Machine Learning and Econometrics

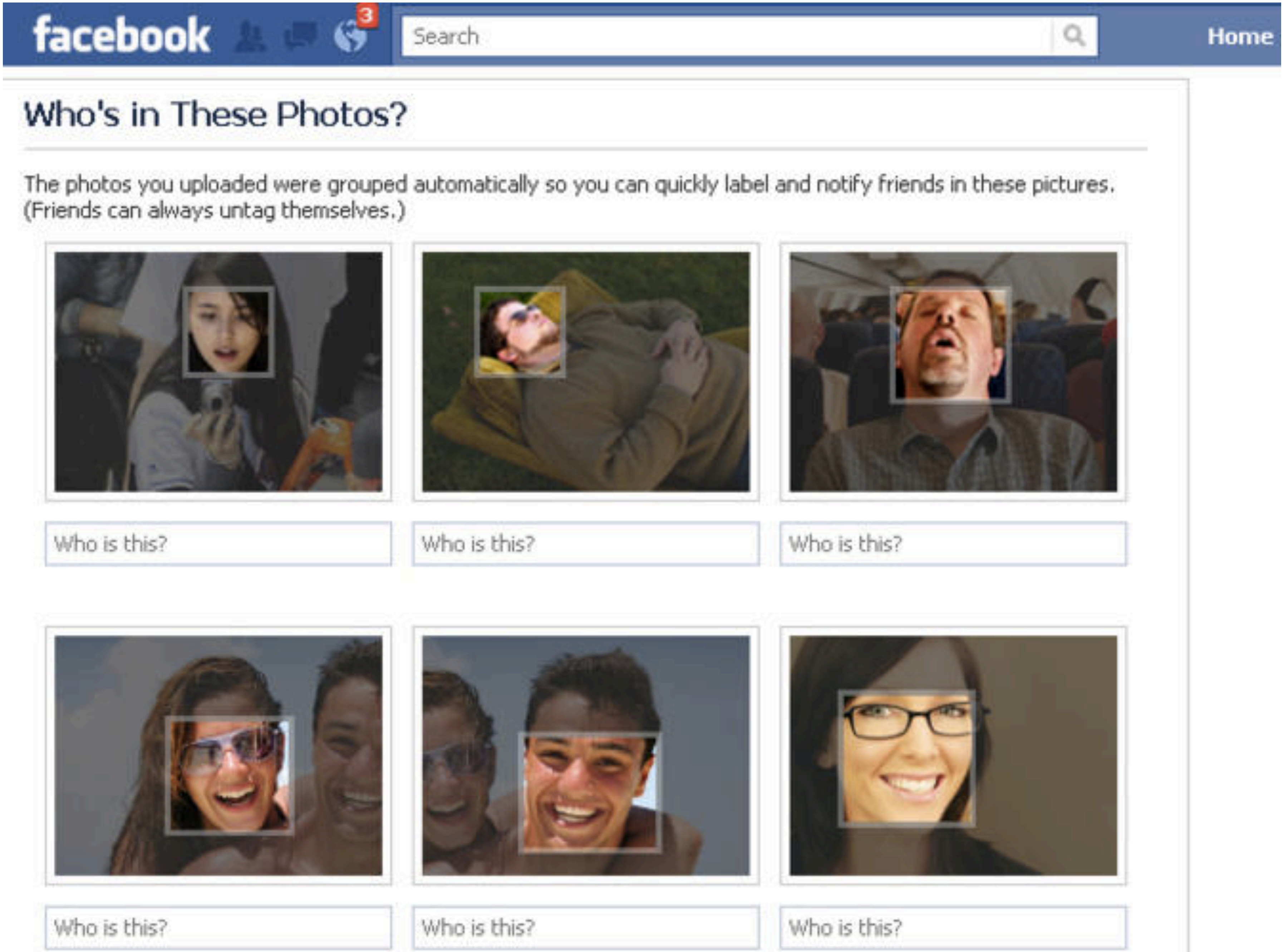
Sendhil Mullainathan
(with Jann Spiess)

Outline

- How did I get interested in this?
- What is the secret sauce of machine learning?
- Where is machine learning useful in economics?

