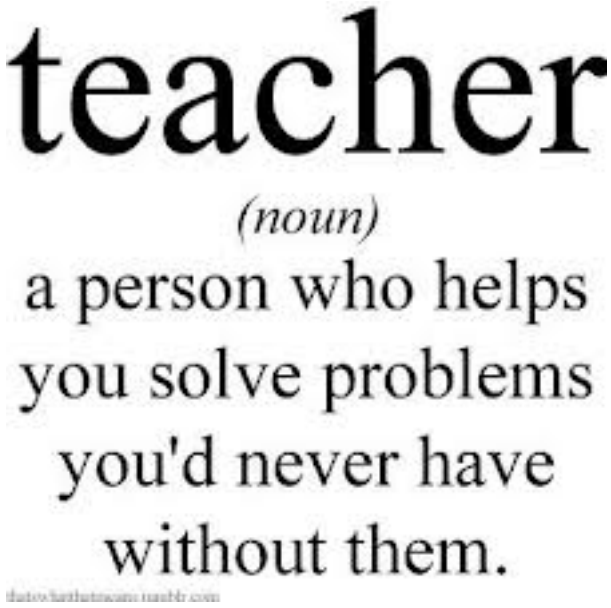
ISTA 455/555
Applied Natural Language Processing

Mihai Surdeanu

Spring 2015

Intros

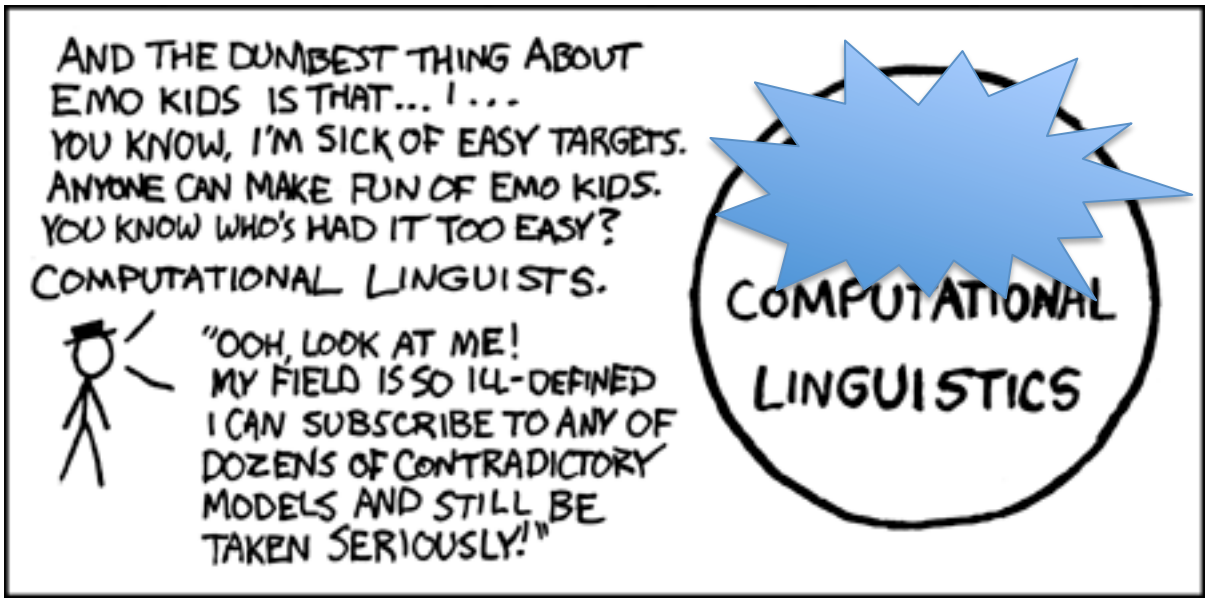

teacher
(noun)
a person who helps
you solve problems
you'd never have
without them.

Survey

- Please give your honest opinion about the knowledge you have on the topics listed in the survey.
- Your answers will not be held against you in any way. They are meant to help the instructor structure the material to better serve you.

