Carnegie Mellon

15-826: Multimedia (Databases) and Data Mining

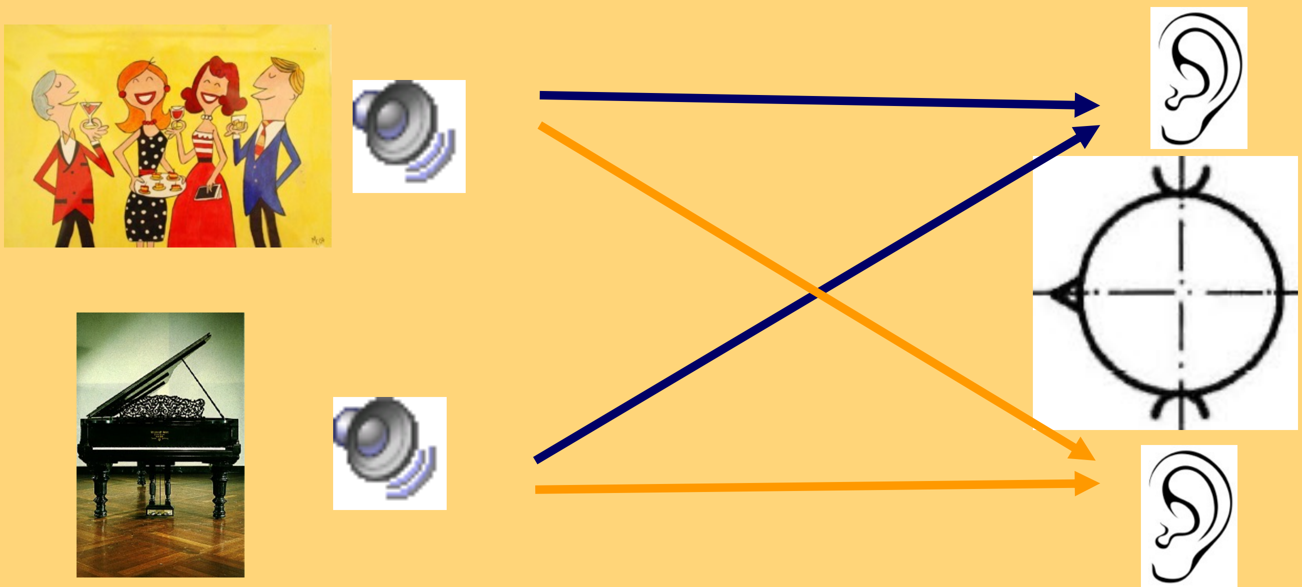
Lecture #20:

Independent Component Analysis (ICA)

Christos Faloutsos

Problem: BSS

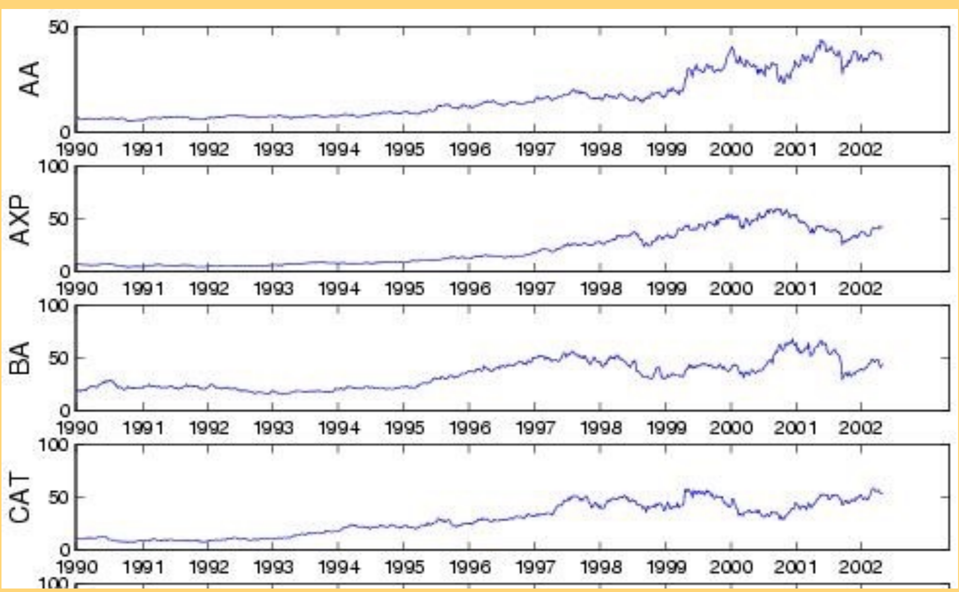
- two sound sources in a cocktail party – separate them

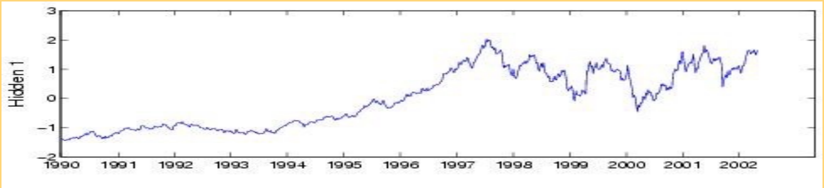


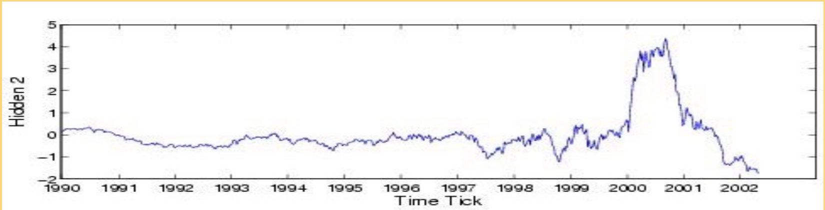
="blind source separation"
(= unknown sources, unknown mixing)

Problem

Q: how to extract **sparse** hidden/latent variables?





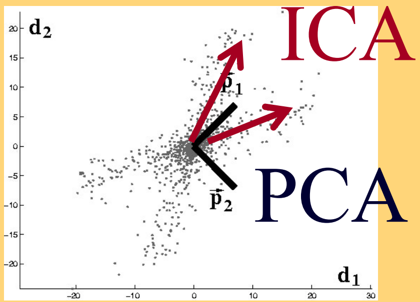


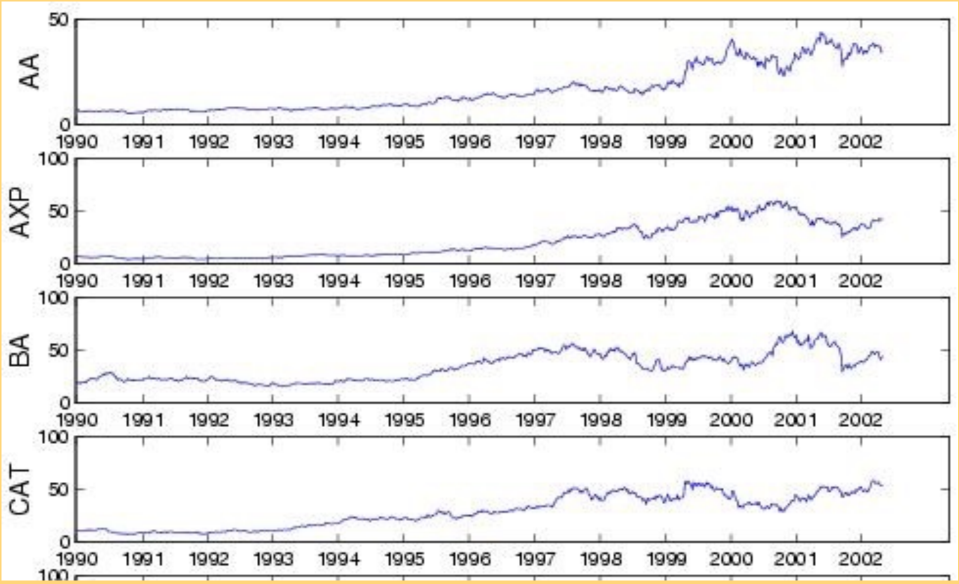
Carnegie Mellon

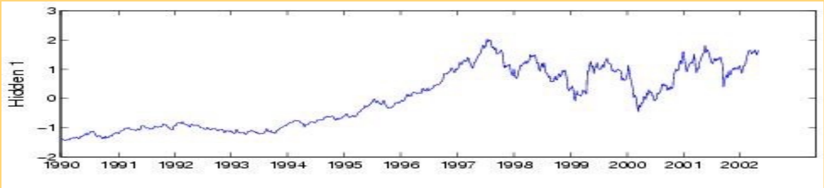
Answer

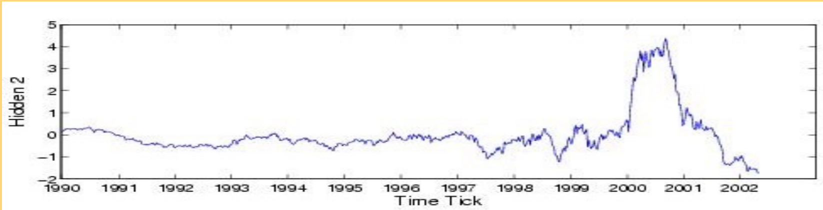
Q: how to extract **sparse** hidden/latent variables?

A: ~~SVD~~ ICA









Must-read Material

- *AutoSplit: Fast and Scalable Discovery of Hidden Variables in Stream and Multimedia Databases,* **Jia-Yu Pan**, Hiroyuki Kitagawa, Christos Faloutsos and Masafumi Hamamoto, PAKDD 2004, Sydney, Australia

