classification

chris.wiggins@columbia.edu

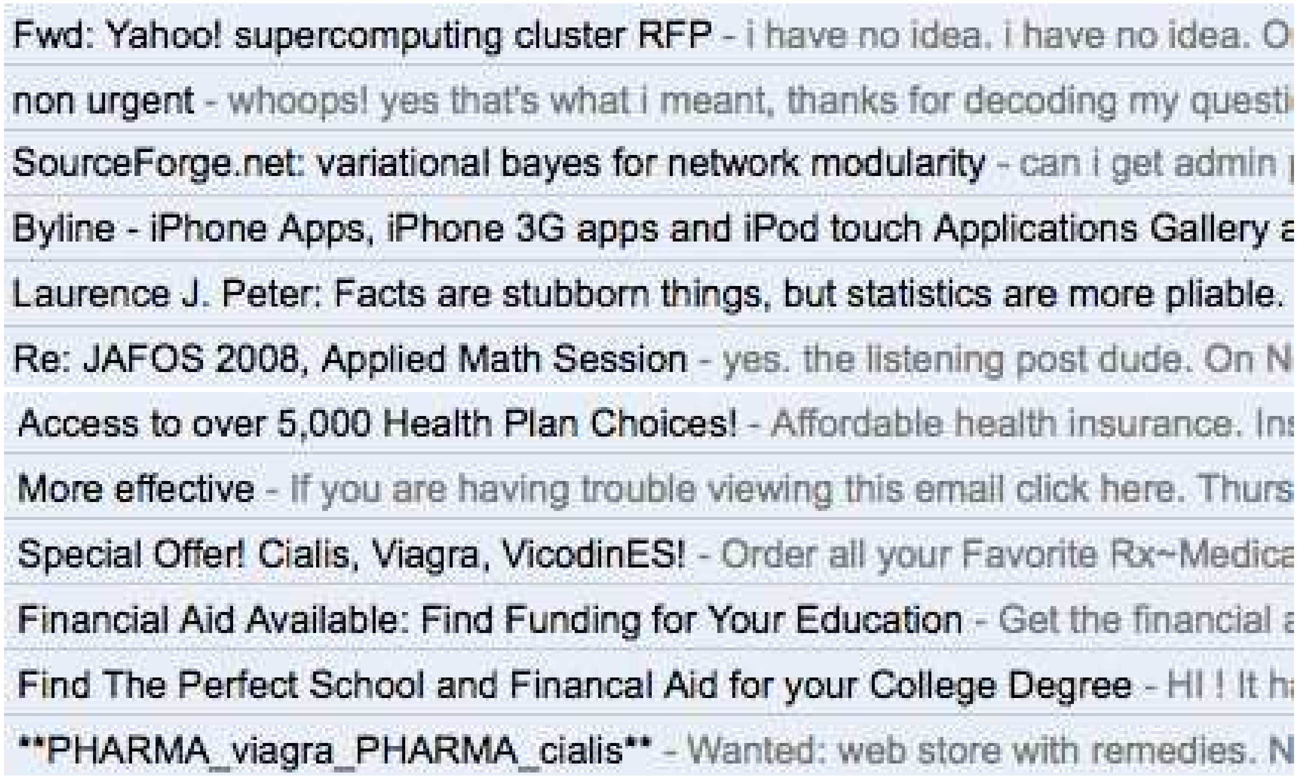
2017-03-10

wat?

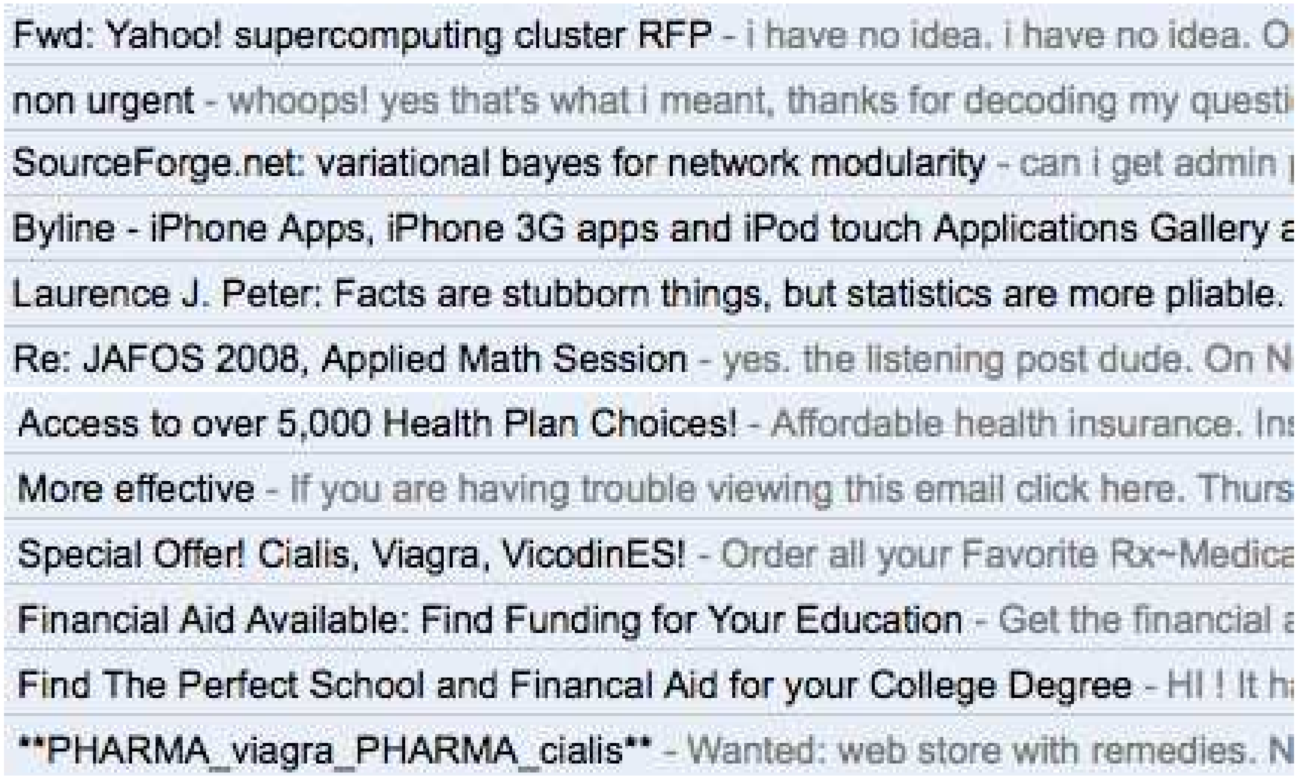
example: spam/ham

(cf. jake's great deck on this)

Learning by example



Learning by example



Fwd: Yahoo! supercomputing cluster RFP - i have no idea. i have no idea. O
non urgent - whoops! yes that's what i meant, thanks for decoding my questi
SourceForge.net: variational bayes for network modularity - can i get admin
Byline - iPhone Apps, iPhone 3G apps and iPod touch Applications Gallery
Laurence J. Peter: Facts are stubborn things, but statistics are more pliable.
Re: JAFOS 2008, Applied Math Session - yes. the listening post dude. On N
Access to over 5,000 Health Plan Choices! - Affordable health insurance. In
More effective - If you are having trouble viewing this email click here. Thurs
Special Offer! Cialis, Viagra, VicodinES! - Order all your Favorite Rx~Medica
Financial Aid Available: Find Funding for Your Education - Get the financial
Find The Perfect School and Financal Aid for your College Degree - HI ! It h
PHARMA_viagra_PHARMA_cialis - Wanted: web store with remedies. N

- How did you solve this problem?
- Can you make this process explicit (e.g. write code to do so)?

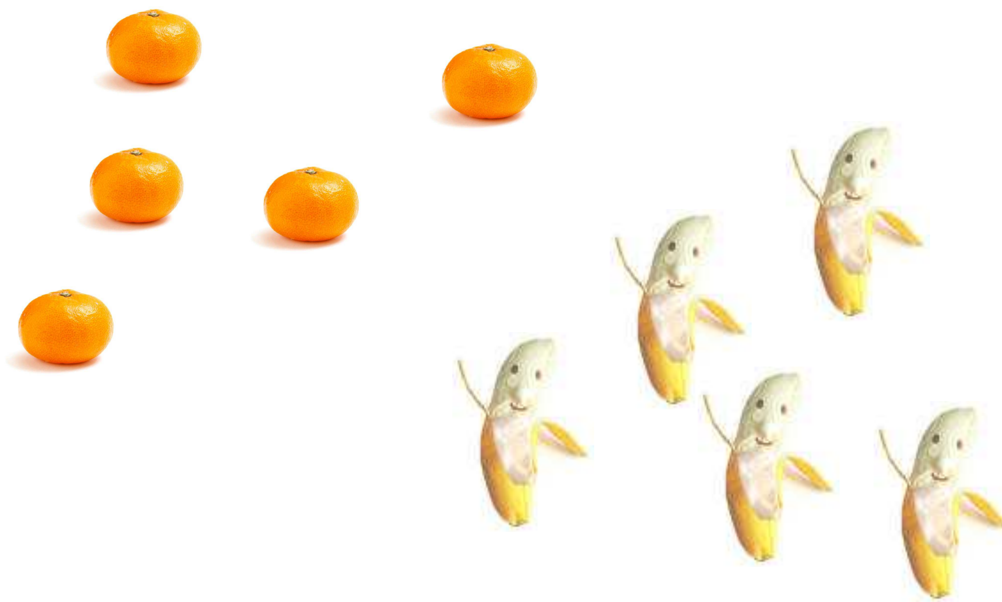
classification?

