A.I. in health informatics
lecture 1 introduction & stuff

kevin small &

byron wallace

what is this class about?

- **health informatics**
  - managing and making sense of biomedical information

- … but mostly from an **artificial intelligence**/machine learning/nlp view
  - accomplishing the above with learning systems

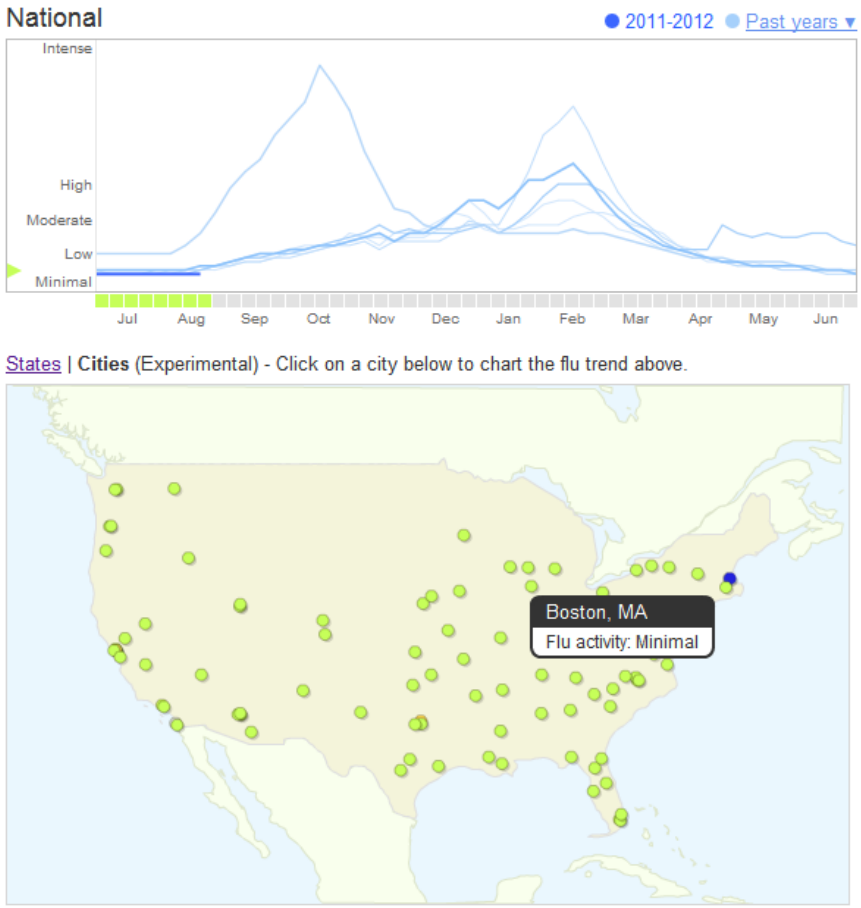
what is this class about?

- by way of example…

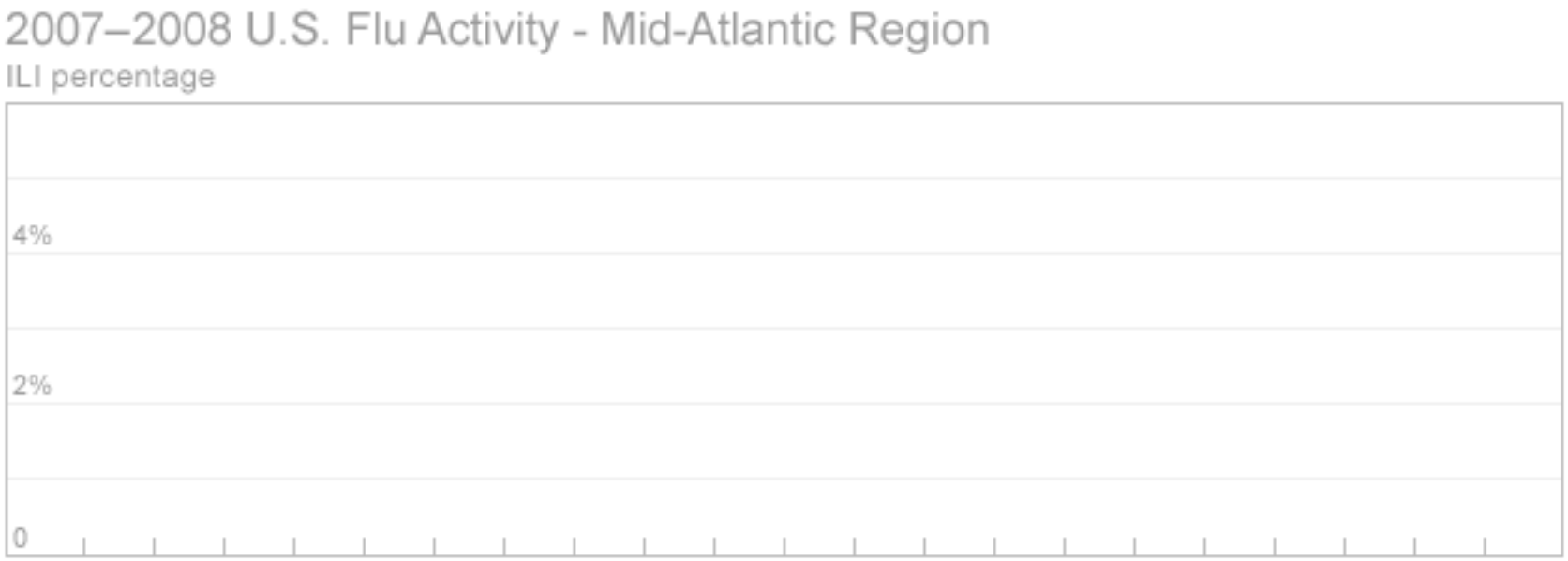
can search queries predict flu outbreaks?

