Lukas Biewald

Scikit-Learn Workshop

Prerequisites for Following Along

- Install scikit-learn package (and pandas and numpy)
  - How I did it (OS X): `pip install scikit-learn`
  - More instructions: http://scikit-learn.org/stable/install.html
- Download some test data
  - http://crowdflower.com/data-for-everyone
    - "Judge Emotion About Brands & Products"
  - http://bit.ly/crowdflower-data

Me

- Not a Machine Learning Algorithms Expert
- Not a Python Expert
- ~~First~~ ~~Second~~ Third! time doing a workshop

Goals

1. Understand basics of Machine Learning
2. Build a real machine learning classifier
3. Give you a framework for understanding scikit-learn documentation

Follow Along

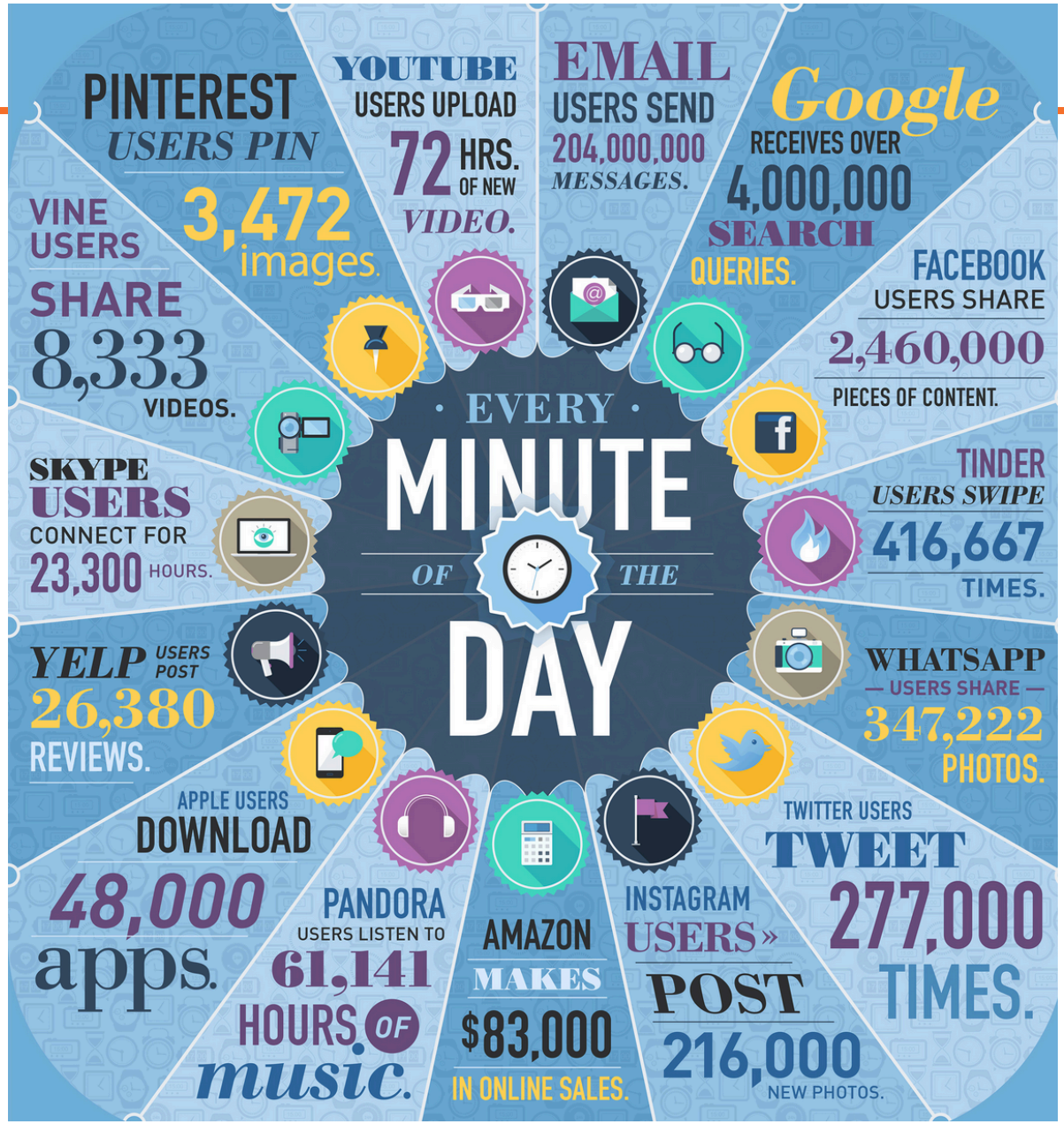
https://github.com/lukas/scikit-class

Gartner Hype Cycle

CrowdFlower

Data Science Ecosystem

CrowdFlower

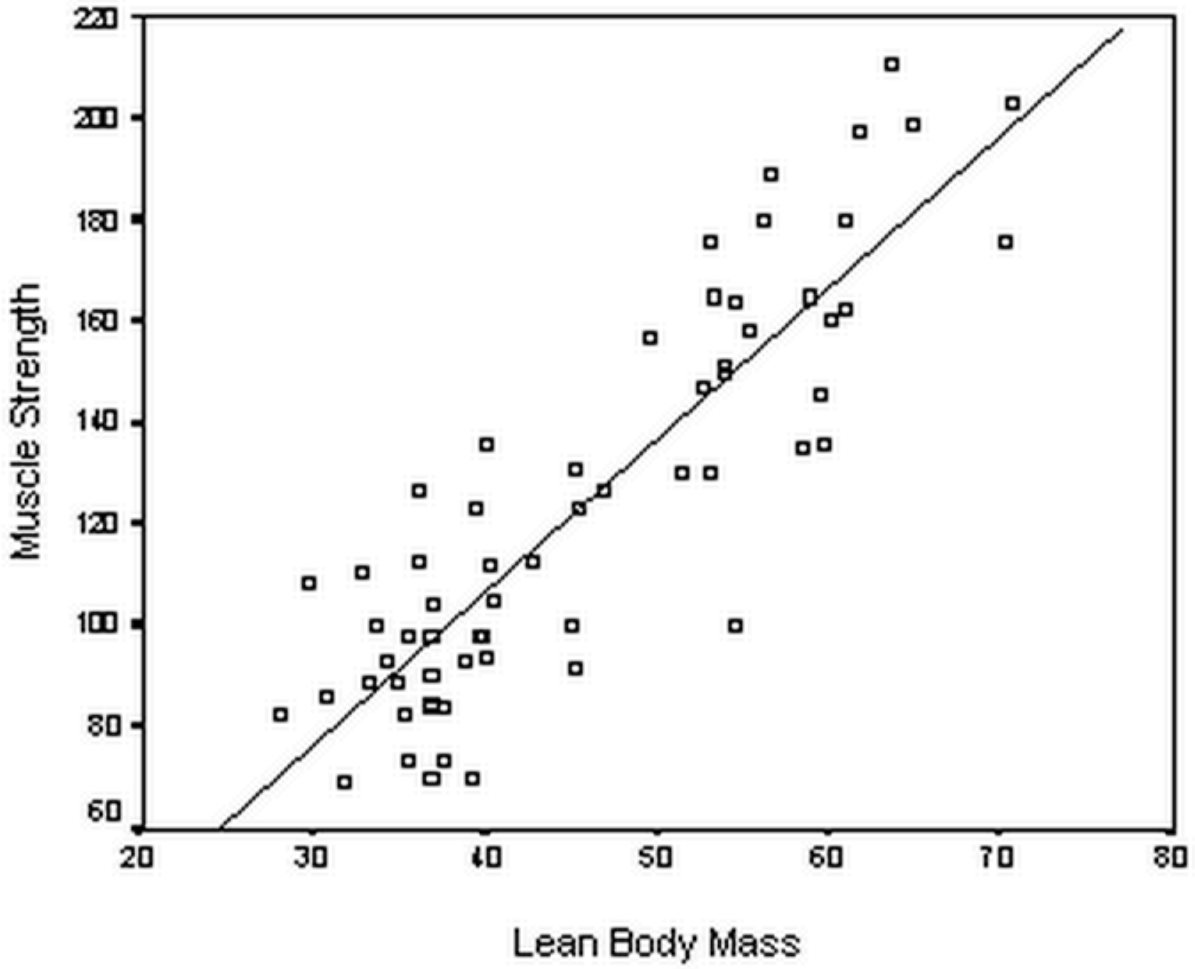


Infographic source: Domo

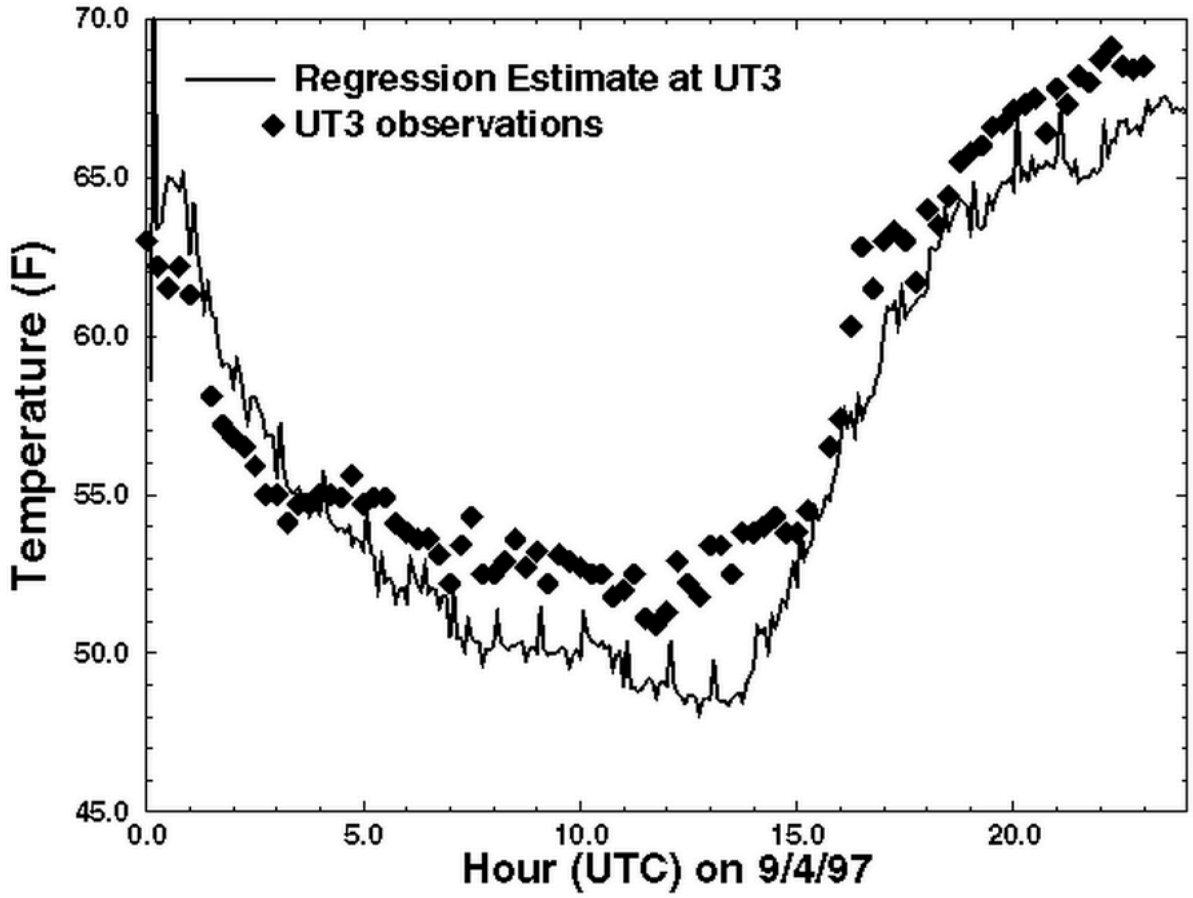
AI

Supervised Learning

Regression



Regression



Multivariable Regression

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Crime Rate	Residential L	Industry	By Charles River	No2									Median Value
2	0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
3	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
4	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
5	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
6	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2
