

CS 591.03

Introduction to Data Mining

Instructor: Abdullah Mueen

---

LECTURE 1: OVERVIEW OF DATA MINING

# John Snow and the Broad St. Pump



**John Snow** (15 March 1813 – 16 June 1858) was an **English physician** and a leader in the adoption of anaesthesia and medical hygiene. He is considered **one of the fathers of modern epidemiology**, in part because of his work in tracing the source of a cholera outbreak in Soho, London, in 1854.

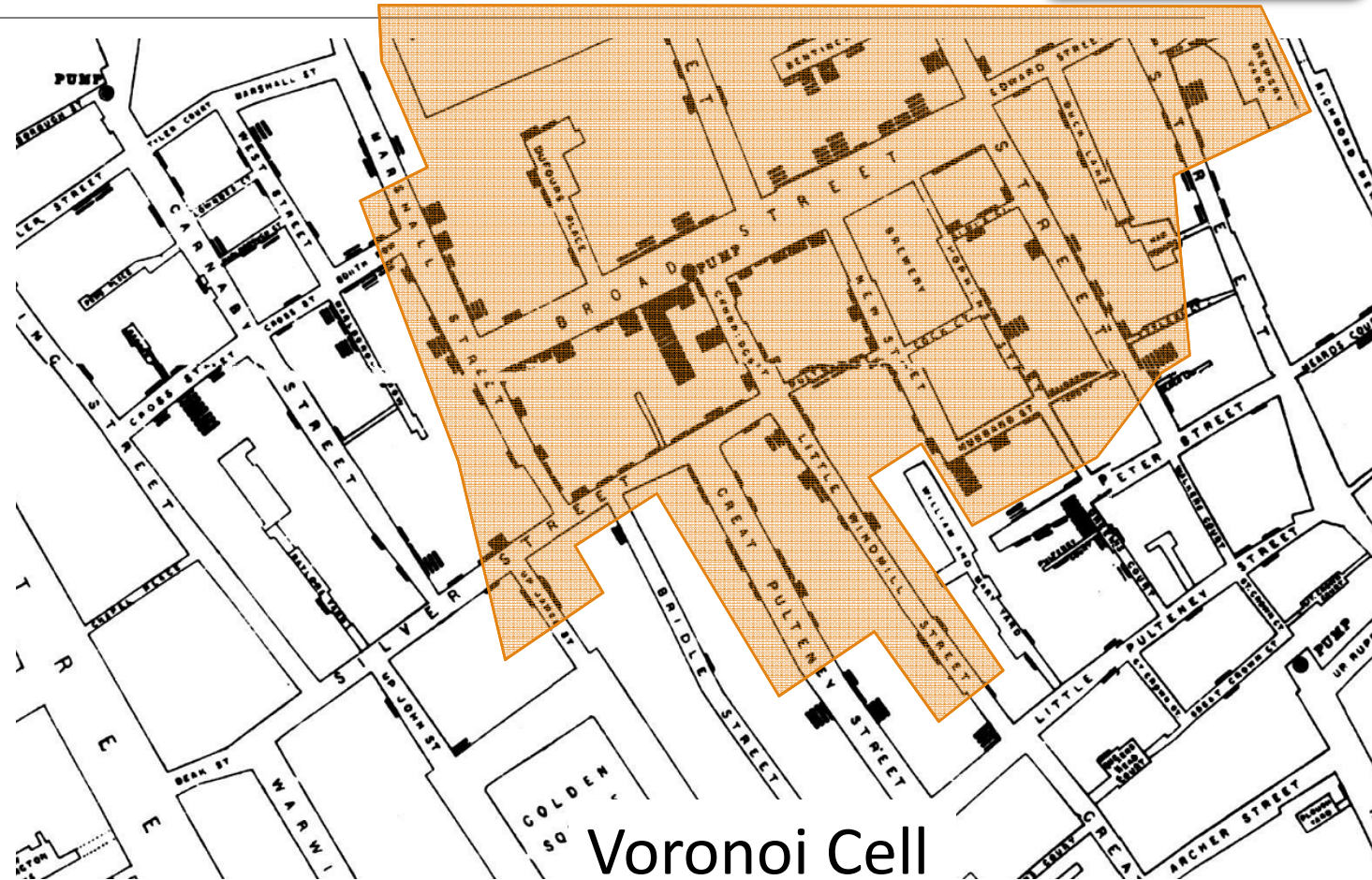
On 31 August 1854, after several other outbreaks had occurred elsewhere in the city, a major outbreak of cholera struck Soho. Over the next three days, 127 people on or near Broad Street died. In the next week, three quarters of the residents had fled the area. By 10 September, 500 people had died and the mortality rate was 12.8 percent in some parts of the city. By the end of the outbreak, 616 people had died.

