BUS 243

Lecture 1: Logistics / Information Retrieval

COURSE LOGISTICS

- Instructor: Yeabin Moon, Ph.D.
- Office hours: After class or online
  - check the syllabus
- Latte Page is the main correspondence
- Lecture materials uploaded before each lecture

LEARNING GOALS

- This is not a full-fledged NLP course
  - CS / Linguistic department offer it
- Want to lean NLP skills in business & finance settings
  - Don’t need to have a degree in statistics to apply stats to your tasks
  - How about NLP?

WHAT DO YOU WANT TO LEARN?

- In statistics / econometrics
  - Learn a regression technique to analyze the elasticity of demand
- In NLP
  - Learn _____ to analyze _____?

WHAT IS YOUR DATA?

- What industry are you interested in?
  - Finance
  - Entertainment and Sports
  - Living
- NLP is necessary?
  - Let's see the latest 10-Q report for Apple: Link

QUESTIONS

- The text has more (or better) information than numbers?
  - Or opposite?
- If so, how could we extract information from text?
  - Wait, do machines understand text?
- Why Machine Learning?
  - Deep learning?

THE SCOPE OF THE COURSE

- Again, this is not the typical introductory course for NLP
- Delve into the text categorization task
- Early part of the class covers the text representation, and then move on to the classification task
  - Frequency-based representation vs. word embeddings
  - Traditional ML models: Naïve Bayes and Logistic Regression
  - Deep learning application
  - Transfer learning
- Let's see the Syllabus

MORE LOGISTICS

- Readings are mandatory
  - Check the syllabus
- Weekly assignment: 55 %
  - 5 in total
  - No accepted more than a day late (10 % off a day late)
- Final exam: 40 % (August 10)
  - **Do not take the class if you are unable to attend**
- Participation / Attendance: 5 %

NOTE FOR ASSIGNMENTS

- Each assignment should be completed individually
- Strongly encourage you to work alone to gain the maximum
- Understand allowed collaboration and how to document it
  - Do not take code off from the web
  - **Must acknowledge** working with other students
  - Must write your own assignment solutions

Let's Begin

DEFINITION?

- NLP is a **principled** approach to processing human language
  - What does it mean?
- It is a subfield of artificial intelligence (AI) that refers to computational approaches to process, understand, and generate human language
  - This is pretty new
- Need to examine some definitions

WHAT DOES IT MEAN FOR A LANGUAGE TO BE NATURAL?

- You might wonder
  - Are there any unnatural languages?
  - Is English natural?
  - Is Spanish more natural than Korean?
- Another tricky term is a formal language
  - Is English formal?

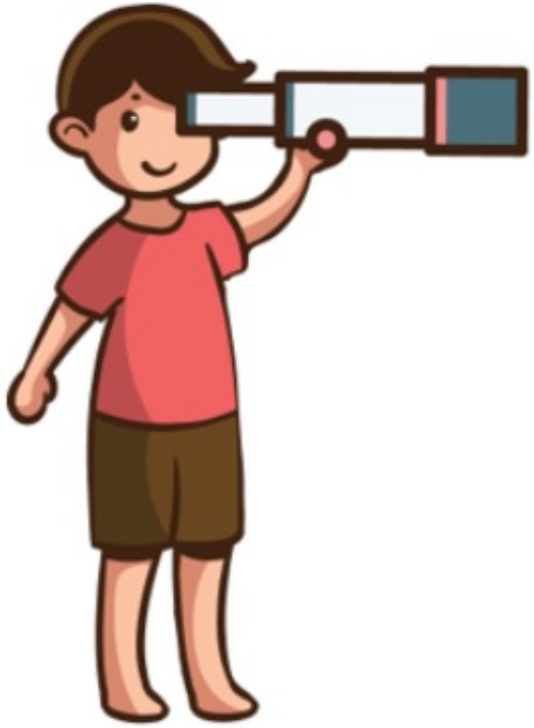
NATURAL VS. FORMAL

- The word *natural* is used to contrast natural languages with formal languages
  - all the languages humans speak are natural
- Formal languages are types of languages that are invented by humans
  - Have strictly and explicitly defined syntax (grammatical rules) and semantics (meaning)
    - Programming languages are examples
  - When you run a compiler on the code, you either get a syntax error or not
  - The behavior of your program is always the same if it's run on the same code

