THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Introduction to Text Mining

May 22, 2018

Heejun Kim
Course materials courtesy of Dr. Jaime Arguello

Outline for Introduction to Text Mining

- What is Text Mining?
- Applications of Text Mining
- A Sample Application: Topic Categorization and Visualization
- A Hands-on Practice
- Predictive Analysis of Text: The Big Picture
- Exploratory Analysis of Text: The Big Picture

What is Text Mining?

Definition of Text Mining

Model

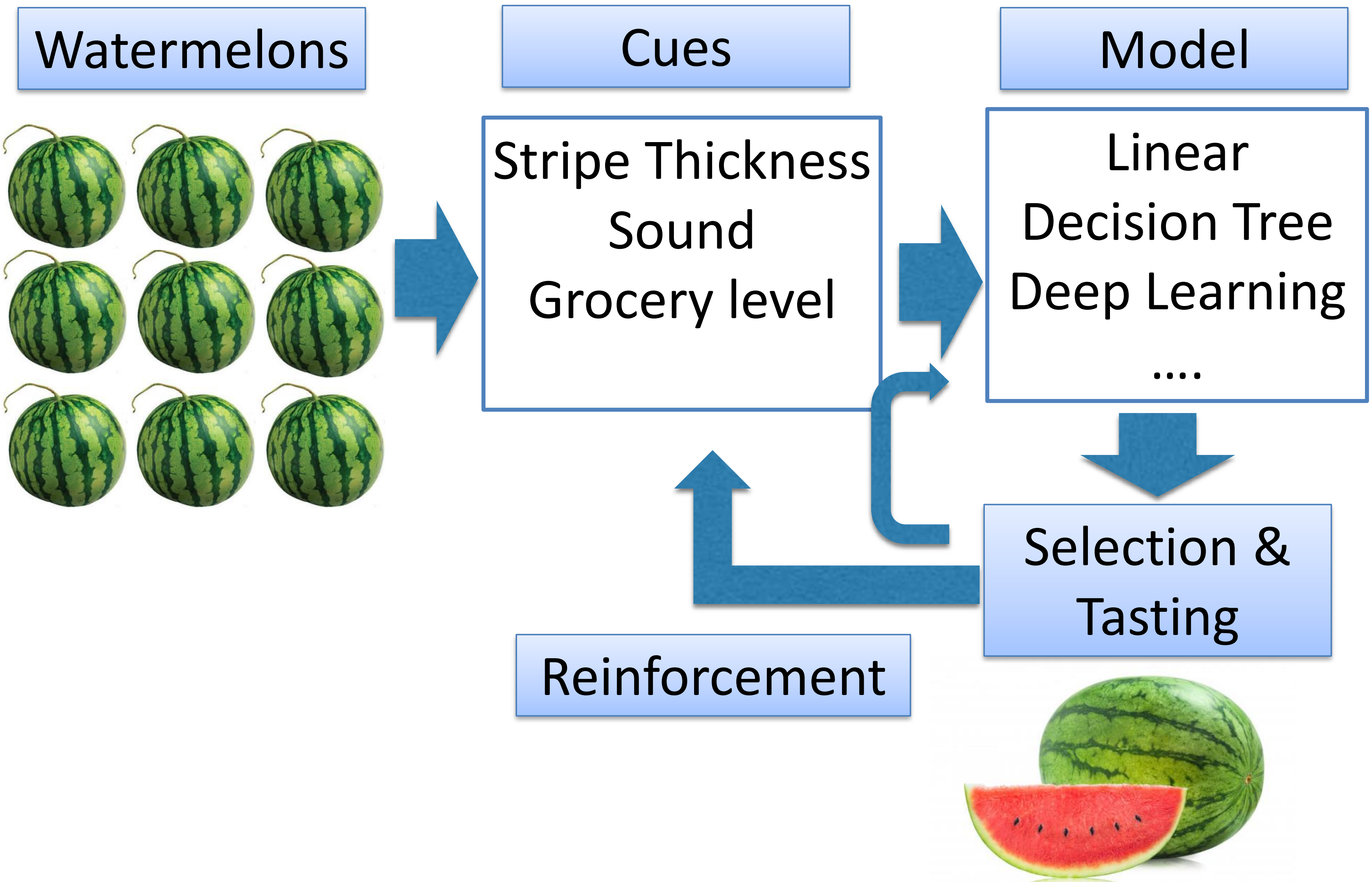
Knowledge

- The science and practice of building and evaluating computer programs that automatically discover useful knowledge or insight in collections of natural language text

Text

What is Data Analytics? (An Analogy)

Text Mining Process (An Analogy)



What is Text Mining?

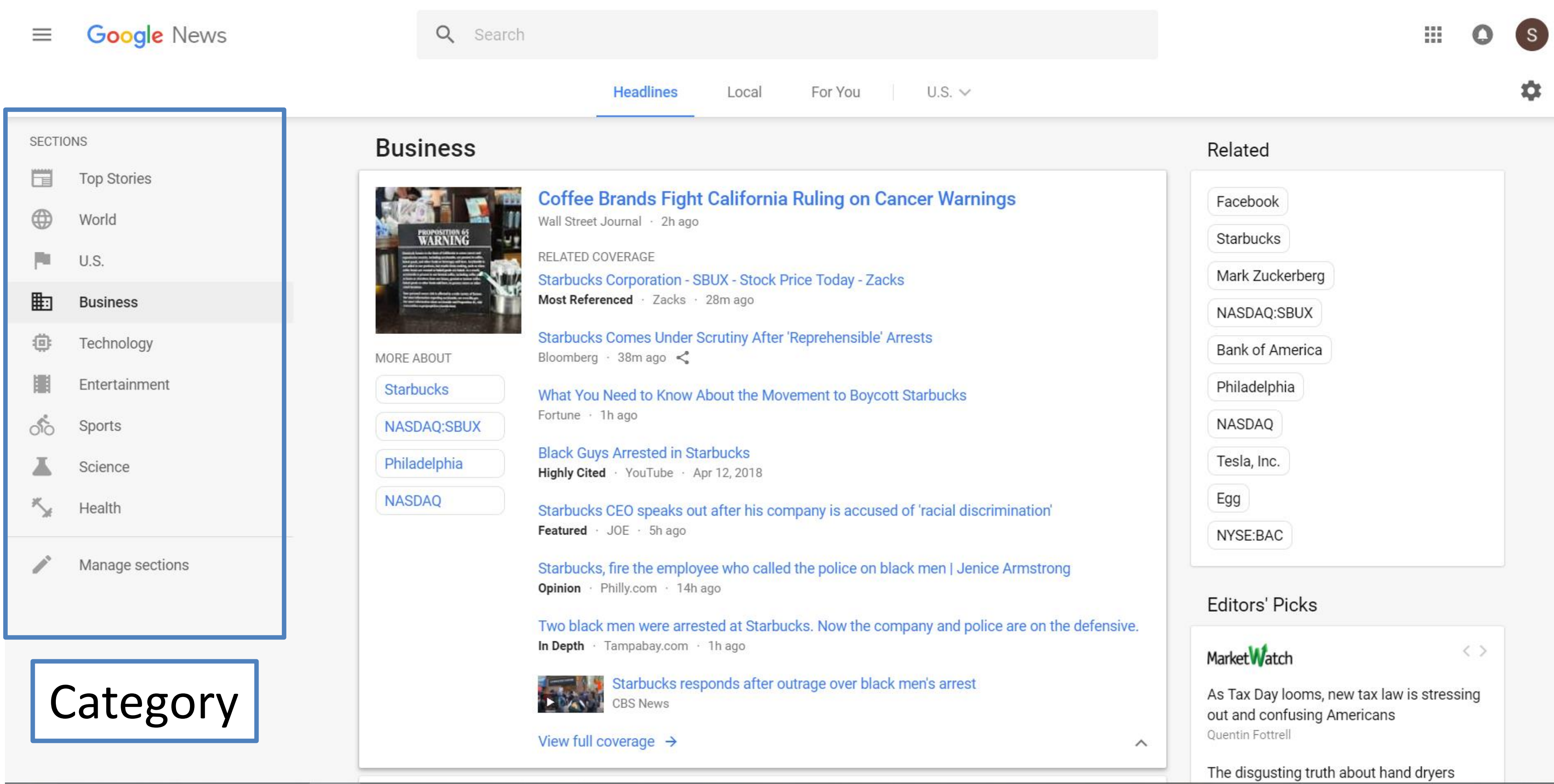
Link

Applications of Text Mining

Applications of Text Mining

- Topic Categorization
- Opinion Mining
- Sentiment Analysis
- Text-Driven Forecasting
- …

Topic Categorization



Topic Categorization

- **Topic Categorization**: automatically assigning documents to a set of pre-defined topical categories

Opinion Mining (Product Reviews)

- This is a great phone with an amazing camera. The facial recognition really blew me away. And the case is thin enough that it can charge on a charging mat (not provided) if that's your fancy.

positive

- DO NOT BUY IT! There is manufacture defect & the seller advise you to deal with the Apple.

negative

- I am a die hard Apple person. All my desktop computers at home are Apple, my other 4 family members all have iPhones and we have laptops that are all Apple. Not really sure why I decided to try out the Samsung Galaxy S9+ but happy that I did so far.

positive

Opinion Mining

- **Opinion Mining**: automatically detecting whether a span of opinionated text expresses a positive or negative opinion about the item being judged

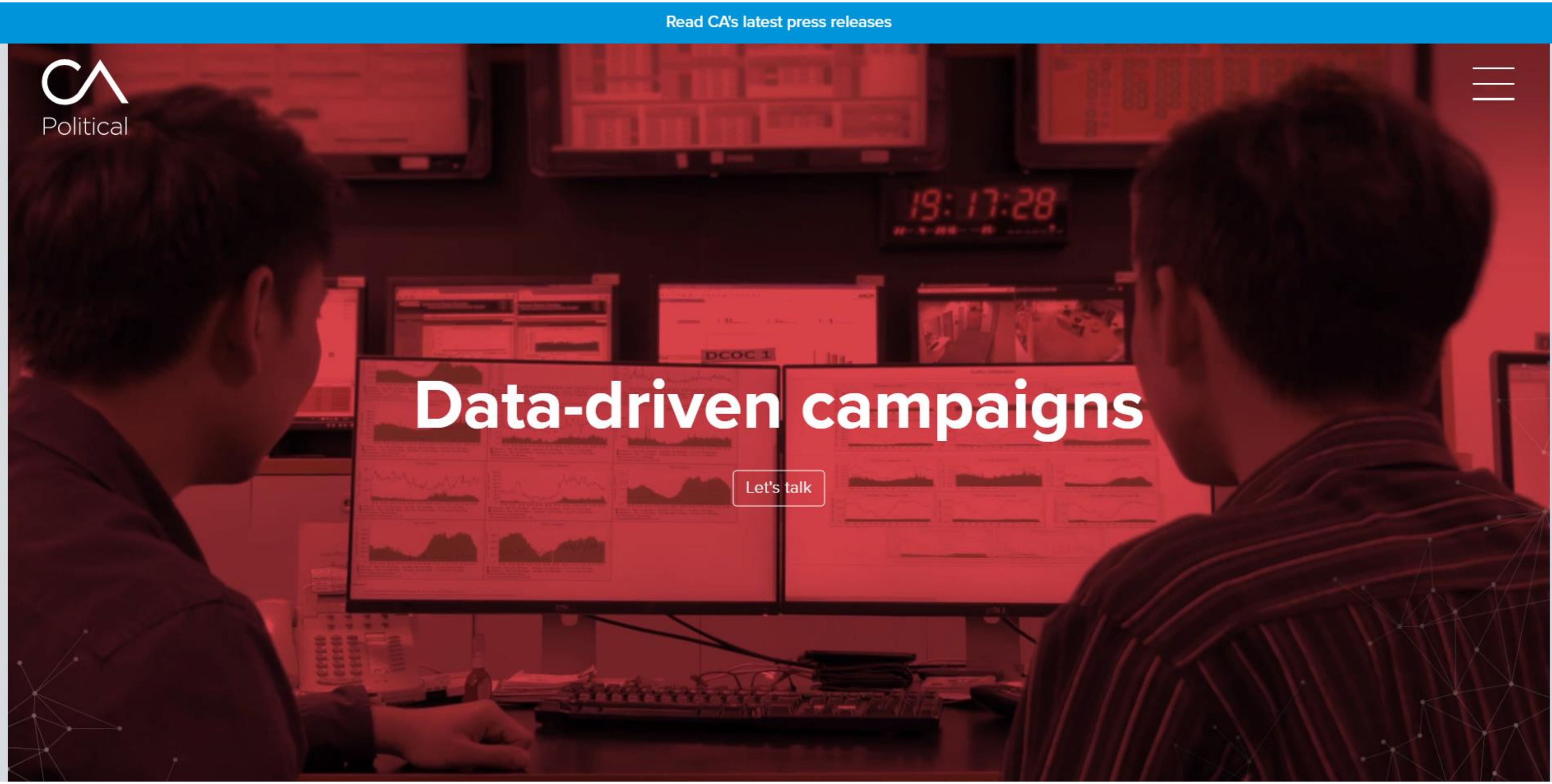
Sentiment Analysis (Support Group Posts)