He identified the source of the outbreak as the public water pump on Broad Street

# John Snow and the Broad St. Pump

Location of each death in the outbreak and locations of the pumps with the help of Rev. Henry Whitehead

Associate pumps with deaths to support the causal relationship



Voronoi Cell

# Components of Data Mining

---

**Data** (Images, Files, Tables, Charts)



**Tools** (Hadoop, Matlab, Algorithms)



**Objective** (Information integration, organization and scientific discovery)



**Data Scientist**





# Web Sensing

---

## Individual Sensing

### Data:

1. Search Query Logs: Mostly Tabular. Query, IP address/Account, Time, Link Clicked
2. Action Sequence: Every Click you make is being recorded across devices
3. Key Sequence: Text, Reviews, Comments, Survey, Instant messaging
4. Voice/Video Data: Video Conferencing
5. Spatio-temporal Data: Check-in Services



# Web Sensing

## Applications Targeted to Individuals

### 1. Targeted advertisement

### 2. Personalized Search Results

Google

mueen

Google

mueen

Web

News

Images

Videos

Shopping

More

Search tools

About 306,000 results (0.41 seconds)

Teaching - Abdullah Mueen

abdullahmueen.com/teaching.html

Spring 2014 CS 464/564 : Introduction to Database Management System

You've visited this page many times. Last visit: 4/22/14

Abdullah Mueen

www.abdullahmueen.com/

Albuquerque, NM, 87131. Phone: (505) 277 1914 mueen(at)cs.unm.edu. Research Interest. My interest is in time series data (i.e. real-valued sequence) mining.

Teaching - Publications - Personal

You've visited this page 2 times. Last visit: 4/23/14

Abdullah Mueen - Google Scholar Citations

scholar.google.com/citations?user=OImDWIoAAAAJ...

Department of Computer Science, University of New Mexico - Verified email at cs.unm.edu

T Rakthanmanon, B Campana, A Mueen, G Batista, B Westover, Q Zhu, ...

Proceedings of the 18th ACM SIGKDD international conference on Knowledge ...

You've visited this page many times. Last visit: 4/22/14

dblp: Abdullah Mueen

www.informatik.uni-trier.de/~leyl/.../Mueen:Abdullah

Mar 19, 2014 - Abdullah Mueen: Time series motif discovery: dimensions and applications. Wiley Interdisc. Rev.: Data Mining and Knowledge Discovery 4(2): ...

You've visited this page many times. Last visit: 4/23/14

Web

News

Images

Videos

Shopping

More

Search tools

About 306,000 results (0.21 seconds)

Abdullah Mueen

www.abdullahmueen.com/

Albuquerque, NM, 87131. Phone: (505) 277 1914 mueen(at)cs.unm.edu. Research Interest. My interest is in time series data (i.e. real-valued sequence) mining.

Teaching - Publications - Personal

Abdullah Mueen - Google Scholar Citations

scholar.google.com/citations?user=OImDWIoAAAAJ...

Department of Computer Science, University of New Mexico - Verified email at cs.unm.edu

T Rakthanmanon, B Campana, A Mueen, G Batista, B Westover, Q Zhu, ...

Proceedings of the 18th ACM SIGKDD international conference on Knowledge ...

Chowdhury Mueen-Uddin - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/Chowdhury\_Mueen-Uddin

Chowdhury Mueen-Uddin (Bengali: চৌধুরী মঈনুদ্দিন; born 27 November 1948), is one of the convicted war criminal for killing Bengali intellectuals in ...


dblp: Abdullah Mueen

www.informatik.uni-trier.de/~leyl/.../Mueen:Abdullah

Mar 19, 2014 - Abdullah Mueen: Time series motif discovery: dimensions and applications. Wiley Interdisc. Rev.: Data Mining and Knowledge Discovery 4(2): ...

Ally Bank® - Member FDIC

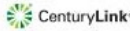
ally.com



Recognized as 2011, 2012 & 2013 "Best Online Bank" by MONEY® Magazine

CenturyLink®


promotions.centurylink.com



Get Internet, TV and Phone for 1 low price when you bundle

Expedia - Find Yours™


expedia.com



Book at Riyadh - Grand Plaza Hotel for as low as \$450 a night.

