Inf2b Learning and Data

http://www.inf.ed.ac.uk/teaching/courses/inf2b/

Lecture 1

Introduction to Learning and Data

Iain Murray, 2013

School of Informatics, University of Edinburgh

Welcome to Inf2b!

Today's Schedule:

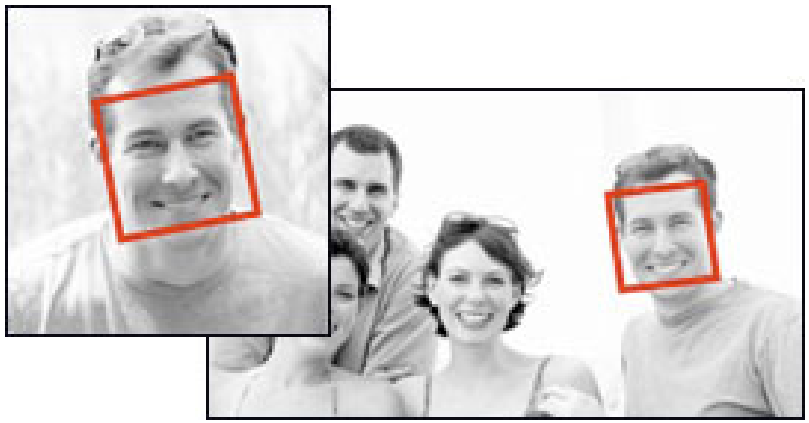
- What is learning? (and why should you care?)
- Administrative stuff
  - How to do well
- Setting up a learning problem

(time allowing)

Face detection

How would you detect a face?

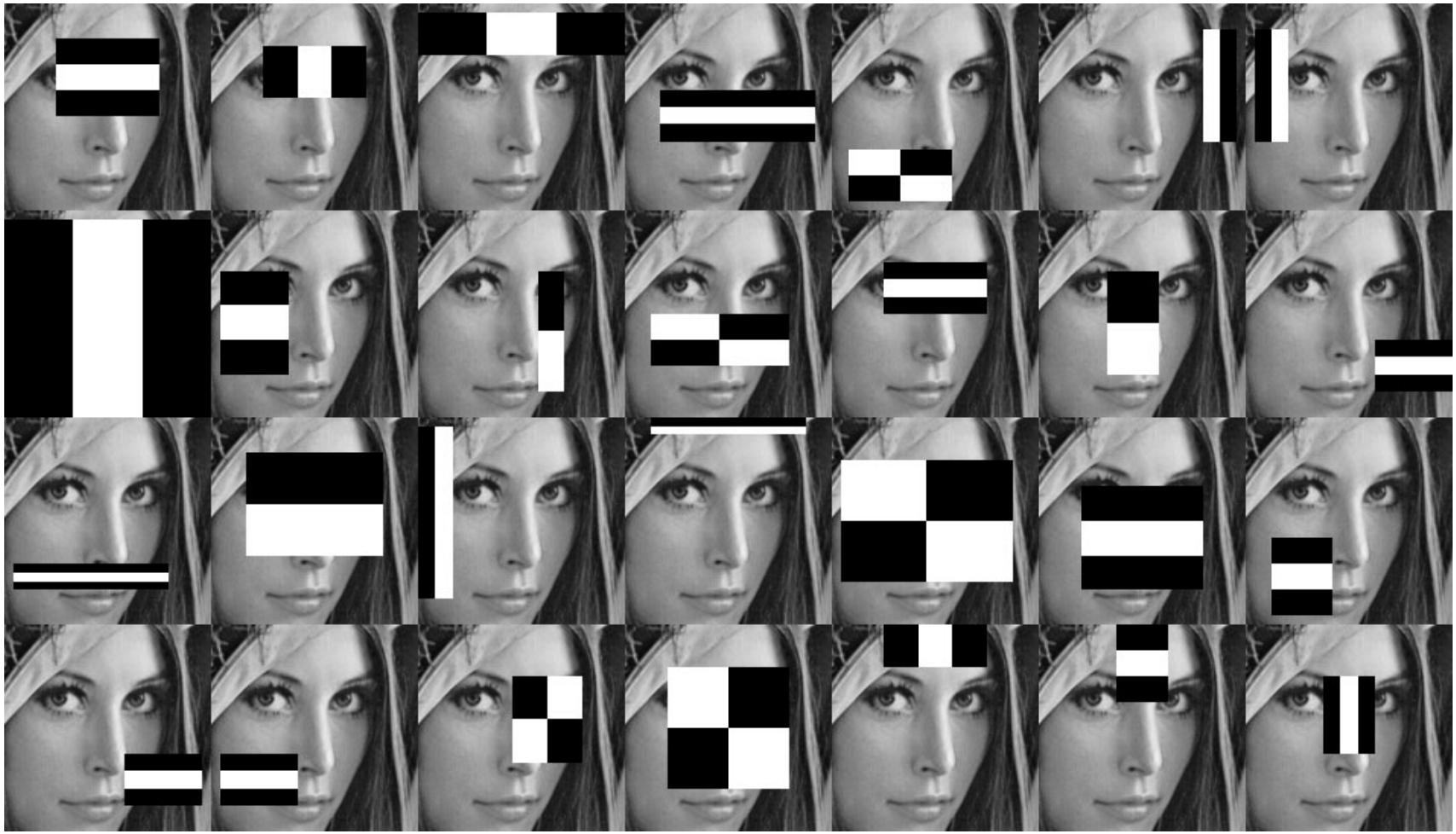
(R. Vaillant, C. Monrocq and Y. LeCun, 1994)



