Unsupervised Learning II

George Konidaris
gdk@cs.brown.edu

Fall 2021

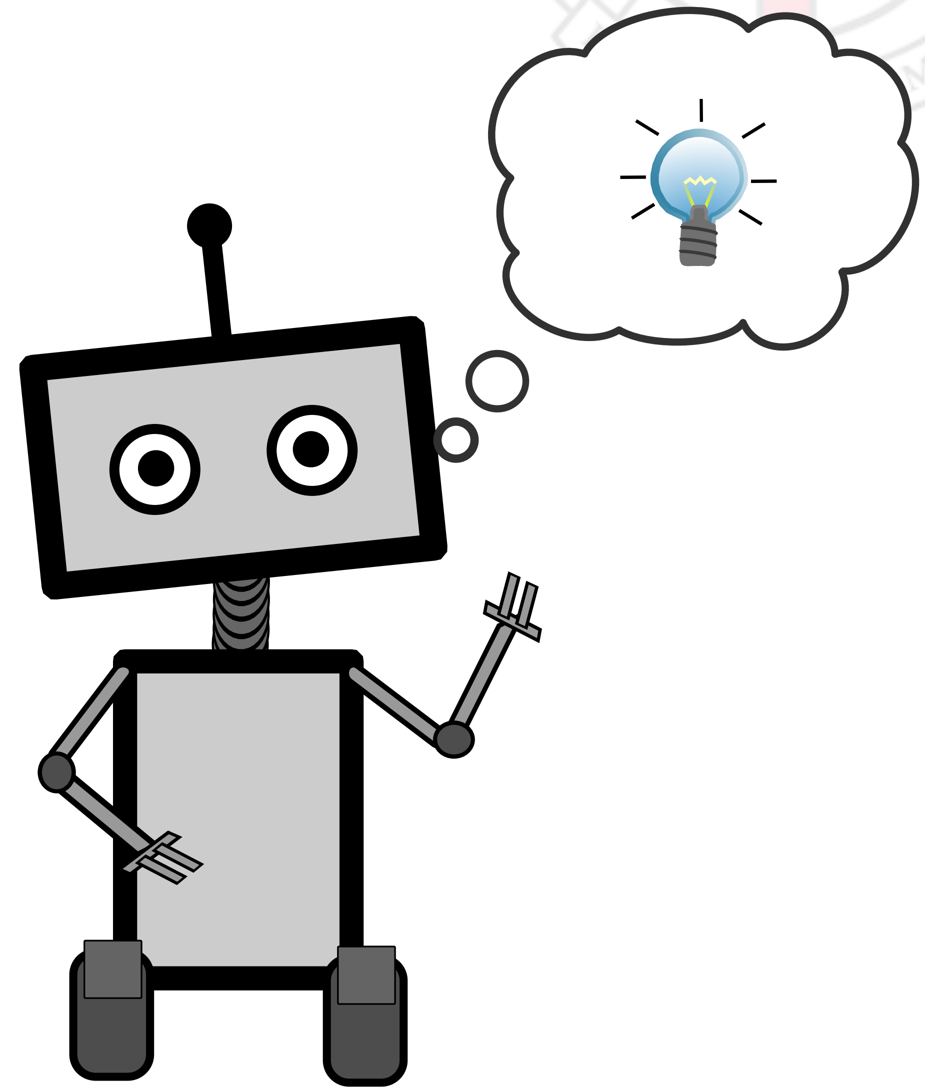
Machine Learning

Subfield of AI concerned with *learning from data*.

Broadly, using:

- ***Experience***
- To Improve ***Performance***
- On Some ***Task***

(Tom Mitchell, 1997)



Unsupervised Learning

Input:

$X = \{x_1, \ldots, x_n\}$ inputs

Try to understand the *structure of the data.*

E.g., how many types of cars? How can they vary?

So Far

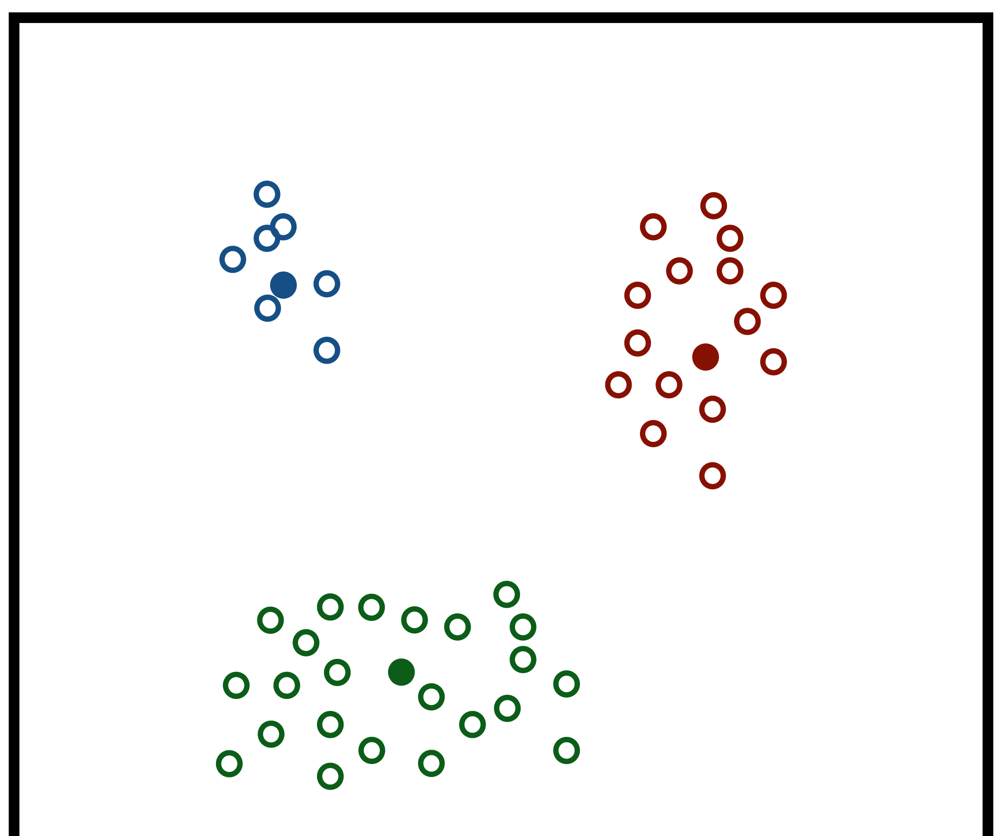
Clustering

Given:

- Data points $X = \{x_1, \ldots, x_n\}$.

Find:

- Number of clusters k
- Assignment function $f(x) = \{1, \ldots, k\}$



So Far

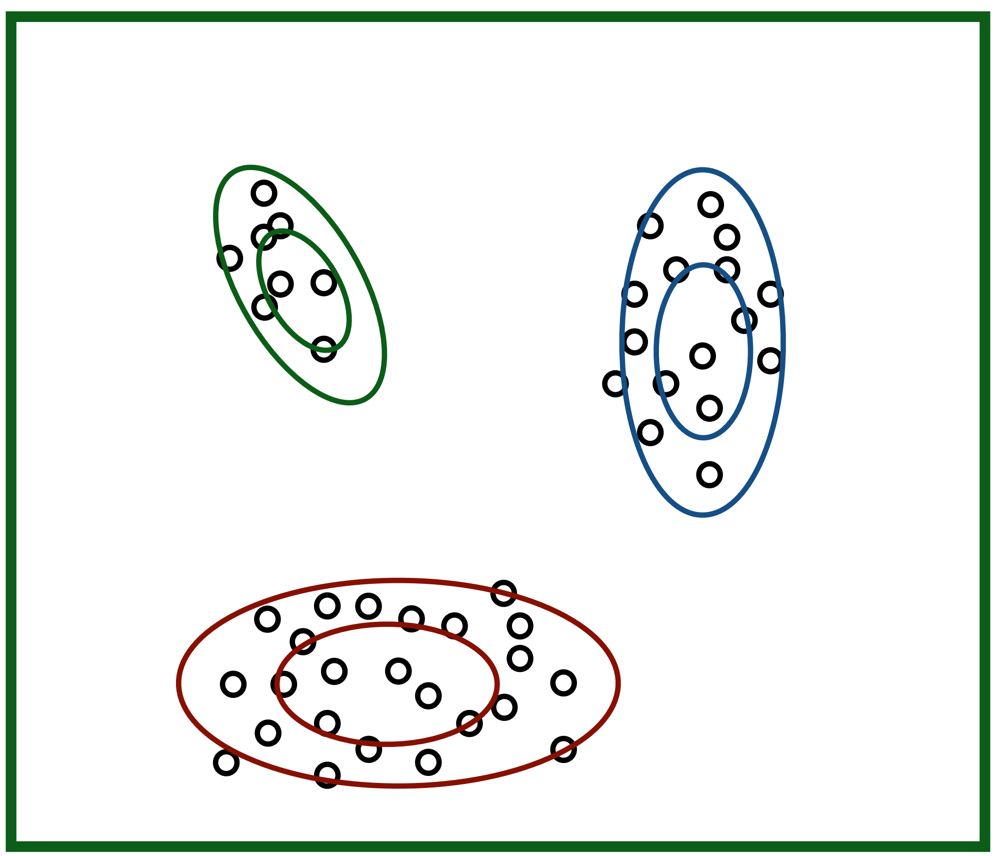
Density Estimation

Given:

- Data points $X = \{x_1, \ldots, x_n\}$.