model probability of flu, given search terms.

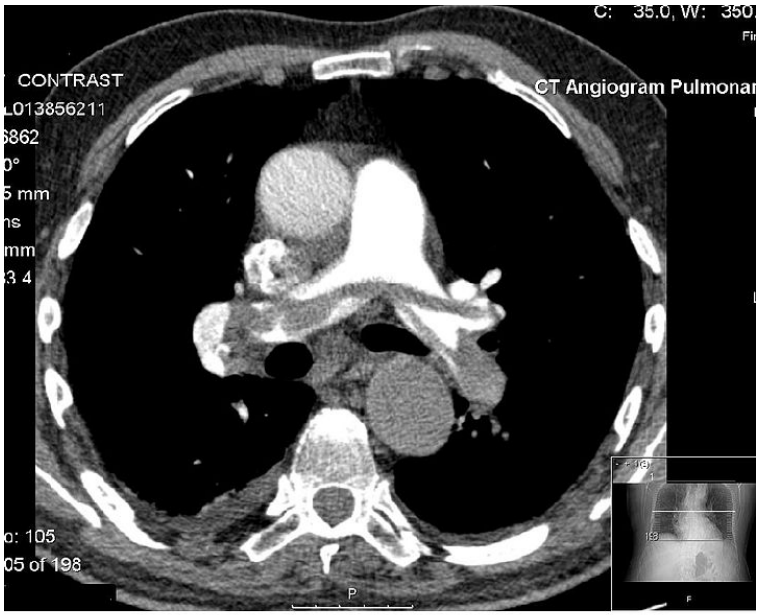
[Ginsberg et al., Nature, 09]