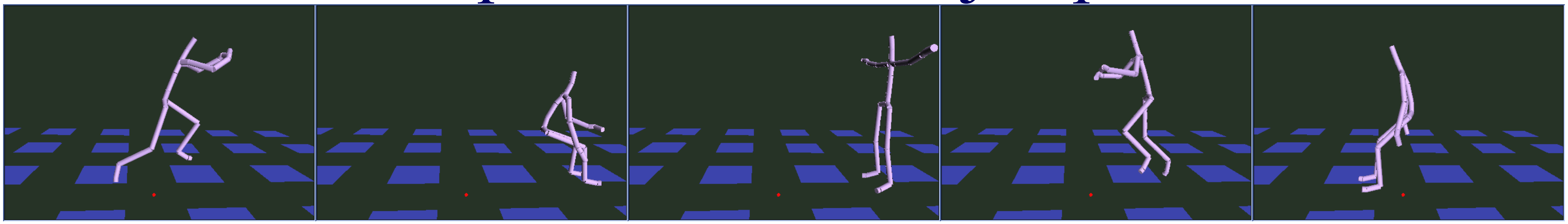
Outline

- Motivation
- Formulation
- PCA and ICA
- Example applications
- Conclusion

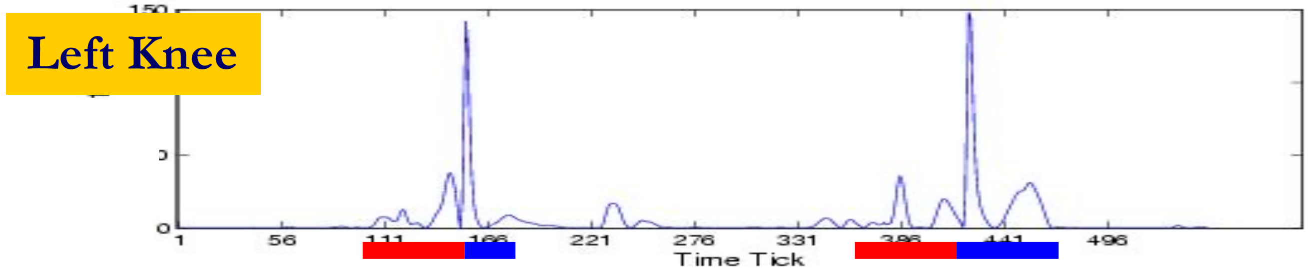
Carnegie Mellon

Motivation: (Q1) Find patterns in data

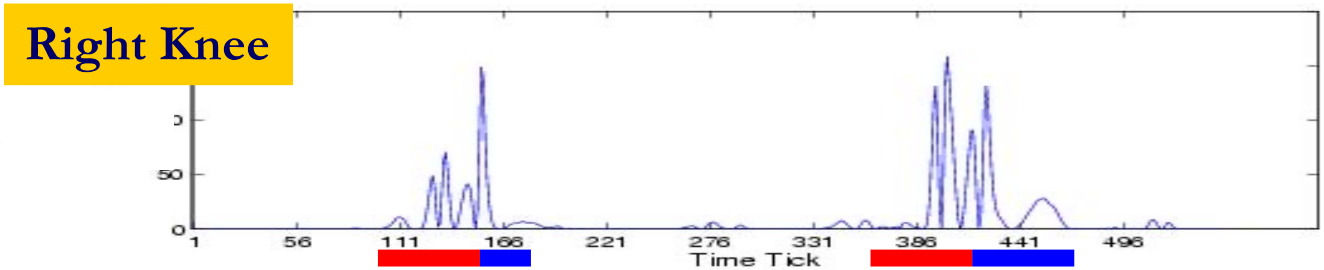
- Motion capture data: broad jumps



Energy exerted

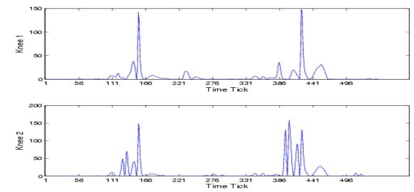


Energy exerted



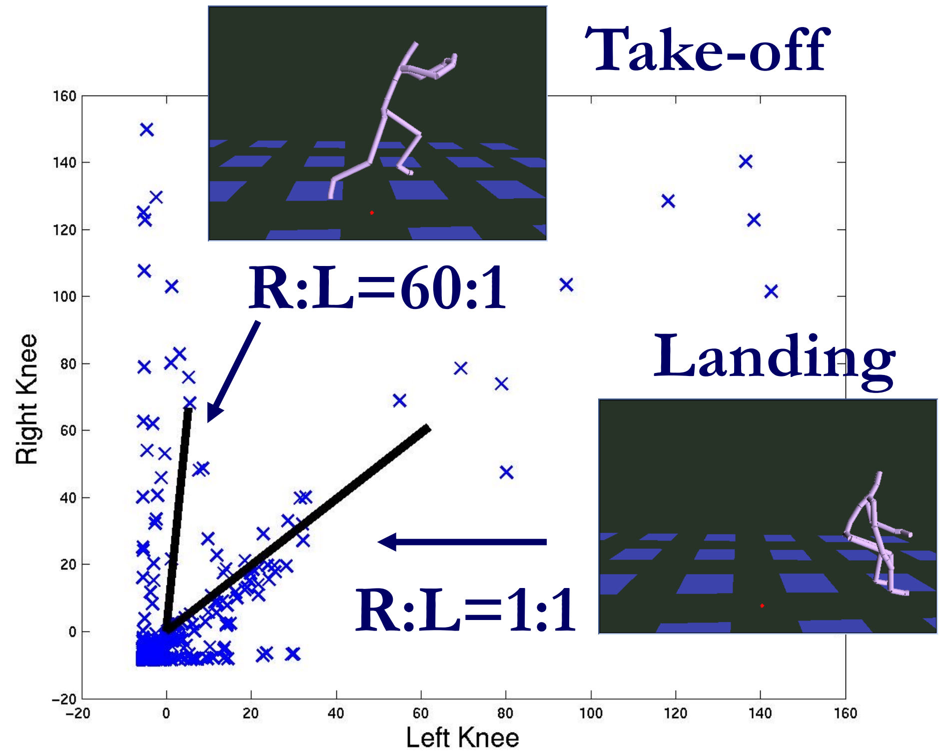
Take-off
Landing

Carnegie Mellon



Motivation:
(Q1) Find patterns in data

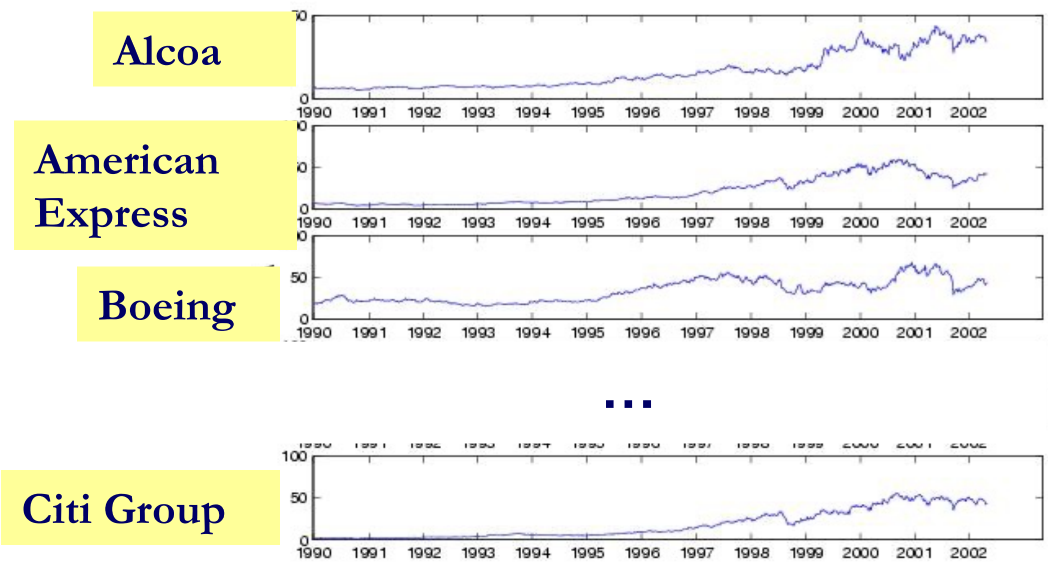
- Human would say
  - Pattern 1: along diagonal
  - Pattern 2: along vertical axis
- How to find these automatically?



Each point is the measurement at a time tick (total 550 points).

Motivation: (Q2) Find hidden variables

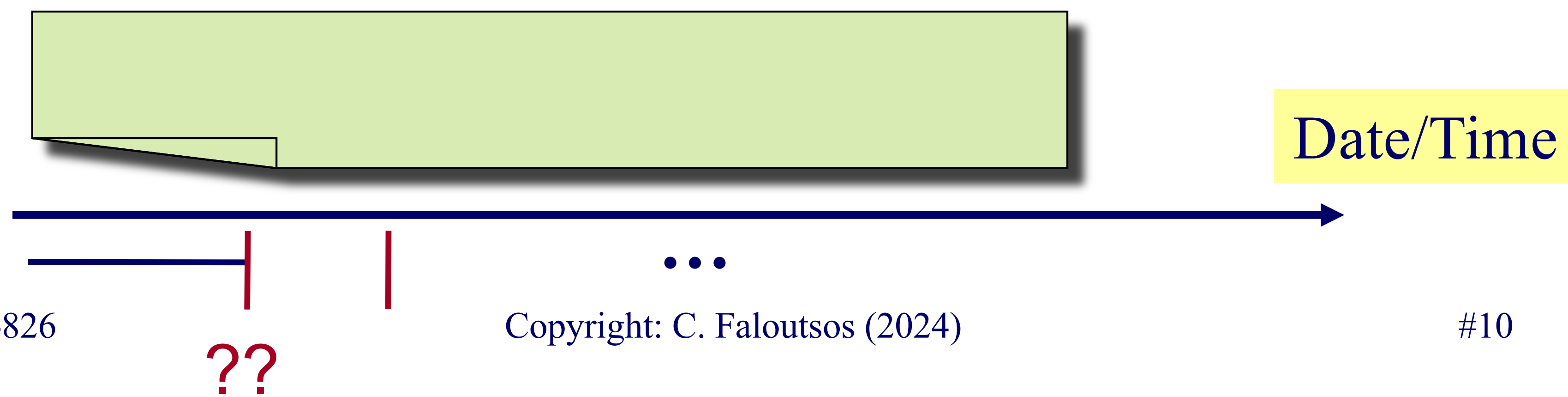
Stock prices



Hidden variables (= 'topics' = concepts)

(Q3): Topic discovery on text streams

- Data: CNN headline news (Jan.-Jun. 1998)
- Documents of 10 topics in one single text stream
  - FIND: the document boundaries
  - AND: the terms of each topic



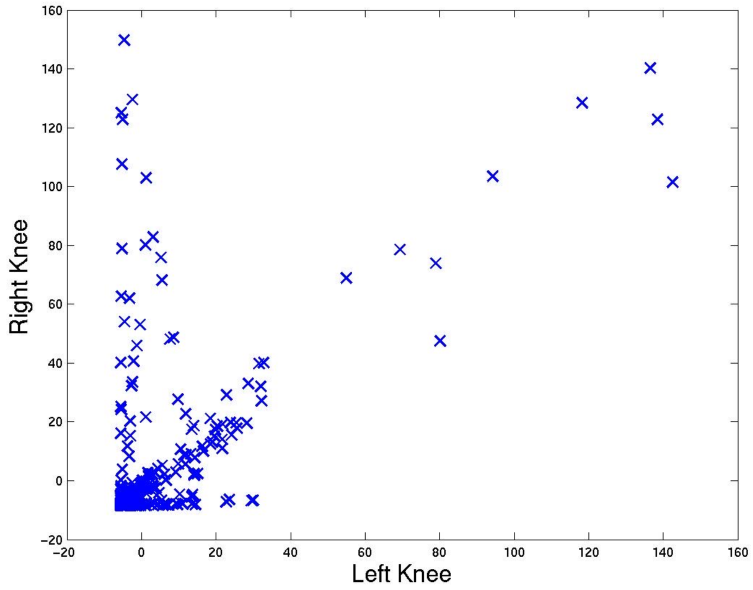
Carnegie Mellon

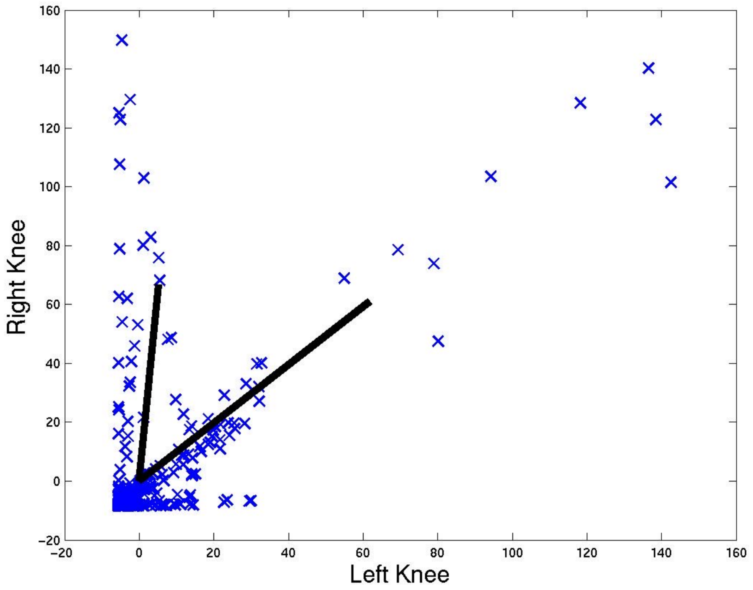
Outline

- Motivation
- Formulation
- PCA and ICA
- Example applications
- Conclusion

Carnegie Mellon

Formulation: Finding patterns

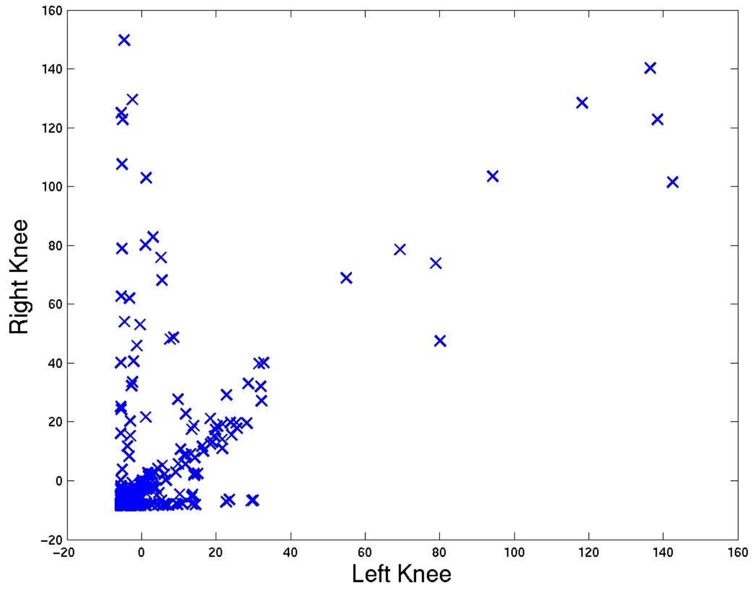


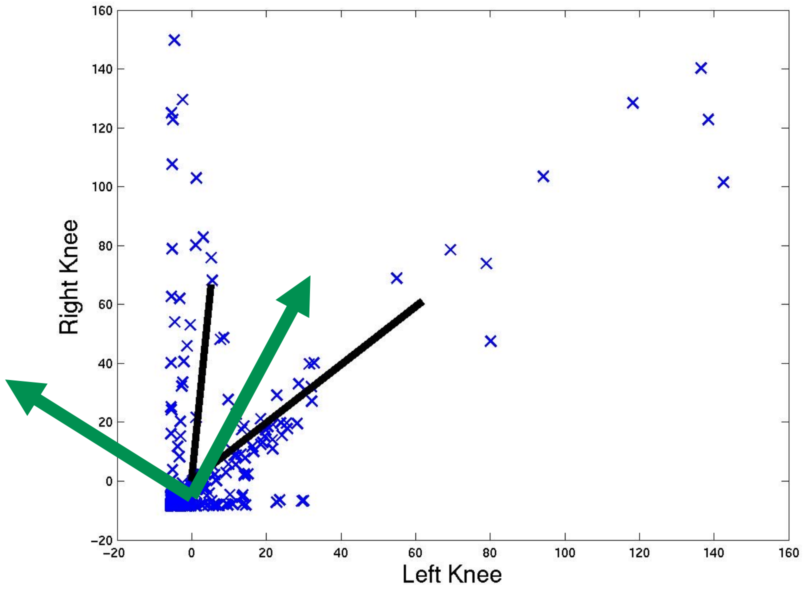


Given n data points, each with m attributes.