7	0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
8	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9
9	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1
10	0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5
11	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9
12	0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15
13	0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	396.9	13.27	18.9
14	0.09378	12.5	7.87	0	0.524	5.889	39	5.4509	5	311	15.2	390.5	15.71	21.7
15	0.62976	0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21	396.9	8.26	20.4
16	0.63796	0	8.14	0	0.538	6.096	84.5	4.4619	4	307	21	380.02	10.26	18.2
17	0.62739	0	8.14	0	0.538	5.834	56.5	4.4986	4	307	21	395.62	8.47	19.9
18	1.05393	0	8.14	0	0.538	5.935	29.3	4.4986	4	307	21	386.85	6.58	23.1
19	0.7842	0	8.14	0	0.538	5.99	81.7	4.2579	4	307	21	386.75	14.67	17.5
20	0.80271	0	8.14	0	0.538	5.456	36.6	3.7965	4	307	21	288.99	11.69	20.2
21	0.7258	0	8.14	0	0.538	5.727	69.5	3.7965	4	307	21	390.95	11.28	18.2
22	1.25179	0	8.14	0	0.538	5.57	98.1	3.7979	4	307	21	376.57	21.02	13.6
23	0.85204	0	8.14	0	0.538	5.965	89.2	4.0123	4	307	21	392.53	13.83	19.6
24	1.23247	0	8.14	0	0.538	6.142	91.7	3.9769	4	307	21	396.9	18.72	15.2
25	0.98843	0	8.14	0	0.538	5.813	100	4.0952	4	307	21	394.54	19.88	14.5
26	0.75026	0	8.14	0	0.538	5.924	94.1	4.3996	4	307	21	394.33	16.3	15.6
27	0.84054	0	8.14	0	0.538	5.599	85.7	4.4546	4	307	21	303.42	16.51	13.9
28	0.67191	0	8.14	0	0.538	5.813	90.3	4.682	4	307	21	376.88	14.81	16.6