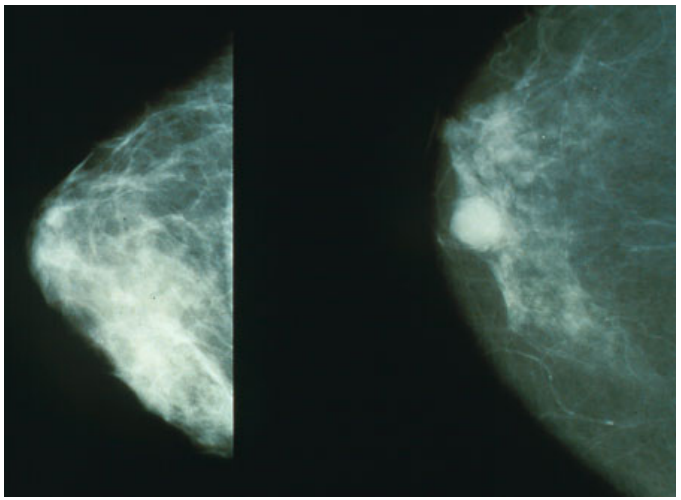


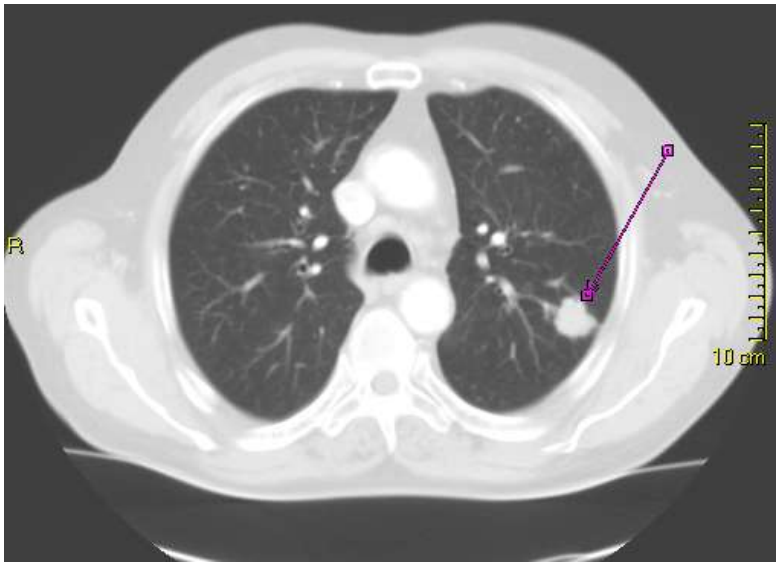
Google flu trends



computer-aided diagnosis







Images from Wikipedia

clinical decision support for $200

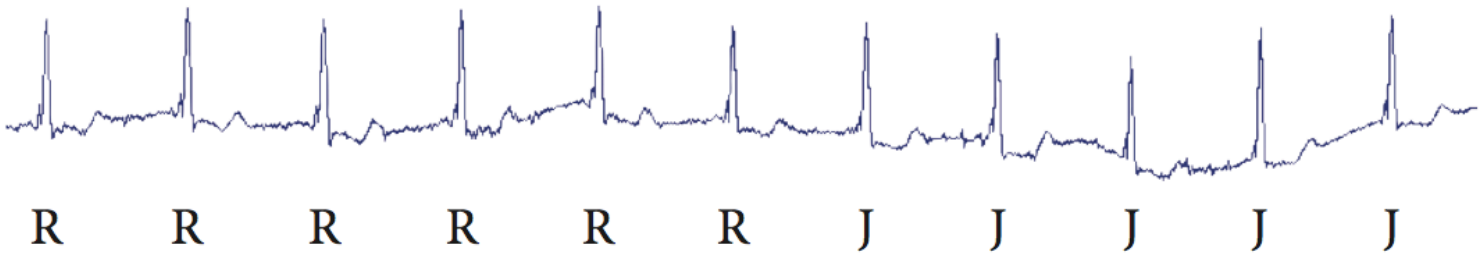
- IBM's Watson is moving into the area of clinical decision support
  - long history of AI in this area
- aim: assist physicians naturally, exploiting huge database of stored knowledge
  - uses natural language processing, machine learning methods

medical question answering

movie time

detection of cardiovascular events

- can we detect cardiac events?



movie time

medical informatics

the scientific field that deals with biomedical information, data, and knowledge – their storage, retrieval, and optimal use for problem solving and decision making.

Shortliffe & Blois

a (very) little history

- 1920s – Hollerith punch cards for public health surveys / epidemiological studies
- 1950s – Data processing for billing
- 1960s – Clinical Support Systems
- 1970s – Hospital Information Systems
- 1980s – Management Information Systems, Computer Diagnostic Imaging
- 1990s – Unified Health Records, Clinical Decision Support Systems

rise of medical informatics

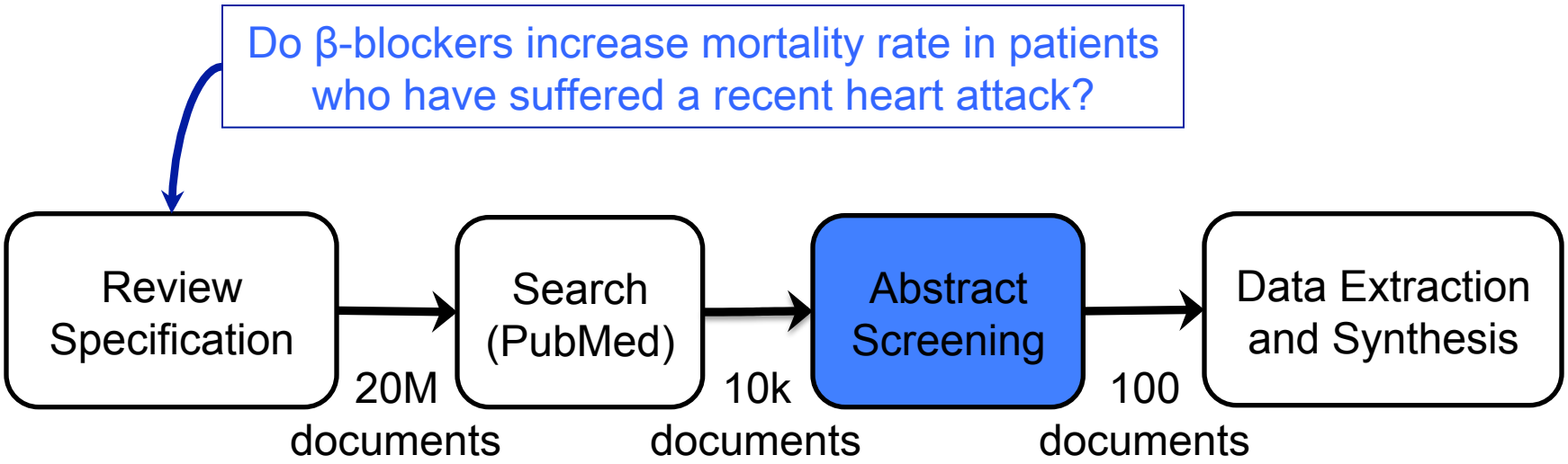
- increased reliance on evidence-based practice guidelines
- too much information – not enough time to analyze
- uncertainty abounds
- lots of patients / patient-centered movement

