Machine Learning part 2

Crowdsourcing and Human Computation

Instructor: Chris Callison-Burch

Website: crowdsourcing-class.org

Building a classifier

- Where does labeled training data come from?
- What transformations should we do to it?
- Code walkthrough for upcoming HW
- Experiment design and evaluation for machine learning experiments

Labeled training data

Pretty awful - very soft and commercial. Confected.	★
Thin and completely uninspiring.	★
Good Syrah character, fruit-driven but not to the point of undrinkability. Pleasant. Scrapes	★★
Very classy, pure, blackberry and apple fruit. Demanding but ripe tannins, very succulent. Really good Dolcetto.	★★★
Fragrant, dry and long. More mineral and complex than the other Ogier wines. Really lovely and should be drunk on its own away from the Gentaz wines that tend to upstage it.	★★★★
An absolute star that could even benefit from another year or two. Tremendous weight, and concentrated minerality but all in balance. Fantastic. Top	★★★★★

Labeled training data

The Kalashnikov family of assault rifles. Alexander Vasilkov/Wikimedia Commons

Recent shootings involving children have rocked two American cities.

Michael Day, 13, died after being caught in the crossfire between two groups in the Edison Neighborhood of Kalamazoo, Mich., on Memorial Day. This wasn't even the first time Day had been a victim of gun violence: On April 6, he was **shot** in the back while leaving a party. He told police he was walking when he heard a gunshot and realized he had been hit.

Victor Manuel Garay, 15, has been **accused** of firing the shot that killed Day. Police had been called earlier in the day to break up the large brawl, but as soon as they left, the fighting continued. If **charged as an adult**, Garay could face life in prison without the possibility of parole.

Kalamazoo County Prosecutor Jeff Getting revealed his anguish at a **press conference** Thursday afternoon. "To talk about the death of a 13-year-old who was shot on one of our streets, allegedly by a 15-year-old, and to think about those as eighth graders and ninth graders…It has an effect on me. I think it has an effect on everyone; it should."

Meanwhile, on a playground at a school in Milwaukee last Wednesday, 10-year-old Sierra Guyton was caught in the crossfire of a **shootout**. She is in stable condition, but as of yesterday, she is not responsive. A **fund** has been created for her family, and a **rally** was held in her honor on Memorial Day.

The suspect is an 18-year-old with a long criminal record, who had been wounded by **gunfire** a week before he allegedly shot Sierra. A witness said 20 shots were fired in the direction of the playground. The suspect told police that he fired his gun until it was empty, then

Labeled training data

http://www.mlive.com/news/kalamazoo/index.ssf/2014/04/kalamazoo_teenager_13_shot_and.html	🔫
http://www.jsonline.com/news/crime/new-developments-in-playground-shooting-to-be-announced-at-430-pm-b99278118z1-260682381.html	🔫
http://www.mlive.com/news/kalamazoo/index.ssf/2014/05/fighting_led_up_to_fatal_shoot.html	🔫
http://www.mlive.com/news/kalamazoo/index.ssf/2014/05/michael_day_kalamazoo.html	🔫
http://www.mlive.com/news/kalamazoo/index.ssf/2014/05/15-year-old_charged_with_murde.html	🔫
http://www.jsonline.com/news/crime/girl-10-on-life-support-after-being-hit-in-playground-shootout-b99275748z1-260251491.html	🔫
http://fox6now.com/2014/05/29/fund-created-for-sierra-guyton-victim-of-shooting-near-playground/	🔫

Collecting data from the web

```
import urllib
import urllib2
from cookielib import CookieJar

def compile_gunreport_urls:
    for year in ["2014", "2013"]:
        for month in range(1, 13):
            for day in range(1, 32):
                url = "http://nocera.blogs.nytimes.com/%s/%s/%s/" %
                        (year, month, day)
                try:
                    cj = CookieJar()
                    opener = urllib2.build_opener
                            (urllib2.HTTPCookieProcessor(cj))
                    site = opener.open(url).read()
```

Collecting data from the web

```
import lxml.etree
import lxml.html
import re

def extract_external_links():
    for url in gunreport_urls:
# The NYTimes redirects you if you don't have cookies set.
cj = CookieJar()
opener = urllib2.build_opener(urllib2.HTTPCookieProcessor(cj))
site = opener.open(url).read()

doc = lxml.etree.HTML(site)

result = doc.xpath("//div[@class='entry-content']//p")
link = re.compile('href="(.*?)"')
for item in result:
      source = lxml.html.tostring(item)
      if link.search(source):
          print link.search(source).group(1)
```

Extracting web page text

Kalamazoo teenager, 13, shot and injured late Saturday while leaving party, police say

5 comments

By **Alex Mitchell | amitche5@mlive.com**
on April 06, 2014 at 9:38 AM, updated April 06, 2014 at 1:01 PM

KALAMAZOO, MI — A 13-year-old Kalamazoo juvenile was shot in the back and injured late Saturday while leaving a party on the south side of Kalamazoo, police say.

