Genome-Wide Association with Digital Phenotypes

Jie Yuan

5/3/2016

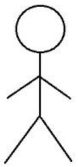
Traditional Genome-Wide Association

Questionnaires,
Medical Records

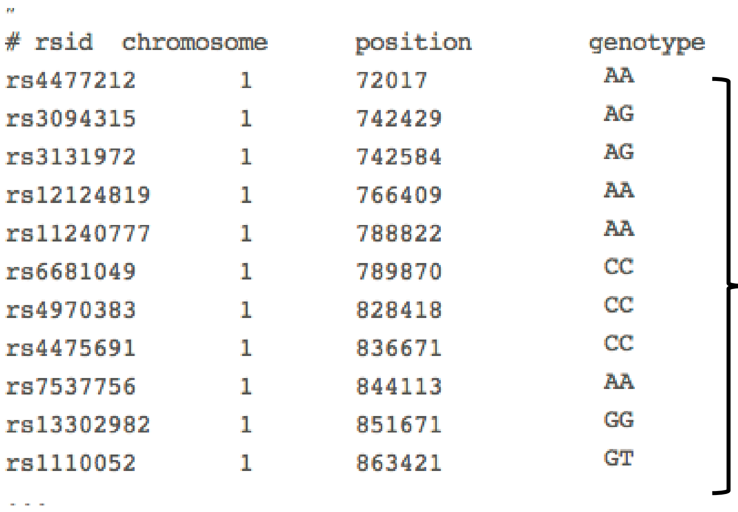
Phenotype
(+/- heart disease)

y

linear/logistic
regression



Genomic Sequencing



```
# rsid   chromosome      position        genotype
rs4477212        1       72017            AA
rs3094315        1       742429           AG
rs3131972        1       742584           AG
rs12124819       1       766409           AA
rs11240777       1       788822           AA
rs6681049        1       789870           CC
rs4970383        1       828418           CC
rs4475691        1       836671           CC
rs7537756        1       844113           AA
rs13302982       1       851671           GG
rs1110052        1       863421           GT
...
```

x

The digital phenotype

Sachin H Jain, Brian W Powers, Jared B Hawkins & John S Brownstein

In the coming years, patient phenotypes captured to enhance health and wellness will extend to human interactions with digital technology.

In 1982, the evolutionary biologist Richard Dawkins introduced the concept of the "extended phenotype"[1], the idea that phenotypes should not be limited just to biological processes, such as protein biosynthesis or tissue growth, but extended to include all effects that a gene has on its environment inside or outside of the body of the individual organism. Dawkins stressed that many delineations of phenotypes are arbitrary. Animals and humans can modify their environments, and these modifications and associated behaviors are expressions of one's genome and, thus, part of their extended phenotype. In the animal kingdom, he cites damn building by beavers as an example of the beaver's extended phenotype[1].

As personal technology becomes increasingly embedded in human lives, we think there is an important extension of Dawkins's theory—the notion of a 'digital phenotype'. Can aspects of our interface with technology be somehow diagnostic and/or prognostic for certain conditions? Can one's clinical data be linked and analyzed together with online activity and behavior data