How does album software tag your friends?

http://demo.pittpatt.com/

Viola–Jones Face detection



Taken from: http://v10.ahprojects.com/art/cv-dazzle

A neat algorithm & data structure

Rectangle intensity:

naively need to add 10^3 to 10^6 pixels

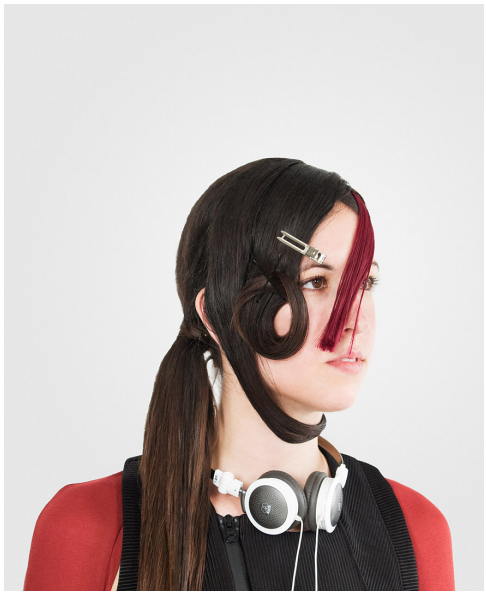
Pre-computation: *Integral Image*,

add/subtract 4 values $\Rightarrow$ rectangle intensity

`http://en.wikipedia.org/wiki/Summed_area_table`

Hiding from the machines



Taken from: http://v10.ahprojects.com/art/cv-dazzle

How does human vision work?

http://brain.mada.org.il/upside-down-e.html

Intro summary

Machine learning:

— Fit numbers in a program to data

— More robust than hand-fitted rules

— Can't approach humans at some tasks (e.g., vision)

— Machines make better predictions in many other cases

Applications of machine learning

Within informatics

— **Vision** as we've seen

— **Graphics** increasingly data driven

— **Robotics** vision, planning, control, . . .

— **Compilers** learning how to optimize

and beyond: data analysis across the sciences

Every day

— Adverts / recommendations all over the web

— Discounts in Tescos

— Speech recognition, Machine Translation, . . .

with self-driving cars 'soon'?

Welcome to Inf2b!