29	0.95577	0	8.14	0	0.538	6.047	88.8	4.4534	4	307	21	306.38	17.28	14.8
30	0.77299	0	8.14	0	0.538	6.495	94.4	4.4547	4	307	21	387.94	12.8	18.4
31	1.00245	0	8.14	0	0.538	6.674	87.3	4.239	4	307	21	380.23	11.98	21
32	1.13081	0	8.14	0	0.538	5.713	94.1	4.233	4	307	21	360.17	22.6	12.7
33	1.35472	0	8.14	0	0.538	6.072	100	4.175	4	307	21	376.73	13.04	14.5

Project for Today

- Automate

Judge Emotion About Brands & Products

Instructions ▲

In this job you will see tweets about several brands and products. Does the tweet include emotion directed at a brand, product, or user experience? If so, is it negative or positive?

Then, please select the brand, product, or user experience that applies. In most cases, there is a single best answer. If the tweet is about an iPad app, there's no need to also check the iPad box.

Brands considered are Apple and Google. **Products** considered are iPad, iPhone, and Android (phones or tablets). Use the "other" category for other products/services, like an emotion directed toward a Google Calendar. **User experiences** are mentions of using an application (App) on either iPad/iPhone or Android.

Note that in some cases links have been replaced with {link} and mentions have been replaced with @mention.

I just noticed DST is coming this weekend. How many iPhone users will be an hour late at SXSW come Sunday morning? #SXSW #iPhone

Is there an emotion directed at a brand or product?