- “[I] also found out that the radiologist is doing the biopsy, not a breast surgeon. I am more scared now than when I ...” fear
- “... My radiologist ‘assured’ me my scan was NOT going to be cancer...she was wrong.” despair
- “ ... My radiologist did my core biopsy. Not a problem and he did a super job of it.” hope

Sentiment Analysis

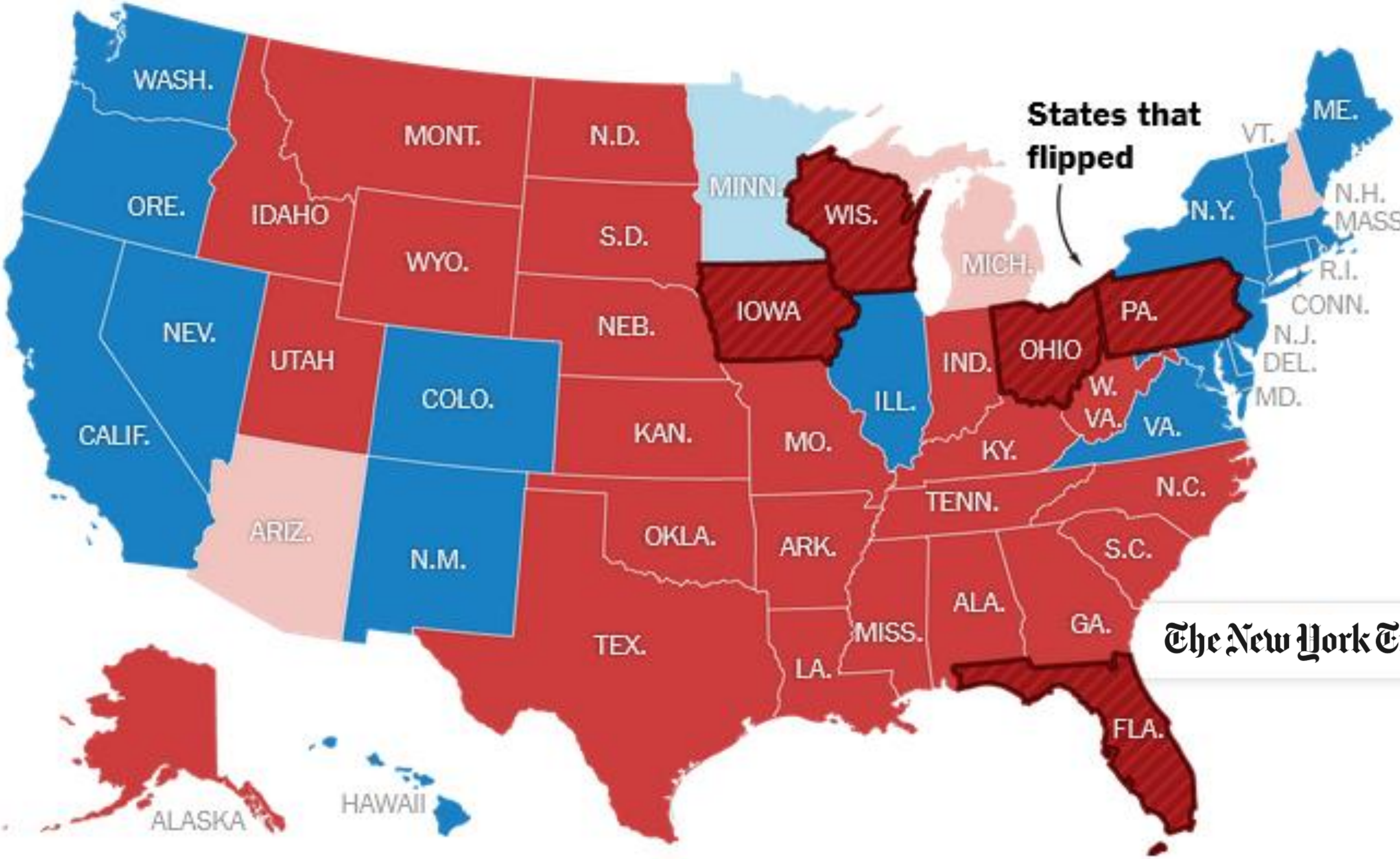
- **Sentiment Analysis**: automatically detecting the emotional state of the author from a span of text (usually from a set of pre-defined emotional states).

Text-based Forecasting



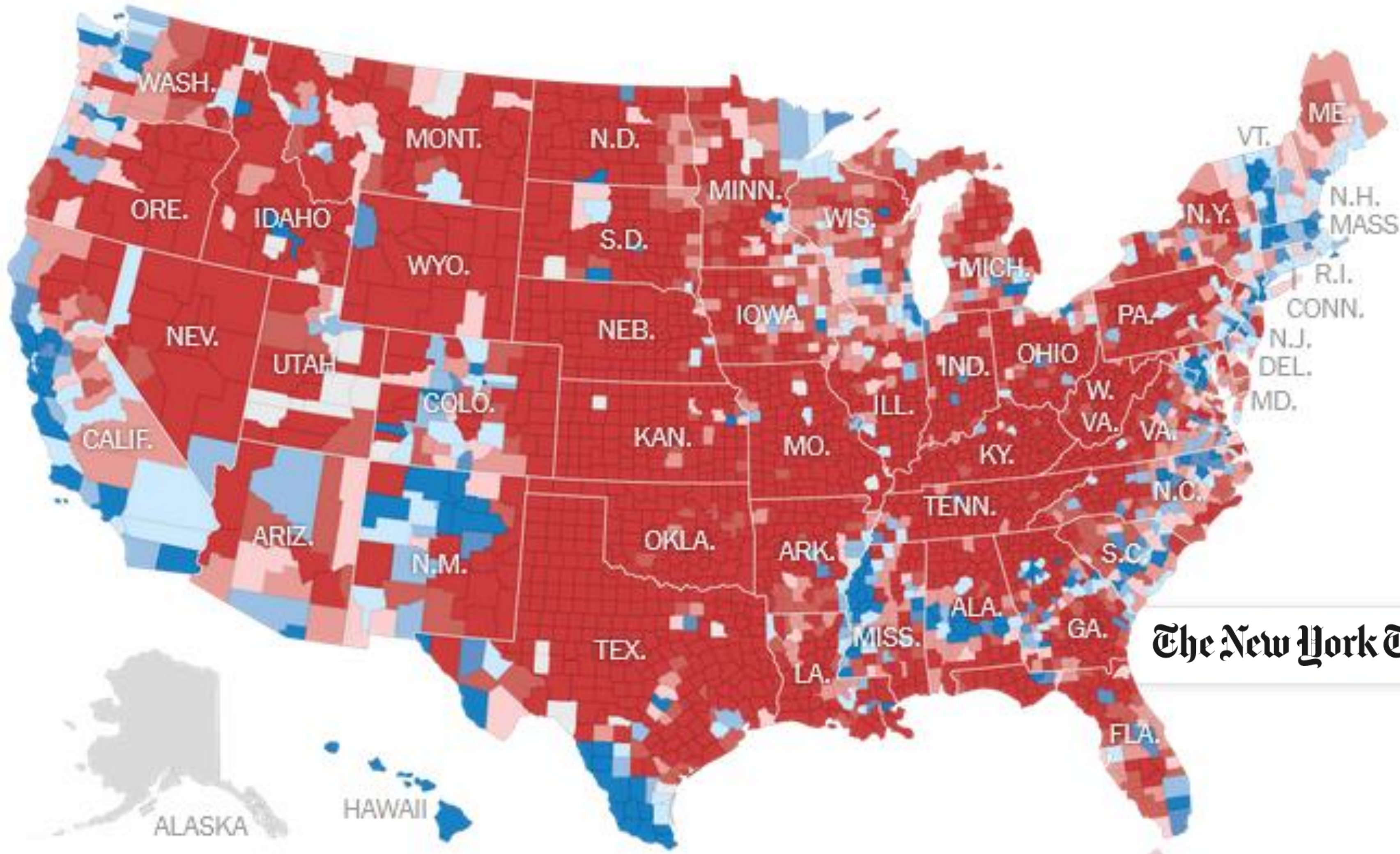
How Trump Reshaped the Election Map

Mr. Trump's victory was a historic rebuke to Democrats by white, blue-collar voters.



The New York Times

Mr. Trump dominated in counties across the rural midsection of the country.



The New York Times

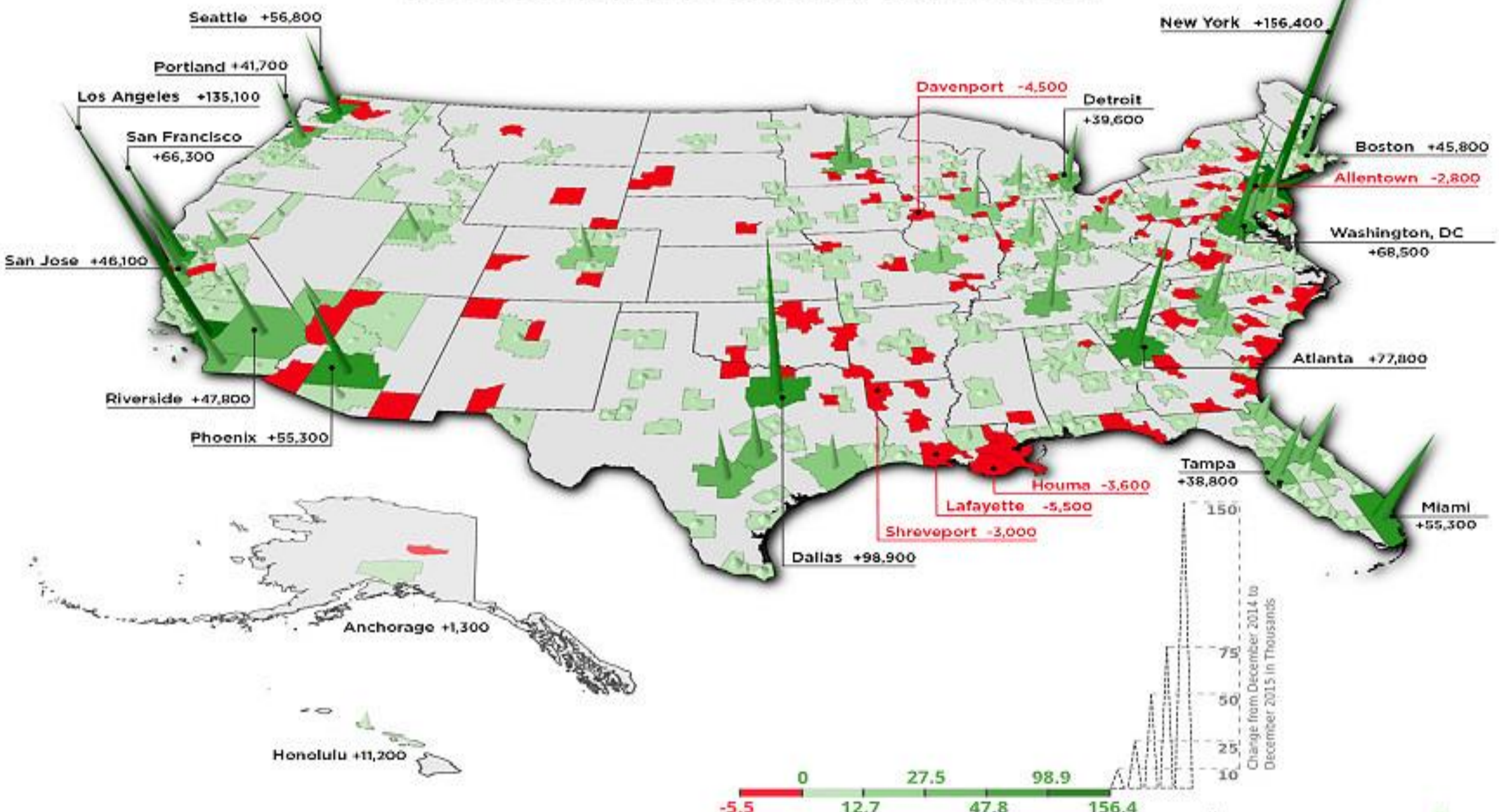
Why 2016 election polls missed their mark

BY ANDREW MERCER, CLAUDIA DEANE AND KYLEY MCGEENEY

Supporters of presidential candidate Hillary Clinton watch televised coverage of the U.S. presidential election at Comet Tavern in the Capitol Hill neighborhood of Seattle on Nov. 8. (Photo by Jason Redmond/AFP/Getty Images)

Job Growth in the U.S. (2015)

