MITACS / CORS 2010 Annual Conference

Data

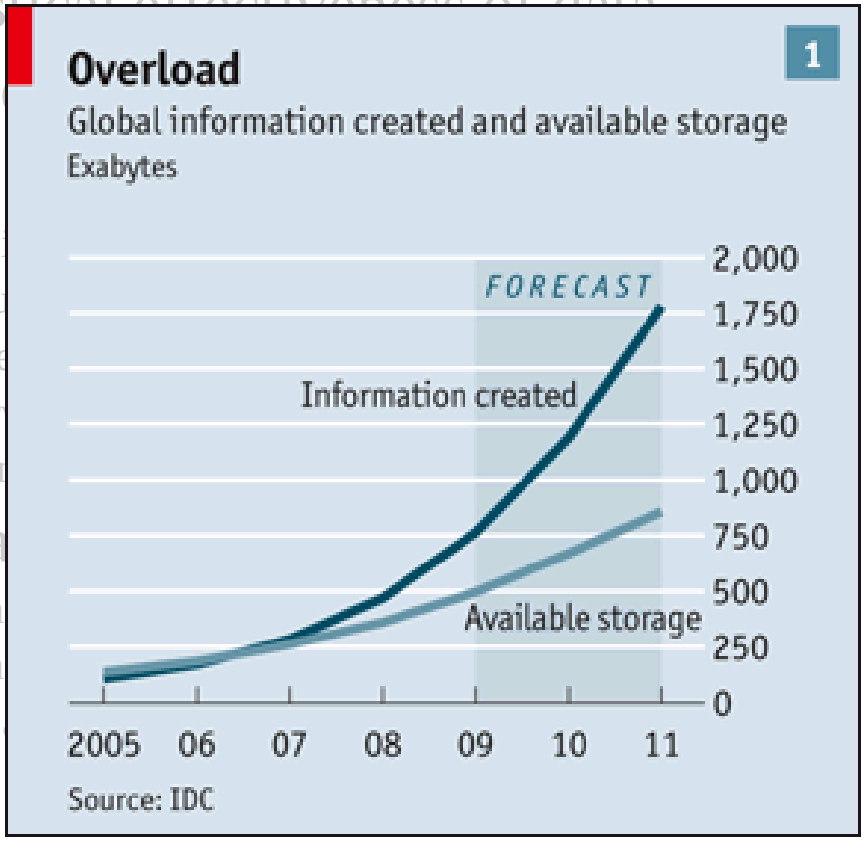
Nando de Freitas
University of British Columbia
May 2010

Outline

1. Big data
2. The opportunities
3. The statistical effectiveness of data
4. Toward semantic understanding
5. Essential tools for big data
   - Probability, statistics and optimization
   - Data structures and compression
   - Online learning
   - Unsupervised learning and feature induction
   - Attention
6. Other challenges
   - Storage and parallel data processing
   - Privacy and security
   - Training and supporting a new generation of data experts

Outline

1. Big data
2. The opportunities
3. The statis
4. Toward s
5. Essential
   - ❑ Probabil
   - ❑ Data stru
   - ❑ Online le
   - ❑ Unsuper
   - ❑ Attentio
6. Other cha
   - ❑ Storage a
   - ❑ Privacy a
   - ❑ Training



Data inflation

2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America's postal service this year will amount to around 5PB. Google processes around 1PB every hour
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of *The Economist*
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence this year is forecast to be around 1.2ZB
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine

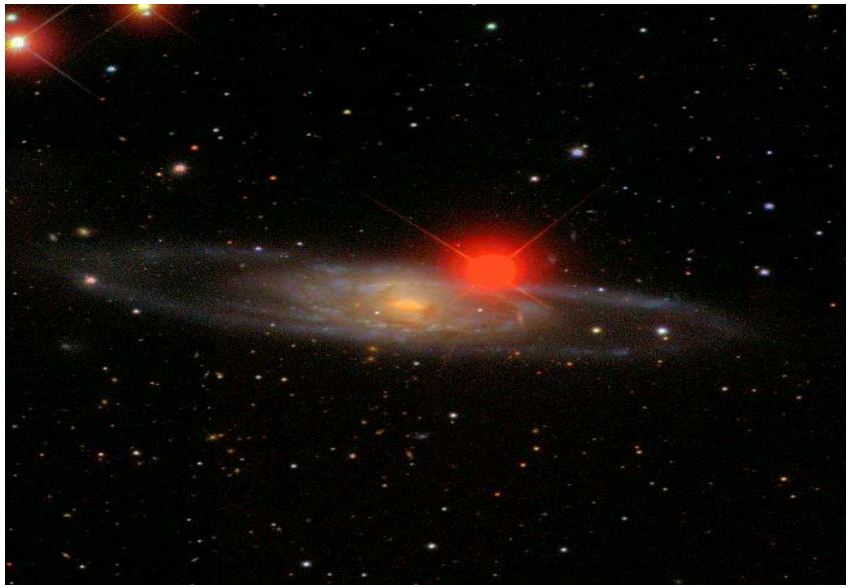
Source: *The Economist*

The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

Wikipedia Current revisions only uncompressed ~112 GB (**896,000,000,000** bits)

Human brain ~**100, 000,000,000** neurons and ~**60,000, 000,000,000** synapses

Big data: Surveying the universe



"When the **Sloan Digital Sky Survey** started work in 2000, its telescope in New Mexico collected more data in its first few weeks than had been amassed in the entire history of astronomy.

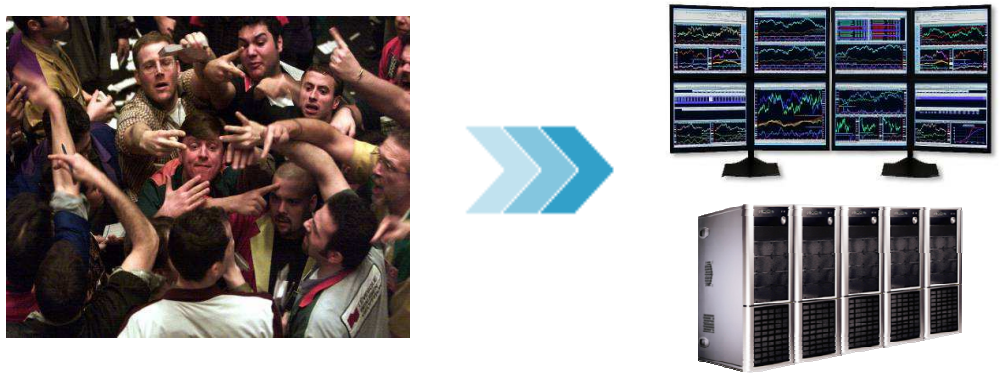
Now, a decade later, its archive contains a whopping **140 terabytes** of information.

A successor, the **Large Synoptic Survey Telescope**, due to come on stream in Chile in 2016, will acquire that quantity of data every five days."

[*The Economist*, February 2010]

Big data: Financial markets

Technology has transformed financial markets.



- Skyrocketing data volumes: 1.5 million messages/sec and growing
- Low latency data feeds and direct market access
- About 70% of volume in US equity markets submitted electronically

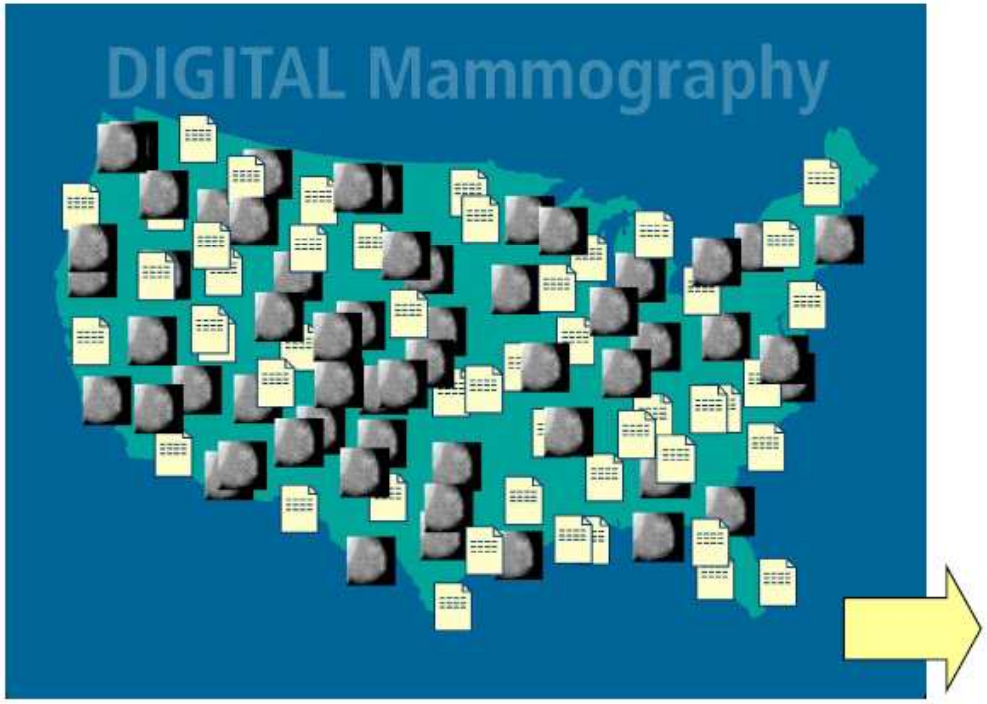
"A 1-millisecond advantage in trading applications can be worth $100 million a year to a major brokerage."

-- The TABB Group

Courtesy of *Alan Wagner, UBC*

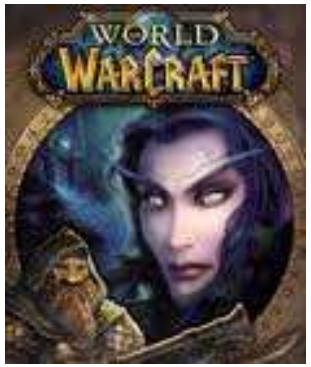
Big data: Medicine

National Digital Mammography Archive: a system designed to include a database growing by 28 PB per year according to IBM sources.



Highly Distributed and Massive Source

Use High Performance Networks, Hierarchical Storage and Indexing

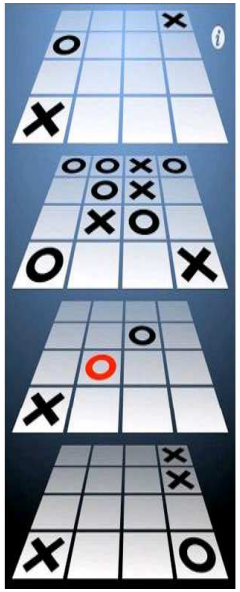


Google

YouTube

- **Library of Congress** text database of ~20 TB
- **AT&T** 323 TB, 1.9 trillion phone call records.
- **World of Warcraft** utilizes 1.3 PB of storage to maintain its game.
- **Avatar** movie reported to have taken over 1 PB of local storage at *Weta Digital* for the rendering of the 3D CGI effects.
- **Google** processes ~24 PB of data per day.
- **YouTube:** 24 hours of video uploaded every minute. More video is uploaded in 60 days than all 3 major US networks created in 60 years. According to *cisco*, internet video will generate over 18 EB of traffic per month in 2013.

Big data: publish, perish and polymath



On January 2009, Fields Medalist Tim Gowers, asked a provocative question: "Is something like massively collaborative mathematics possible?"

Density Hales-Jewett and Moser numbers, by D.H.J. Polymath. 49 pages. To appear, Szemeredi birthday conference proceedings.

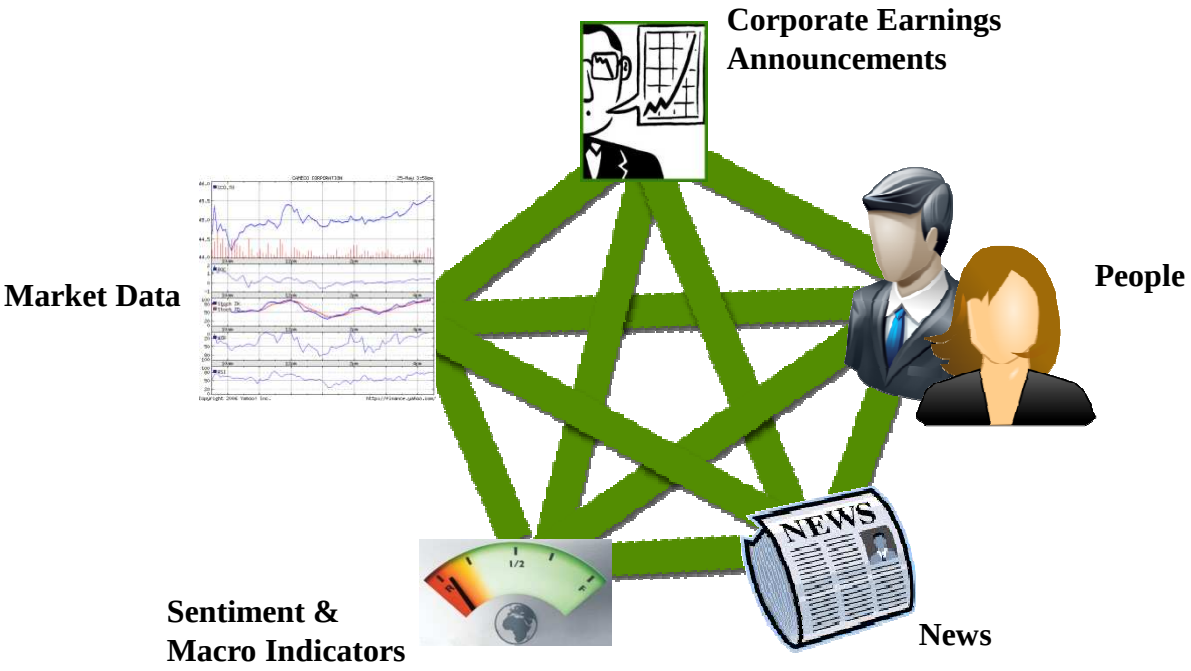
Outline

1. Big data
2. The opportunities
3. The statistical effectiveness of data
4. Toward semantic understanding
5. Essential tools for big data
   - Probability, statistics and optimization
   - Data structures and compression
   - Online learning
   - Unsupervised learning and feature induction
   - Attention
6. Other challenges
   - Storage and parallel data processing
   - Privacy and security
   - Training and supporting a new generation of data experts

Opportunities

Business

- Mining correlations, trends, spatio-temporal predictions.
- Efficient supply chain management.
- Opinion mining and sentiment analysis.
- Recommender systems.
- …



With *Alan Wagner, UBC*

Optemo

Home Products Technology Company Partners Contact Us

Optemo's solutions enhance the online shopping experience and increase your online retail business.