Today's Schedule:

- What is learning? (and why should you care?)
- **Administrative stuff**
  - How to do well
- Setting up a learning problem

(time allowing)

Course structure

Website:

`http://www.inf.ed.ac.uk/teaching/courses/inf2b/`

Constituents:

- 30 lectures (including review)
- Tutorials starting in week 2
- 2 assessed assignments

Equal split into two threads:

- Algorithms & Data Structures — KK (Kyriakos Kalorkoti)
- Learning and Data — Iain

Private study

$\sim$2 hours private study per lecture,

in addition to tutorials & assignments!

No required textbook for Inf2b

There are notes. See those for recommended books.

Come to lectures! (really, skipping lectures is a *bad* idea)

Feedback:
ask questions, use tutorials, NB (for learning only), class reps

Class reps

WANTED: Inf2b class reps (for ADS & learning)

Email: `i.murray@ed.ac.uk`

your name, degree, email address.

Two hours study this week?

Start to familiarize yourself with MATLAB (or OCTAVE)

Introductory worksheet on the course website

Many others at the end of a web search

Love Python? Learn NUMPY+SCIPY+MATPLOTLIB
(instead, or as well)

Vital skills:

— add, average, multiply vectors and matrices
— plot data stored in vectors
— save/read data to/from files

Welcome to Inf2b!

Today's Schedule:

- What is learning? (and why should you care?)
- Administrative stuff
  - How to do well
- Setting up a learning problem

(time allowing)

The Netflix Prize

"We're quite curious, really. To the tune of one million dollars.

It's "easy" really. We provide you with a lot of anonymous rating data, and a prediction accuracy bar that is 10% better than what Cinematch can do on the same training data set."

`http://www.netflixprize.com`, October 2006.

Kaggle

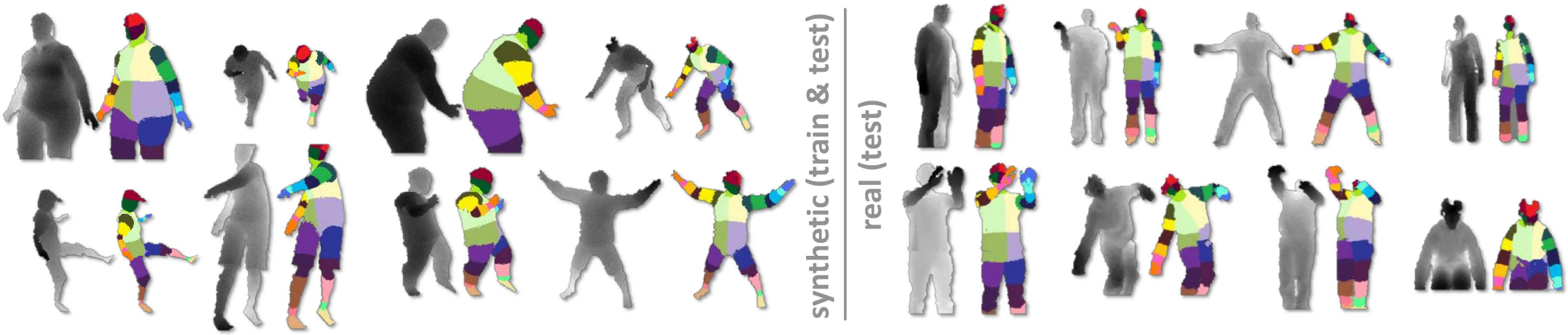
Crowd-sourcing data-science solutions:

`http://kaggle.com/`

Creating training data

Microsoft Kinect (Shotton et al., CVPR 2011)

`http://research.microsoft.com/apps/pubs/default.aspx?id=145347`



Random forest applied to fantasies

Summary of setting

Each challenge has:

— A measure of success

Objective function, cost function, metric, . . .

