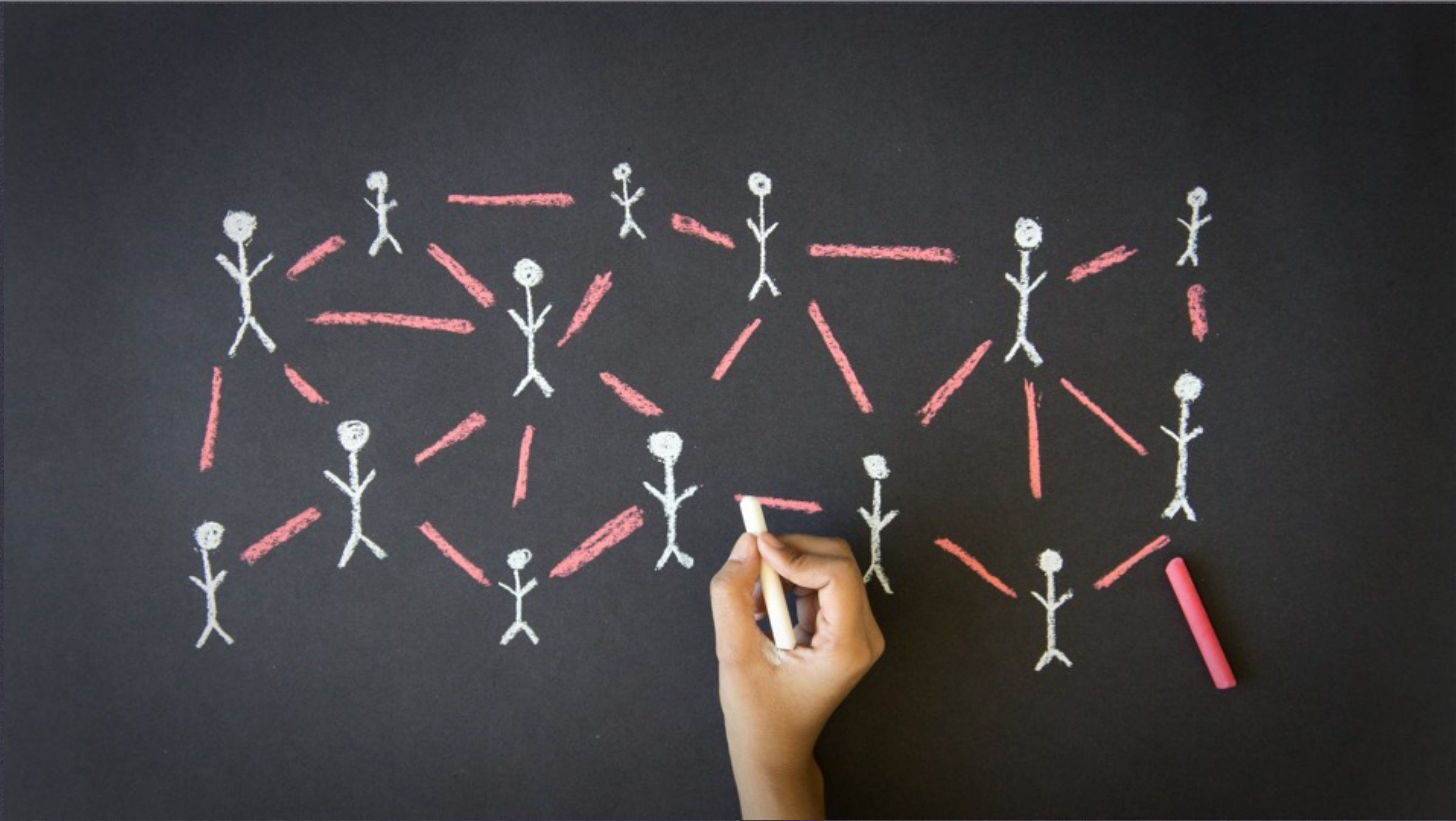


PROJECT

MAKING FASTER & BETTER INTRODUCTIONS

RECOMMENDER SYSTEM FOR BETTER CONNECTIONS

PROBLEM STATEMENTS

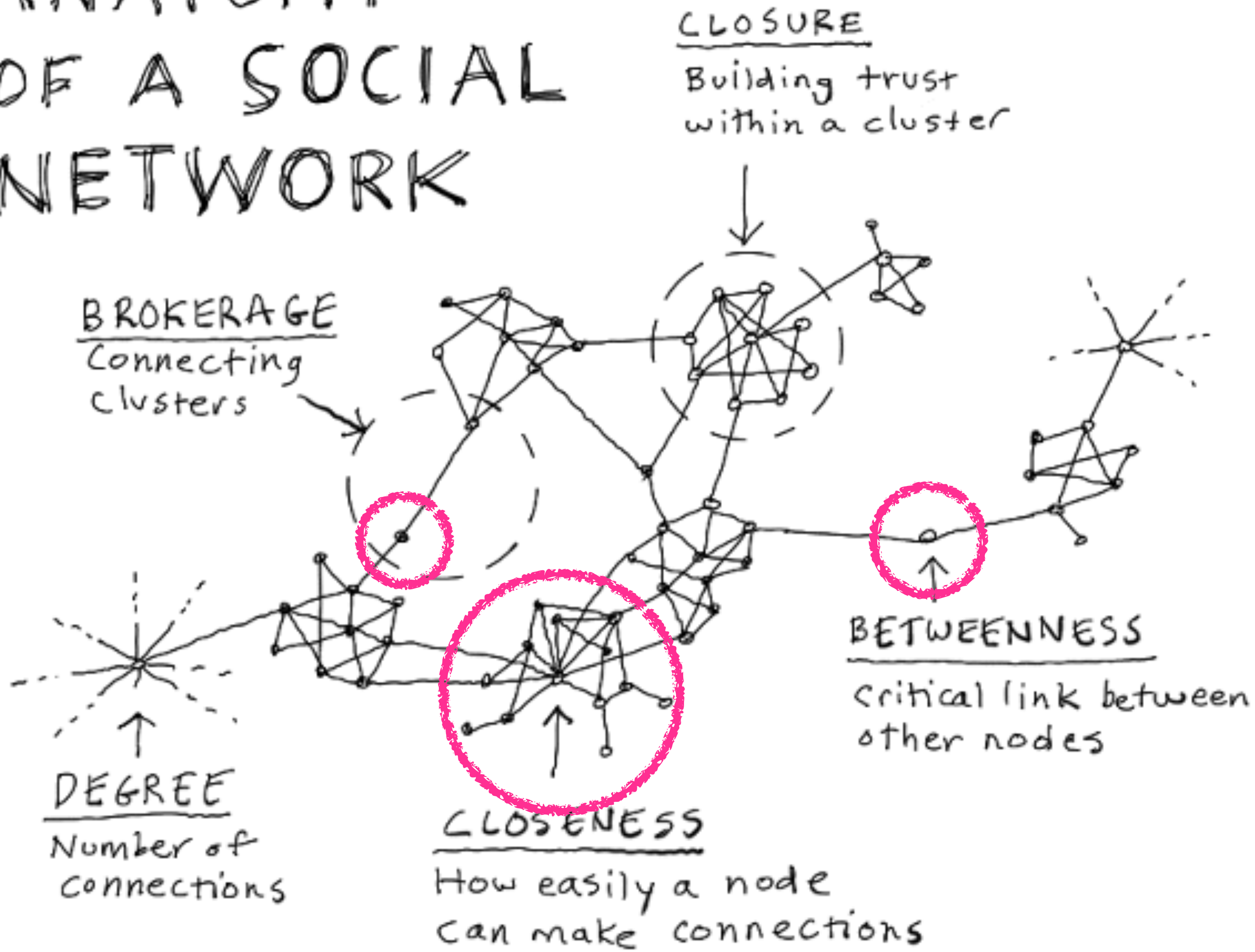
- Making introductions are important. From dating to business, most of the introductions tend to rely heavily on human brains.
- Targeted Audience: Super Connectors

WHO ARE SUPER CONNECTORS?