Take a Look Where the Jobs Were Created



Article & Sources: http://howmuch.net/articles/job-growth-usa-2015

howmuch.net

pennsylvania results

presidential results

president

key race 20 electoral votes

projected winner

trump ✓

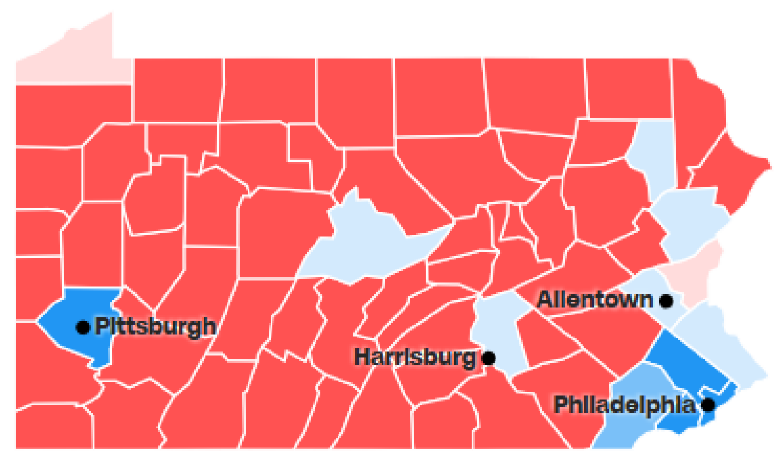
candidate	%	votes
trump	**48.8%**	**2,912,351**
clinton	47.7%	2,844,339
johnson	2.4%	142,623
stein	0.8%	48,998

est. 97% in

updated 9:50 am ET, Nov. 9

full details

county results map



size of lead

0%-5% 5%-10% 10%+ tie / no data

Text-based Forecasting

MAGELLAN STRATEGIES

Home
Services
Blog
Podcast
In the News
Our Team
Testimonials
Contact

303-861-8585

Every product and service Magellan Strategies offers is built on our decades of experience working with and analyzing political data. Political data is the foundation everything we do, but more importantly, we know how to use the data to help our clients make better decisions and generate relevant information so they have a competitive advantage in the political arena.

National Voter Registration Database of 190 Million Voters

Our in-house national voter registration database empowers us to work with candidates and political organizations anywhere in the country. We use the voter registration data for an array of political analysis projects, predictive data modeling, survey sample, and survey quotas. Our national voter file is updated on a regular basis, in some states monthly.

National Precinct Election Return Database

Precinct level election return data is incredibly helpful for political analysis, predictive modeling, and survey research projects, especially in states that do not register voters by party. This dataset helps our clients be more efficient in their ground operations and voter contact programs. Mapping the data is also very helpful for decision makers to understand past political performance in local areas.

Current and Historical Polling and Modeling Data

Magellan Strategies has historical benchmark, voter id, and predictive modeling survey data on more than one million registered voters across the country. We use this data to bring value to our clients who want to utilize data for their campaign from the beginning. This gives them a competitive edge on their opponent because they know what voters to contact at the start of the campaign.

"When it comes to voter data analysis, Magellan Strategies is one of the best. Their staff is extremely knowledgeable and highly responsive. The firm's seasoned team is quickly and effectively able to leverage their proprietary national database to address complex voter identification needs."

Paul Hanley, Senior Vice President,
George K. Baum & Company

CALL US TODAY
(303) 861-8585

START THE CONVERSATION

Text-based Forecasting

Link

Text-based Forecasting

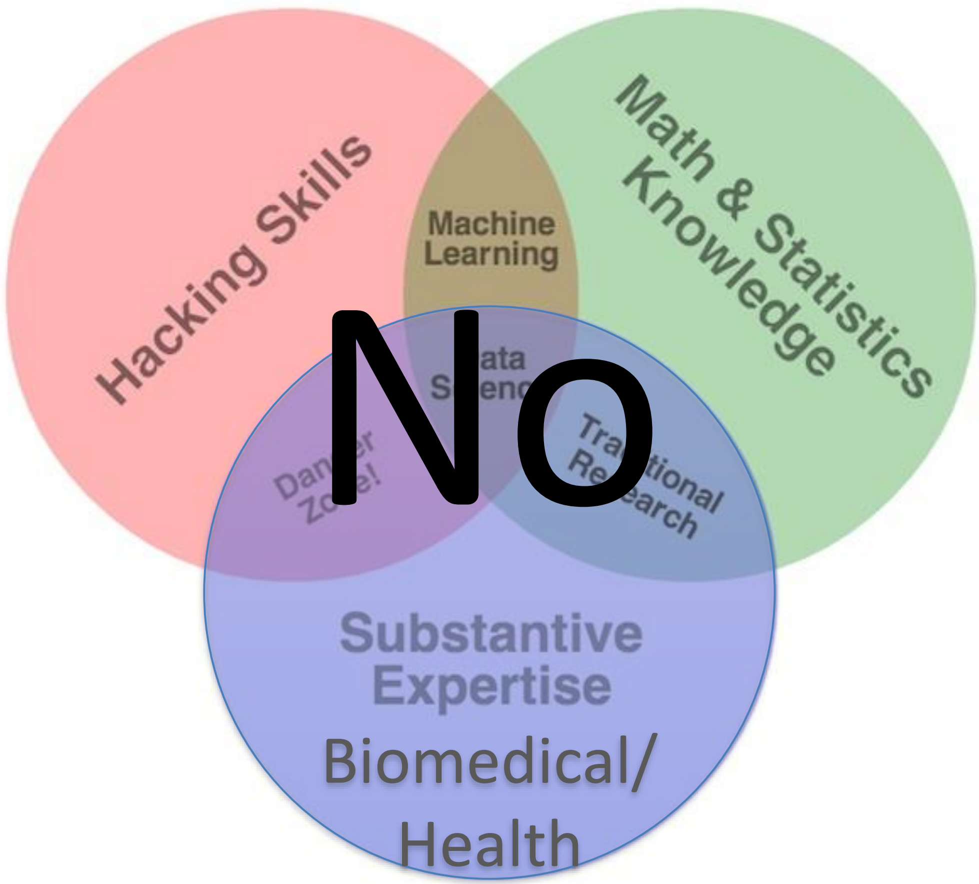
- Text-based Forecasting: monitoring incoming text (e.g., tweets) and making predictions about external, real-world events or trends, for example:
  - a presidential candidate’s poll rating
  - a company’s stock value change
  - a movie’s box office earnings
  - side-effects for a particular drug
  - Google Flu Trend

Course Curriculum

Learning Objectives

- Understand the power of a large amount of text data
- Learn underlying theories and techniques of text mining
- Make practices with real-world data and issues
- Learn data resources and tools related to biomedical and health informatics

Prerequisite?



Useful Resources

- Introduction to modern information retrieval. Gerard Salton and Michael J. McGill. 1986. Mcgraw Hill. ISBN:0070544840
- Fundamentals of Predictive Text Mining. Sholom M. Weiss, Nitin Indurkhya, and Tong Zhang. 2nd Edition. 2015. Springer. ISBN:144716749X

Roadmap

- Predictive Analysis of Text
  - Supervised machine learning principles
  - Text representation
  - Basic machine learning algorithms
  - Experimentation and evaluation
  - Feature selection
- Exploratory Analysis of Text
  - Clustering
  - Co-occurrence statistics

Roadmap

- Biomedical/health informatics related data resources
- Tools for text analytics
- Introduction to applications of text mining
- Is there anything that you would like to learn more about?

Course Resources

- **Class website**
  - Schedule and lecture slides https://enable.unc.edu/hidav-curriculum-resources/
- **Piazza**
  - Class discussion and support forum
  - Please register now!
  - piazza.com/unc/summer2018/hidav_text

Seeking Help

- Best Option: **Piazza**

- E-Mail: heejunk**@email.unc.edu**
  - It will likely take **<u>12-36 hours</u>** to get a response, often even longer
  - Use public posts if possible.
  - **<u>Try Piazza first!</u>**

- Office Hours:
  - After Tuesday class
  - **<u>Or by appointment</u>**

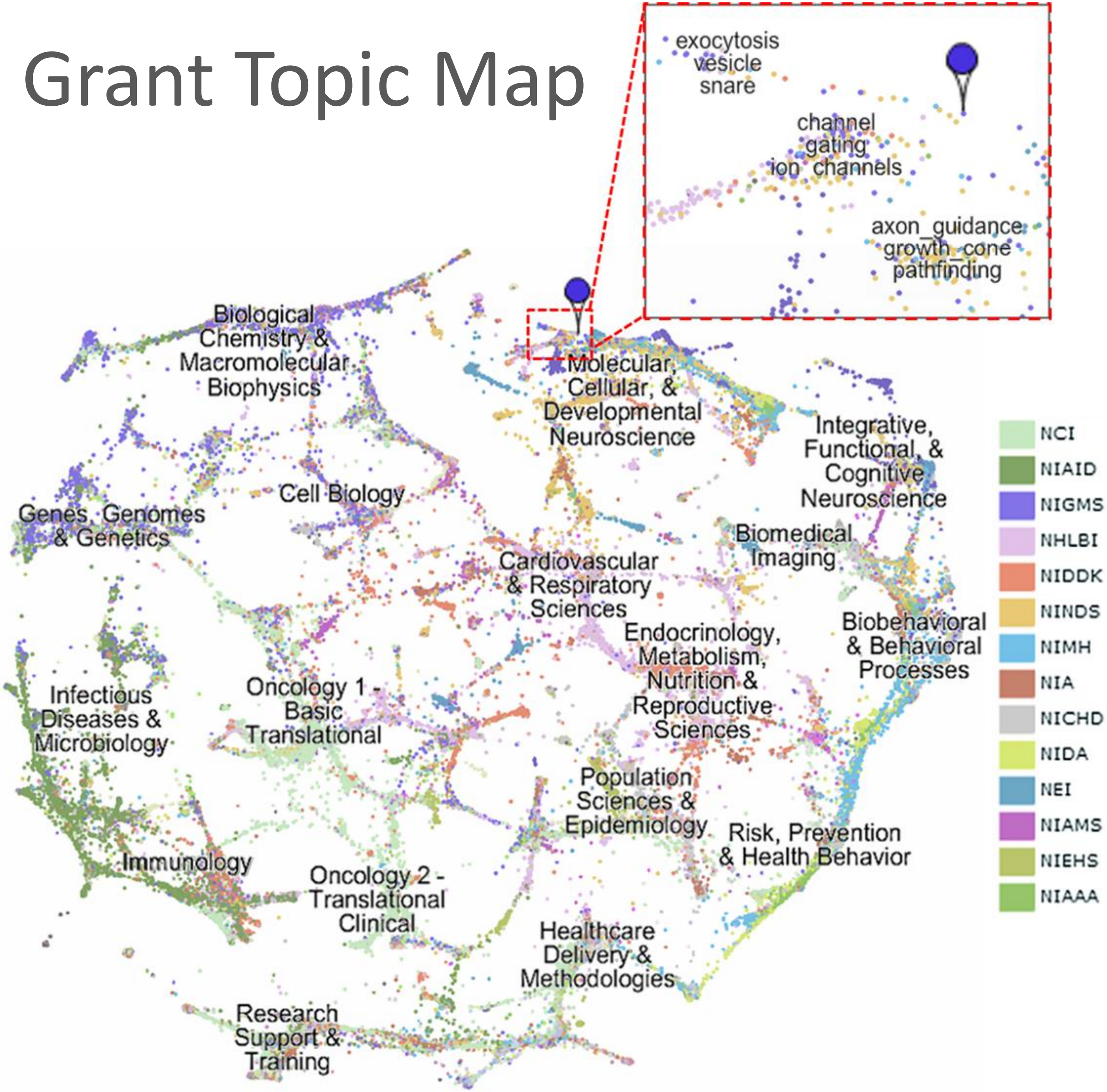
Course Tips

- Work hard
- Be patient and have reasonable expectations
- you’re not supposed to understand everything we cover in class during class
- Seek help sooner rather than later
- Remember the golden rule: no pain, no gain

Any Questions?