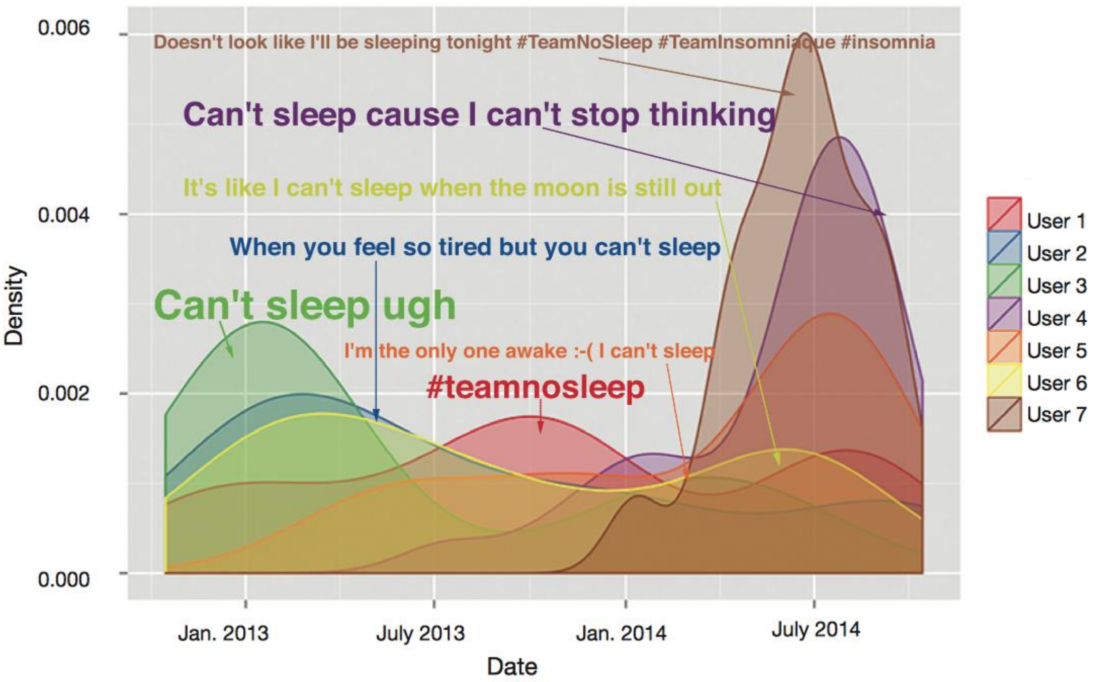


Figure 1 Timeline of insomnia-related tweets from representative individuals. Density distributions (probability density functions) are shown for seven individual users over a two-year period. Density on the *y* axis highlights periods of relative activity for each user. A representative tweet from each user is shown as an example.