- (•) Positive emotion
- ( ) Negative emotion
- ( ) I can't tell
- ( ) No emotion toward brand or product

Emotion in tweet is directed at

- [ ] Apple
- [ ] iPad
- [ ] iPhone
- [ ] iPad or iPhone App

What the Data Looks Like

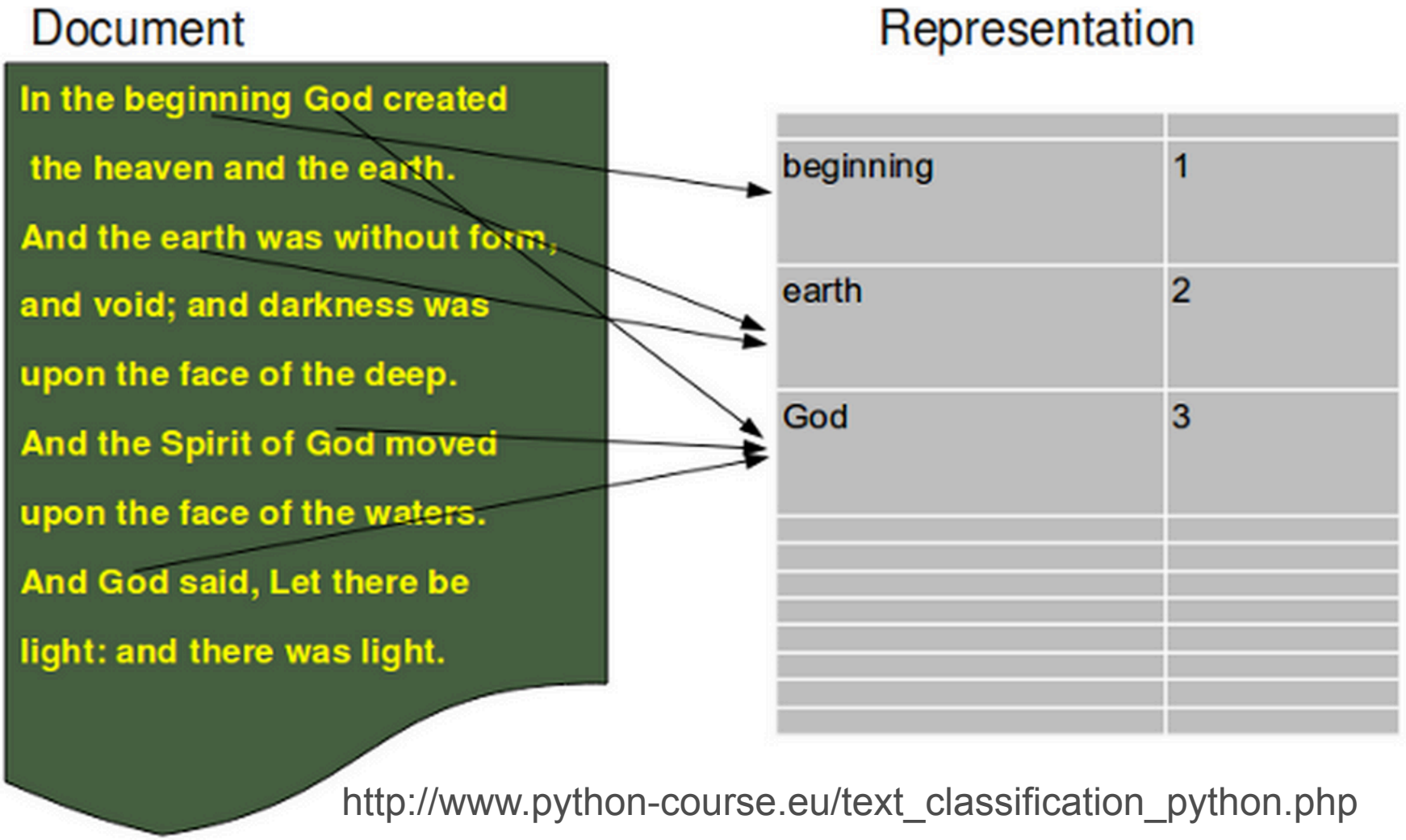
A	B	C	D	E	F	G
tweet_text	emotion_in_tweet_is_directed_at	is_there_an_emotion_directed_at_a_brand_or_product				
.@wesley83 I have a 3G iPhone. After 3 hrs tweeting at #RISE_Austin, it was dead! I need to	iPhone	Negative emotion				
@jessedee Know about @fludapp ? Awesome iPad/iPhone app that you'll likely appreciate f	iPad or iPhone App	Positive emotion				
@swonderlin Can not wait for #iPad 2 also. They should sale them down at #SXSW.	iPad	Positive emotion				
@sxsw I hope this year's festival isn't as crashy as this year's iPhone app. #sxsw	iPad or iPhone App	Negative emotion				
@sxtxstate great stuff on Fri #SXSW: Marissa Mayer (Google), Tim O'Reilly (tech books/conf	Google	Positive emotion				
@teachntech00 New iPad Apps For #SpeechTherapy And Communication Are Showcased At	The #SXSW Conference http://ht.ly/49n4M #iear #edchat	No emotion toward brand or product				
		No emotion toward brand or product				
#SXSW is just starting, #CTIA is around the corner and #googleio is only a hop skip and a jum	Android	Positive emotion				
Beautifully smart and simple idea RT @madebymany @thenextweb wrote about our #hollei	iPad or iPhone App	Positive emotion				
Counting down the days to #sxsw plus strong Canadian dollar means stock up on Apple gear	Apple	Positive emotion				
Excited to meet the @samsungmobileus at #sxsw so I can show them my Sprint Galaxy S stil	Android	Positive emotion				
Find & Start Impromptu Parties at #SXSW With @HurricaneParty http://bit.ly/gVLrIn I (	Android App	Positive emotion				
Foursquare ups the game, just in time for #SXSW http://j.mp/grN7pK) - Still prefer @Gowall	Android App	Positive emotion				
Gotta love this #SXSW Google Calendar featuring top parties/ show cases to check out. RT @	Other Google product or service	Positive emotion				
Great #sxsw ipad app from @madebymany: http://tinyurl.com/4nqv92l	iPad or iPhone App	Positive emotion				
haha, awesomely rad iPad app by @madebymany http://bit.ly/hTdFim #hollergram #sxsw	iPad or iPhone App	Positive emotion				
Holler Gram for iPad on the iTunes App Store - http://t.co/kfN3f5Q (via @marc_is_ken) #sxsw		No emotion toward brand or product				
I just noticed DST is coming this weekend. How many iPhone users will be an hour late at SX	iPhone	Negative emotion				
Just added my #SXSW flights to @planely. Matching people on planes/airports. Also downlo	iPad or iPhone App	Positive emotion				
Must have #SXSW app! RT @malbonster: Lovely review from Forbes for our SXSW iPad app	iPad or iPhone App	Positive emotion				
Need to buy an iPad2 while I'm in Austin at #sxsw. Not sure if I'll need to Q up at an Austin A	iPad	Positive emotion				
Oh. My. God. The #SXSW app for iPad is pure, unadulterated awesome. It's easier to browse	iPad or iPhone App	Positive emotion				
Okay, this is really it: yay new @Foursquare for #Android app!!!!11 kthxbai. #sxsw	Android App	Positive emotion				
Photo: Just installed the #SXSW iPhone app, which is really nice! http://tumblr.com/x6t1pi6	iPad or iPhone App	Positive emotion				
Really enjoying the changes in Gowalla 3.0 for Android! Looking forward to seeing what else	Android App	Positive emotion				
RT @LaurieShook: I'm looking forward to the #SMCDallas pre #SXSW party Wed., and hopin	iPad	Positive emotion				
RT haha, awesomely rad iPad app by @madebymany http://bit.ly/hTdFim #hollergram #sxsv	iPad or iPhone App	Positive emotion				
someone started an #austin @PartnerHub group in google groups, pre-#sxsw. great idea	Other Google product or service	Positive emotion				
The new #4sq3 looks like it is going to rock. Update for iPhone and Android should push ton	iPad or iPhone App	Positive emotion				
They were right, the @gowalla 3 app on #android is sweeeeet! Nice job by the team there. #	Android App	Positive emotion				
Very smart from @madebymany #hollergram iPad app for #sxsw! http://t.co/A3xvWc6 (ma	iPad or iPhone App	Positive emotion				
You must have this app for your iPad if you are going to #SXSW http://itunes.apple.com/us/	iPad or iPhone App	Positive emotion				
Attn: All #SXSW frineds, @mention Register for #GDGTLive and see Cobra iRadar for Android. {link}		No emotion toward brand or product				
Anyone at #sxsw want to sell their old iPad?		No emotion toward brand or product				
Anyone at #SXSW who bought the new iPad want to sell their older iPad to me?		No emotion toward brand or product				
At #sxsw, Oooh. RT @mention Google to Launch Major New Social Network Called Circles, Possibly Today {link}		No emotion toward brand or product				