Find out what Optemo's Discovery Browser can do for your company by signing up for a demo.

SIGN UP NOW

Guaranteed to Improve Conversion Rates

Optemo has created a suite of novel and intelligent e-commerce solutions for retailers. Optemo's patent-pending technologies allow shoppers to browse thousands of items in a couple of minutes. Our solutions also organize and personalize product information in such a way that shoppers can efficiently make an informed purchase. This will revolutionize the online shopping experience!

Bestbuy.ca

Best Buy Canada had an 80% increase in sales with shoppers using the Discovery Browser. Download the data sheet to find out more about the Discovery Browser.

Technology used on Printers

LaserPrinterHub.com is showcasing Optemo's technology. It allows shoppers to compare the different laser printers available from different online retailers.

Recent News

The Vancouver Sun

Optemo has been awarded one of 14 prestigious Precarn T-GAP commercialization grants in Canada to improve online shopping with its novel technology.

straight

Georgia Straight, Vancouver's weekly, recently had an interview with Optemo's co-founder about the company and the Discovery Browser.

Opportunities

Science

- Astronomy
- Biology
- Medicine
- Ecology
- Brain Science
- …

AQUATIC INFORMATICS™

Safety

- Crime stats
- Emergency response
- …

Government and institutional accountability

Outline

1. Big data
2. The opportunities
3. The statistical effectiveness of data
4. Toward semantic understanding
5. Essential tools for big data
   - Probability, statistics and optimization
   - Data structures and compression
   - Online learning
   - Unsupervised learning and feature induction
   - Attention
6. Other challenges
   - Storage and parallel data processing
   - Privacy and security
   - Training and supporting a new generation of data experts

Big data: text

“Large” text dataset:

- **1,000,000 words in 1967**
- **1,000,000,000,000 words in 2006**

Success stories:

- **Speech recognition**
- **Machine translation**

What is the common thing that makes both of these work well?

- **Lots of labeled data**
- **Memorization is a good policy**

[Halevy, Norvig & Pereira, 2009]

Machine translation

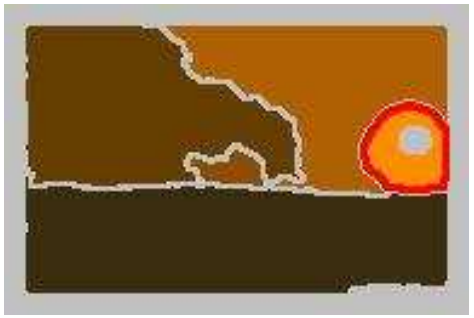
I love you **I love chocolate** **I am**

Yo te amo **Yo amo el chocolate** **Yo soy**

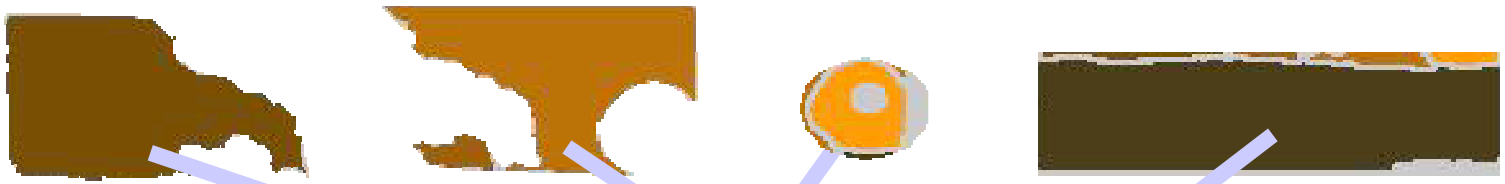
1. **Get many sentence pairs – easy.**
2. **Compute correspondences**
3. **Compute translation table: P(*Spanish*|*English*)**
4. **Repeat steps 2 and 3 till convergence**

Machine translation

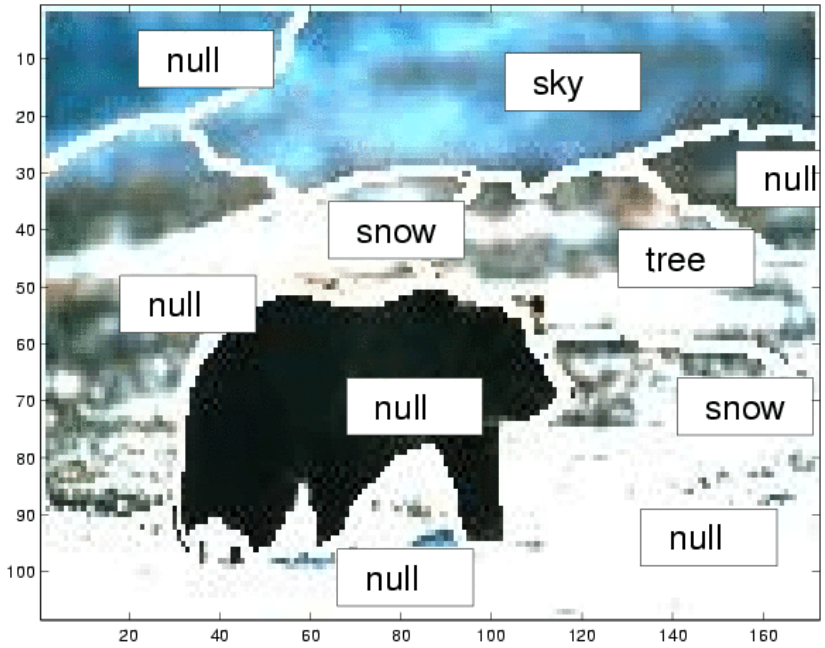
“Gorgeous red sea, sun and sky”

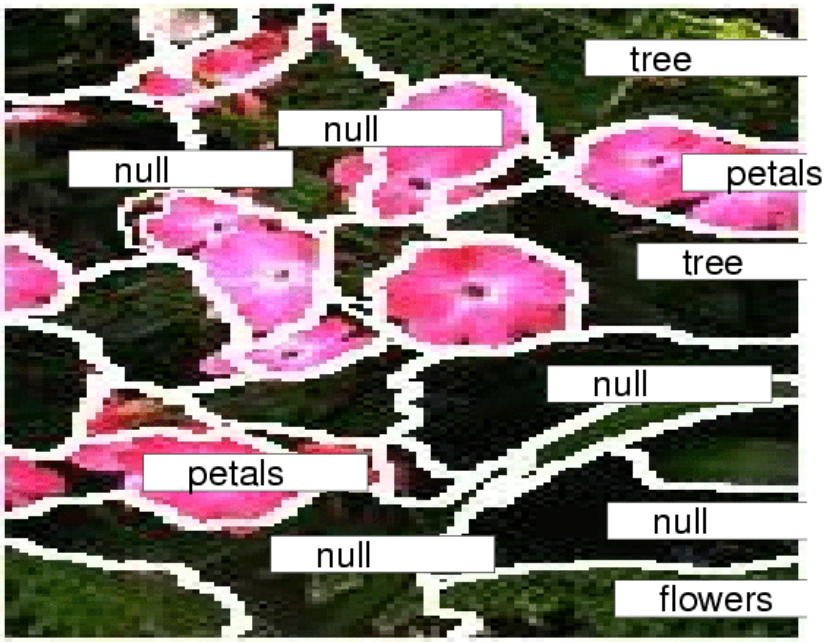


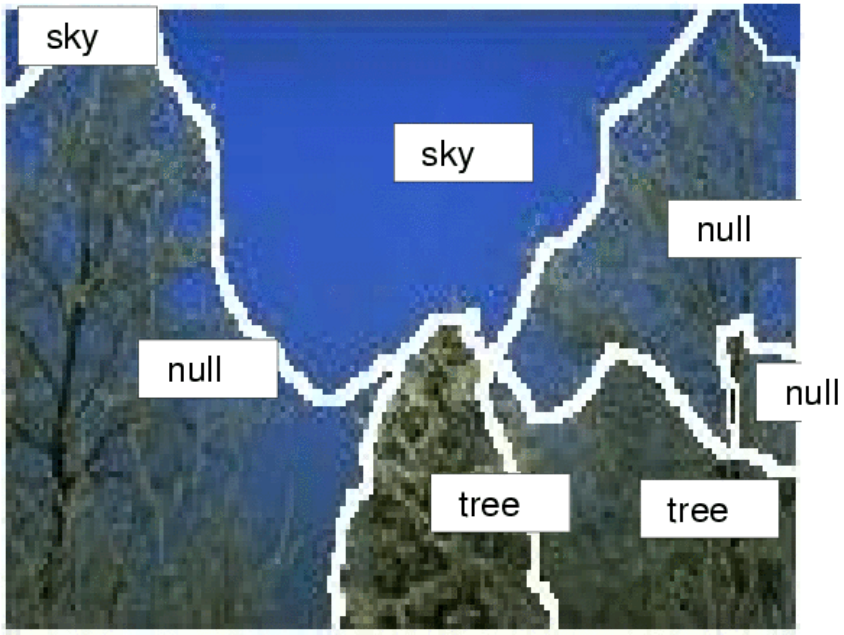
sun sea sky

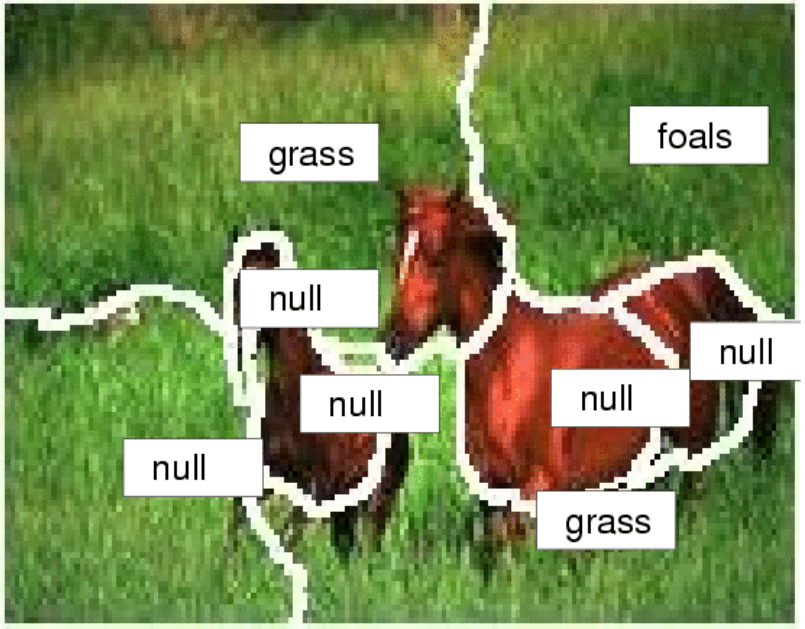


sun sea sky


null
sky
null
snow
tree
null
null
snow
null
null
10
20
30
40
50
60
70
80
90
100
20
40
60
80
100
120
140
160


tree
null
null
petals
tree
null
petals
null
null
flowers


sky
sky
null
null
null
tree
tree


grass
foals
null
null
null
null
null
grass

Text to images: auto-illustration

Text Passage (Moby Dick)

"The large importance attached to the harpooneer's vocation is evidenced by the fact, that originally in the old Dutch Fishery, two centuries and more ago, the command of a whale-ship …"

Retrieved Images

PRINT NAVAL BATTLE JAPANESE SHIP CHINESE BEING SHIP WATER

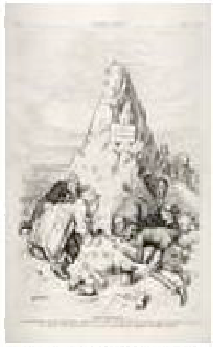
PRINT SHIP SURROUNDED ICE SEVERAL SHIP SEEN WHALE OTHER CURRIER

PRINT ATTACK WAGON ROAD FOREST CALLOT

PRINT WAR FRIGATE UNITED STATE ENGLISH SHIP AMERICAN SHIP CURRIER

PRINT SMALL BOAT APPROACHING BLOWING WHALE SHIP MOUNTAIN BACKGROUND CURRIER

PLAY BOAT PRINT KUNISADA



PRINT MEN SMALL MOUNTAIN HAS COME SEVERAL SMALL FOREGROUND POLITICAL

PRINT WHITE HOUSE GROUNDS BACKGROUND POLITICAL TYPE INDIAN ARMS TREE

Images to text: auto-annotation

Curator labels:

KUSATSU SERIES STATION TOKAIDO
GOJUSANTSUGI PRINT HIROSHIGE

Predicted labels:

tokaido print hiroshige object artifact series
ordering gojusantsugi station facility arrangement
minakuchi

Poems to songs

Input poem

The Waste Land

T S Eliot

For Ezra Pound,
***il miglior fabbro*.**

I. The Burial of the Dead

April is the cruelest month, breeding
Lilacs out of the dead land, mixing
Memory and desire, stirring
Dull roots with spring rain.
Winter kept us warm, covering
Earth in forgetful snow, feeding
A little life with dried tubers.
Summer surprised us, coming over the Starnbergersee
With a shower of rain; we stopped in the colonnade
And went on in sunlight, into the Hofgarten,
And drank coffee, and talked for an hour.
Bin gar keine Russin, stamm' aus Litauen, echt deutsch.
And when we were children, staying at the arch-duke's,
My cousin's, he took me out on a sled,
And I was frightened. He said, Marie,
Marie, hold on tight. And down we went.
In the mountains, there you feel free.
I read, much of the night, and go south in winter.
…