a brief illustrative task: *abstract screening*

- or, a shameless instance of rampant self-promotion,
- or, our day job

abstract screening

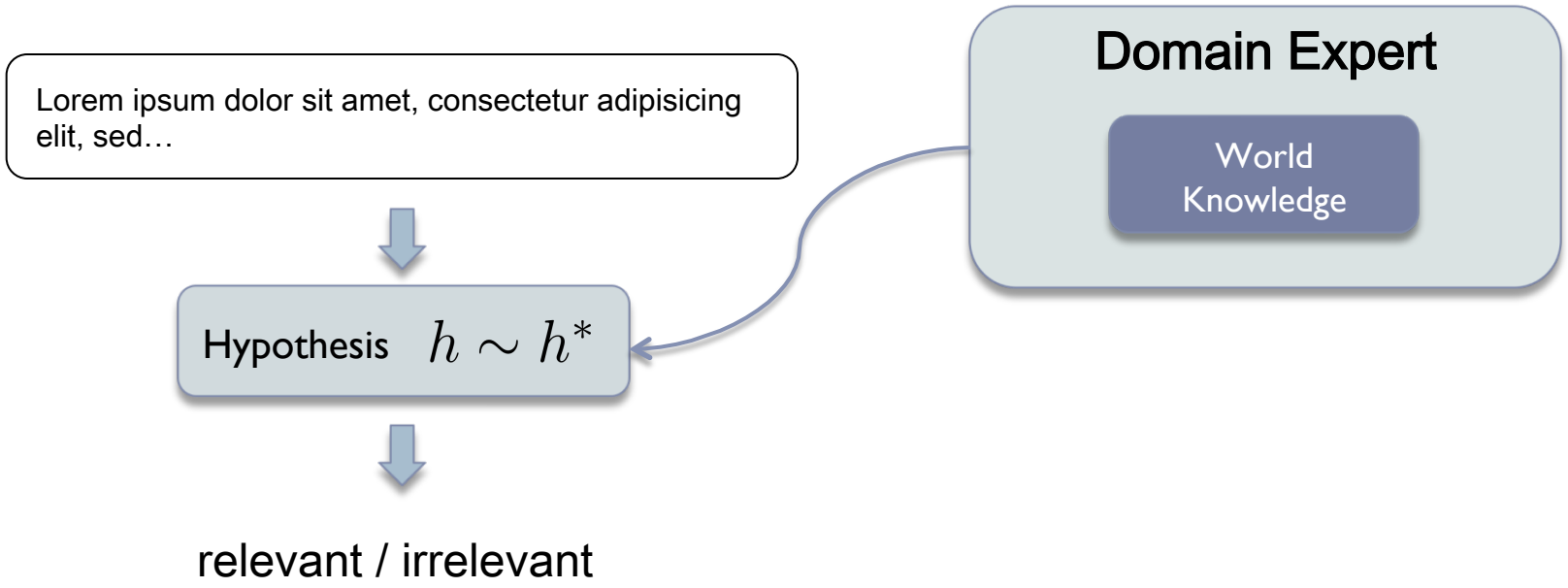
- **Systematic review**: an exhaustive assessment of existing published evidence regarding a precise clinical question