"Super connectors are people with more than just a strong social media followings and lots of friends. They're people who are making high-level connections on a regular basis through methodical and well thought out — albeit "simple" — introductions."

– *thenextweb.com*

ANATOMY OF A SOCIAL NETWORK

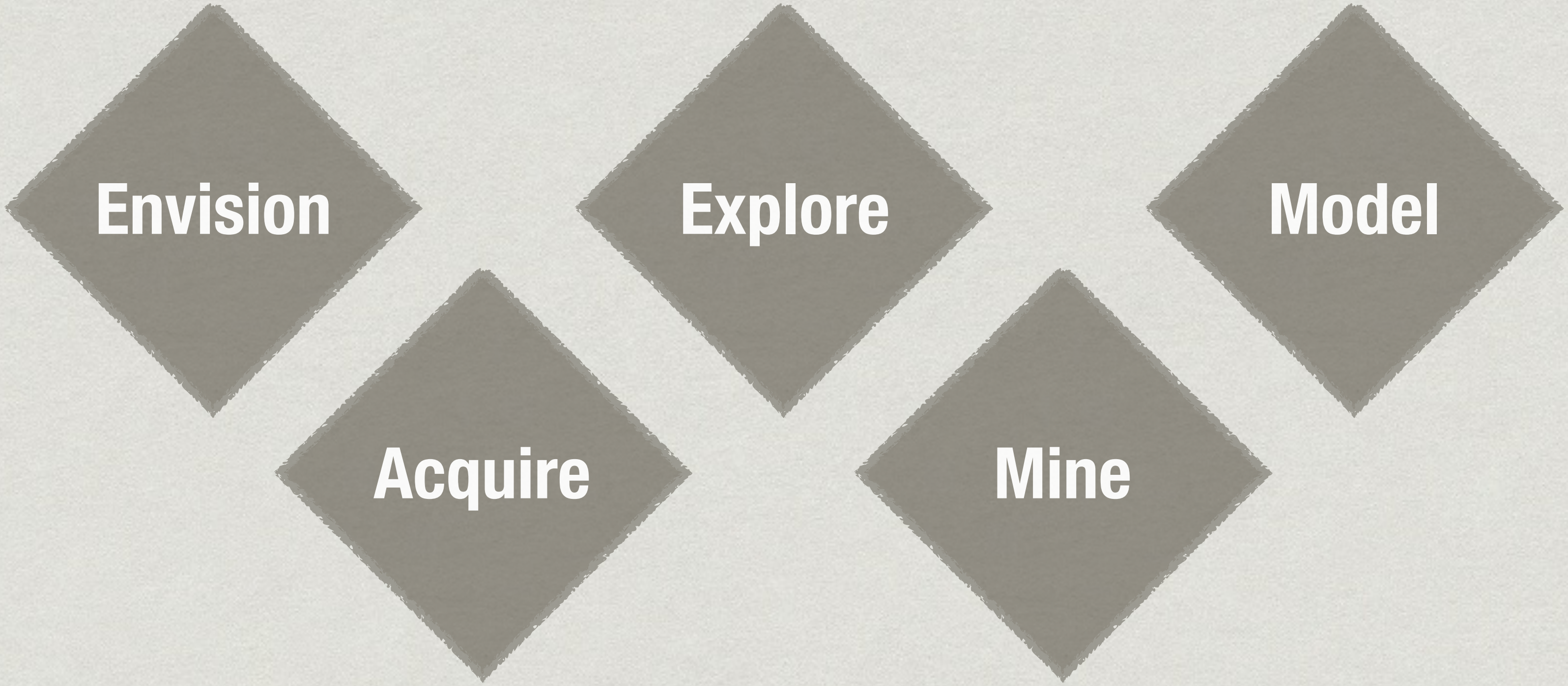


HOW DO HUMAN MAKE INTROS?

- **Brain**: “person 1 is thinking about this a lot or into this a lot, and person 2 is also! Let me intro them!”
- **Tools:** evernote, human memory, spreadsheet…
- Wait! But what if you have… over 4000 friends?

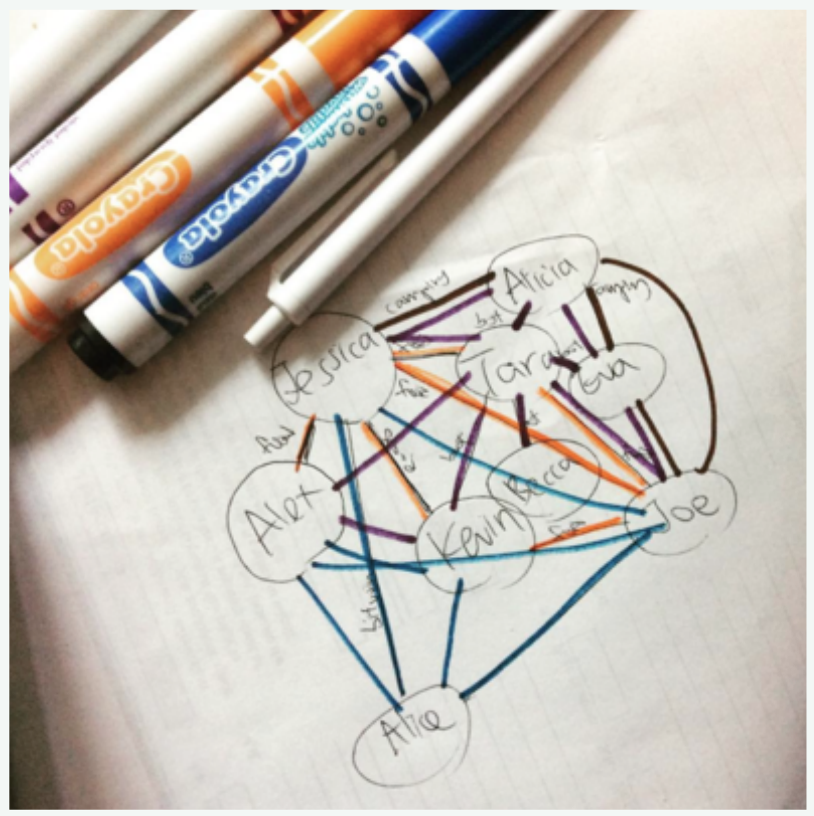
By connecting people based on their past conversations with a super-connector, he/she can systematically make efficient introductions faster than a human brain