FIGURE 1.2. *Examples of handwritten digits from U.S. postal envelopes.*

build a theory of 3's?

1-slide summary of classification

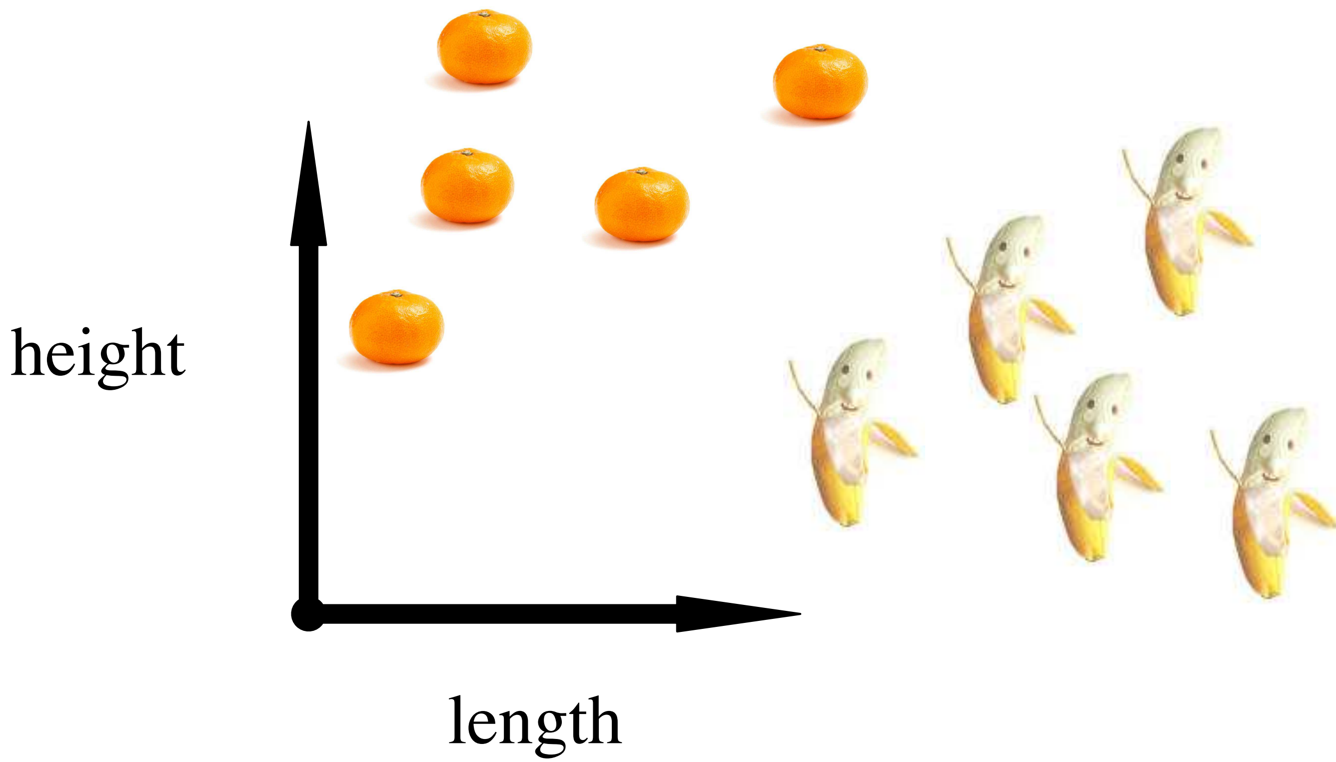
- banana or orange?



what would Gauss do?

1-slide summary of classification

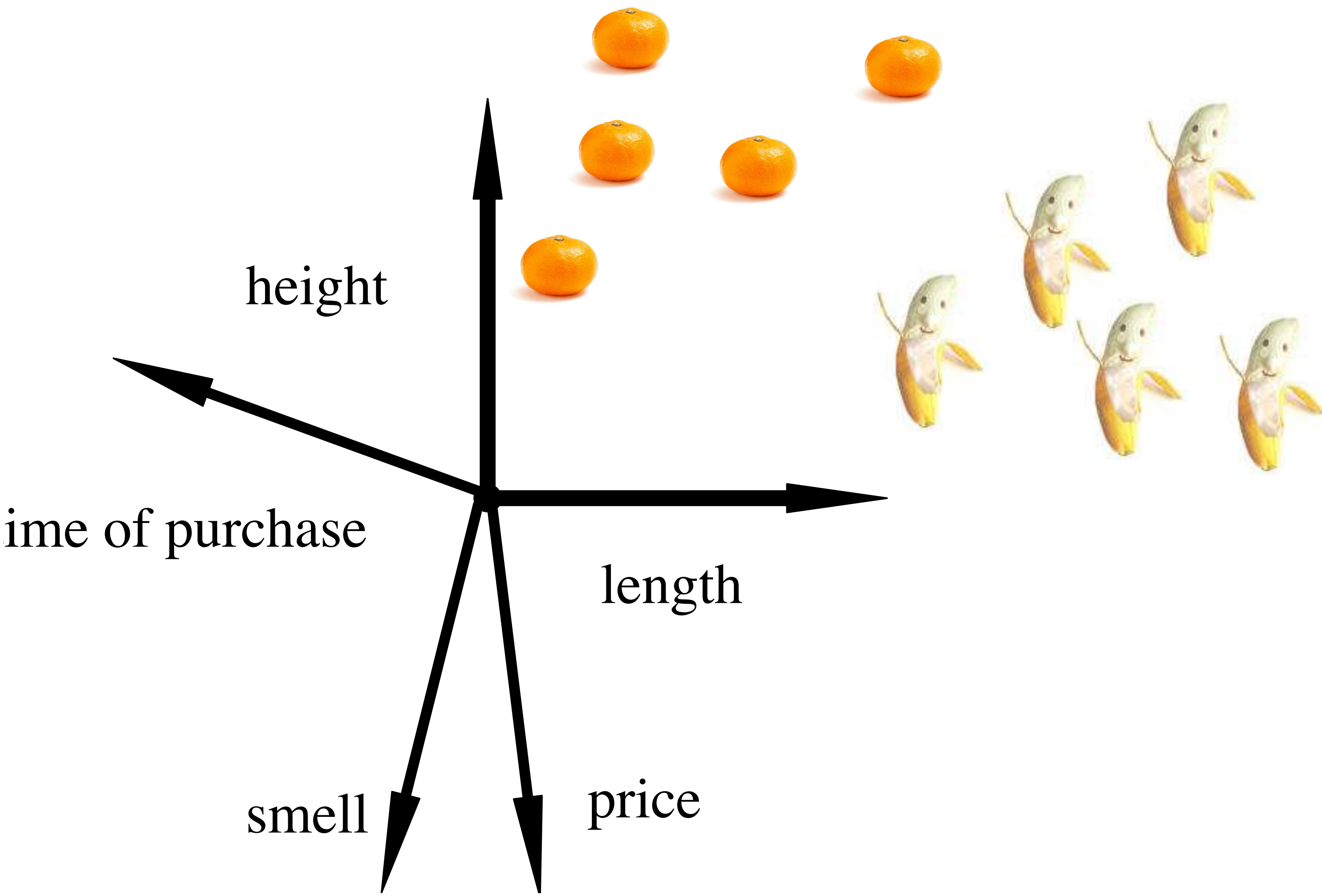
- banana or orange?



what would Gauss do?