Gazette File

Officers responded around 11:40 p.m. to the area of Lake Street and Maywood Avenue to a report of a subject that had been shot and discovered a teenage male suffering from a gunshot wound in his back, according to a press release issued by the **Kalamazoo Department of Public Safety**.

The victim, whose name has not been released, told police he had just left a party in the 600 block of Carr Street. The teen said he was walking near Lake and Maywood when he heard a

Extracting web page text

AlchemyAPI™

Resources Products Developers Company Blog Contact Sales

Products » Demo » AlchemyLanguage

Try the AlchemyLanguage API

Load a URL or paste some text below.

Free API Key
Get Started with AlchemyAPI Today!

Load a sample URL	Load a text sample	Enter your own URL	Enter your own text
click for a new web page	*click for a new text passage*	*click to clear URL field*	*click to clear text field*

http://www.mlive.com/news/kalamazoo/index.ssf/2014/04/kalamazoo_teenager_13_shot_and.html

TITLE: Kalamazoo teenager, 13, shot and injured late Saturday while leaving party, police say
LANGUAGE: English

KALAMAZOO, MI — A 13-year-old Kalamazoo juvenile was shot in the back and injured late Saturday while leaving a party on the south side of Kalamazoo, police say. Officers responded around 11:40 p.m. to the area of Lake Street and Maywood Avenue to a report of a subject that had been shot and discovered a teenage male suffering from a gunshot wound in his back, according to a press release issued by the Kalamazoo Department of Public Safety. The victim, whose name has not been released, told police he had just left a party in the 600 block of Carr Street. The teen said he was walking near Lake and Maywood when he heard a gunshot and realized he had been struck in the back, police said. While waiting for an ambulance to arrive, the victim was transported to Bronson Methodist Hospital for treatment of non-life threatening injuries by an acquaintance, officers said. He is currently in stable condition. Police said no arrests related to this

Extracting web page text

Kalamazoo teenager, 13, shot and injured late Saturday while leaving party, police say

KALAMAZOO, MI — A 13-year-old Kalamazoo juvenile was shot in the back and injured late Saturday while leaving a party on the south side of Kalamazoo, police say. Officers responded around 11:40 p.m. to the area of Lake Street and Maywood Avenue to a report of a subject that had been shot and discovered a teenage male suffering from a gunshot wound in his back, according to a press release issued by the Kalamazoo Department of Public Safety. The victim, whose name has not been released, told police he had just left a party in the 600 block of Carr Street. The teen said he was walking near Lake and Maywood when he heard a gunshot and realized he had been struck in the back, police said. While waiting for an ambulance to arrive, the victim was transported to Bronson Methodist Hospital for treatment of non-life threatening injuries by an acquaintance, officers said. He is currently in stable condition. Police said no arrests related to this incident have been made at this time. This was the second reported shooting in Kalamazoo Saturday. A 21-year-old Kalamazoo man was also shot in the back while leaving a party in the Northside neighborhood around 2:30 a.m. and was treated at Bronson Hospital for non-life threatening injuries, police said.

Representing data with Features

- In machine learning, we represent the training data as a vector of labels (**y**) and a matrix of training items (**X**)
  - Each training item is itself represented as a vector
  - The vector specifies what **features** that item has

Representing data with Features

y	X?
★	Pretty awful - very soft and commercial. Confected.
★★★★★	An absolute star that could even benefit from another year or two. Tremendous weight, and concentrated minerality but all in balance. Fantastic. Top
★★★	Very classy, pure, blackberry and apple fruit. Demanding but ripe tannins, very succulent. Really good Dolcetto.
★★	Good Syrah character, fruit-driven but not to the point of undrinkability. Pleasant. Scrapes
★	Thin and completely uninspiring.
★★★★	Fragrant, dry and long. More mineral and complex than the other Ogier wines. Really lovely and should be drunk on its

Representing data with Features

raw input

Pretty awful - very **soft** and commercial. Confected.