WHY YOU SHOULD TAKE THIS COURSE

What the world (well, xkcd) thinks about NLP

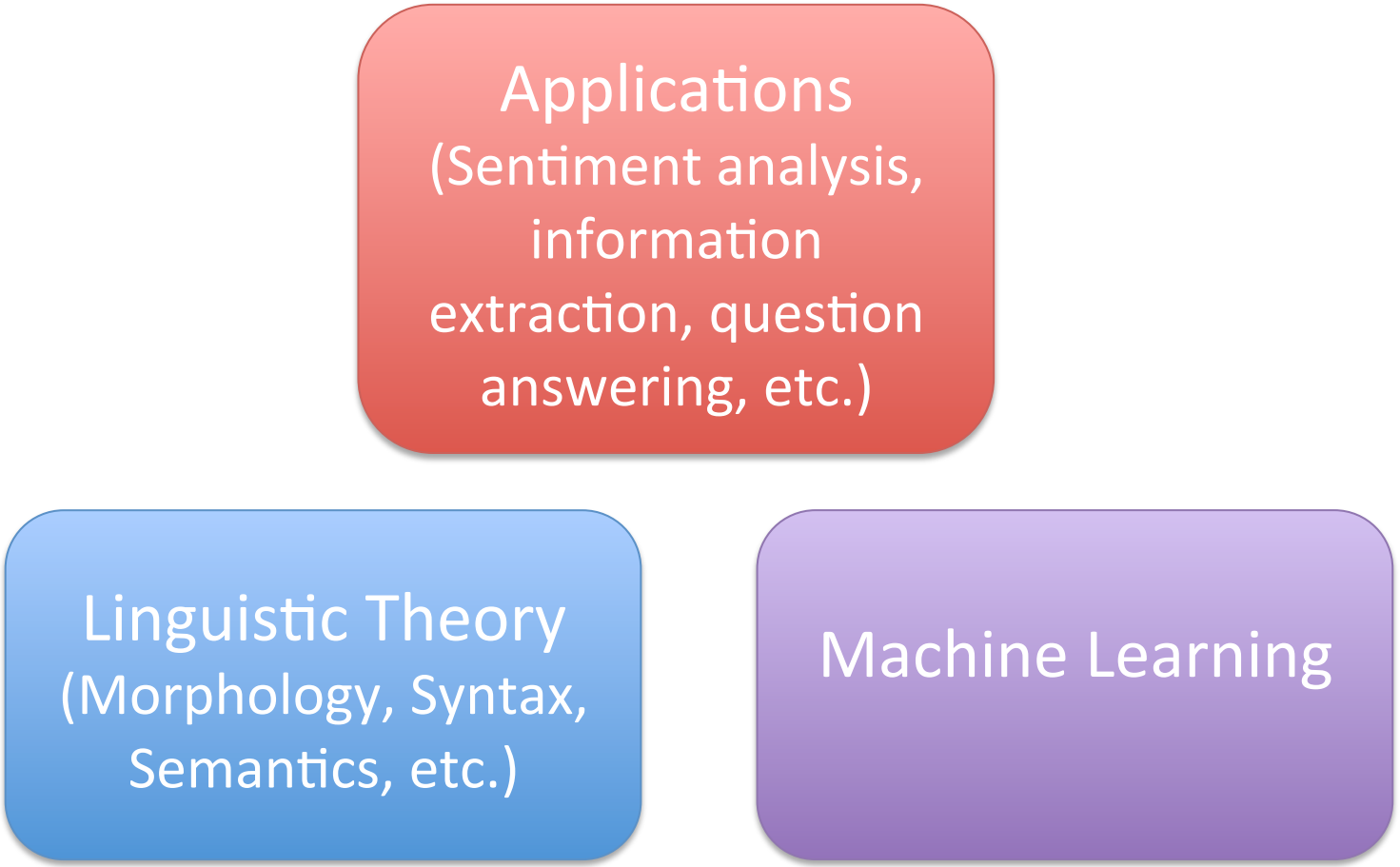


Uncensored version: http://xkcd.com/114/

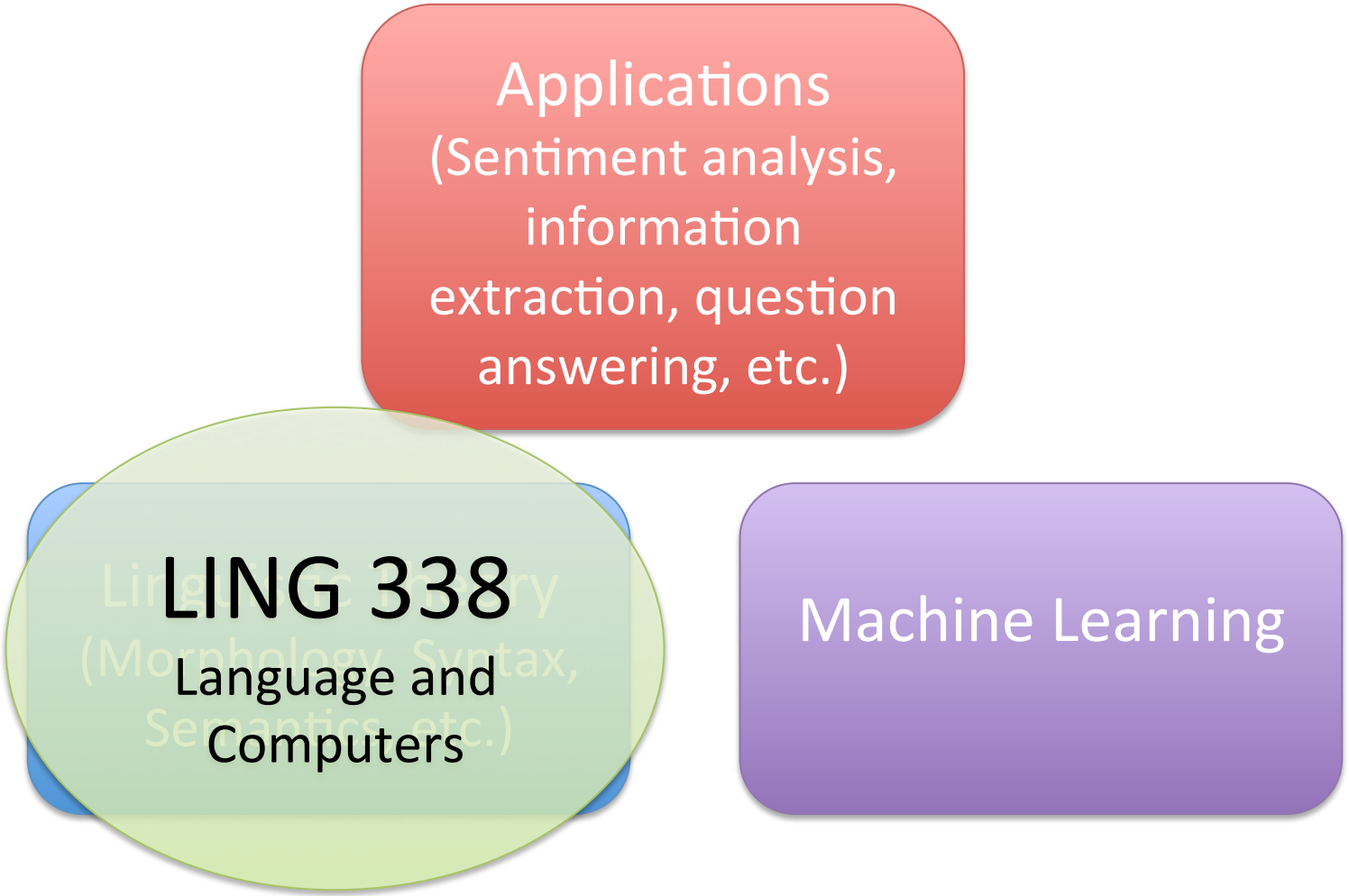
This course will prove that…

- NLP is not ill defined ☺
- NLP is extremely useful in the real world
  - Better jobs
  - Fun jobs
  - Meaningful jobs