Find patterns that describe data properties the best.

Formulation: Finding patterns

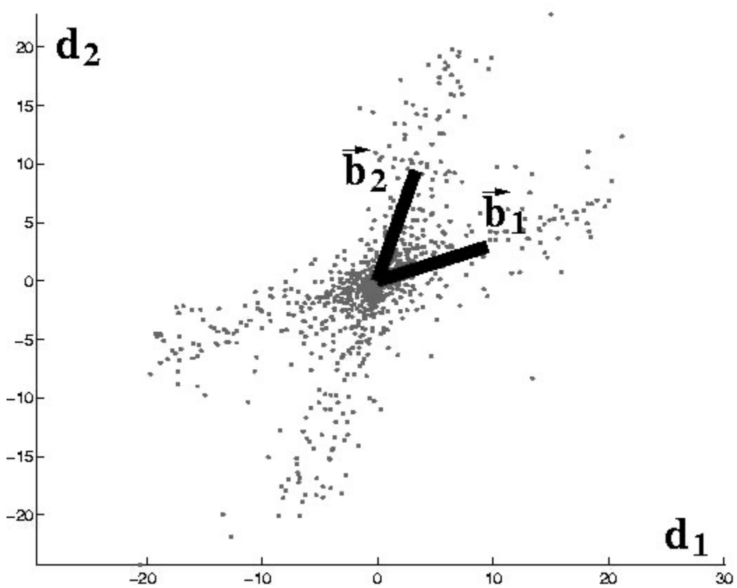


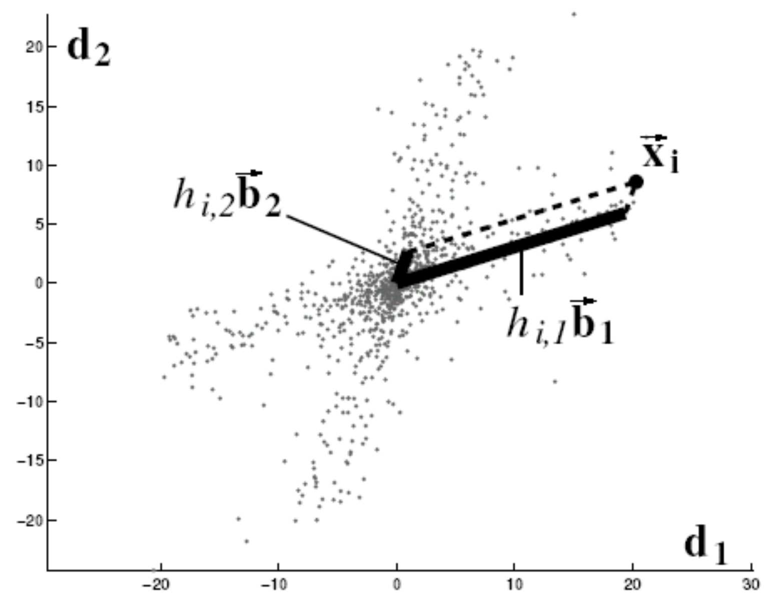


Given n data points,
each with m attributes.

SVD/PCA: ORTHOGONAL vectors

Linear representation

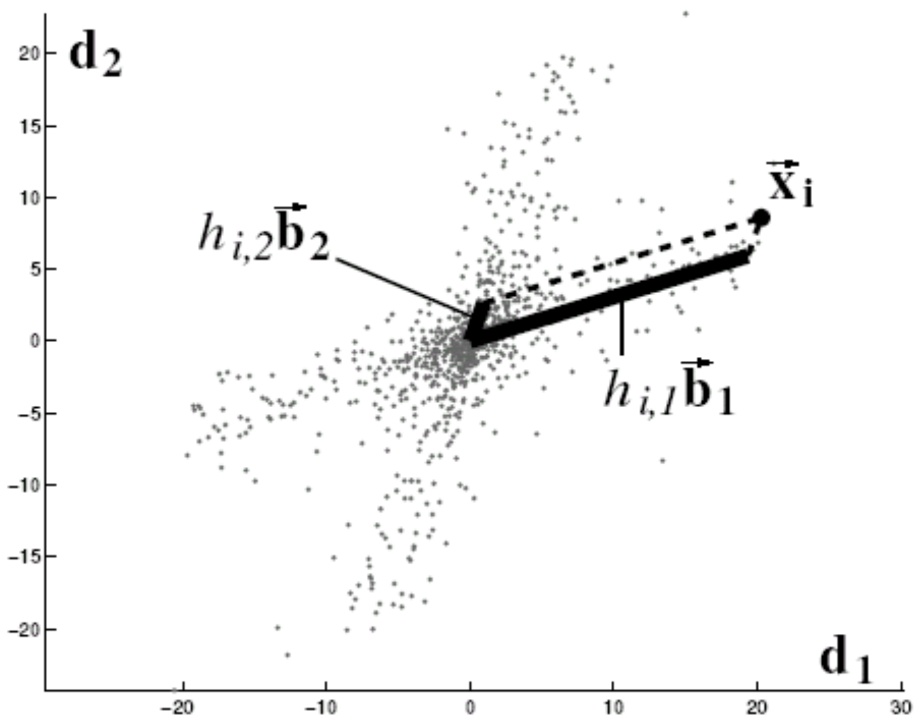




- Find vectors that describe the data set the best.
- Each point: linear combination of the vectors (patterns):

$$\vec{\mathbf{x}}_{\mathbf{i}} = h_{i,1}\vec{\mathbf{b}}_{\mathbf{1}} + h_{i,2}\vec{\mathbf{b}}_2$$

Patterns as data "vocabulary"



(a) ICA representation of $\vec{\mathbf{x}}_i$

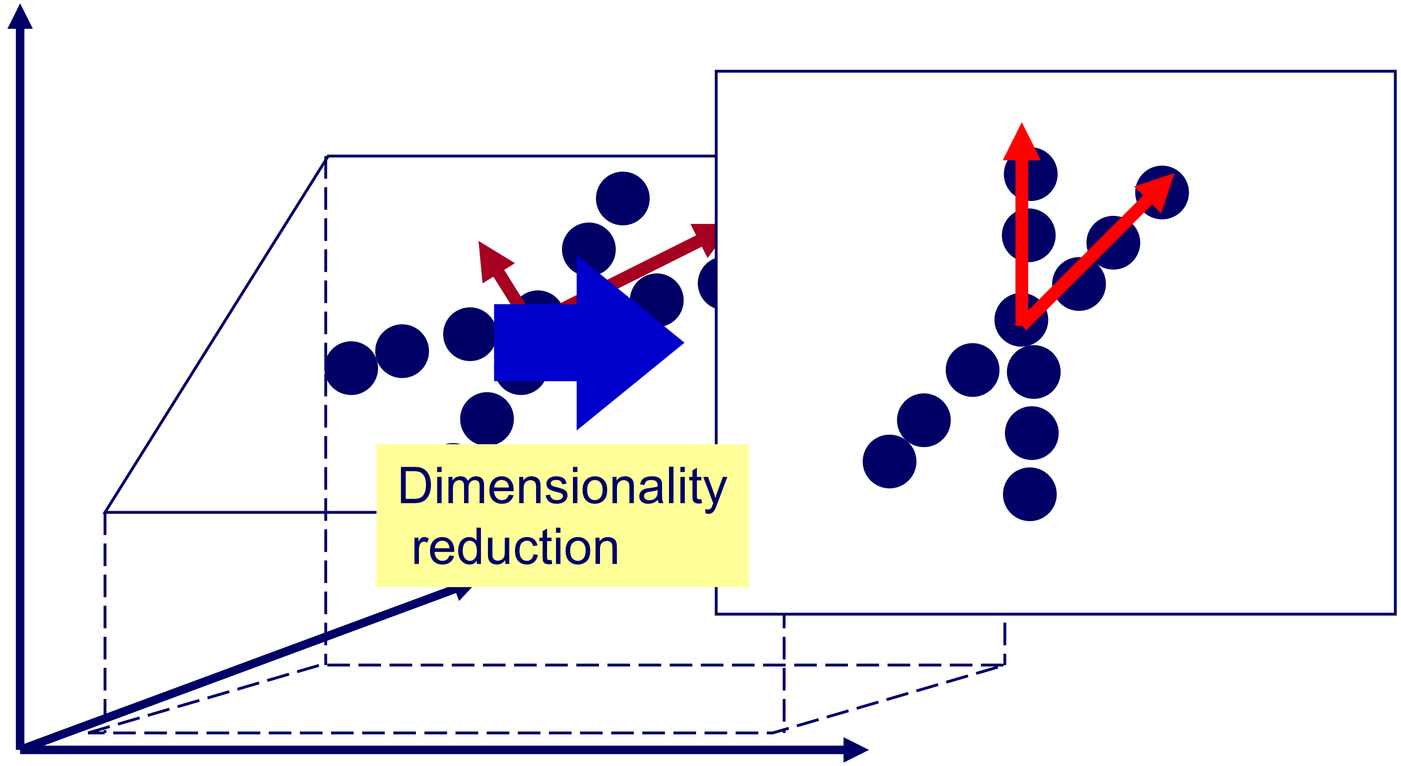
$$\vec{\mathbf{x}}_i = h_{i,1}\vec{\mathbf{b}}_1 + h_{i,2}\vec{\mathbf{b}}_2$$

Good pattern
≈ sparse coding

b_1 alone, can describe x_i.

Carnegie Mellon

PCA: first step of ICA



PCA finds the hyperplane.

ICA finds the correct patterns.

Software

- Open source software: ‘fastICA’ http://research.ics.aalto.fi/ica/fastica/

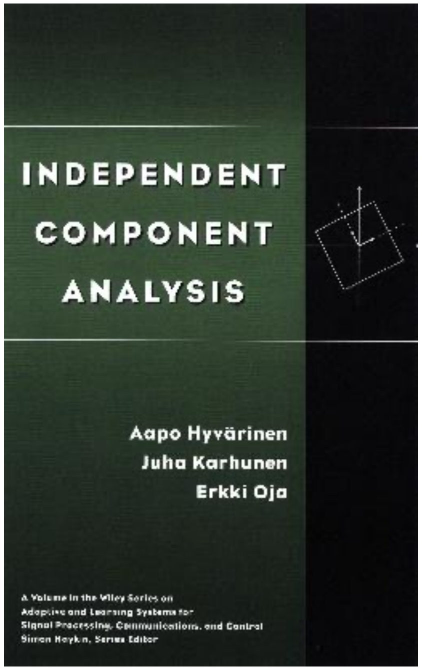
- Or ‘autosplit’:

www.cs.cmu.edu/~jypan/software/autosplit_cmu.tar.gz

Carnegie Mellon

References

- Aapo Hyvärinen, Juha Karhunen, Erkki Oja: *Independent Component Analysis*, John Wiley & Sons, 2001



Carnegie Mellon

Outline

- Motivation
- Formulation
- PCA and ICA
- Example applications
  - Hidden variables in stock prices
  - Find topics in documents
- Conclusion

Carnegie Mellon

Motivation: Find hidden variables

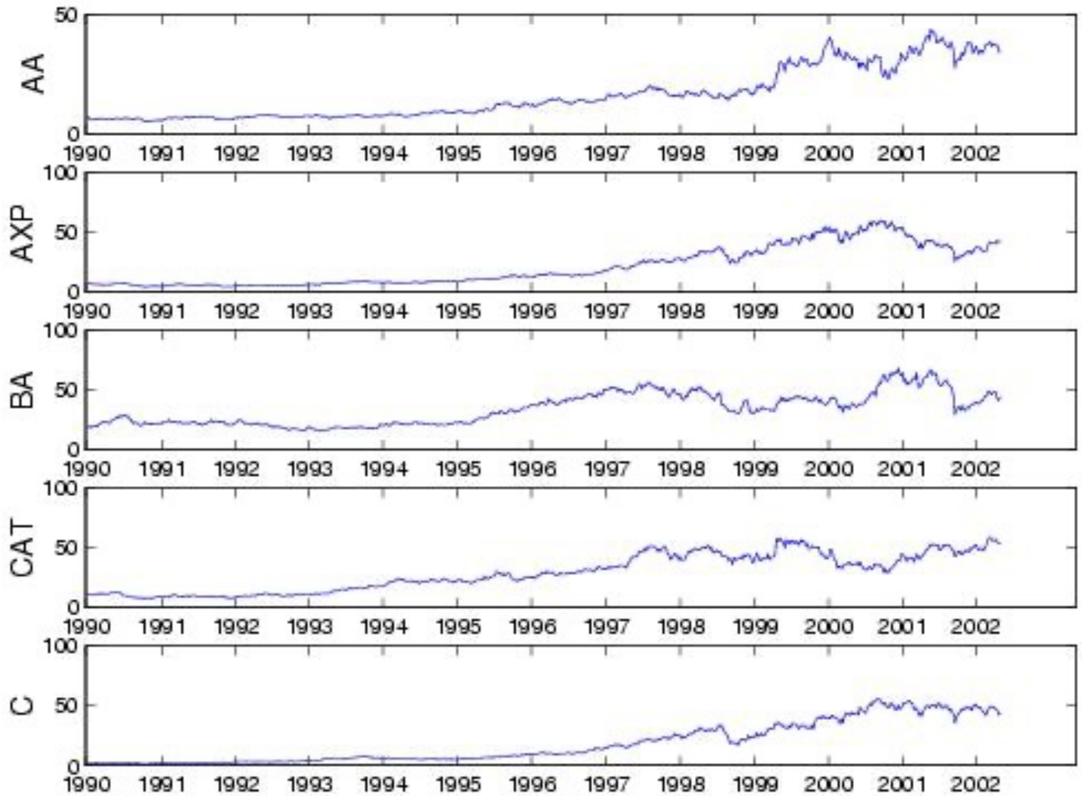
Alcoa

American Express

Boeing

Caterpillar

Citi Group



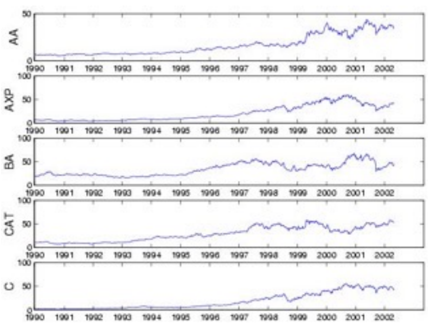
Dow Jones Industrial Average

Find common hidden variables, and weights.