An absolute **star** that could even **benefit** from another year or two. **Tremendous** weight, and concentrated minerality but all in balance. **Fantastic**. **Top**

Very **classy**, **pure**, blackberry and apple fruit. **Demanding** but ripe tannins, very succulent. Really **good** Dolcetto.

Good Syrah character, fruit-driven but not to the point of undrinkability. **Pleasant**.

Thin and completely **uninspiring**.

subjectivity lexicon

Word	Polarity	Strength
abandoned	negative	weak
abandonmen	negative	weak
abandon	negative	weak
abase	negative	strong
abasement	negative	strong
abash	negative	strong
abate	negative	weak
abdicate	negative	weak
aberration	negative	strong
aberration	negative	strong
...		...
zest	positive	strong

Representing data with Features

raw input

Pretty awful - very **soft** and commercial. Confected.
An absolute **star** that could even **benefit** from another year or two. **Tremendous** weight, and concentrated minerality but all in balance. **Fantastic**. **Top**
Very **classy**, **pure**, blackberry and apple fruit. **Demanding** but ripe tannins, very succulent. Really **good** Dolcetto.
Good Syrah character, fruit-driven but not to the point of undrinkability. **Pleasant**.
Thin and completely **uninspiring**.

feature matrix X

Strong Neg	Neg	Pos	Strong Pos
1	1		1
		2	3
	1	3	
		2	
1			

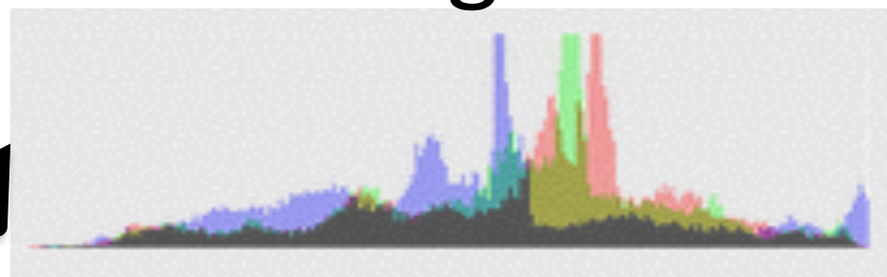
Other types of data

"cockatoo"

"red tailed hawk"

Representing data with Features

Color histogram features



SIFT keypoint features

→ into **visual feature vectors**: **<3, 0, 0, 0, 7, 0, … 0, 0, 19, 0, 0, 38, …>**

Training classifiers in Python

Experimental design in machine learning

- Splitting data into training / test sets
- Baselines
- Evaluation

Training/test spit

- Typically we have a fixed set of labeled data that we run experiments on
- In our experiments we typically split the data into a training set, and a disjoint test set
- Why?

It is generalization that counts

- The fundamental goal of machine learning is to generalize beyond the examples in the training set
- No matter how much data we have, at test time we are unlikely to see exactly the same items

The problem of overfitting

- Sometimes our classifier *overfits* the data
- It encodes random quirks of the data instead of learning good generalizations
- Symptom: your learner creates a classifier that is 100% accurate on the training data but only 50% accurate on test data

n-fold cross validation

- Splitting the data reduces the amount of available data for training
- Mitigated through *cross-validation*: randomly dividing your training data into 10 pieces, train on 9 test on 1, average results

Precision and Recall

	Actual Class	
Predicted class	True positive: Correct result	False positive: Unexpected result
	False negative: Missing result	True negative: Correct absence of result

$$\text{Precision} = \frac{tp}{tp + fp} \qquad \text{Recall} = \frac{tp}{tp + fn}$$

Precision and Recall

	Actual Class	
Predicted class	True positive: Correct result	False positive: Unexpected result
	False negative: Missing result	True negative: Correct absence of result

$$\text{Precision} = \frac{tp}{tp + fp}$$

Precision and Recall

	Actual Class	
Predicted class	True positive: Correct result	False positive: Unexpected result
	False negative: Missing result	True negative: Correct absence of result

$$\text{Recall} = \frac{tp}{tp + fn}$$

Accuracy

	Actual Class	
Predicted class	True positive: Correct result	False positive: Unexpected result
	False negative: Missing result	True negative: Correct absence of result

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

Evaluating a system

- Say your system gets 93% accuracy, is that good?