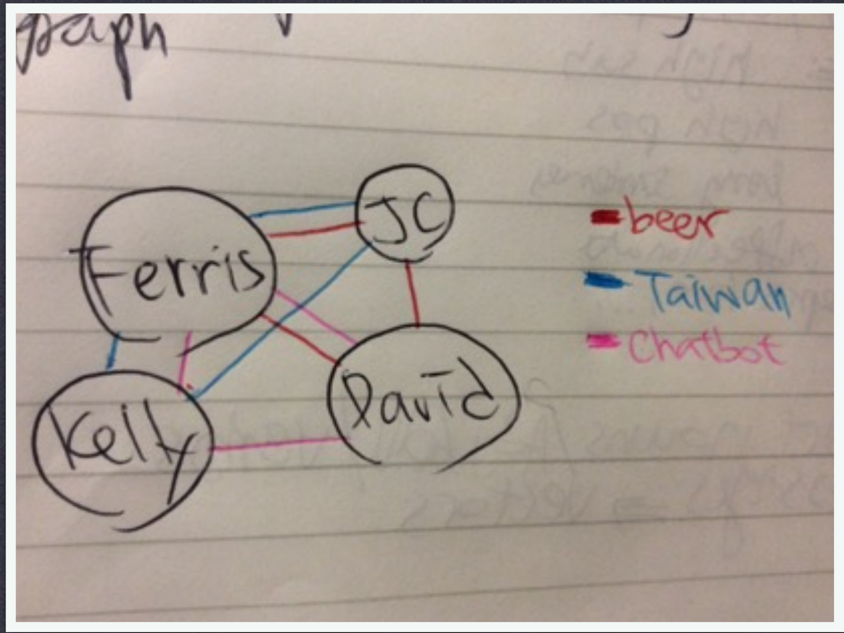
APPROACH

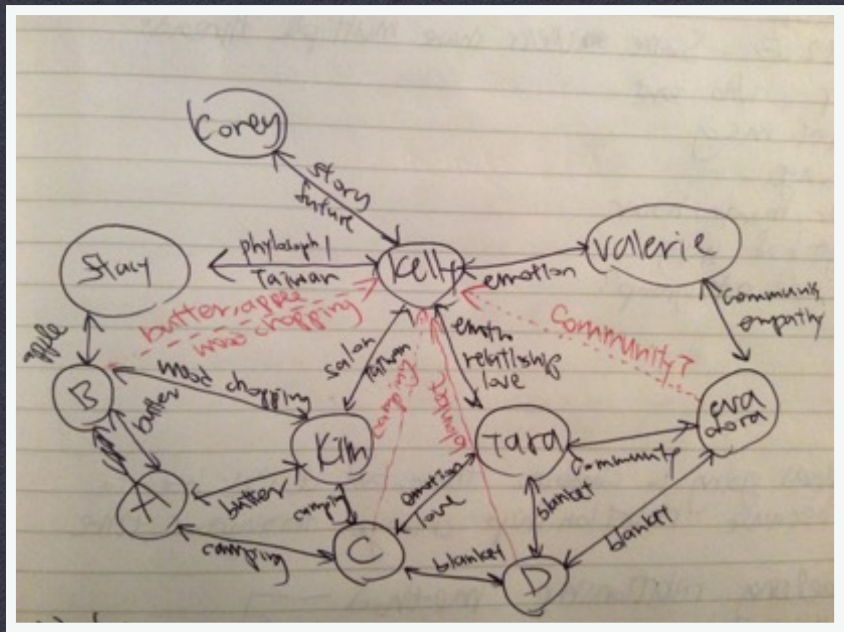


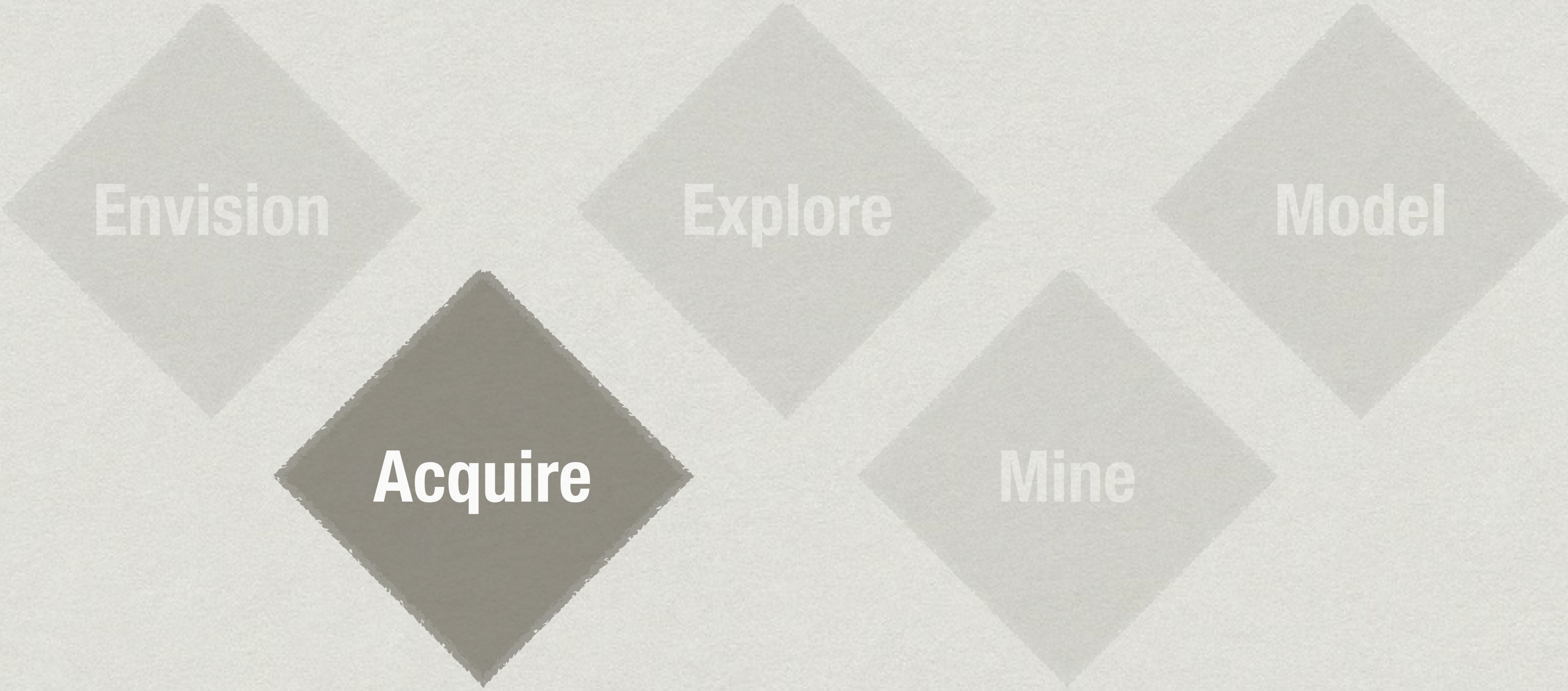

Envision
Explore
Model
Acquire
Mine

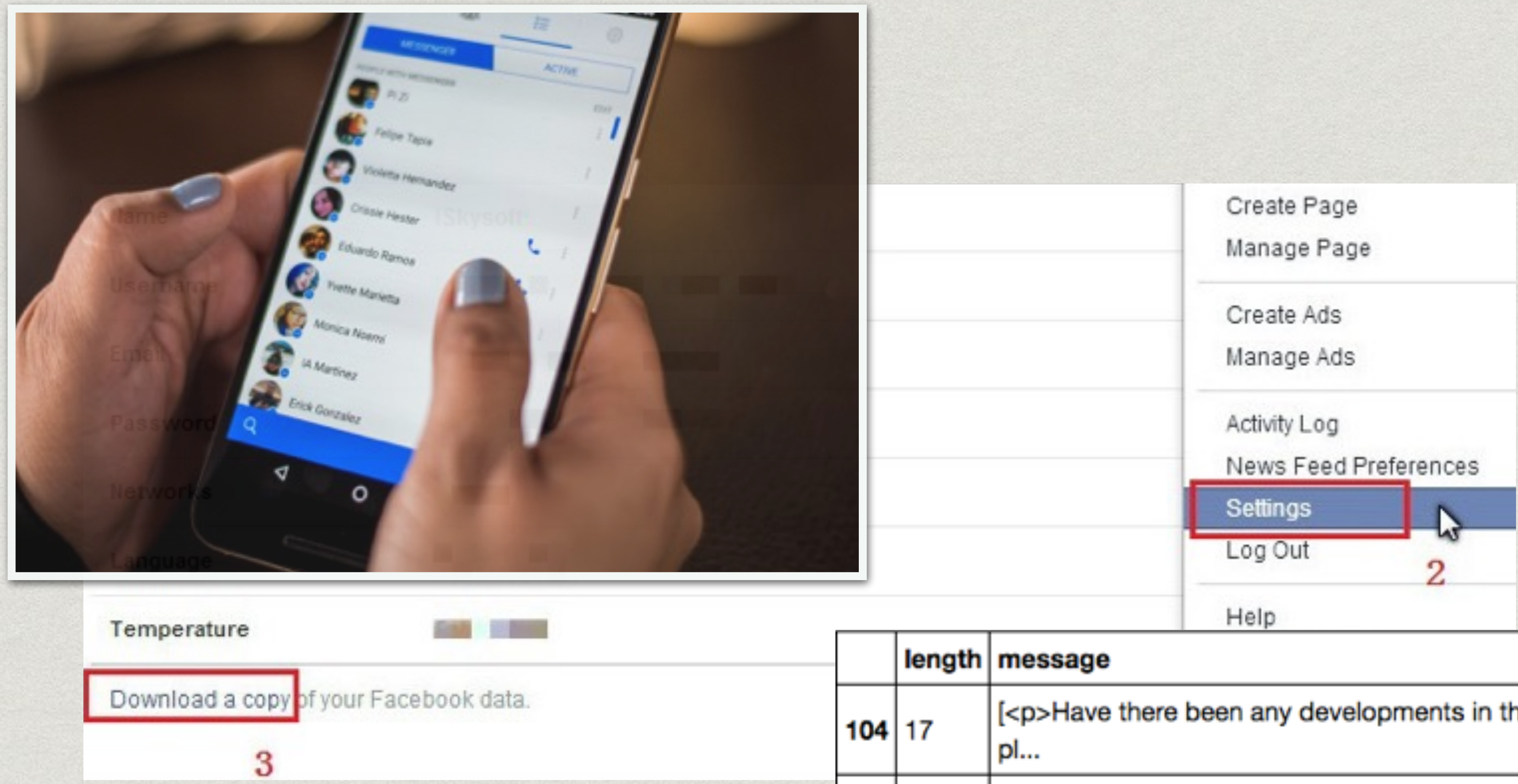
Envision through fast prototype





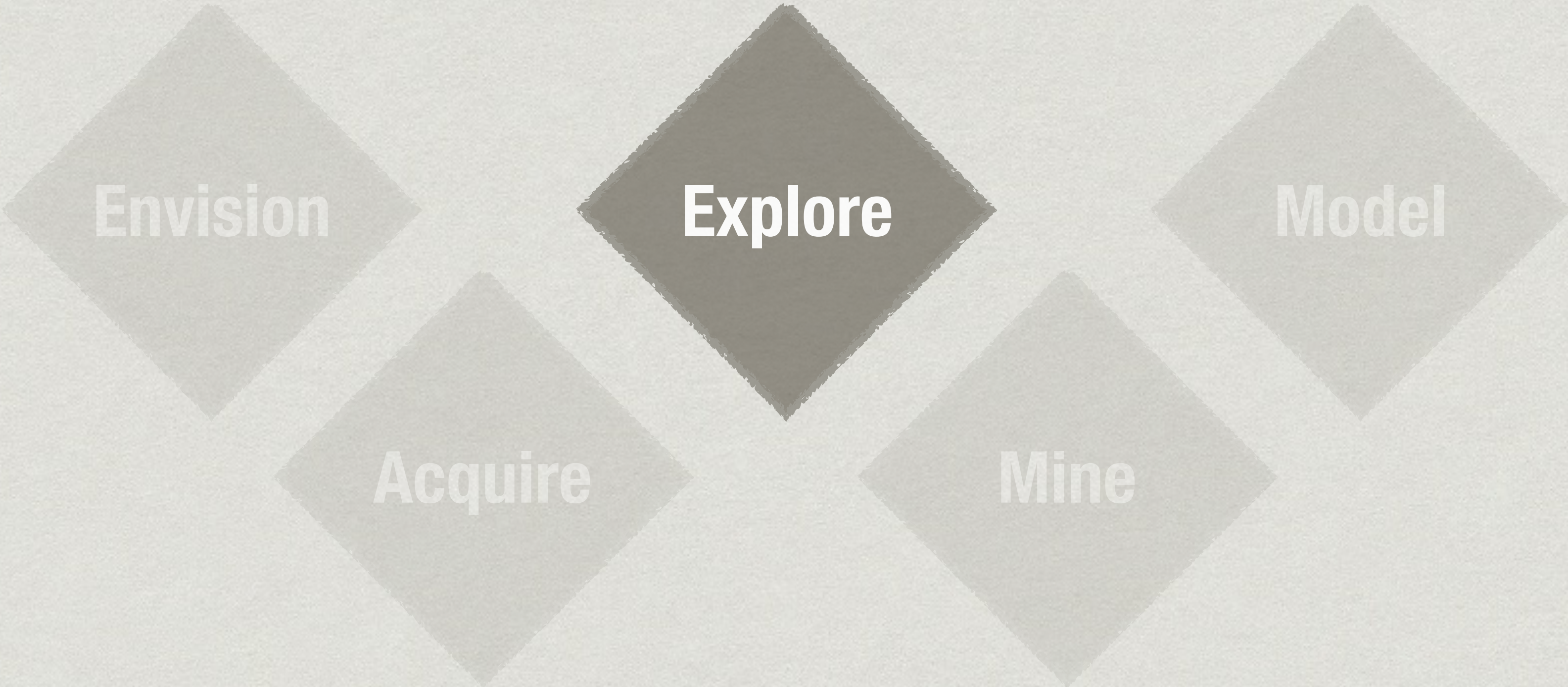



Envision
Explore
Model
Acquire
Mine