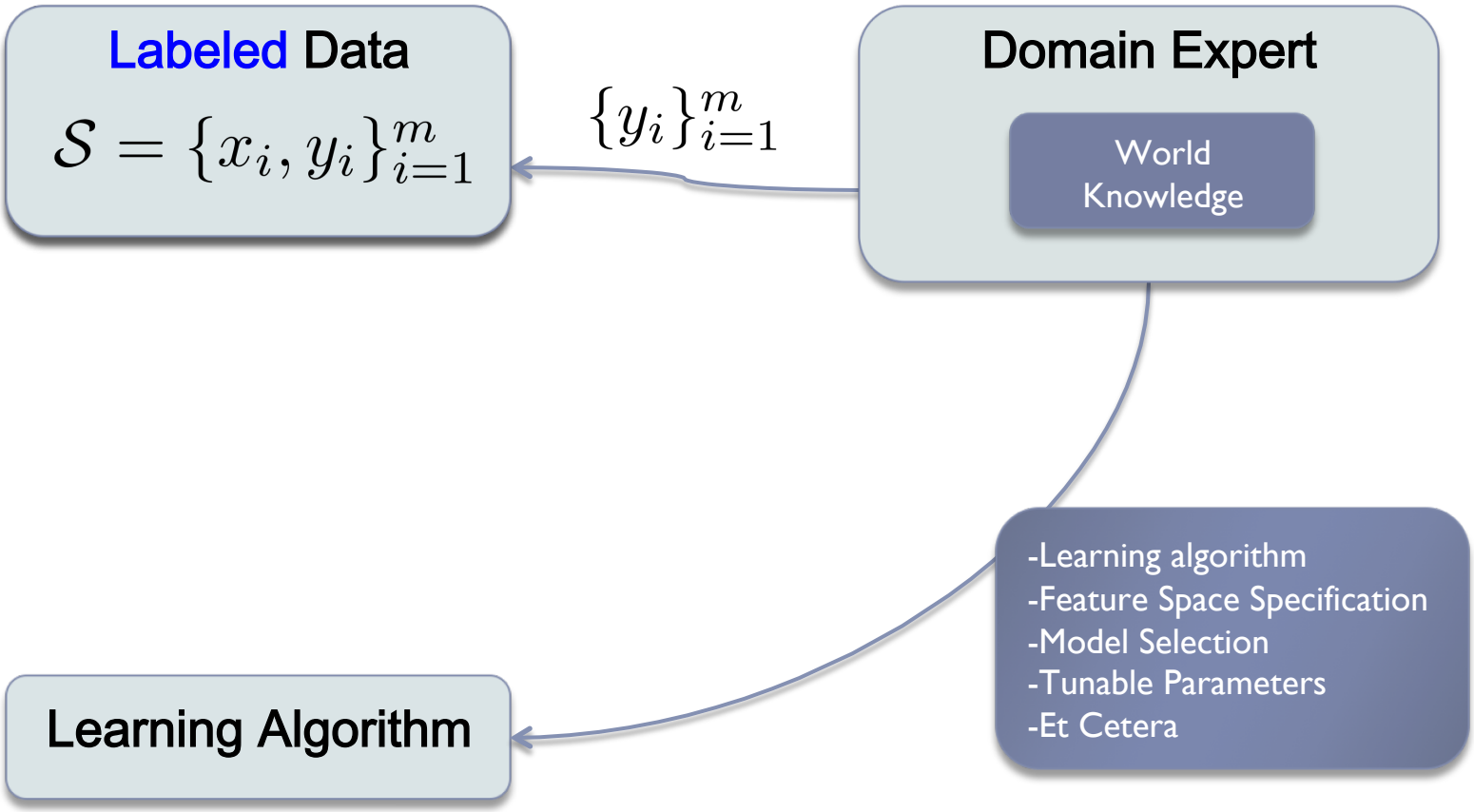
- Goal is to have doctors screen a small number of abstracts (e.g. 100s) and have a classifier do the remainder automatically [Wallace et al.; KDD 10]