Find:

- PDF $P(x)$



So Far

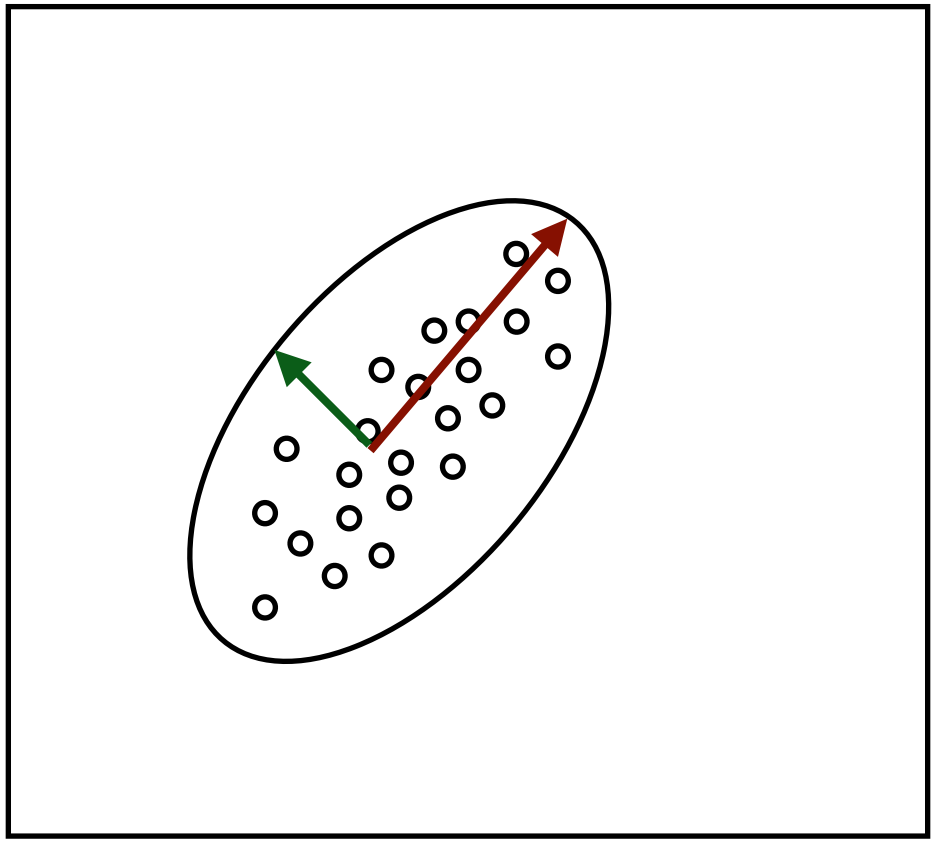
Dimensionality Reduction

Given:

- Data points $X = \{x_1, \ldots, x_n\}$.

Find:

- $f : X \rightarrow X'$
- $|X'| << |X|$



PCA

- Gather data $X_1, \ldots, X_m$.
- Adjust data to be zero-mean:

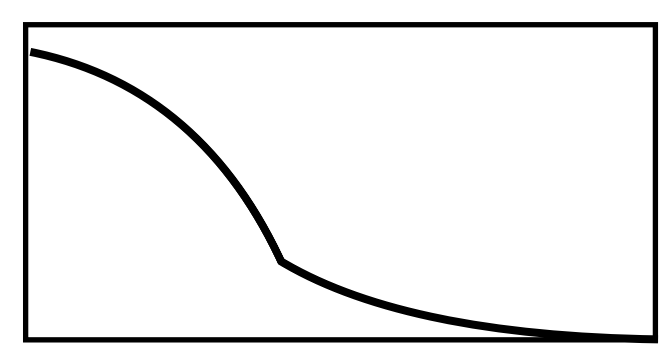
$$X_i = X_i - \sum_j \frac{X_j}{m}$$

- Compute covariance matrix C.
- Compute unit eigenvectors V_i and eigenvalues v_i of C.

Each V_i is a direction, and each v_i is its importance - the amount of the data's variance it accounts for.

New data points:

$$\hat{X}_i = [V_1, ..., V_p] X_i$$



PCA

Reconstruction:

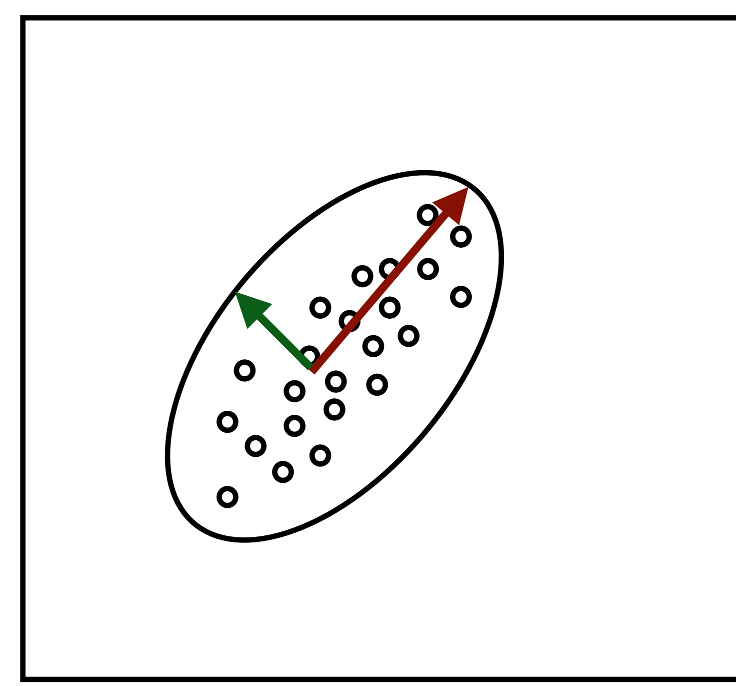
$$\bar{X}_i = V_1 \hat{X}_i[1] + V_2 \hat{X}_i[2] + ... + V_p \hat{X}_i[p]$$

orthogonal axes

real valued numbers

Every data point is expressed as a point in a new coordinate frame.

Equivalently: weighted sum of basis (eigenvector) functions.



Autoencoders

Fundamental issue with PCA:
Linear reconstruction.

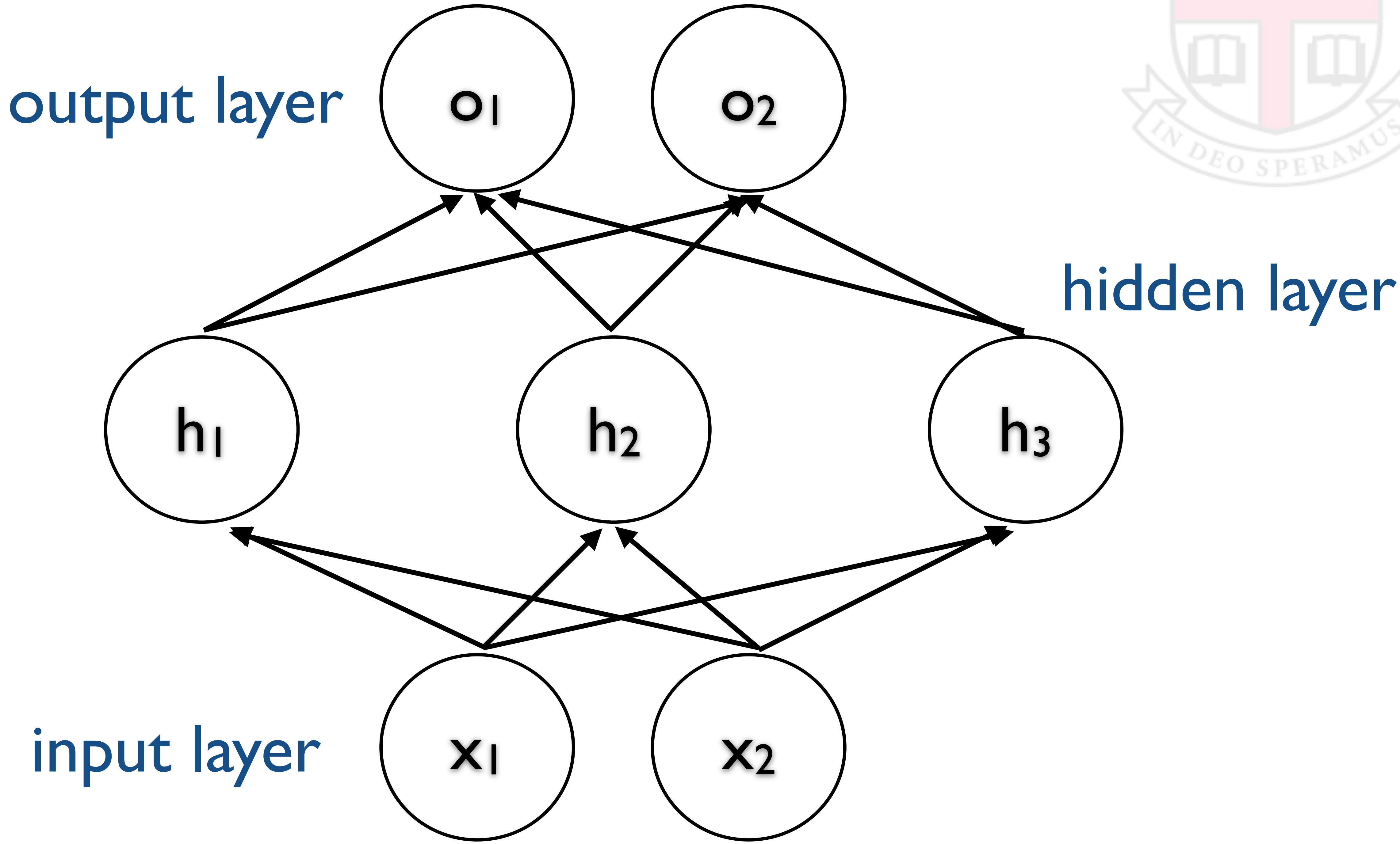
Can we use a nonlinear method for construction?