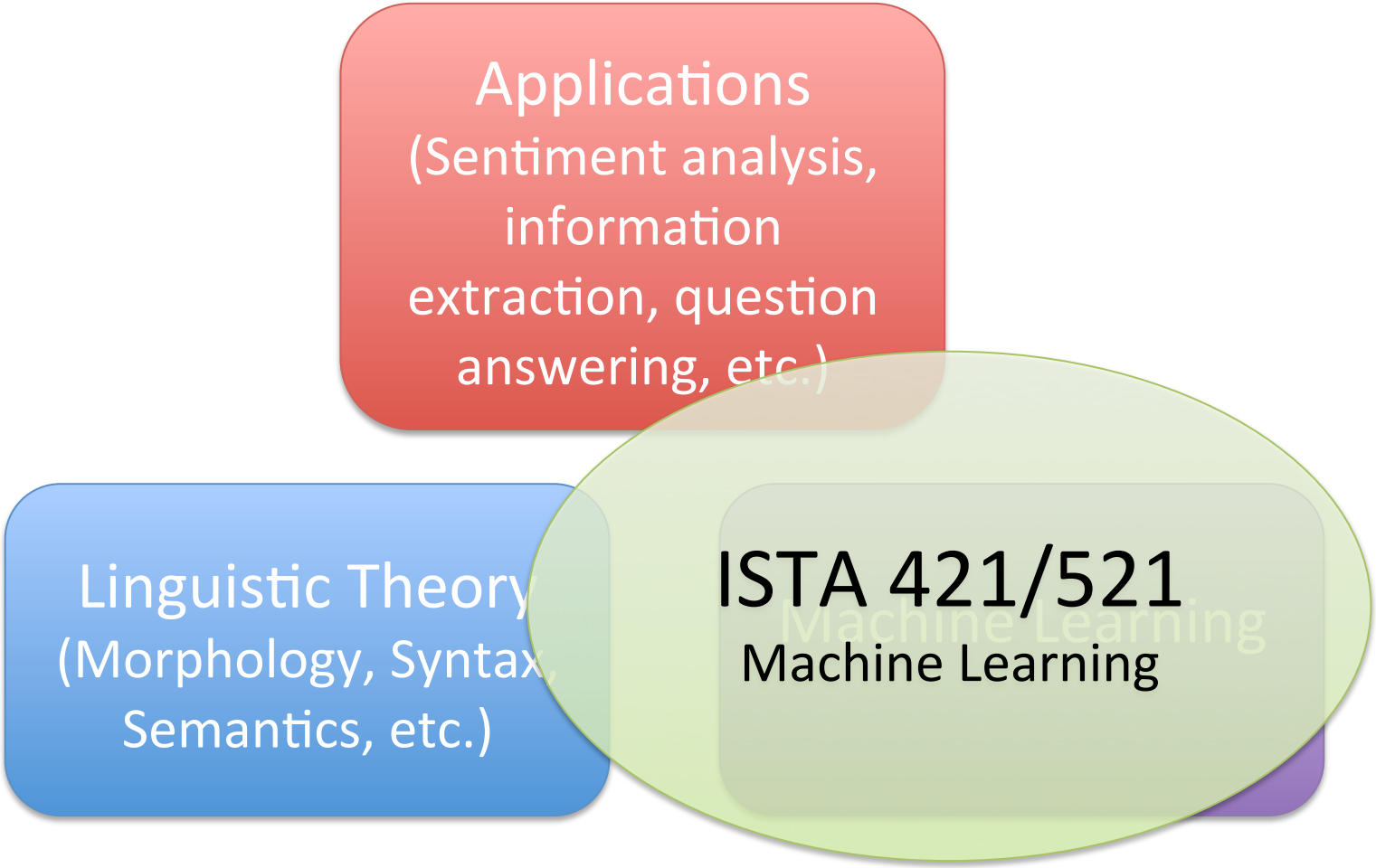
A simplified view of NLP



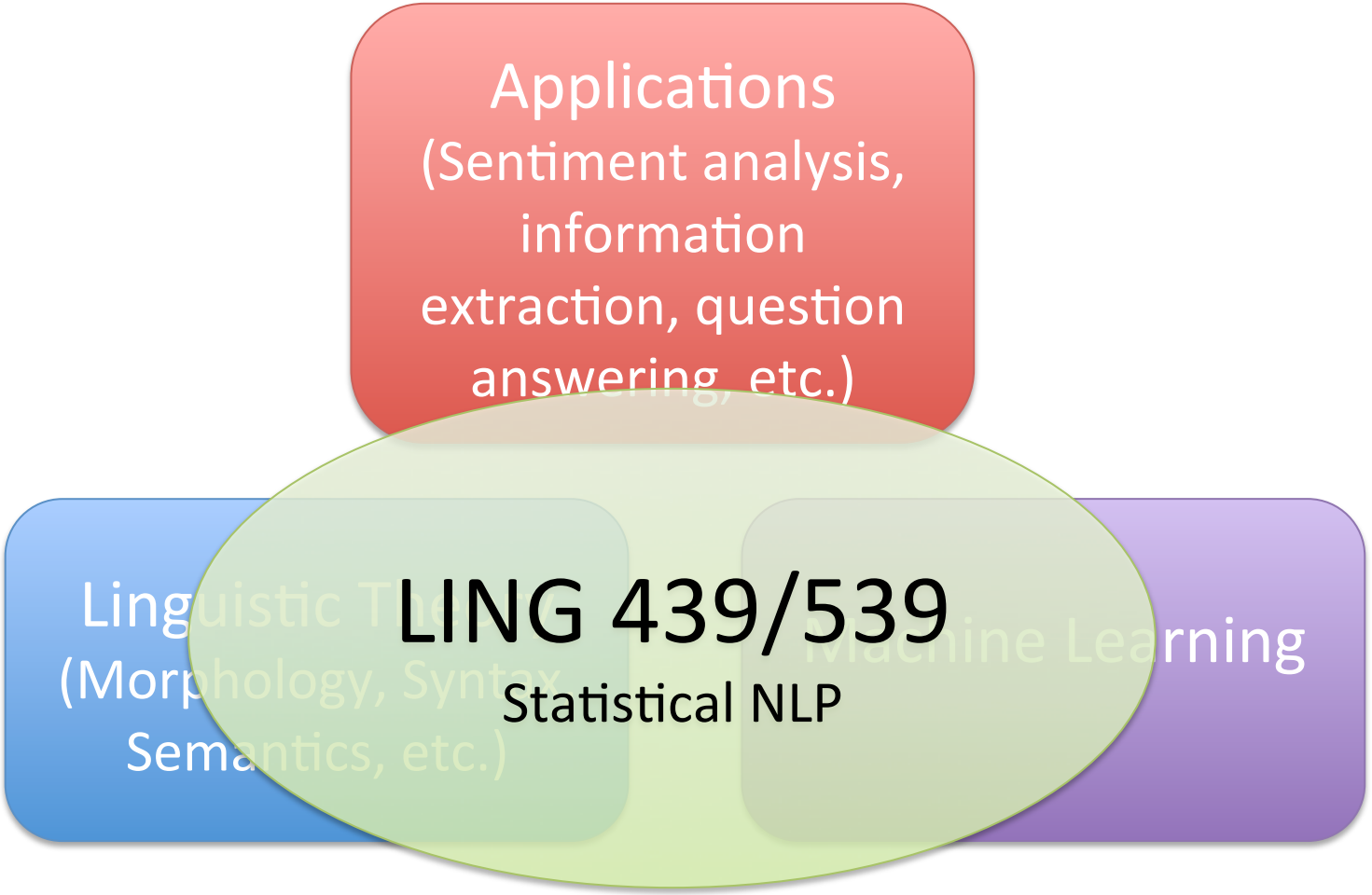
Other UA courses



Other UA courses



Other UA courses



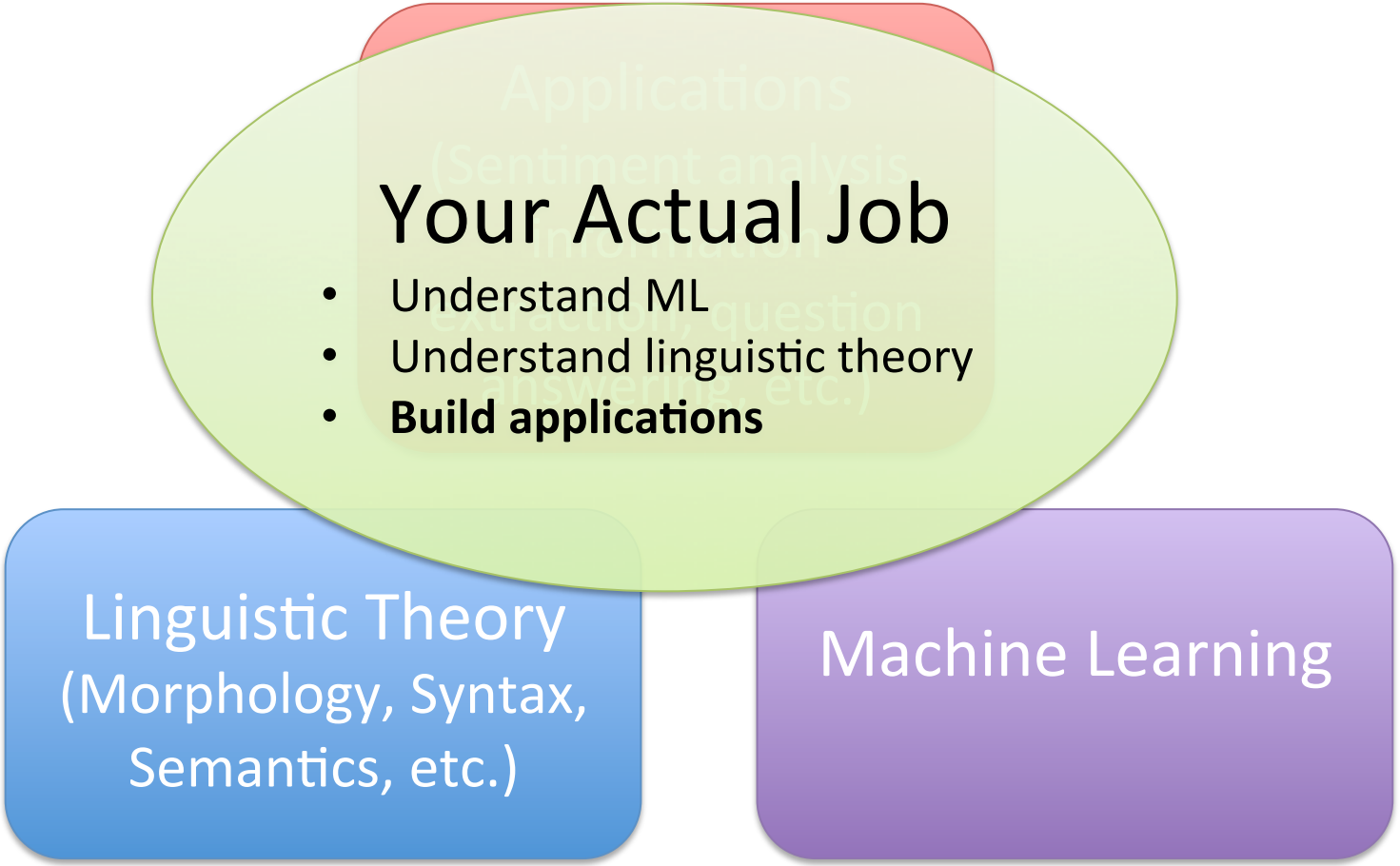