Load the Data (load_data.py)

```
1  import pandas as pd
2  import numpy as np
3
4
5  df = pd.read_csv('tweets.csv')
6  target = df['is_there_an_emotion_directed_at_a_brand_or_product']
7  text = df['tweet_text']
8
9  print target[0:5]
10 print text[0:5]
```

How do we turn the text into numbers?

- Bag of words



http://www.python-course.eu/text_classification_python.php

How do we turn the text into numbers? (feature_extraction.py)

load_data.py | feature_extraction.py | feature_extraction_... | feature_extraction_... | tweets.csv

```
1  import pandas as pd
2  import numpy as np
3
4
5  df = pd.read_csv('tweets.csv')
6  target = df['is_there_an_emotion_directed_at_a_brand_or_product']
7  text = df['tweet_text']
8
9  text = text[pd.notnull(text)]
10 target = target[pd.notnull(text)]
11
12 from sklearn.feature_extraction.text import CountVectorizer
13 count_vect = CountVectorizer()
14 count_vect.fit(text)
15
16 print count_vect.vocabulary_.get(u'3g')
```

Handling weird input data (feature_extraction_2.py)

load_data.py | feature_extraction.py | feature_extraction_2.py | feature_extraction_... | tweets.csv

```
1  import pandas as pd
2  import numpy as np
3
4
5  df = pd.read_csv('tweets.csv')
6  target = df['is_there_an_emotion_directed_at_a_brand_or_product']
7  text = df['tweet_text']
8
9  fixed_text = text[pd.notnull(text)]
10 fixed_target = target[pd.notnull(text)]
11
12 from sklearn.feature_extraction.text import CountVectorizer
13 count_vect = CountVectorizer()
14 count_vect.fit(fixed_text)
15
16 print count_vect.vocabulary_.get(u'3g')
```