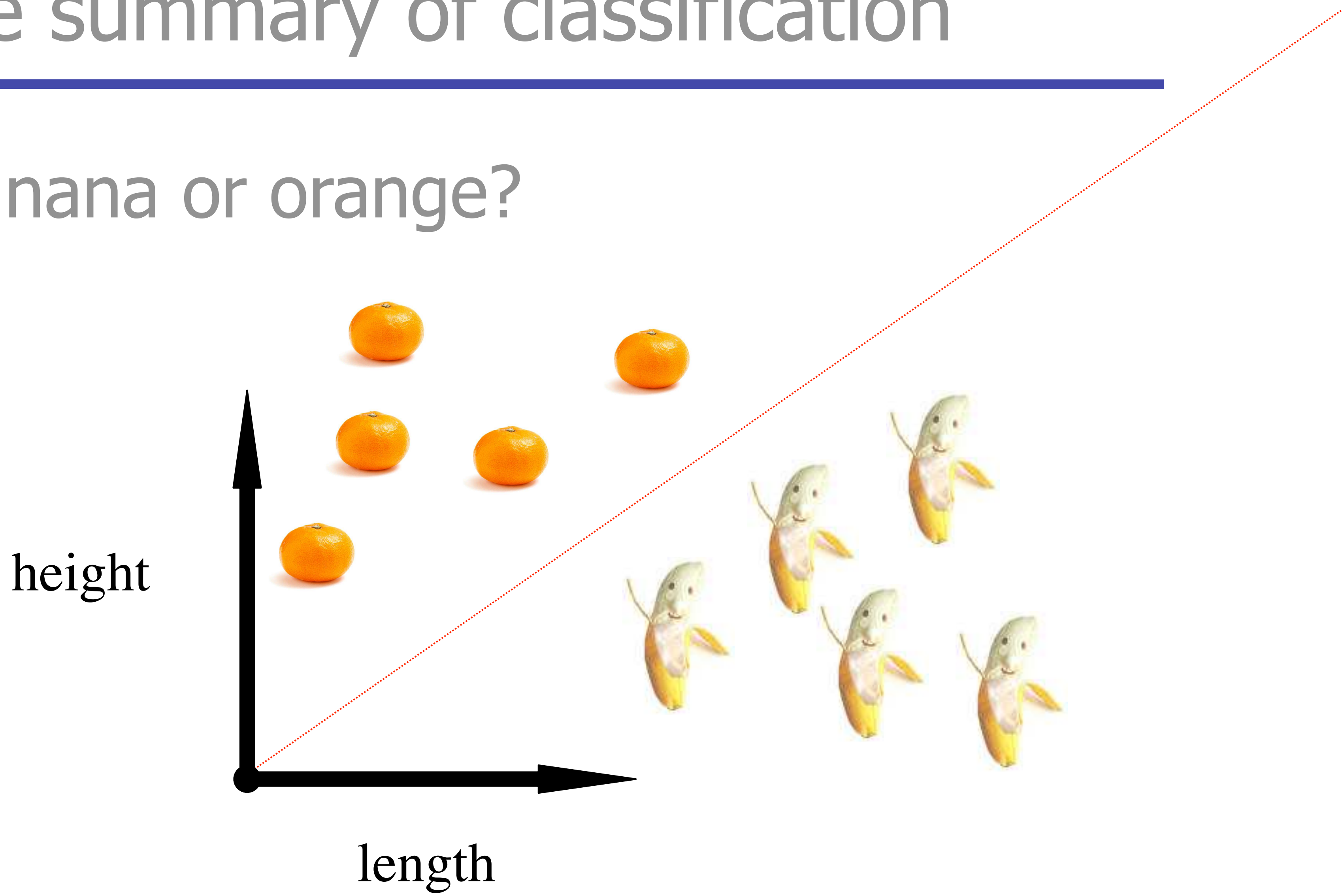
1-slide summary of classification

- banana or orange?



1-slide summary of classification

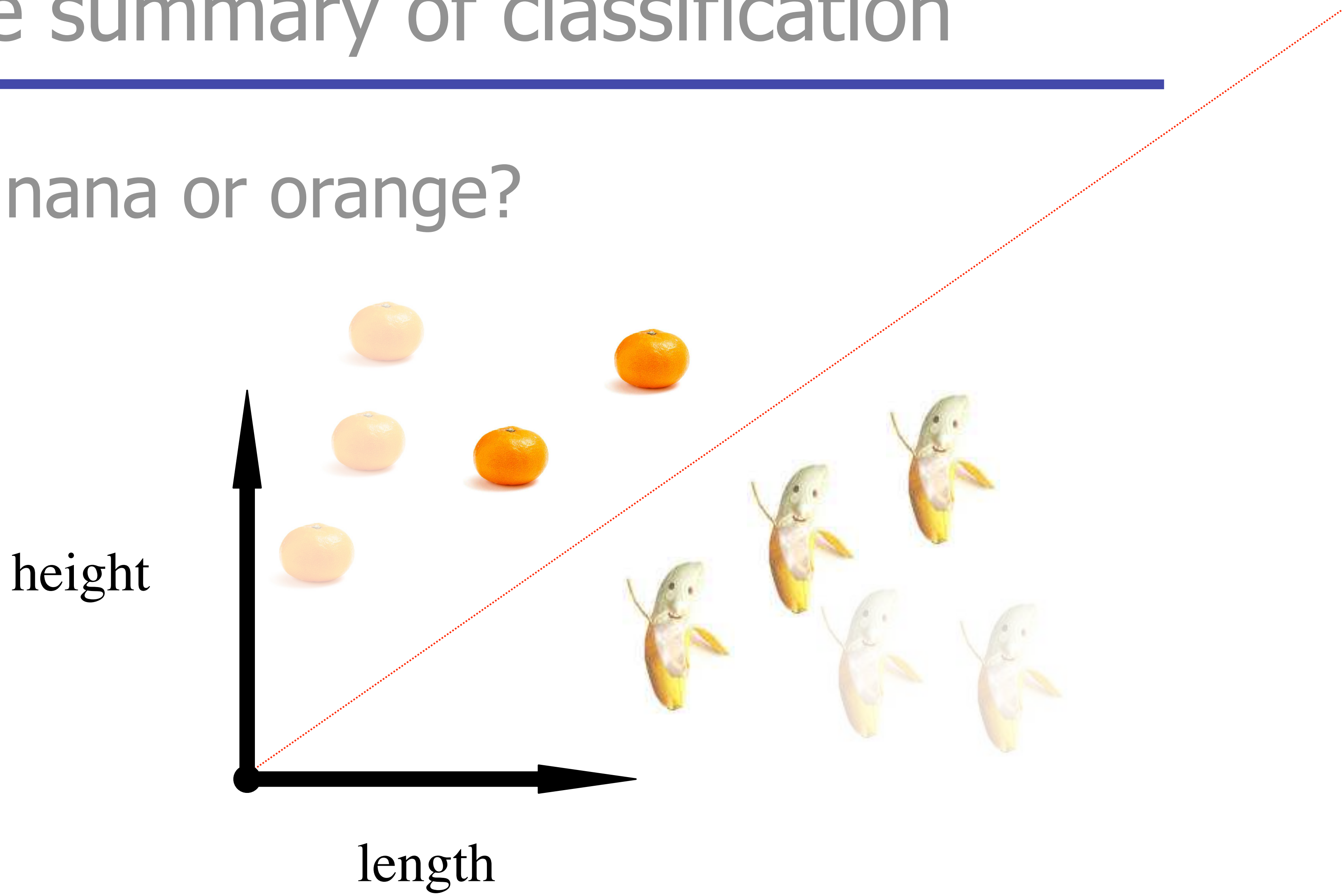
- banana or orange?



game theory:
“assume the worst”

1-slide summary of classification

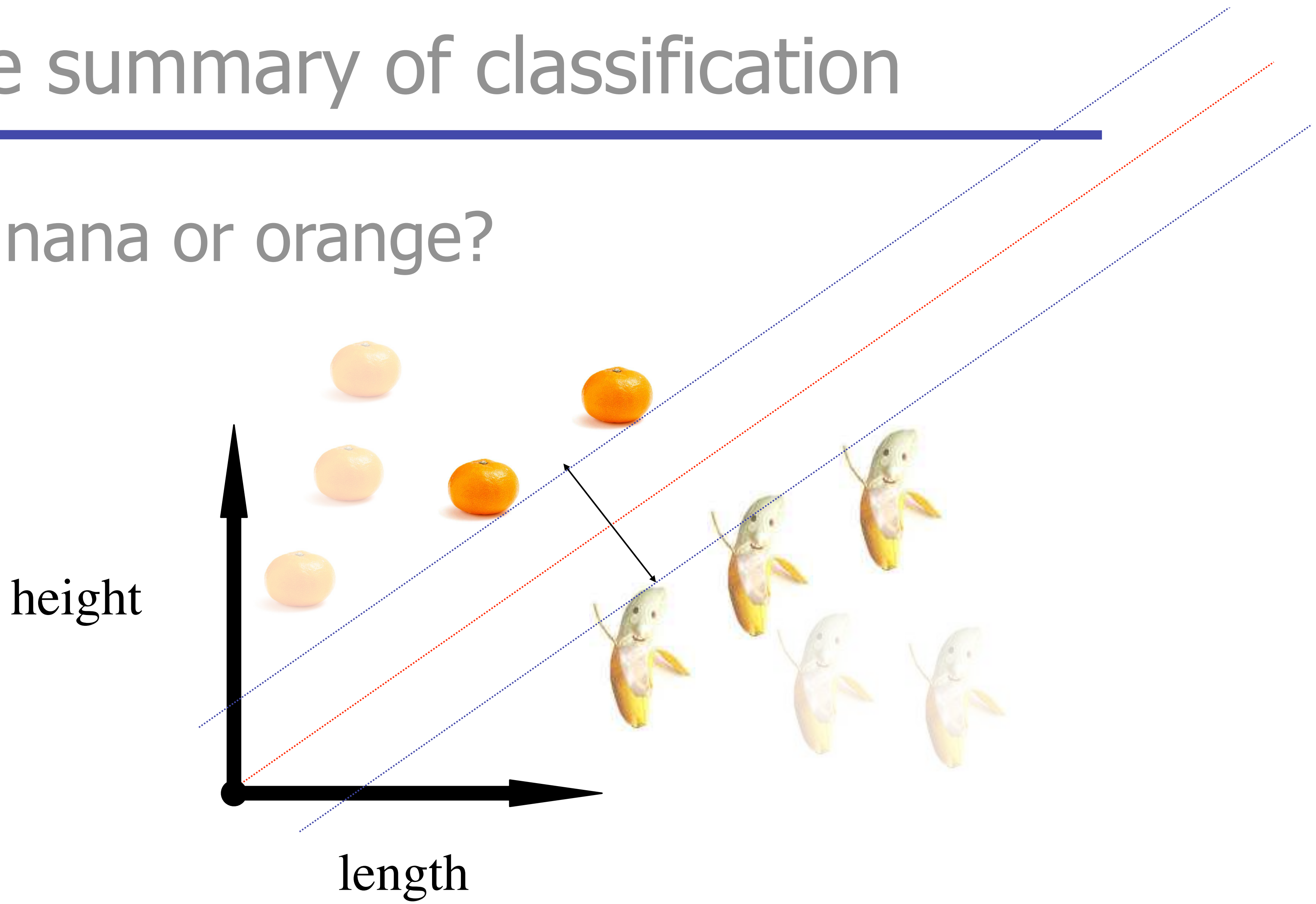
- banana or orange?