Outline

- How did I get interested in this?
- What is the secret sauce of machine learning?
- Where is machine learning useful in economics?


facebook
3
Search
Home
Who's in These Photos?
The photos you uploaded were grouped automatically so you can quickly label and notify friends in these pictures.
(Friends can always untag themselves.)
Who is this?
Who is this?
Who is this?
Who is this?
Who is this?
Who is this?


984

Describes without errors	Describes with minor errors	Somewhat related to the image	Unrelated to the image
 A person riding a motorcycle on a dirt road.	 Two dogs play in the grass.	 A skateboarder does a trick on a ramp.	 A dog is jumping to catch a frisbee.
 A group of young people playing a game of frisbee.	 Two hockey players are fighting over the puck.	 A little girl in a pink hat is blowing bubbles.	 A refrigerator filled with lots of food and drinks.
 A herd of elephants walking across a dry grass field.	 A close up of a cat laying on a couch.	 A red motorcycle parked on the side of the road.	 A yellow school bus parked in a parking lot.

Magic?

- Hard not to be wowed
- But what makes them tick?
- Could that be used elsewhere? In my own work?
- Look at something simpler than vision

Click to open expanded view

Hutzler 571 Banana Slicer

by Hutzler

4,548 customer reviews

| 368 answered questions

List Price: ~~$4.99~~
Price: $3.85 & **FREE Shipping** on orders over $35. Details
You Save: $1.14 (23%)

In Stock.

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

Want it Wednesday, Dec. 4? Order within **20 hrs 53 mins** and choose **One-Day Shipping** at checkout. Details

- Faster, safer than using a knife
- Great for cereal
- Plastic, dishwasher safe
- Slice your banana with one quick motion
- Kids love slicing their own bananas

Confusing

…ptember 11, 2012

There is no way to tell if this is a standard or metric banana slicer. Additional markings on it would help greatly.

endeavor was spent cleaning this implement. It is not easy to clean--you have to scrub between every rung to thoroughly clean it.

AI Approach

- We do it perfectly.
- How do we do it?
- Introspect
- Let's program that up.

Programming

- For each review make a vector of words

- Figure out whether it has positive words and negative words

- Count

Trying to Copy Humans

Brilliant

Dazzling

Cool

Gripping

Moving

60%

Suck

Cliched

Slow

Awful

Bad

What is so hard?

- Decide what words makes for a positive review
  - What combination of words do you look for?
    - “Some people say this talk was great”

- This problem was endemic to every problem
  - Driving a car: What is a tree?
  - Language: Which noun does this pronoun modify?

This Approach Stalled

- “Trivial” problems proved impossible
  – Marvin Minsky once assigned "the problem of computer vision" as a summer project

- Forget about the more complicated problems like language

What is the magic trick?

- Make this an empirical exercise
  - Collect some data
- Example dataset:
  - 2000 movie reviews
  - 1000 good and 1000 bad reviews
- Now just ask what combination of words predicts being a good review

Learning not Programming

STILL

Love

Superb

Great

95%

Bad

Stupid

Worst

?

!

Pang, Lee and Vaithyanathan

Machine learning

- Turn any "intelligence" task into an empirical learning task
  - Specify what is to be predicted
  - Specify what is used to predict it

ML drives many innovations…

- Every domain
  - Post office uses machines to read addresses
  - Voice recognition (Siri)
  - Spam filters
  - Recommender systems
  - Driverless cars
- Not a coincidence that ML and big data arose together