NATURAL LANGUAGE IS HARDLY FORMAL

- You can write a sentence that is *maybe* grammatical
  - Subjective, and worse, time varying
  - There are some grammar topics where even experts disagree with each other
- This is what makes human languages interesting but challenging, and why the entire field of NLP even exists
- Bottom line: human languages are ambiguous

HE SAW A GIRL WITH A TELESCOPE



HERE COMES A NEW CHALLENGER

- Now let's consider the following scenario and think how you'd approach this problem
- You are working as a junior data scientist at a midsized company
- You got a giant TSV file containing all the responses to the survey questions about the product from the marketing team:
    1. How did you know about our product?
    2. How do you like our product?
    3. A free-response question, where our customers can write whatever they feel about our product

- The marketing team realized there was a bug in the online system and the answers to the second question were not recorded in the database at all
- Your task is whether you could recover the lost data
- Fortunately, data structure is fairly standard
  - It has several fields such as timestamps and submission IDs
  - At the end of each line is a lengthy field for the free-response question

FIRST TRY

- Example responses:
  - A very good product!
  - Very bad. It crashes all the time!

```
def get_sentiment(text):
    """Return 1 if text is positive, -1 if negative.
       Otherwise, return 0."""
    if 'good' in text:
        return 1
    elif 'bad' in text:
        return -1
    return 0
```

AGAIN, NATURAL LANGUAGE IS AMBIGUOUS

- The code filtered a decent amount of data
- Alas, my code returns
  - I can't think of a single good reason to use this product: positive
  - Not bad: negative
- Right. Negation!

```
1  def get_sentiment(text):
2      """Return 1 if text is positive, -1 if negative.
3         Otherwise, return 0."""
4      sentiment = 0
5      if 'good' in text:
6          sentiment = 1
7      elif 'bad' in text:
8          sentiment = -1
9      if 'not' in text or "n't" in text:
10         sentiment *= -1
11     return sentiment
```