… is a lot of work

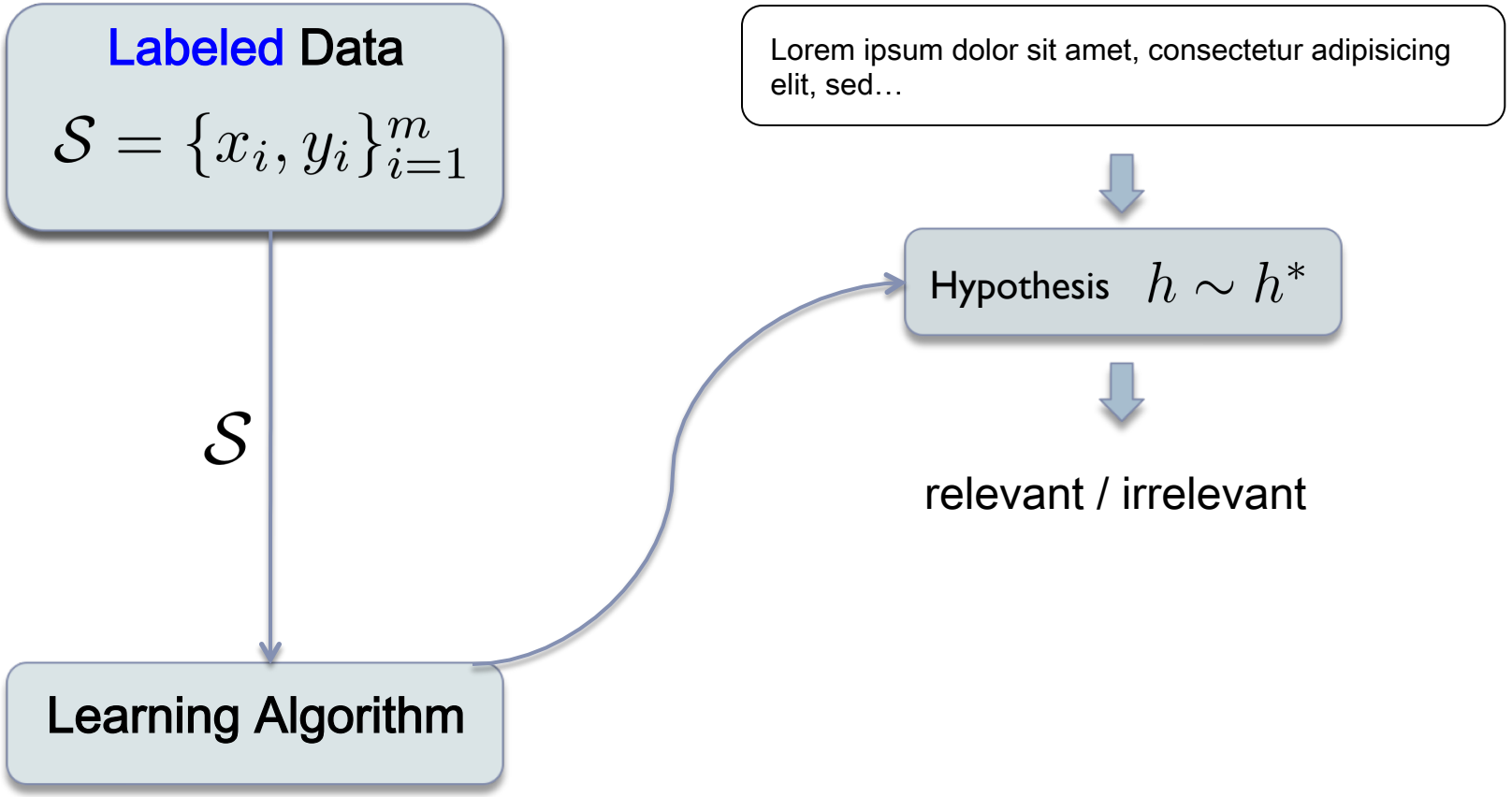
predictive models



machine learning



machine learning



abstract screening, redux

- need to derive a suitable representation for the input data (text)

- need to select an appropriate learning algorithm

bag-of-words representation

- classification algorithms operate on vectors
- *feature space*: an n-dimensional representation of things
  - … but how to vectorize text?
- bag-of-words: map documents to indicator vectors

a bag-of-words example

let’s say we want to encode two sentences

S_1= “Boston drivers are frequently aggressive”
S_2= “The Boston Red Sox frequently hit line drives”

eliminate stopwords

S_1= “Boston drivers are frequently aggressive”
S_2= “The Boston Red Sox frequently hit line drives”

remove case information

S_1= “boston drivers are frequently aggressive”
S_2= “the boston red sox frequently hit line drives”

stemming

S_1= “boston drivers are frequently aggressive”
S_2= “the boston red sox frequently hit line drives”

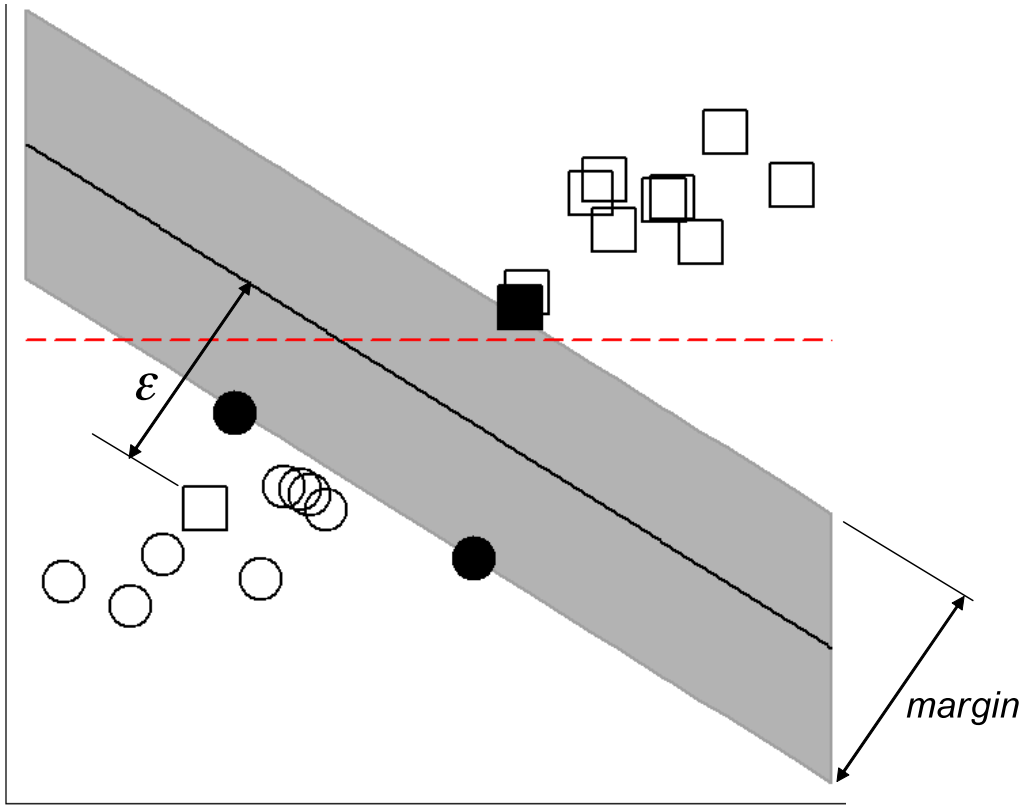
feature vectors

	hit	red	sox	line	boston	frequent	drive	aggressive
x_1 =	0	0	0	0	1	1	1	1
x_2 =	1	1	1	1	1	1	1	0

a new sentence, S_3, comes along it reads: "I hate the red sox". to which sentence is it most similar?