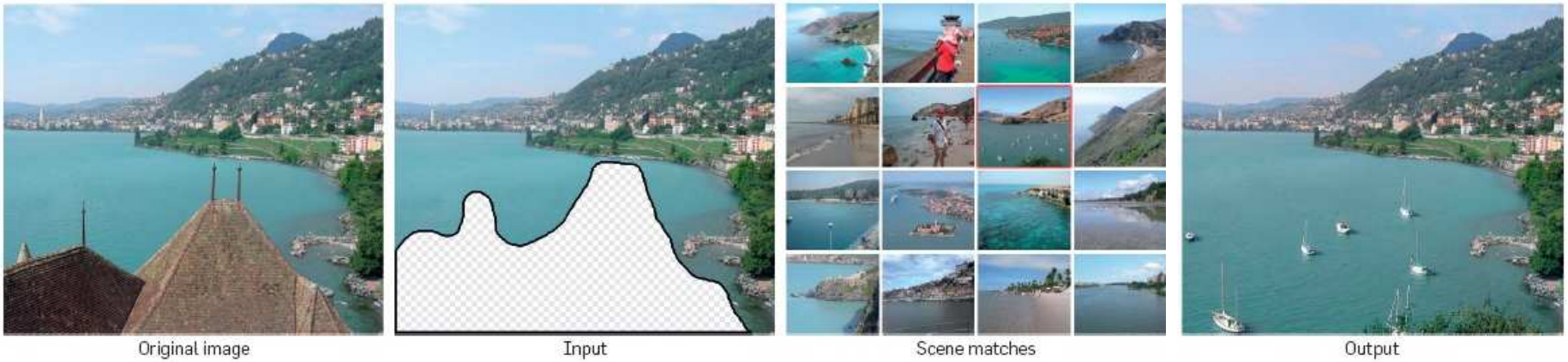
Closest song match

One Hundred Years

The Cure

It doesn't matter if we all die
Ambition in the back of a black car
In a high building there is so much to do
Going home time
A story on the radio
Something small falls out of your mouth
And we laugh
A prayer for something better
Please love me
Meet my mother
But the fear takes hold
Have we got everything?
She struggles to get away
The pain
And the creeping feeling
A little black haired girl
Waiting for Saturday
The death of her father pushing her
Pushing her white face into the mirror
Aching inside me
…

Scene completion: more data is better



Given an input image with a missing region,
Efros uses matching scenes from a large
collection of photographs to complete the image

[Efros, 2008]

Outline

1. Big data
2. The opportunities
3. The statistical effectiveness of data
4. Toward semantic understanding
5. Essential tools for big data
   - Probability, statistics and optimization
   - Data structures and compression
   - Online learning
   - Unsupervised learning and feature induction
   - Attention
6. Other challenges
   - Storage and parallel data processing
   - Privacy and security
   - Training and supporting a new generation of data experts

The semantic challenge

"We've already solved the sociological problem of building a network infrastructure that has encouraged hundreds of millions of authors to share a trillion pages of content.

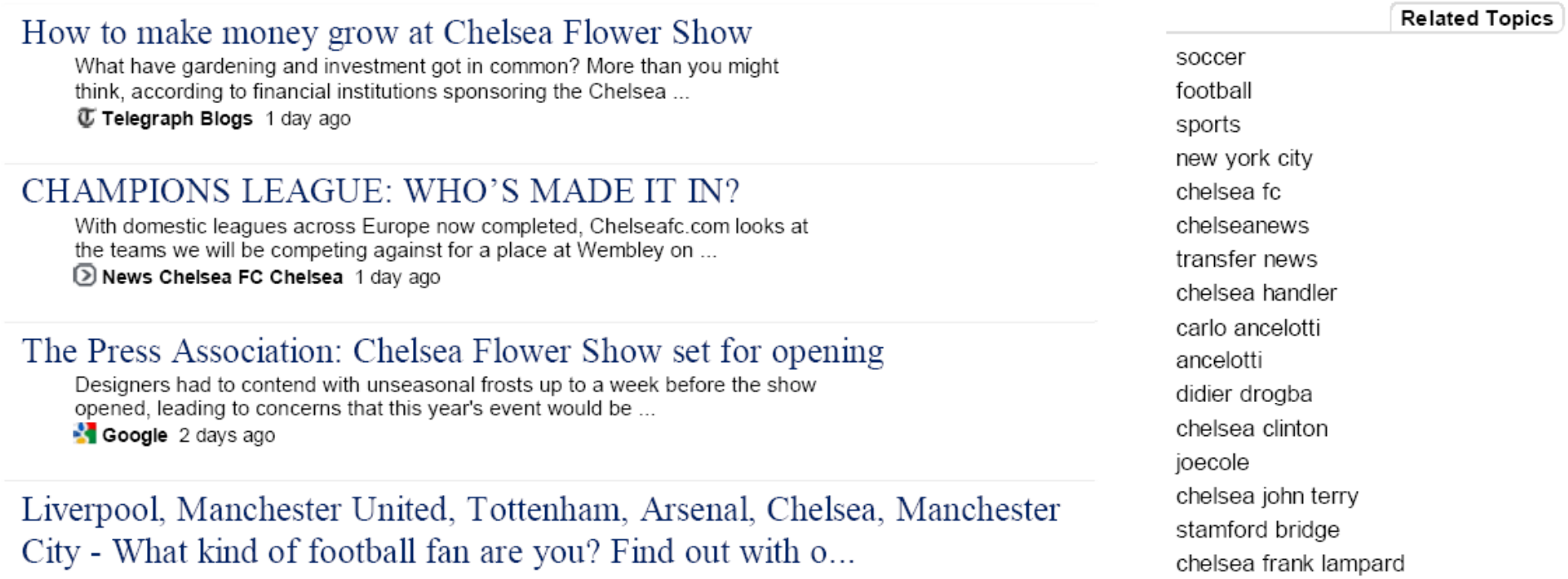
We've solved the technological problem of aggregating and indexing all this content.

But we're left with a scientific problem of interpreting the content"

Probability (*fact* given *evidence*) = ?

[Halevy, Norvig & Pereira, 2009]

The semantic challenge: **Zite**



To go beyond this, we need to improve our natural language processing techniques for semantic role labeling, parsing, analogy extraction and other structured inference tasks.

search topics or source

ICWSM 2010 Tutorial on Large-scale social media analysis with Hadoop

Over the last several years there has been a rapid increase in the number, variety, and size of readily available social media data.

jakehofman.com 2 days ago in machine learning

Google Pac-Man Might've Cost Us $120,483,800

Last week, the Google logo was turned into a game of Pac-Man and we all took breaks to play.

Gizmodo 13 hours ago from like-minded users

Ethnicity and Geography of Facebook Users - Data Mining: Text Mining, Visualization and Social Media

ePluribus: Ethnicity on Social Networks, by the Facebook data science team includes some interesting estimates of the geographic ...

Data Mining: Text Mining 1 day ago in text mining

FluidDB Aims To Become The Wikipedia Of Databases

A few years ago, Terry Jones sold his London apartment so that he could single-mindedly pursue a rather radical idea.

TechCrunch 20 hours ago from like-minded users

Huge Gap Remains Between Mainstream Media and the Social Web [REPORT]

A study underlines the large disconnect between what the mainstream media and what social media users consider newsworthy.

mashable.com 15 hours ago from like-minded users

3M Technology Optimizes Your Ads For Maximum Effectiveness

Penn Olson - Your Social, Media, Brand Catalog.

Penn Olson 2 days ago in visual attention

The Tragic Cost of Google Pac-Man – 4.82 million hours

When Google launched it's Pac-Man logo on Friday, we immediately heard amused groans in our tweet-streams.

blog.rescuetime.com 21 hours ago from like-minded users

The Secret Lives of Professors

I came to Harvard 7 years ago with a fairly romantic notion of what it meant to be a professor -- I imagined unstructured days spent ...

matt-welsh.blogspot.com 1 day ago from like-minded users

The powerful vagueness of sustainability « UBC Vancouver Sustainability Initiative

What does sustainability mean? Is it so vague a term that it has no real content, and can be high-jacked by anyone in the service of ...

blogs.ubc.ca 19 hours ago in ubc

Follow List

Edit

apriori algorithm
bayesian
boltzmann machines, deep learning
booyah, mytown
cognitive science
compressed sensing
computational neuroscience, attention brain (2 more)
computer vision
grid cells, grid neurons (2 more)
information retrieval
jeff hawkins, hierarchical temporal memory
john hopfield
kitsilano
machine learning
markov decision processes
monte carlo
mpi parallel computing
natural image statistics
policy gradient
pomdp, pomdps
random projections
semantic role labeling
sequential monte carlo, markov chain monte carlo (2 more)
sparse coding
statistical natural language processing, computational linguistics (1 more)
stochastic approximation, stochastic optimization
submodular functions, submodularity
ubc
vancouver restaurants
visual attention
worio
zite

add topics or source...

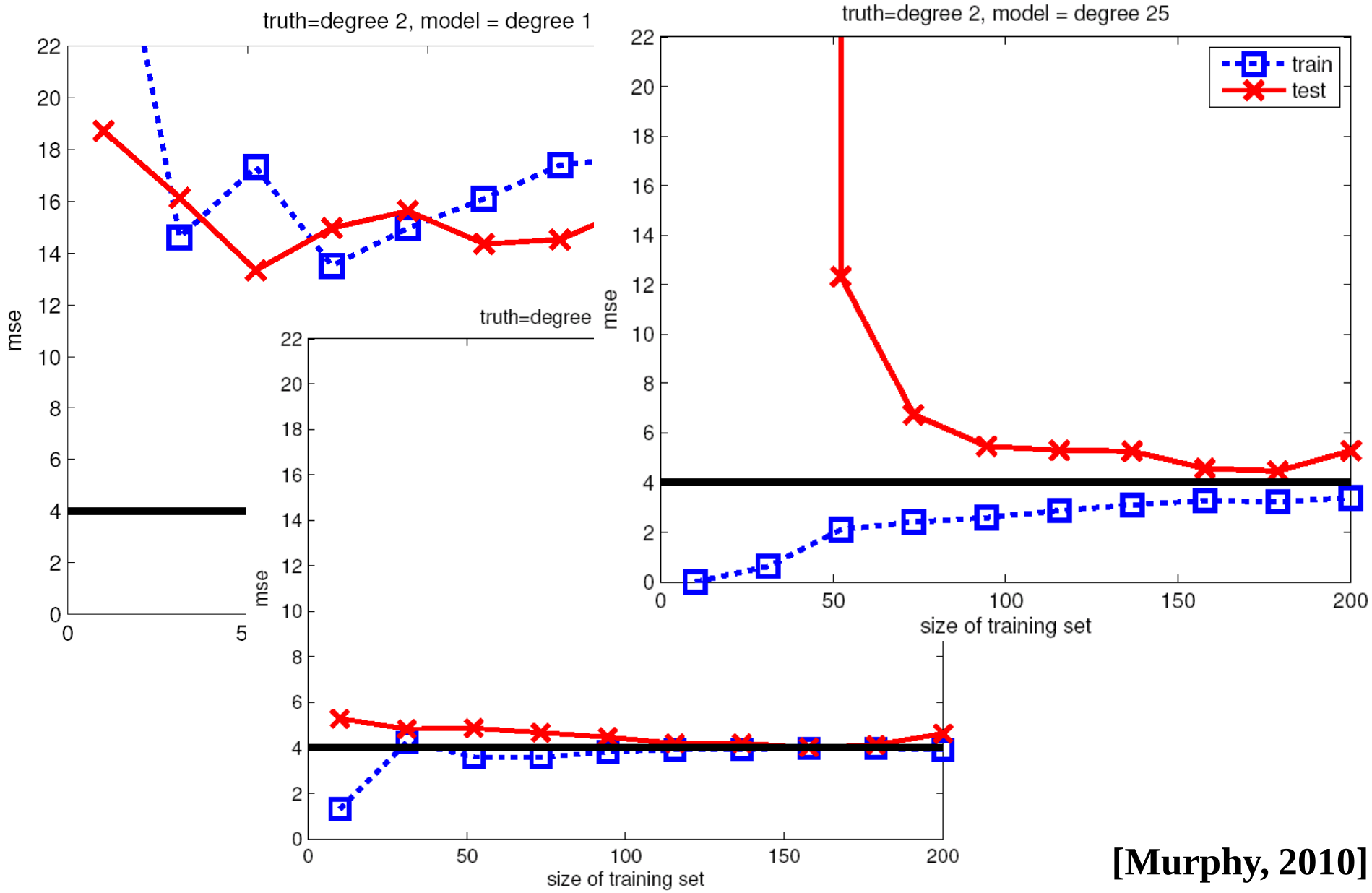
Outline

1. Big data
2. The opportunities
3. The statistical effectiveness of data
4. Toward semantic understanding
5. Essential tools for big data
   - ❑ Probability, statistics and optimization
   - ❑ Data structures and compression
   - ❑ Online learning
   - ❑ Unsupervised learning and feature induction
   - ❑ Attention
6. Other challenges
   - ❑ Storage and parallel data processing
   - ❑ Privacy and security
   - ❑ Training and supporting a new generation of data experts

Outline

1. Big data
2. The opportunities
3. The statistical effectiveness of data
4. Toward semantic understanding
5. Essential tools for big data
   - ❑ Probability, statistics and optimization
   - ❑ Data structures and compression
   - ❑ Online learning
   - ❑ Unsupervised learning and feature induction
   - ❑ Attention
6. Other challenges
   - ❑ Storage and parallel data processing
   - ❑ Privacy and security
   - ❑ Training and supporting a new generation of data experts

Approximation, stats and optimization



[Murphy, 2010]

Approximation, stats and optimization

$$
\begin{aligned}
E(\tilde{f}_n) - E(f^*) &= E(f^*_{\mathcal{F}}) - E(f^*) && \text{Approximation error} \\
&+ E(f_n) - E(f^*_{\mathcal{F}}) && \text{Estimation error} \\
&+ E(\tilde{f}_n) - E(f_n) && \text{Optimization error}
\end{aligned}
$$