Ultimately, you should know all these topics. But…

Applications
(Sentiment analysis, information extraction, question answering, etc.)

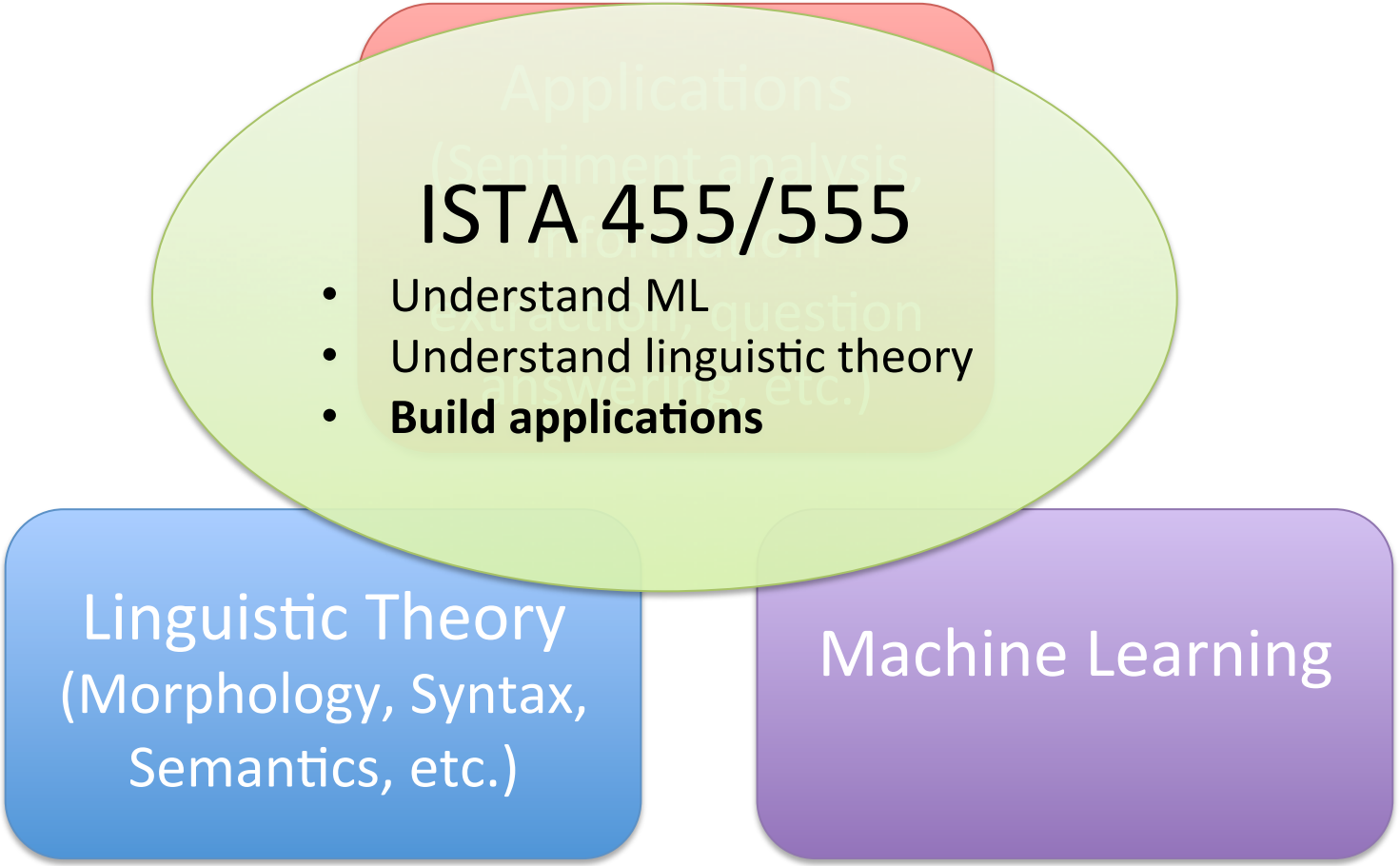
Linguistic Theory
(Morphology, Syntax, Semantics, etc.)