Run the feature extraction (feature_extraction_3.py)

load_data.py | feature_extraction.py | feature_extraction_... | feature_extraction_3.py | tweets.csv

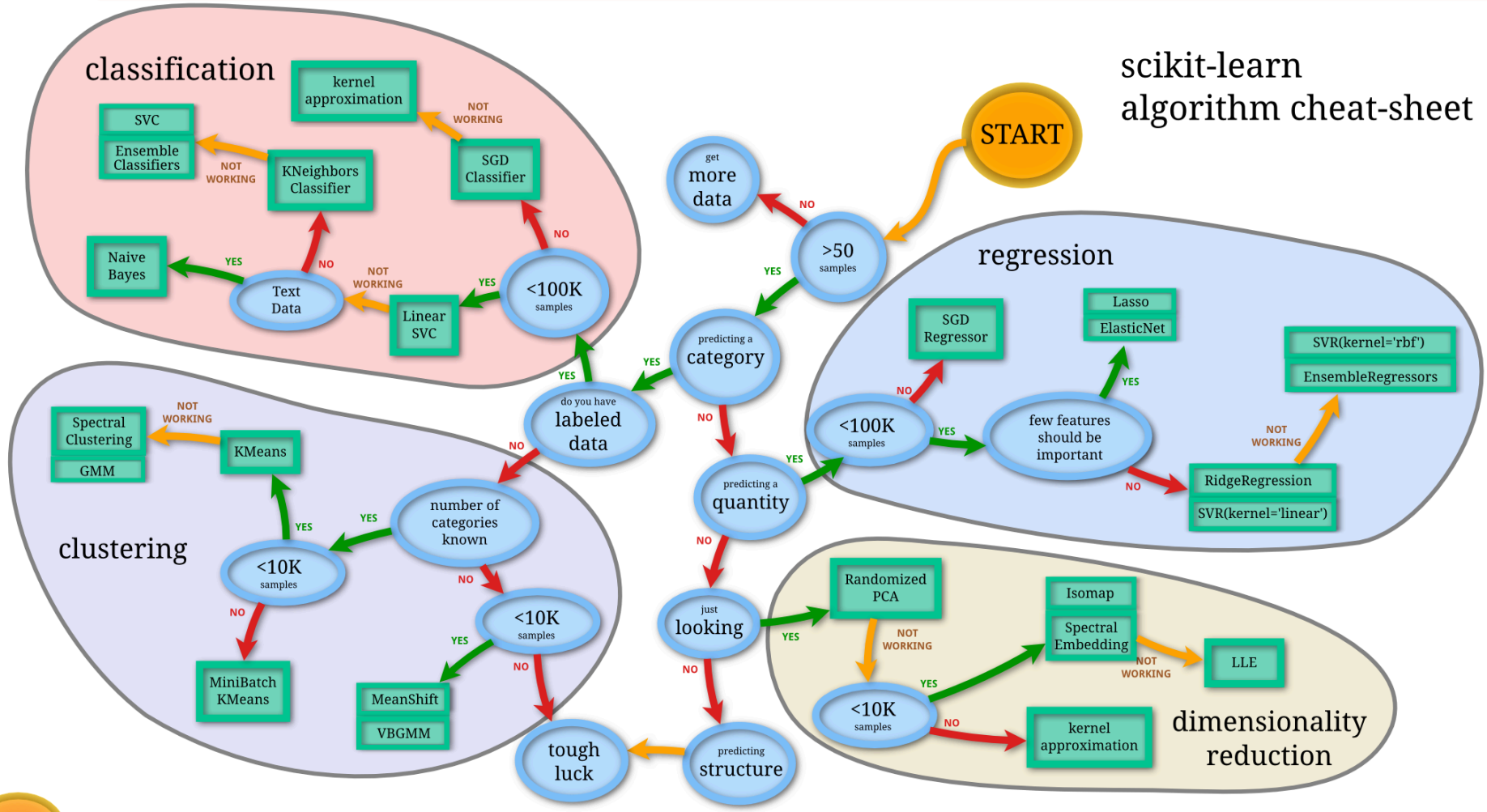
```
1  import pandas as pd
2  import numpy as np
3
4
5  df = pd.read_csv('tweets.csv')
6  target = df['is_there_an_emotion_directed_at_a_brand_or_product']
7  text = df['tweet_text']
8
9  fixed_text = text[pd.notnull(text)]
10 fixed_target = target[pd.notnull(text)]
11
12 from sklearn.feature_extraction.text import CountVectorizer
13 count_vect = CountVectorizer()
14 count_vect.fit(fixed_text)
15
16 counts = count_vect.transform(fixed_text)
17
18 print counts
```

Lots of important choices already!