Problem:

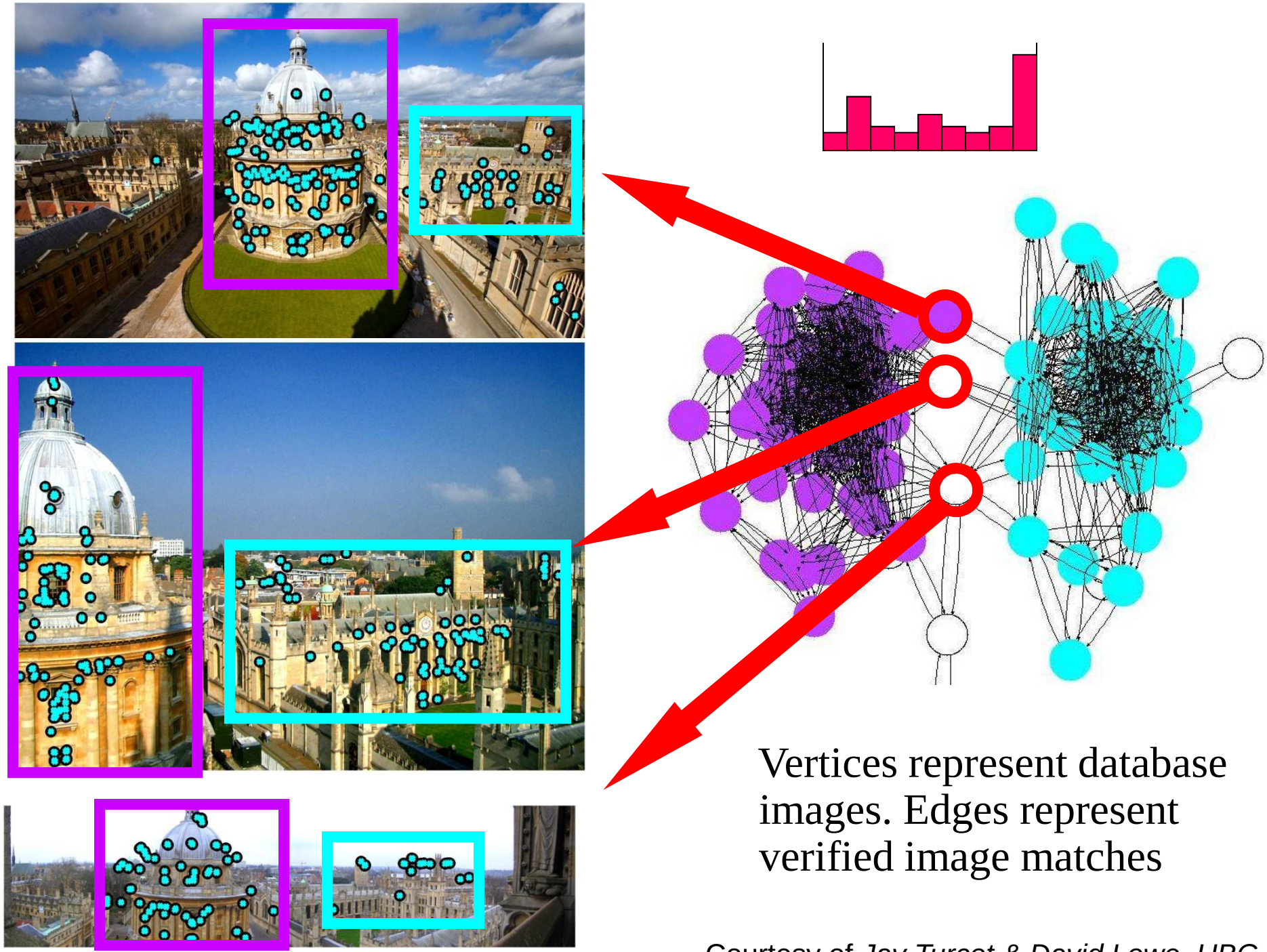
Choose $\mathcal{F}$, n, and ρ to make this as small as possible,

subject to budget constraints $\begin{cases} \text{maximal number of examples } n \\ \text{maximal computing time } T \end{cases}$

[Bottou, 2008]

Outline

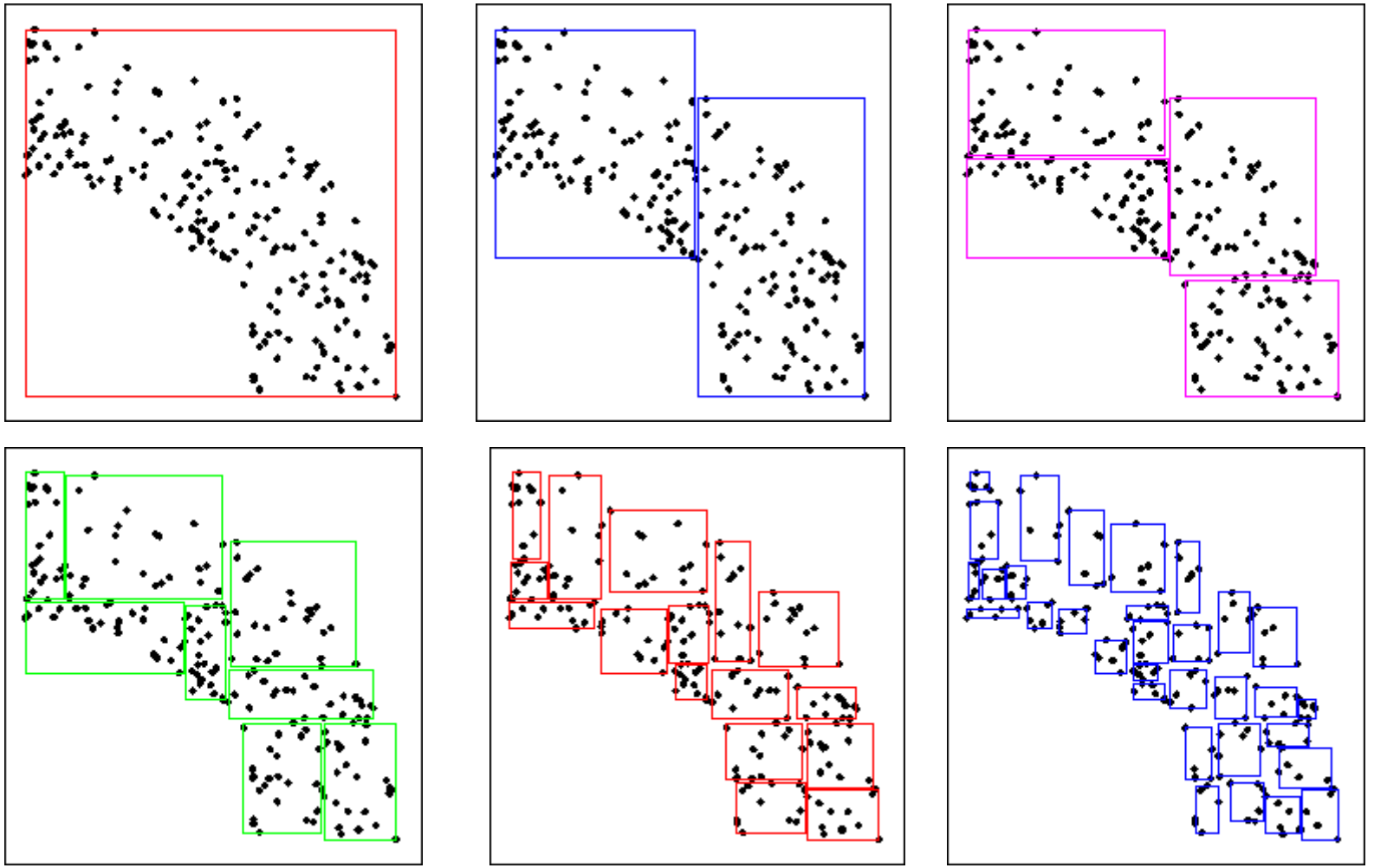
1. Big data
2. The opportunities
3. The statistical effectiveness of data
4. Toward semantic understanding
5. Essential tools for big data
   - ❑ Probability, statistics and optimization
   - ❑ Data structures and compression
   - ❑ Online learning
   - ❑ Unsupervised learning and feature induction
   - ❑ Attention
6. Other challenges
   - ❑ Storage and parallel data processing
   - ❑ Privacy and security
   - ❑ Training and supporting a new generation of data experts



Vertices represent database images. Edges represent verified image matches

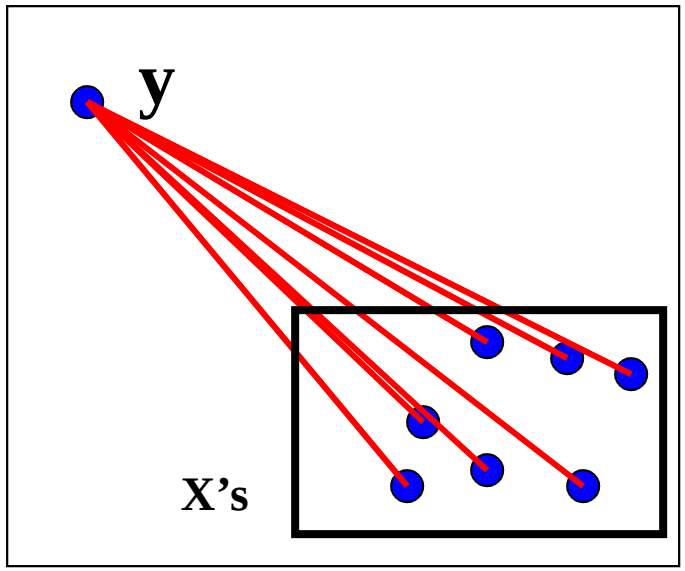
Courtesy of *Jay Turcot & David Lowe, UBC*

Tree recursions: We start by partitioning points using kd-trees or any metric trees



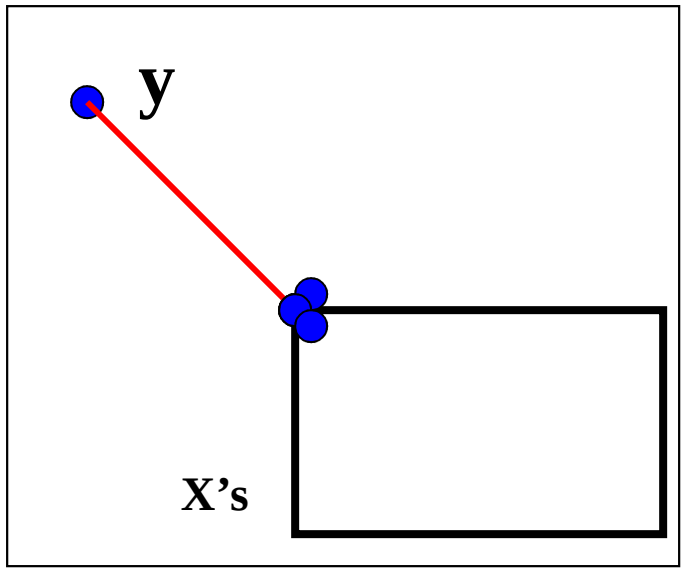
(Gray and Moore, 2000)

Far away groups of points are replaced by two single points (upper and lower bound)



$$f_j = \sum_{i=1}^{N} w_i \; K_i \left(\| x_i - y_j \| \right)$$

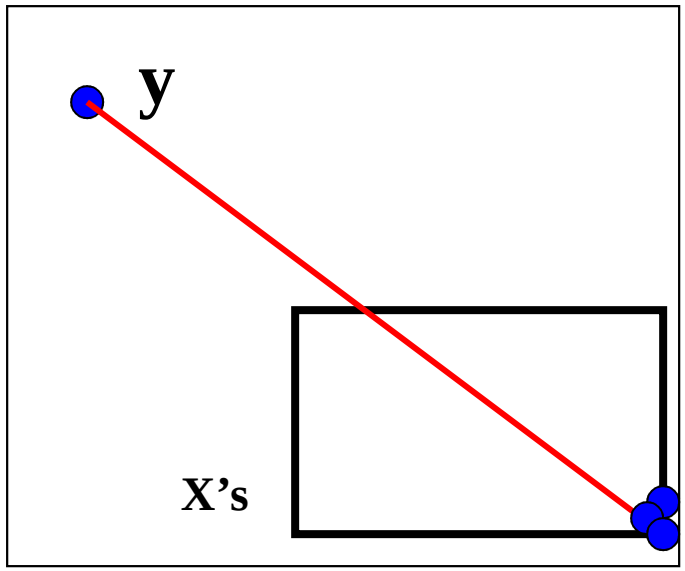
Far away groups of points are replaced by two single points (upper and lower bound)



$$f_j = \sum_{i=1}^{N} w_i \; K_i \left(\| x_i - y_j \| \right)$$

$$f^{(upper)}(X,Y) \quad = \quad K\left(d^{(lower)}(X,Y) \right) \sum_{i \in X} w_i$$

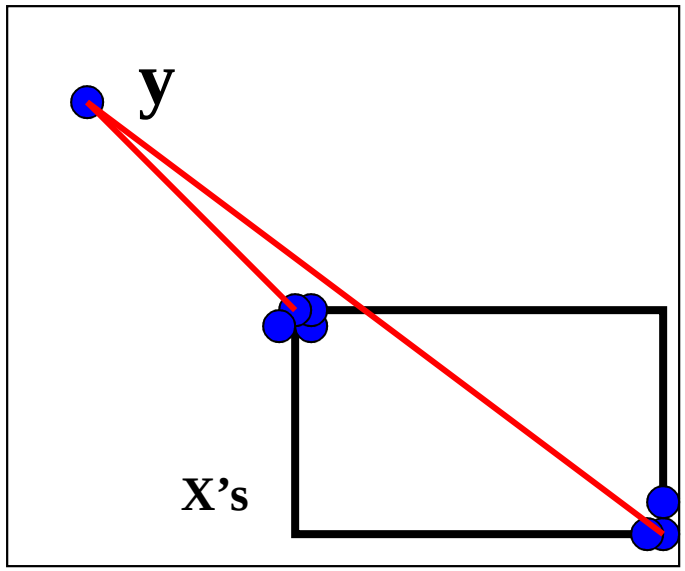
Far away groups of points are replaced by two single points (upper and lower bound)



$$f_j = \sum_{i=1}^{N} w_i \; K_i \left(\| x_i - y_j \| \right)$$

$$f^{(lower)}(X,Y) \quad = \quad K\left(d^{(upper)}(X,Y) \right) \; \sum_{i \in X} w_i$$