Machine Learning

Your job is likely to be in this space



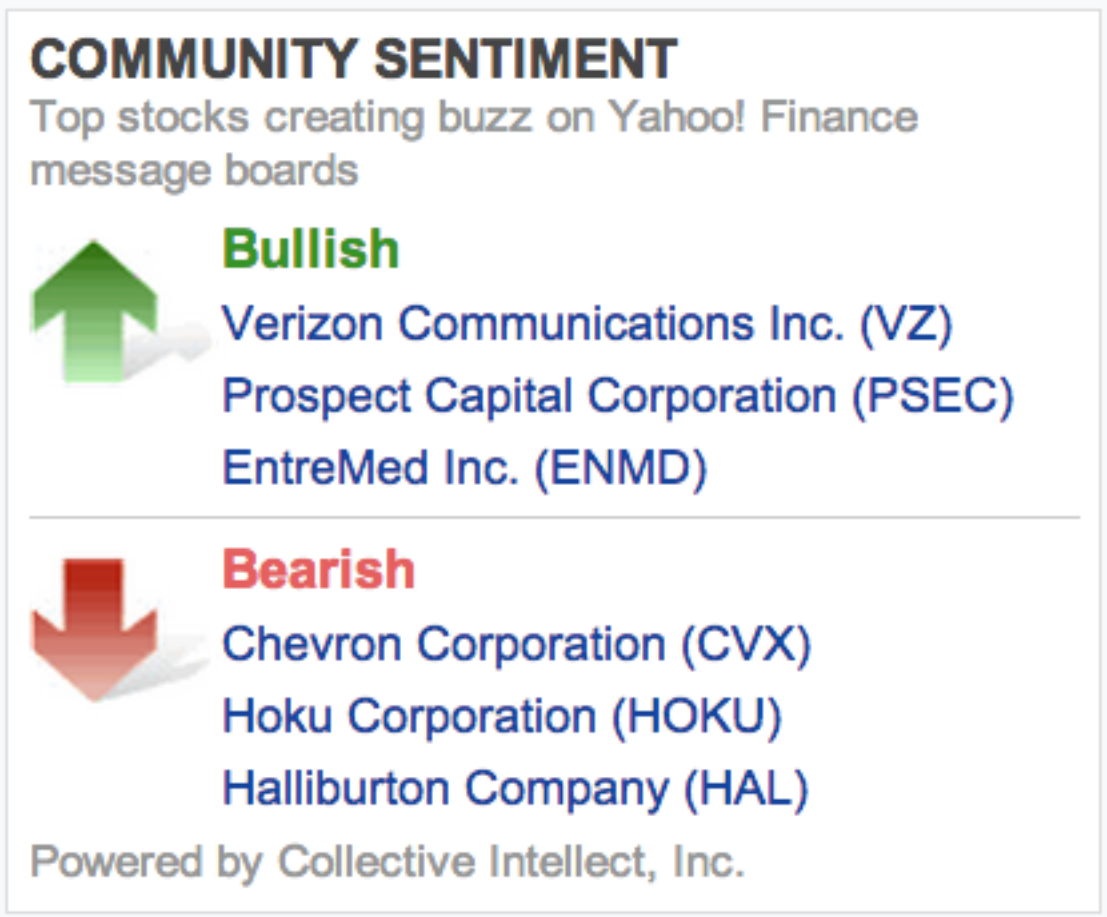
This course


ISTA 455/555
• Understand ML
• Understand linguistic theory
• **Build applications**
Linguistic Theory
(Morphology, Syntax,
Semantics, etc.)
Machine Learning

NLP Applications
Sentiment Analysis

- Voice of the voter
- Capital markets modeling/voice of the market
- Voice of the customer
- Voice of the employee
- Brand reputation management