- Extract more complex relationships within the data.
- Remove "linear reconstruction" property.

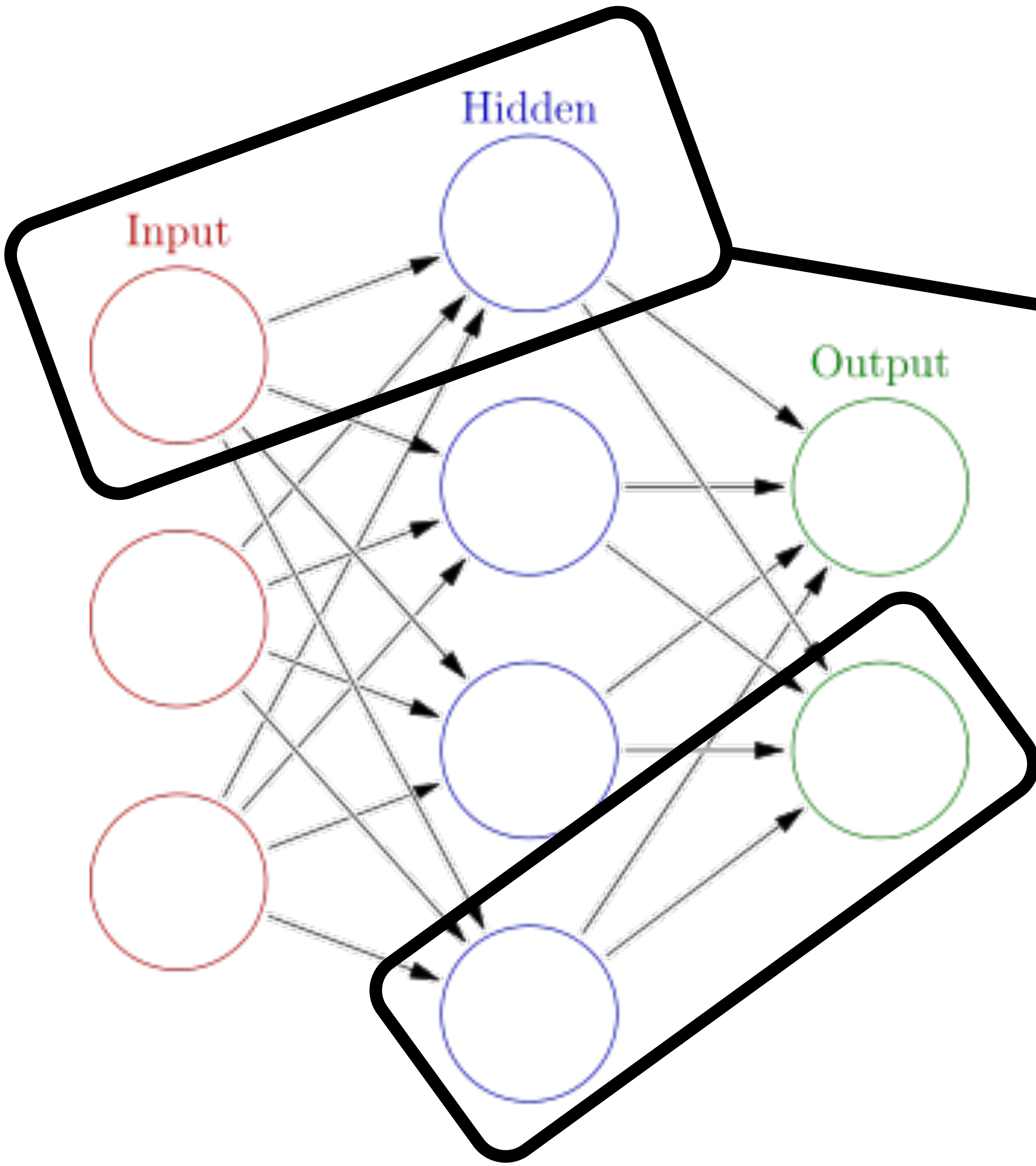
Yes, there are several.

- Let's talk about neural nets.

