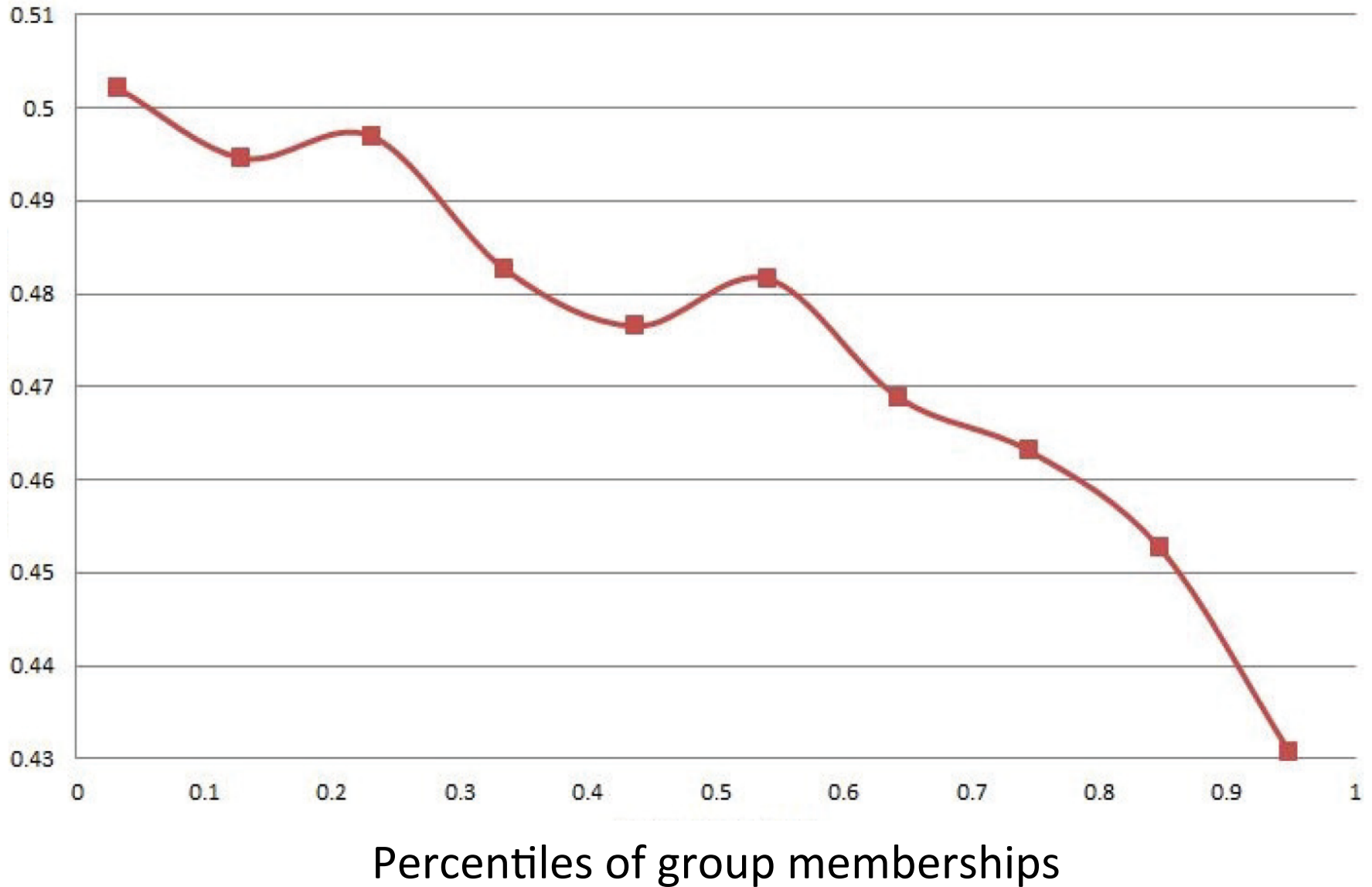
Table 1. Facebook profile features used in this study.

Correlation with Openness



Percentiles of group memberships

Conscientious



Predictive Performance

Trait	R^2	RMSE
Openness	0.11	0.29
Conscientiousness	0.17	0.28
Extraversion	0.33	0.27
Agreeableness	0.01	0.29
Neuroticism	0.26	0.28

- Predictions for percentiles
 - “user at the 85-% of agreeableness”
- Baseline: always predict 0.5 (= the median)
 - $\int_{0.0}^{1.0} (x-0.5)^2 dx = \frac{(x-0.5)^3}{3} \Big|_{0.0}^{1.0} = 0.08333$
- RMSE: $\sqrt{0.08333} = 0.289$

Reminder:
Competition

Timeline of the Competition

- Before Thu. 11h00 (and after Wed. 14h00):
Cast your vote for one submitted proposal:

<http://tinyurl.com/RuSSIR-Proposal-Voting>

- **During Thu. lecture: the top three proposals (according to online votes) are announced**
- During Fri. lecture: the top three proposals are presented in person (2 min each, max of 3 slides)

And the Top Three Proposals are ...

(not necessarily in order of number of votes)

- Xxx
- Xxx
- Xxx

Please prepare a short presentation for tomorrow
Improvise if you prefer, otherwise three slides max

Questions?

End of Day 4

ingmar@yahoo-inc.com