ICA: Like SVD, but sparse U

1^{st} behavior

t_1 t_n



AA
AXP

$\sim$

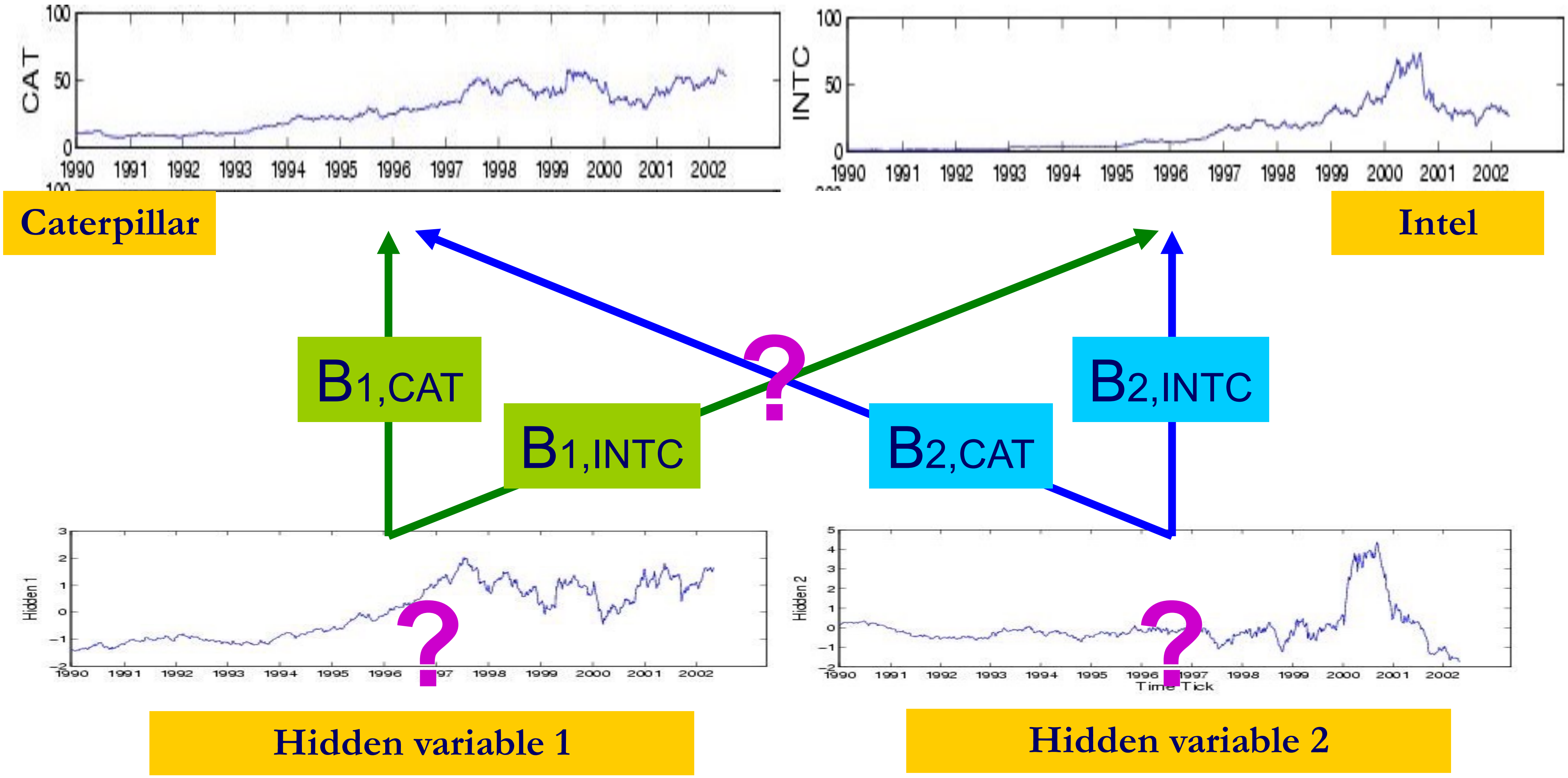
2^{nd} behavior

Participation weight of row *i* to behavior *j*

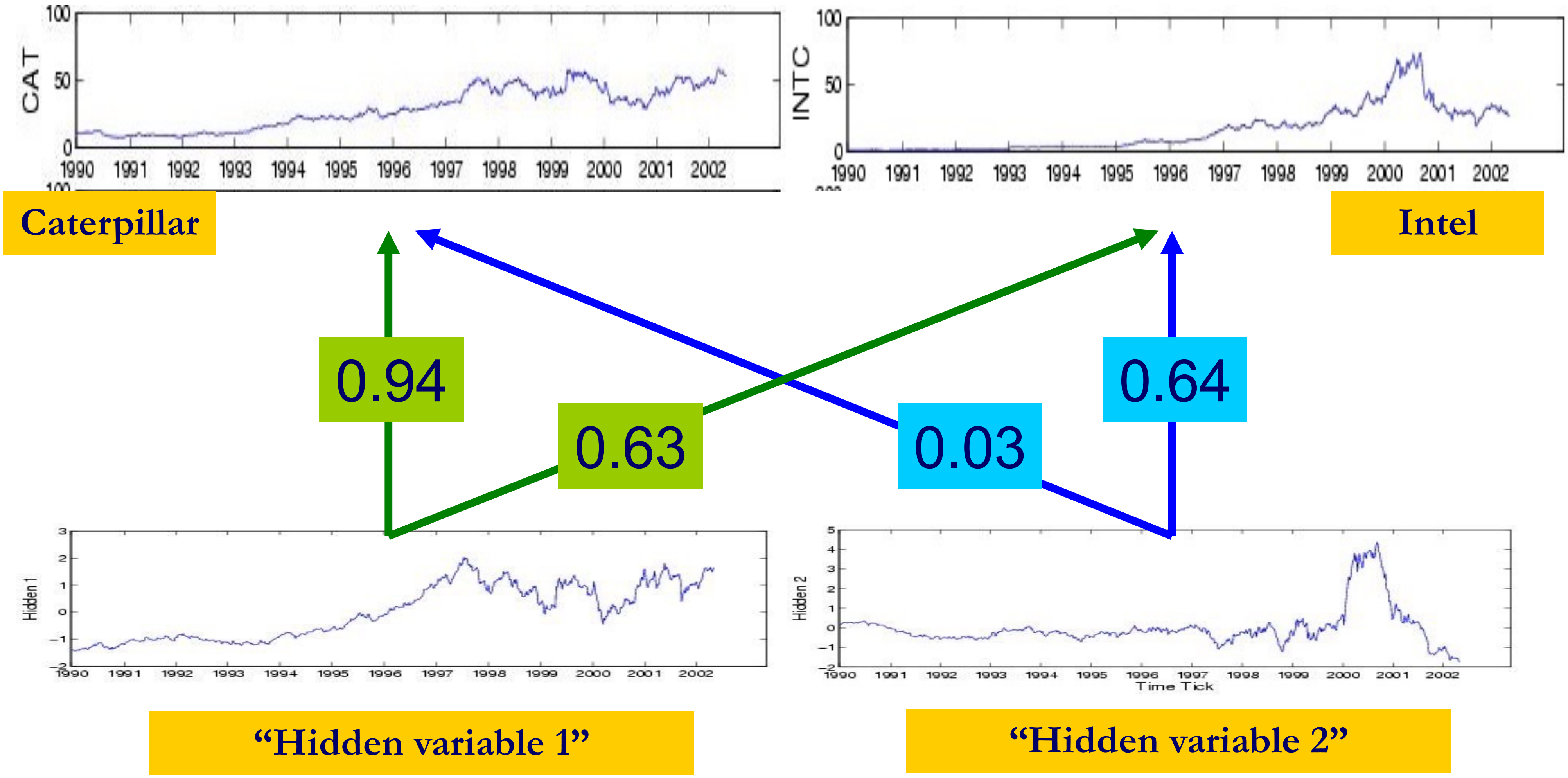
U $\cancel{\Sigma}$ $\mathbf{V}^T$

Motivation: Find hidden variables

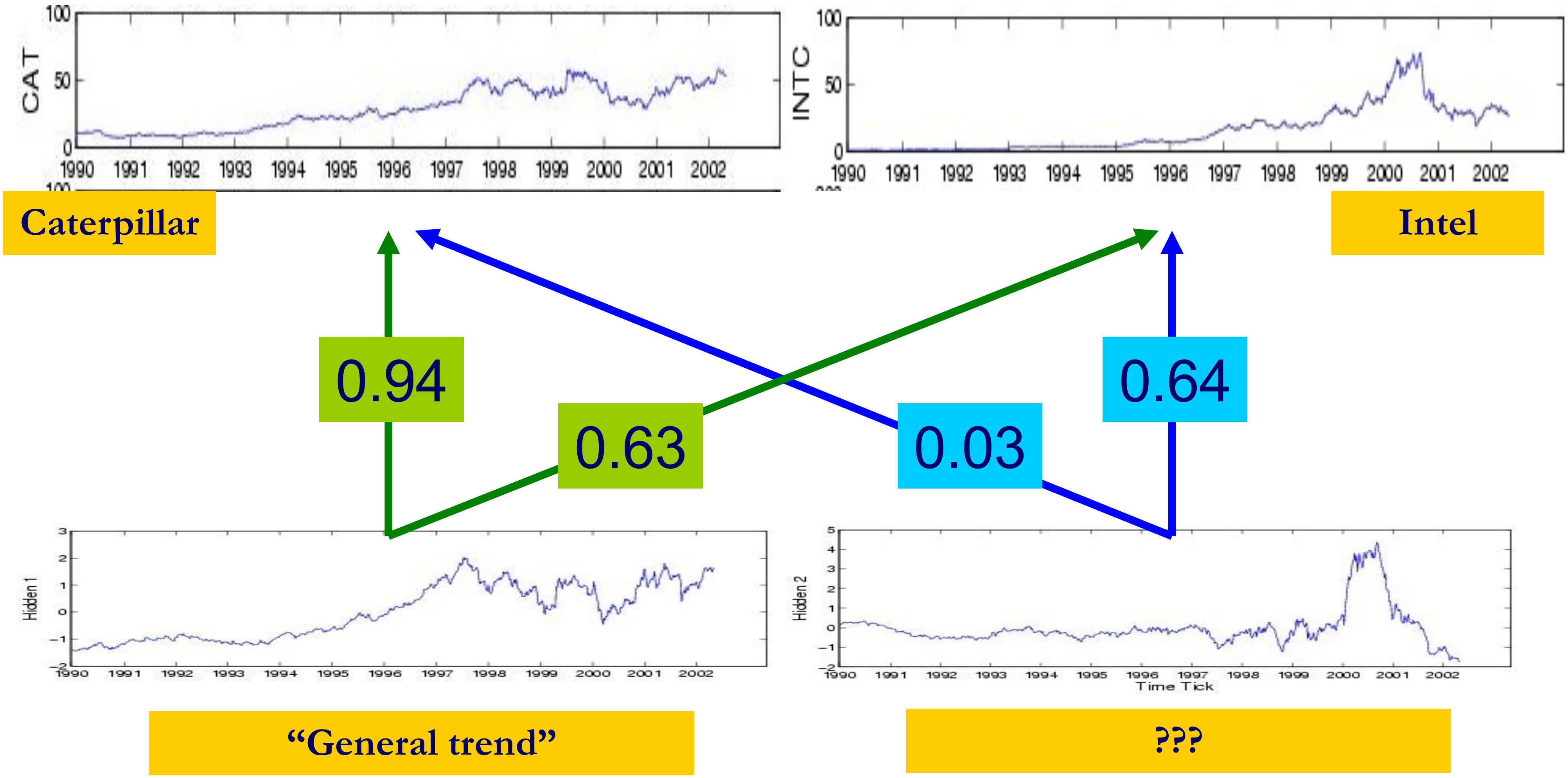
= behaviors



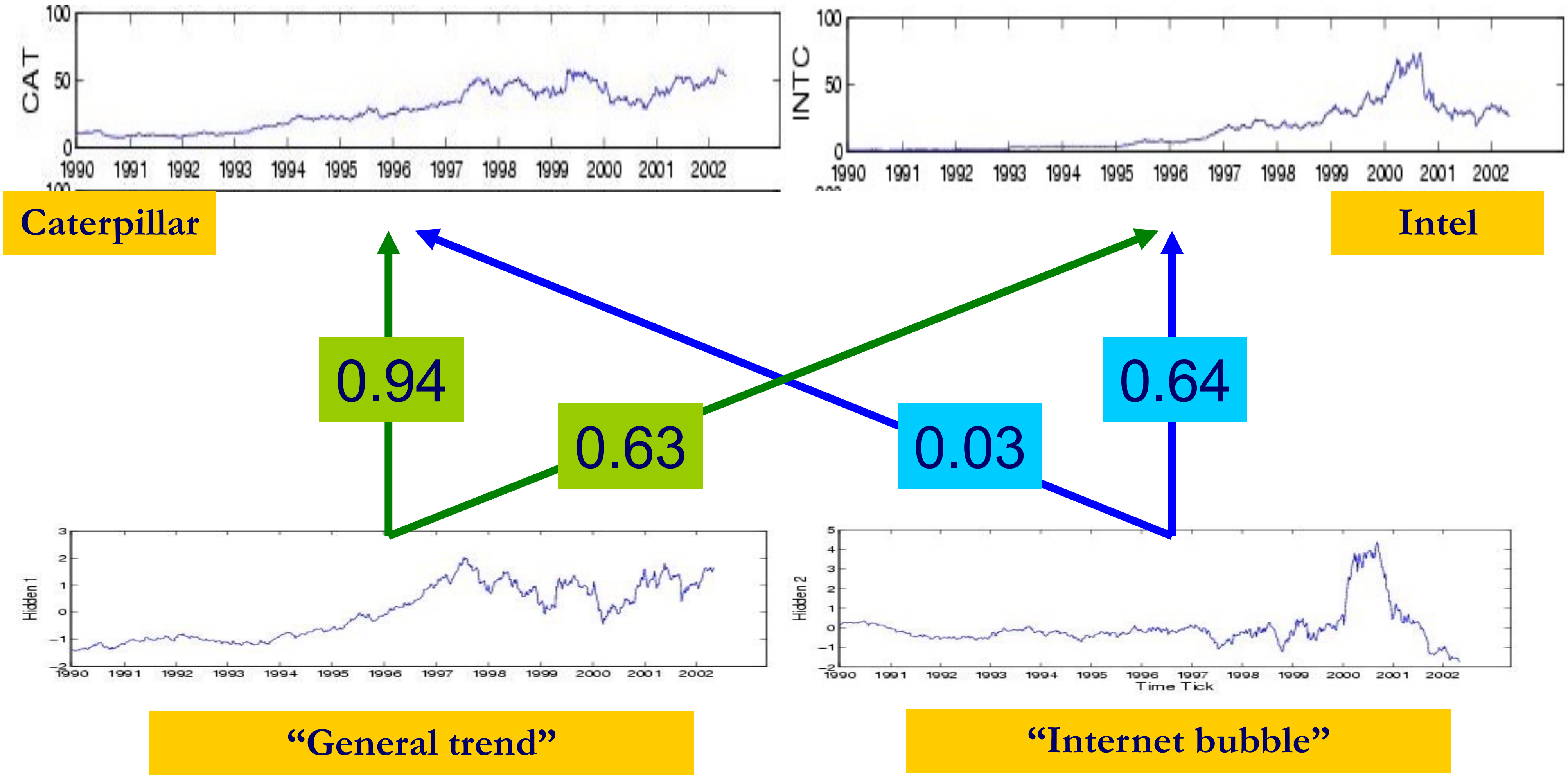
Motivation: Find hidden variables



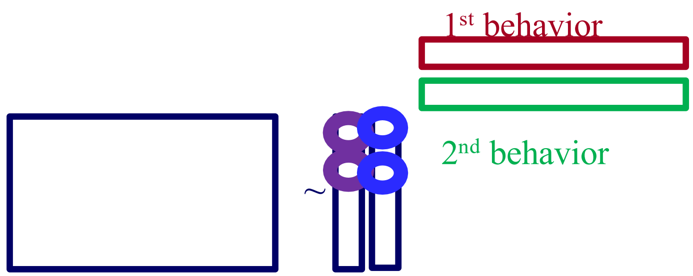
Motivation: Find hidden variables



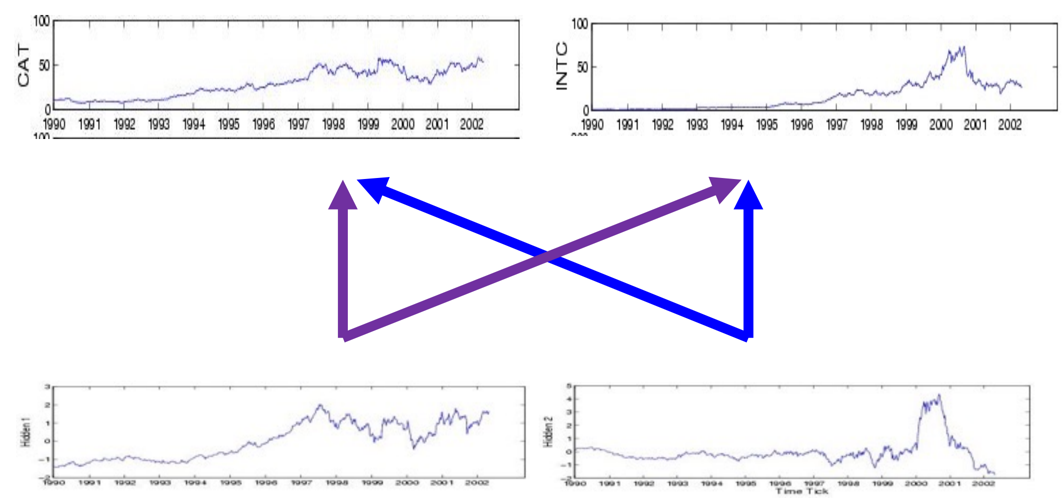
Motivation:
Find hidden variables



ICA: Like SVD, but sparse

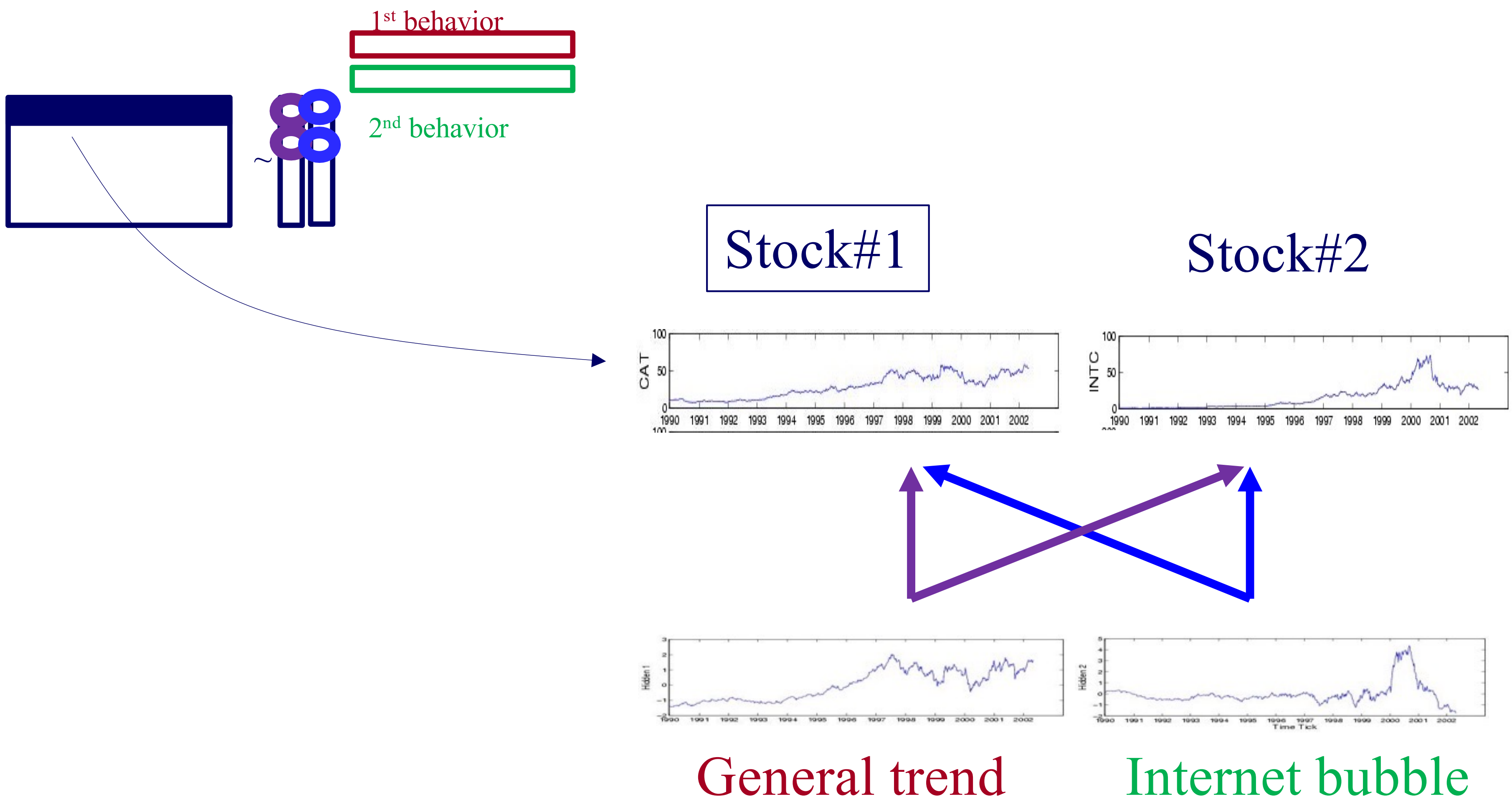


Stock#1 Stock#2

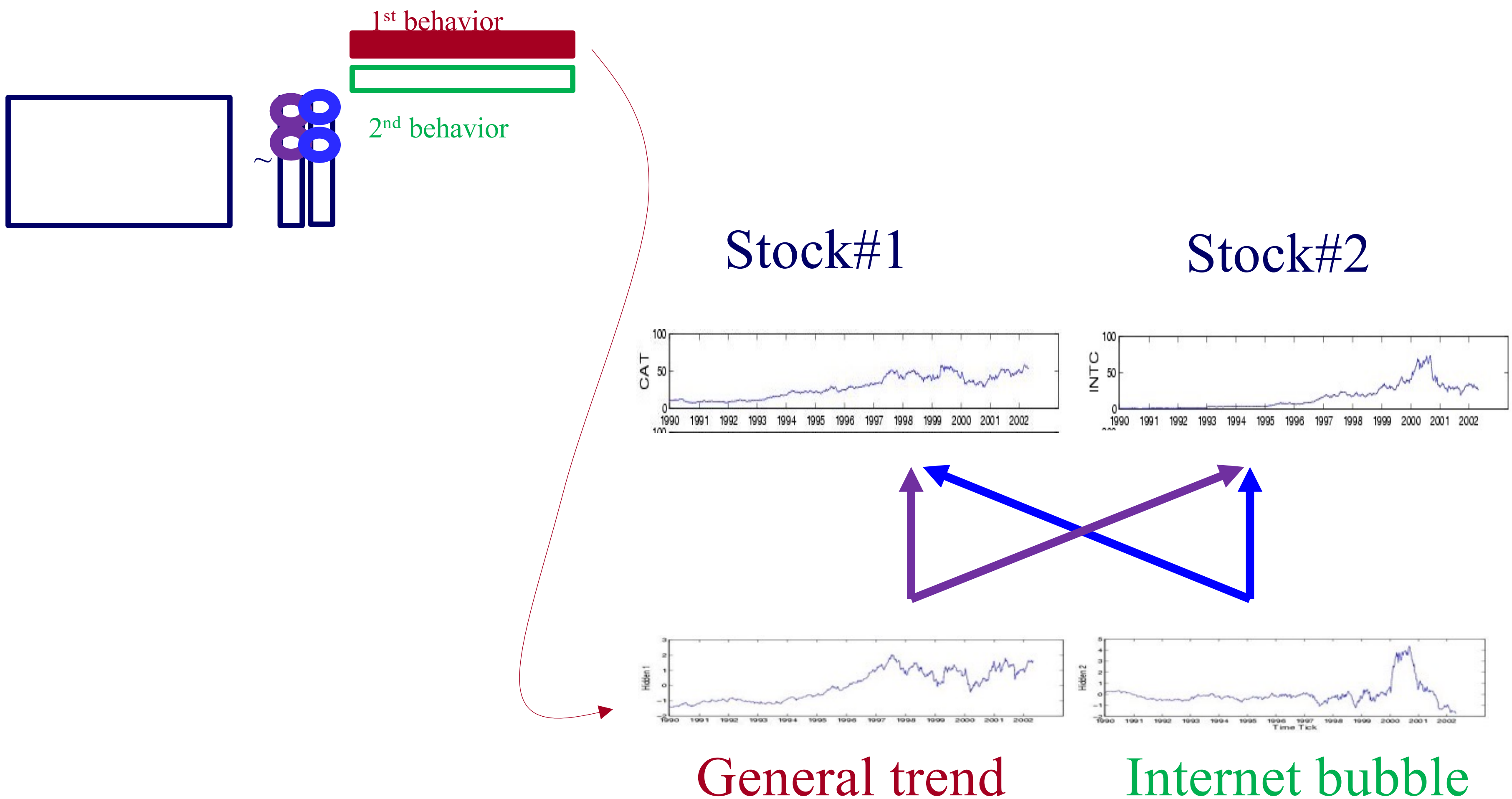


General trend Internet bubble

ICA: Like SVD, but sparse