Neural Network Regression



Neural Network Regression

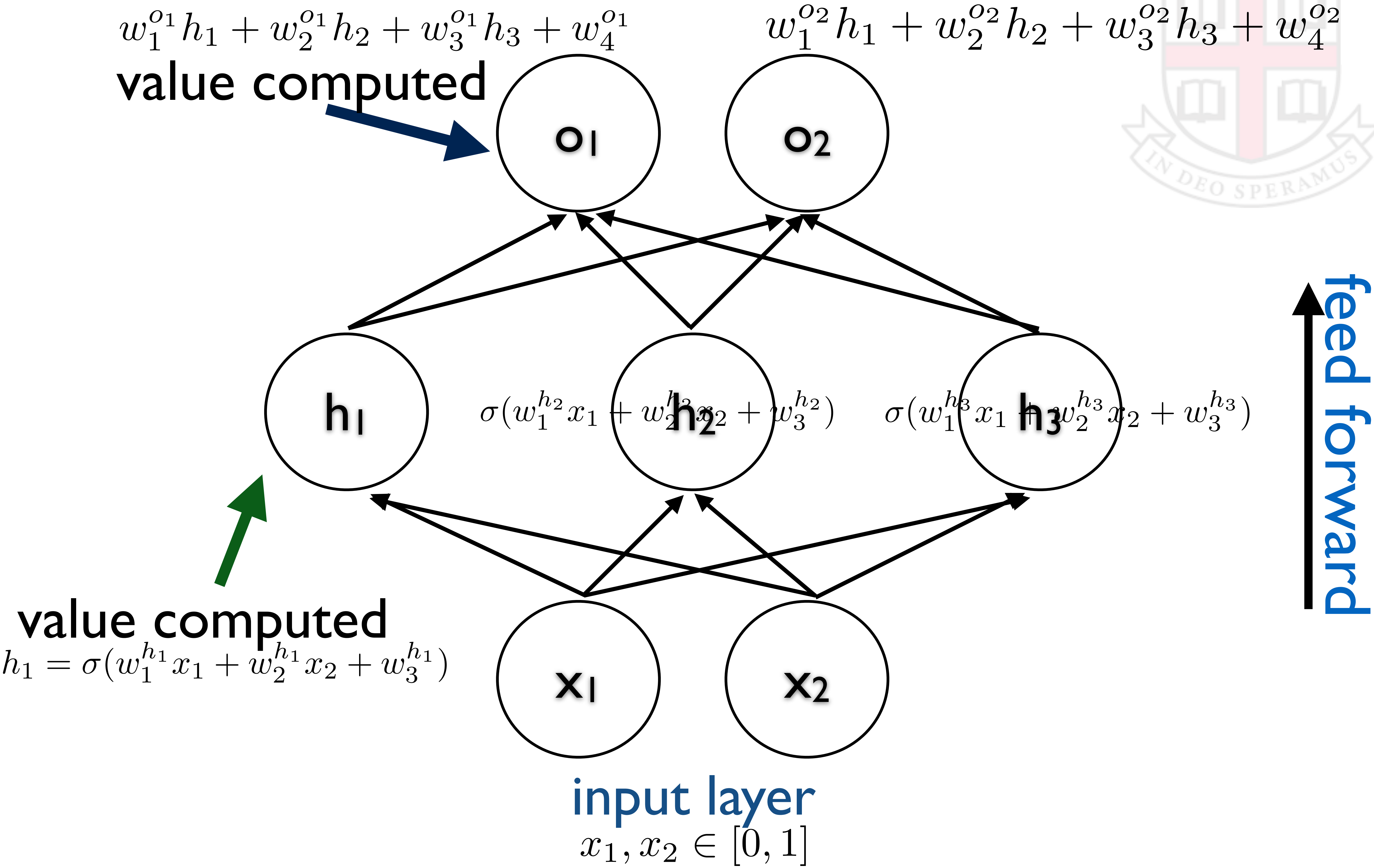


$\sigma(w \cdot x + c)$

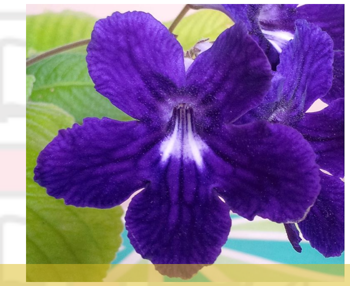
$w \cdot x + c$

regression

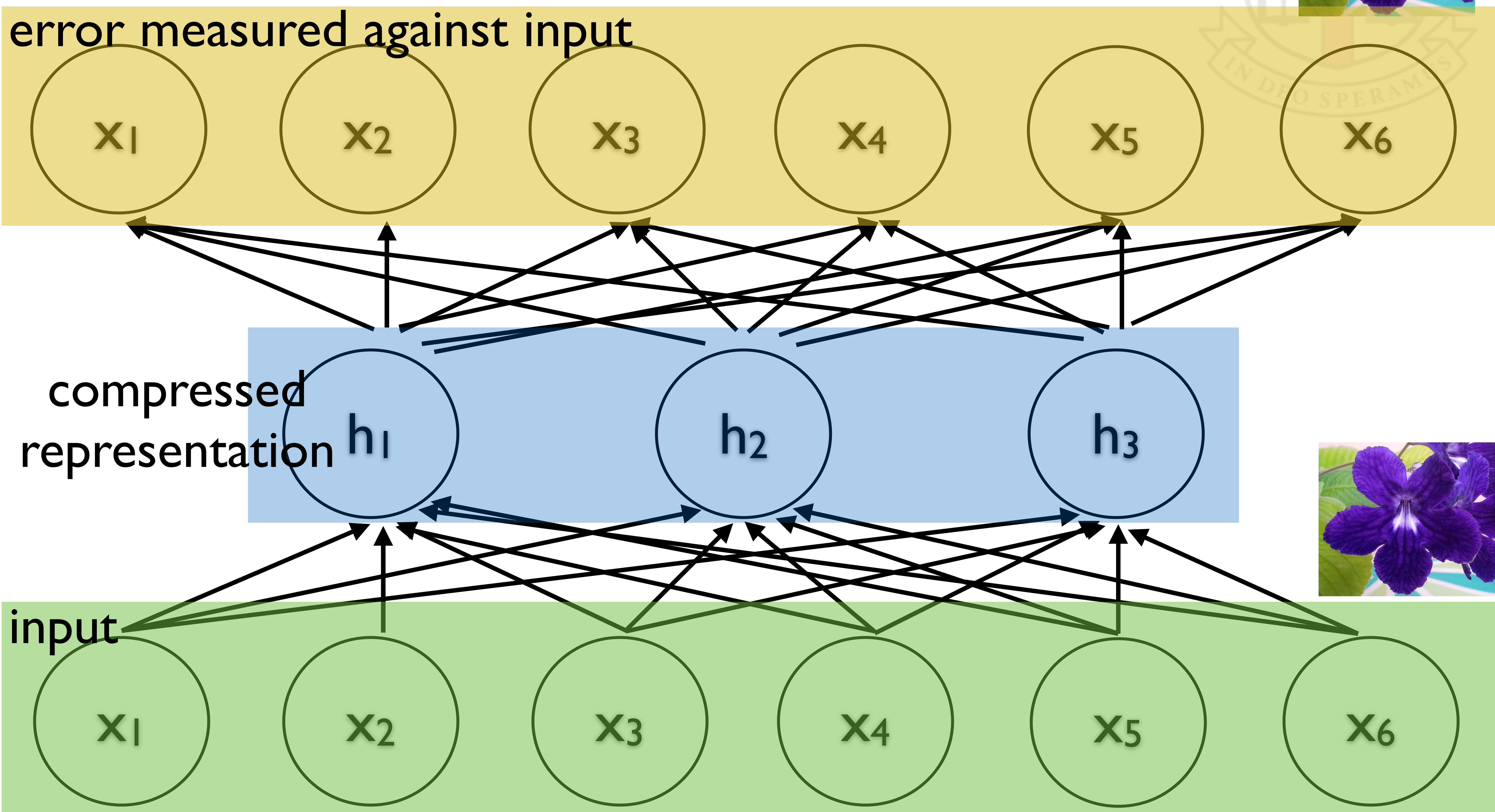
Neural Network Regression



Autoencoders

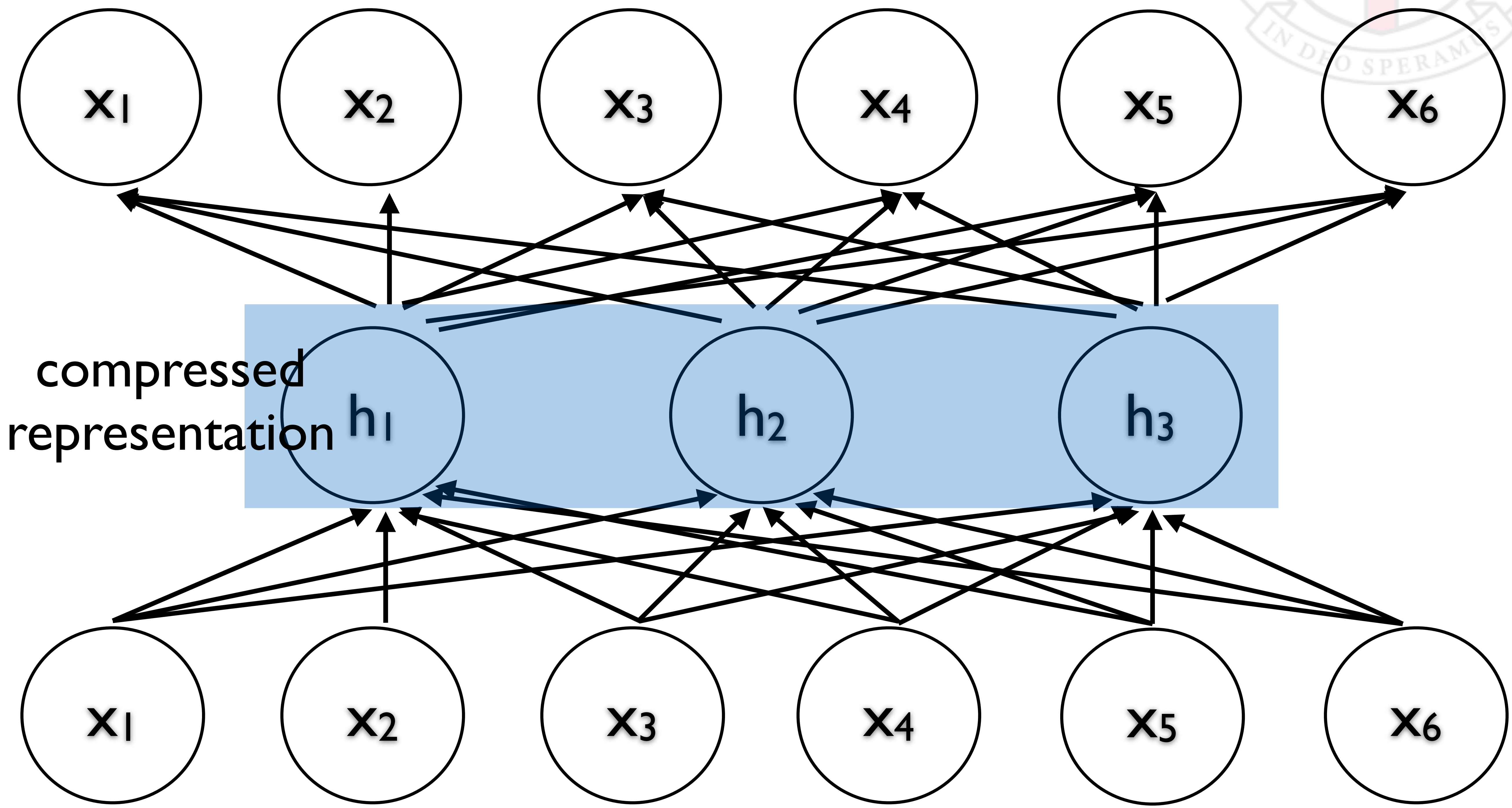


Idea: train the network to reproduce the output.

