Ling 583
Statistical Methods in Text Analysis

Th 4:00am–6:40pm / SHW-243

Rob Malouf
619.594.7111
rmalouf@mail.sdsu.edu

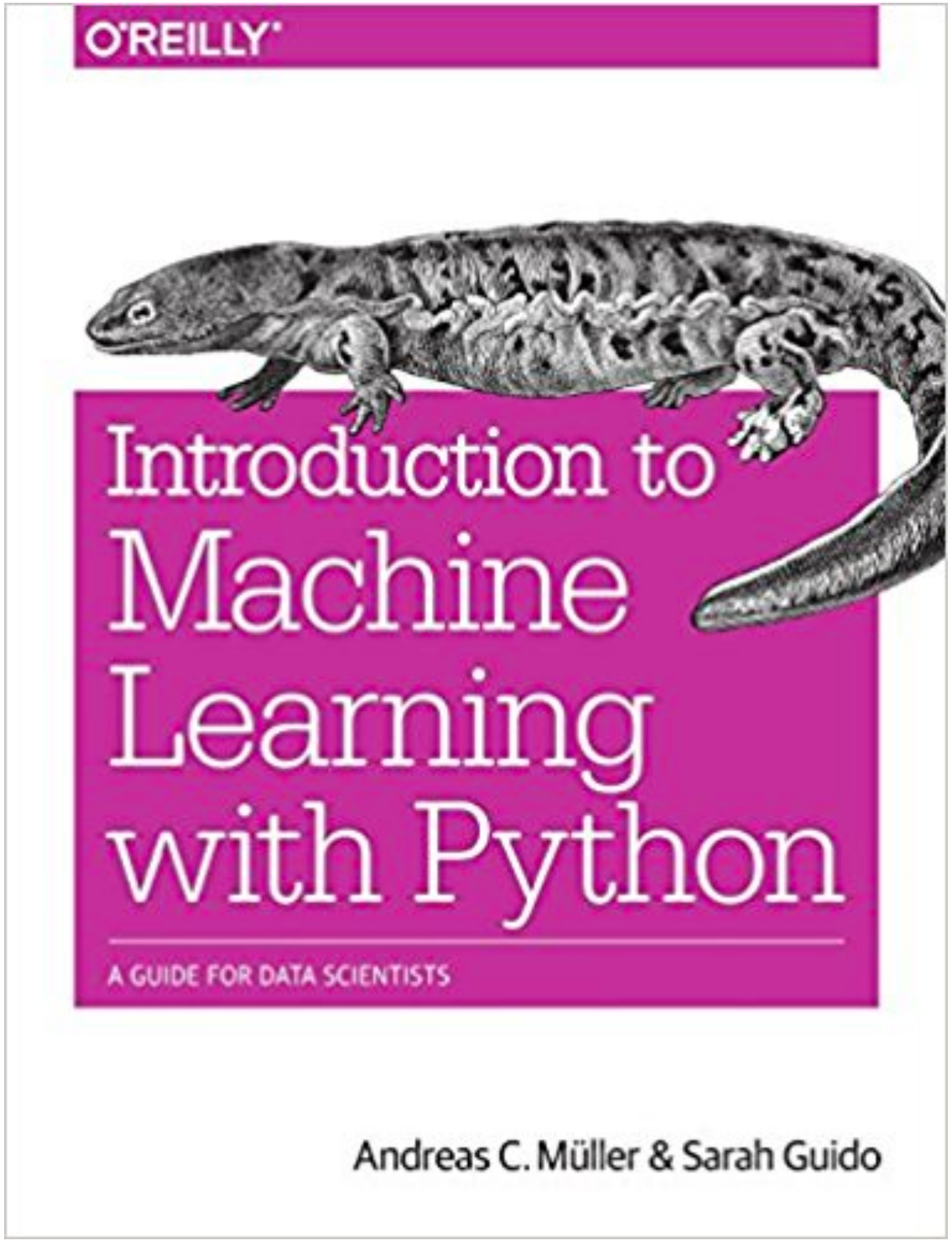
Office hours
Tu 8:00–9:00, Th 1:00–2:00, or by appt
SHW-244