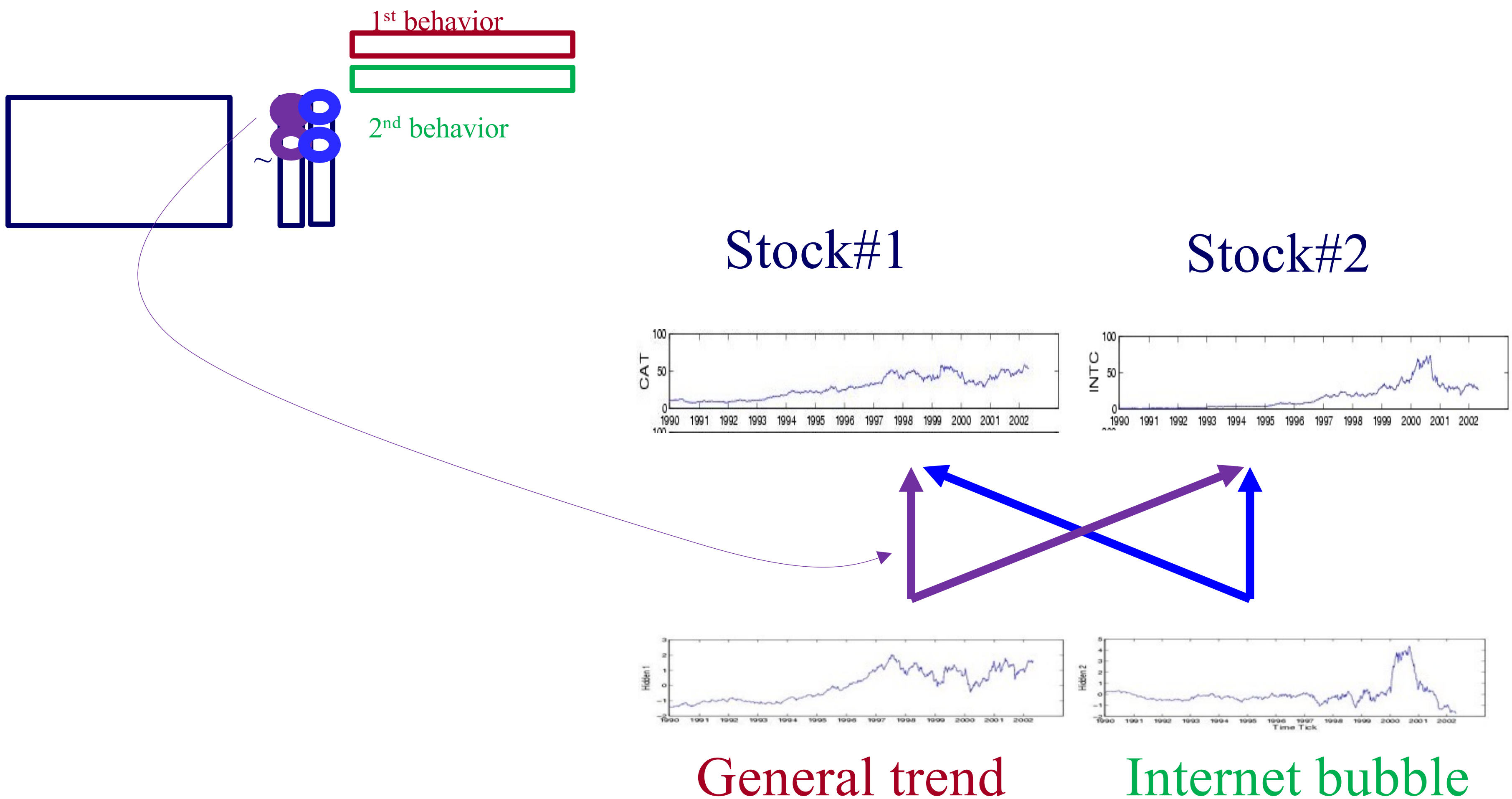


ICA: Like SVD, but sparse



Carnegie Mellon

ICA: Like SVD, but sparse



What else can ICA tell us?

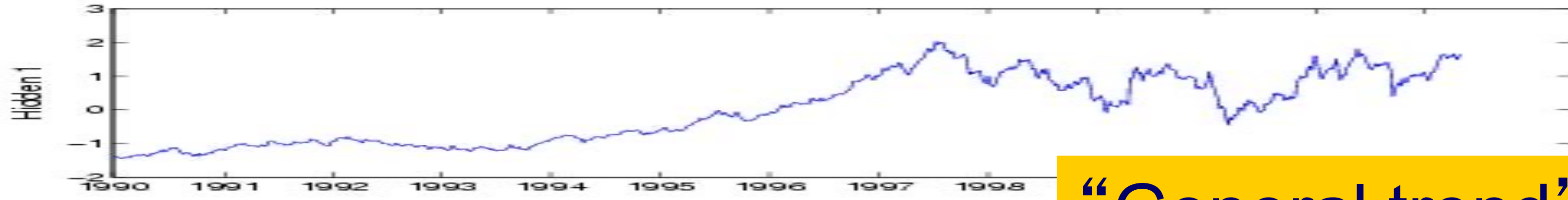
Carnegie Mellon

Companies related to hidden variable 1

$B_{1,j}$			
Highest		Lowest	
Caterpillar	0.938512	AT&T	0.021885
Boeing	0.911120	WalMart	0.624570
MMM	0.906542	Intel	0.638010
Coca Cola	0.903858	Home Depot	0.647774
Du Pont	0.900317	Hewlett-Packard	0.658768

All companies are affected by the "general trend" variable (with weights 0.6~0.9), except AT&T.

General trend (and outlier)



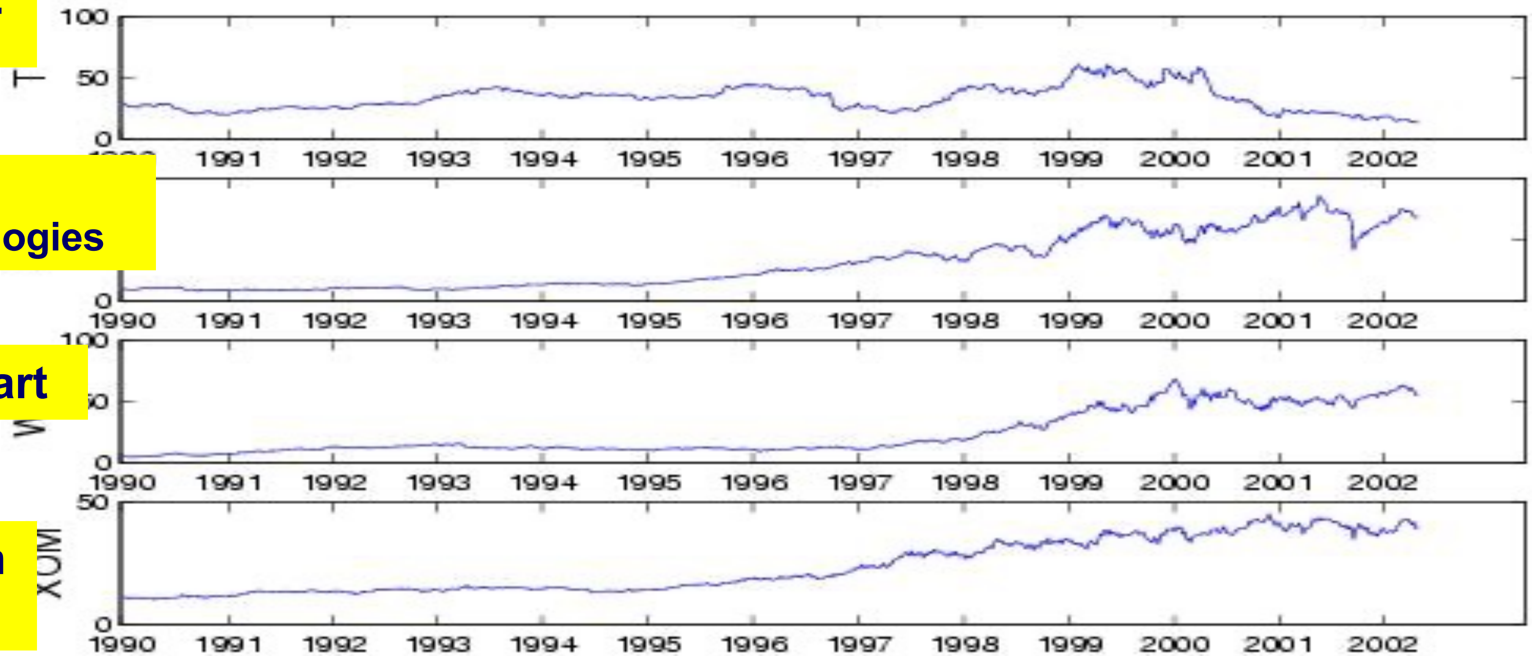
"General trend"

AT&T

United Technologies

Walmart

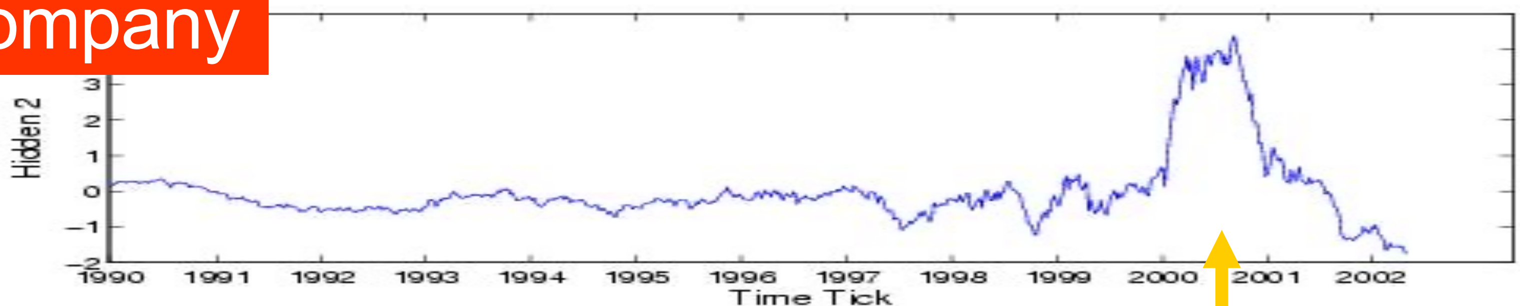
Exxon Mobil



Companies related to hidden variable 2