x_3 =	0	1	1	0	0	0	0	0

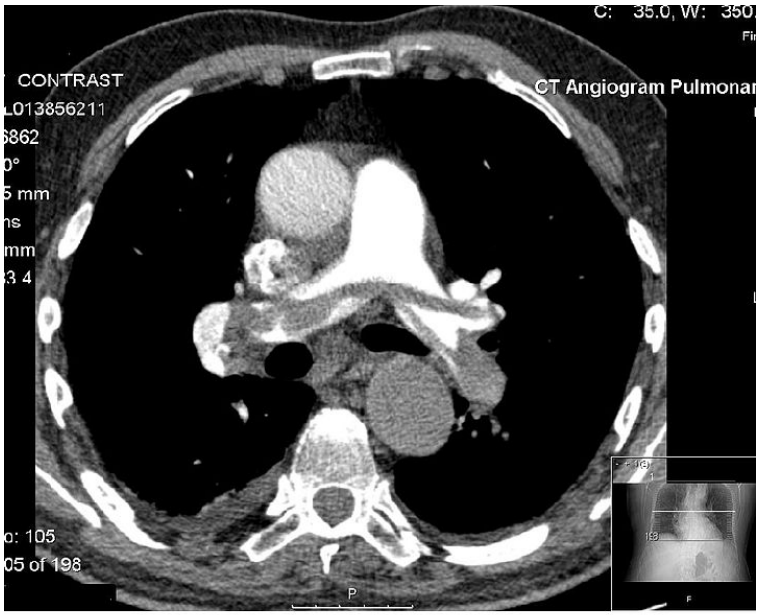
support vector machines

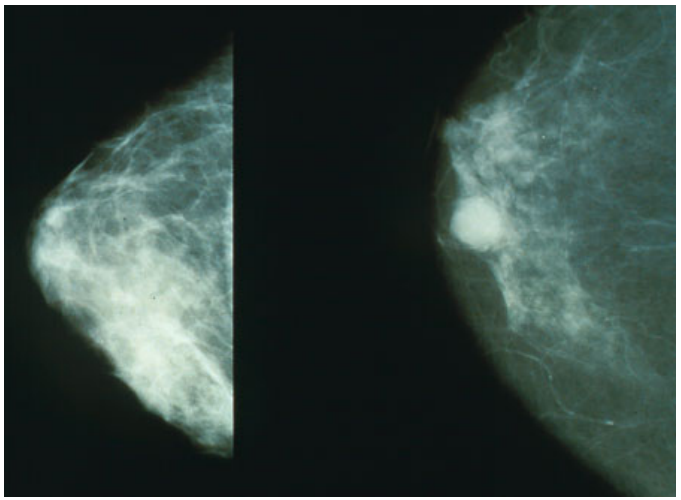


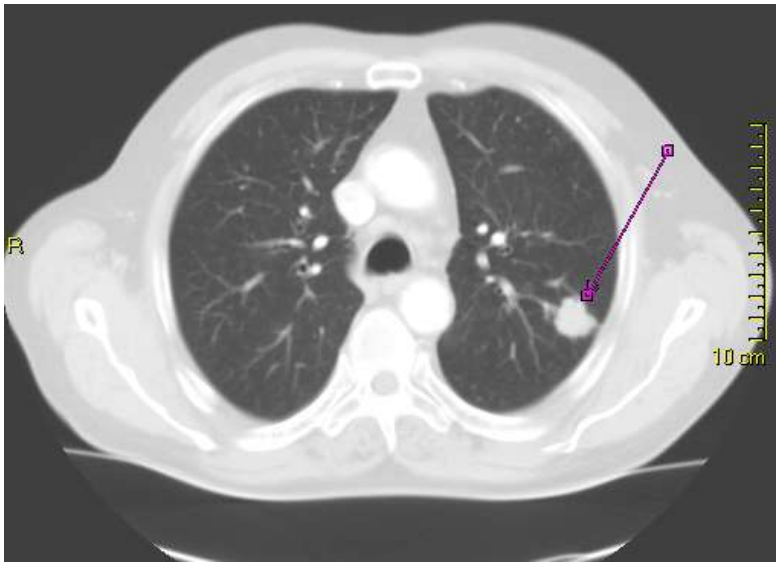
$$\min_{\vec{w},\vec{\varepsilon}} \left(\underbrace{\frac{1}{2}\vec{w}^T\vec{w}}_{\text{inversely related to margin between support vectors}} + \underbrace{C\sum_{i=1}^{l}\varepsilon_i}_{\text{cost of mis-classifications}} \right)$$

$$\hat{y} = h(x) = \underset{y' \in \mathcal{Y}}{\operatorname{argmax}}\, f(\mathbf{x}, y')$$