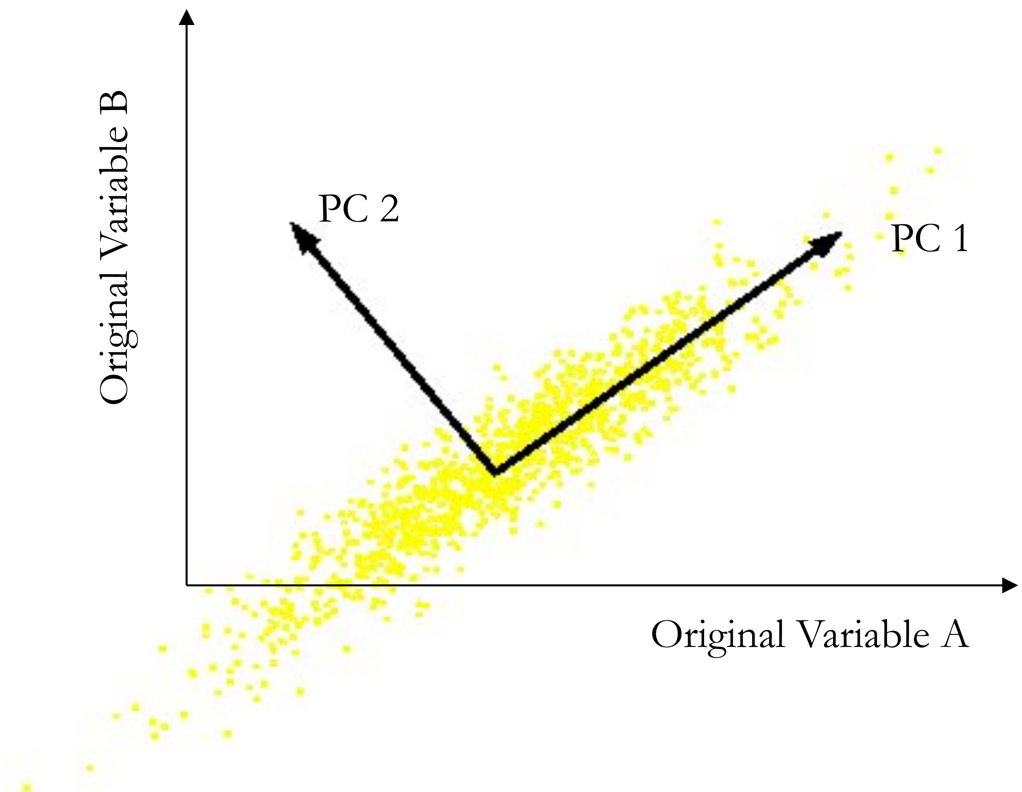
Wonderful

- Great that they discovered the 100+ year old field of statistics!

- We've been estimating functions from data for a long time

- KEY: This is in part definitely true

Principal Component Analysis Example



• Orthogonal directions of greatest variance in data
• Projections along PC1 discriminate the data most along   any one axis

PCA

- The intuitions usually come from two dimensions
- But in very high dimensions thought this can get very interesting…

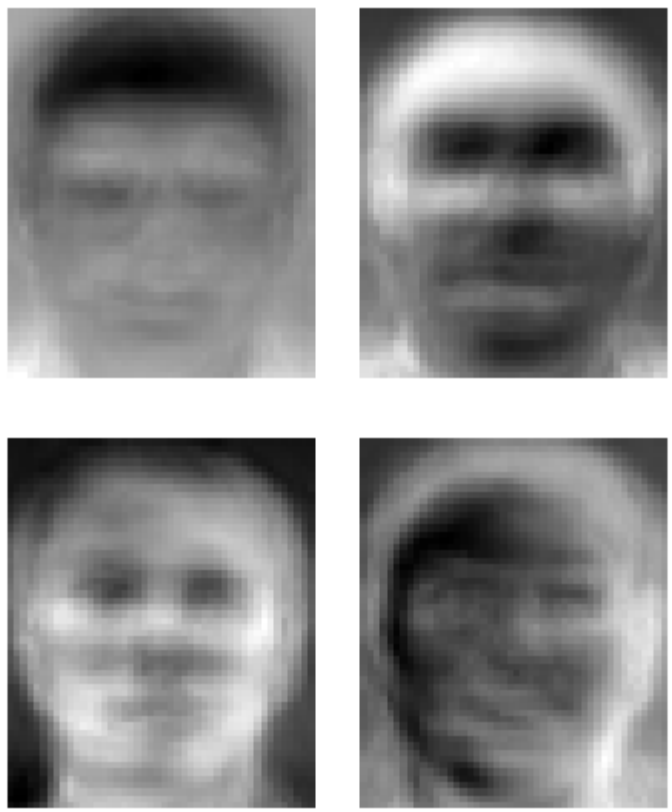
PCA applications -Eigenfaces

1. Large set of digitized images of human faces is taken under the same lighting conditions.
2. The images are normalized to line up the eyes and mouths.
3. The eigenvectors of the covariance matrix of the statistical distribution of face image vectors are then extracted.
4. These eigenvectors are called eigenfaces.