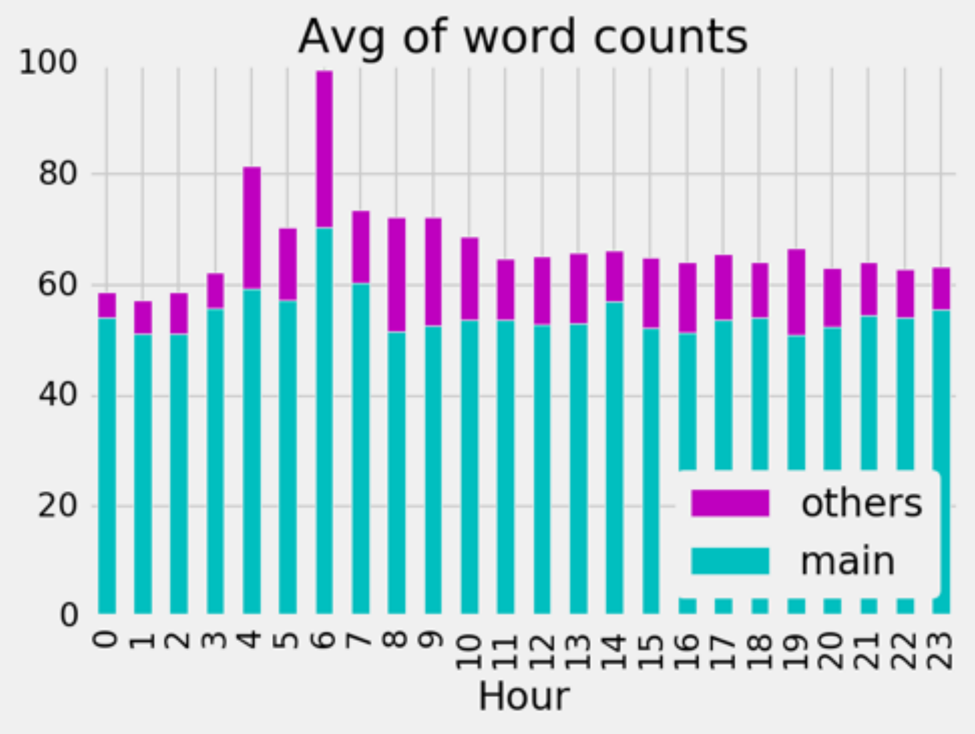
	length	message	thread	user
104	17	[<p>Have there been any developments in the pl...	104	[10387 50353
105	4	[<p>Hiya, Don't even worry about it - I'll tak...	105	[10495 10635

- **161MB** of one super connector's Facebook chat history since **2006**
- Scraping from an exported **HTML** file