Text Analytics

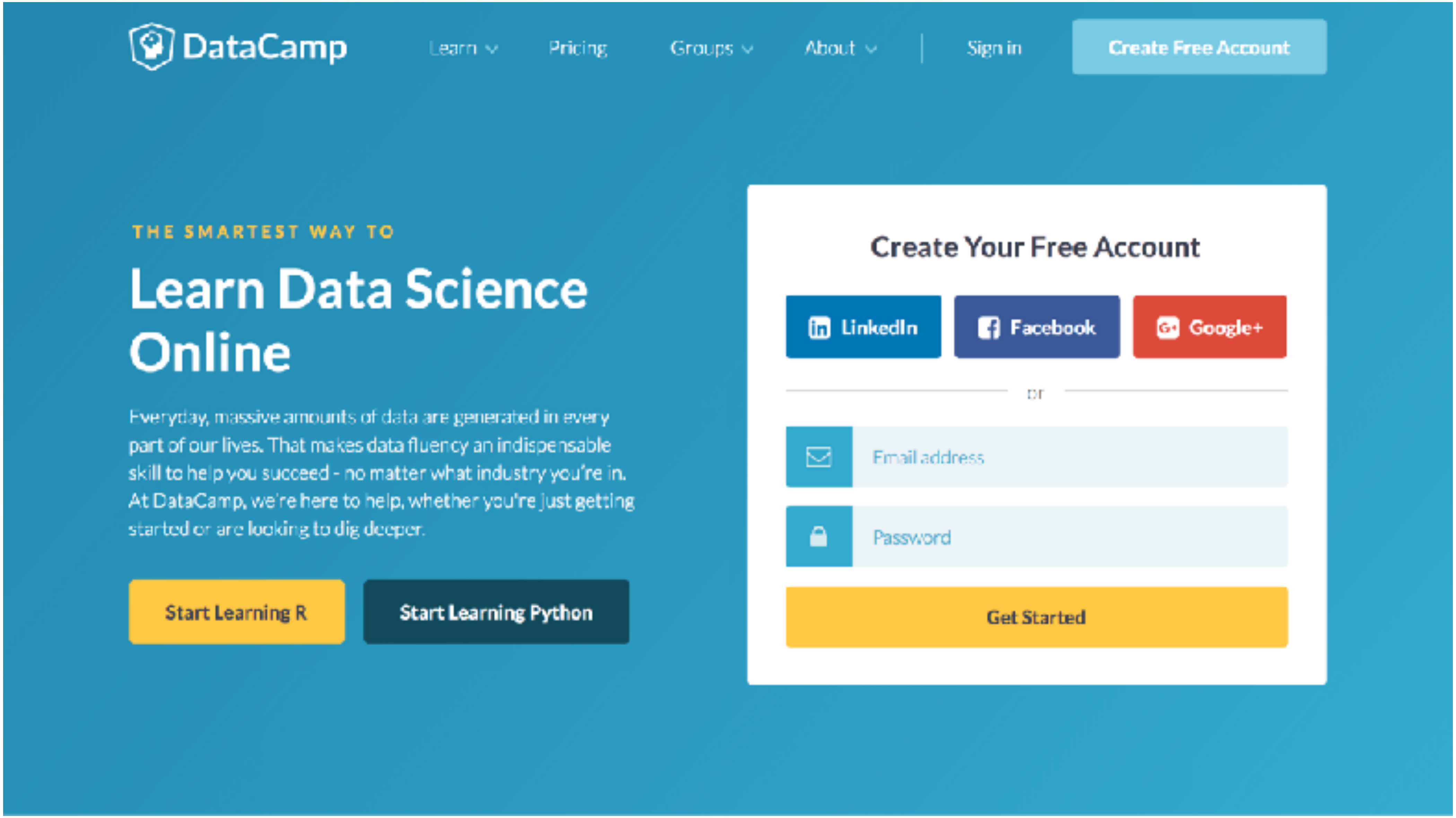
- According to industry lore, up to 80% of the information owned by large organizations is in the form of ‘unstructured’ text documents
- Some processes are inherently textual (litigation support), others generate text as a byproduct (help desk)
- How do we get at and make use of that information?

Text Analytics

- Outcomes
    - Apply techniques for collecting and preparing text data for computational analysis.
    - Describe a variety of text analysis steps, understand their interdependencies, and identify algorithms for each step.
    - Choose and apply appropriate text classification and clustering algorithms
    - Choose and apply appropriate semantic analysis techniques
    - Integrate knowledge and apply skills acquired over the sequence of text analysis courses to solve real world problems.


O'REILLY
Introduction to Machine Learning with Python
A GUIDE FOR DATA SCIENTISTS
Andreas C. Müller & Sarah Guido

Andreas C. Müller and Sarah Guido. 2017. *Introduction to Machine Learning with Python: A Guide for Data Scientists.* O'Reilly.



THE SMARTEST WAY TO

Learn Data Science Online

Everyday, massive amounts of data are generated in every part of our lives. That makes data fluency an indispensable skill to help you succeed - no matter what industry you're in. At DataCamp, we're here to help, whether you're just getting started or are looking to dig deeper.

Create Your Free Account

	Undergrads	Grads
Labs	10%	10%
Project 1	15%	15%
Project 2	15%	15%
Project 3	30%	20%
Project 4	30%	20%
Paper	—	20%

- Week 1 **Introduction**
  Background · Uses for text analytics · Using the Computational Linguistics Lab

- Week 2 **Text repositories**
  Linguistic corpora · Pubmed · Open data collections

- Week 3 **Web scraping**
  Spiders · HTML and XPATH · Data cleanup

  Project 1: Gathering text due

- Week 4 **Annotation**
  Planning annotation schemes · Manual annotation · Inter-annotator agreement

- Week 5–6 **Sequence models**
  Sequence annotation · Language models · Hidden Markov Models · Conditional random fields

- Week 7–8 **Classifiers**
  Text classifiers · Naive Bayes · Maximum Entropy models · Support Vector Machines · Sentiment analysis