Vectorization key building block

PCA applications -Eigenfaces

- The principal eigenface looks like a bland androgynous average human face



http://en.wikipedia.org/wiki/Image:Eigenfaces.png

```
coefficients =
 {-6.85693, 23.7498, -11.4515, -3.43352, 5.24749, -7.1615,
  8.09015, -9.7205, -0.660834, -2.4148, -10.3942, 3.33424,
  2.94988, -2.75981, 3.02687, -2.4499, -2.09885, -5.98832,
  -4.22564, -0.65014, 2.20144, -5.43782, -9.61821, -3.25227,
  7.49413, -0.145002, 7.61483, -0.696994, -3.7731, 3.23569,
  -1.78853, 0.0400116, -3.86804, -2.02456, 2.20949, -1.86902,
  1.23445, 0.140996, 0.698304, -0.420466, 2.30691, 3.70434,
  1.02417, 0.382809, 0.413049, -0.994902, 0.754145, 0.363418,
  -0.383865, 1.46379, 1.96381, -2.90388, -2.33381, -0.438939,
  -0.30523, -0.105925, 0.665962, -0.729409, -1.28977, 0.150497,
  0.645343, 0.30724, -1.04942, 1.0462, -0.60808, 0.333288,
  1.09659, -1.38876, 0.33875, 0.278604, 1.0632, -0.0446148,
  0.24526, -0.283482, -0.236843, 0.312122};
```

Wonderful

- Great that they discovered the 100+ year old field of statistics!
- We've been estimating functions from data for a long time
- KEY: This is in part definitely true
- But in important ways *not* true

	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	**78.7**	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	**82.9**
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	**82.7**
(4)	bigrams	16165	pres.	77.3	**77.4**	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	**81.9**
(6)	adjectives	2633	pres.	77.0	**77.7**	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	**81.4**
(8)	unigrams+position	22430	pres.	81.0	80.1	**81.6**

ASIDE: Vectorization

Pang, Lee and Vaithyanathan

NOTE: <u>Large</u> sets of variables

Why high dimensional data analysis should not really be possible

- Easiest to see in linear case

- If you have n data points and k variables, then $X'X$ is not invertible
  - Size $k+1$ by $k+1$
  - But rank is at most $n < k+1$

Face Recognition

- Very simple problem

$$
\begin{aligned}
Y &= \underbrace{\{0,1\}}_{Face?} \\
X &= \underbrace{\{0,1,..,g\}}_{\text{gray scale}}{}^{24*24} \\
\hat{f} &= argmin_f E[L(f(x),y)]
\end{aligned}
$$

Vectorization Again

Some (possibly asymmetric) loss
for correctly or incorrectly guessing

Face Recognition Dataset

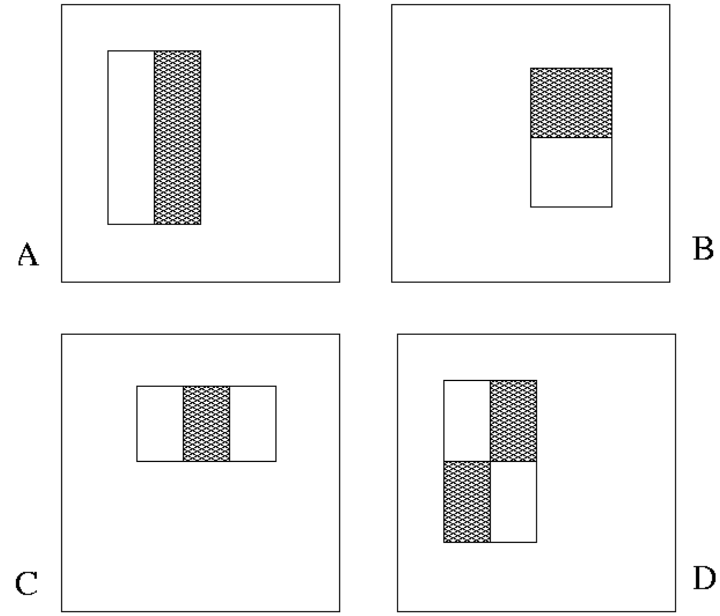
- Sample size:
  - 5000 face photos (+ many non-face photos)
- Number of variables:
  - 24x24 pixel array
  - So 576 variables (with values ranging up to g)
  - Or 576*g variables if we allow gray scale
- A bit tight on sample size…

Functions non-linear in these dummies

- But that's only if we use binary variables
- Obviously a face is not going to be well approximated by a linear function of these binary inputs….