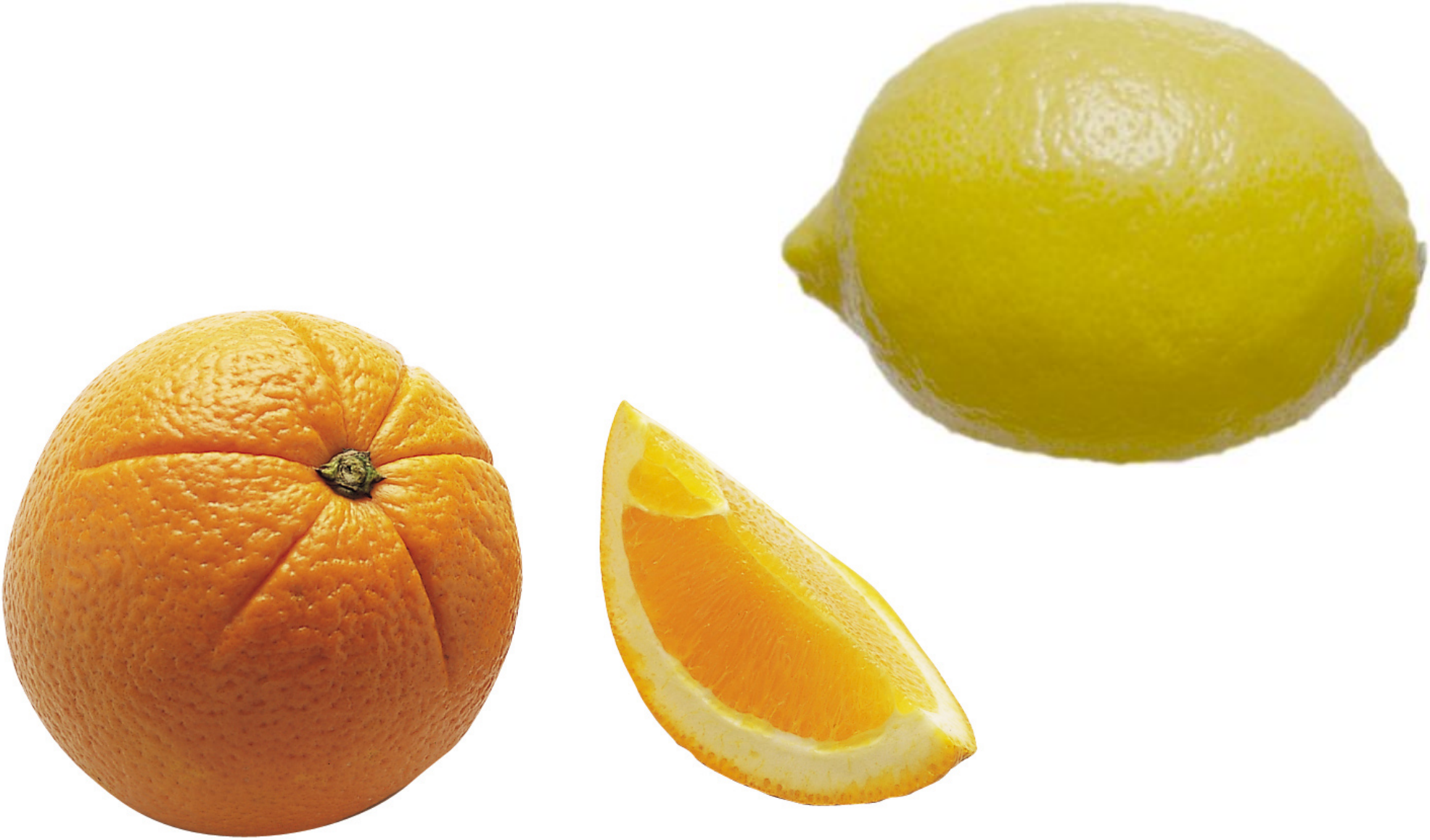
— Data is useful (but needs to be available)