  Project 2: Annotation due

- Week 9–10 **Topic models**
  Vector spaces · Latent Semantic Analysis · Latent Dirichlet Allocation · Embeddings

- Week 11 **Clustering**
  K-means · Hierarchical clustering · Visualization

  Project 3: Extracting topic clusters

- Week 12 **Dependency parsing**
  Dependency vs. constituents · Projective and non-projective algorithms

- Week 13–14 **Relation detection**
  Entities and relations · Taggers and classifiers for relations · 'Deep' learning and neural nets

- Week 15 **Project presentations**

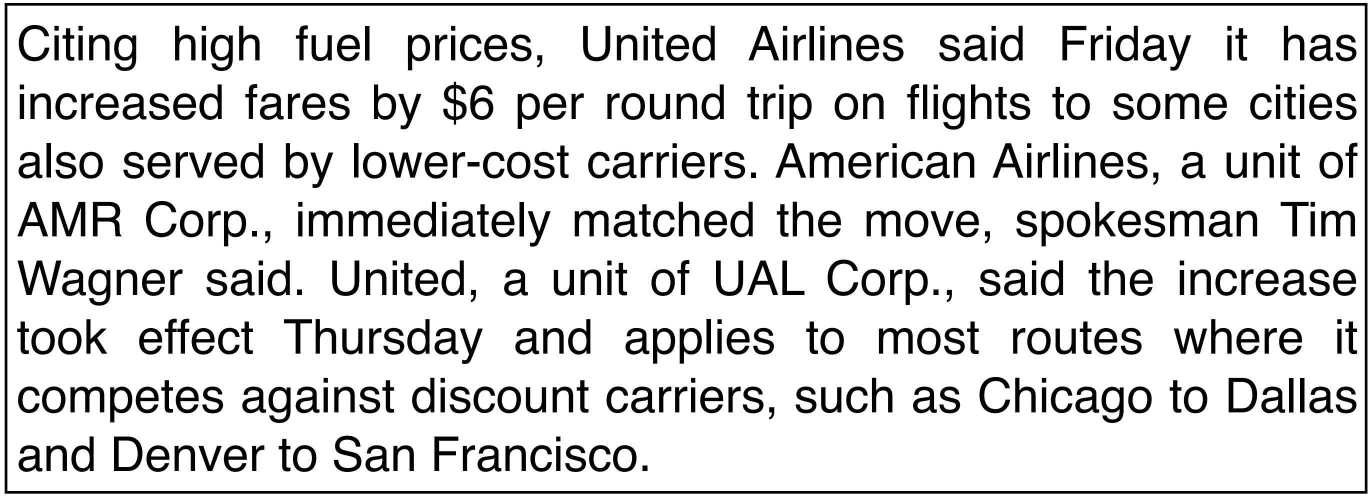
  Final projects and papers due

Text mining

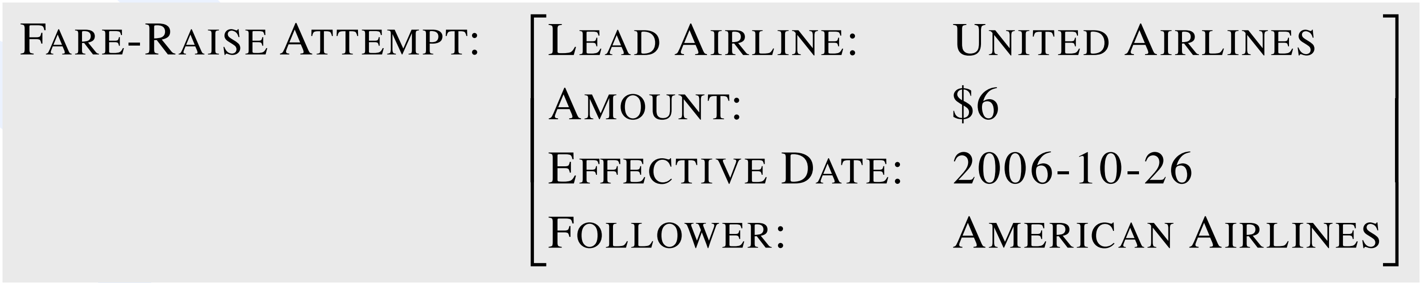
- Information retrieval: find documents which are relevant to a query
- Information extraction: given a relevant document, pull out a relevant segment
- Text mining: discover something new about the world which is not explicitly stated in any one text (or even known to the authors)

Information extraction

- A cascade of sequence labelers can be assembled into an **information extraction** system
- Start with a relevant text:



Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR Corp., immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL Corp., said the increase took effect Thursday and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Denver to San Francisco.



FARE-RAISE ATTEMPT:	LEAD AIRLINE:	UNITED AIRLINES
	AMOUNT:	$6
	EFFECTIVE DATE:	2006-10-26
	FOLLOWER:	AMERICAN AIRLINES

San Salvador, 19 Apr 89 (ACAN-EFE) — [TEXT] Salvadoran President-elect Alfredo Cristiani condemned the terrorist killing of Attorney General Roberto Garcia Alvarado and accused the Farabundo Marti National Liberation Front (FMLN) of the crime. . . . Garcia Alvarado, 56, was killed when a bomb placed by urban guerrillas on his vehicle exploded as it came to a halt at an intersection in downtown San Salvador. . . .Vice President-elect Francisco Merino said that when the attorney general’s car stopped at a light on a street in downtown San Salvador, an individual placed a bomb on the roof of the armored vehicle. . . . According to the police and Garcia Alvarado’s driver, who escaped unscathed, the attorney general was traveling with two bodyguards. One of them was injured.