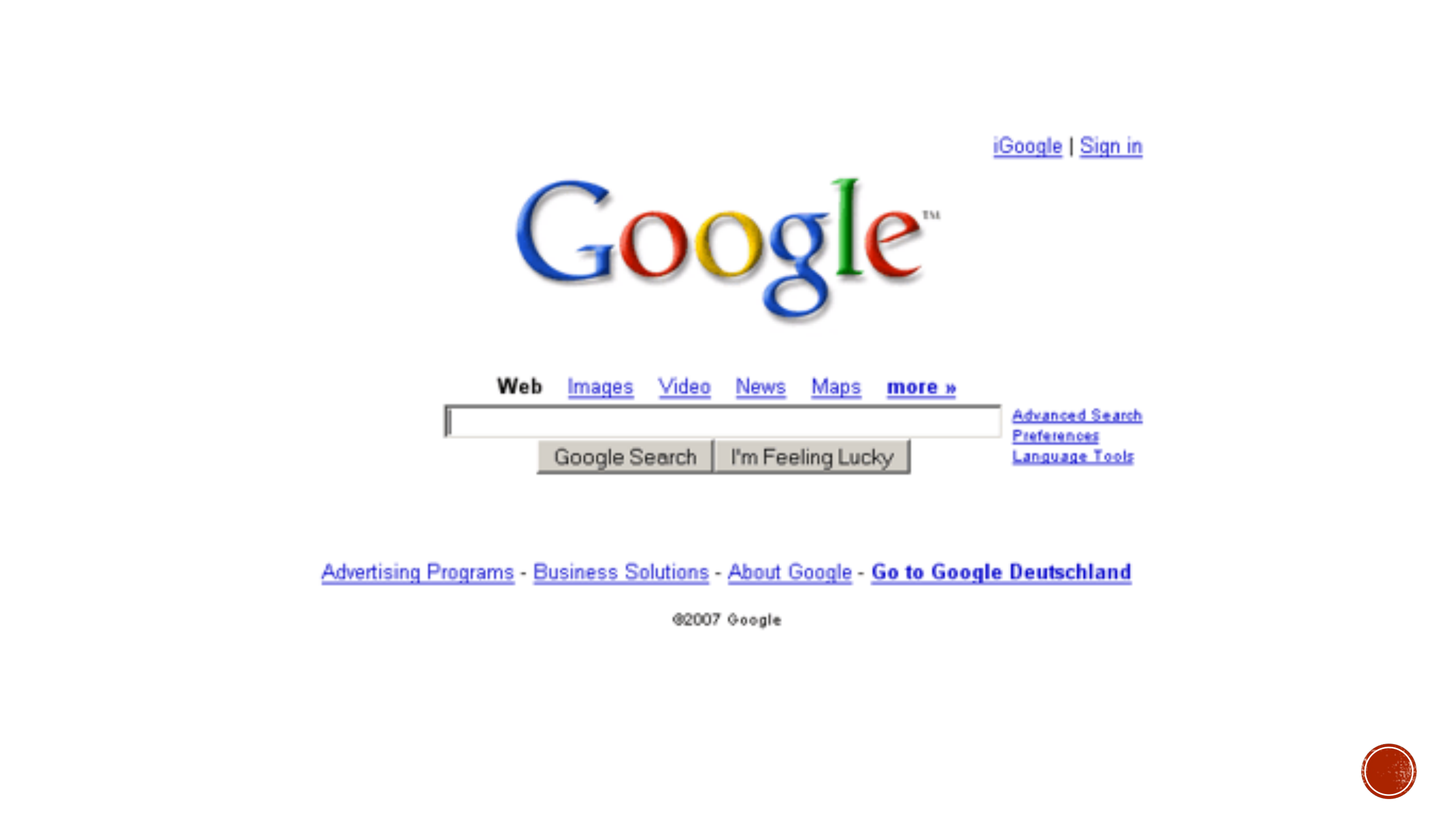
. . . .

- The product is not only cheap but also very good!: negative
- Worse
  - I always wanted this feature badly!
  - It's very badly made
- How could a single word in a language have two completely opposite meanings?
- This course will save you

THIS COURSE WILL SAVE YOU?

- What does it mean? Another ambiguity
- This course will save you because it will teach you either
  1. How to deal with the problems described above
  2. They are impossible to solve

Information Retrieval


iGoogle | Sign in
Google
Web Images Video News Maps more »
Advanced Search
Preferences
Language Tools
Google Search I'm Feeling Lucky
Advertising Programs - Business Solutions - About Google - Go to Google Deutschland
©2007 Google

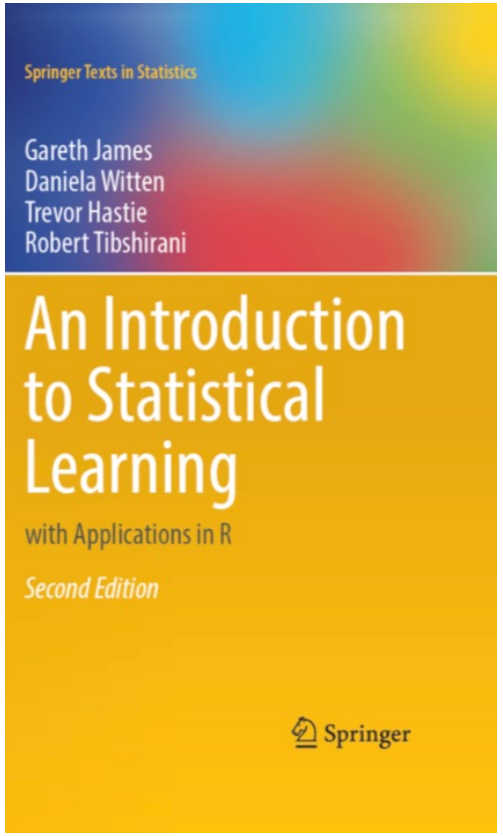
- Information retrieval can money
  - Search engines are the most visited websites in most countries
    - Google, Bing, Baidu, Yahoo, AOL, Naver
  - Discussion platforms
    - Reddit, Quora, Stack Exchange
  - Shared knowledge
    - Wikipedia
    - Have you heard of Britannica?