- over **4700** friends, **~3192** threads, and over 1700 unknown names due to the nature of Facebook data


Envision
Explore
Model
Acquire
Mine

EXPLORATORY DATA ANALYSIS

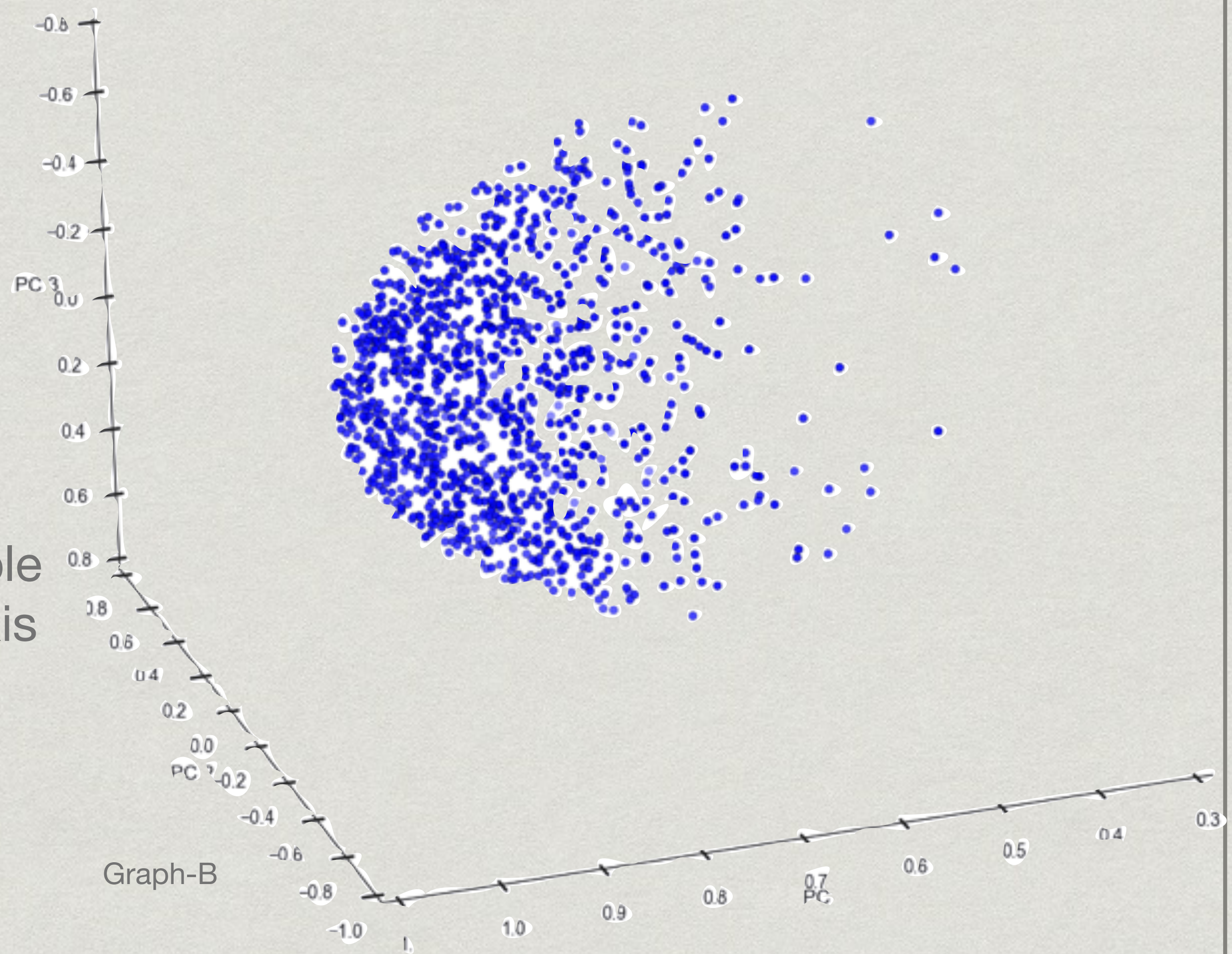
- From “7 networking tips successful super connectors”, super-connectors tend to ask more questions than answering one which explains Graph-A