Example of Interactions

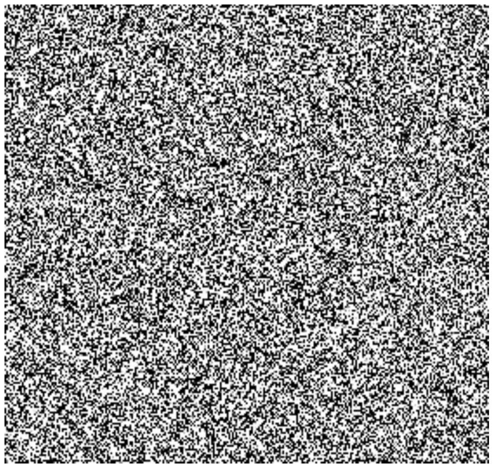
"Rectangle filters"



Value =

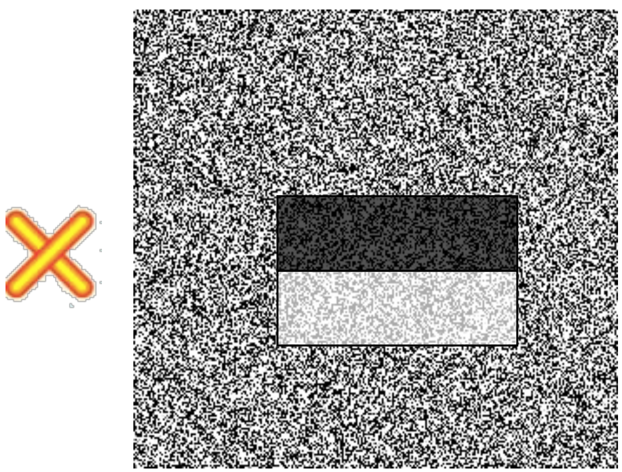
∑ (pixels in white area) –
∑ (pixels in black area)

Example



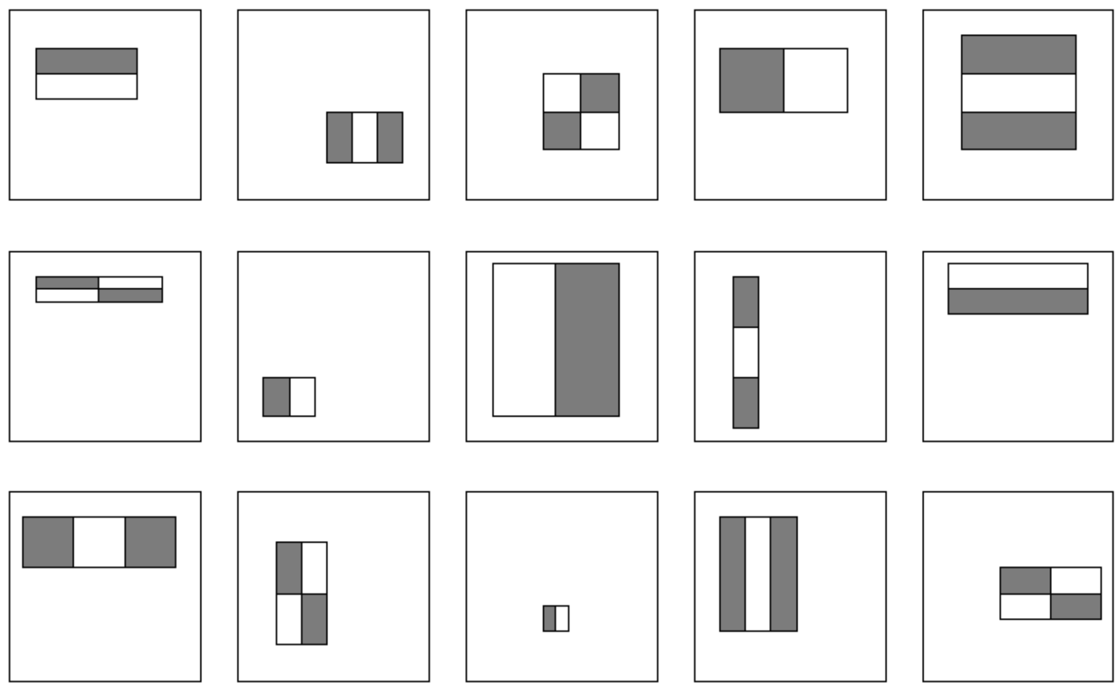
Source

Result



How many variables do we have now?