A Sample Application: Topic Categorization and Visualization

NIH Grant Topic Map



Topic Maps (Simple Case)

BRCA

- Doc1: Genotype-Phenotype Correlations in BRCA Mutation Carriers
- Doc2: Breast cancer following ovarian cancer in BRCA mutation carriers
- Doc3: Breast cancer, BRCA mutations, and attitudes regarding pregnancy
- Doc4: Surgical management of breast cancer in BRCA-mutation carriers
- Doc5: Cancer risk management decision making for BRCA women

- Doc6: Inverse association between cancer and neurodegenerative disease
- Doc7: Molecular neurodegeneration: basic biology and disease pathways
- Doc8: Mechanisms of neurodegeneration and axonal dysfunction
- Doc9: Dysfunction of neuronal calcium signaling in neuroinflammation and neurodegeneration
- Doc10: Epigenetic mechanisms of neurodegeneration in Huntington's disease

Neurodegeneration

Bag of Words Representation

Genotype-Phenotype Correlations in BRCA Mutation Carriers

Breast cancer following ovarian cancer in BRCA mutation carriers

Breast cancer, BRCA mutations, and attitudes regarding pregnancy

Surgical management of breast cancer in BRCA-mutation carriers

Cancer risk management decision making for BRCA women

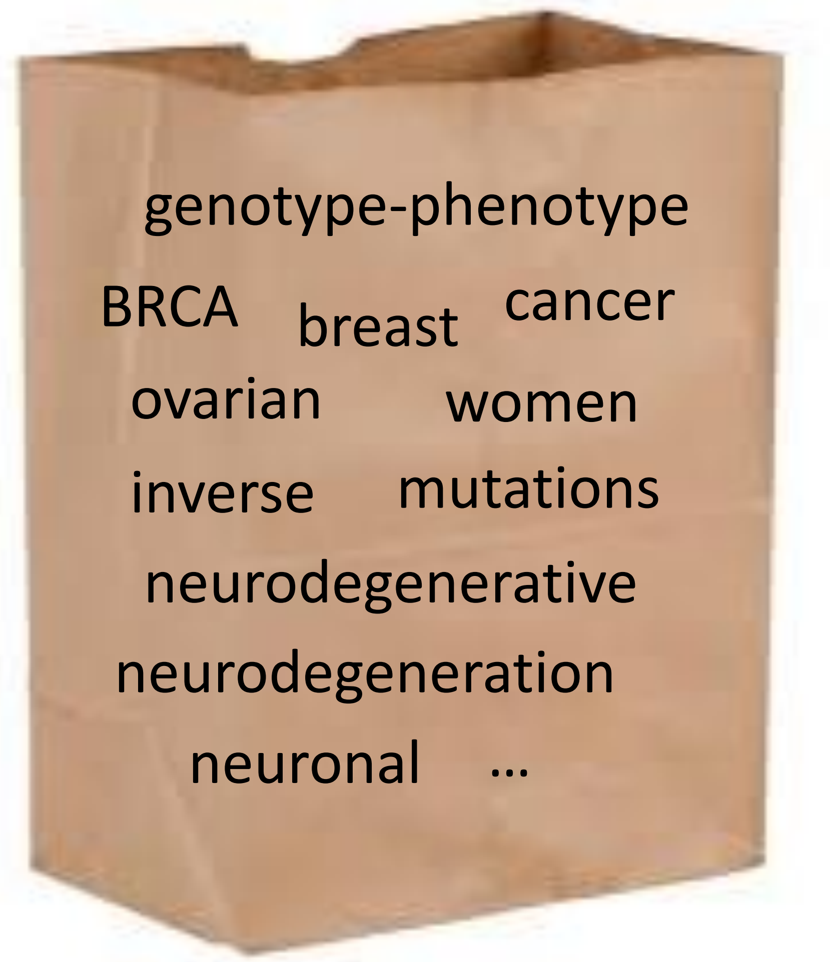
Inverse association between cancer and neurodegenerative disease

Molecular neurodegeneration: basic biology and disease pathways

Mechanisms of neurodegeneration and axonal dysfunction

Dysfunction of neuronal calcium signaling in neuroinflammation and neurodegeneration

Epigenetic mechanisms of neurodegeneration in Huntington's disease



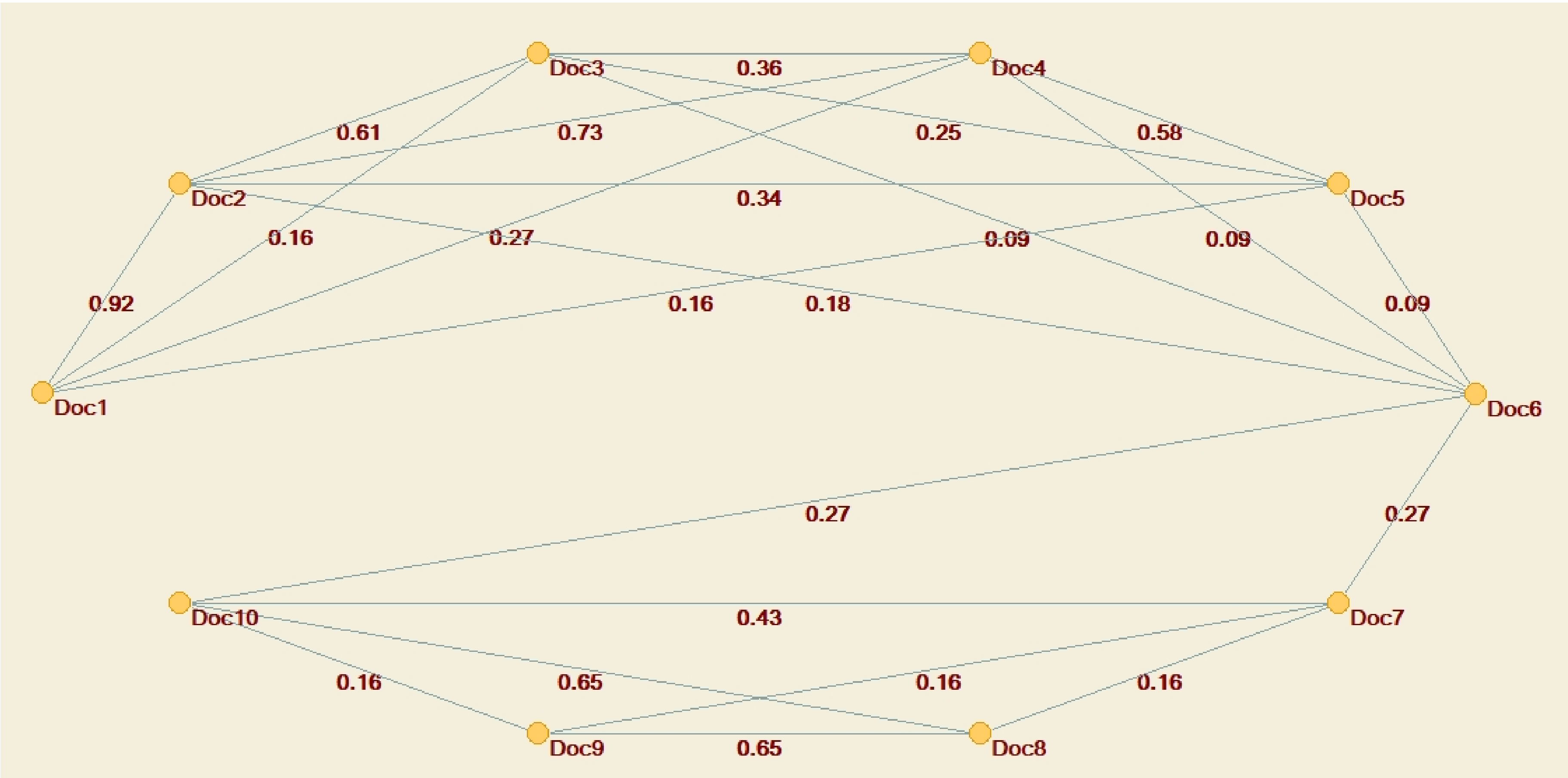
Document-Term Matrix

	brca	breast	cancer	mutation	neuro degeneration	neur onal	neuro degenerative	…
Doc1	1	0	0	1	0	0	0	
Doc2	1	1	2	1	0	0	0	
Doc3	1	1	1	0	0	0	0	
Doc4	0	1	1	1	0	0	0	
Doc5	1	0	1	0	0	0	0	
Doc6	0	0	1	0	0	0	1	
Doc7	0	0	0	0	1	0	0	
Doc8	0	0	0	0	1	0	0	
Doc9	0	0	0	0	1	1	0	
Doc10	0	0	0	0	1	0	0	

Similarity Among Documents

	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10
Doc1	1.000	0.364	0.078	0.111	0.068	0.000	0.000	0.000	0.000	0.000
Doc2	0.364	1.000	0.294	0.337	0.167	0.120	0.000	0.000	0.000	0.000
Doc3	0.078	0.294	1.000	0.159	0.110	0.056	0.000	0.000	0.000	0.000
Doc4	0.111	0.337	0.159	1.000	0.182	0.059	0.000	0.000	0.000	0.000
Doc5	0.068	0.167	0.110	0.182	1.000	0.049	0.000	0.000	0.000	0.000
Doc6	0.000	0.120	0.056	0.059	0.049	1.000	0.092	0.000	0.000	0.110
Doc7	0.000	0.000	0.000	0.000	0.000	0.092	1.000	0.082	0.058	0.168
Doc8	0.000	0.000	0.000	0.000	0.000	0.000	0.082	1.000	0.245	0.300
Doc9	0.000	0.000	0.000	0.000	0.000	0.000	0.058	0.245	1.000	0.069
Doc10	0.000	0.000	0.000	0.000	0.000	0.110	0.168	0.300	0.069	1.000

Topic Map (Visualization)



THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Tokenization

- Token
  - A unit of text analysis. Usually a word or other atomic parse element (i.e., symbol, term, etc.)
- Tokenization
  - Splitting text into terms of tokens

Tokenization

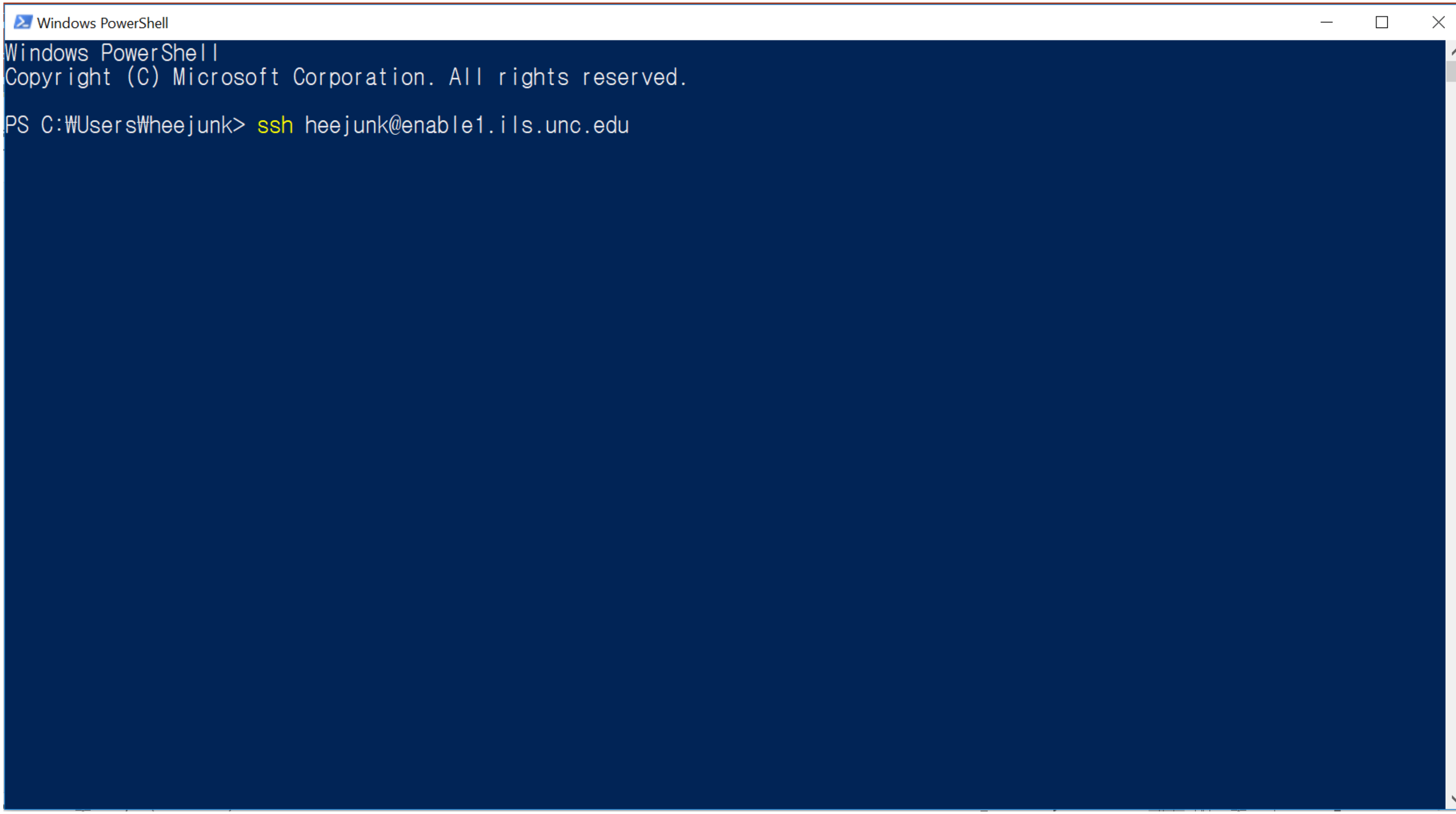
- Bag of Words Text Representation
  - Making a set of distinct terms appearing at least once in text corpus
  - No duplication is allowed, position information and word order is lost

A Hands-on Practice

Starting Windows Powershell

Right-click on the Windows icon at the bottom left of Windows screen and select "Windows Powershell."