Facebook Likes are predictive of a variety of demographic traits

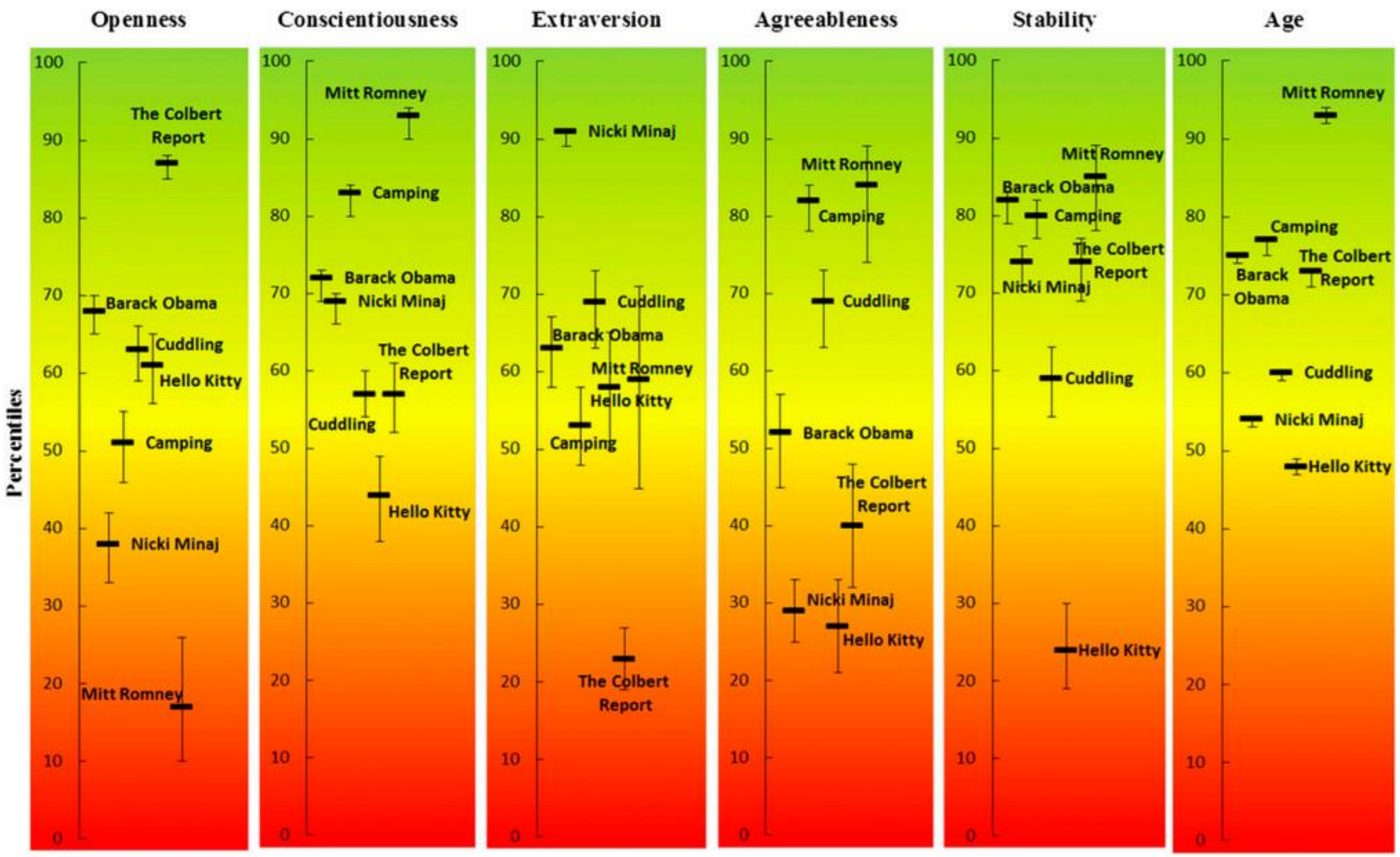


Fig. S1. Average levels of five personality traits and age of the users associated with selected Likes presented on the percentile scale. For example, the average extraversion of users associated with "The Colbert Report" was relatively low: it was lower only for 23% of other Likes in the sample. Error bars signify 95% confidence intervals of the mean.

(But still requires questionnaires for labels)

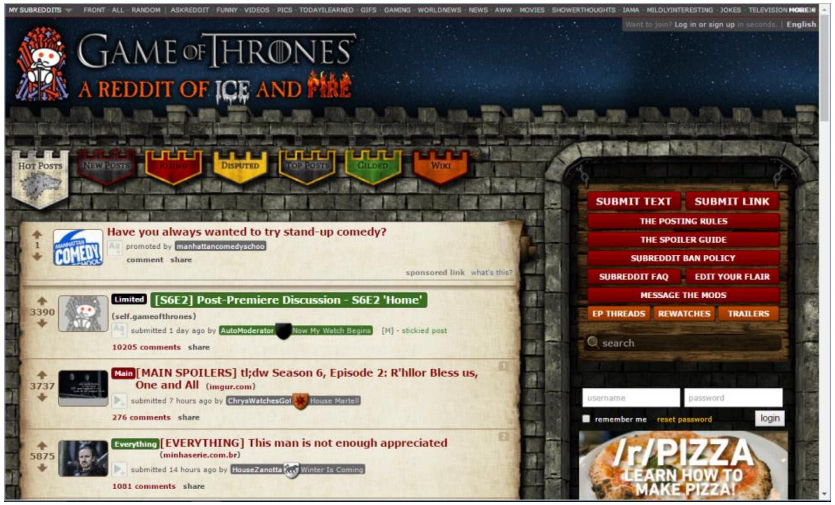
Kosinski et al. (2013)

Problem:

- Can we identify traits and word-trait associations from text, without access to external labels (e.g. questionnaire results)?
  - An extension of Dr. Mark Dredze’s sLDA classification problem

- Given these inferred traits, can we identify genome-trait associations?

Reddit is divided into communities (subreddits) with highly specialized interests

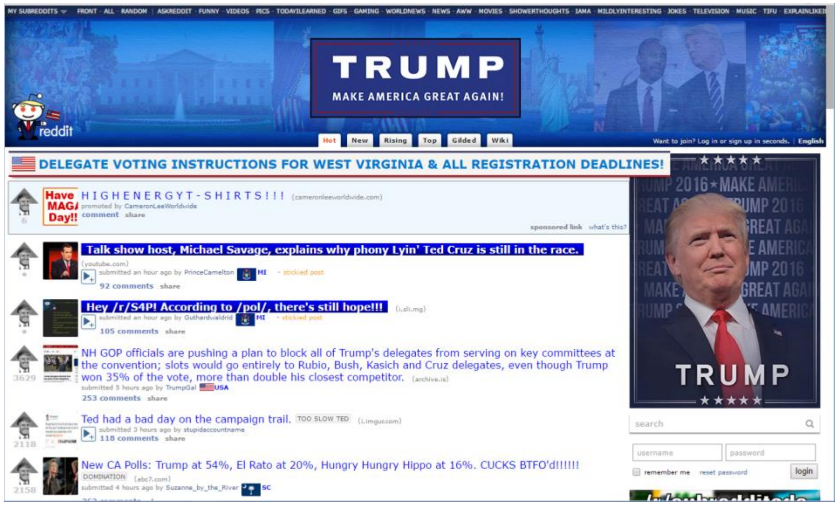


1 Have you always wanted to try stand-up comedy?
promoted by manhattancomedyschoo
comment share
sponsored link what's this?