Far away groups of points are replaced by two single points (upper and lower bound)



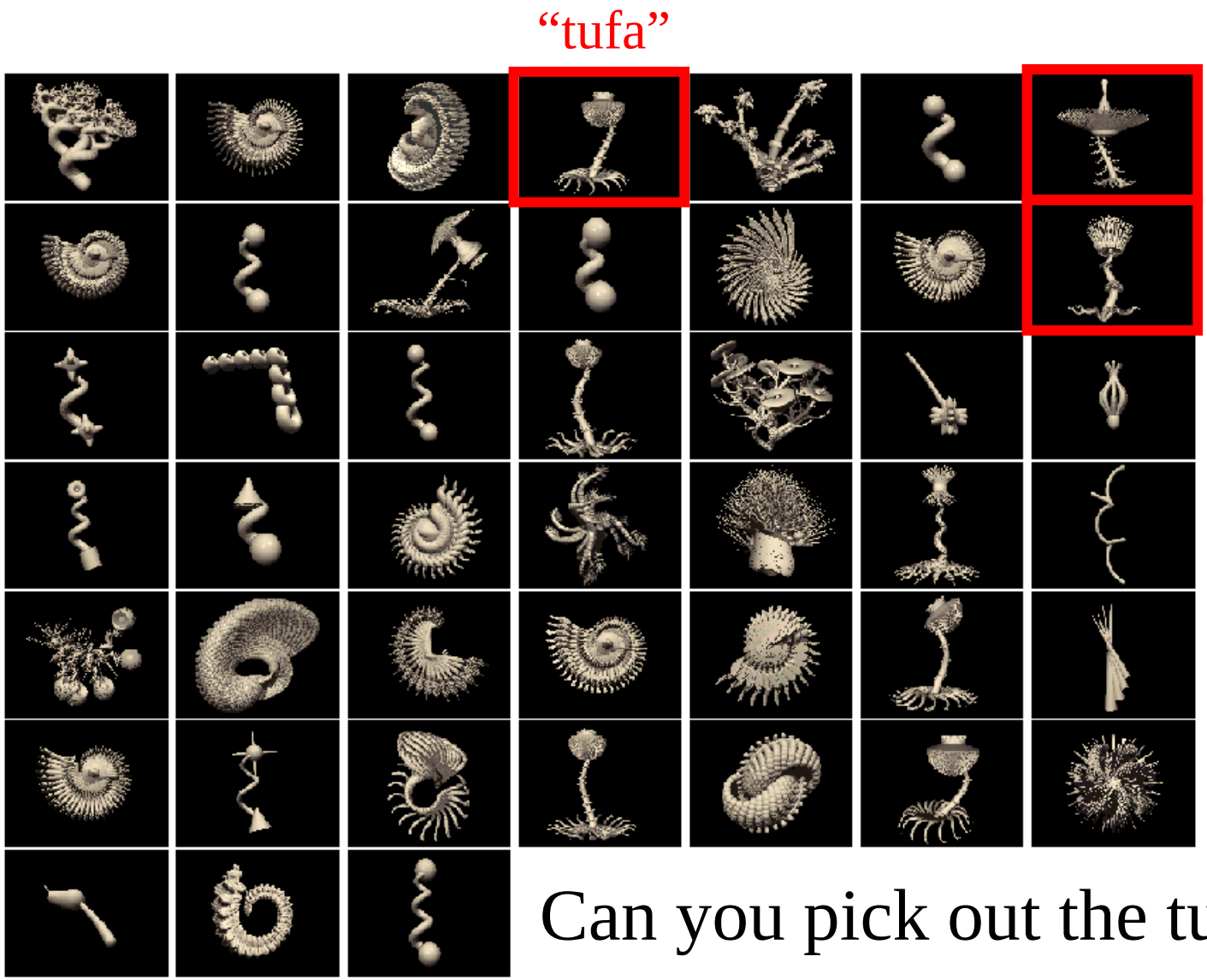
$$f_j = \sum_{i=1}^{N} w_i \; K_i \left(\| x_i - y_j \| \right)$$

$$\tilde{f}(X,Y) \; = \; \frac{1}{2} \; \left(f^{(upper)}(X,Y) + f^{(lower)}(X,Y) \right)$$

$$e(X,Y) \; = \; \frac{1}{2} \; \left(f^{(upper)}(X,Y) - f^{(lower)}(X,Y) \right)$$

Outline

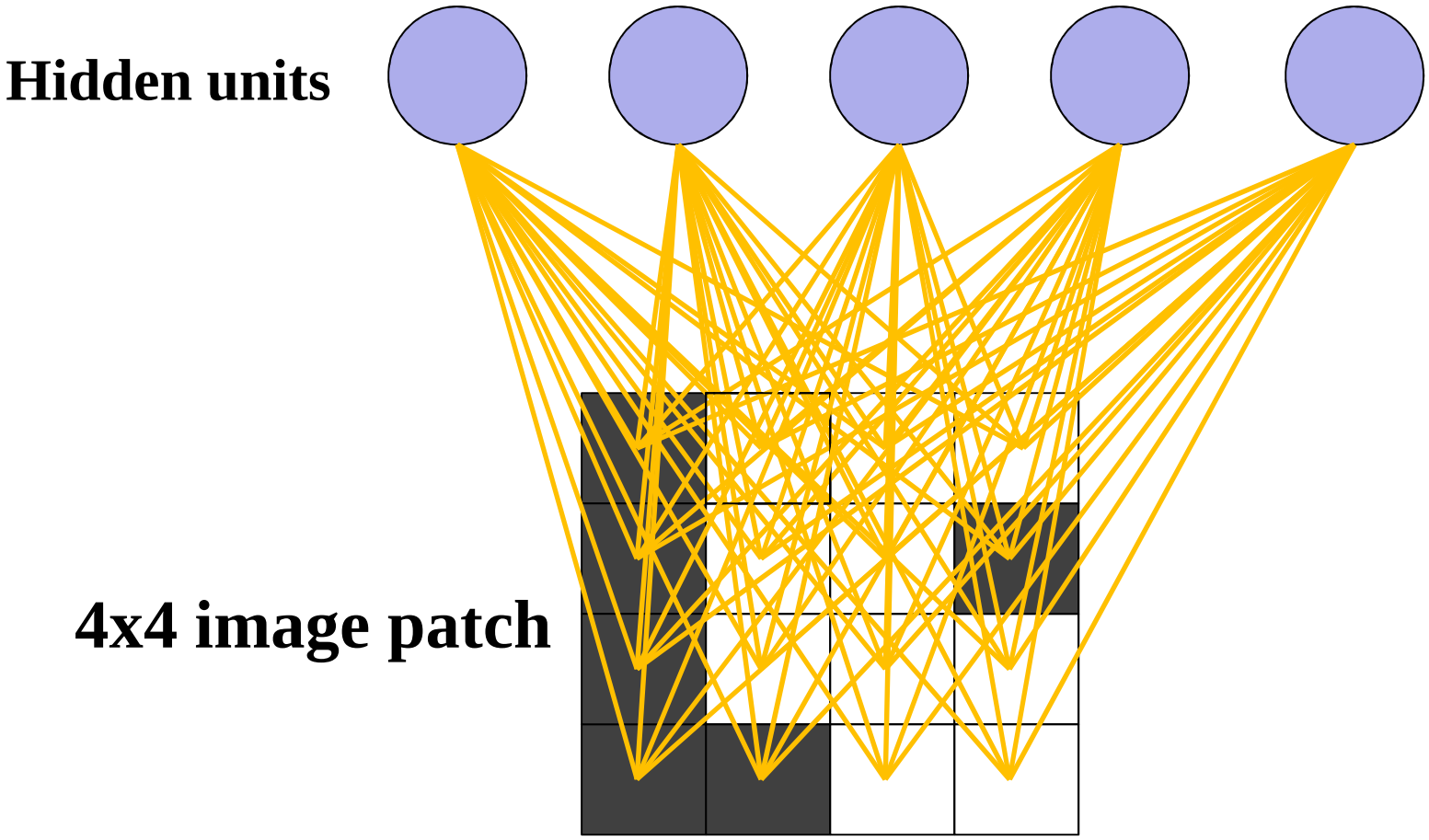
1. Big data
2. The opportunities
3. The statistical effectiveness of data
4. Toward semantic understanding
5. Essential tools for big data
   - ❑ Probability, statistics and optimization
   - ❑ Data structures and compression
   - ❑ Online learning
   - ❑ Unsupervised learning and feature induction
   - ❑ Attention
6. Other challenges
   - ❑ Storage and parallel data processing
   - ❑ Privacy and security
   - ❑ Training and supporting a new generation of data experts



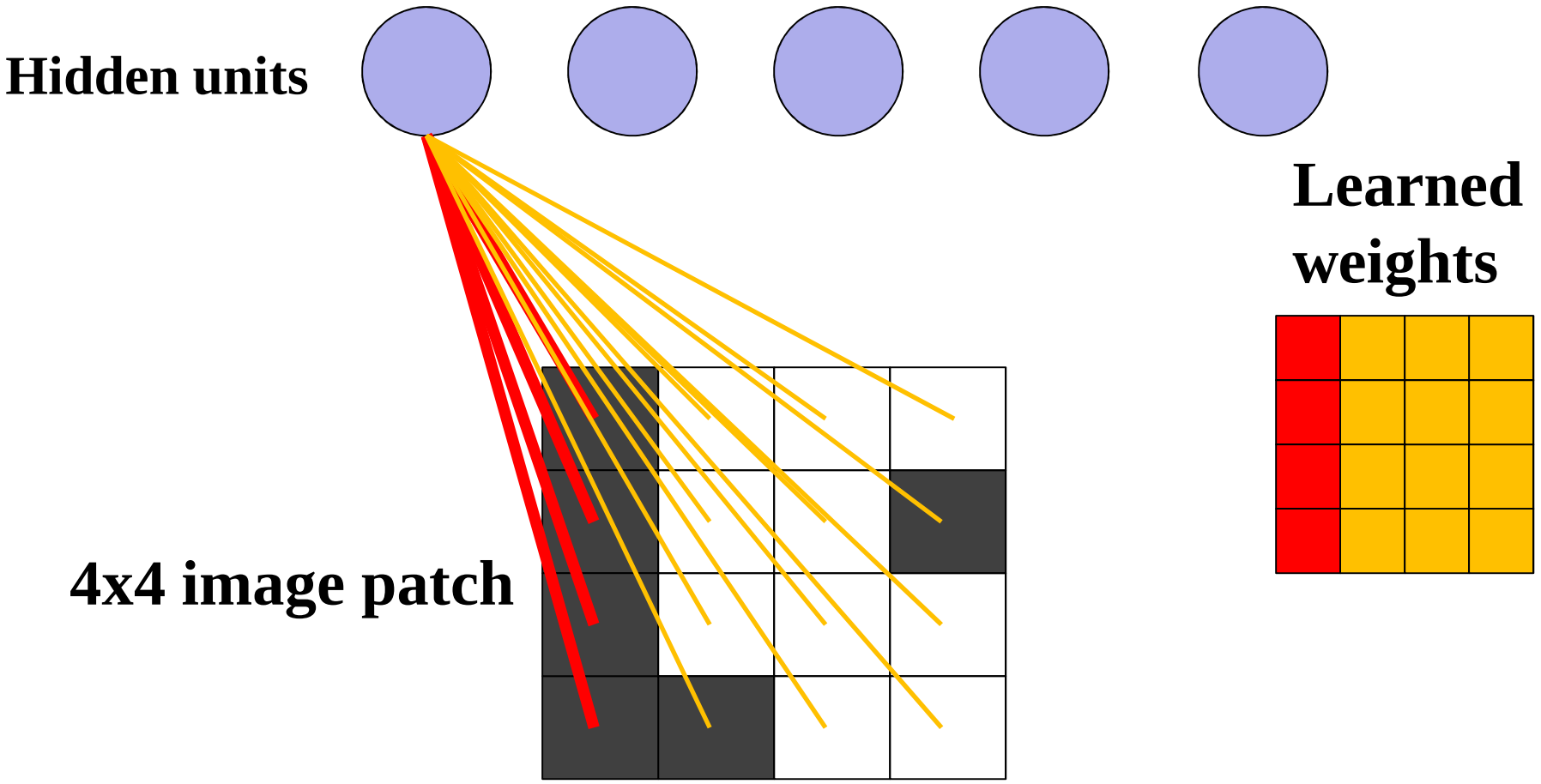
Can you pick out the tufas?