large deviation theory:
"maximum margin"

1-slide summary of classification

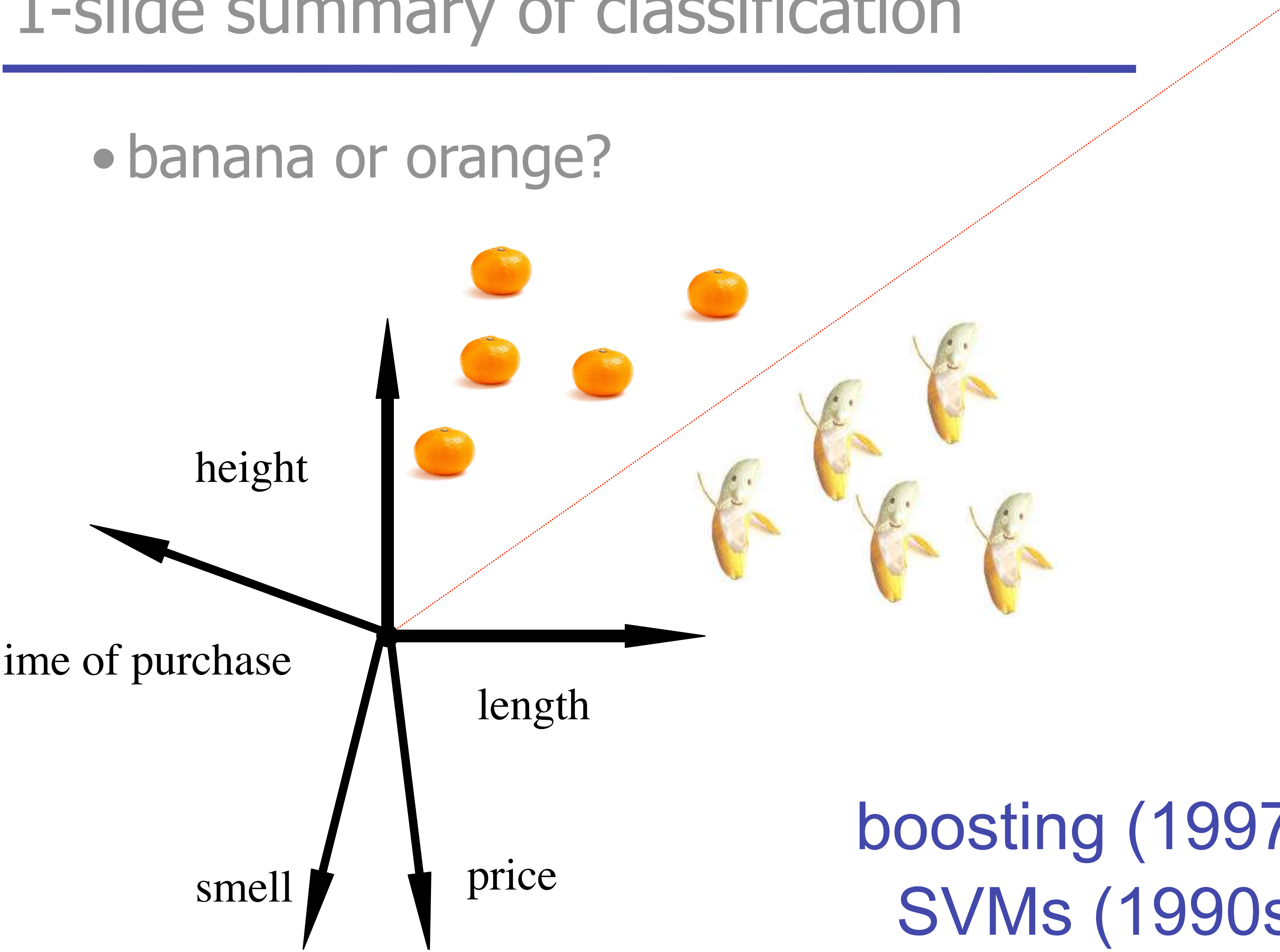
- banana or orange?



large deviation theory:
"maximum margin"

1-slide summary of classification

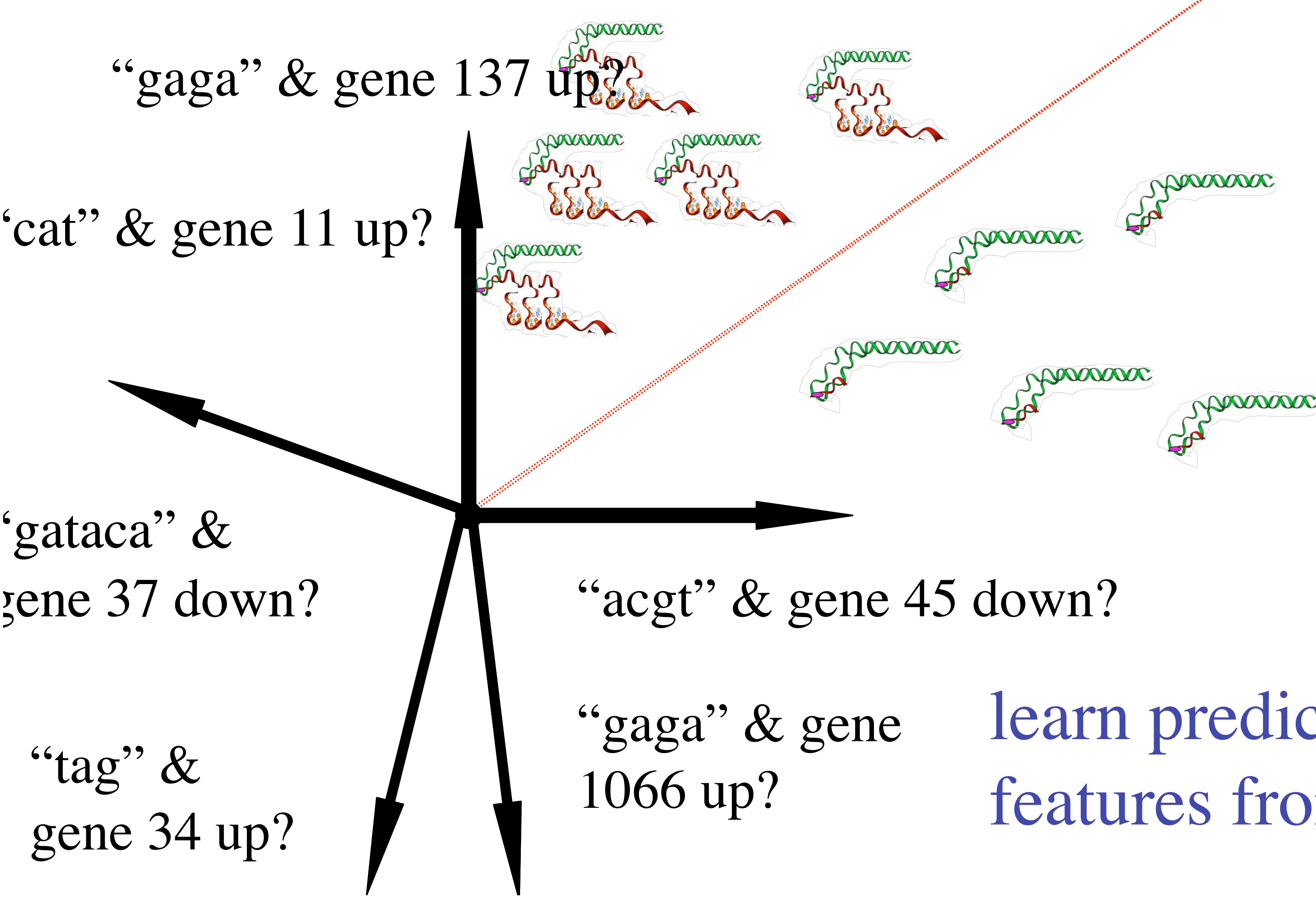
- banana or orange?