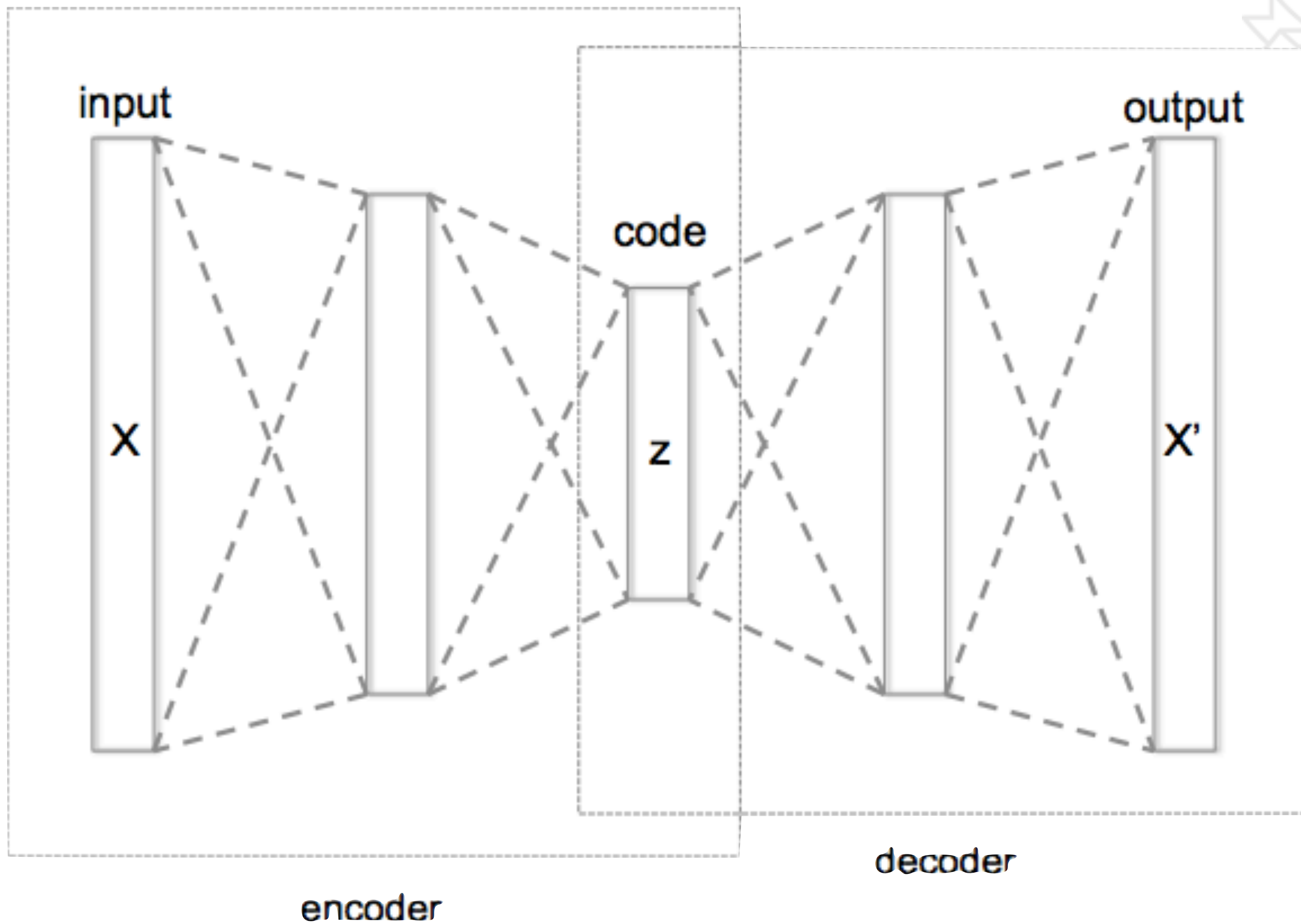


Autoencoders

The compressed representation is *sufficient* to reproduce input.