Source: *Josh Tenenbaum*

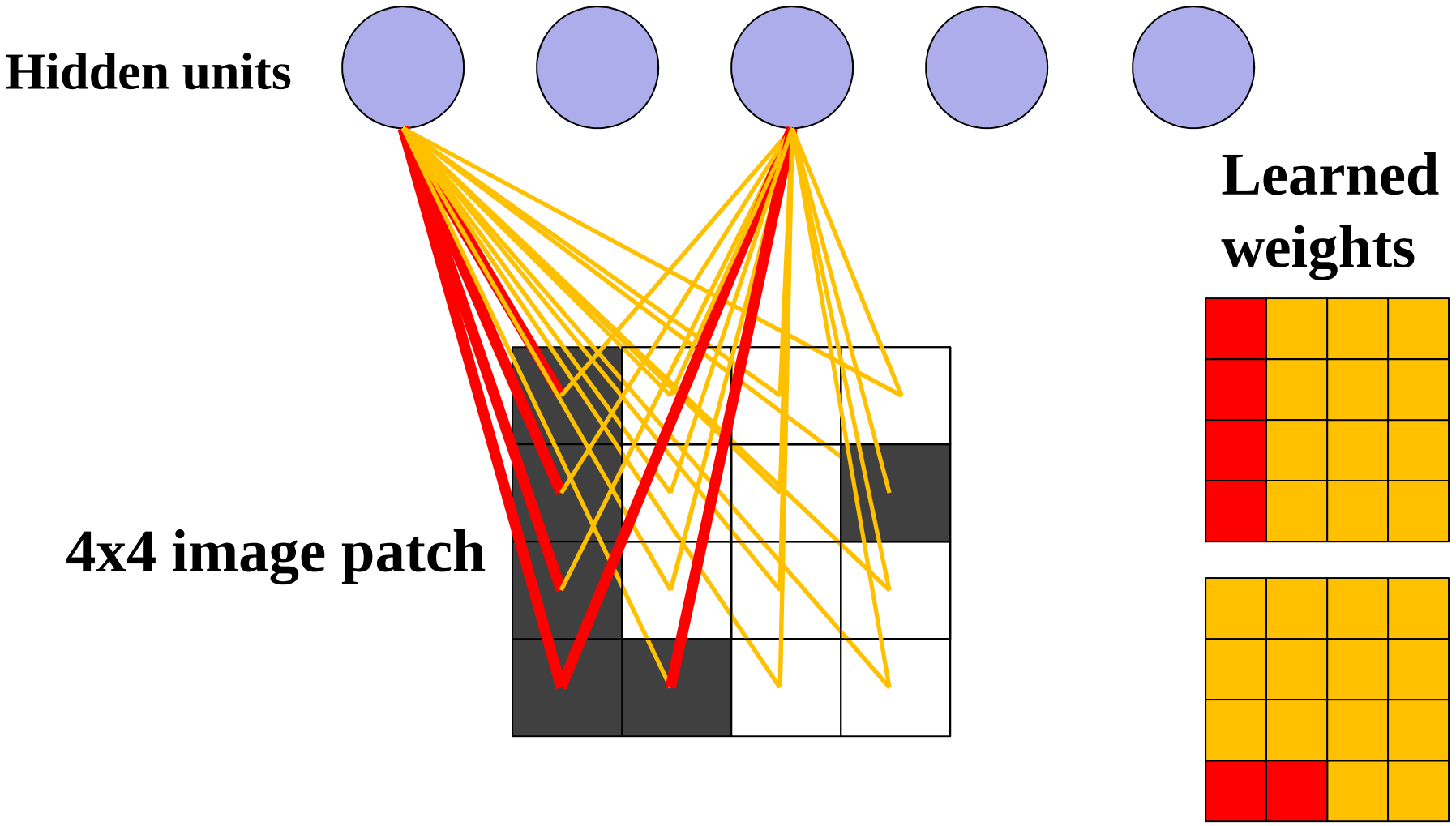
Distributed representation



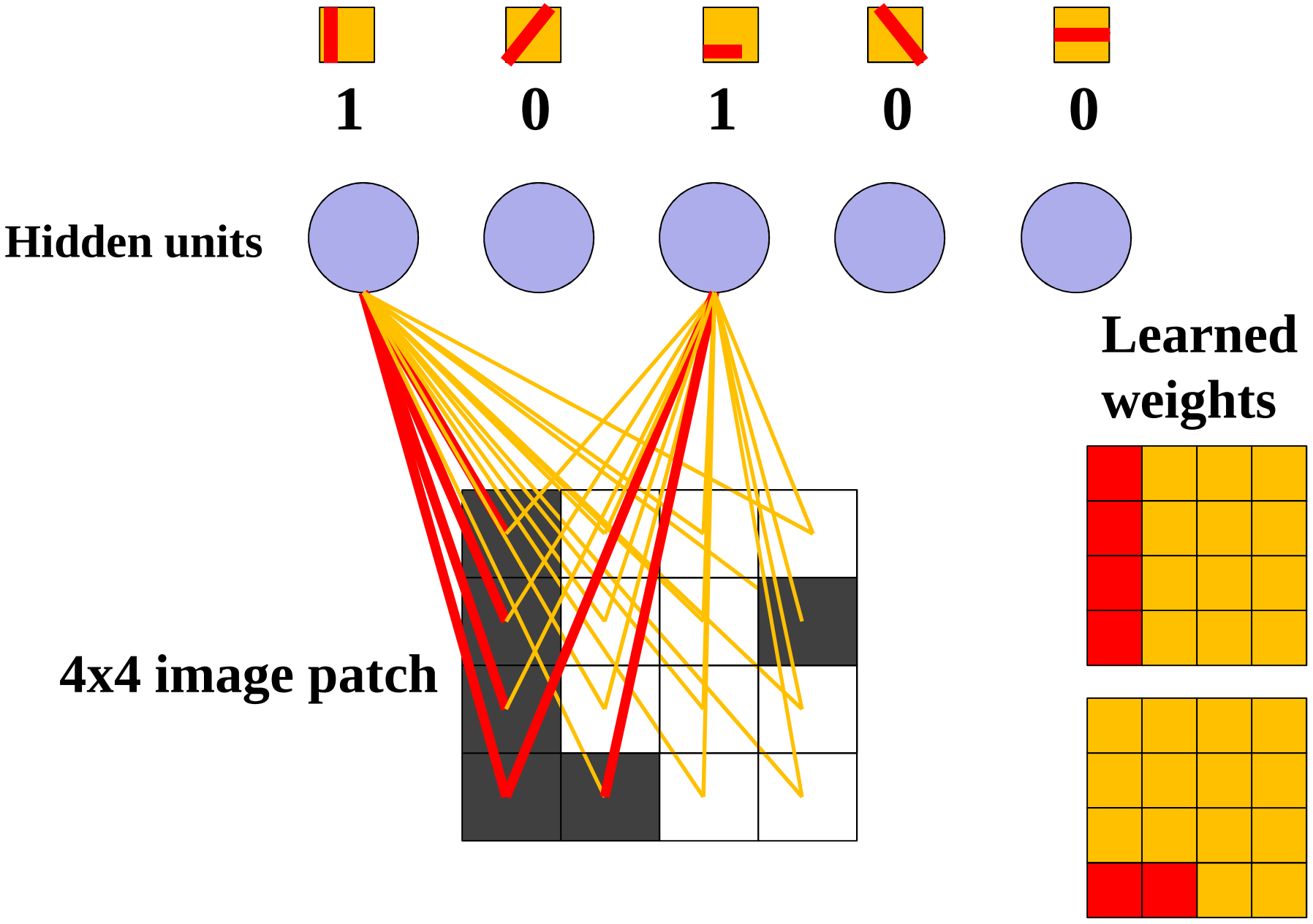
Distributed representation



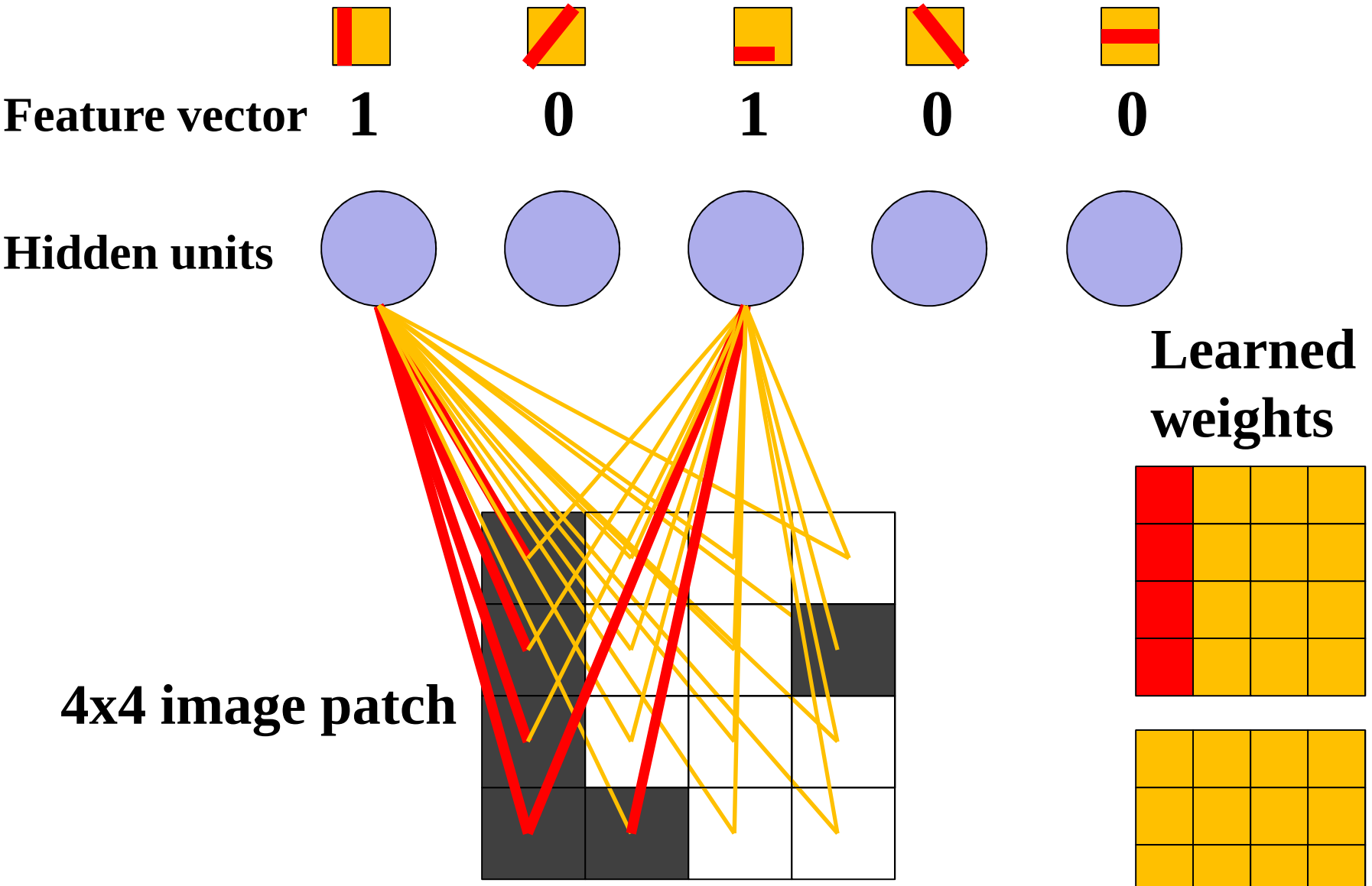
Distributed representation



Distributed representation



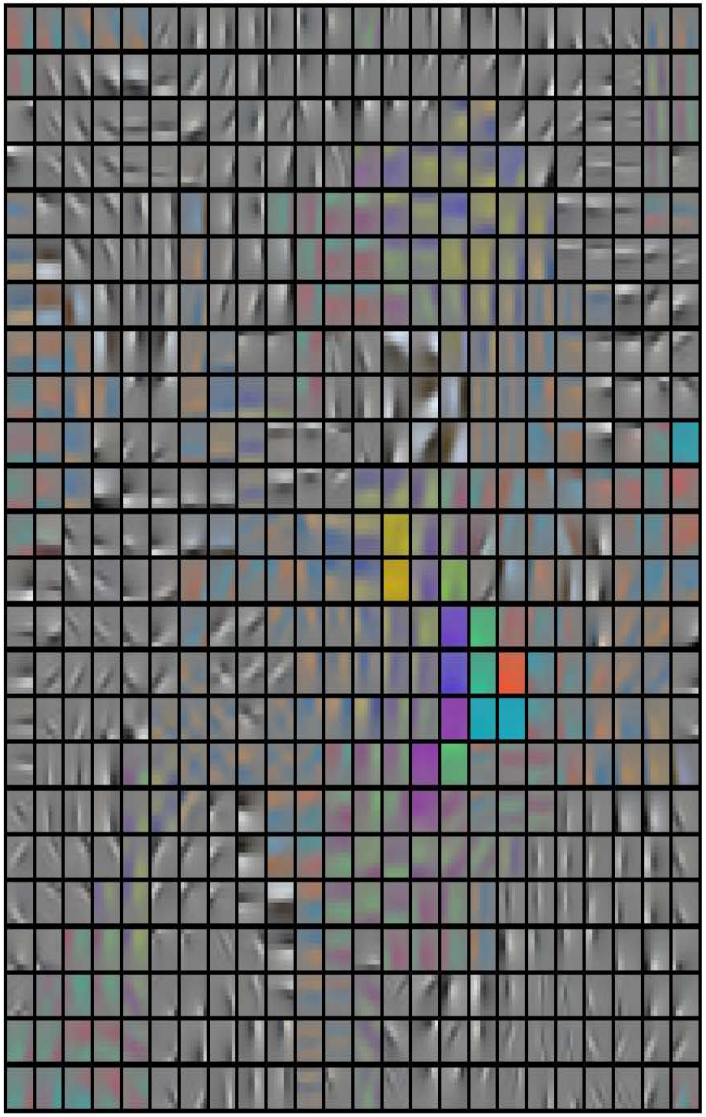
Distributed representation

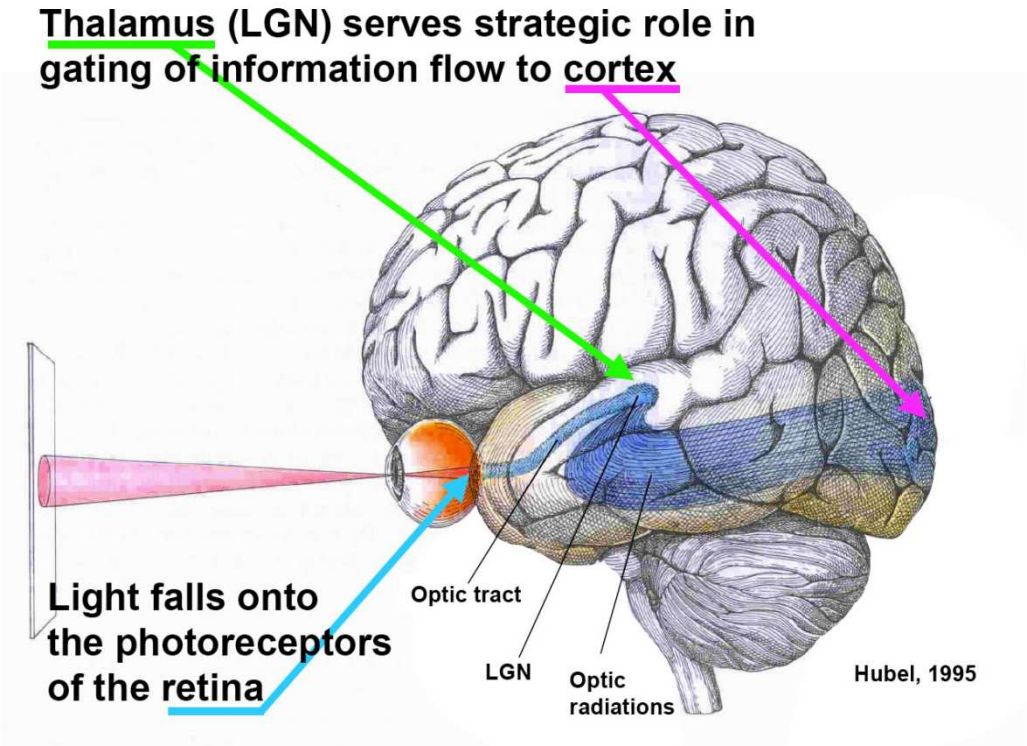


Insight: We're assuming edges occur often in nature, but dots don't

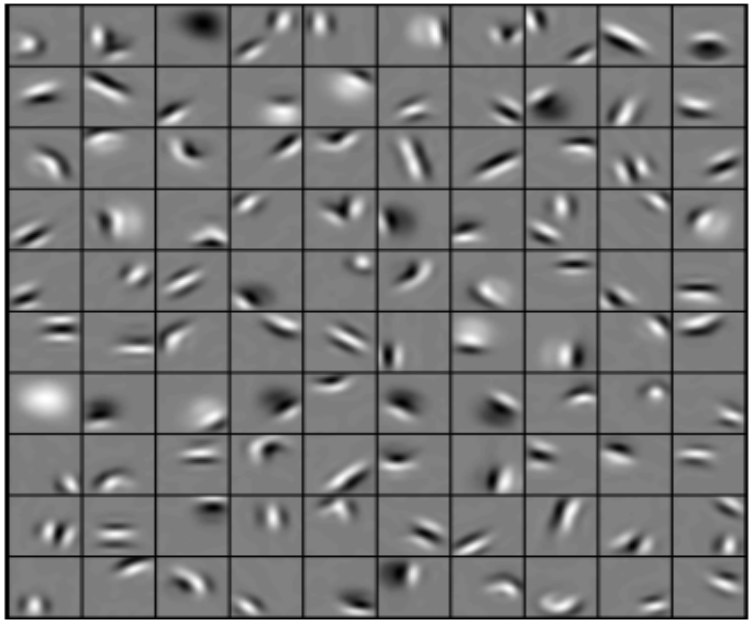
We learn the regular structures in the world

Automatically learned features to describe images match features measured in V1 area of brain



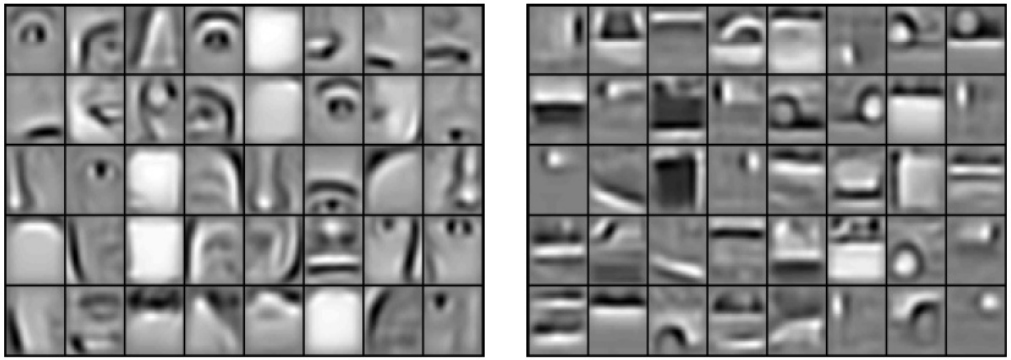


Layer 1

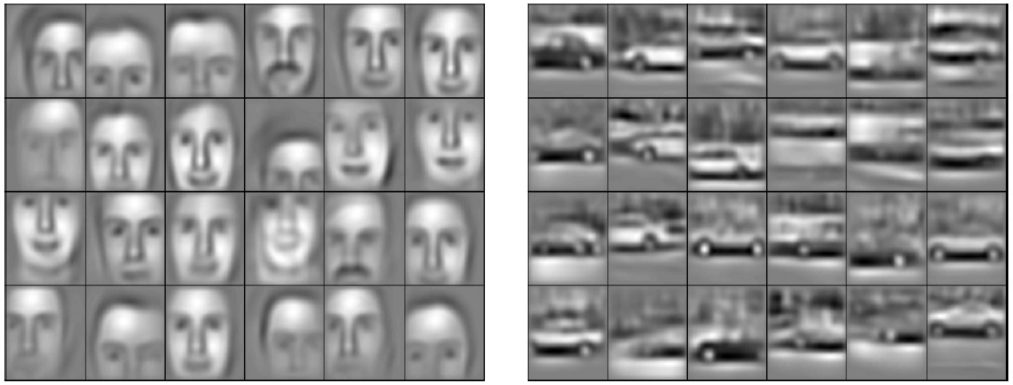


faces cars

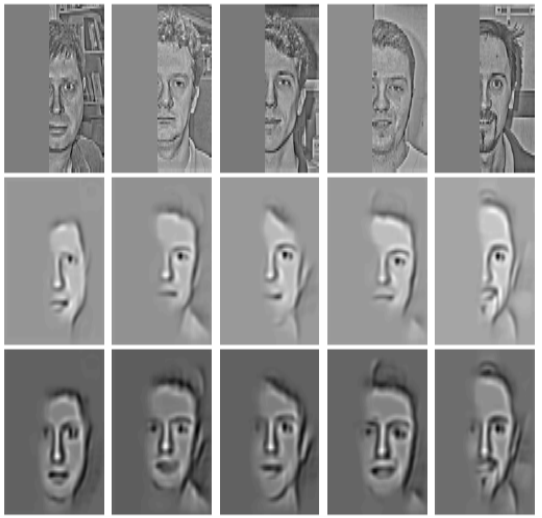