Connecting to a Server



Replace your own ID with *heejunk* and type the Onyen password.

Running a Code for Tokenization

```
Select heejunk@enable1: ~/clustering
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

PS C:\Users\heejunk> ssh heejunk@enable1.ils.unc.edu
heejunk@enable1.ils.unc.edu's password:
Warning: Your password will expire in 60 days on Sun Jun 17 06:14:20 2018
Welcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.4.0-116-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

25 packages can be updated.
0 updates are security updates.


*** System restart required ***
Last login: Tue Apr 17 16:06:06 2018 from 152.23.48.198
heejunk@enable1:~$ cd clustering/
heejunk@enable1:~/clustering$ python3 "tokens.py"
Within your current directory, please provide the filename: sample.txt
```

Move into a “clustering” directory, run “tokens.py” and select input file.

Examining Tokens

```
heejunk@enable1: ~/clustering
ical', 'decision', 'support', 'intervention', 'designed', 'reduce', 'urinary', 'catheter', 'utilization', 'catheter-asso
ciated', 'tract', 'infections', 'testing', 'avoiding', 'diabetes', 'action', 'plan', 'targeting', '(adapt)', 'integratin
g', 'care-based', 'counseling', 'pre-diabetes', 'evaluation', 'pharmacogenomics', 'aids', 'knowledge', 'resources', 'pro
vider', 'order', 'entry', 'system:', 'mixed', 'methods', 'developing', 'software', ' "track', 'catch" ', 'missed', 'foll
ow-up', 'abnormal', 'test', 'results', 'complex', 'sociotechnical', 'environment', 'comparison', 'alphabetical', 'versus
', 'categorical', 'display', 'format', 'medication', 'simulated', 'touch', 'screen', 'anesthesia', 'information', 'manag
ement', 'experiment', 'clinician-computer', 'interaction', 'increasing', 'efficacy', 'primary', 'prevention:', 'rational
e', 'design', 'adapt', '(avoiding', 'targeting)', 'trial', 'dashboard', 'physician', 'effi-', 'ciency', 'accuracy', 'acce
ssing', 'data', 'needed', 'high-quality', 'care', 'socio-technical', 'considerations', 'epilepsy', 'patient', 'implement
ation', 'taste', 'individualized', 'medicine:', 'physicians' ', 'reactions', 'automated', 'genetic', 'interpretations',
'point-of-care', 'documentation:', 'assessment', 'bladder', 'cancer', 'informatics', 'tool', '(ecancercarebladder):', 'r
andomized', 'controlled', 'study', 'efficacy,', 'efficiency', 'user', 'friendliness', 'compared', 'standard', 'medical',
 'records', 'effect', 'interfaces', 'intensive', 'task', 'load,', 'errors', 'cognition,', 'performance', 'iso', '13606',
 'archetype', 'based', 'visualization', 'method', 'development', 'computerised', 'decisions', 'system', 'renal', 'risk',
 'drugs', 'healthcare', 'v-model:', 'perspective', 'ehr-based', 'phenotyping', 'shared', 'e-decision', 'portal', 'pediat
ric', 'asthma', 'trauma', 'record:', 'tablet-based', 'injury', 'surveillance', 'low', 'resource', 'settings', 'ws', 'app
lication', 'applying', 'adoption', 'framework', 'evaluate', 'ambulatory', 'interoperability', 'practices:', 'discrete-ev
ent', 'simulation', 'problem', 'list', 'terminologies', 'acceptance,', 'safety,', 'effectiveness', 'computer-based', 'co
lonoscopy', 'preparation', 'facilitate', 'reconciliation', 'hospital', 'discharge', 'technical', 'solution', 'improving'
, 'palliative', 'hospice', 'consultation', 'alternative', 'referral', 'patients', 'chronic', 'kidney', 'disease:', 'netw
orked', 'improve', 'accessibility', 'documentation', 'time', 'paper-based', 'optometrists', 'eye', 'south', 'india:', 't
ime-motion', 'making', 'pharmacogenomic-based', 'prescribing', 'alerts', 'effective:', 'scenario-based', 'pilot', 'physi
cians', 'determining', 'differences', 'expert', 'novice', 'doctors', '(ehr)', 'knowledge-based', 'personal', 'empower',
'outpatients', 'mellitus', 'means', 'p4', 'medicine', 'clinicians0', 'computer-assisted', 'summarization', 'essential',
'questions:', 'accuracy,', 'perceptions', 'drag/drop', 'user-composable', 'workflow', 'systems,', 'digital', 'dictation'
, 'think-aloud', 'protocol', 'analysis', ' "near-live" ', 'simulations', 'evaluating', 'cognitive', 'highly', 'configura
ble', 'web', '2.0', 'ehr', 'interface', 'advice', 'makers', 'program', 'all-inclusive', 'elders', 'site', 'prototype',
practice:', 'secondary', 'stroke', 'prevention', 'veterans', 'facility', 'importance', 'selecting', 'implementing', 'sys
tem.', 'ahlta', 'clinics', 'military', 'center.', 'standards-based', 'interoperable', 'architecture', 'context', 'korean
', 'creation', 'emr-based', 'paper', 'summaries', 'hiv-care', 'uganda,', 'africa', 'mobile', 'ict', 'settings:', 'method
ological', 'practical', 'challenges', 'harvest,', 'longitudinal', 'summarizer', 'histories', 'home']
heejunk@enable1:~/clustering$
```

The script automatically generates tokens, displays those tokens on the screen, and also stores it in a file (tokens.txt).

Any Questions?

Two Paradigms in Text Mining

Two Paradigms in Text Mining

- <u>Predictive</u> Analysis of Text
  - Predict: “to say that an event or action will happen in the future, especially <u>as a result of knowledge</u> or experience” (Cambridge Dictionary)
- <u>Exploratory</u> Analysis of Text
  - Explore: “<u>travel in</u> or through (an unfamiliar country or area) in order <u>to learn about</u> or <u>familiarize</u> oneself with it.” (Wikipedia)

Two Paradigms in Text Mining

Explore

Predict

What makes the difference?

Knowledge,
Experience

Two Paradigms in Text Mining

Explore

Predict

So how can computers have knowledge/experience?

Two Paradigms in Text Mining

Exploratory Analysis

Doc No.	Content
Doc1	Genotype-Phenotype Correlations in BRCA Mutation Carriers
Doc2	Breast cancer following ovarian cancer in BRCA mutation carriers
Doc3	Inverse association between cancer and neurodegenerative disease
Doc4	Epigenetic mechanisms of neurodegeneration in Huntington's disease

Predictive Analysis

Doc No.	Content	Label
Doc1	Genotype-Phenotype Correlations in BRCA Mutation Carriers	BRCA
Doc2	Breast cancer following ovarian cancer in BRCA mutation carriers	BRCA
Doc3	Inverse association between cancer and neurodegenerative disease	Neuro degen-eration
Doc4	Epigenetic mechanisms of neurodegeneration in Huntington's disease	Neuro degen-eration

Two Paradigms in **Machine Learning**

Unsupervised Learning

Doc No.	Content
Doc1	Genotype-Phenotype Correlations in BRCA Mutation Carriers
Doc2	Breast cancer following ovarian cancer in BRCA mutation carriers
Doc3	Inverse association between cancer and neurodegenerative disease
Doc4	Epigenetic mechanisms of neurodegeneration in Huntington's disease

Supervised Learning

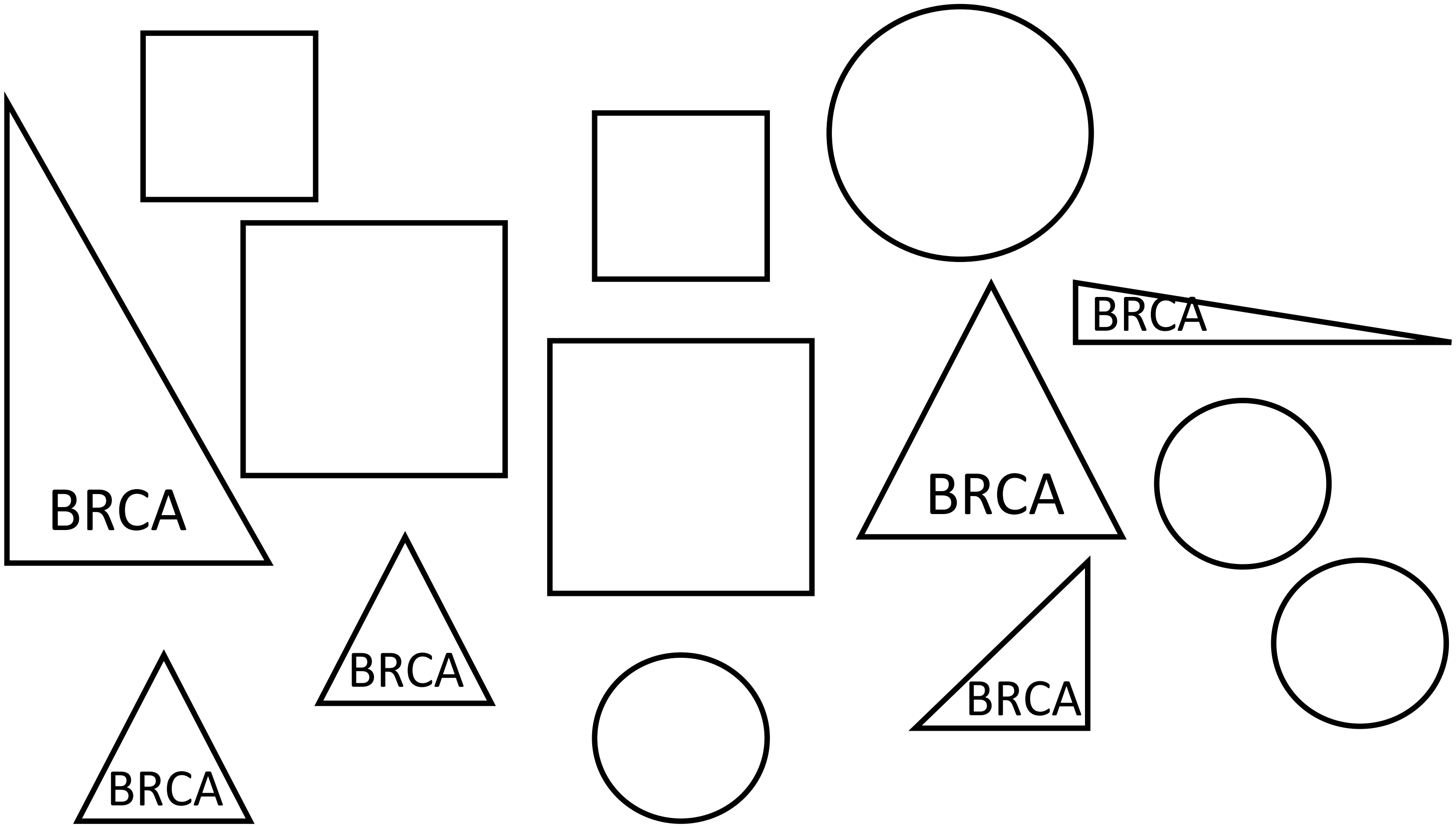
Doc No.	Content	Label
Doc1	Genotype-Phenotype Correlations in BRCA Mutation Carriers	BRCA
Doc2	Breast cancer following ovarian cancer in BRCA mutation carriers	BRCA
Doc3	Inverse association between cancer and neurodegenerative disease	Neuro degen-eration
Doc4	Epigenetic mechanisms of neurodegeneration in Huntington's disease	Neuro degen-eration

Two Paradigms in Text Mining

- Predictive Analysis of Text
  - developing computer programs that automatically <u>predict</u> a particular <u>concept</u> within a span of text
- Exploratory Analysis of Text
  - developing computer programs that automatically <u>discover</u> interesting and useful <u>patterns or trends</u> in text collections

Predictive Analysis of Text: The Big Picture

Predictive Analysis example: recognizing triangles (e.g.,BRCA)