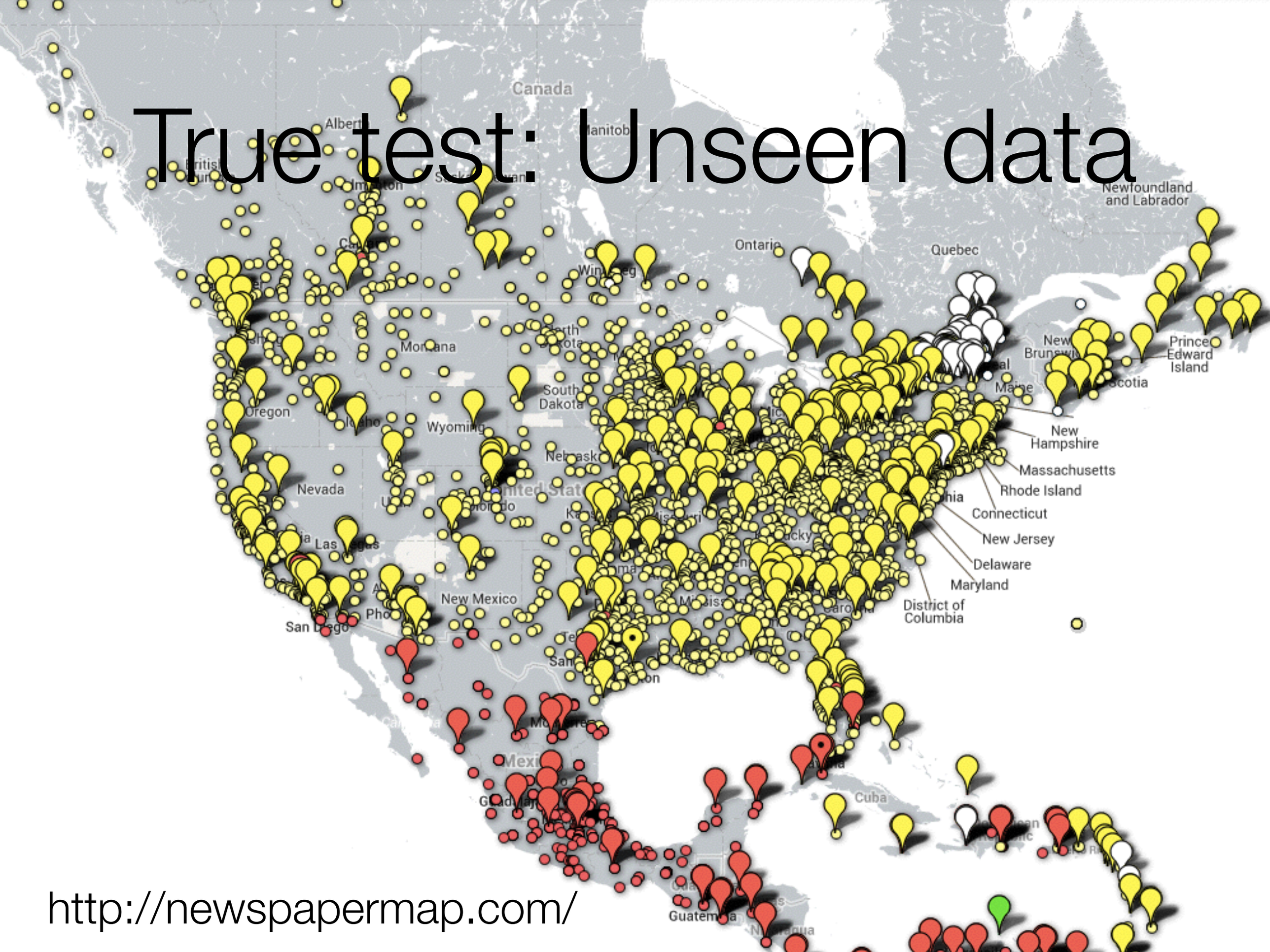
Baseline

- Our training data is imbalanced:
  - 62694 negative examples
  - 8973 positive examples
  - A system that always guessed "not a gun related article" would get 87% accuracy
  - This is the "majority class baseline"
  - The rule based system that guess + iff "shooting" occurs in the article and - otherwise gets 93%

Our data may be too easy

- Jennifer Mascia described how she wrote the Gun Report for the NYTimes in an NPR interview
- JENNIFER MASCIA: Well, I would google "**shooting**," "**man shot**," "**woman shot**," "**child shot**," "**teen shot**" and "**accidentally shot**". You know, this was all day one coverage of shootings, so a lot of times the details aren't flushed out. If there was no name and scant details, I had to skip over those. So each day, there'd be about 35 to 40 shootings that I would present.

True test: Unseen data



http://newspapermap.com/

HW4 is now available.

Due: 1 week from now