Information extraction

- Database entry:

Incident: Date: - *19 Apr 89*
Incident: Location: *El Salvador: San Salvador* (CITY)
Incident: Type: *Bombing*
Perpetrator: Individual ID : *“urban guerrillas”*
Perpetrator: Organization ID: *“FMLN”*
Perpetrator: Organization: Suspected or Accused by Authorities: *“FMLN”*
Physical Target: Description: *“vehicle”*
Physical Target: Effect Some Damage: *“vehicle”*
Human Target: Name: *“Roberto Garcia Alvarado”*
Human Target: Description : *“attorney general”* : *“Roberto Garcia Alvarado”, “driver”, “bodyguards”*
Human Target: Effect: Death: *“Roberto Garcia Alvarado”,* No Injury: *“driver”*, Injury: *“bodyguards”*

Text mining

- Altavista vs. Google
- Given a query, Altavista returned documents ranked by *relevance*
- Relevance scores computed using vector space models
- Google ranks relevant pages by *prestige*
- Prestige is derived from patterns of cross-citation between documents, not from any one document

Text mining

- **Text mining** moves beyond information extraction, to find facts about the word which aren’t mentioned explicitly in the text
- Swanson (1986): “undiscovered public knowledge”
- A large field like medicine may have several contemporary, complementary, but disjoint literatures (e.g., clinical vs. preclinical)
- If we find that A causes B and B causes C, it may be worth investigating whether A causes C
- Vos (1991): “Drugs in search of diseases” (Rogaine, Viagra)

Literature-based discovery

- Raynaud’s syndrome is characterized by a loss of blood flow to the fingers and toes
- Raynaud’s is associated in the literature with platelet aggregation, blood viscosity, and vasoconstriction (among other things)
- Dietary fish oil is associated with platelet aggregation, blood viscosity, and vasoconstriction (among other things)
- Hypothesis: dietary fish oil as a potential treatment for Raynaud’s (confirmed by clinical studies in 1989)

Literature-based discovery

- Biomedical research is particularly open to mining for "undiscovered public knowledge"
- Comprehensive, public literature databases (MEDLINE, PubMed) with a standardized ontology of topic keywords (MeSH, UMLS)
- More recently used to explore gene co-expression patterns
- Other indices have also been successfully mined using titles and abstracts

Software

- Anaconda (get Python 3.6 version)