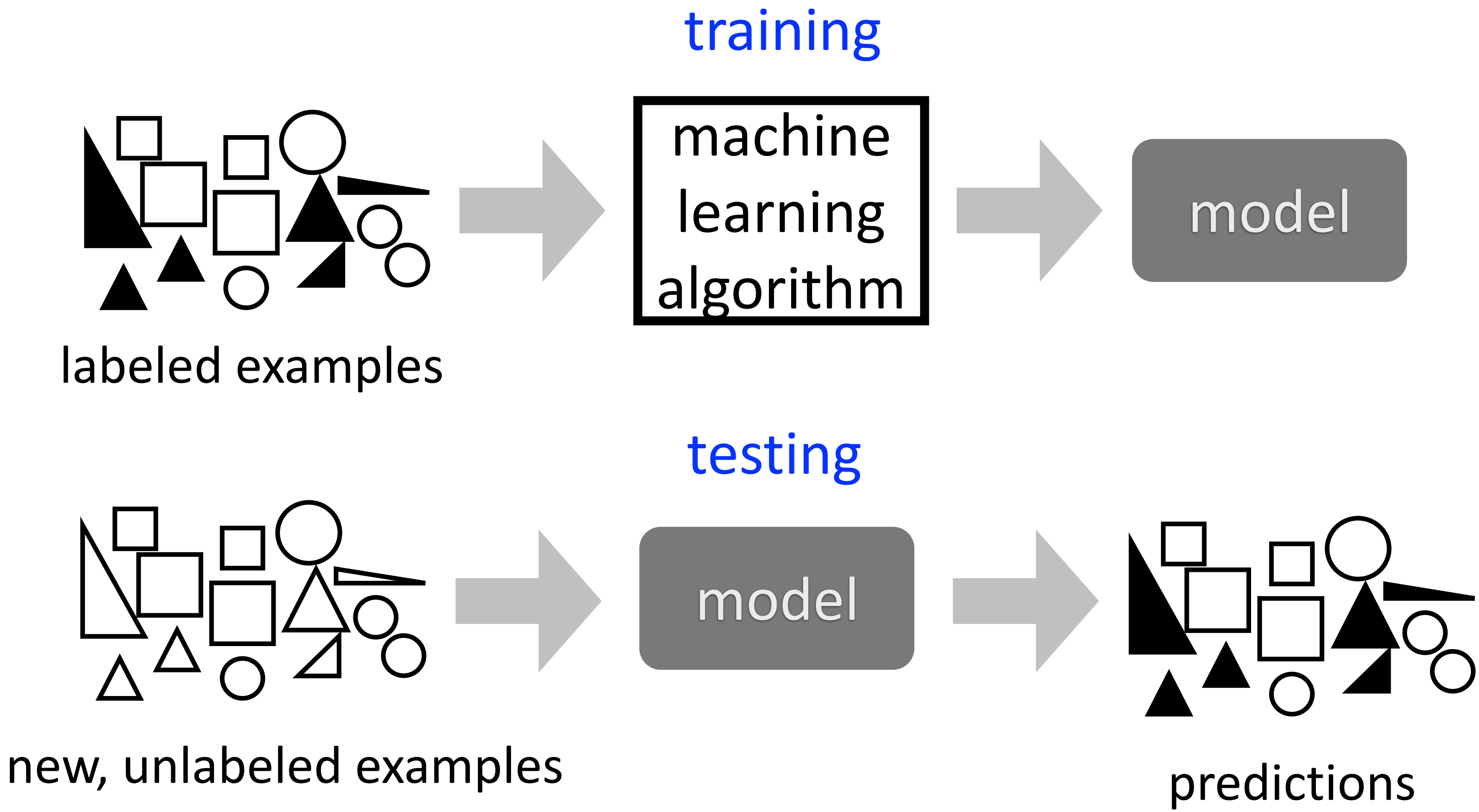


ENABLE THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Predictive Analysis example:
recognizing triangles

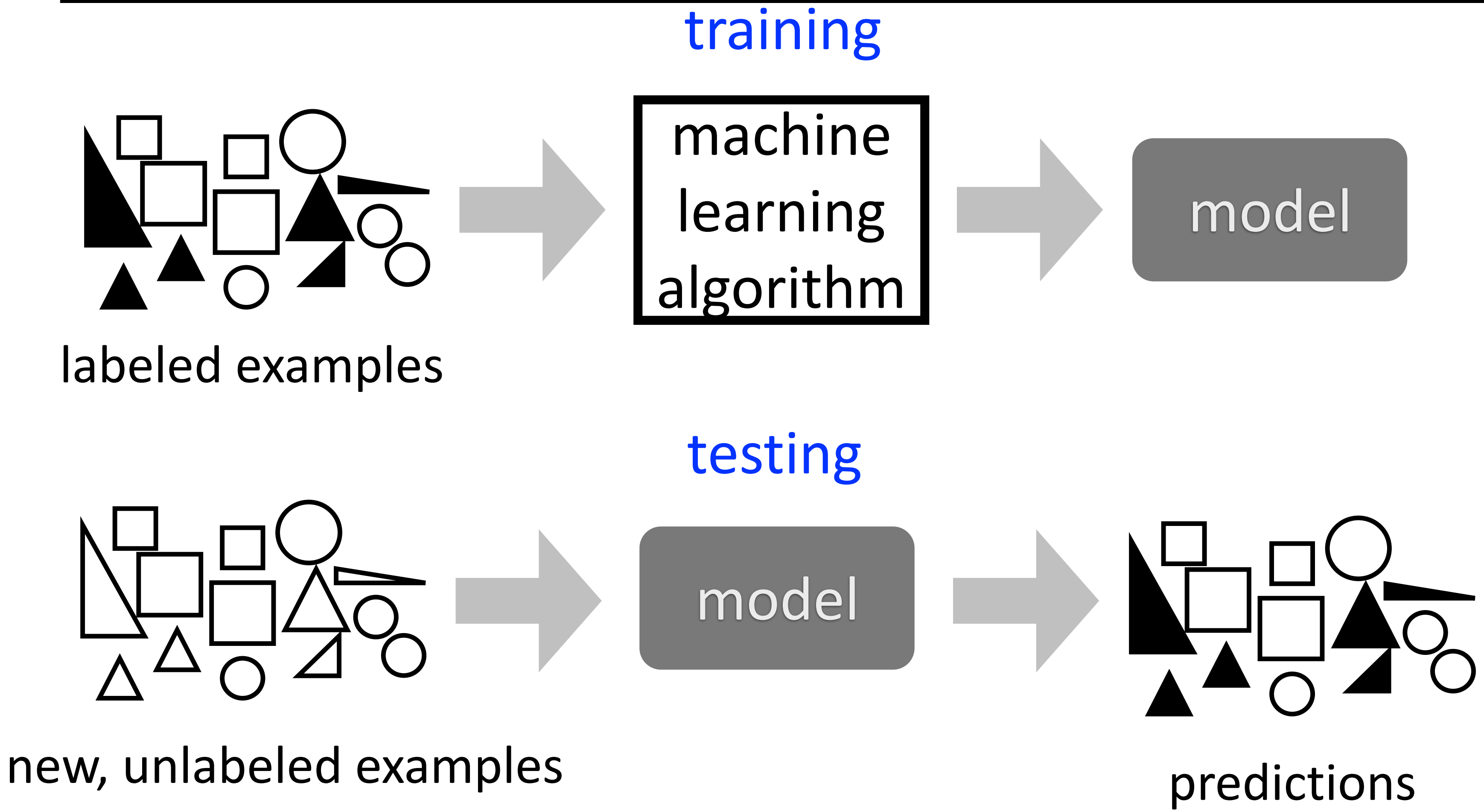
- We could imagine writing a “triangle detector” by hand:
  - if shape has three sides, then shape = triangle.
  - otherwise, shape = other.
- Alternatively, we could use supervised machine learning!

Predictive Analysis example:
recognizing triangles



What is the part that is missing?

HINT: It’s what most of this summer will be about!



Predictive Analysis

representation: features

color	size	# slides	equal sides	...	label
red	big	3	no	...	yes
green	big	3	yes	...	yes
blue	small	inf	yes	...	no
blue	small	4	yes	...	no
⋮	⋮	⋮	⋮	⋮	⋮
red	big	3	yes	...	yes

Predictive Analysis example:
recognizing triangles

color	size	# slides	Equal sides	...	label
red	big	3	no	...	yes
green	big	3	yes	...	yes
blue	small	inf	yes	...	no
blue	small	4	yes	...	no
⋮	⋮	⋮	⋮	⋮	⋮
red	big	3	yes	...	yes

labeled examples

training

machine learning algorithm

model

color	size	# slides	Equal sides	...	label
red	big	3	no	...	?
Green	big	3	yes	...	?
blue	small	inf	yes	...	?
blue	small	4	yes	...	?
⋮	⋮	⋮	⋮	⋮	⋮
red	big	3	yes	...	?

new, unlabeled examples

testing

model

color	size	# slides	Equal sides	...	label
red	big	3	no	...	yes
green	big	3	yes	...	yes
blue	small	inf	yes	...	no
blue	small	4	yes	...	no
⋮	⋮	⋮	⋮	⋮	⋮
red	big	3	yes	...	yes

predictions

Predictive Analysis: basic ingredients

Highly Influential

- **Training data**: a set of examples of the concept we want to automatically recognize
- **Representation**: a set of features that we believe are useful in recognizing the desired concept
- **Learning algorithm**: a computer program that uses the training data to learn a predictive model of the concept

Predictive Analysis:
basic ingredients

- **Model**: a (mathematical) function that describes a predictive relationship between the feature values and the presence/absence of the concept
- **Test data**: a set of previously unseen examples used to estimate the model's effectiveness
- **Performance metrics**: a set of statistics used to measure the predictive effectiveness of the model

Common Mistakes in Predictive Analysis

Feature Representation: what could possibly go wrong?

color	size	90 deg. angle	equal sides	...	label
red	big	yes	no	...	yes
green	big	no	yes	...	yes
blue	small	no	yes	...	no
blue	small	yes	yes	...	no
⋮	⋮	⋮	⋮	⋮	⋮
red	big	no	yes	...	yes

Feature Representation:
what could possibly go wrong?

color	size	90 deg. angle	equal sides	...	label
red	big	yes	no	...	yes
green	big	no	yes	...	yes
blue	small	no	yes	...	no
blue	small	yes	yes	...	no
⋮	⋮	⋮	⋮	⋮	⋮
red	big	no	yes	...	yes

1. bad feature representation!

Selective Attention Test

Link

Monkey Business Illusion

Link

Training data + Feature Representation: what could possibly go wrong?

color	size	# slides	equal sides	...	label
blue	big	3	no	...	yes
blue	big	3	yes	...	yes
red	small	inf	yes	...	no
green	small	4	yes	...	no
⋮	⋮	⋮	⋮	⋮	⋮
blue	big	3	yes	...	yes

ENABLE

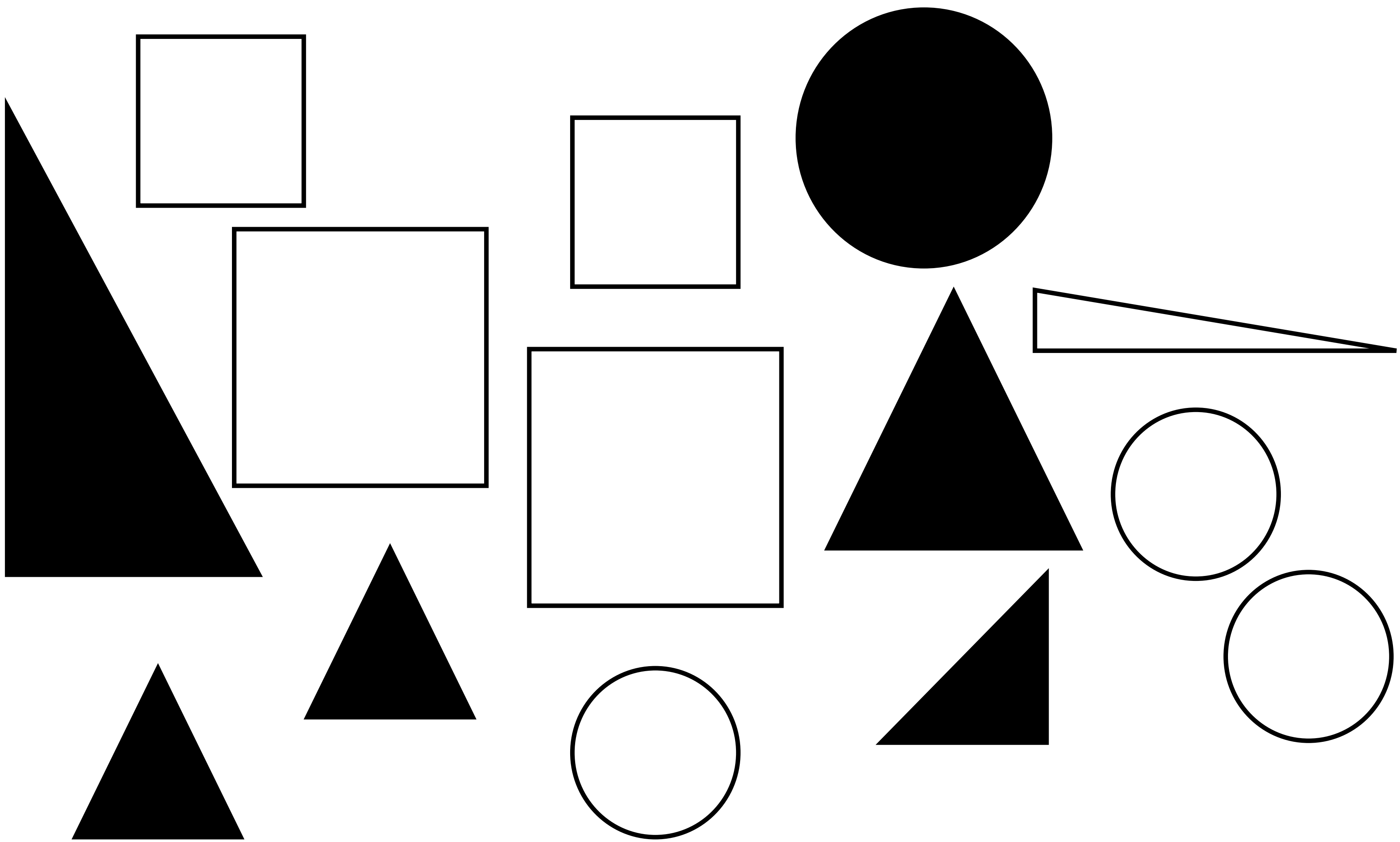
THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Training data + Feature Representation: what could possibly go wrong?

color	size	# slides	equal sides	...	label
blue	big	3	no	...	yes
blue	big	3	yes	...	yes
red	small	inf	yes	...	no
green	small	4	yes	...	no
⋮	⋮	⋮	⋮	⋮	⋮
blue	big	3	yes	...	yes

2. bad data + misleading correlations!