- For a 24x24 detection region, the number of possible rectangle features is ~160,000!



Something pretty interesting…

- **High dimensional** prediction
  - What does high dimensional mean?
    - **Not** (just) about more variables than data.

  - Really about "effective" number of variables given the functions *f*
    - In linear world = dimensions of function class equals number of variables
- Able to search through many (MANY) possible predictors

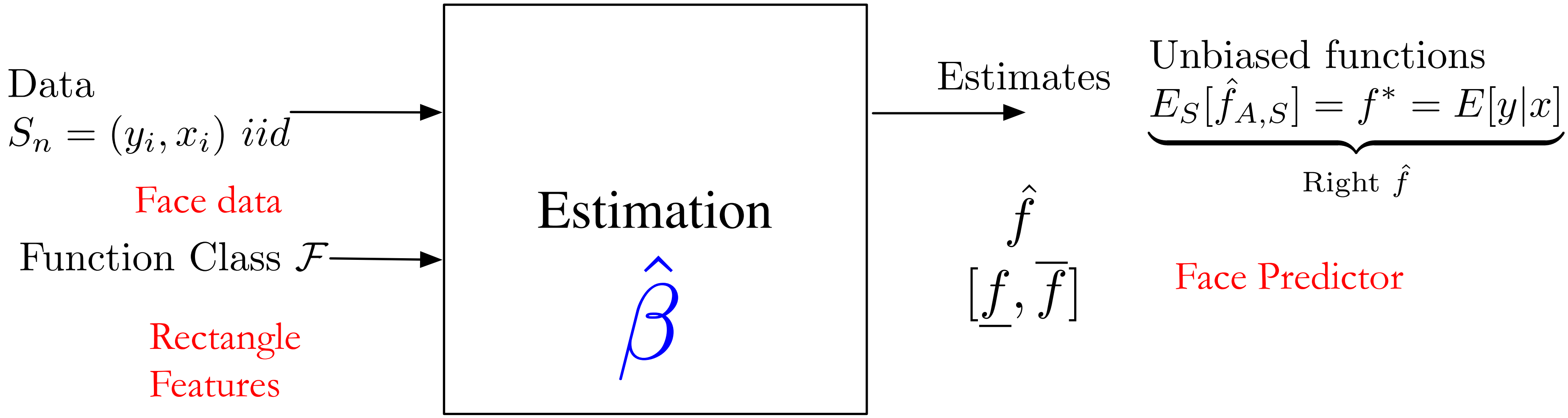
So…

Estimation

- Fit Y with X
- Low dimensional

Machine Learning

- Fit Y with X <u>out of sample</u>
- High dimensional
- JUST BETTER?