3390 Limited [S6E2] Post-Premiere Discussion - S6E2 'Home' (self.gameofthrones)
submitted 1 day ago by AutoModerator Now My Watch Begins [M] - stickied post
10205 comments share

3737 Main [MAIN SPOILERS] tl;dw Season 6, Episode 2: R'hllor Bless us, One and All (imgur.com)
submitted 7 hours ago by ChrysWatchesGoT House Martell
276 comments share

5875 Everything [EVERYTHING] This man is not enough appreciated (minhaserie.com.br)
submitted 14 hours ago by HouseZanotta Winter Is Coming
1081 comments share


MY SUBREDDITS · FRONT · ALL · RANDOM | ASKREDDIT · FUNNY · VIDEOS · PICS · TODAYILEARNED · GIFS · GAMING · WORLDNEWS · NEWS · AWW · MOVIES · SHOWERTHOUGHTS · IAMA · MILDLYINTERESTING · JOKES · TELEVISION · MUSIC · TIFU · EXPLAINLIKEI
TRUMP
MAKE AMERICA GREAT AGAIN!
reddit
Hot
New
Rising
Top
Gilded
Wiki
Want to join? Log in or sign up in seconds. | English
TRUMP 2016 ★ MAKE AMERICA GREAT AGAIN
TRUMP
search
username
password
remember me
reset password
login

DELEGATE VOTING INSTRUCTIONS FOR WEST VIRGINIA & ALL REGISTRATION DEADLINES!

6 Have MAGA Day!! HIGHENERGYT-SHIRTS!!! (cameronleeworldwide.com)
promoted by CameronLeeWorldwide
comment share
sponsored link what's this?

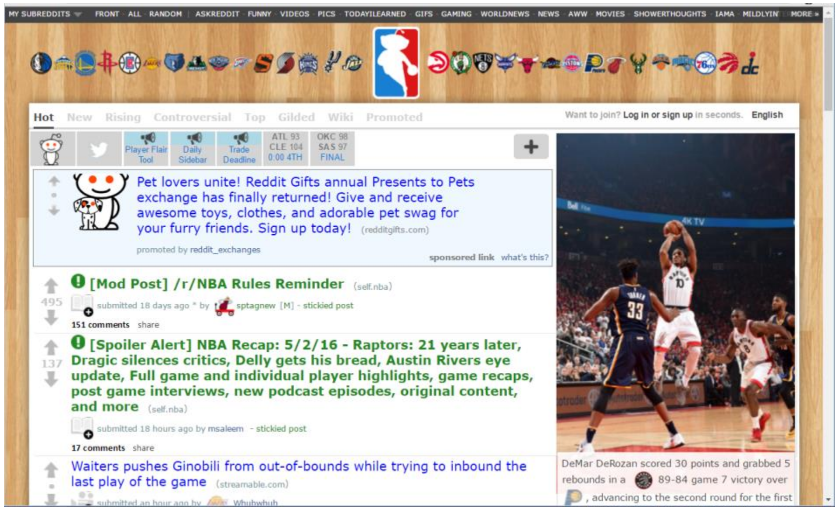
Talk show host, Michael Savage, explains why phony Lyin' Ted Cruz is still in the race. (youtube.com)
submitted an hour ago by PrinceCamelton MI - stickied post
92 comments share

Hey /r/S4P! According to /pol/, there's still hope!!! (i.sli.mg)
submitted an hour ago by Guthardwaldrid MI - stickied post
105 comments share

3629 NH GOP officials are pushing a plan to block all of Trump's delegates from serving on key committees at the convention; slots would go entirely to Rubio, Bush, Kasich and Cruz delegates, even though Trump won 35% of the vote, more than double his closest competitor. (archive.is)
submitted 5 hours ago by TrumpGal USA
253 comments share

2118 Ted had a bad day on the campaign trail. TOO SLOW TED (i.imgur.com)
submitted 3 hours ago by stupidaccountname
118 comments share

2158 New CA Polls: Trump at 54%, El Rato at 20%, Hungry Hungry Hippo at 16%. CUCKS BTFO'd!!!!!! DOMINATION (abc7.com)
submitted 4 hours ago by Suzanne_by_the_River SC