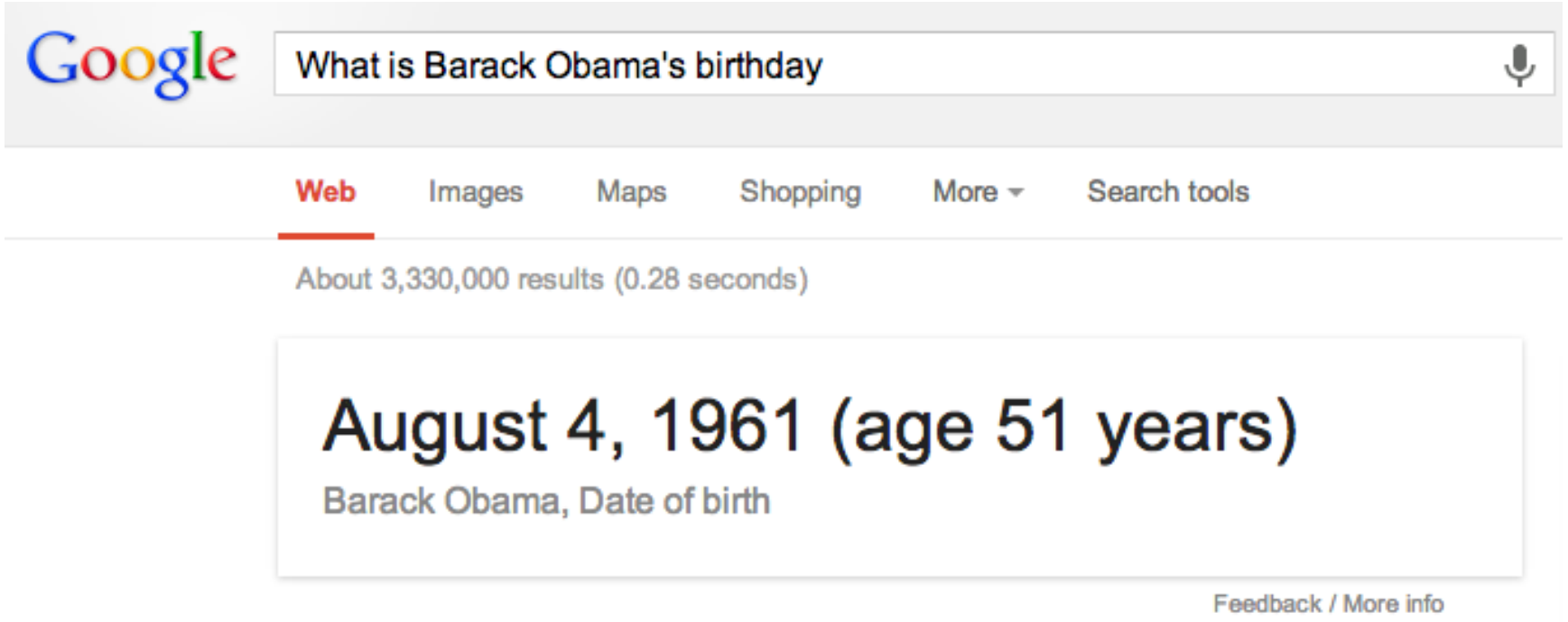
NLP Applications
Sentiment Analysis



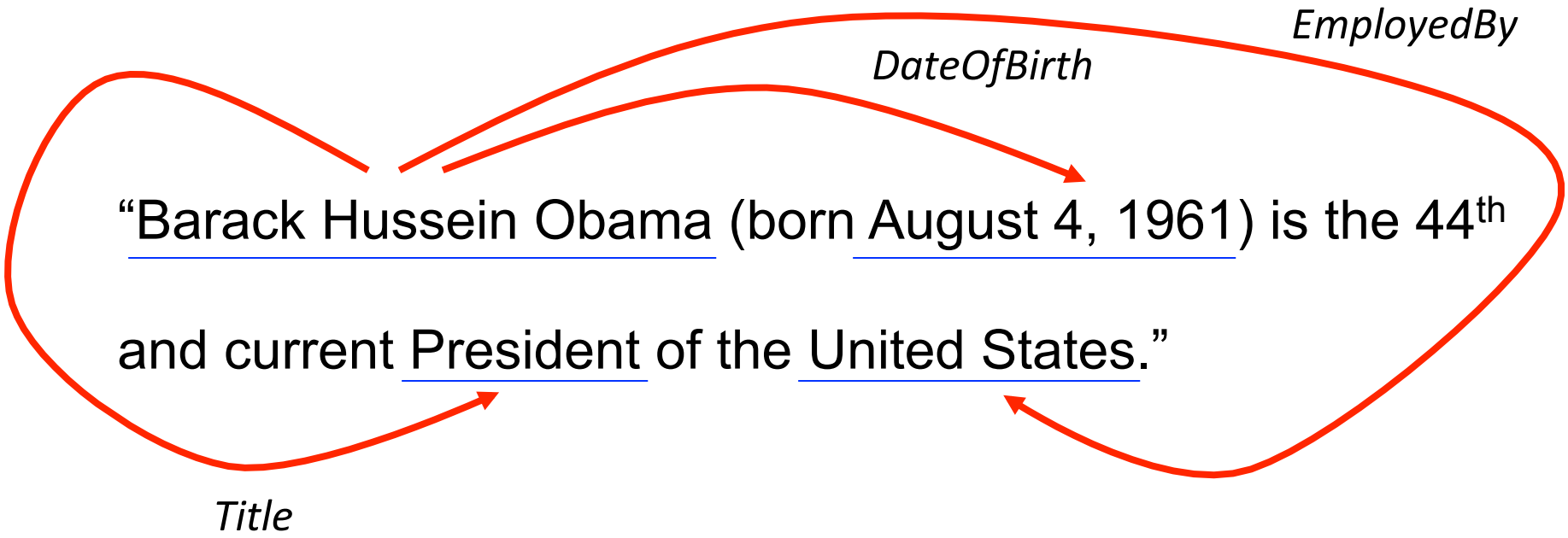
NLP Applications
Question Answering

NLP Applications
Question Answering

NLP Applications
Question Answering

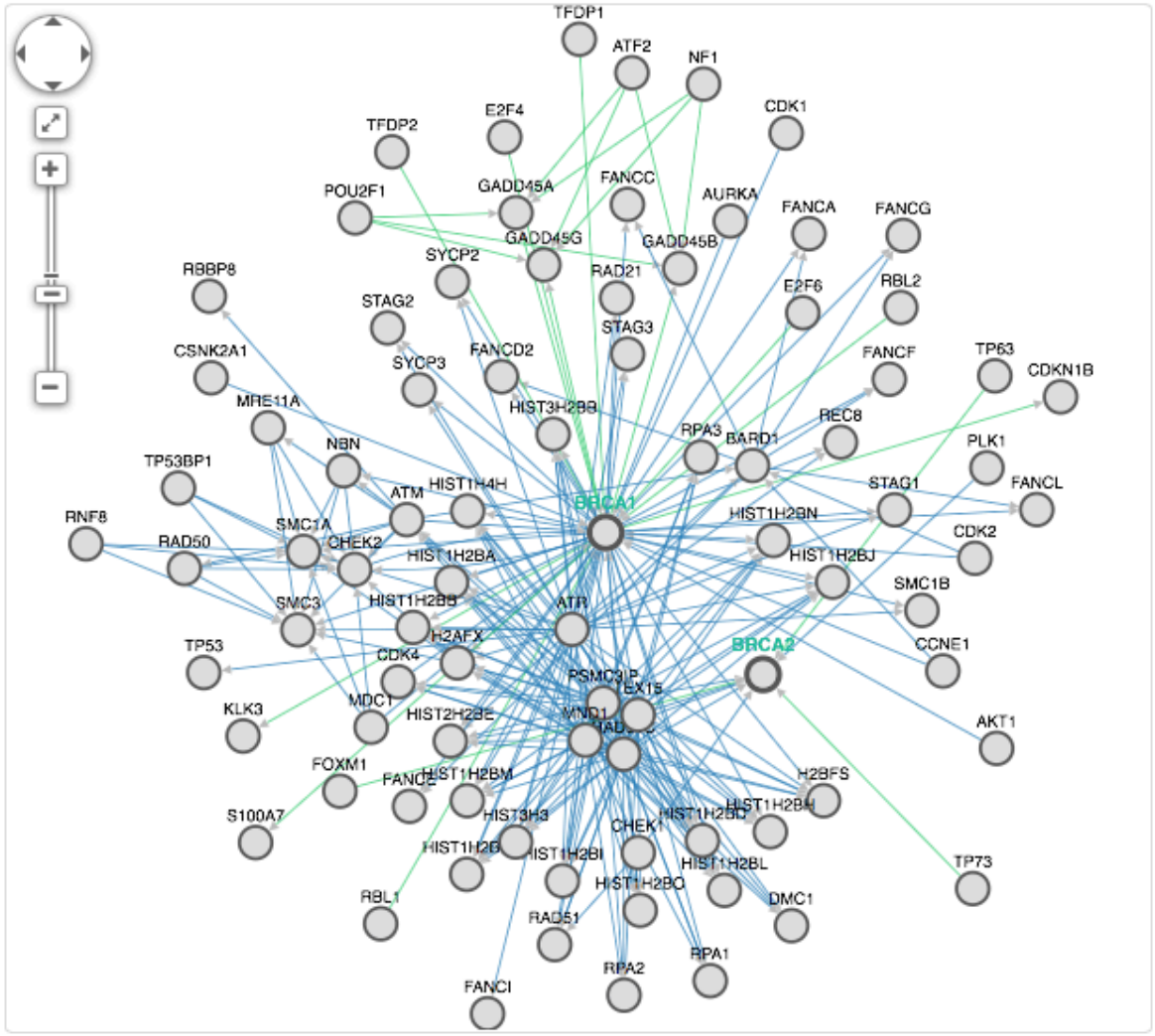


NLP Applications
Information Extraction

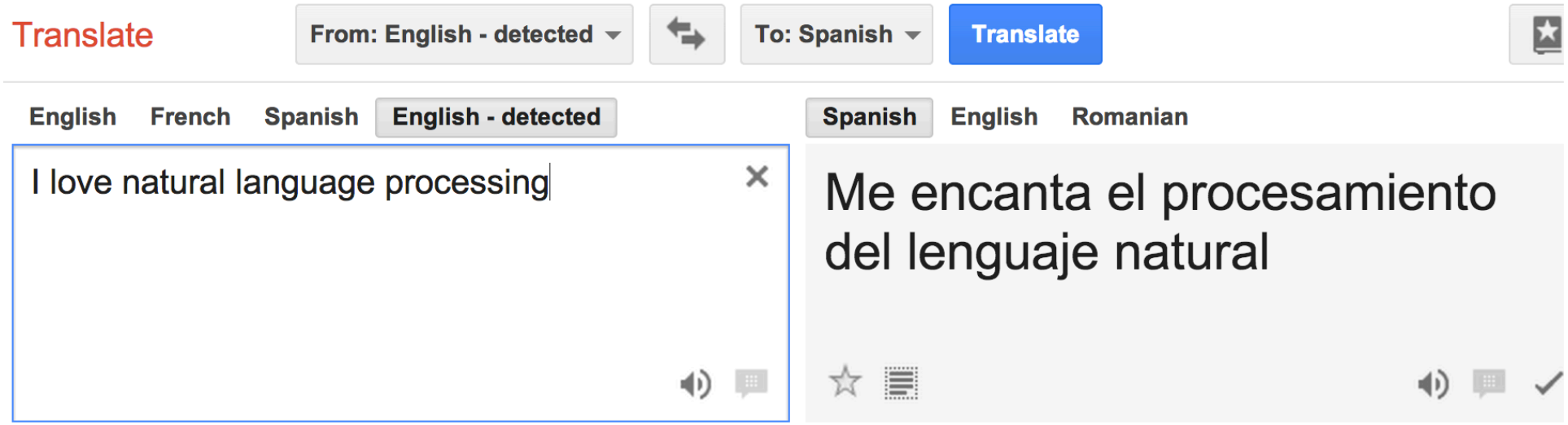


NLP Applications
Information Extraction

NLP Applications
Information Extraction



NLP Applications
Machine Translation



NLP Applications
Health Informatics

Goal

- Teach NLP through the **design and implementation of real-world applications**
  - Sentiment analysis
  - Information extraction
  - Question answering
  - Information retrieval

Prerequisites

- **Two programming courses** at the level of ISTA 130 or higher
- Recommended: one NLP course (LING 338 or LING 439/539) and one ML course (ISTA 421/521).
  - If you don't have any of the recommended courses, you will need the instructor's permission to continue.