Trade in and Trade Up

bestbuy.com




Trade in and trade up to Samsung GS5. Get up to \$200 back. Click to learn more!

Ferdous Kawsar likes this

Download MariaDB 10 GA


skysql.com



New version comes with superior performance, enhanced replication & NoSQL capabilities!

Faculty Travel Free


landing.efcollegestudytours.com



We're going to London. You should come. EF College Study Tours.

New Spring 2014 Catalog!

overstockart.com



Browse our New Spring 2014 Catalog and Enjoy an Extra 20% Off your entire purchase!

# Web Sensing

## Social/Community Sensing

### Data:

Networks: Friend Net, Call Net, Follower Net,

Text: News, Reviews, Comments, Twits

Census Data

### Applications:

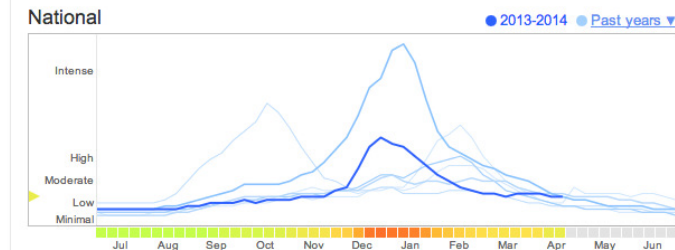
[Flue Trends](#)

[BoxOffice Prediction](#)

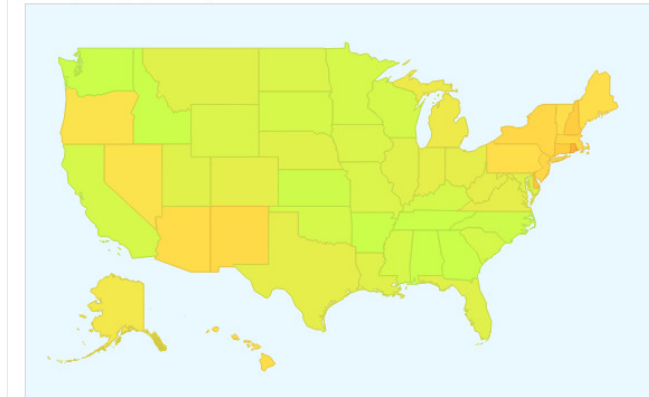


### Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. [Learn more »](#)



States | [Cities](#) (Experimental)



Estimates were made using a model that proved accurate when compared to historic official flu activity data. Data current through April 21, 2014.

# Business

---

Stock market

Banks

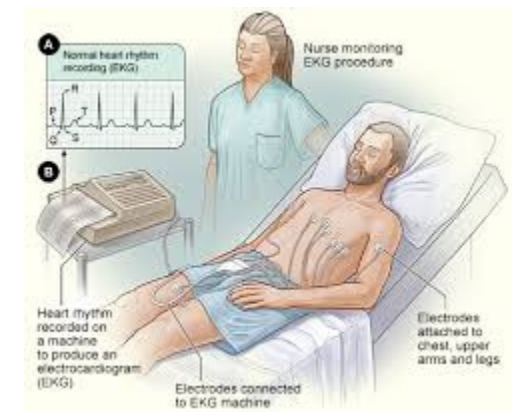
Insurance...



# Health and Medicine

Patient Records (Clinical, Pathological etc.)

Sequencing Data...