Training data + Feature Representation: what could possibly go wrong?



Training data + Feature Representation: what could possibly go wrong?

color	size	# slides	equal sides	...	label
white	big	3	no	...	yes
white	big	3	no	...	no
white	small	inf	yes	...	yes
white	small	4	yes	...	no
⋮	⋮	⋮	⋮	⋮	⋮
white	big	3	yes	...	yes

Training data + Feature Representation: what could possibly go wrong?

color	size	# slides	equal sides	...	label
white	big	3	no	...	yes
white	big	3	no	...	yes
white	small	inf	yes	...	yes
white	small	4	yes	...	No
⋮	⋮	⋮	⋮	⋮	⋮
white	big	3	yes	...	yes

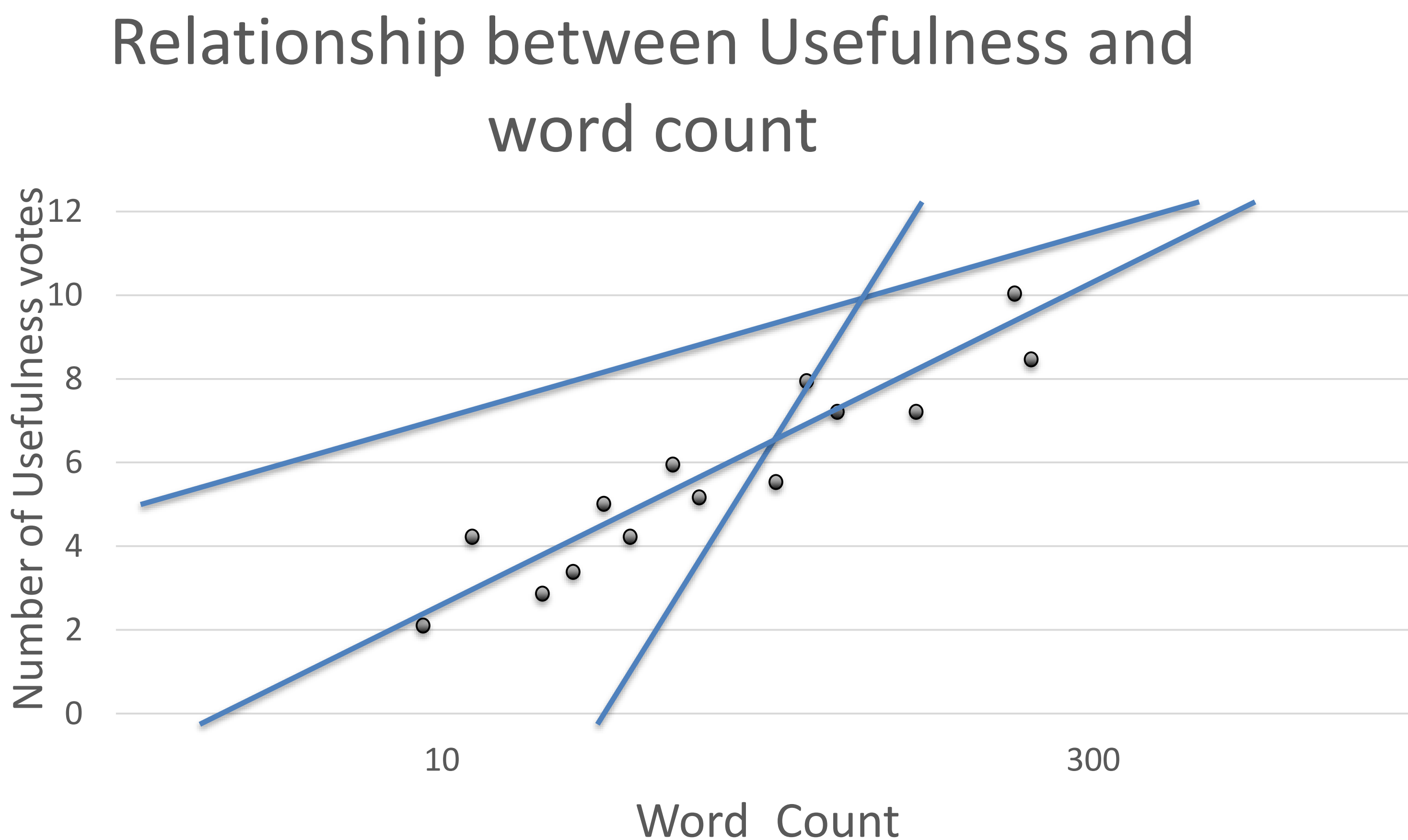
3. Noisy training data!

Learning Algorithm + Model: what could possibly go wrong?

- Linear classifier

$$y = \begin{cases} 1 & \text{if } w_0 + \sum_{j=1}^{n} w_j x_j > 0 \\ 0 & \text{otherwise} \end{cases}$$

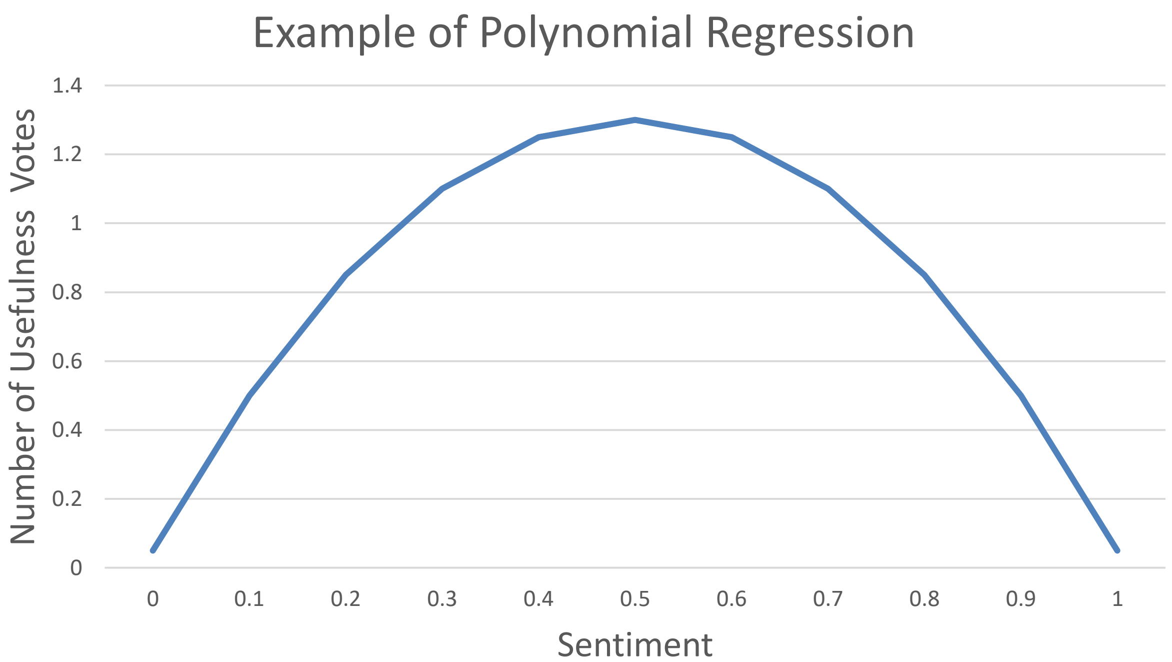
Learning Algorithm + Model: what could possibly go wrong?



Learning Algorithm + Model: what could possibly go wrong?

x_1

- Polynomial model: $y = a - b_1 * (x_1 - b_2)^2$



4. Bad learning algorithm

Evaluation Metrics:
what could possibly go wrong?

- Most evaluation metrics can be understood using a contingency table

		true	
		triangle	other
predicted	triangle	A	B
	other	C	D

- What number(s) do we want to maximize?
- What number(s) do we want to minimize?

Evaluation Metrics:
what could possibly go wrong?

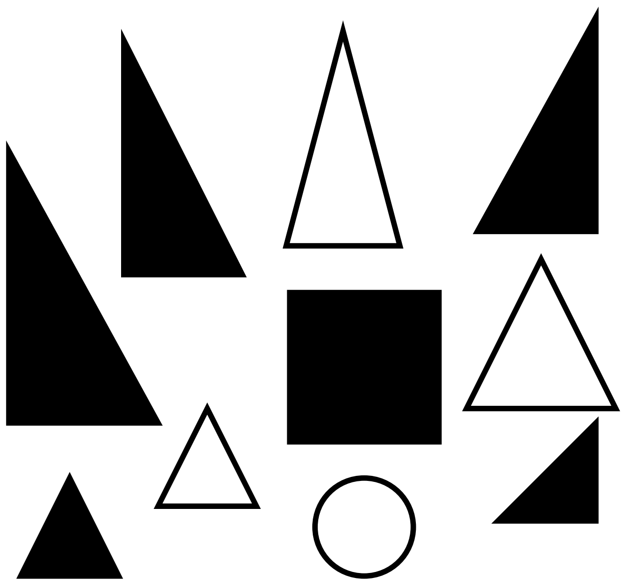
- Accuracy: percentage of predictions that are correct

$$\frac{(A + D)}{(A + B + C + D)}$$

predicted \ true	triangle	other
triangle	A	B
other	C	D

Evaluation Metrics:
what could possibly go wrong?

- Accuracy: percentage of predictions that are correct



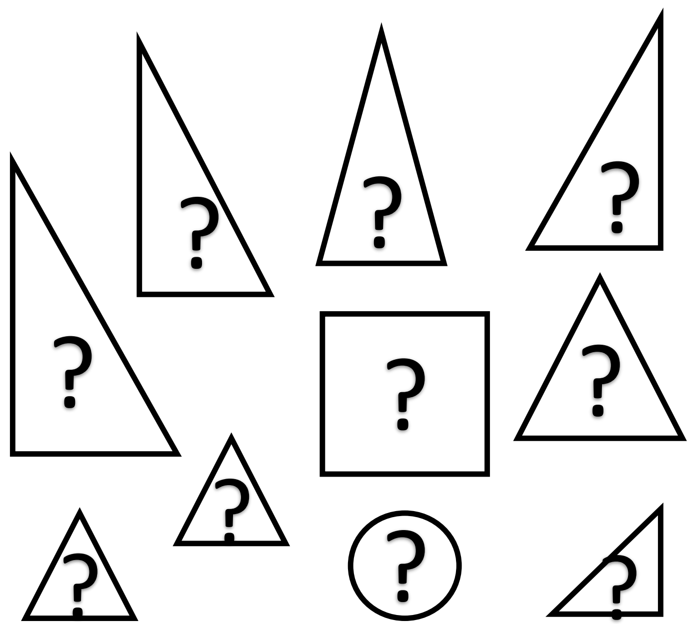
		true	
		triangle	other
predicted	triangle	5	1
	other	3	1

$$\frac{(5 + 1)}{(5 + 1 + 3 + 1)}$$

- What is the accuracy of this model?

Evaluation Metrics:
what could possibly go wrong?

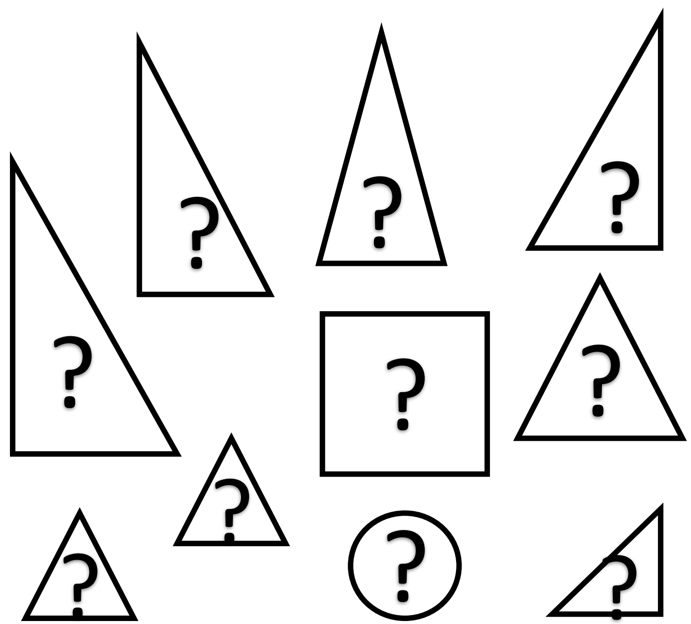
- Interpreting the value of a metric on a particular data set requires some thinking ...