I assume you know

- How to program
- How to write regular expressions
- What are the following (although we will review all this next)
  - POS tagger
  - NER
  - Syntactic parser (constituency-based, dependency-based)
  - Coreference resolution
  - Supervised classification

ESSENTIAL INFORMATION

Location and time

- Lectures
  - Tuesday/Thursday: 9:30 – 10:45
  - Social Sciences, Room 308
- Laboratory
  - NO LAB this semester

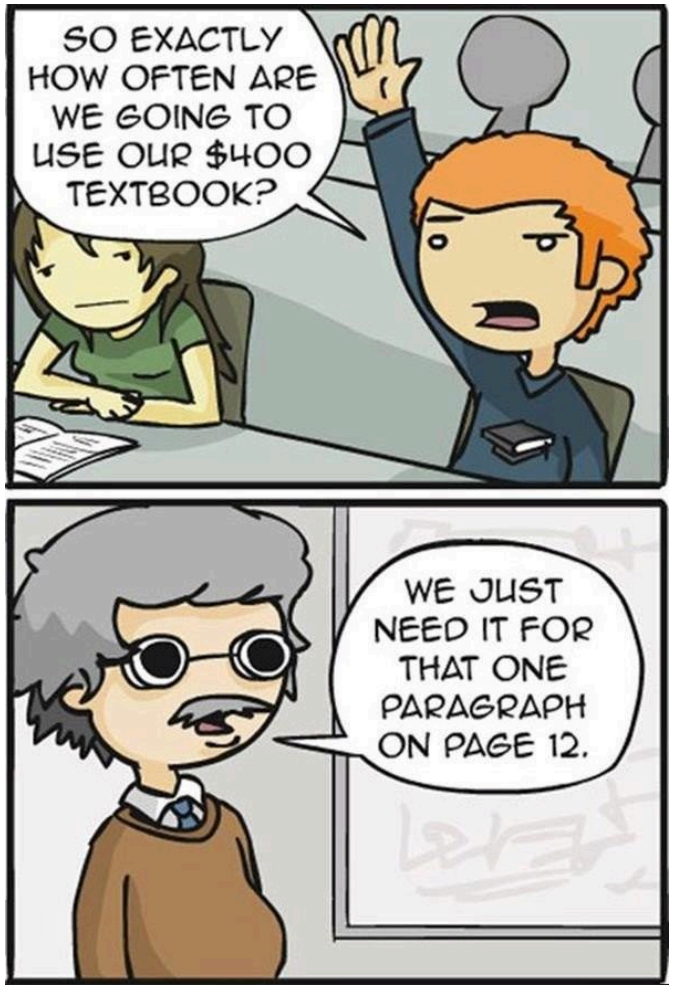
Instructor

- Mihai Surdeanu
- Email: msurdeanu@email.arizona.edu
- Office: Gould-Simpson 811
- Office hours: by appointment

Course information

- http://surdeanu.info/mihai/teaching/ista555-spring15/
  - Syllabus
  - Readings
  - First lecture
  - First project
- D2L
  - All lectures
  - Project descriptions
  - Grades

Textbooks


SO EXACTLY HOW OFTEN ARE WE GOING TO USE OUR $400 TEXTBOOK?
WE JUST NEED IT FOR THAT ONE PARAGRAPH ON PAGE 12.

Textbooks

- This course does not follow any particular NLP or ML book, but several are recommended below.
- You should read at least one NLP and one ML book.
- Most of the lectures will follow research papers, which will be posted on the website.

Recommended books

- Daniel Jurafsky and James H. Martin. 2008. *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Second Edition*. Pearson/Prentice-Hall. http://www.cs.colorado.edu/~martin/slp.html (in the bookstore)
  - Most recent NLP textbook. Not free…
- Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. 6th printing with corrections, 2003. The MIT Press. http://nlp.stanford.edu/fsnlp/ (available for **free** electronically through UA library)
  - Still a great foundational book.
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press. Available for **free** at http://nlp.stanford.edu/IR-book/
  - Excellent coverage of IR. We'll use it in the last part of the class.

Recommended books

- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer. Available for **free** at http://www-stat.stanford.edu/~tibs/ElemStatLearn/
  - Great introductory ML book
- Christopher M. Bishop. 2009. *Pattern Recognition and Machine Learning*. Springer.
  - Advanced ML. Not free…

Recommended books

- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media. Available for **free** at http://nltk.org/book/
  - Used for most programming in this class
- Applied machine learning in Python with scikit-learn. Available for **free** at http://scikit-learn.github.io/scikit-learn-tutorial/
  - Used for most ML in this class

Getting the Manning and Schütze book

- Click here: http://bit.ly/16fkUhS
- Click “Show all links from other libraries”
- Click “cognet.mit.edu”

Grading

- No written assignments and exams
- **4 programming projects**. Each includes code and a written paper (mimicking a research publication)
- Two in-class presentations
  - Presentation 1: 10 minutes; project 2 or 3
  - Presentation 2: 20 minutes: final project

Projects


DO YOU HAVE AN IDEA FOR YOUR PROJECT YET?
NO, I'M WAITING FOR INSPIRATION.
YOU CAN'T JUST TURN ON CREATIVITY LIKE A FAUCET. YOU HAVE TO BE IN THE RIGHT MOOD.
WHAT MOOD IS THAT?
LAST-MINUTE PANIC.

Projects

- Project 1
  - Getting your feet wet
  - "Analyzing happiness"
- Project 2
  - Choice #1
    - Sentiment analysis
    - "Predicting movie ratings"
  - Choice #2
    - Predicting the quality of a presentation
- Project 3
  - Information extraction
  - "Relation extraction"
- Project 4 (final)
  - Your choice! (must validate with instructor)
  - Must be more complex than projects 1 – 3
  - Instructor will offer one or more backup projects

Grading

Component	Weight
Project 1	20 pts
Project 2	20 pts
Project 3	20 pts
Project 4 (final)	20 pts
Presentation 1	5 pts
Presentation 2	10 pts
In-class participation	5 pts
Total	**100 pts**

Grade	Point range
A	90 – 100
B	80 – 89
C	70 – 79
D	60 – 69
E	0 – 59

Grade disputes

- Disputes about project grades will be entertained for 2 weeks from the day the project is due, or 1 day before grades are due, whichever is sooner.
- These will be resolved by re-grading the entire project. Note that this can result in a lower grade in the event that new mistakes are discovered.
- **No negotiations about individual students' letter grades will be entertained once final grades are assigned.**

Collaboration policy

- **Projects must be individually implemented**. Copying is not permitted and will be treated as academic dishonesty.
- Students are encouraged to discuss problems and general approaches to solutions.