boosting (1997)
SVMs (1990s)

1-slide summary of classification

- up- or down- regulated?



learn predictive
features from data

example: bad bananas

example@NYT in CAR (computer assisted reporting)

www.nytimes.com/2014/09/12/business/air-bag-flaw-long-known-led-to-recalls.html?_r=1

HOME SEARCH The New York Times

BUSINESS DAY

Air Bag Flaw, Long Known to Honda and Takata, Led to Recalls

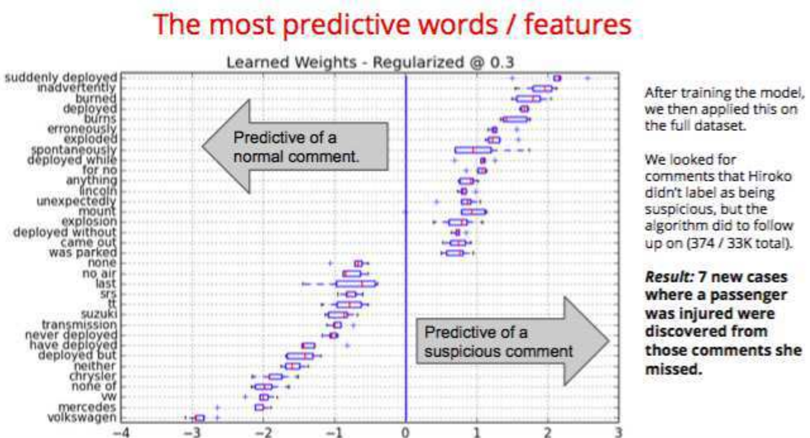
By HIROKO TABUCHI SEPT. 11, 2014

The air bag in Jennifer Griffin's Honda Civic was not among the recalled vehicles in 2008. Jim Keely

Figure 1: Tabuchi article

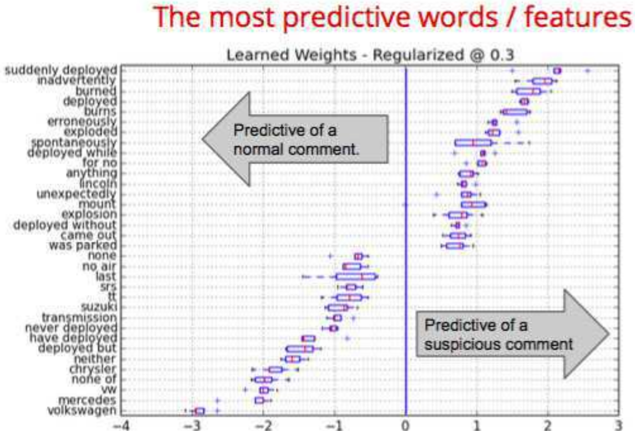
example in CAR (computer assisted reporting)

- cf. Friedman's "Statistical models and Shoe Leather"[1]



example in CAR (computer assisted reporting)

- cf. Friedman's "Statistical models and Shoe Leather"[1]
- Takata airbag fatalities



After training the model, we then applied this on the full dataset.

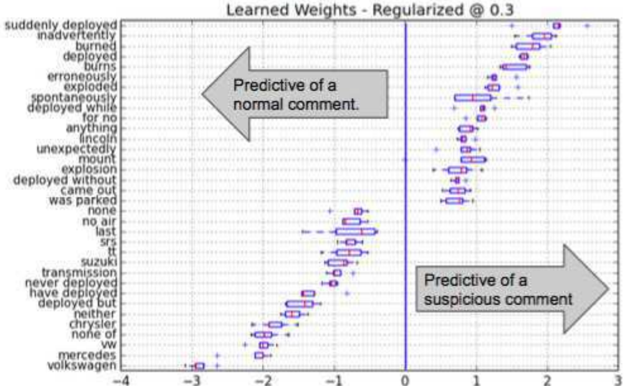
We looked for comments that Hiroko didn't label as being suspicious, but the algorithm did to follow up on (374 / 33K total).

***Result:* 7 new cases where a passenger was injured were discovered from those comments she missed.**

example in CAR (computer assisted reporting)

- cf. Friedman's "Statistical models and Shoe Leather"[1]
- Takata airbag fatalities
- 2219 labeled[2] examples from 33,204 comments

The most predictive words / features



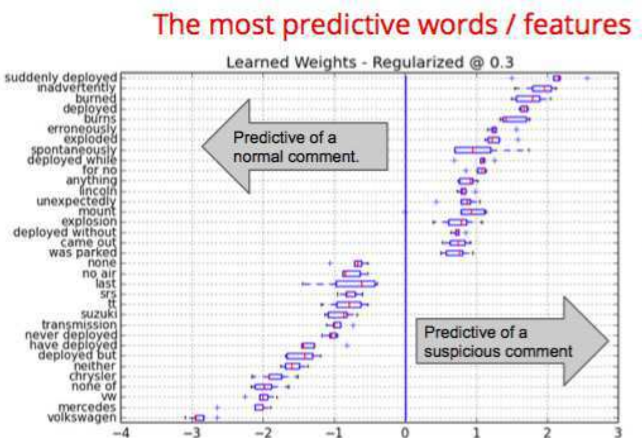
After training the model, we then applied this on the full dataset.

We looked for comments that Hiroko didn't label as being suspicious, but the algorithm did to follow up on (374 / 33K total).

***Result:* 7 new cases where a passenger was injured were discovered from those comments she missed.**

example in CAR (computer assisted reporting)

- cf. Friedman's "Statistical models and Shoe Leather"[1]
- Takata airbag fatalities
- 2219 labeled[2] examples from 33,204 comments
- cf. Box's "Science and Statistics"[3]



After training the model, we then applied this on the full dataset.

We looked for comments that Hiroko didn't label as being suspicious, but the algorithm did to follow up on (374 / 33K total).

Result: **7 new cases where a passenger was injured were discovered from those comments she missed.**

computer assisted reporting

- Impact

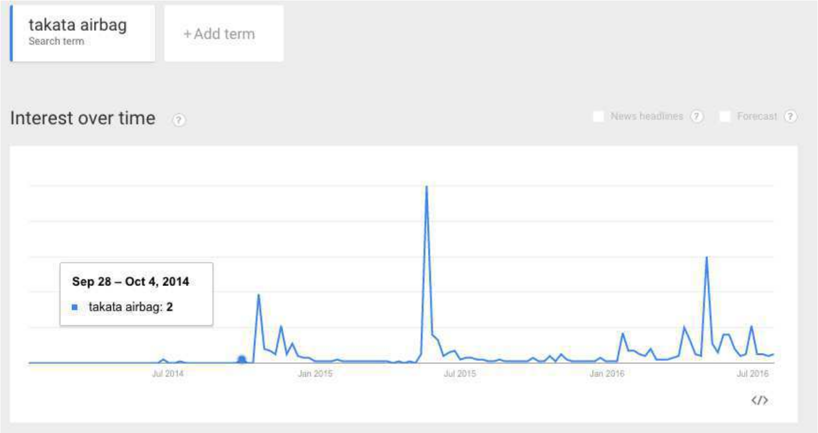


Figure 3: impact

conjecture: cost function?