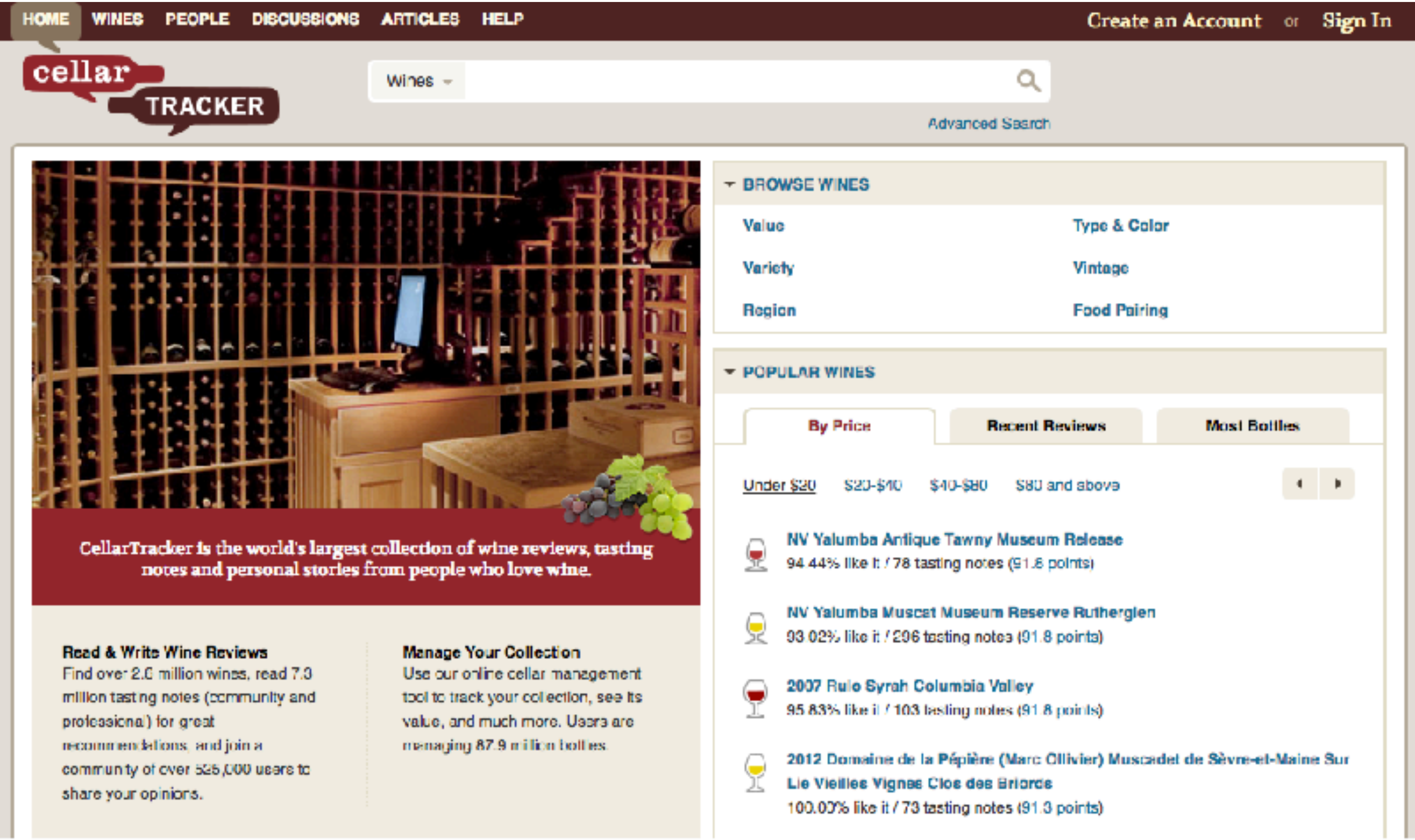
  https://www.anaconda.com/download

- (windows only) git version control

  https://git-scm.com/book/en/v2/Getting-Started-Installing-Git

- (windows only) PuTTY ssh client

  http://www.putty.org/



CellarTracker is the world's largest collection of wine reviews, tasting notes and personal stories from people who love wine.

Read & Write Wine Reviews
Find over 2.0 million wines, read 7.0 million tasting notes (community and professional) for great recommendations, and join a community of over 525,000 users to share your opinions.

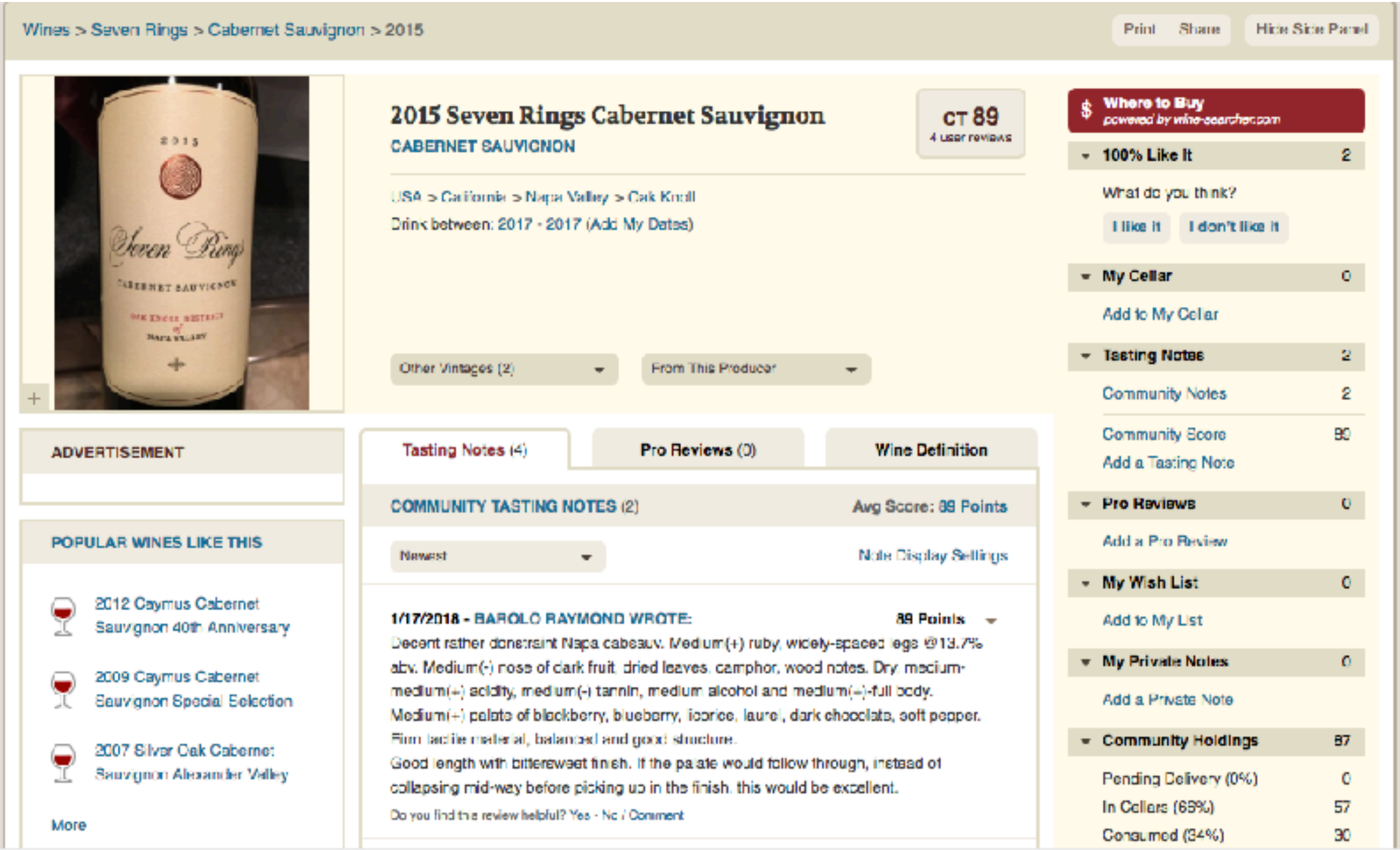
Manage Your Collection
Use our online cellar management tool to track your collection, see its value, and much more. Users are managing 87.9 million bottles.