WE START HERE

- Why IR?
- Simple to understand
  - But hard to implement
- Important to the world

- *Information retrieval is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)*

- Suppose you want to learn *Gradient Descent*
- Suppose you love a textbook treatment



Control + F: *Gradient Descent*

Assumption?
Why is this preferred?

Comedy	History	Tragedy	Poetry
All's Well That Ends Well As You Like It The Comedy of Errors Cymbeline Love's Labours Lost Measure for Measure The Merry Wives of Windsor The Merchant of Venice A Midsummer Night's Dream Much Ado About Nothing Pericles, Prince of Tyre Taming of the Shrew The Tempest Troilus and Cressida Twelfth Night Two Gentlemen of Verona Winter's Tale	Henry IV, part 1 Henry IV, part 2 Henry V Henry VI, part 1 Henry VI, part 2 Henry VI, part 3 Henry VIII King John Richard II Richard III	Antony and Cleopatra Coriolanus Hamlet Julius Caesar King Lear Macbeth Othello Romeo and Juliet Timon of Athens Titus Andronicus	*The Sonnets* *A Lover's Complaint* *The Rape of Lucrece* *Venus and Adonis* *Funeral Elegy by W.S.*

- Which plays of Shakespeare contain the words *Brutus* AND *Caesar* AND NOT *Calpurnia?*

- Let's index each play
- Use *Terms* (think it as a word for now)

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	…
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
…							

- This is called a term-document incidence matrix

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	…
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
…							

- To answer the question:
  - 110100 AND 110111 AND 101111 = 100100
  - What's the answer?

BOOLEAN RETRIEVAL

- The Boolean retrieval model is an IR model
  - Formulate queries using **Boolean expressions**
    - AND, OR, and NOT
    - Can be arbitrarily nested
  - The model views each document as just a set of words
  - Any given query divides the collection into two sets: retrieved, not-retrieved
    - Pure Boolean systems do not define an ordering of the results
      - What does it mean?

STRENGTHS AND WEAKNESSES

- Strengths
  - Precise, if you know the right strategies
  - Precise, if you have an idea of what you're looking for
  - Implementations are fast and efficient
- Weaknesses
  - Users must learn Boolean logic
  - Boolean logic insufficient to capture the richness of language
  - No control over size of result set: either too many hits or none
  - When do you stop reading? All documents in the result set are considered "equally good"
  - What about partial matches? Documents that "don't quite match" the query may be useful also

TERMINOLOGY ALERT

- Write your own definition and list the corresponding examples in the Shakespeare's works
  - Terms
  - Documents
  - Corpus