- `class sklearn.feature_extraction.text.CountVectorizer(input=u'content', encoding=u'utf-8', decode_error=u'strict', strip_accents=None, lowercase=True, preprocessor=None, tokenizer=None, stop_words=None, token_pattern=u'(?u)\b\w\w+\b', ngram_range=(1, 1), analyzer=u'word', max_df=1.0, min_df=1, max_features=None, vocabulary=None, binary=False, dtype=<type 'numpy.int64'>)`

- Should we remove really rare words?
- Should we remove really common words?
- Should we remove “stop words”?
- Should we lower case all the words?
- What is a word?
  - For those at #SXSW: Apple sets up 5,000-square-foot temporary store at SXSW to sell new iPads, test potential traffic
  - If ur not at the #google #aclu 80's party....u should be! #sxsw
  - My iPhone battery at 100%. #winning at #SXSW

Choose an algorithm



Back

http://scikit-learn.org/stable/tutorial/machine_learning_map/

Run the Algorithm (classifier.py)

```
13  count_vect = CountVectorizer()
14  count_vect.fit(fixed_text)
15
16  counts = count_vect.transform(fixed_text)
17
18  from sklearn.naive_bayes import MultinomialNB
19  nb = MultinomialNB()
20  nb.fit(counts, fixed_target)
21
22  print nb.predict(count_vect.transform(["I love my iphone!!!"]))
```

Lots of scary choices!

- MultinomialNB vs GaussianNB vs BernoulliNB

Attributes:	**class_prior_** : array, shape (n_classes,) probability of each class. **class_count_** : array, shape (n_classes,) number of training samples observed in each class. **theta_** : array, shape (n_classes, n_features) mean of each feature per class **sigma_** : array, shape (n_classes, n_features) variance of each feature per class

- Do I want to weight one class more than the other?
- Do I want to monkey around with the algorithm?

How well is the algorithm working? (test_algorithm_1.py)

```
21
22  predictions = nb.predict(counts)
23  sum(predictions == fixed_target)
24
```

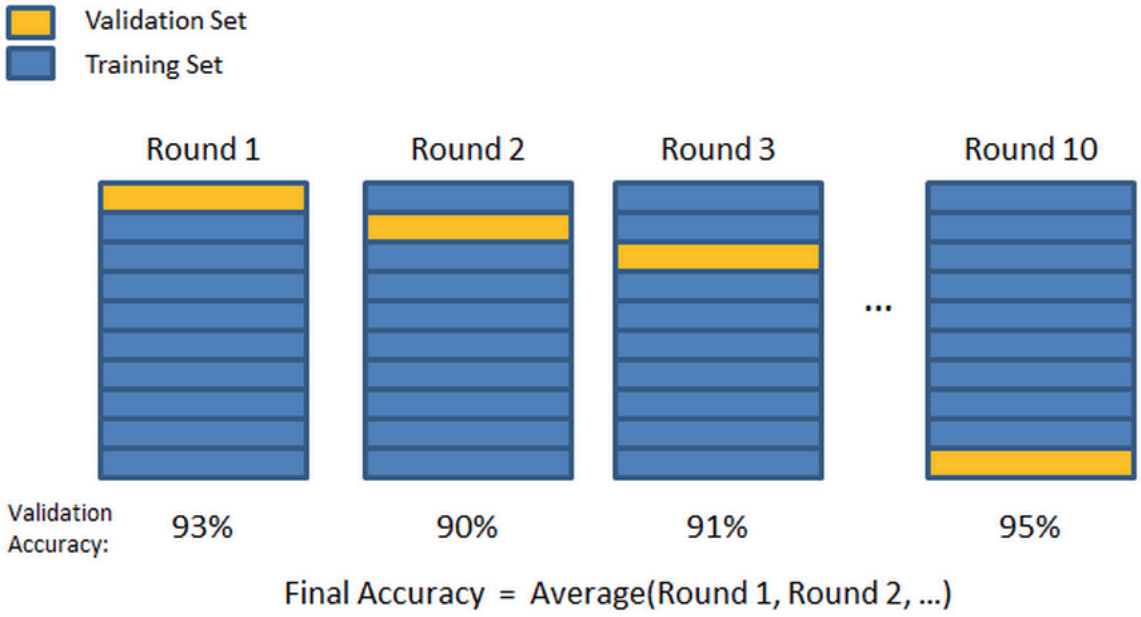
Test/Train Split (test_algorithm_2.py)

```
21  nb.fit(counts[0:6000], fixed_target[0:6000])
22
23  predictions = nb.predict(counts[6000:9092])
24  print sum(predictions == fixed_target[6000:9092])
```

Baselines (test_algorithm_dummy.py)

```
19  from sklearn.dummy import DummyClassifier
20
21  nb = DummyClassifier(strategy='most_frequent')
22
23  nb.fit(counts[0:6000], fixed_target[0:6000])
24
25  predictions = nb.predict(counts[6000:9092])
26  print sum(predictions == fixed_target[6000:9092])
27
```

Cross Validation



https://chrisjmccormick.wordpress.com/2013/07/31/k-fold-cross-validation-with-matlab-code/

Cross Validation (test_algorithm_cross_validation.py)

```
20
21  from sklearn import cross_validation
22
23  scores = cross_validation.cross_val_score(nb, counts, fixed_target, cv=10)
24  print scores
25  print scores.mean()
26
```

Cross Validation With Dummy (test_algorithm_cross_validation_dummy.py)