$B_{2,j}$			
Highest		Lowest	
Intel	0.641102	Philip Morris	-0.194843
Hewlett-Packard	0.621159	International Paper	-0.089569
GE	0.509164	Caterpillar	0.031678
American Express	0.504871	Procter and Gamble	0.109576
Disney	0.490529	Du Pont	0.133337

Tech company



2000-2001 “Internet bubble”

Companies related to hidden variable 2

$B_{2,j}$			
Highest		Lowest	
Intel	0.641102	Philip Morris	-0.194843
Hewlett-Packard	0.621159	International Paper	-0.089569
GE	0.509164	Caterpillar	0.031678
American Express	0.504871	Procter and Gamble	0.109576
Disney	0.490529	Du Pont	0.133337

Tech company

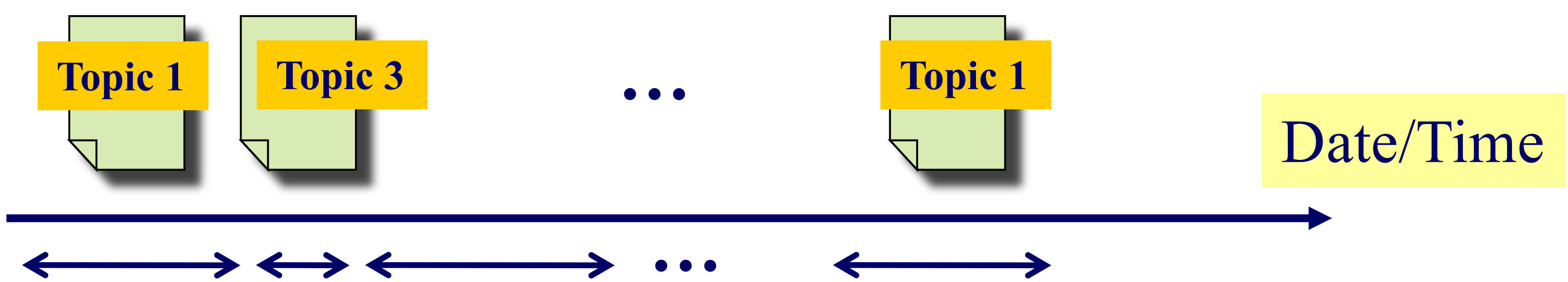
Companies affected by the “internet bubble” variable (with weights 0.5~0.6) are tech-related. Other companies are un-related (weights < 0.15).