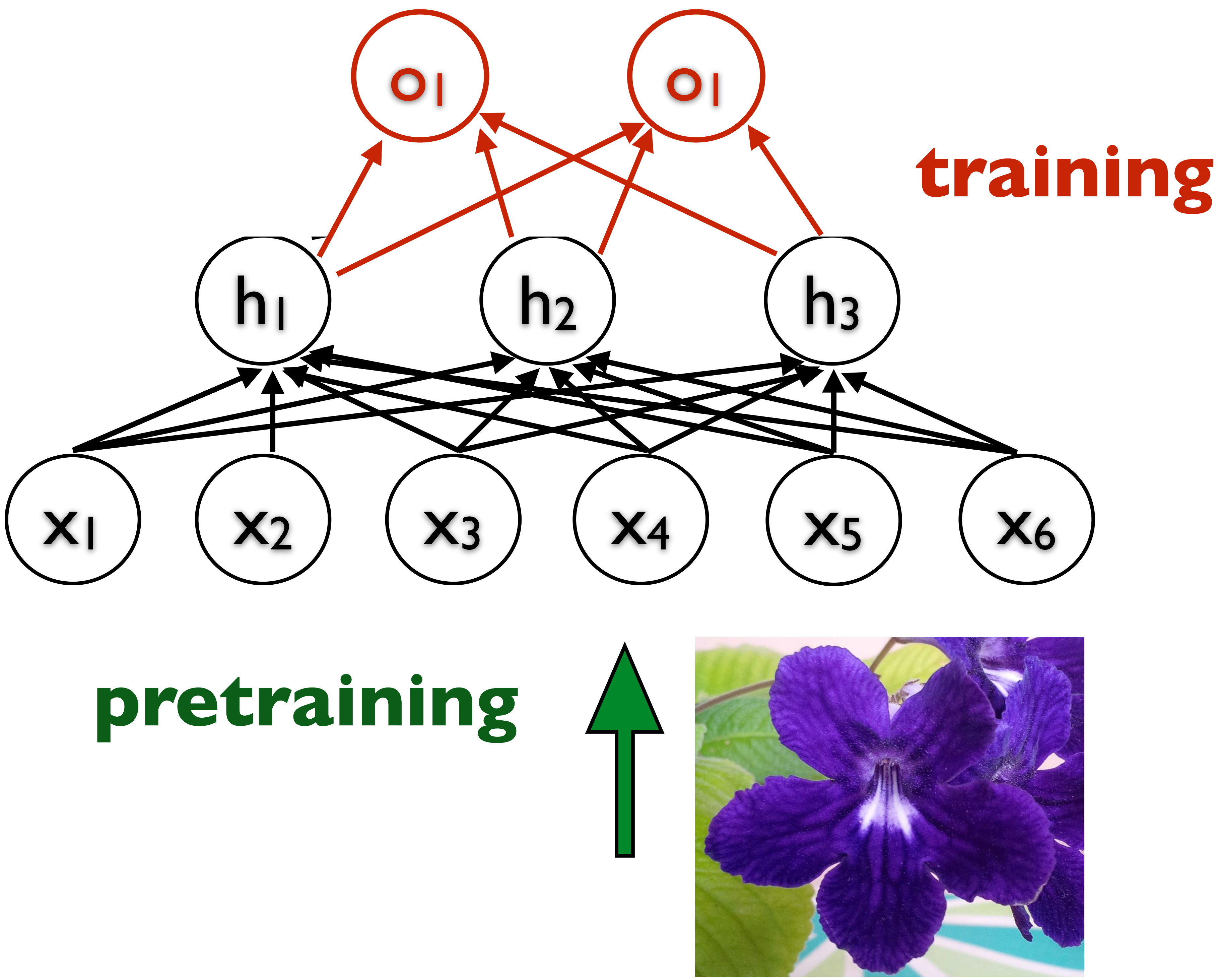


Autoencoders



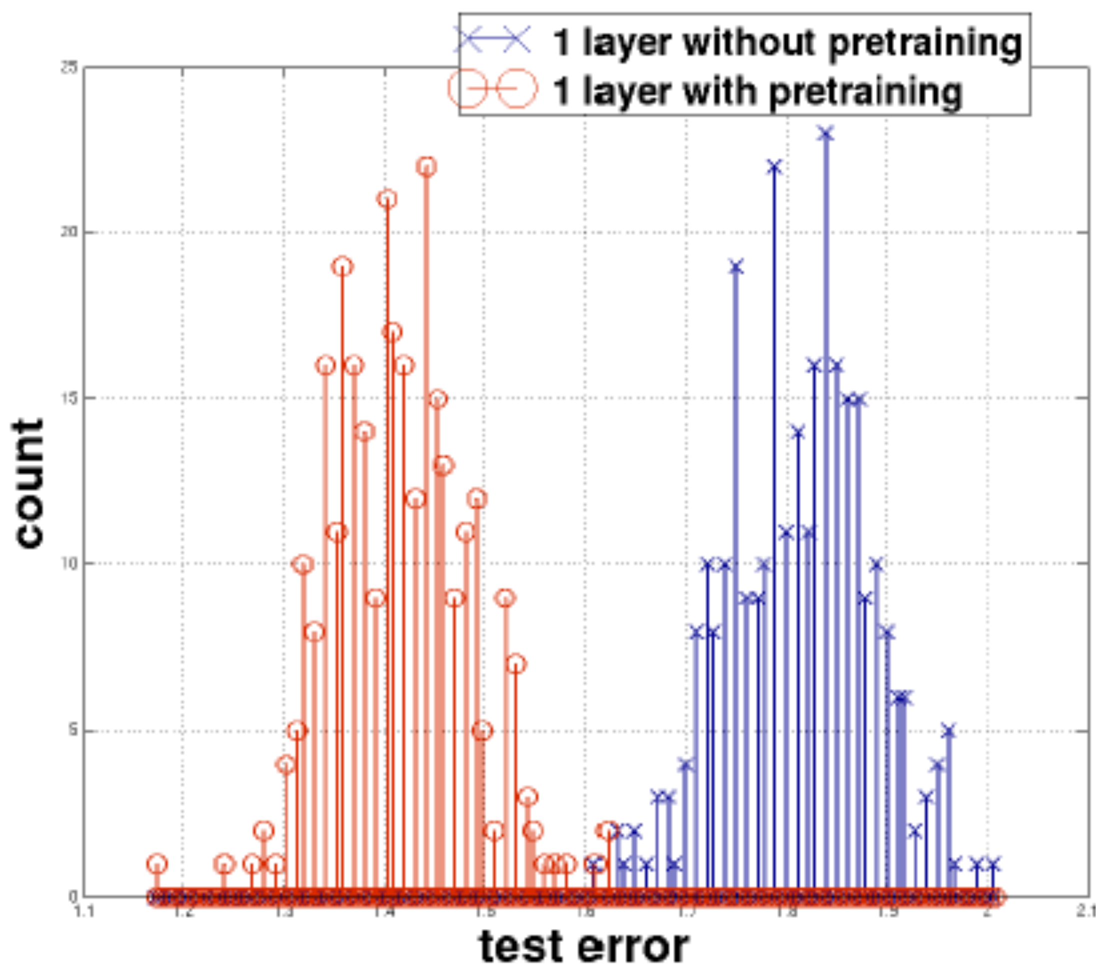
(wiki)

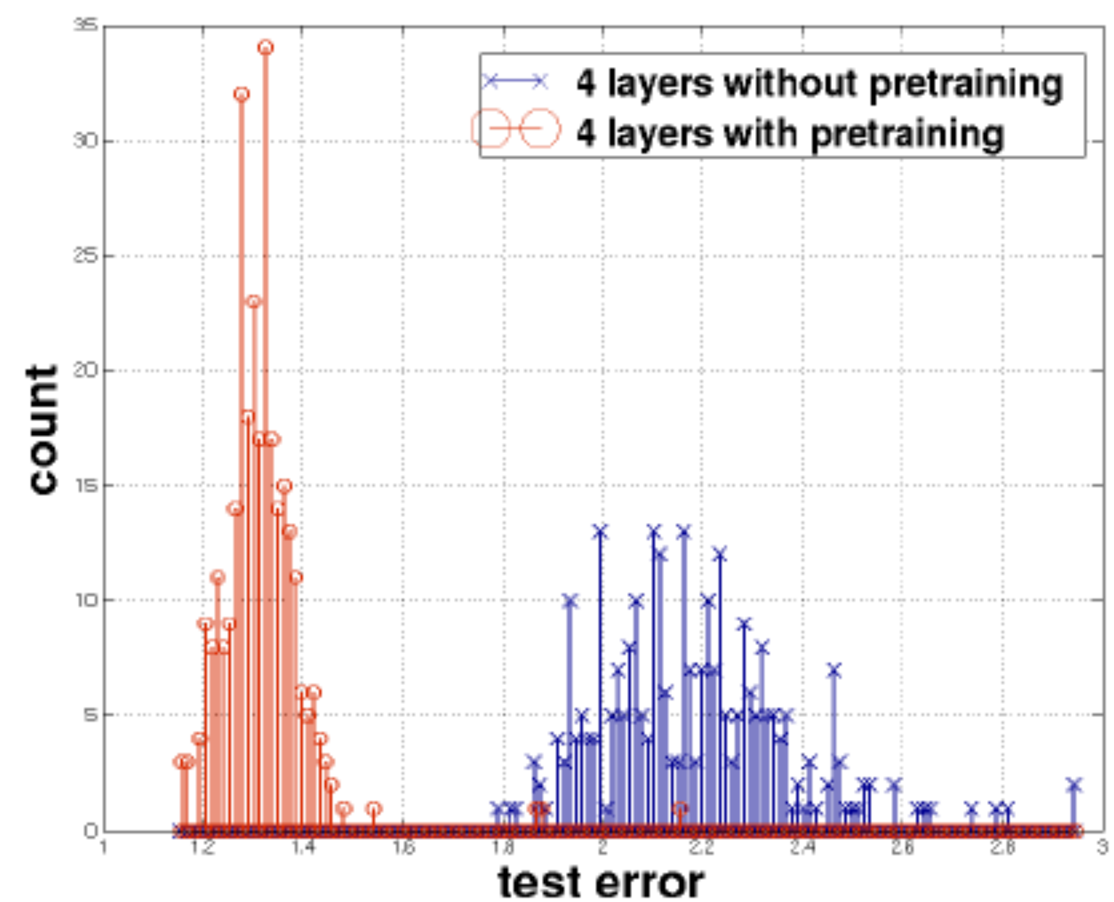
Autoencoders for Classification



Autoencoders

How helpful is this for classification?





[Erhan et al., 2010]

Fun with Autoencoders

Denoising Autoencoders

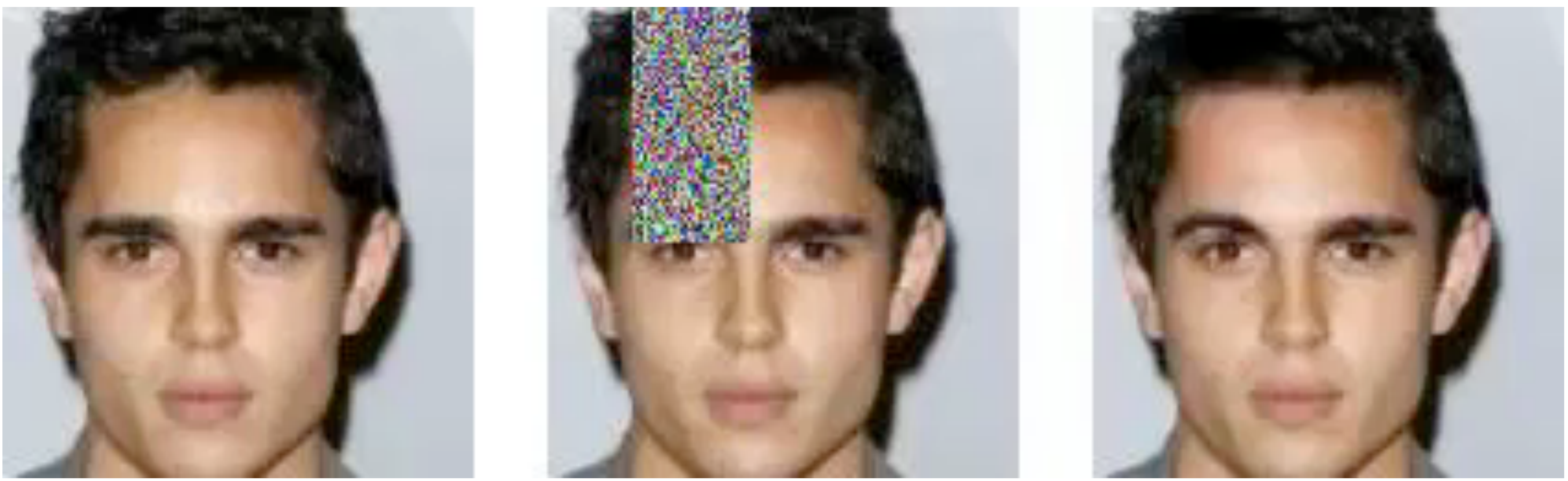
- Input noisy version of the image
- Optimize error with respect to original image
- Deep autoencoder learns to "clean"

via OpenDeep.org

Fun with Autoencoders

Image completion

- Train with parts of the image deleted
- Measure error on the completed image



via Yijun Li

Unsupervised Learning

Yet another type!

Latent Structure Learning
What hidden structure explains the data?

<u>Given</u>:

- Data points $X = \{x_1, \ldots, x_n\}$.

<u>Find:</u>

- Latent variables *Z*.
- PDF $P(X|Z)$

Topic Modeling

Common problem in *Natural Language Processing*.