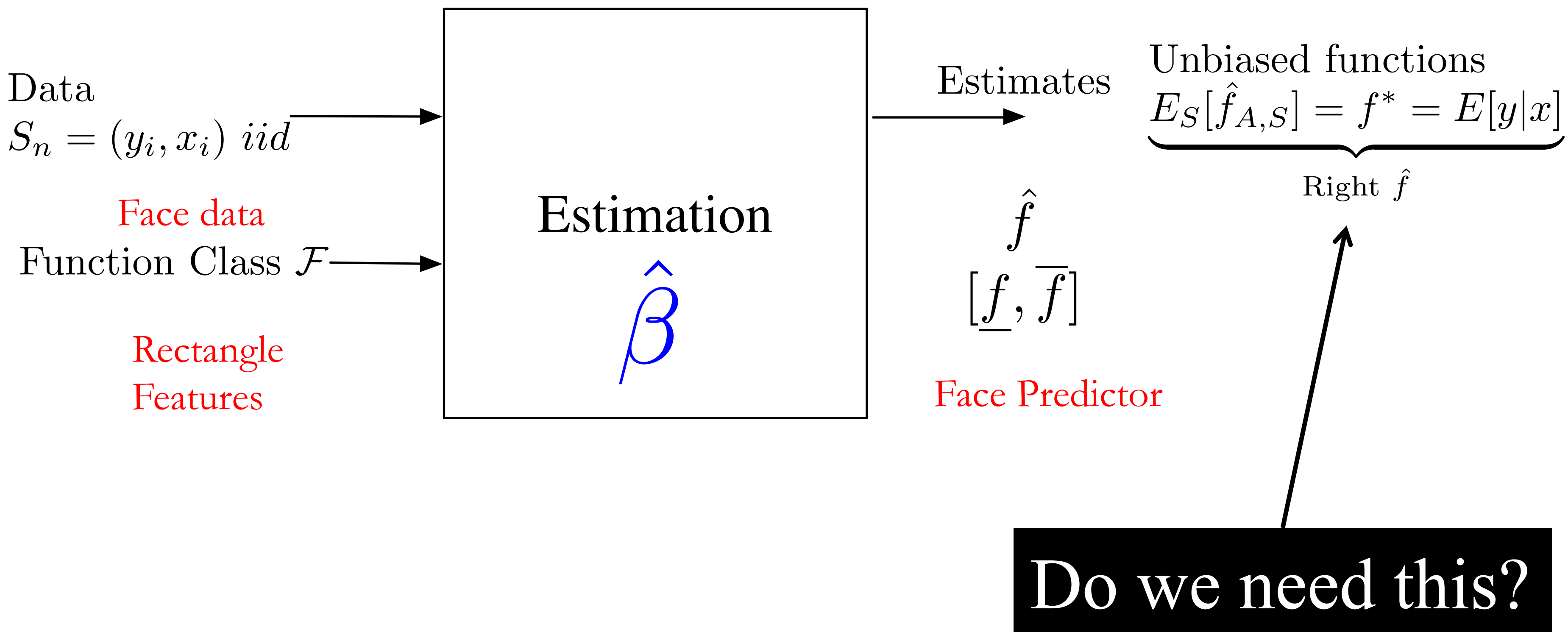
Data size
Information going in

Thousands?

Estimates
Information coming out

Hundreds of Thousands

How can we get more information out than we're putting in?

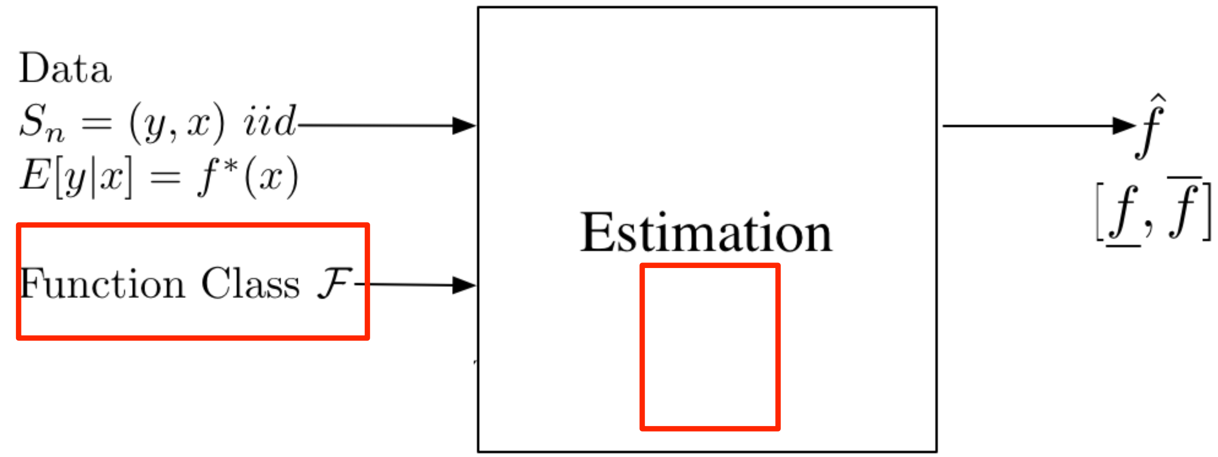

Data
$S_n = (y_i, x_i)$ iid
Face data
Function Class $\mathcal{F}$
Rectangle
Features
Estimation
$\hat{\beta}$
Estimates
$\hat{f}$
$[\underline{f}, \overline{f}]$
Face Predictor
Unbiased functions
$\underbrace{E_S[\hat{f}_{A,S}] = f^* = E[y|x]}_{\text{Right } \hat{f}}$
Do we need this?