fallback: probability

review: regression as probability

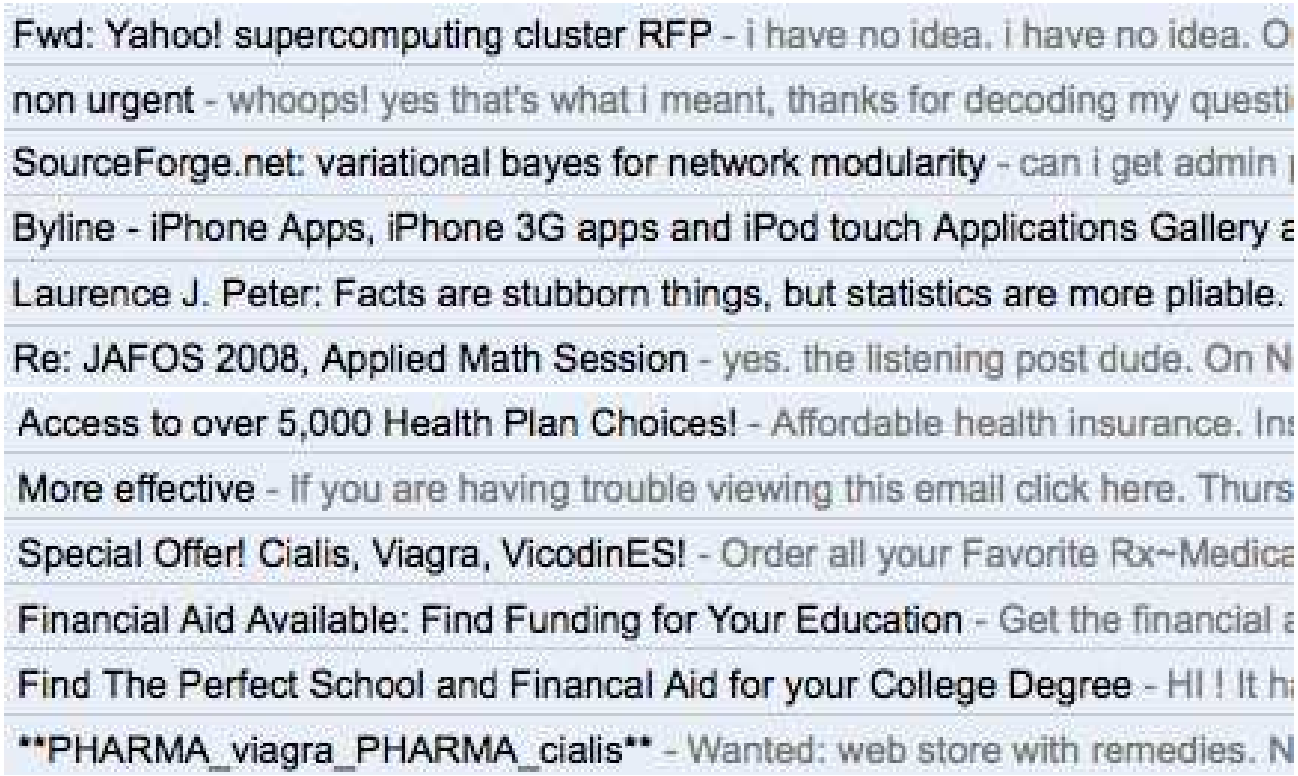
classification as probability

binary/dichotomous/boolean features + NB

digression: bayes rule

generalize, maintain linerarity

Learning by example



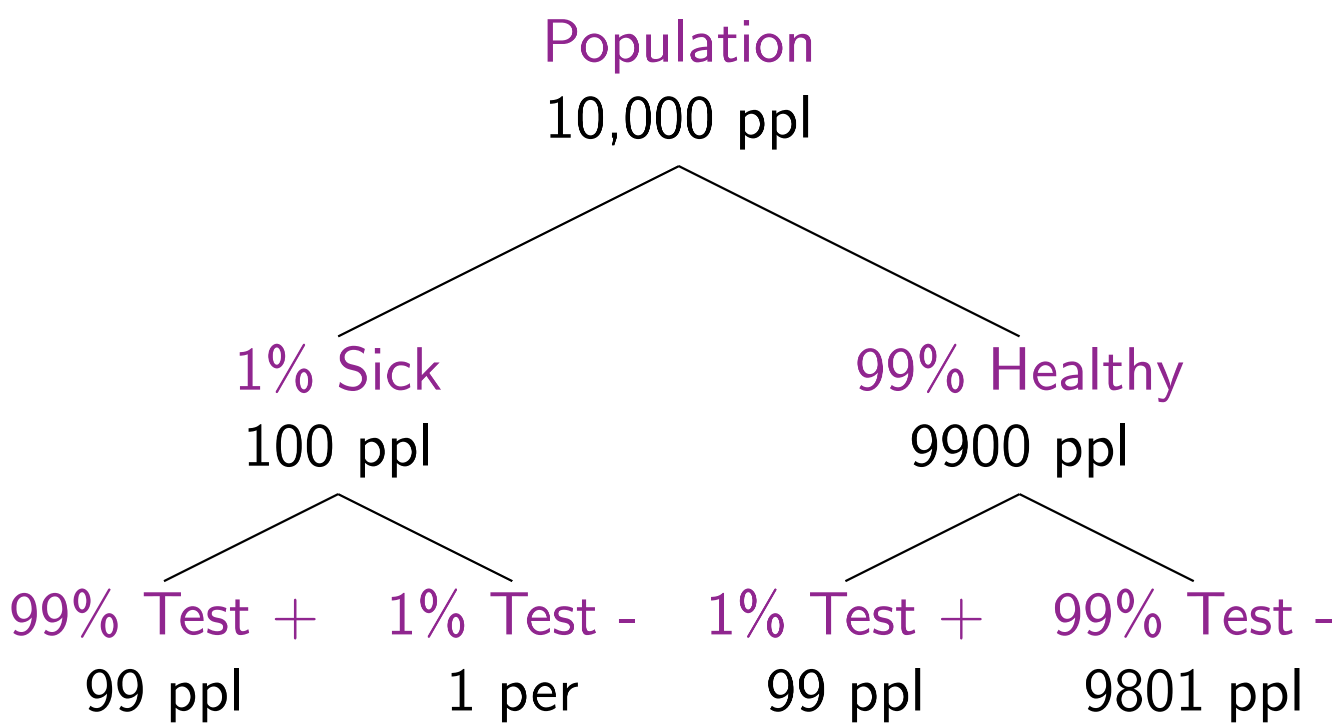
- How did you solve this problem?
- Can you make this process explicit (e.g. write code to do so)?

Diagnoses a la Bayes[1]

- You're testing for a rare disease:
  - 1% of the population is infected
- You have a highly sensitive and specific test:
  - 99% of sick patients test positive
  - 99% of healthy patients test negative
- Given that a patient tests positive, what is probability the patient is sick?

[1] Wiggins, SciAm 2006

Diagnoses a la Bayes

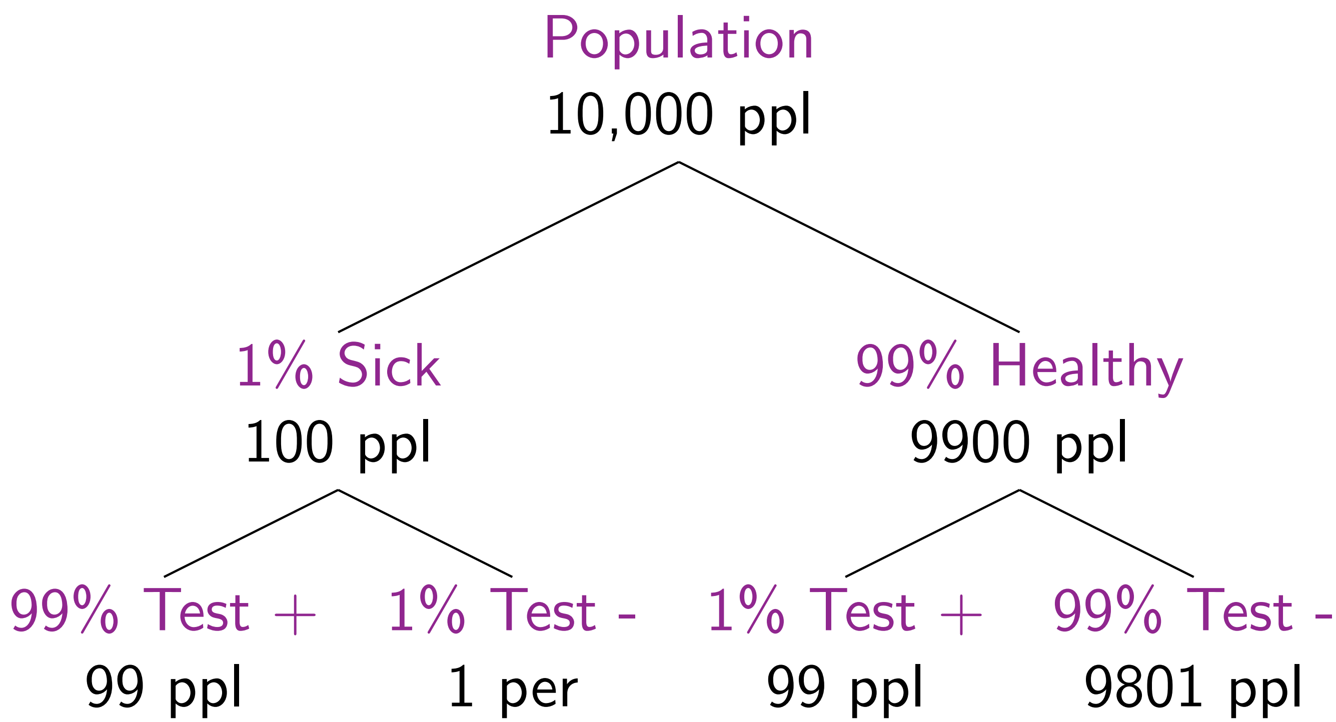


Diagnoses a la Bayes

- Population
  10,000 ppl
  - 1% Sick
    100 ppl
    - 99% Test +
      99 ppl
    - 1% Test -
      1 per
  - 99% Healthy
    9900 ppl
    - 1% Test +
      99 ppl
    - 99% Test -
      9801 ppl