MAIN TASKS IN IR

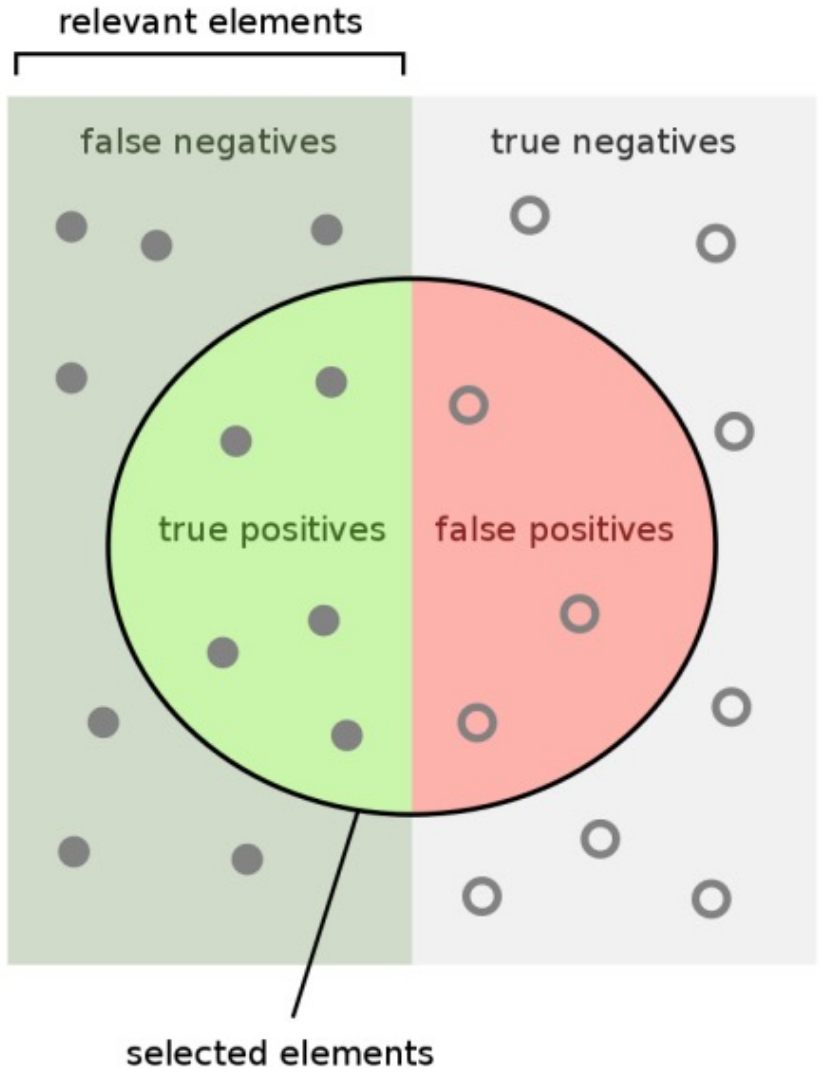
- Information need
- Query Formulation
- Assessment

EVALUATING RESULTS: INTRINSIC

- TREC: set of 'gold' relevant documents
- How many of the documents found?
- Annual bake-off of IR systems

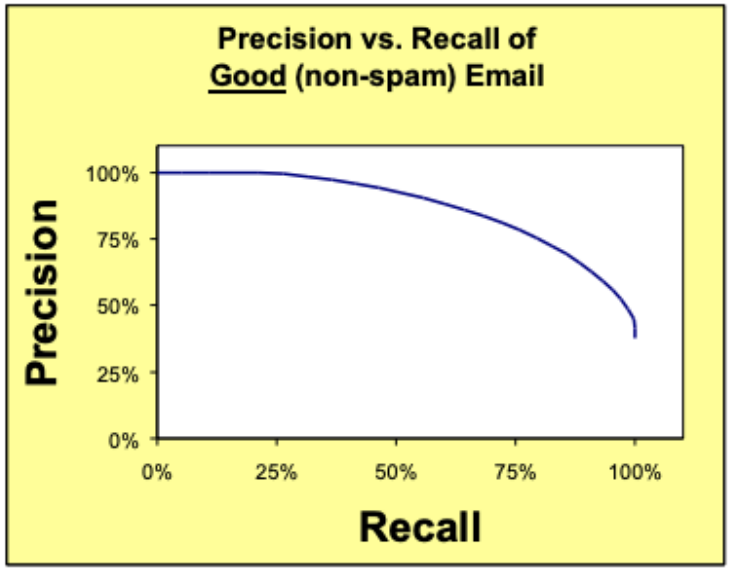
RELEVANCE TERMINOLOGY

- TP: True Positive
- FP: False Positive
- TN: True Negative
- FN: False Negative



PRECISION AND RECALL

- Precision: What fraction of the returned results are relevant to the information need?
  - $P = \frac{|TP|}{|TP|+|FP|}$
- Recall: What fraction of the relevant documents in the collection were returned by the system?
  - $P = \frac{|TP|}{|TP|+|FN|}$
- F-measure: geometric mean



EVALUATING RESULTS: EXTRINSIC

- IR is often used to find answers to questions
- But it takes a human to read the results
  - If you know what answer is, you can search for **similar questions**
  - Build machines to read the answer
- We will see both later
- Programming?