[Success Stories in Data/Text Mining](#) by Christophe Giraud-Carrier

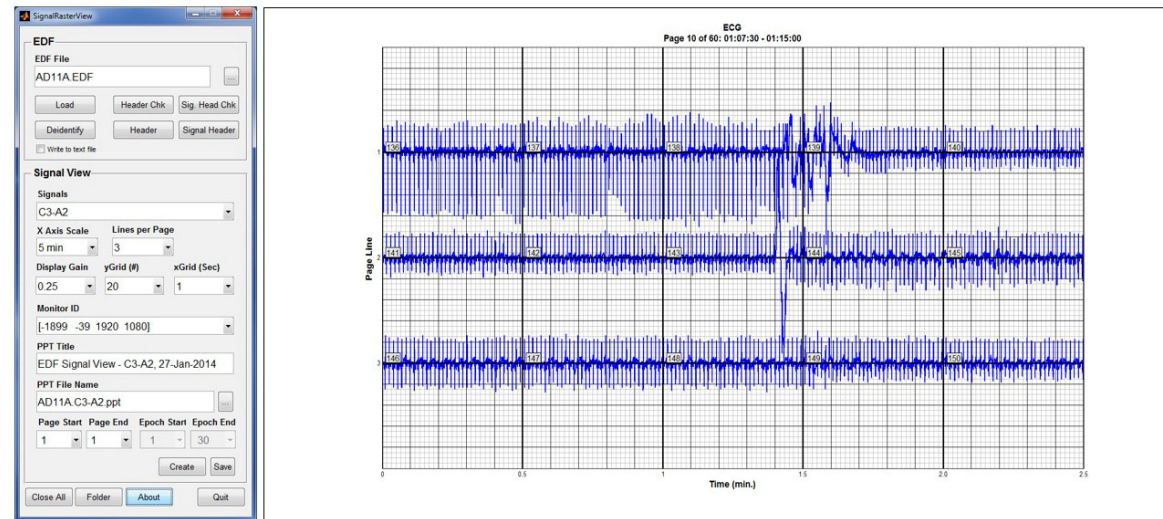


# Medical

Electro-physiological data

Signals <http://www.physionet.org/>

Images (microarray)

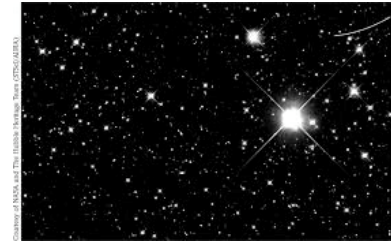


# Remote Sensing

---

From Earth to the Outer Space

From Space to the Earth



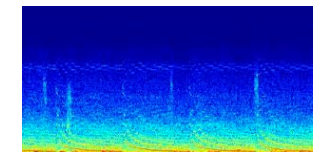
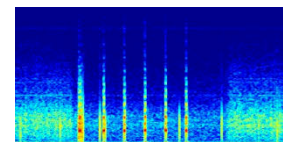
**Data:**

[Images and spectrograms](#)

**Derived Data:**

Vegetation Index

Sea-surface Height





# Remote Sensing

---

## **Applications** in Space Exploration

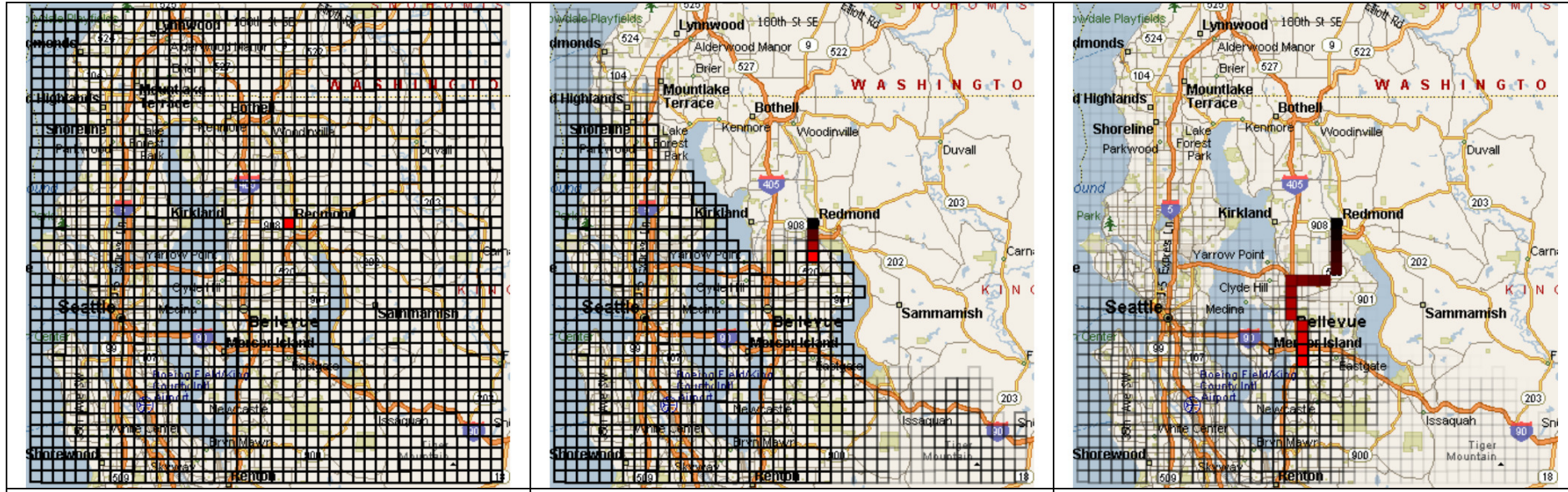
1. Detecting, Tracking, categorizing asteroids
  - [TopCoder Contest](#)
2. Categorizing stars based on types and their remaining life using [light curves](#)