computer-aided diagnosis



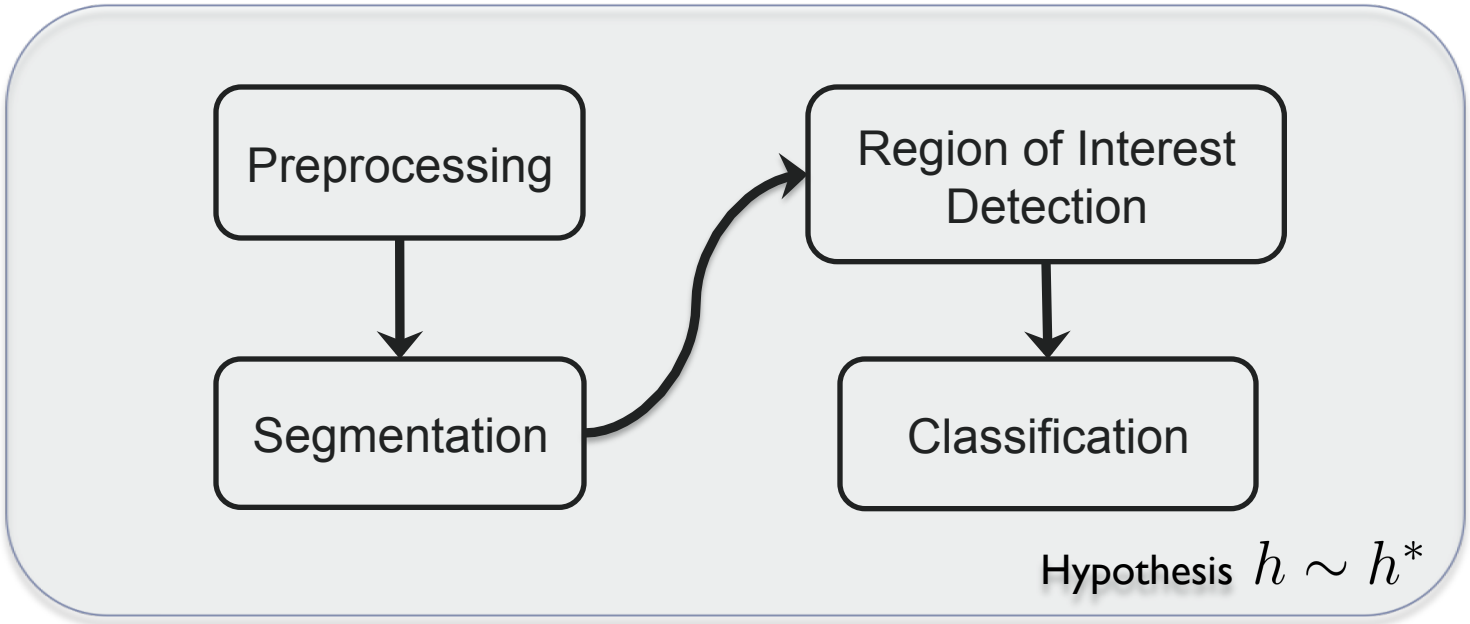




Images from Wikipedia

pipeline model

- decomposes complex task into sequential stages of simpler tasks



- drawbacks?

inference

- actionable intelligence may require multiple classifiers and domain knowledge
  - important for structured information
- how do we effectively assemble this information?
- how do we get system users to trust the results?

unique issues

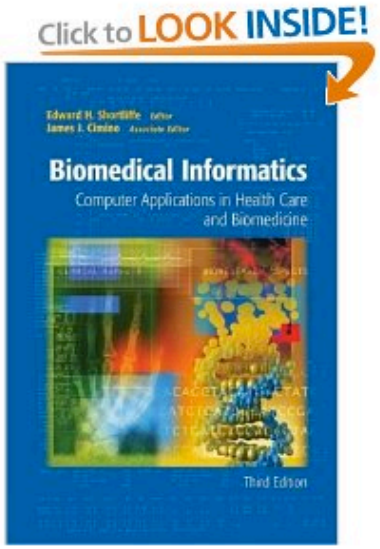
- low prevalence, asymmetric loss
- value of engineering
- tons of available data
- analytic frameworks & formal reasoning systems already exist

course goals, expectations & logistics

what are our goals?

- a survey course on the application of ai and ml to health informatics
- a competence level of such that you will understand research papers and implement ideas
  - …ideally at a level at which you can conduct your own research
- this is *not* a bioinformatics course

useful textbook



Biomedical Informatics: Computer Applications in Health Care and Biomedicine (Health Informatics) [Hardcover]

Edward H. Shortliffe (Editor), James J. Cimino (Editor)

(11 customer reviews) | Like (3)

List Price: ~~$99.00~~

Price: $72.52 & this item ships for **FREE with Super Saver Shipping**. Details

You Save: $26.48 (27%)

Special Offers Available

In Stock.

Ships from and sold by **Amazon.com**. Gift-wrap available.

Want it delivered Thursday, September 8? Order it in the next 7 hours and 9 minutes, and choose **One-Day Shipping** at checkout. Details

16 new from $72.52 **29 used** from $62.92

expectations & logistics

- read class material **before** class
- ask questions
- grading
  - 25% homework (4-5 written/programming)
  - 10% reaction papers (6-8 one page)
  - 25% midterm
  - 40% final project (collaborative, per approval)

coordinates

- http://www.cs.tufts.edu/comp/150AIH/
- ksmall@cs.tufts.edu
- byron.wallace@gmail.com