Collection of documents

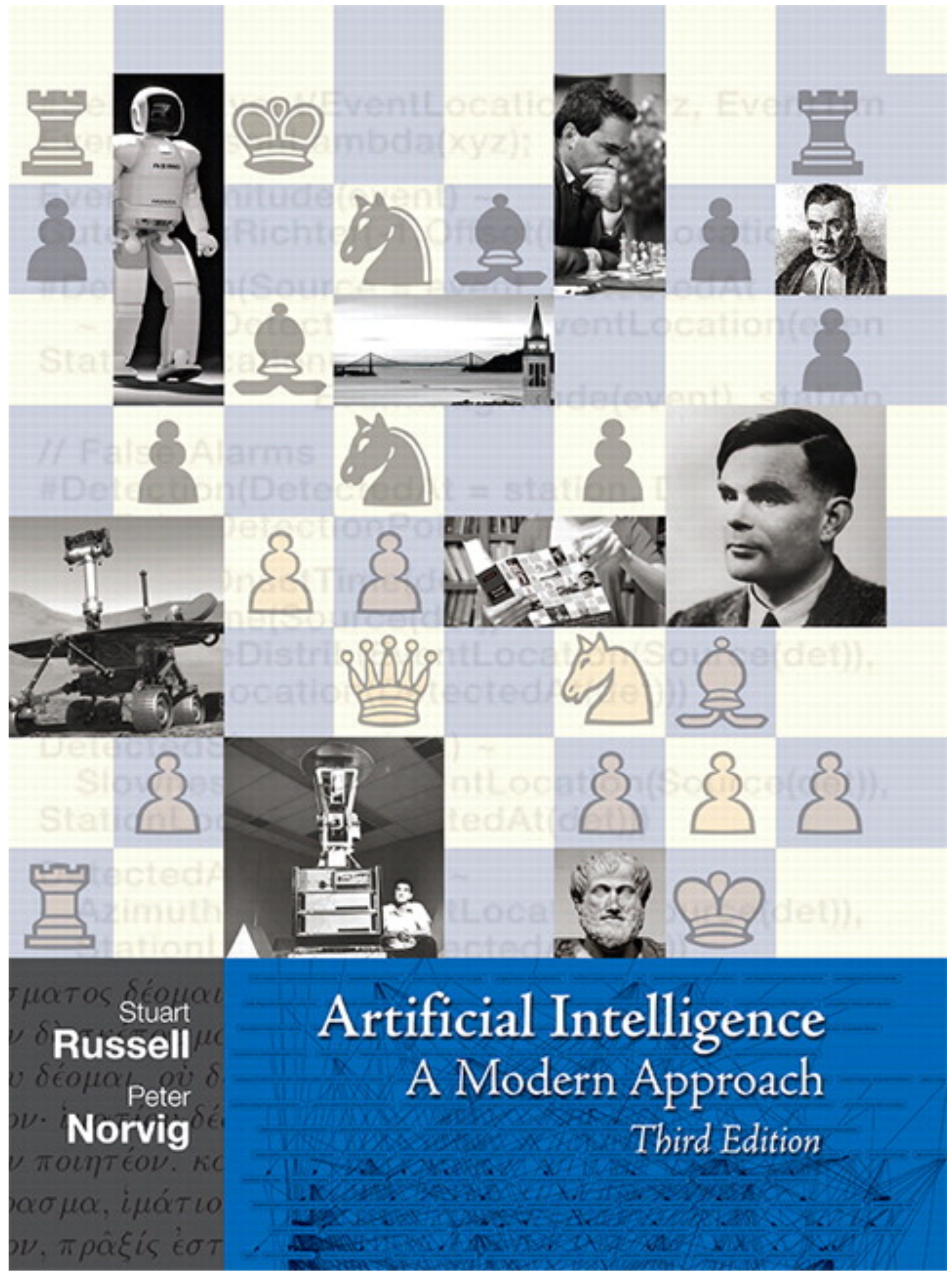
- $X = \{x_1, \ldots, x_n\}$
- Each x_i is a sequence of words

Assume that they are *about something*.

Specifically:

- Latent topics Z.
- Each topic z generates similar language across documents.

Topics


Stuart
Russell
Peter
Norvig
Artificial Intelligence
A Modern Approach
Third Edition

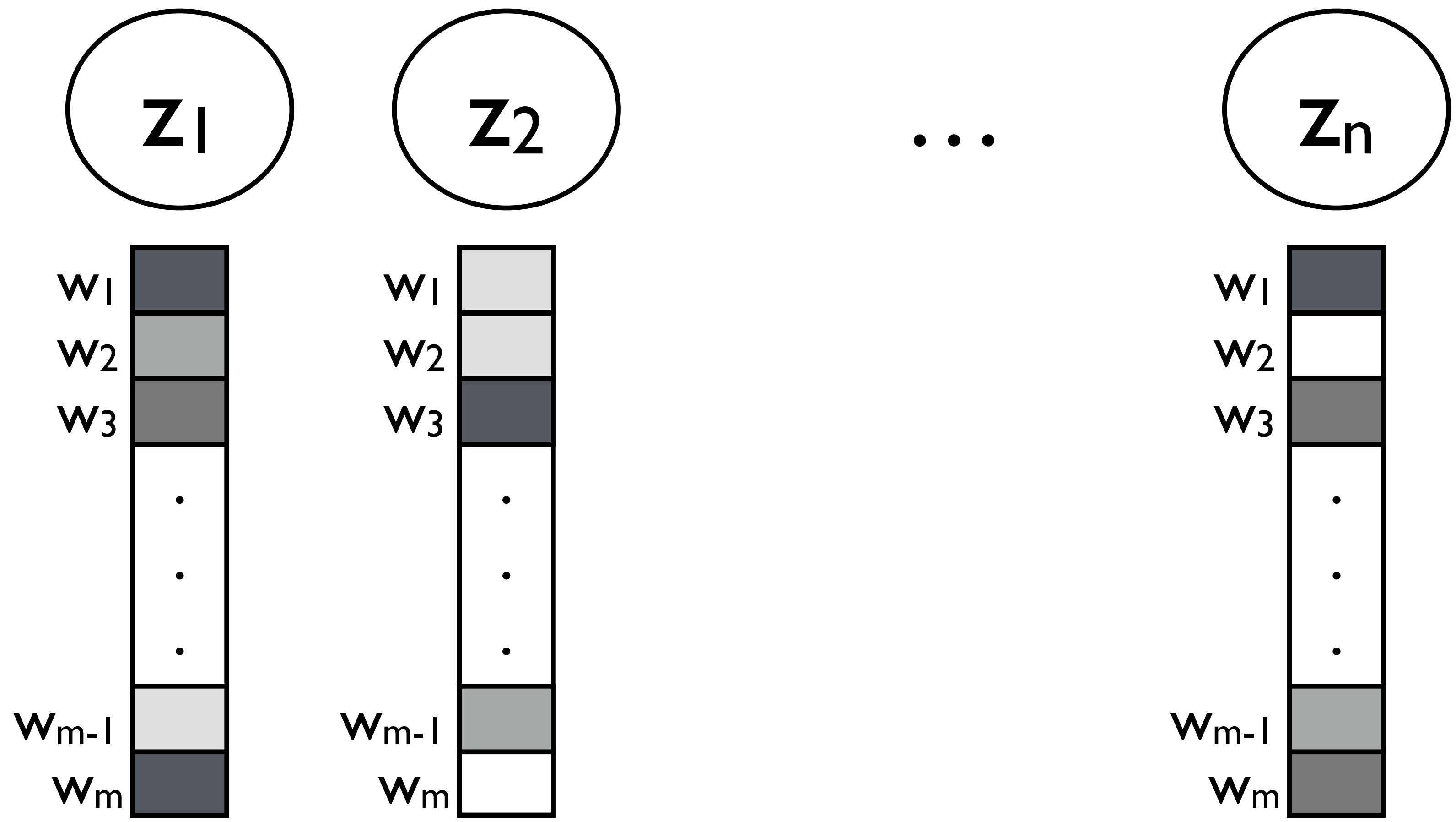
Topics

MELVIL DEWEY.