Face Recognition

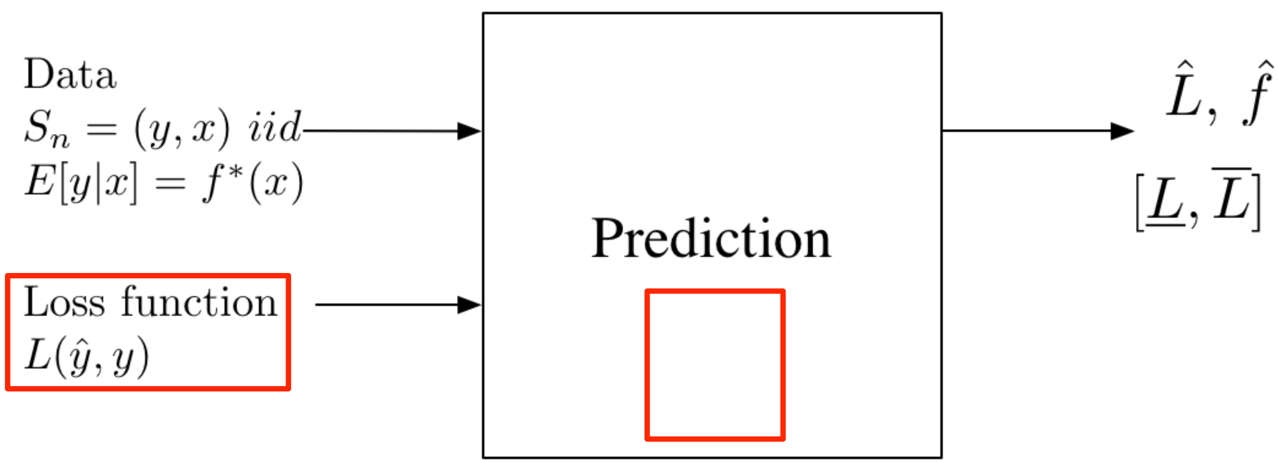
- Problem:

$$
\begin{aligned}
Y &= \underbrace{\{0,1\}}_{Face?} \\
X &= \underbrace{\{0,1,..,g\}}_{\text{gray scale}}^{24*24} \\
\hat{f} &= argmin_f E[L(f(x),y)]
\end{aligned}
$$

Only need good predictions



Gets more out? Put more in



Estimation vs Prediction

Estimation

- Strict assumptions about data generating process
- Back out coefficients
- Low dimensional

$$\hat{\beta}$$

Prediction

- Allow for flexible functional forms
- Get high quality predictions
- Give up on adjudicating between **observably** similar functions (variables)

$$\hat{y}$$

But How?

- This tells us that there’s no free lunch
- But does not tell us mechanically how machine learning works..

Outline

- How did I get interested in this?

- What is the secret sauce of machine learning?

- Where is machine learning useful in economics?

Outline

- How did I get interested in this?
- What is the secret sauce of machine learning?
- Where is machine learning useful in economics?

Understand OLS

AVERAGES NOTATION

$$\hat{\beta}^{\text{OLS}} = \arg\min_{\beta} \mathbb{E}_{S_n}(\beta' x - y)^2$$

$$\beta^*_{\text{prediction}} = \arg\min_{\beta} \mathrm{E}_{(y,x)}(\beta' x - y)^2$$

- The real problem here is minimizing the "wrong" thing: In-sample fit vs out-of-sample fit

Overfit problem

- OLS looks good with the sample you have
  - It's the best you can do *on this sample*

- Problem is OLS by construction overfits
  - We overfit in estimation
  - <u>Where does overfit show up?</u>
  - But in low-dimensional this is not a major problem

This problem is exactly why wide data is troubling

- Why are we worried about having so many variables?

- We'll fit very well (perfectly if $k > n$) **in sample**

- But arbitrarily badly **out of sample**

Understanding overfit

- Let’s consider a general class of algorithms

A General Class of Algorithms

- Consider algorithms of the form

$$\hat{f}_A = \arg\min_{f \in \mathcal{F}_\mathcal{A}} \mathbb{E}_H L(f(x), y)$$

  - Like OLS *empirical loss minimizers*

- So algorithms are equivalent to the function class they choose from
- For estimation what we typically do…
  - Show that empirical loss minimizers generate unbiasedness