MY SUBREDDITS · FRONT · ALL · RANDOM | ASKREDDIT · FUNNY · VIDEOS · PICS · TODAYILEARNED · GIFS · GAMING · WORLDNEWS · NEWS · AWW · MOVIES · SHOWERTHOUGHTS · IAMA · MILDLYIN · MORE »
Hot
New
Rising
Controversial
Top
Gilded
Wiki
Promoted
Want to join? Log in or sign up in seconds. English
Player Flair Tool
Daily Sidebar
Trade Deadline
ATL 93 CLE 104 0:00 4TH
OKC 98 SAS 97 FINAL

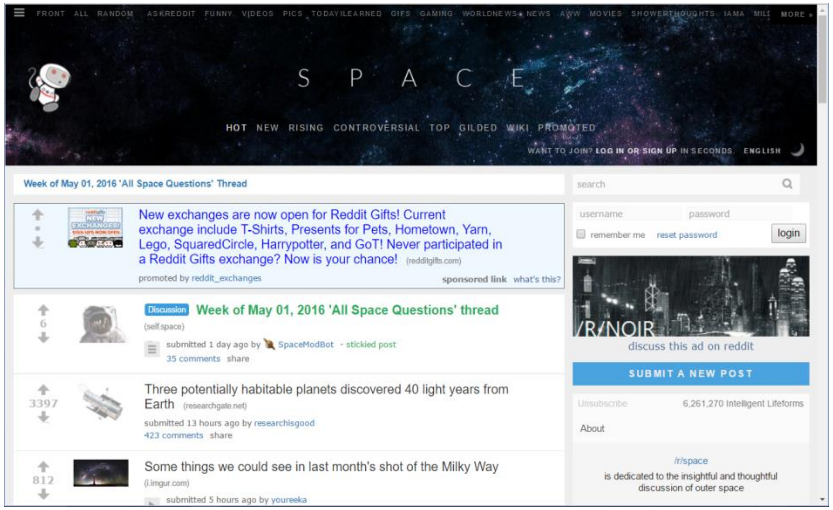
Pet lovers unite! Reddit Gifts annual Presents to Pets exchange has finally returned! Give and receive awesome toys, clothes, and adorable pet swag for your furry friends. Sign up today! (redditgifts.com)
promoted by reddit_exchanges
sponsored link what's this?

495 [Mod Post] /r/NBA Rules Reminder (self.nba)
submitted 18 days ago * by sptagnew [M] - stickied post
151 comments share

137 [Spoiler Alert] NBA Recap: 5/2/16 - Raptors: 21 years later, Dragic silences critics, Delly gets his bread, Austin Rivers eye update, Full game and individual player highlights, game recaps, post game interviews, new podcast episodes, original content, and more (self.nba)
submitted 18 hours ago by msaleem - stickied post
17 comments share

Waiters pushes Ginobili from out-of-bounds while trying to inbound the last play of the game (streamable.com)
submitted an hour ago by Whubwhuh

DeMar DeRozan scored 30 points and grabbed 5 rebounds in a 89-84 game 7 victory over , advancing to the second round for the first


FRONT · ALL · RANDOM · ASKREDDIT · FUNNY · VIDEOS · PICS · TODAYILEARNED · GIFS · GAMING · WORLDNEWS · NEWS · AWW · MOVIES · SHOWERTHOUGHTS · IAMA · MILI · MORE »
SPACE
HOT
NEW
RISING
CONTROVERSIAL
TOP
GILDED
WIKI
PROMOTED
WANT TO JOIN? LOG IN OR SIGN UP IN SECONDS. ENGLISH
search
username
password
remember me
reset password
login
/R/NOIR
discuss this ad on reddit
SUBMIT A NEW POST
Unsubscribe
About

Week of May 01, 2016 'All Space Questions' Thread

New exchanges are now open for Reddit Gifts! Current exchange include T-Shirts, Presents for Pets, Hometown, Yarn, Lego, SquaredCircle, Harrypotter, and GoT! Never participated in a Reddit Gifts exchange? Now is your chance! (redditgifts.com)
promoted by reddit_exchanges
sponsored link what's this?