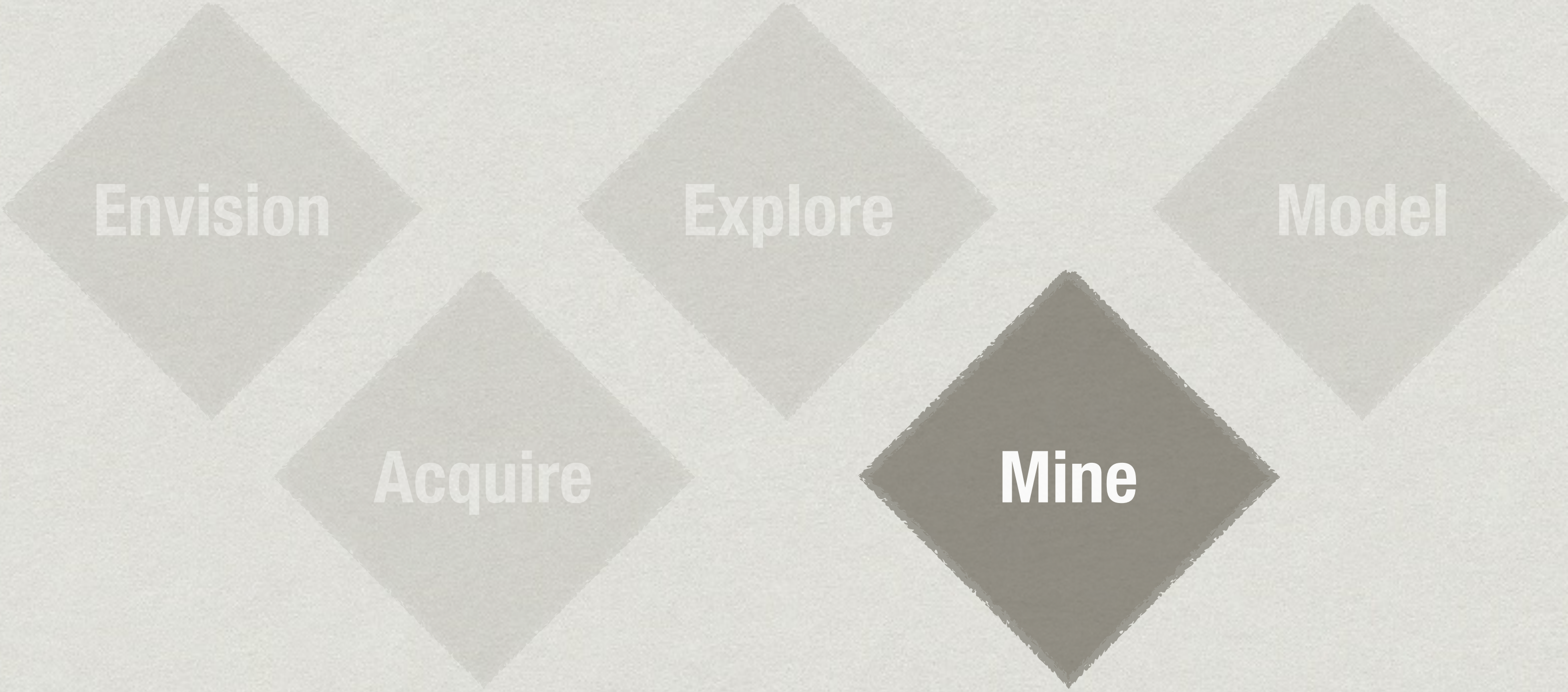
Graph-A

LATENT SEMANTIC ANALYSIS

- Observed groups of people who are similar on one axis and very different on another



Graph-B

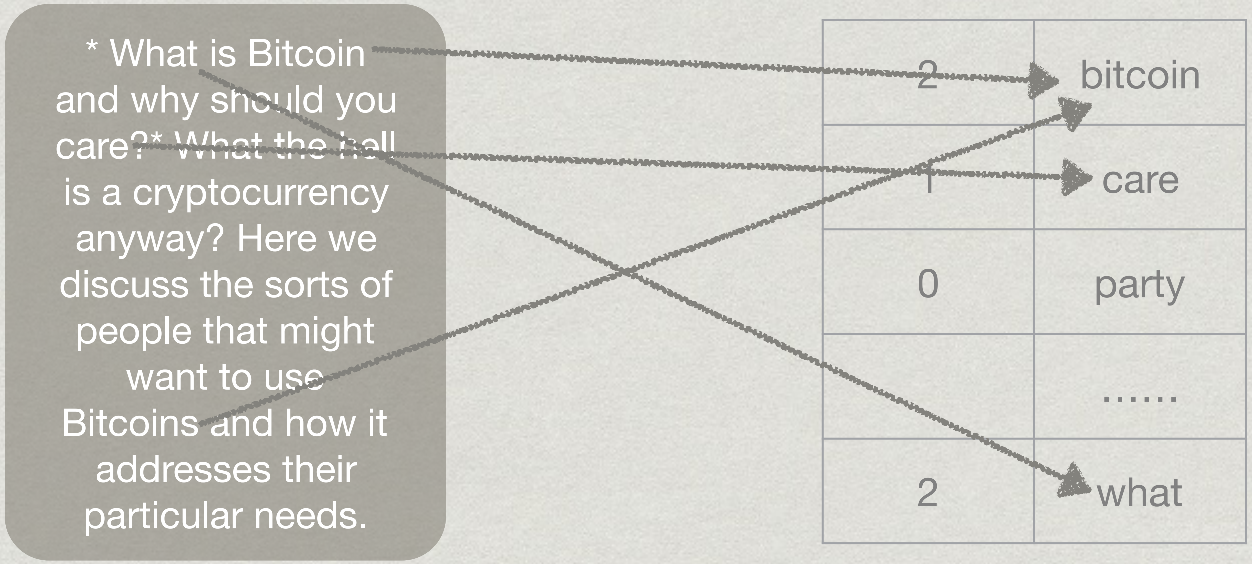

Envision
Explore
Model
Acquire
Mine

FORMAT, CLEAN, SLICE, COMBINE, FEATURE ENGINEERING

- **Format:** HTML to a list of dictionaries to DataFrame
- **Clean:** unicode, emojis, links, numbers, emails, stop words, one letter words, SnowStemmer, and drop messages sent before 2011 and users who have sent less than 20 messages
- **Slice & Combine:** group messages by senders and turn them into a text document per sender

FORMAT, CLEAN, SLICE, COMBINE, FEATURE ENGINEERING

- **Features:** converting text into vectors via Count-Vectorizer and Tf-idf Transformer (bag of words)



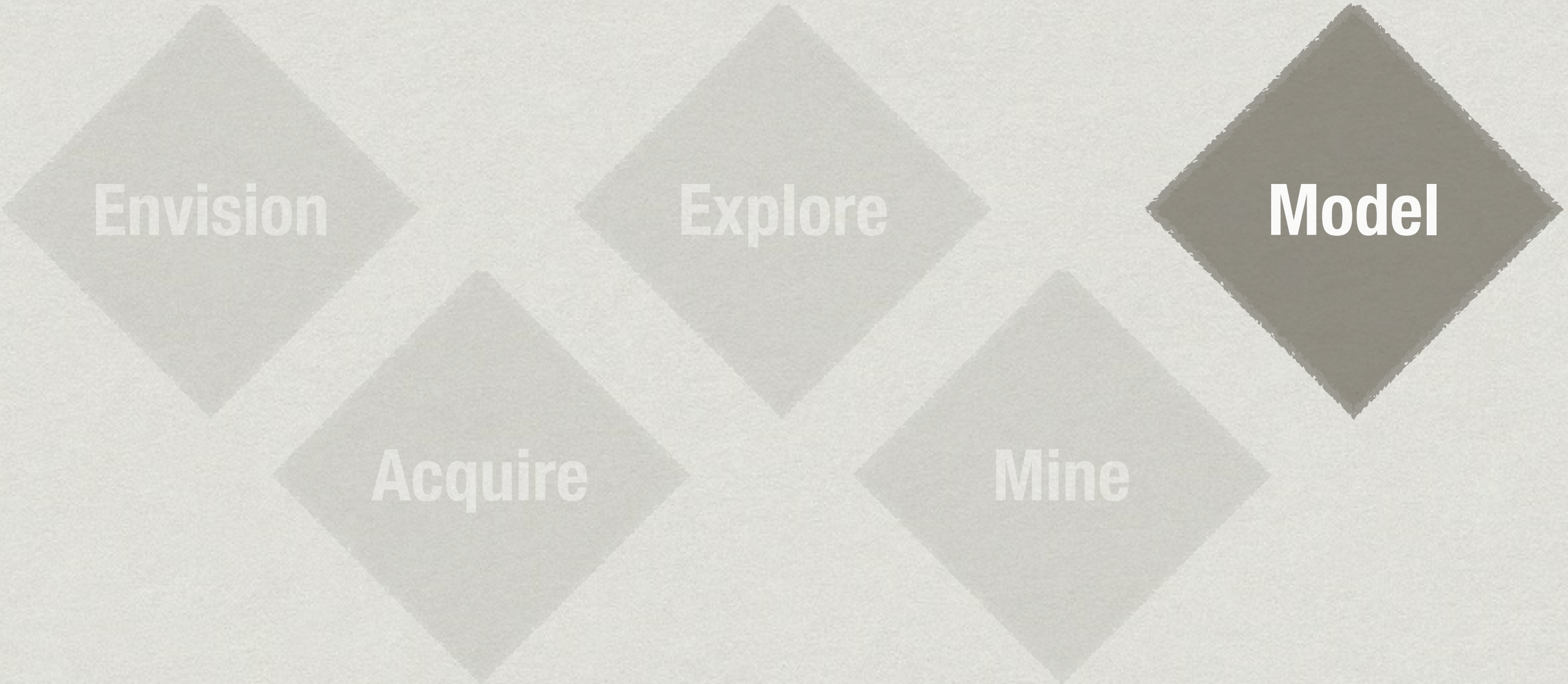
FORMAT, CLEAN, SLICE, COMBINE, FEATURE ENGINEERING