NV Yalumba Antique Tawny Museum Release
94.44% like it / 78 tasting notes (91.6 points)

NV Yalumba Muscat Museum Reserve Rutherglen
93.02% like it / 206 tasting notes (91.8 points)

2007 Rulo Syrah Columbia Valley
95.83% like it / 103 tasting notes (91.8 points)

2012 Domaine de la Pépière (Marc Ollivier) Muscadet de Sèvre-et-Maine Sur Lie Vieilles Vignes Clos des Briords
100.00% like it / 73 tasting notes (91.3 points)



Wines > Seven Rings > Cabernet Sauvignon > 2015

2015 Seven Rings Cabernet Sauvignon

CABERNET SAUVIGNON

CT 89
4 user reviews

USA > California > Napa Valley > Oak Knoll

Drink between: 2017 - 2017 (Add My Dates)

ADVERTISEMENT

POPULAR WINES LIKE THIS

- 2012 Caymus Cabernet Sauvignon 40th Anniversary
- 2009 Caymus Cabernet Sauvignon Special Selection
- 2007 Silver Oak Cabernet Sauvignon Alexander Valley

More

COMMUNITY TASTING NOTES (2) Avg Score: 89 Points

1/17/2018 - BAROLO RAYMOND WROTE: **89 Points**

Decent rather constraint Napa cabsauv. Medium(+) ruby, widely-spaced legs @13.7% abv. Medium(-) nose of dark fruit, dried leaves, camphor, wood notes. Dry, medium-medium(+) acidity, medium(-) tannin, medium alcohol and medium(+)-full body. Medium(+) palate of blackberry, blueberry, licorice, laurel, dark chocolate, soft pepper. Firm tactile material, balanced and good structure.

Good length with bittersweet finish. If the palate would follow through, instead of collapsing mid-way before picking up in the finish, this would be excellent.

Do you find this review helpful? Yes - No / Comment

Where to Buy
powered by wine-searcher.com

100% Like It	**2**
What do you think?	
I like it / I don't like it	
My Cellar	**0**
Add to My Cellar	
Tasting Notes	**2**
Community Notes	2
Community Score	89
Add a Tasting Note	
Pro Reviews	**0**
Add a Pro Review	
My Wish List	**0**
Add to My List	
My Private Notes	**0**
Add a Private Note	
Community Holdings	**87**
Pending Delivery (0%)	0
In Cellars (65%)	57
Consumed (34%)	30

Mutual information

- What kind of evaluative language do wine reviewers use?
- Are there words that occur more often in positive reviews? Negative reviews?
- **Pointwise mutual information** (PMI) measures how far the joint probability of two events is from what would be expected if they were independent:

$$I(x, y) = \log_2 \frac{p(x, y)}{p(x)\, p(y)}$$

Mutual information

- Corpus linguistics, lexicography, sentiment analysis
- $I(w, BAD)$>0 means w occurs more often in negative reviews than you would expect by chance
- $I(w, BAD)$<0 means w occurs less often in negative reviews than you would expect by chance
- $I(w, BAD)$≈0 means w occurs about as often in negative reviews than you would expect by chance
- For word counts:

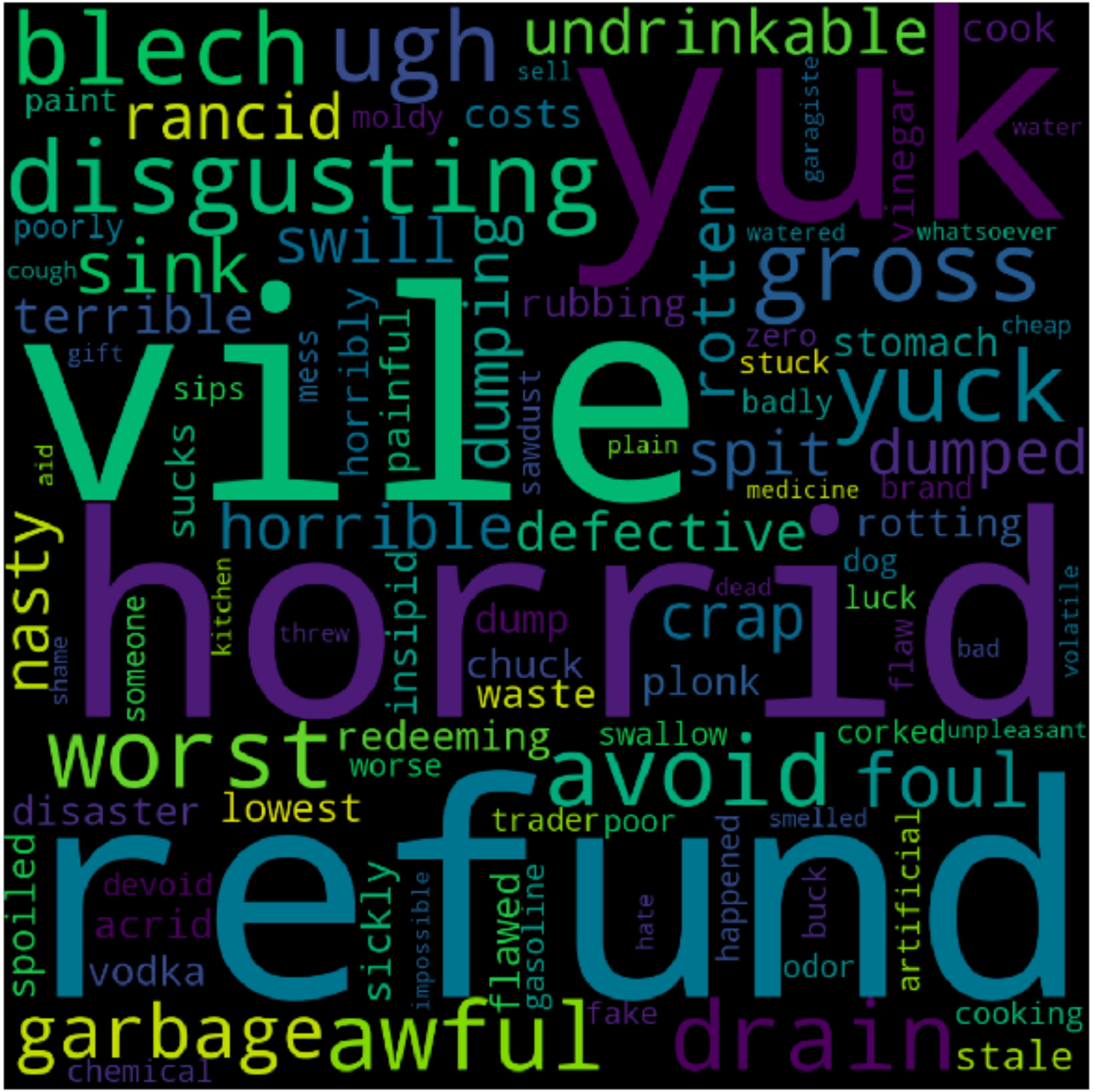
$$I(w, BAD) = \log_2 \frac{p(w, BAD)}{p(w) \times p(BAD)} = \log_2 \frac{f(w, BAD) \times N}{f(w) \times N(BAD)}$$