6 Discussion Week of May 01, 2016 'All Space Questions' thread (self.space)
submitted 1 day ago by SpaceModBot - stickied post
35 comments share

3397 Three potentially habitable planets discovered 40 light years from Earth (researchgate.net)
submitted 13 hours ago by researchisgood
423 comments share

812 Some things we could see in last month's shot of the Milky Way (i.imgur.com)
submitted 5 hours ago by youreeka

6,261,270 Intelligent Lifeforms

/r/space is dedicated to the insightful and thoughtful discussion of outer space

Discovering Word Associations

- Can we find overrepresented words within subreddits?
  - Calculate $\frac{f-g}{f+g}$, where f and g are % frequencies in the subreddit and rest of Reddit, respectively
  - Frequency cutoff filter to remove rare words (e.g. typos) appearing in the subreddit

A sampling of medically relevant subreddits and their overrepresented words

r/depression	r/diabetes	r/stopsmoking	r/insomnia
therapist	insulin	nicotine	insomnia
medication	diabetic	quitting	melatonin
meds	glucose	cigarettes	mg
suicidal	carb	smoker	sleepy
antidepressants	dexcom	congratulations	ambien
scared	diagnosed	badge	pills
counselor	basal	smokers	caffeine
mg	keto	nonsmoker	meds
diagnosed	endo	addicted	addictive
sadness	lantus	carrs	medication

- Create a matrix of traits, and word frequencies (or sLDA topics)

	depression	insomnia	“sleepy”	“caffeine”	(topic 1)	(topic 2)
user	0	1	15	10	0.20	0.05
user	1	1				
user	0	0			.	
user*	?	?			.	
user*	?	?			.	
user*	?	?				

+/- phenotype
(is the user a
frequent
poster on
relevant
subreddit)

frequencies for
overrepresented
words, and/or
significant topics
from sLDA

- Create a matrix of traits, and word frequencies (or sLDA topics)

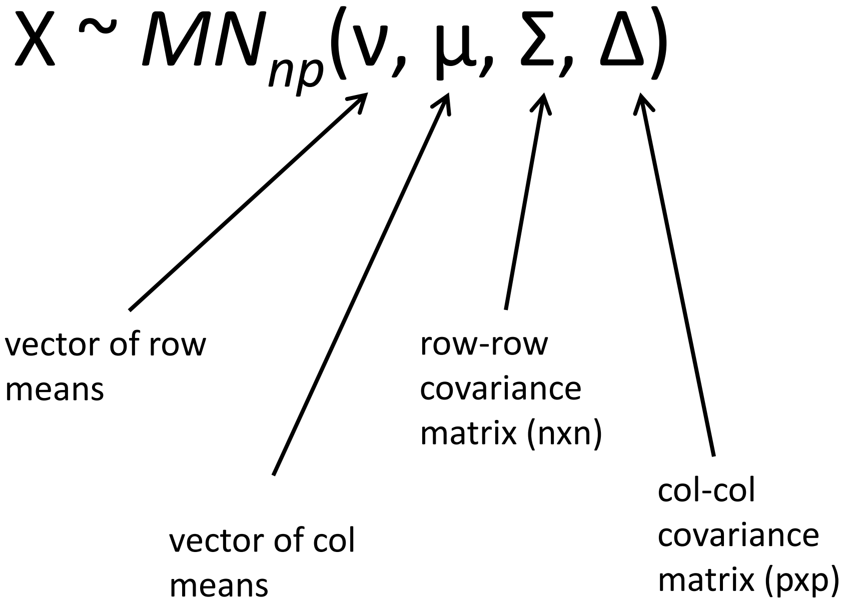
	depression	insomnia	“sleepy”	“caffeine”	(topic 1)	(topic 2)
user	0	1	15	10	0.20	0.05
user	1	1				
user	0	0			.	
user*	?	?			.	
user*	?	?			.	
user*	?	?				

user: a Reddit user whose digital phenotypes are the set of frequently visited subreddits

user*: a user with unlabeled social media text (e.g. from Facebook), whose digital phenotypes we want to impute

- Can we impute the “?” entries

Matrix imputation using Transposable Regularized Covariance Model (TRCM)



$X \sim MN_{np}(\nu, \mu, \Sigma, \Delta)$

vector of row means

vector of col means

row-row covariance matrix (nxn)

col-col covariance matrix (pxp)

- Use EM to calculate row and col parameters
- Plug in mean estimates for imputed values

For a missing value in row *i* and col *j*

$$x_{ij} = \nu_i + \mu_i + \varepsilon_{ij}$$

$$\varepsilon_{ij} \sim N(0, \Sigma_{ii}\Delta_{jj})$$

Allen, Tibshirani. (2010)