- **More Features:** creating numerical metrics for each person
  - time orientation
  - us vs them
  - sentiment score (negative <—> positive)
  - subjectivity vs objectivity
  - certainty vs uncertainty
  - Regressive Imagery Dictionary (psychoanalysis from text)
  - Entity extraction with or without POS
  - Using TextBlob, Pattern Library by CLiPS, and RID
  - Did not end up using them because of time constraint

an example of how other people use RID. Screenshot from 750words.com

CLiPS
COMPUTATIONAL LINGUISTICS & PSYCHOLINGUISTICS RESEARCH CENTER

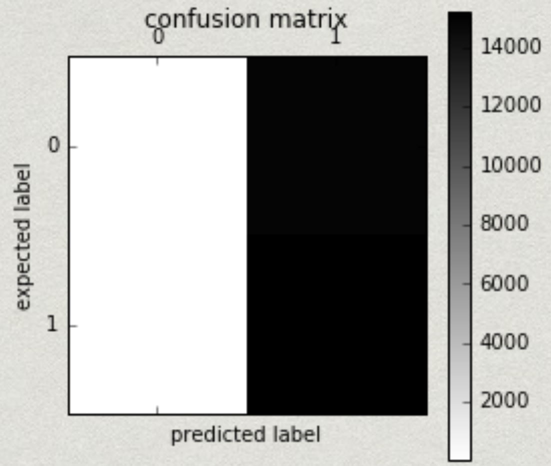
TextBlob


Envision
Explore
Model
Acquire
Mine

MODELLING TRIALS AND ERRORS:

Trial ONE: Predict private conversations from group conversations

- very difficult to make more accurate predictions due to the nature of Facebook user naming system



the Multinomial Naive Bayes model classified almost every conversation as private chat

MODELLING TRIALS AND ERRORS:

Trial TWO: Using **LSA** to match people based on text similarity, then use **LDA** to give conversational topic suggestions

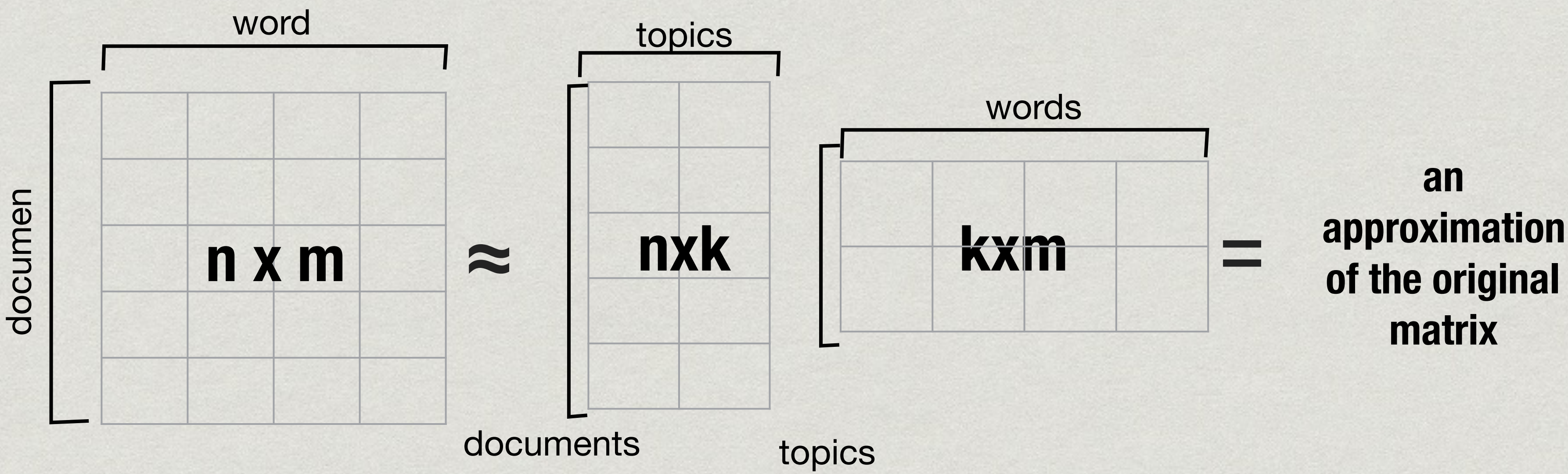
- metrics on LSA: rank similarities of 1199 people with the super connector and manually rated the accuracy of top 257 people **Results:** 34 super close, 52 really close, 39 close, and 132 not close. 48% of the rated relationships were accurate.
- Tried LDA for 5 days while tuning different hyper-parameters and did not get enough representable topics.
- But the results helped to make a more customised stopwords list

MODELLING TRIALS AND ERRORS:

use decomposition to extract topics

Trail THREE: Non-Negative Matrix Factorization

0~positive word frequencies



MODELLING TRIALS AND ERRORS:

use decomposition to extract topics

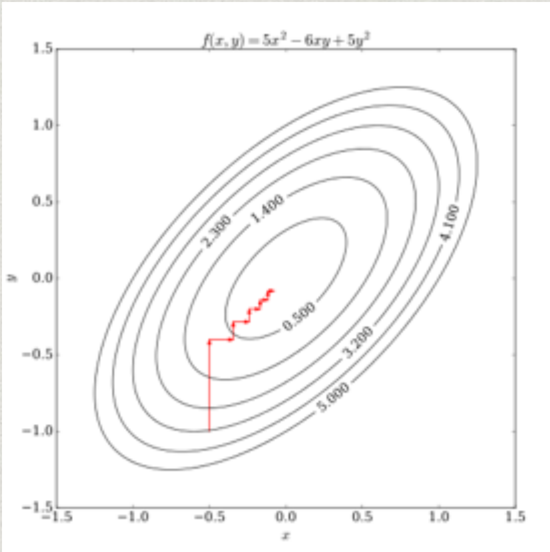
3. Non-Negative Matrix Factorization

0~positive word frequencies

- **Frobenius Norm:** the square root of the sum of the absolute squares of a matrix's elements

$$\|A\|_F \equiv \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$$

- **Coordinate Descent** with 0.1 learning rate to minimise the squared error of (actual - predict) ^2 and to find the right factors