— Nothing is certain

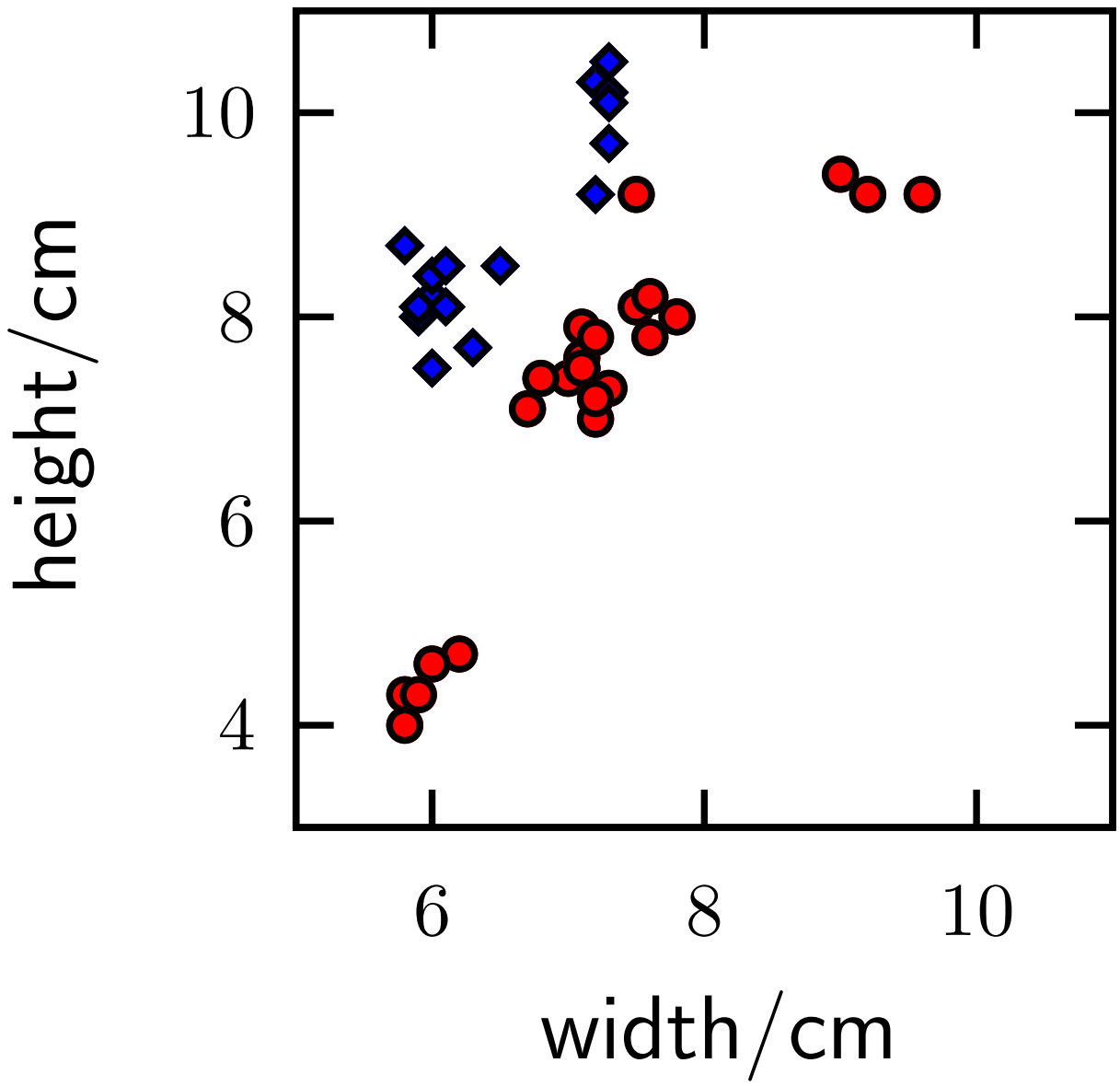
we will use probability a lot

How does a machine use the data?

Oranges and Lemons

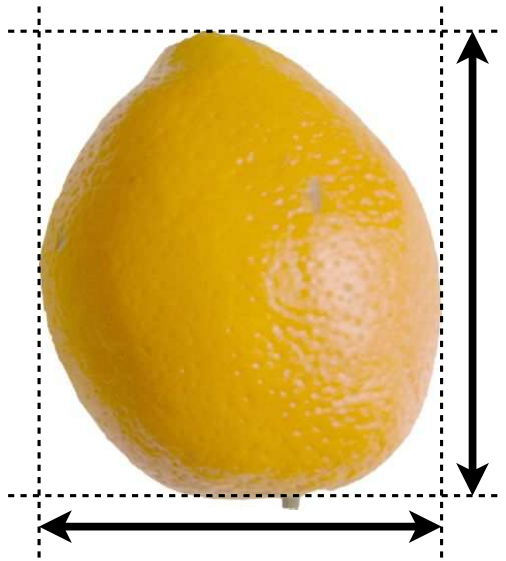


A two-dimensional space



Oranges:

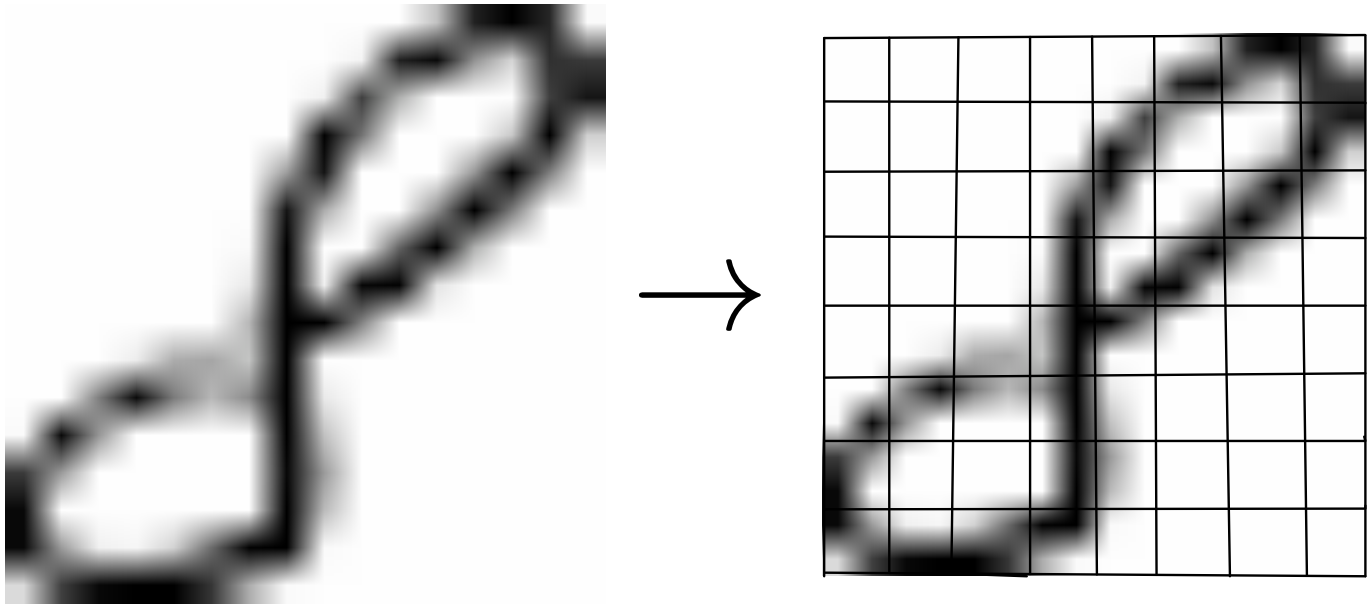
Lemons:



Handwritten digits

http://alex.seewald.at/digits/

A 64-dimensional space



Turn each cell into a number (somehow, see notes)

Unravel into a column vector, a **feature vector**

$\Rightarrow$ represented digit as point in $64D$

`http://alex.seewald.at/digits/`

Euclidean distance

Distance between $2D$ vectors: (x, y) and (x', y')

$$r_2 = \sqrt{(x - x')^2 + (y - y')^2}$$

Distance between D-dimensional vectors: $\mathbf{x}$ and $\mathbf{x}'$

$$r_2(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{d=1}^{D} (x_d - x'_d)^2}$$

Measures similarities between feature vectors

i.e., similarities between digits, movies, sounds, galaxies, . . .

Question

Have high-resolution scans of digits.

How many pixels should be sample?

What are pros and cons of:

2×2, 4×4, 16×16, or 100×100?