Late policy

- Projects are due electronically via D2L by the stated deadline.
- Permission for an extension must be granted by the lab instructor in advance of the deadline in order to receive full credit for a late submission.
- The first request by a given student is likely to be granted; the probability decreases with each subsequent request.
- No project will be accepted once solutions are posted online.

455 vs. 555

- To differentiate between graduate and undergraduate students, the instructor will require graduate students to implement more complex, state-of-the-art algorithms for the assigned projects.
- Projects will be graded separately for undergraduate and graduate students, as described in each project's description.

Tentative schedule

Topic	Lectures	Topics
Introduction	1	Syllabus, first project!
NLP Overview	2	Tokenization, part-of-speech tagging, named entity recognition, parsing, coreference resolution
ML Overview	3	Probabilistic vs. geometric models, ensemble models, avoiding overfitting, statistical significance
Sentiment Analysis	4	Lexicon-based methods; supervised classification; distant supervision; using sentiment analysis for event forecasting
Information Extraction	6	IE as text segmentation; relation extraction: distant supervision, rule-based, supervised, unsupervised; event extraction: rule-based, supervised, unsupervised
Question Answering	6	Factoid QA; non-factoid QA; web-based QA
Optional Topics	6	Summarization, topic modeling, distributional similarity, machine translation, deep learning

Project and Presentation Deadlines

Assignment	**Due Date**
Project 1	February 1
Project 2	February 22
Presentation 1	**February 26**
Project 3	March 29
Project 4	April 26
Presentation 2	**April 30 and May 5**

University policies

- Classroom behavior
  - You are encouraged to ask/answer questions
  - Class participation credit will be assigned based on this
- Attendance
  - Is mandatory-ish
  - Students who miss more than 1/3 of classes may be dropped
- For more, see syllabus

CHOOSING A PROGRAMMING LANGUAGE

My recommendations

- Pick one of these
  - Python
  - Scala
  - Java
  - C/C++
- The “official” programming language in this class will be Python
  - But I don’t use it for my own research…

Python

- Pros
  - Clean syntax
  - Popular: many NLP/ML libraries exist
  - Clean exception handling
- Cons
  - Slow
  - Dynamically typed
  - No great IDE

Java

- Pros
  - Pretty fast
  - Probably the most common language for NLP
  - Clean exception handling
  - Statically typed
  - Several great IDEs
- Cons
  - Syntax too verbose
  - Inconsistent semantics due to enforced backwards compatibility (primitive types vs. objects, equality, etc.)

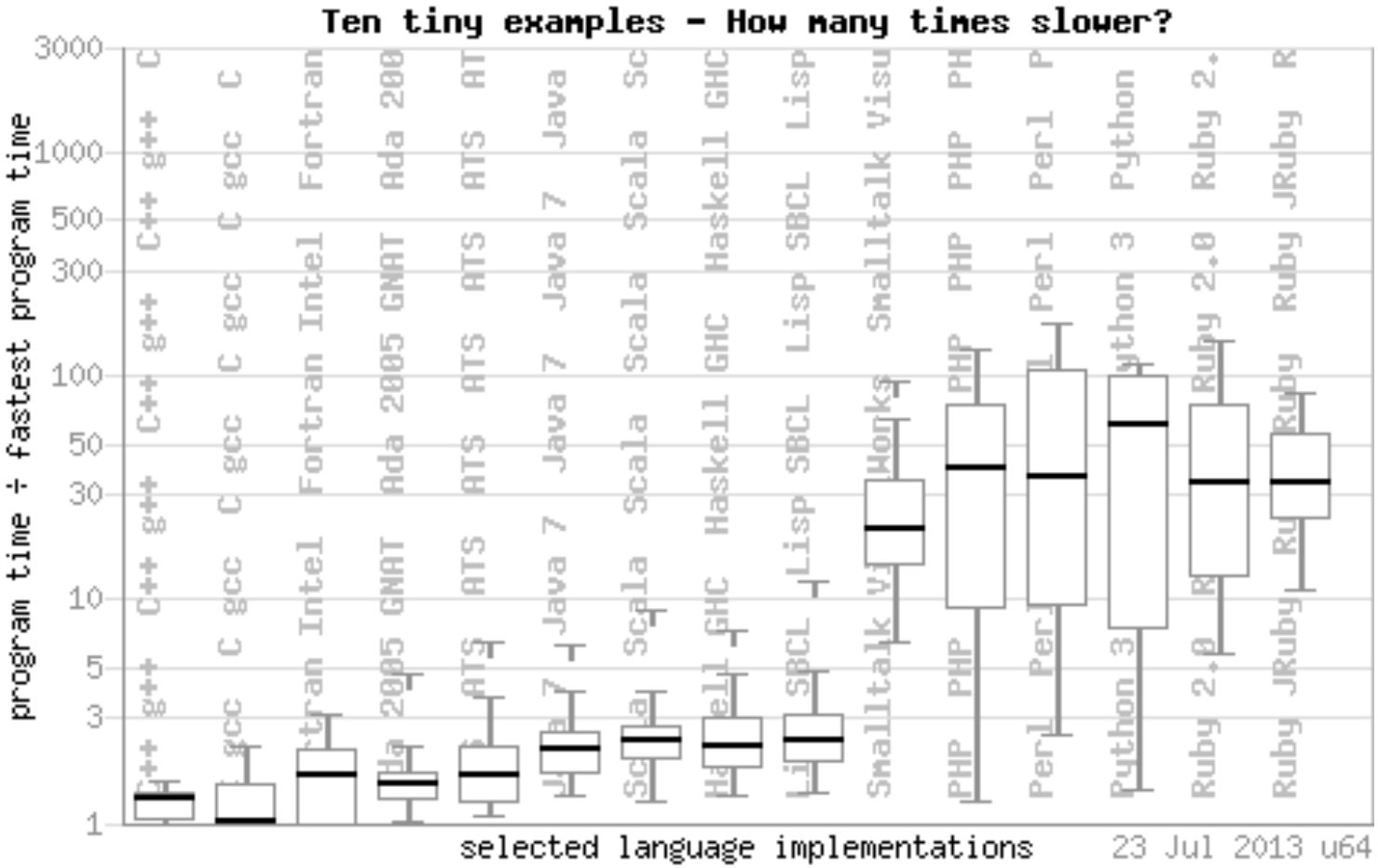
Scala

- Pros
  - Pretty fast
  - “Hot” language for NLP, ML, web development
  - Clean exception handling
  - Clean syntax
  - Consistent semantics
  - Statically typed
  - At least one great IDE
  - Fully compatible with Java (use all Java libraries)
- Cons
  - It has some “dark corners”
  - Backwards compatibility not guaranteed

C/C++

- Pros
  - Fast
- Cons
  - Too many to list

Comparison



More benchmarks: http://benchmarksgame.alioth.debian.org/u64/benchmark.php?test=all&lang=all&data=u64

THINGS TO DO NOW

Project 1 due in two weeks!