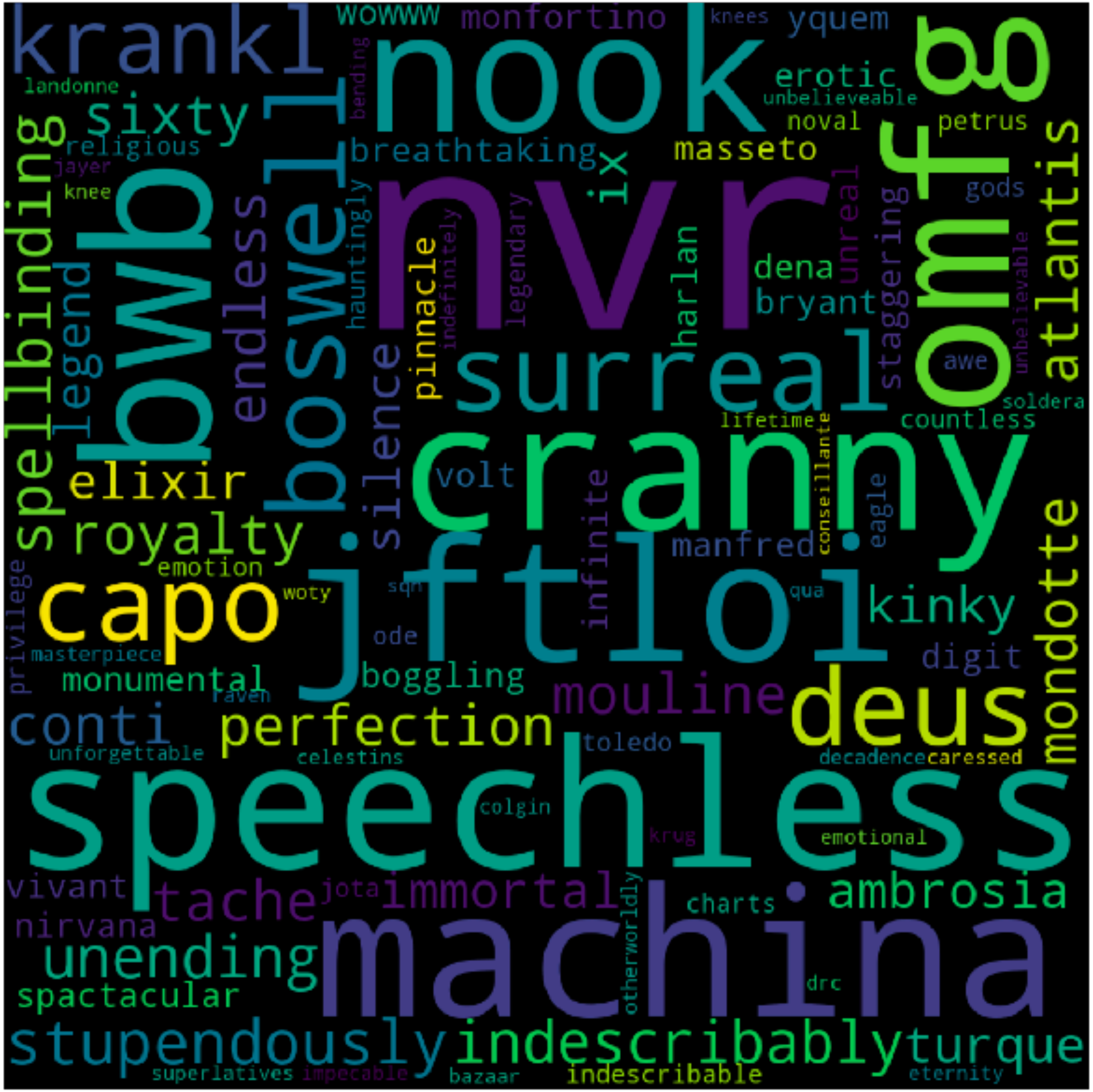
Mutual information

- Strategy:
  - Gather bad reviews
  - Normalize and tokenize text

    ‘Olive,␣horse␣sweat,␣dirty␣saddle,␣and␣smoke.␣This␣actually␣got␣quite␣a␣bit␣more␣spicy . . .’

    ‘olive’, ‘horse’, ‘sweat’, ‘dirty’, ‘saddle’, ‘and’, ‘smoke’, ’this’, ‘actually’, ‘got’, ‘quite’, ‘a’, ‘bit’, ‘more’, ‘spicy’, . . .
  - Count words
  - Compute PMI
  - Rank words


blech
ugh
undrinkable
cook
yuk
sell
paint
rancid
moldy
costs
garagiste
water
disgusting
vinegar
poorly
swill
dumping
rotten
watered
whatsoever
cough
sink
gross
terrible
rubbing
cheap
vile
zero
stomach
gift
stuck
mess
horribly
painful
sips
yuck
badly
sawdust
aid
plain
spit
dumped
sucks
medicine
brand
horrible
defective
rotting
nasty
dog
horrid
kitchen
dead
luck
volatile
threw
insipid
crap
shame
someone
dump
flaw
bad
chuck
plonk
waste
worst
redeeming
swallow
corked
unpleasant
avoid
foul
worse
disaster
lowest
trader
poor
smelled
spoiled
refund
devoid
happened
artificial
impossible
sickly
flawed
gasoline
hate
buck
acrid
vodka
odor
fake
cooking
garbage
awful
drain
stale
chemical



krankl
wowww
monfortino
knees
yquem
nook
g
landonne
bending
erotic
unbelieveable
sixty
noval
petrus
religious
breathtaking
masseto
jayer
ix
knee
omf
atlantis
gods
bwb
boswell
endless
nvr
spellbinding
legend
hauntingly
pinnacle
indefinitely
legendary
harlan
unreal
dena
staggering
bryant
surreal
awe
unbelievable
soldera
lifetime
countless
elixir
cranny
silence
volt
royalty
infinite
conseillante
eagle
mondotte
manfred
emotion
privilege
woty
sqn
qua
capo
jftloi
kinky
ode
masterpiece
digit
monumental
boggling
raven
mouline
deus
conti
perfection
toledo
unforgettable
celestins
decadence
caressed
colgin
speechless
krug
emotional
vivant
tache
jota
immortal
otherworldly
charts
ambrosia
nirvana
unending
machina
drc
spactacular
stupendously
indescribably
turque
superlatives
impecable
bazaar
indescribable
eternity

Lab

- Log in with your last name as username and password
- Right-click and open terminal
- yppasswd
  - password doesn’t echo
  - “A minimum of 10 characters. Your password must include at least 1 character from each of the following sets: uppercase, lowercase, numerical, and special character.”
- System tools > settings
  - Mouse speed
  - Internet accounts

Lab

- Go to github.com and create an account
- Fill in course registration form here:

  https://goo.gl/forms/lA525b6Y7X8eTUpr2

- When you get the invitation, sign up for datacamp.com
- For next week, do “Intro to Python for Data Science”

Lab

- Right-click and open terminal

```
source activate ling583
jupyter-notebook
```