Outline

- Motivation
- Formulation
- PCA and ICA
- Example applications
  - Hidden variables in stock prices
  - Find topics in documents
- Conclusion

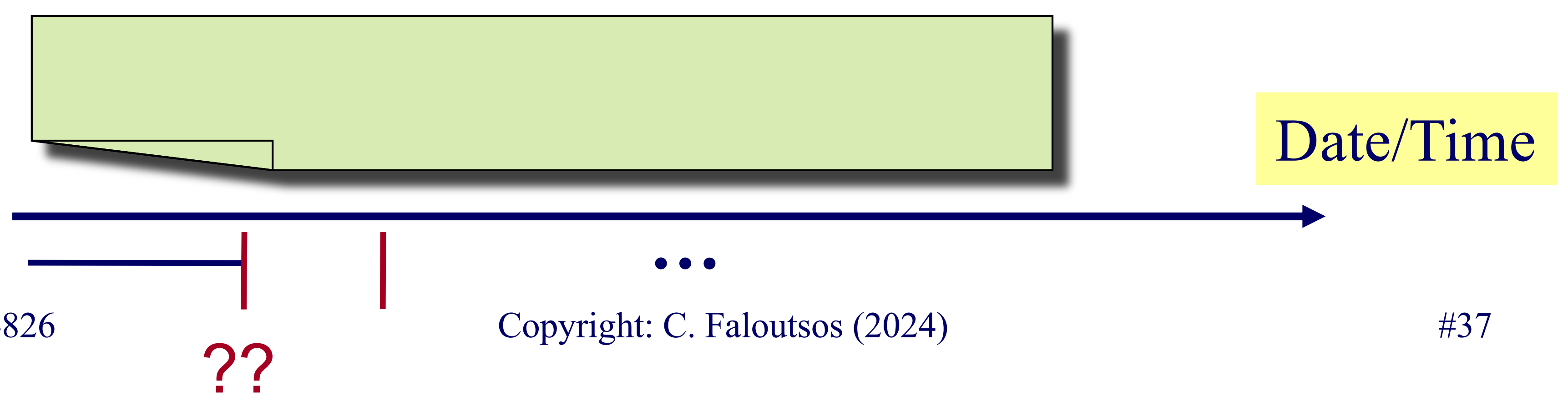
Topic discovery on text streams

- Data: CNN headline news (Jan.-Jun. 1998)
- Documents of 10 topics in one single text stream
  - Documents are sorted by date/time
  - Subsequent documents may have different topics



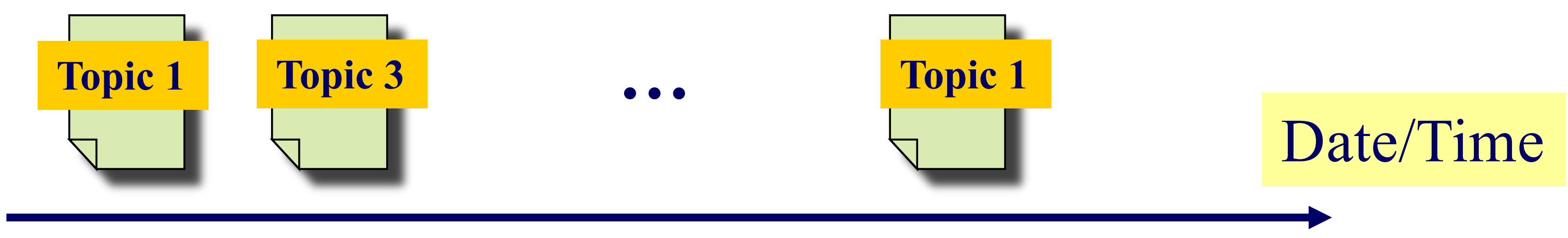
Topic discovery on text streams

- Data: CNN headline news (Jan.-Jun. 1998)
- Documents of 10 topics in one single text stream
  - FIND: the document boundaries
  - AND: the terms of each topic

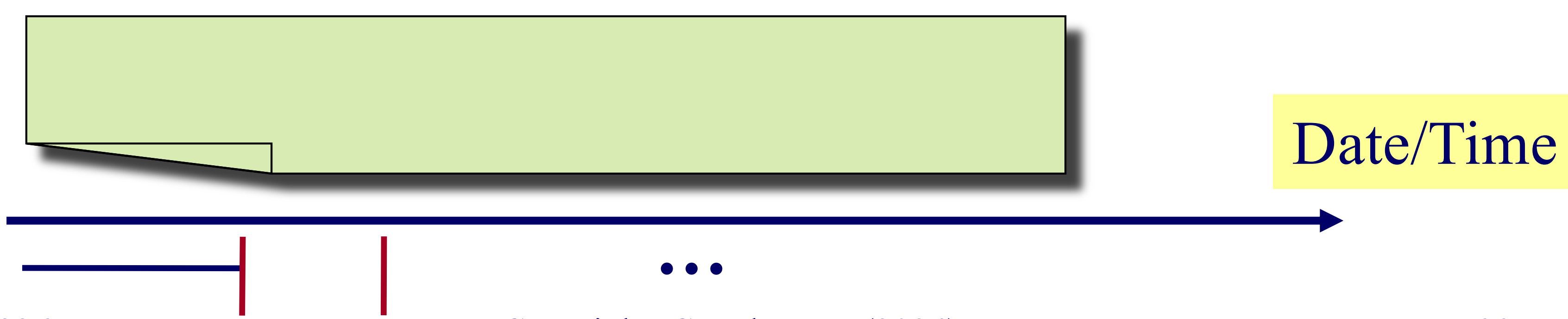


Topic discovery on text streams

- Known: number of topics = 10
- Unknown: (1) topic of each document (2) topic description

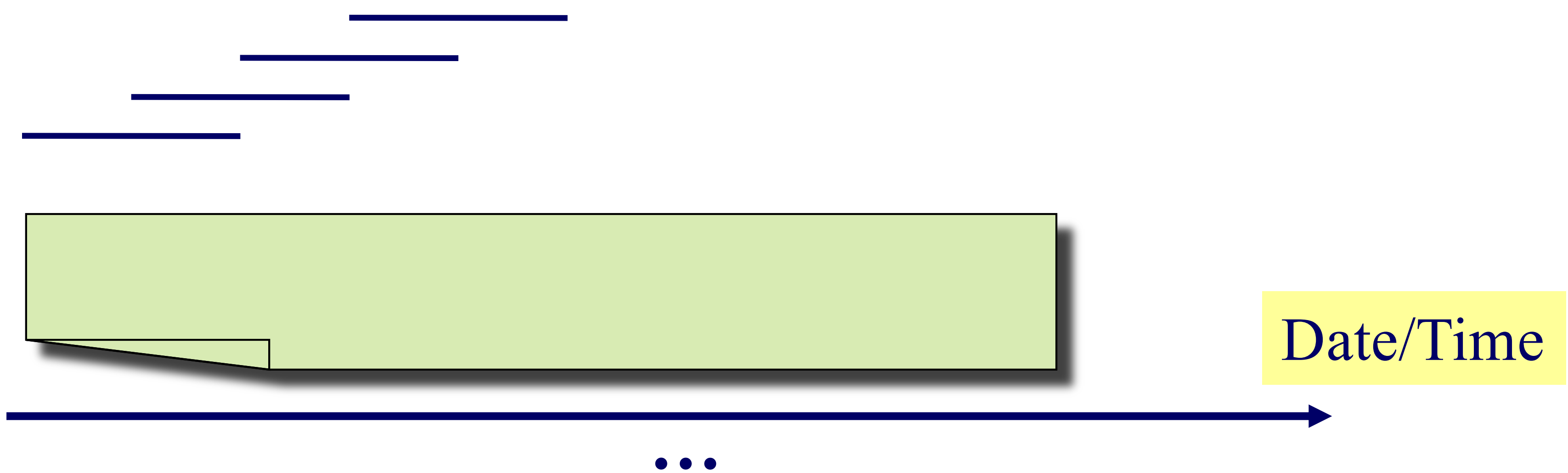


How to proceed?

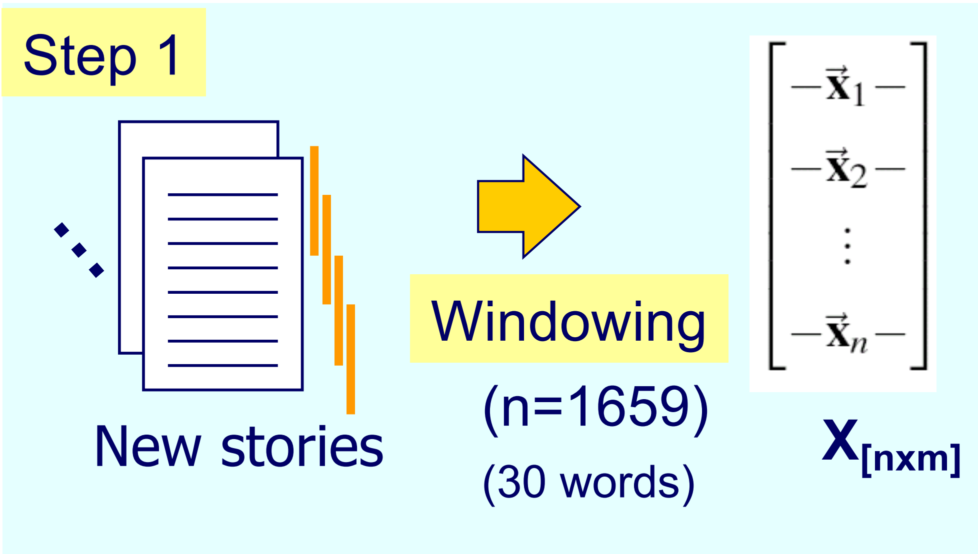


How to proceed?

- A: Sliding windows



Topic discovery in documents

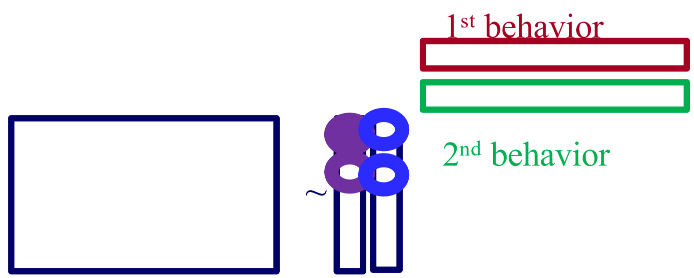


aaron

zoo

x_i = [1, 5, …, 0]

m=3887 (dictionary size)



Carnegie Mellon

Step 3: Interpret the patterns

$$\begin{bmatrix} -\mathbf{b}_1' - \\ -\mathbf{b}_2' - \\ \vdots \\ -\mathbf{b}_{10}' - \end{bmatrix}$$

aaron animal zoo

b’ $_i$ = [0, 0.7, …, 0.6]

m=3887 (dictionary size)

Top words : “animal”, “zoo”, …

A hidden topic!

Topics found

ID	Sorted word list				
A	Mckinne	Sergeant	sexual	Major	Armi
B	bomb	Rudolph	Clinic	Atlanta	Birmingham
C	Winfrei	Beef	Texa	Oprah	Cattl
D	Viagra	Drug	Impot	Pill	Doctor
E	Zamora	Graham	Kill	Former	Jone
F	Medal	Olymp	Gold	Women	Game
G	Pope	Cube	Castro	Cuban	Visit
H	Asia	Economi	Japan	Econom	Asian
I	Super	Bowl	Game	Team	Re
J	Peopl	Tornado	Florida	Re	bomb

Step 3: Evaluate the patterns

ID	True Topic
1	Sgt. Gene Mckinney is on trial for alleged sexual misconduct
2	A bomb explodes in a Birmingham, AL abortion clinic
3	The Cattle Industry in Texas sues Oprah Winfrey for defaming beef
4	New impotency drug Viagra is approved for use
5	Diane Zamora is convicted of helping to murder her lover's girlfriend

ID	Sorted word list				
A	mckinne	sergeant	sexual	major	armi
B	bomb	rudolph	clinic	atlanta	birmingham
C	winfrei	beef	texa	oprah	cattl
D	viagra	drug	Impot	pill	doctor
E	zamora	graham	kill	former	jone

AutoSplit finds correct topics.