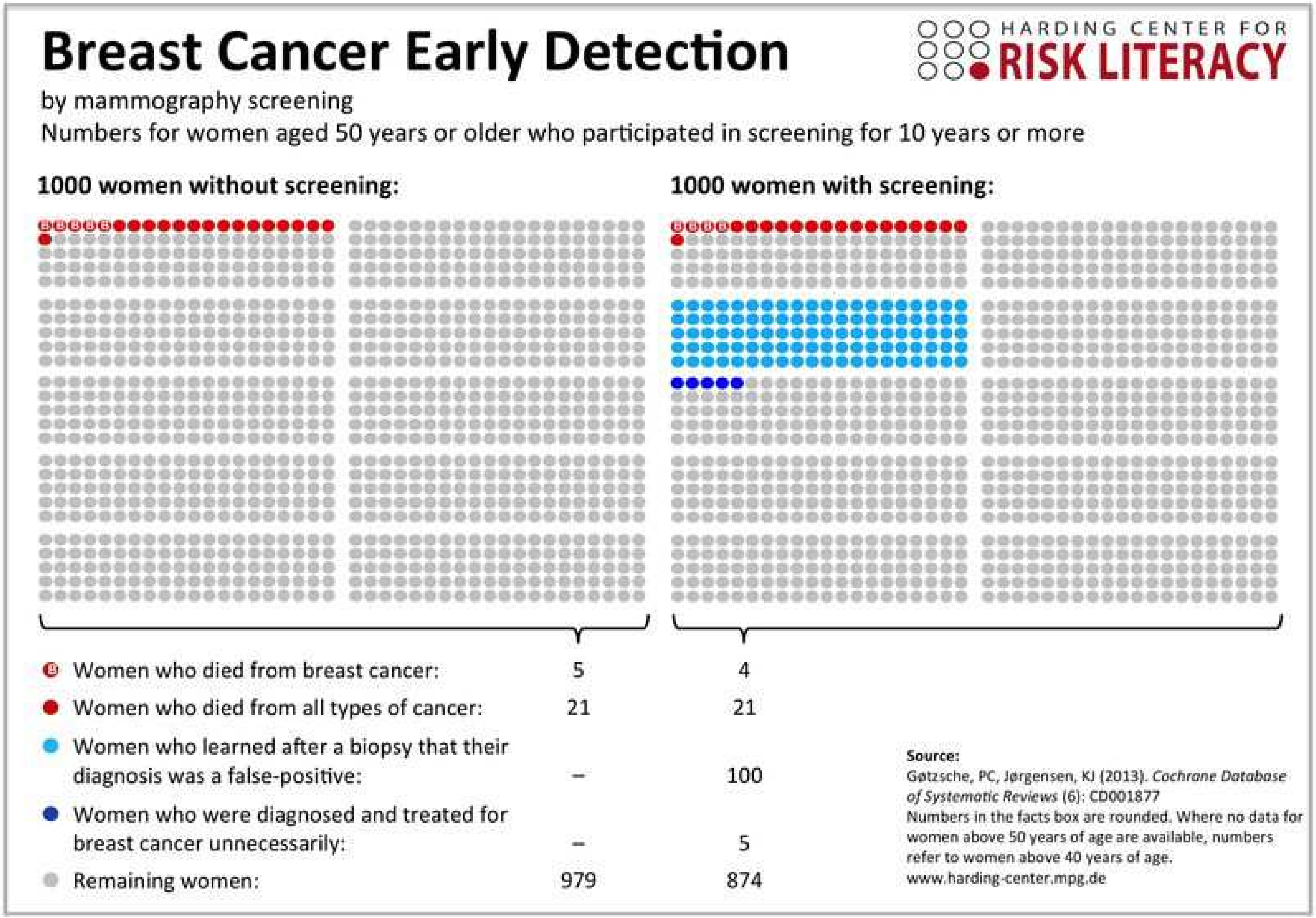
So given that a patient tests positive (198 ppl), there is a 50% chance the patient is sick (99 ppl)!

Diagnoses a la Bayes



The small error rate on the large healthy population produces many false positives.

Natural frequencies a la Gigerenzer[2]



[2] http://bit.ly/ggbbc

Inverting conditional probabilities

Bayes' Theorem

Equate the far right- and left-hand sides of product rule

$$p\left(y|x\right)p\left(x\right) = p\left(x, y\right) = p\left(x|y\right)p\left(y\right)$$

and divide to get the probability of y given x from the probability of x given y:

$$p\left(y|x\right) = \frac{p\left(x|y\right)p\left(y\right)}{p\left(x\right)}$$

where $p\left(x\right) = \sum_{y \in \Omega_Y} p\left(x|y\right)p\left(y\right)$ is the normalization constant.

Diagnoses a la Bayes

Given that a patient tests positive, what is probability the patient is sick?

$$p(sick|+) = \frac{\overbrace{p(+|sick)}^{99/100}\overbrace{p(sick)}^{1/100}}{\underbrace{p(+)}_{99/100^2+99/100^2=198/100^2}} = \frac{99}{198} = \frac{1}{2}$$

where $p(+) = p(+|sick)\,p(sick) + p(+|healthy)\,p(healthy)$.

(Super) Naive Bayes

We can use Bayes' rule to build a one-word spam classifier:

$$p\left(spam|word\right) = \frac{p\left(word|spam\right) p\left(spam\right)}{p\left(word\right)}$$

where we estimate these probabilities with ratios of counts:

$$\begin{aligned}
\hat{p}(word|spam) &= \frac{\text{\# spam docs containing word}}{\text{\# spam docs}} \\
\hat{p}(word|ham) &= \frac{\text{\# ham docs containing word}}{\text{\# ham docs}} \\
\hat{p}(spam) &= \frac{\text{\# spam docs}}{\text{\# docs}} \\
\hat{p}(ham) &= \frac{\text{\# ham docs}}{\text{\# docs}}
\end{aligned}$$

(Super) Naive Bayes

```
$ ./enron_naive_bayes.sh meeting
1500 spam examples
3672 ham examples
16 spam examples containing meeting
153 ham examples containing meeting

estimated P(spam) = .2900
estimated P(ham) = .7100
estimated P(meeting|spam) = .0106
estimated P(meeting|ham) = .0416

P(spam|meeting) = .0923
```

(Super) Naive Bayes

```
$ ./enron_naive_bayes.sh money
1500 spam examples
3672 ham examples
194 spam examples containing money
50 ham examples containing money

estimated P(spam) = .2900
estimated P(ham) = .7100
estimated P(money|spam) = .1293
estimated P(money|ham) = .0136

P(spam|money) = .7957
```

(Super) Naive Bayes

```
$ ./enron_naive_bayes.sh enron
1500 spam examples
3672 ham examples
0 spam examples containing enron
1478 ham examples containing enron

estimated P(spam) = .2900
estimated P(ham) = .7100
estimated P(enron|spam) = 0
estimated P(enron|ham) = .4025

P(spam|enron) = 0
```

Naive Bayes

Represent each document by a binary vector $\vec{x}$ where $x_j = 1$ if the j-th word appears in the document ($x_j = 0$ otherwise).

Modeling each word as an *independent* Bernoulli random variable, the probability of observing a document $\vec{x}$ of class c is:

$$p(\vec{x}|c) = \prod_j \theta_{jc}^{x_j} (1 - \theta_{jc})^{1-x_j}$$

where θ_{jc} denotes the probability that the j-th word occurs in a document of class c.

Naive Bayes

Using this likelihood in Bayes' rule and taking a logarithm, we have:

$$
\begin{aligned}
\log p(c|\vec{x}) &= \log \frac{p(\vec{x}|c)\, p(c)}{p(\vec{x})} \\
&= \sum_j x_j \log \frac{\theta_{jc}}{1-\theta_{jc}} + \sum_j \log(1-\theta_{jc}) + \log \frac{\theta_c}{p(\vec{x})}
\end{aligned}
$$

where θ_c is the probability of observing a document of class c.