- On this dataset, what would be the expected accuracy of a model that does NO learning

Evaluation Metrics:
what could possibly go wrong?

- Interpreting the value of a metric on a particular data set requires some thinking ...



5. Misleading interpretation of a metric value!

What could possibly go wrong?

- Bad feature representation
- Bad data + misleading correlations
- Noisy labels for training and testing
- Bad learning algorithm
- Misleading evaluation metric

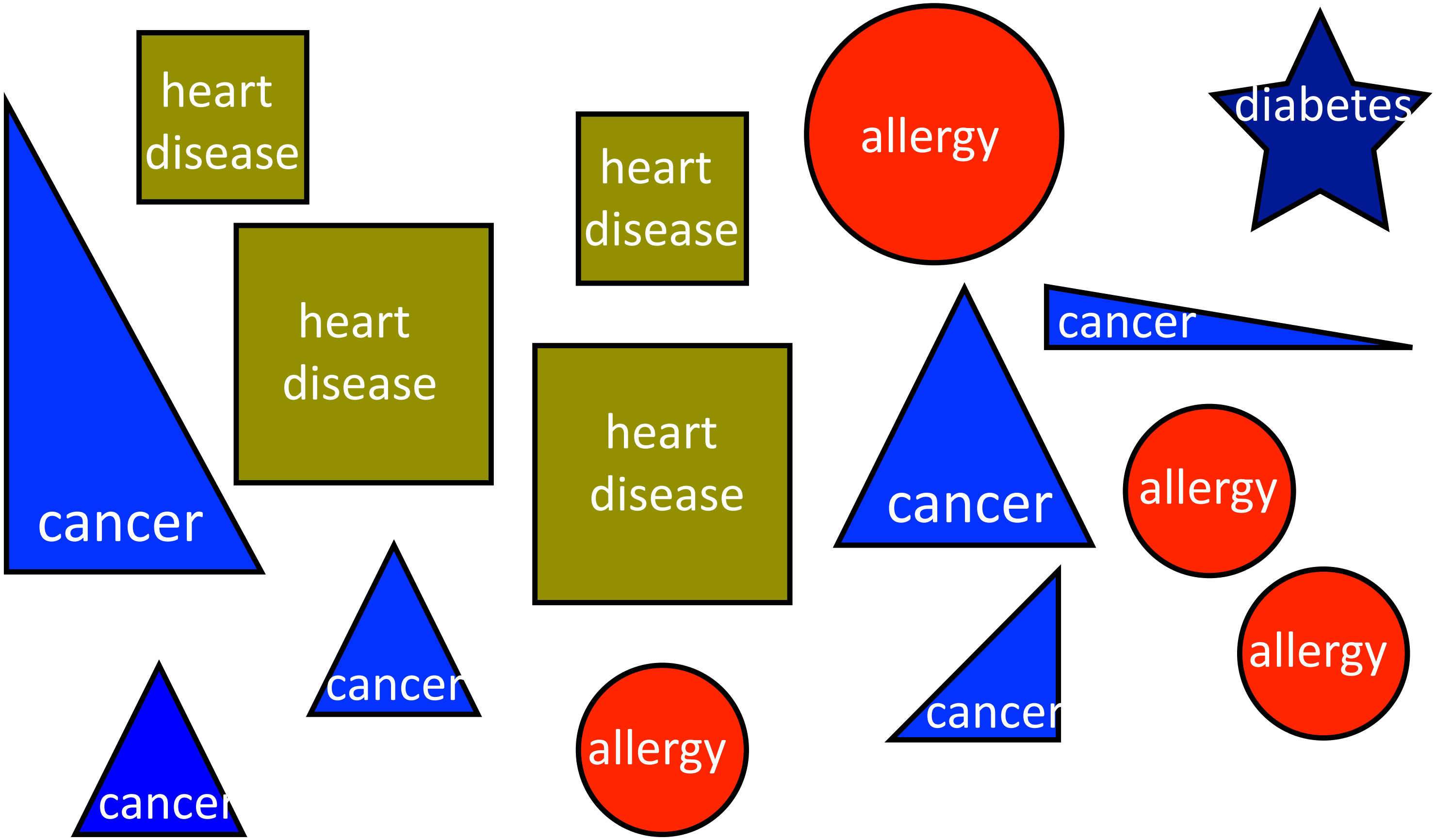
Exploratory Analysis of Text: The Big Picture

Two Paradigms in Text Mining

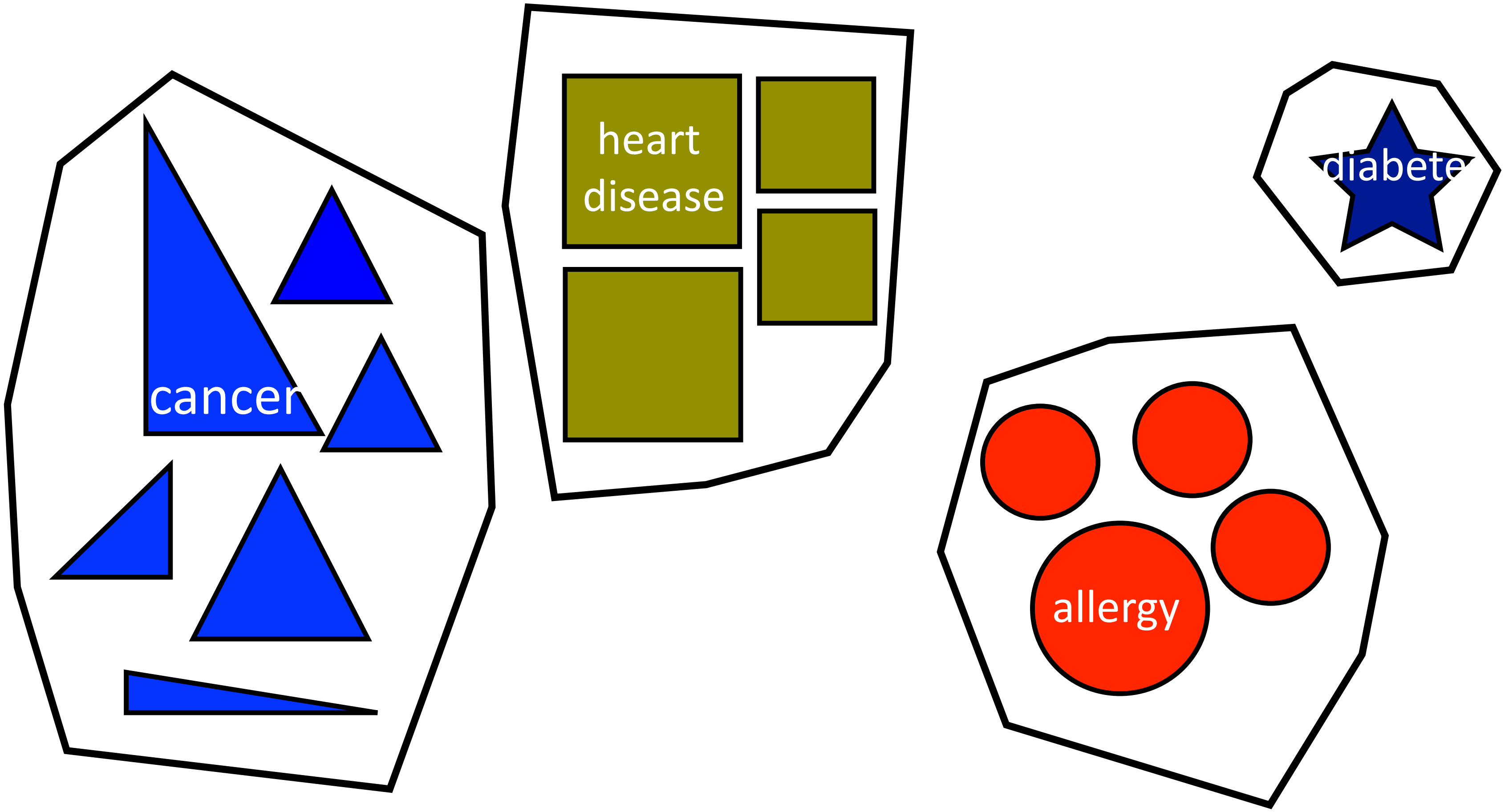
- Predictive Analysis of Text
  - developing computer programs that automatically predict a particular concept within a span of text
- Exploratory Analysis of Text
  - developing computer programs that automatically discover interesting and useful patterns or trends in text collections

Exploratory Analysis

example: clustering shapes

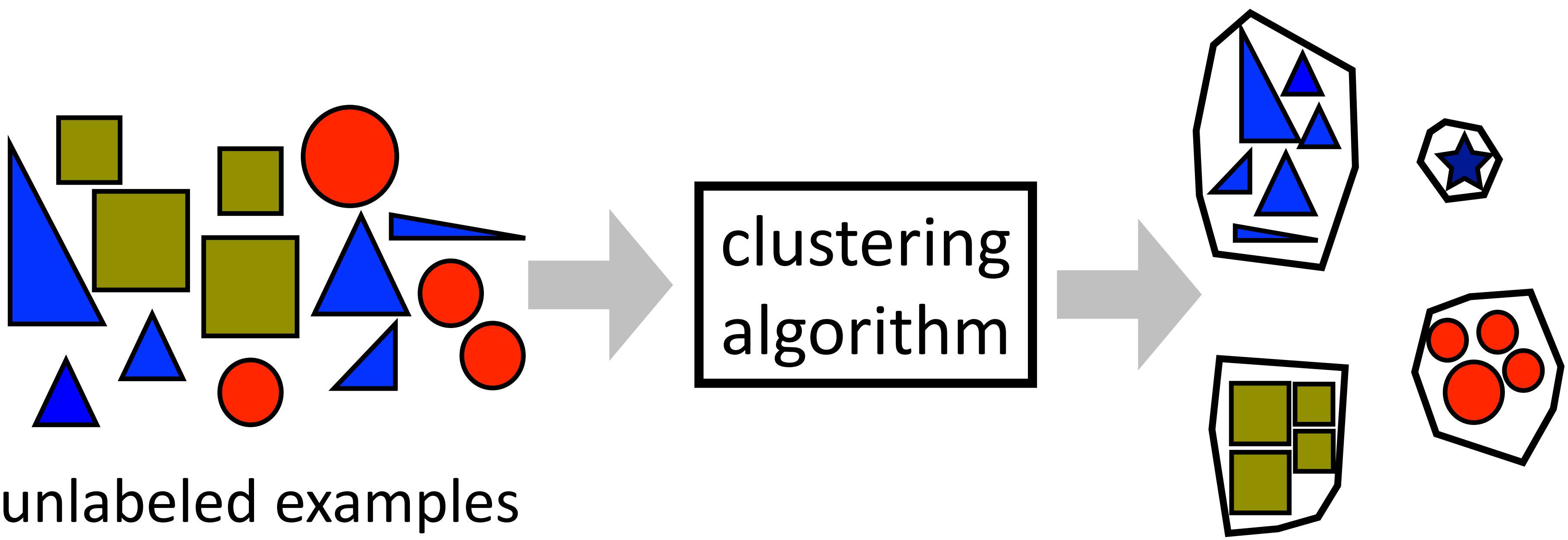


Exploratory Analysis
example: clustering shapes



Exploratory Analysis

example: clustering shapes



unlabeled examples

Exploratory Analysis
representation: features

No label

color	size	# slides	equal sides	...	shape
blue	big	3	no	...	triangle
blue	big	3	yes	...	triangle
red	small	inf	yes	...	circle
green	small	4	yes	...	square
⋮	⋮	⋮	⋮	⋮	⋮
Blue	big	3	yes	...	triangle

Exploratory Analysis
representation: features

color	size	# slides	equal sides	...
blue	big	3	no	...
blue	big	3	yes	...
red	small	inf	yes	...
green	small	4	yes	...
⋮	⋮	⋮	⋮	⋮
Blue	big	3	yes	...

Exploratory Analysis basic ingredients

- **Data**: a set of examples that we want to automatically analyze in order to discover interesting trends
- **Representation**: a set of features that we believe are useful in describing the data (i.e., its main attributes)
- **Similarity Metric**: a measure of similarity between two examples that is based on their feature values
- **Clustering algorithm**: an algorithm that assigns items with similar feature values to the same group

Any Questions?

Predictive Analysis of Text: Concepts, Features, and Instances

Next Class