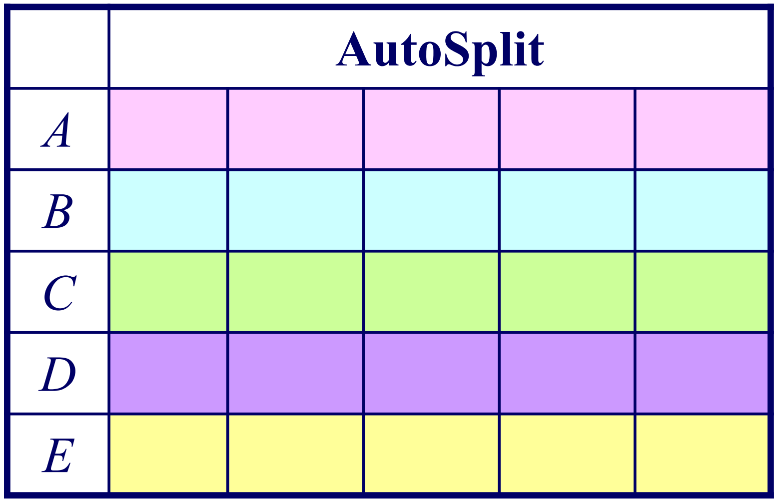
Step 3: Evaluate the patterns

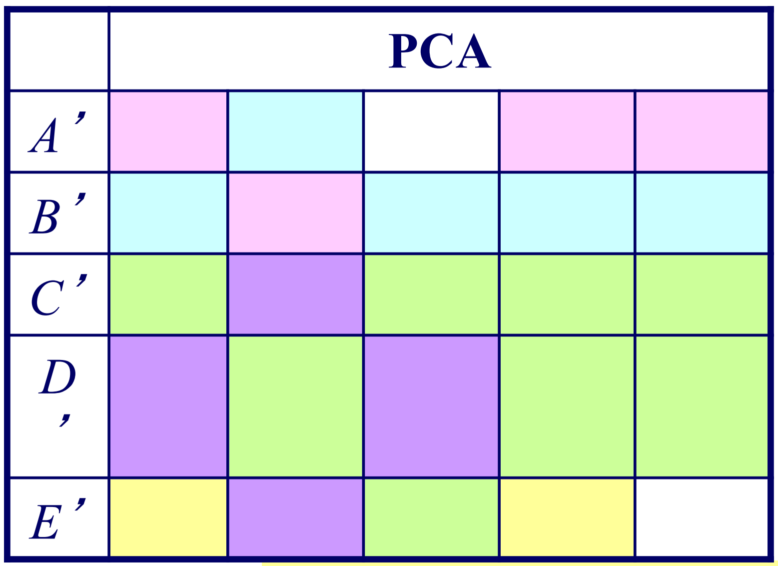
ID	AutoSplit				
A	mckinne	sergeant	sexual	major	armi
B	bomb	rudolph	clinic	atlanta	birmingham
C	winfrei	beef	texa	oprah	cattl
D	viagra	drug	Impot	pill	doctor
E	zamora	graham	kill	former	jone

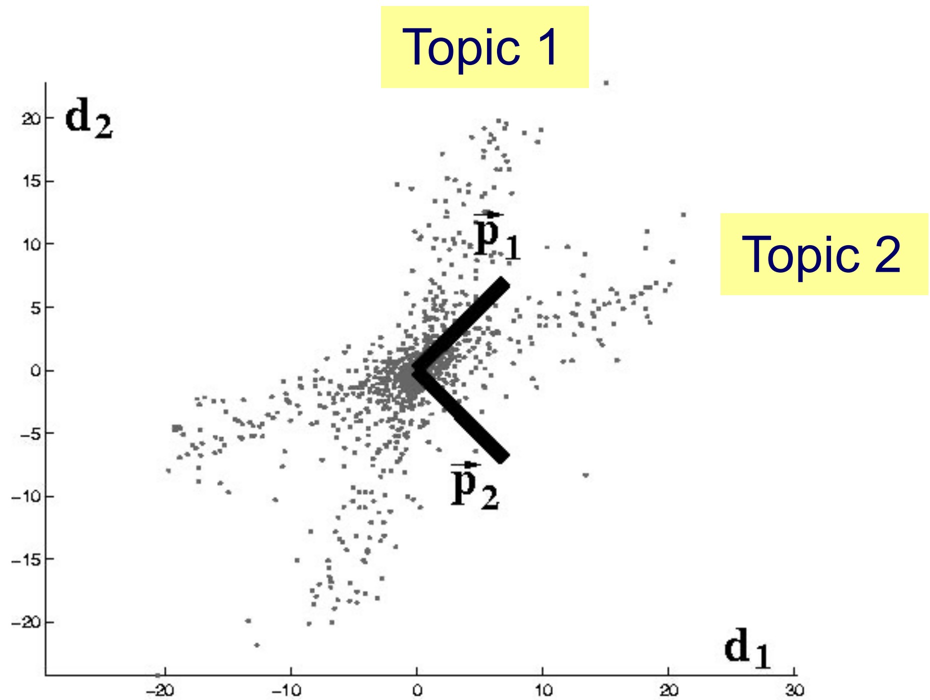
ID	PCA				
A'	mckinne	bomb	women	sexual	sergeant
B'	bomb	mckinne	rudolph	clinic	atlanta
C'	winfrei	viagra	texa	beef	oprah
D'	viagra	winfrei	drug	texa	beef
E'	zamora	viagra	winfrei	graham	olymp

AutoSplit's topics are better than PCA.

Step 3: Evaluate the patterns



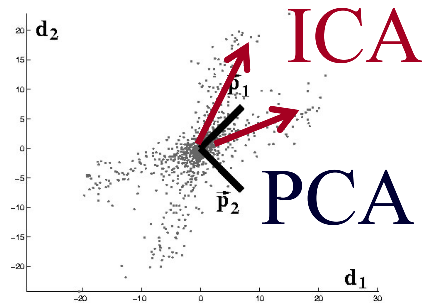




PCA vectors mix the topics.

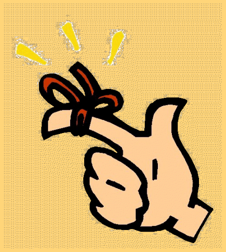
AutoSplit's topics are better than PCA.

Carnegie Mellon



Conclusion

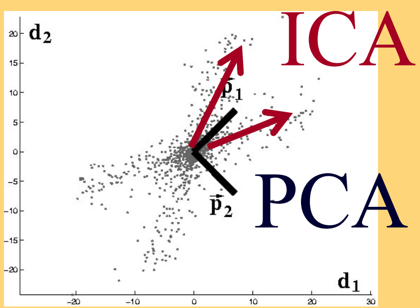
- ICA: more flexible than PCA in finding patterns.
- Many applications
  - Find hidden variables in time series (e.g., stock prices)
  - Blind source separation
- Rule of thumb: plot after PCA;
  - if ‘chicken-feet’, try ICA

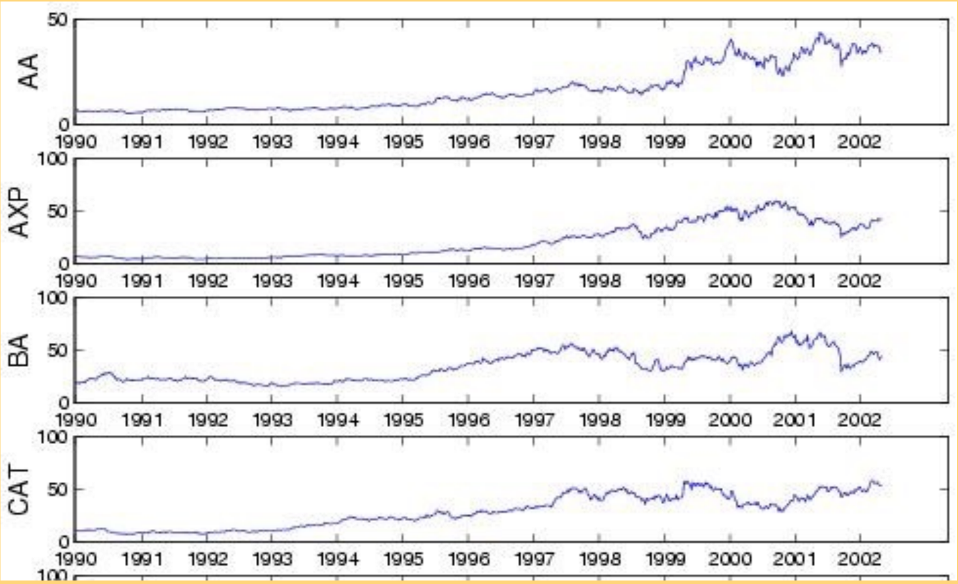


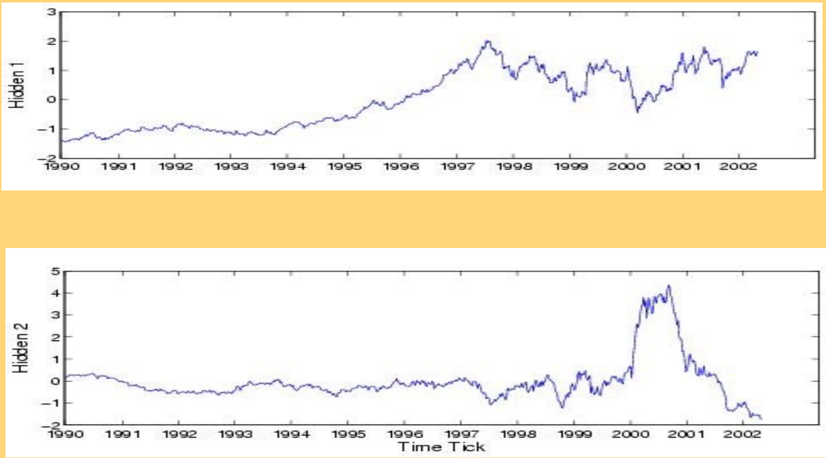
Answer

Q: how to extract **sparse** hidden/latent variables?

A: ~~SVD~~ ICA





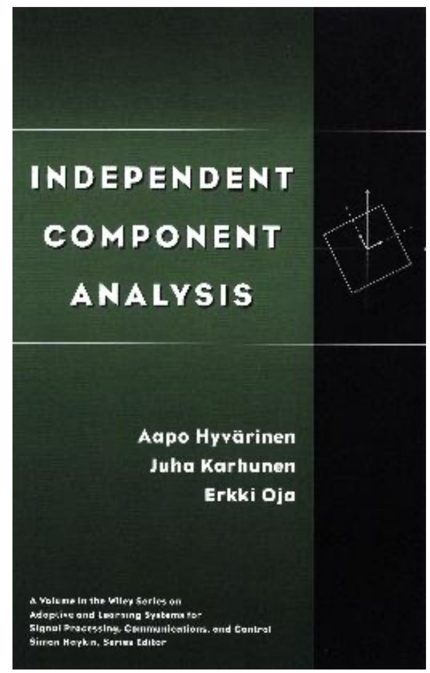


SVD, ICA: special cases of Artificial Neural Networks

- Aapo Hyvärinen, Juha Karhunen, Erkki Oja: *Independent Component Analysis*

A.N.N. autoencoder with
- Linear transfer function
- 1 hidden layer
- L2 penalty

= SVD
(page 136, sec. 6.2.4)



Citation

- *AutoSplit: Fast and Scalable Discovery of Hidden Variables in Stream and Multimedia Databases*, **Jia-Yu Pan**, Hiroyuki Kitagawa, Christos Faloutsos and Masafumi Hamamoto

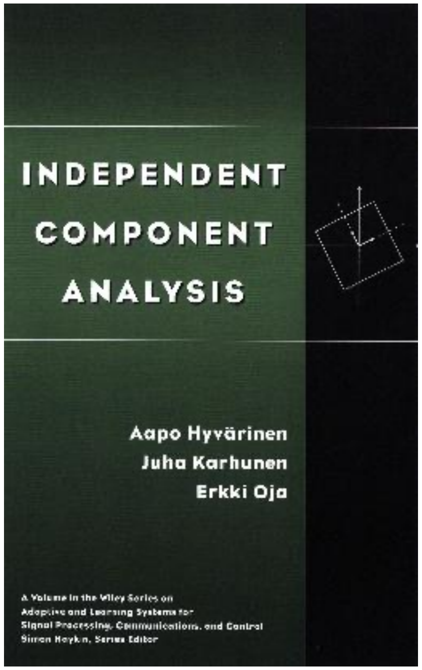
PAKDD 2004, Sydney, Australia

References

- Jia-Yu Pan, Andre Guilherme Ribeiro Balan, Eric P. Xing, Agma Juci Machado Traina, and Christos Faloutsos. *Automatic Mining of Fruit Fly Embryo Images*. *KDD,* 2006.
- Arnab Bhattacharya, Vebjorn Ljosa, Jia-Yu Pan, Mark R. Verardo, Hyungjeong Yang, Christos Faloutsos, and Ambuj K. Singh. *ViVo: Visual Vocabulary Construction for Mining Biomedical Images*. *ICDM,* 2005.
- Jia-Yu Pan, Hiroyuki Kitagawa, Christos Faloutsos, and Masafumi Hamamoto. *AutoSplit: Fast and Scalable Discovery of Hidden Variables in Stream and Multimedia Databases*. *PAKDD*, 2004.

References

- Aapo Hyvärinen, Juha Karhunen, Erkki Oja: *Independent Component Analysis*, John Wiley & Sons, 2001



Carnegie Mellon

Software

- Open source software: 'fastICA' http://research.ics.aalto.fi/ica/fastica/
- Also on scikit-learn: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.FastICA.html
- Or 'autosplit':

www.cs.cmu.edu/~jypan/software/autosplit_cmu.tar.gz