(a) big picture: surrogate convex loss functions

general

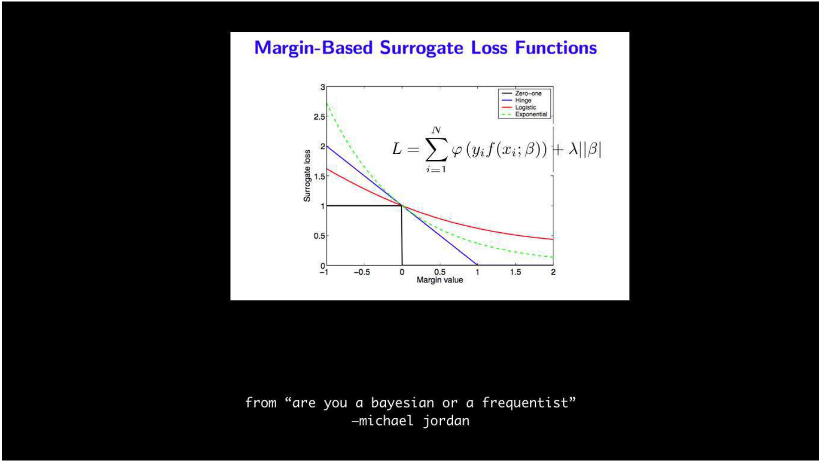


Figure 4: Reminder: Surrogate Loss Functions

boosting

A decision-theoretic generalization of on-line learning and an application to boosting*

Yoav Freund Robert E. Schapire

AT&T Labs
180 Park Avenue
Florham Park, NJ 07932
{yoav, schapire}@research.att.com

December 19, 1996

Figure 5: 'Cited by 12599'

tangent: logistic function as surrogate loss function

- define $f(x) \equiv \log p(y=1|x)/p(y=-1|x) \in R$

tangent: logistic function as surrogate loss function

- define $f(x) \equiv \log p(y=1|x)/p(y=-1|x) \in R$
- $p(y=1|x) + p(y=-1|x) = 1 \to p(y|x) = 1/(1+\exp(-yf))$

tangent: logistic function as surrogate loss function

- define $f(x) \equiv \log p(y=1|x)/p(y=-1|x) \in R$
- $p(y=1|x) + p(y=-1|x) = 1 \rightarrow p(y|x) = 1/(1+\exp(-yf))$
- $-\log_2 p(\{y\}_1^N) = \sum_i \log_2 \left(1 + \mathrm{e}^{-y_i f(x_i)}\right) \equiv \sum_i \ell(y_i f(x_i))$

tangent: logistic function as surrogate loss function

- define $f(x) \equiv \log p(y=1|x)/p(y=-1|x) \in R$
- $p(y=1|x) + p(y=-1|x) = 1 \rightarrow p(y|x) = 1/(1+\exp(-yf))$
- $-\log_2 p(\{y\}_1^N) = \sum_i \log_2 \left(1 + \mathrm{e}^{-y_i f(x_i)}\right) \equiv \sum_i \ell(y_i f(x_i))$
- $\ell'' > 0$, $\ell(\mu) > 1[\mu < 0]\ \forall \mu \in R$.

tangent: logistic function as surrogate loss function

- define $f(x) \equiv \log p(y=1|x)/p(y=-1|x) \in R$
- $p(y=1|x) + p(y=-1|x) = 1 \rightarrow p(y|x) = 1/(1+\exp(-yf))$
- $-\log_2 p(\{y\}_1^N) = \sum_i \log_2 \left(1 + \mathrm{e}^{-y_i f(x_i)}\right) \equiv \sum_i \ell(y_i f(x_i))$
- $\ell'' > 0$, $\ell(\mu) > 1[\mu < 0] \ \forall \mu \in R$.
- $\therefore$ maximizing log-likelihood is minimizing a surrogate convex loss function for classification

tangent: logistic function as surrogate loss function

- define $f(x) \equiv \log p(y = 1|x)/p(y = -1|x) \in R$
- $p(y = 1|x) + p(y = -1|x) = 1 \rightarrow p(y|x) = 1/(1 + \exp(-yf))$
- $-\log_2 p(\{y\}_1^N) = \sum_i \log_2 \left(1 + \mathrm{e}^{-y_i f(x_i)}\right) \equiv \sum_i \ell(y_i f(x_i))$
- $\ell'' > 0$, $\ell(\mu) > 1[\mu < 0] \; \forall \mu \in R$.
- $\therefore$ maximizing log-likelihood is minimizing a surrogate convex loss function for classification
- but $\sum_i \log_2 \left(1 + \mathrm{e}^{-y_i w^T h(x_i)}\right)$ not as easy as $\sum_i \mathrm{e}^{-y_i w^T h(x_i)}$

boosting 1

L exponential surrogate loss function, summed over examples:

- $L[F] = \sum_i \exp\left(-y_i F(x_i)\right)$

boosting 1

L exponential surrogate loss function, summed over examples:

- $L[F] = \sum_i \exp\left(-y_i F(x_i)\right)$
- $= \sum_i \exp\left(-y_i \sum_{t'}^{t} w_{t'} h_{t'}(x_i)\right) \equiv L_t(\mathbf{w}_t)$

boosting 1

L exponential surrogate loss function, summed over examples:

- $L[F] = \sum_i \exp\left(-y_i F(x_i)\right)$
- $= \sum_i \exp\left(-y_i \sum_{t'}^{t} w_{t'} h_{t'}(x_i)\right) \equiv L_t(\mathbf{w}_t)$
- Draw $h_t \in \mathcal{H}$ large space of rules s.t. $h(x) \in \{-1, +1\}$

boosting 1

L exponential surrogate loss function, summed over examples:

- $L[F] = \sum_i \exp\left(-y_i F(x_i)\right)$
- $= \sum_i \exp\left(-y_i \sum_{t'}^{t} w_{t'} h_{t'}(x_i)\right) \equiv L_t(\mathbf{w}_t)$
- Draw $h_t \in \mathcal{H}$ large space of rules s.t. $h(x) \in \{-1, +1\}$
- label $y \in \{-1, +1\}$

boosting 1

L exponential surrogate loss function, summed over examples:

- $L_{t+1}(\mathbf{w}_t; w) \equiv \sum_i d_i^t \exp\left(-y_i w h_{t+1}(x_i)\right)$

Punchlines: sparse, predictive, interpretable, fast (to execute), and easy to extend, e.g., trees, flexible hypotheses spaces, L_1, L_∞[4], ...

boosting 1

L exponential surrogate loss function, summed over examples:

- $L_{t+1}(\mathbf{w}_t; w) \equiv \sum_i d_i^t \exp\left(-y_i w h_{t+1}(x_i)\right)$
- $= \sum_{y=h'} d_i^t \mathrm{e}^{-w} + \sum_{y \neq h'} d_i^t \mathrm{e}^{+w} \equiv \mathrm{e}^{-w} D_+ + \mathrm{e}^{+w} D_-$

Punchlines: sparse, predictive, interpretable, fast (to execute), and easy to extend, e.g., trees, flexible hypotheses spaces, L_1, L_∞[4], ...

---

boosting 1

L exponential surrogate loss function, summed over examples:

- $L_{t+1}(\mathbf{w}_t; w) \equiv \sum_i d_i^t \exp\left(-y_i w h_{t+1}(x_i)\right)$
- $= \sum_{y=h'} d_i^t \mathrm{e}^{-w} + \sum_{y \neq h'} d_i^t \mathrm{e}^{+w} \equiv \mathrm{e}^{-w} D_+ + \mathrm{e}^{+w} D_-$
- $\therefore w_{t+1} = \mathrm{argmin}_w L_{t+1}(w) = (1/2) \log D_+ / D_-$

Punchlines: sparse, predictive, interpretable, fast (to execute), and easy to extend, e.g., trees, flexible hypotheses spaces, L_1, L_∞[4], ...

---

boosting 1

L exponential surrogate loss function, summed over examples:

- $L_{t+1}(\mathbf{w}_t; w) \equiv \sum_i d_i^t \exp\left(-y_i w h_{t+1}(x_i)\right)$
- $= \sum_{y=h'} d_i^t \mathrm{e}^{-w} + \sum_{y \neq h'} d_i^t \mathrm{e}^{+w} \equiv \mathrm{e}^{-w} D_+ + \mathrm{e}^{+w} D_-$
- $\therefore w_{t+1} = \mathrm{argmin}_w L_{t+1}(w) = (1/2) \log D_+ / D_-$
- $L_{t+1}(\mathbf{w}_{t+1}) = 2\sqrt{D_+ D_-} = 2\sqrt{\nu_+(1-\nu_+)}/D$, where $0 \leq \nu_+ \equiv D_+/D = D_+/L_t \leq 1$

Punchlines: sparse, predictive, interpretable, fast (to execute), and easy to extend, e.g., trees, flexible hypotheses spaces, L_1, L_∞[4], ...

boosting 1

L exponential surrogate loss function, summed over examples:

- $L_{t+1}(\mathbf{w}_t; w) \equiv \sum_i d_i^t \exp\left(-y_i w h_{t+1}(x_i)\right)$
- $= \sum_{y=h'} d_i^t \mathrm{e}^{-w} + \sum_{y \neq h'} d_i^t \mathrm{e}^{+w} \equiv \mathrm{e}^{-w} D_+ + \mathrm{e}^{+w} D_-$
- $\therefore w_{t+1} = \mathrm{argmin}_w L_{t+1}(w) = (1/2) \log D_+ / D_-$
- $L_{t+1}(\mathbf{w}_{t+1}) = 2\sqrt{D_+ D_-} = 2\sqrt{\nu_+(1 - \nu_+)}/D$, where $0 \leq \nu_+ \equiv D_+/D = D_+/L_t \leq 1$
- update example weights $d_i^{t+1} = d_i^t \mathrm{e}^{\mp w}$

Punchlines: sparse, predictive, interpretable, fast (to execute), and easy to extend, e.g., trees, flexible hypotheses spaces, L_1, L_∞[4], ...

[4] Duchi + Singer "Boosting with structural sparsity" ICML '09

svm