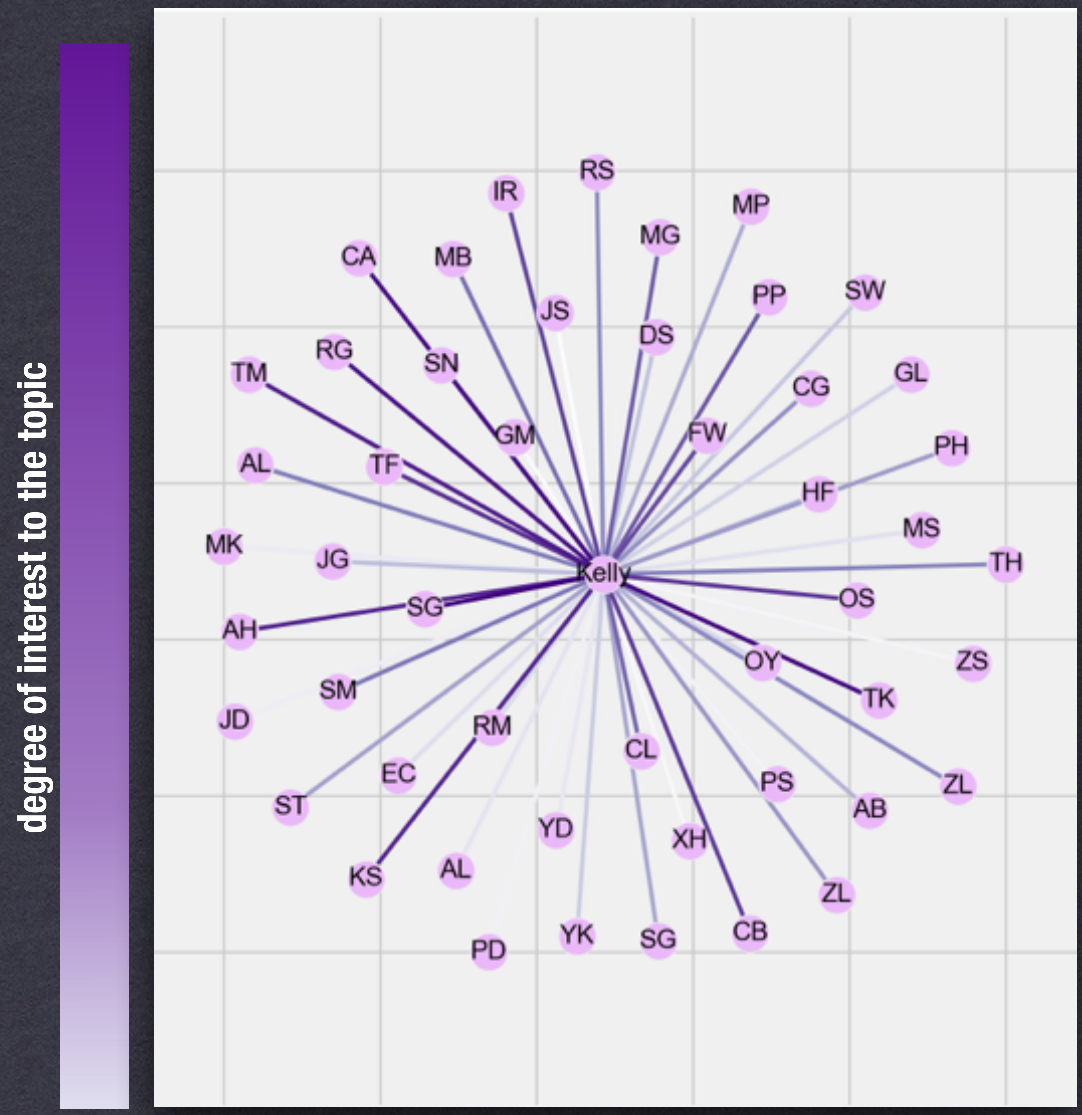
NNMF RESULTS

self-care	startups	photo-graphy	rationality	crypto-currency	design-thinking	mindful-ness	life-problems	chatbots	writing
feel	company	shoot	CFAR	bitcoin	design	life	mom	bot	post
talk	market	picture	berkeley	coin	idea	mind	live	data	group
thoght	job	wedding	workshop	btc	project	meditation	job	sensay	blog
share	develop	photo-graphy	meetup	currency	graphic	emotion	parent	chat	write
emotion	startup	cosplay	bay	buy	experience	self	money	code	read
hard	business	assist	rationalist	card	product	learn	kid	twitter	twitter
converse	hire	light	MIRI	sell	meet	experiece	family	hack	idea

- 50 topics in total, here are some legible examples with manually labeled topics

WHO SHOULD I INTRODUCE TO KELLY?

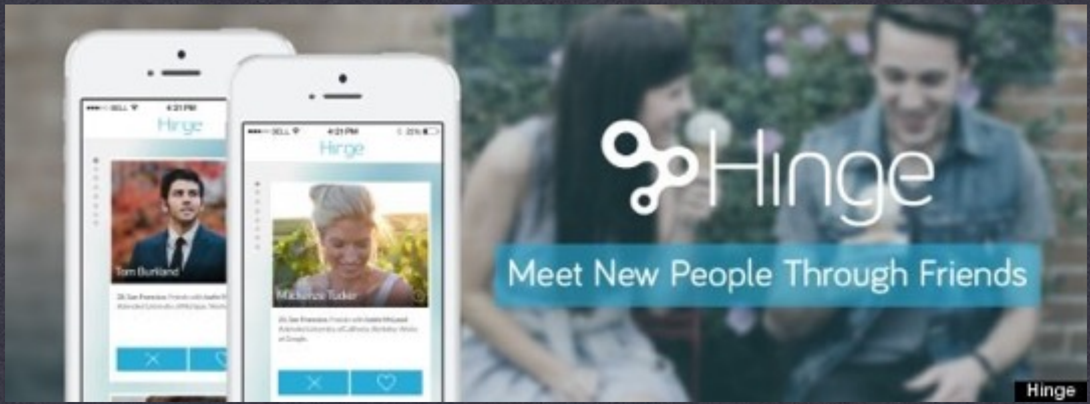
- choose one topic of interest and chose one person to make intros to
- a node = person’s initial
- an edge = interest level

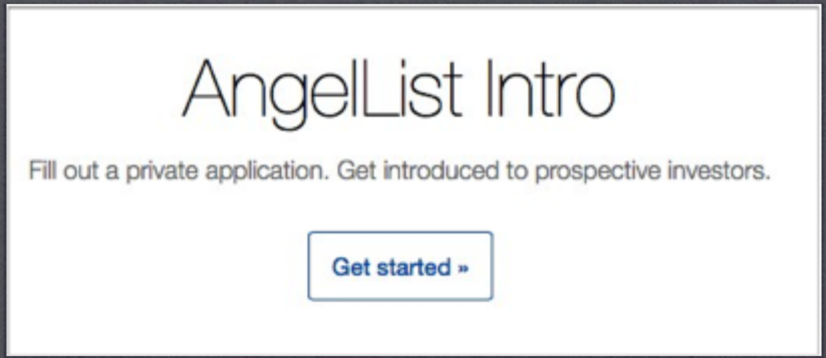


LIMITATIONS & ASSUMPTIONS

- challenging to evaluate the accuracy of a model qualitatively besides using subjective human evaluations
- assuming that people can be represented by keywords
- assuming that emojis, links, time, punctuations contain no useful information
- the model might recommend you to the same person because Facebook gives you a new ID every time you changed your name
- the model assumes that people's interests are static and don't change

WHO WILL BENEFIT?





WHAT'S NEXT?

- Incorporate generated numeric data
- use Gephi graph to see the entire network zoomed out
- spend more time optimize hyper-parameters
- build a robust recommender chatbot
- recommend topics that a friend might be interested in but have yet talked about in the past (discovering new passion!)