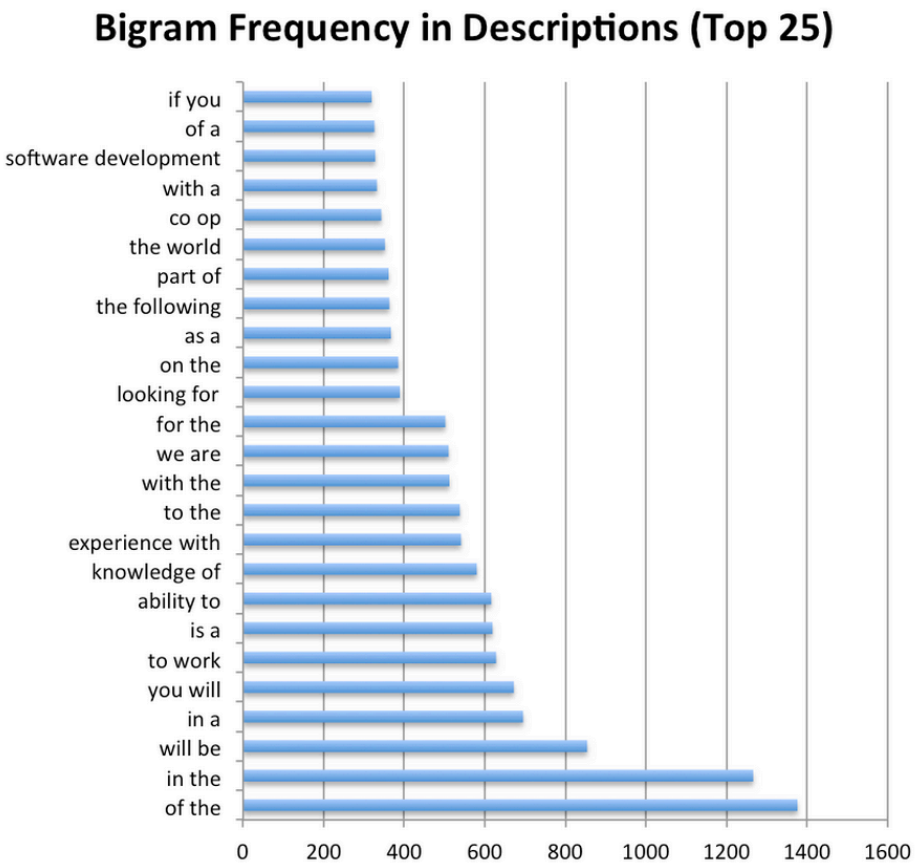
```
21  nb = DummyClassifier(strategy='most_frequent')
22
23  from sklearn import cross_validation
24
25  scores = cross_validation.cross_val_score(nb, counts, fixed_target, cv=10)
26  print scores
27  print scores.mean()
```

Pipelines (pipeline.py)

```
12  from sklearn.feature_extraction.text import CountVectorizer
13  from sklearn.naive_bayes import MultinomialNB
14  from sklearn.pipeline import Pipeline
15
16  p = Pipeline(steps=[('counts', CountVectorizer()),
17                      ('multinomialnb', MultinomialNB())])
18
19  p.fit(fixed_text, fixed_target)
20  print p.predict(["I love my iphone!"])
21
```

N-Grams

- Great vs. “Oh, great”



CrowdFlower

Bigrams (pipeline_bigrams.py)

```
p = Pipeline(steps=[('counts', CountVectorizer(ngram_range=(1, 2))),
                    ('multinomialnb', MultinomialNB())])

p.fit(fixed_text, fixed_target)
print p.named_steps['counts'].vocabulary_.get(u'garage sale')
print len(p.named_steps['counts'].vocabulary_)
```

Bigrams Accuracy (pipeline_bigrams_cross_validation.py)

```
p = Pipeline(steps=[('counts', CountVectorizer(ngram_range=(1, 2))),
                    ('multinomialnb', MultinomialNB())])

p.fit(fixed_text, fixed_target)

from sklearn import cross_validation

scores = cross_validation.cross_val_score(p, fixed_text, fixed_target, cv=
print scores
print scores.mean()
```

Feature Selection (feature_selection.py)

```
p = Pipeline(steps=[('counts', CountVectorizer(ngram_range=(1, 2))),
                    ('feature_selection', SelectKBest(chi2, k=10000)),
                    ('multinomialnb', MultinomialNB())])

p.fit(fixed_text, fixed_target)

from sklearn import cross_validation

scores = cross_validation.cross_val_score(p, fixed_text, fixed_target, cv
print scores
print scores.mean()
```

Grid Search (grid_search.py)

```
parameters = {
    'counts__max_df': (0.5, 0.75, 1.0),
    'counts__min_df': (1, 2, 3),
    'counts__ngram_range': ((1,1), (1,2)),
#     'feature_selection__k': (1000, 10000, 100000)
    }

grid_search = GridSearchCV(p, parameters, n_jobs=1, verbose=1, cv=10)

grid_search.fit(fixed_text, fixed_target)

print("Best score: %0.3f" % grid_search.best_score_)
print("Best parameters set:")
best_parameters = grid_search.best_estimator_.get_params()
for param_name in sorted(parameters.keys()):
    print("\t%s: %r" % (param_name, best_parameters[param_name]))
```

Make Your Own Features!

- Overwrite “fit” (optional) and “transform”
- Some Ideas
  - # of exclamation points
  - Emoji
  - Length of tweet
  - Language of tweet

Get More Data

crowdflower.com/data-for-everyone

Thanks!

Lukas Biewald
@L2K
LUKAS@CROWDFLOWER.COM