Layer 2



Layer 3



Completing scenes



[Honglak Lee et al 2009]

Geoff Hinton, Yoshua Bengio and Yann LeCun have lead the way in this field

Inference

(i) Given a training image, the binary state h_j of each feature detector j is set to 1 with probability

$$\frac{1}{1+\exp(-b_j-\sum_i v_i w_{ij})}$$

(ii) Given a hidden configuration, imagine visible unit v_i by setting it to 1 with probability

$$\frac{1}{1+\exp(-b_i-\sum_j w_{ij} h_j)}$$

Learning

data

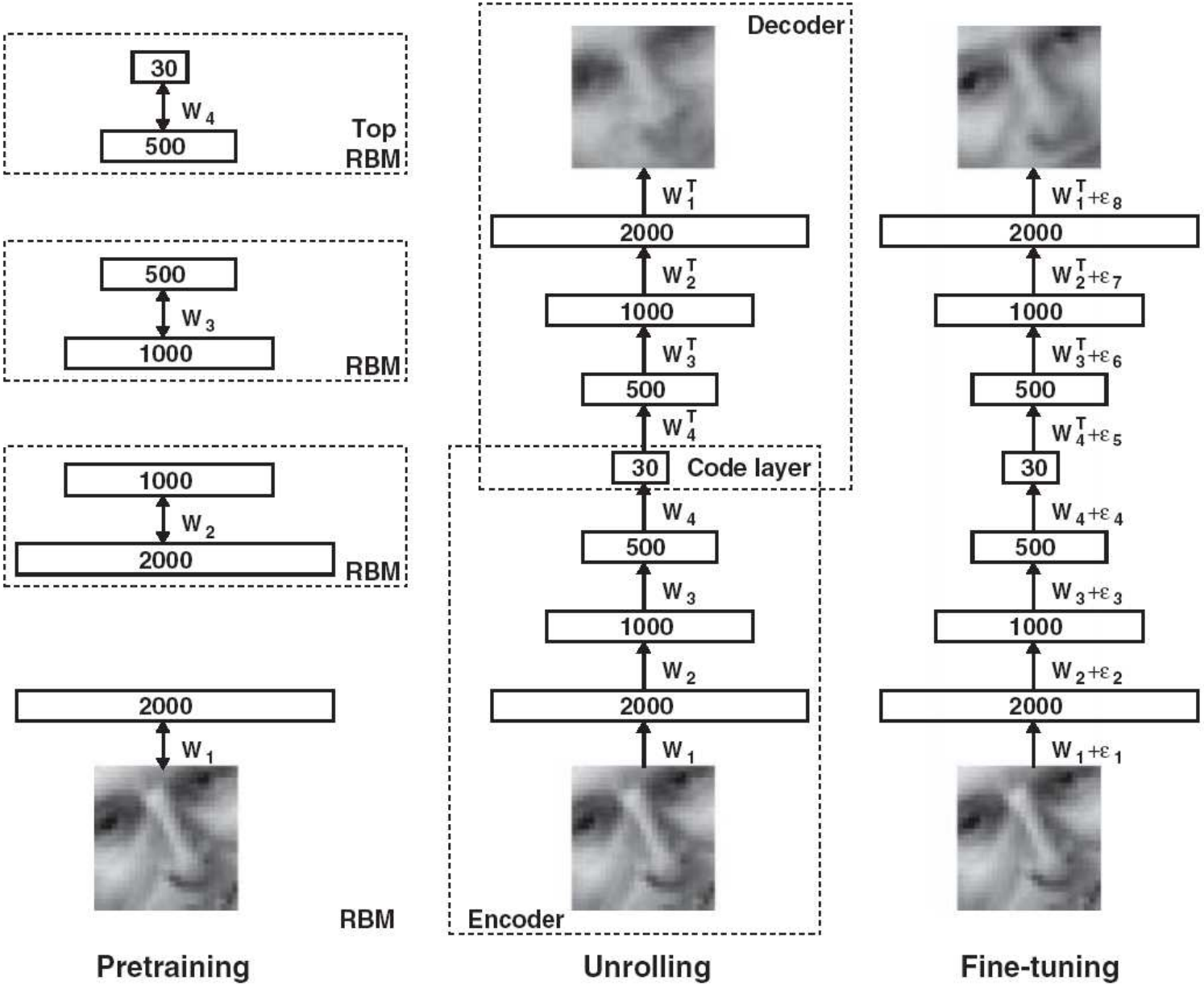
confabulation

$$\Delta w_{ij} = \varepsilon\left(\langle v_i h_j\rangle_{\text{data}} - \langle v_i h_j\rangle_{\text{recon}}\right)$$

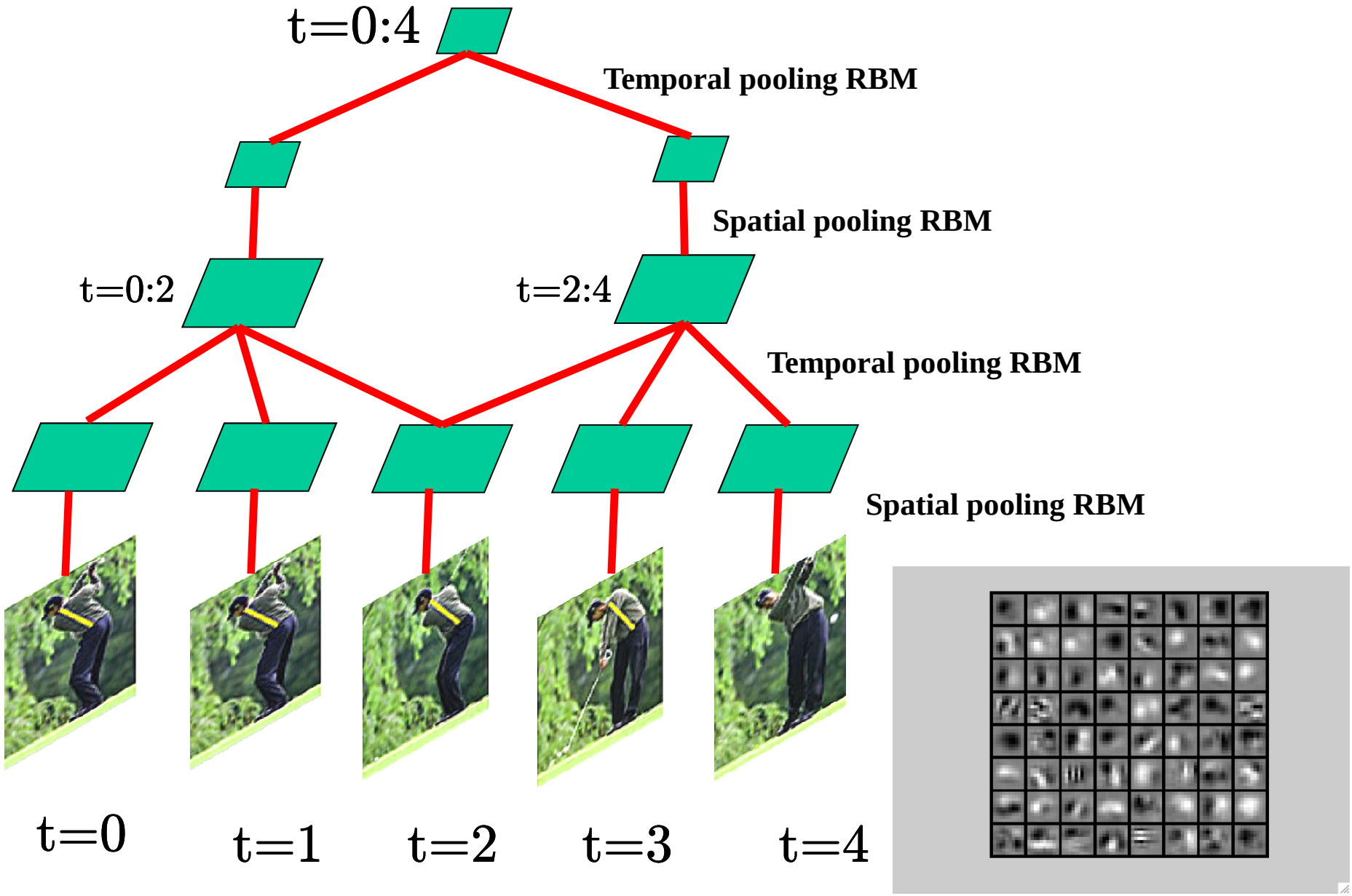
Advantages of these distributed feature representations