LDA

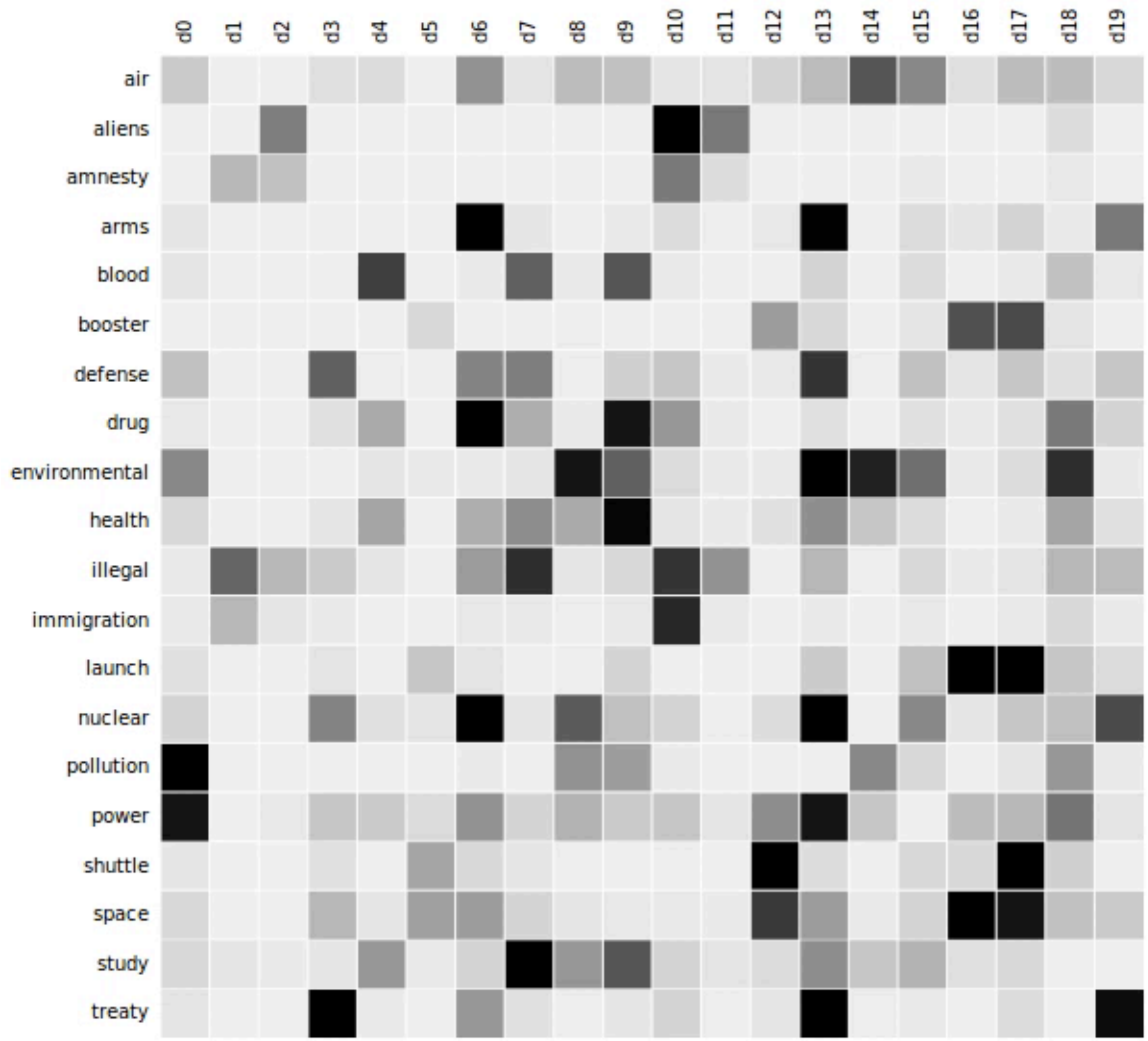
Bayes Net for describing topic models.

There is a set of hidden topics, *Z*, and a set of words, *W*.



Each topic z_i has a conditional probability of each word w_j appearing in a document: $P(w_j \mid z_i)$

Topic Modeling



(wiki)

LDA

Each document is modeled as …

A combination of topics

- Expressed as a distribution over topics
- The probability that *each word is drawn from each topic.*

A collection of words

- Each word is drawn at random from a topic.
- Order doesn't matter (anywhere).

obviously wrong

Goal:

- Infer number of topics, distribution
- Infer per-topic distribution over words
- Describe each document as mixture of topics

LDA

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

AP corpus: 16k articles

Data Mining

Most common application of unsupervised learning.

Given large corpus of data, what can be learned?

Lots of subproblems:

- Database management
- Privacy
- Visualization
- Unsupervised learning

Any unsupervised method can be applied in principle.

Most common in industry:

- Learning associations and patterns.

Data Mining

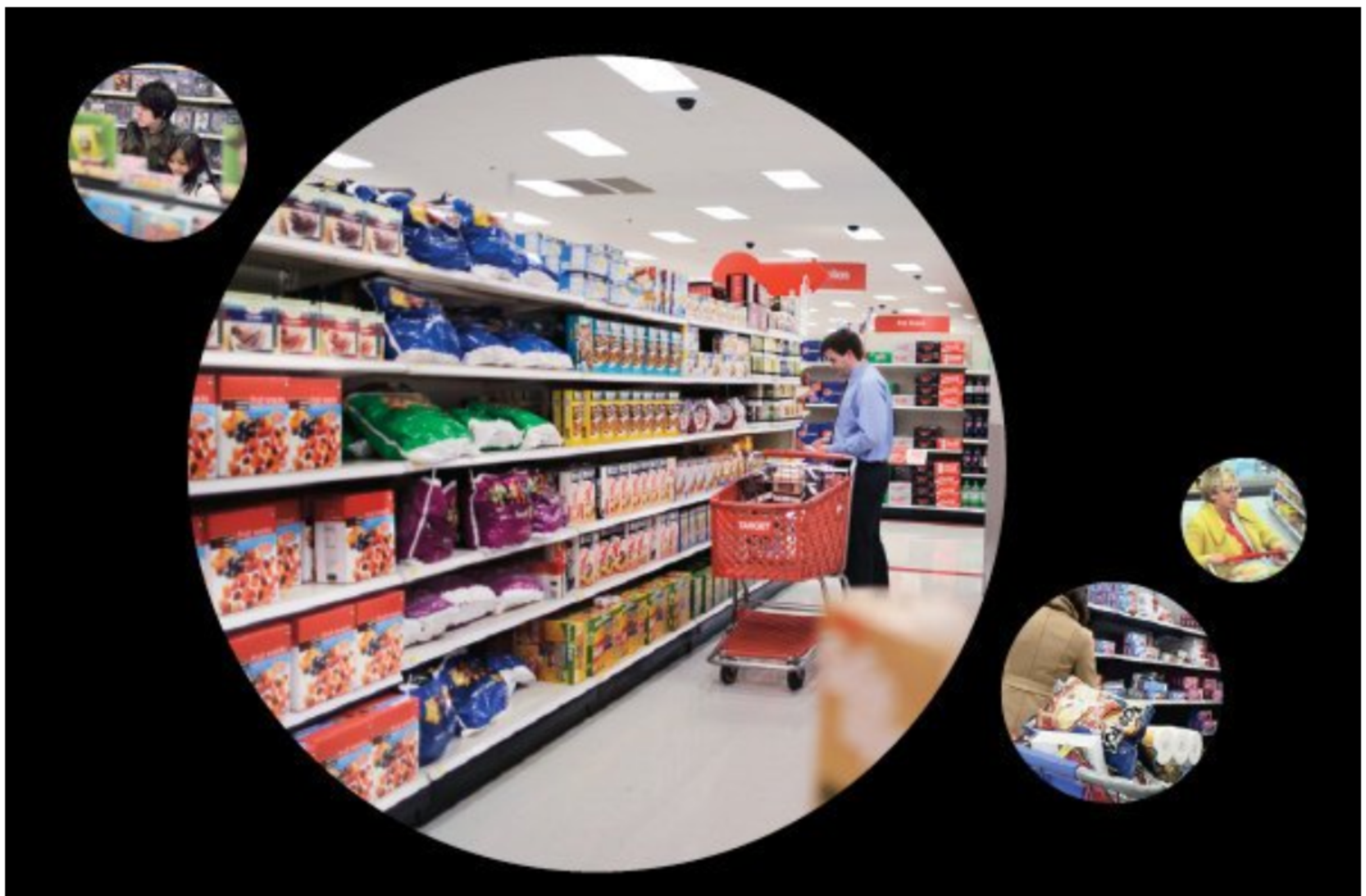

Stop&Shop


CharlieCard
Massachusetts Bay Transportation Authority

Data Mining

How Companies Learn Your Secrets

By CHARLES DUHIGG FEB. 16, 2012



Data Mining

"As Pole's computers crawled through the data, he was able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a "pregnancy prediction" score. More important, he could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.

One Target employee I spoke to provided a hypothetical example. Take a fictional Target shopper named Jenny Ward, who is 23, lives in Atlanta and in March bought cocoa-butter lotion, a purse large enough to double as a diaper bag, zinc and magnesium supplements and a bright blue rug. There's, say, an 87 percent chance that she's pregnant and that her delivery date is sometime in late August."

Your Smartphone

So far, Jebara says, Sense Networks has categorized 20 types, or “tribes,” of people in cities, including “young and edgy,” “business traveler,” “weekend mole,” and “homebody.” These tribes are determined using three types of data: a person’s “flow,” or movements around a city; publicly available data concerning the company addresses in a city; and demographic data collected by the U.S. Census Bureau. If a person spends the evening in a certain neighborhood, it’s more likely that she lives in that neighborhood and shares some of its demographic traits.

https://www.technologyreview.com/s/412529/mapping-a-citys-rhythm/

Spurious Correlations

http://www.tylervigen.com/spurious-correlations