1. Unsupervised learning of features.
2. Lend themselves to transfer learning (self-taught learning).
3. Are memory efficient: Parts can be used in compositional models (e.g. deep nets).
4. Good generalization: *Blue animal with "big teeth" likely to be dangerous.*
5. Robust to occlusion and detection failures.
6. Follow an ecological-statistical stance.
7. Inspired by a biological system that works**.**

Deep learning (Hinton and collaborators)



Hierarchical spatio-temporal feature learning

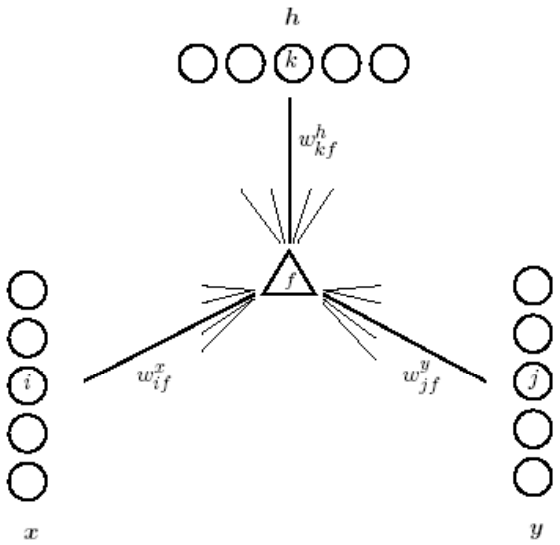


Hierarchical spatio-temporal feature learning

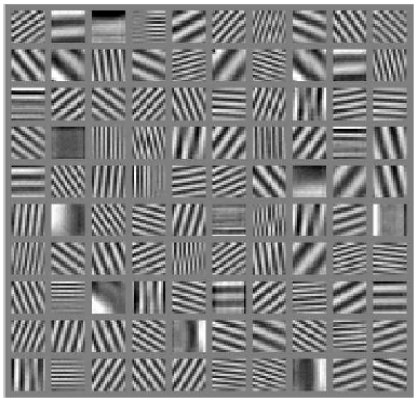
Observed gaze sequence

Model predictions

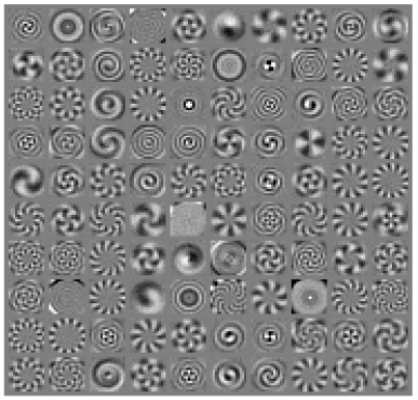
Learning image transformations and analogy



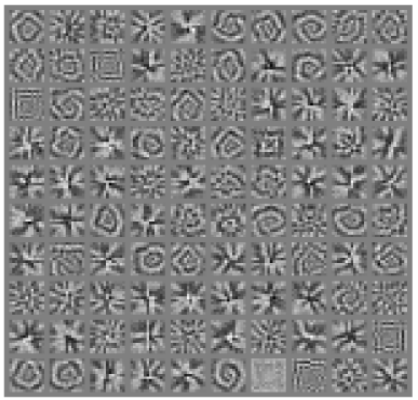
Translation



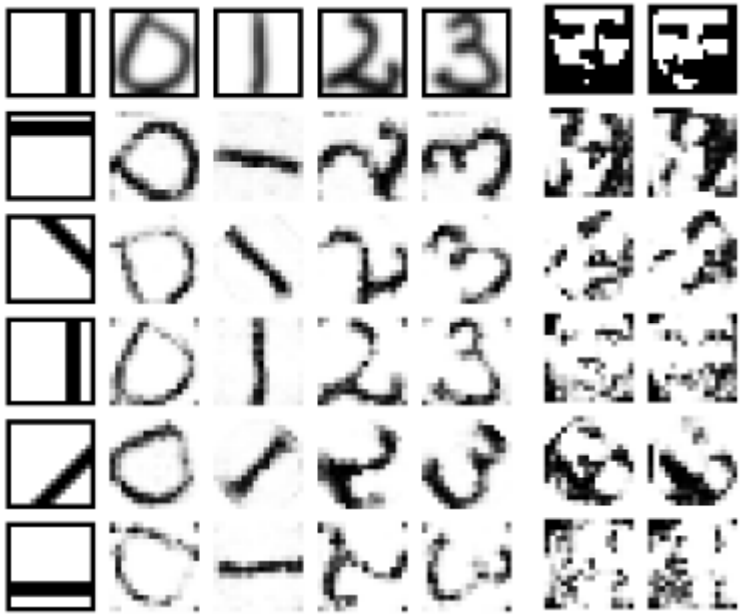
Rotation



Scaling

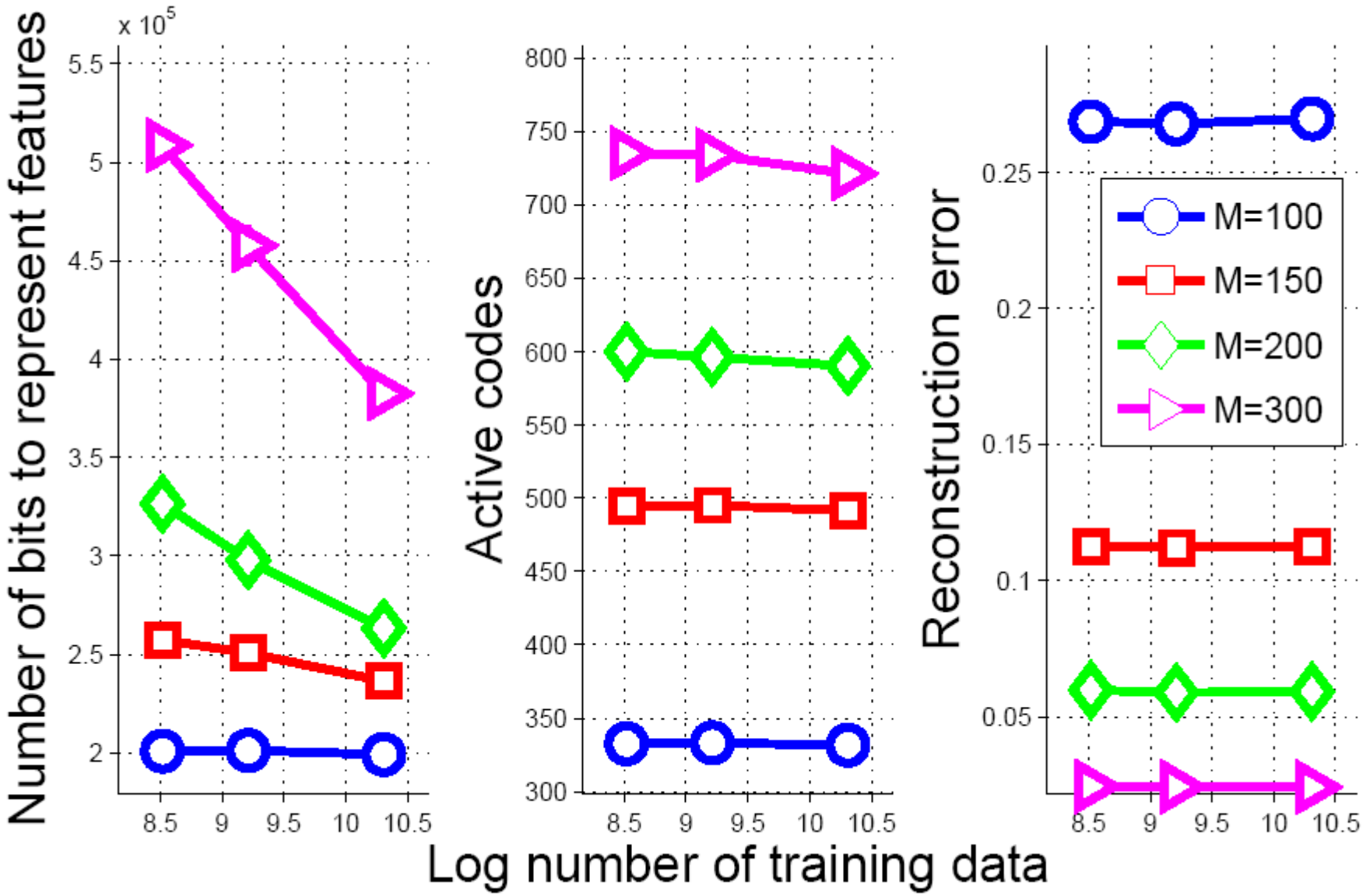


Learning by analogy



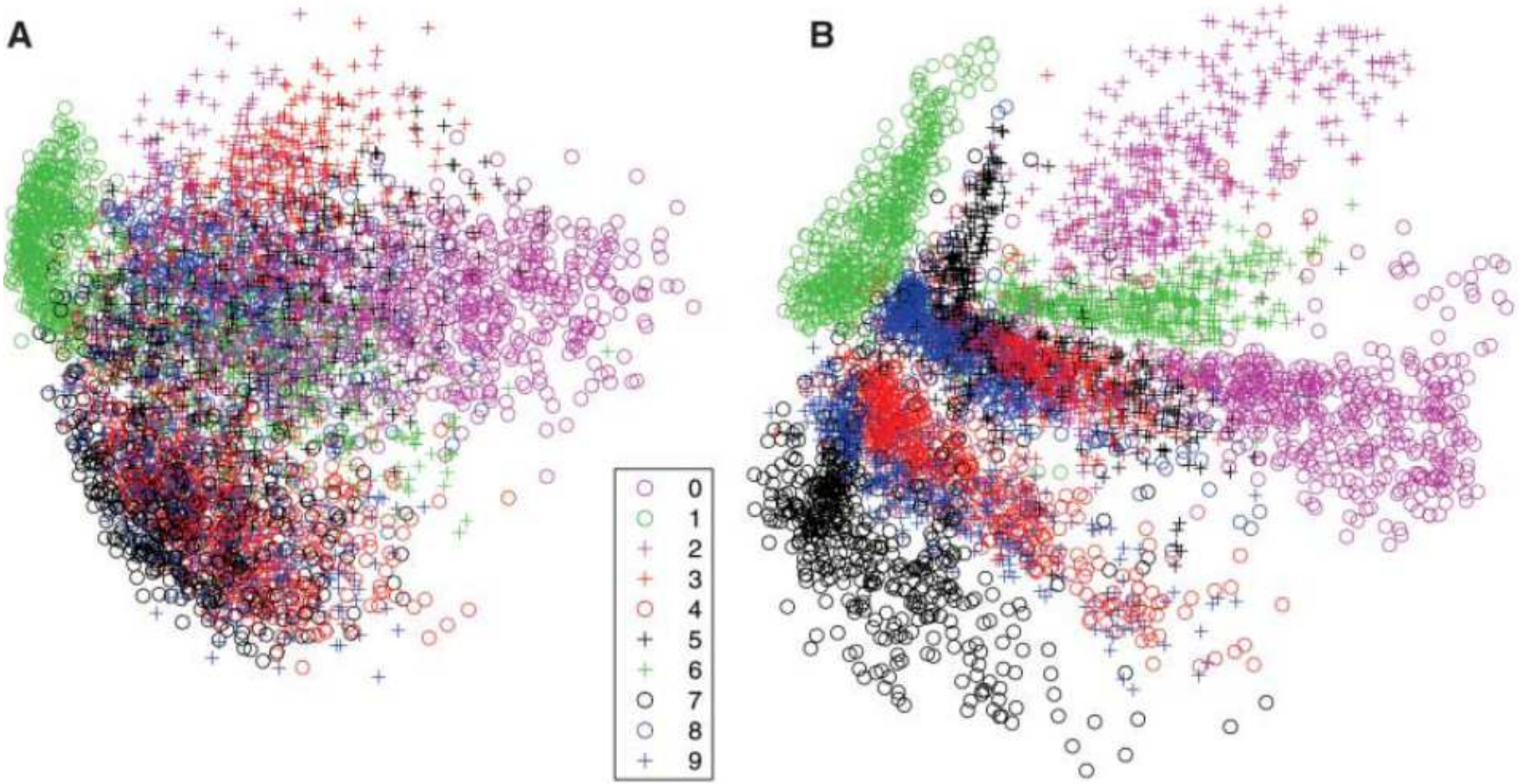
[Memisevic et al 2009]

The effect of dataset size



Deep net encodings for digits

(A) The two-dimensional codes for 500 digits of each class produced by taking the first two principal components of all 60,000 training images.
(B) The two-dimensional codes found by a 784-1000-500-250-2 autoencoder.



Challenges

- Storage and parallel data processing.
  - **Parallel** data processing (e.g., Hadoop MapReduce)
  - **Cloud** computing (e.g., Amazon's EC2)
  - Graphic processing units (**GPU**s)
- Privacy and other **social** phenomena.
- Data **security**.
- **Training** and supporting a new generation of data analysis and prediction experts.
- **Semantic understanding** of text, images, video, weather, medical, environmental and other data.

Thank you