## **Applications** in Observing Earth

1. Modeling and Validating [Climate Changes](#)
2. Predicting storm formation
3. Detecting forest fire, deep ocean eddies, air pollution, etc. [[Expedition](#)]



# Movement Sensing



**Data:** GPS Traces of [Human](#) and [Animals](#), Maps  
**Applications**

1. Traffic based route planning
2. Destination Prediction
3. Opportunistic Crowdsourcing



# Government Data

---

## Data:

Transportation Data

Environmental Data

Utility Data

Police Data

 <http://www.cabq.gov/abq-data>

## Applications:

[Smart City Applications](#)

Energy Efficient [Building](#), [Transportation](#) etc.

# Anthropology



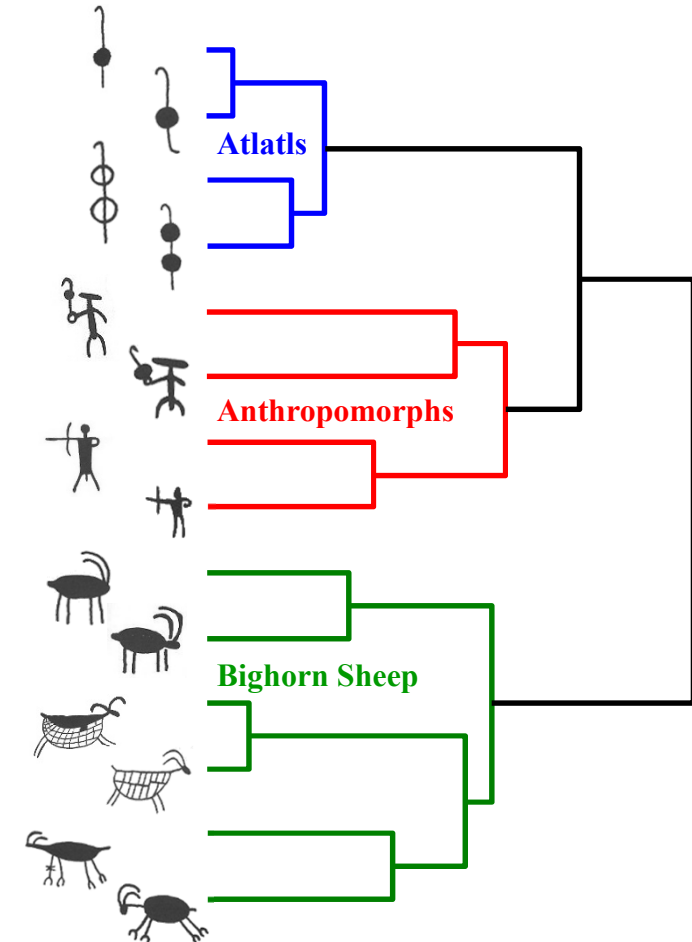
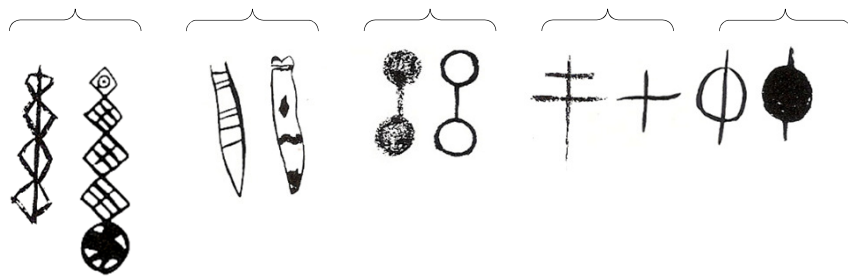
## Data:

Images and Shapes of the Petroglyphs and Petrographs

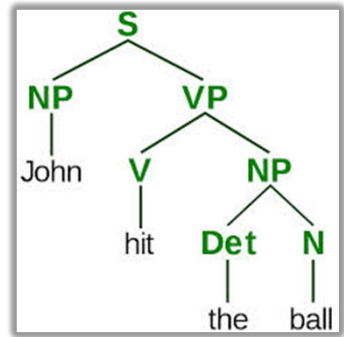
## Applications:

Clustering Petroglyphs

Finding repeated Petroglyphs across states or countries



# Linguistics



## Data:

Text Data: Books and News

Audio: [Audio Corpus](#)

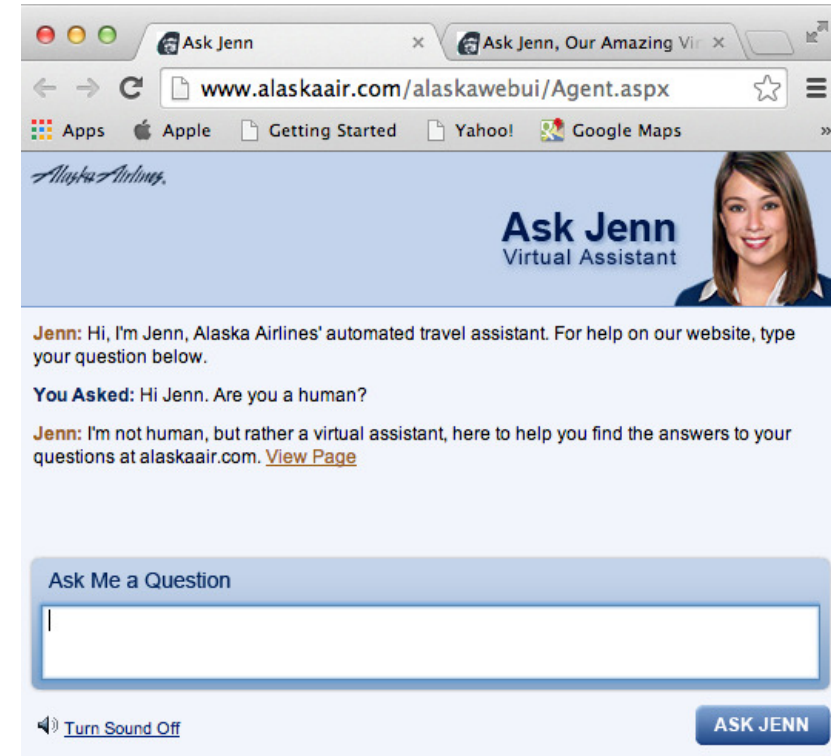
## Applications

Machine Translation

Dialogue Processing

NLP for assistive technologies

[IBM Watson](#)



# Data Mining Algorithms

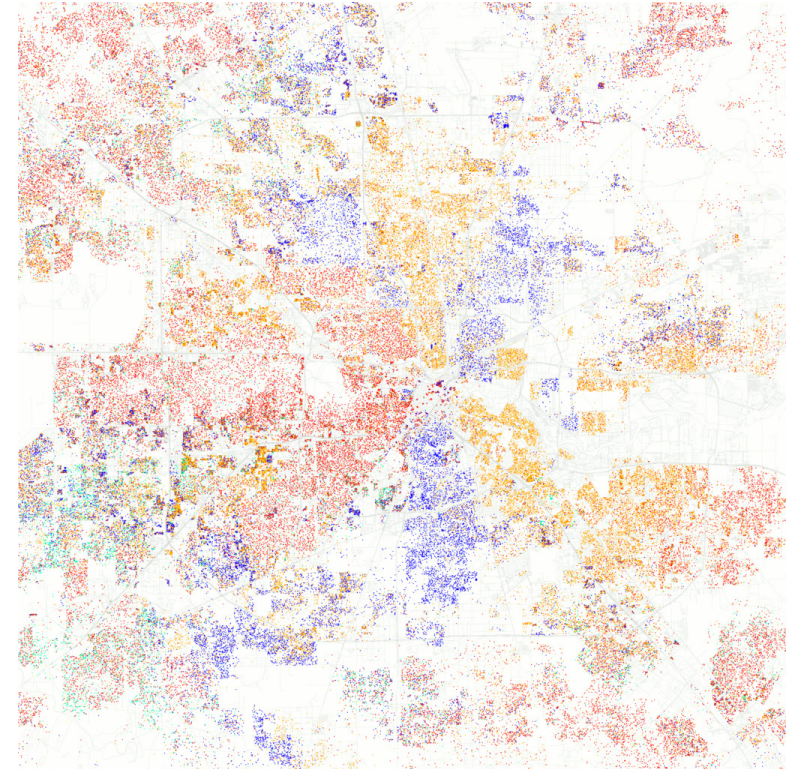
---



# Clustering

---

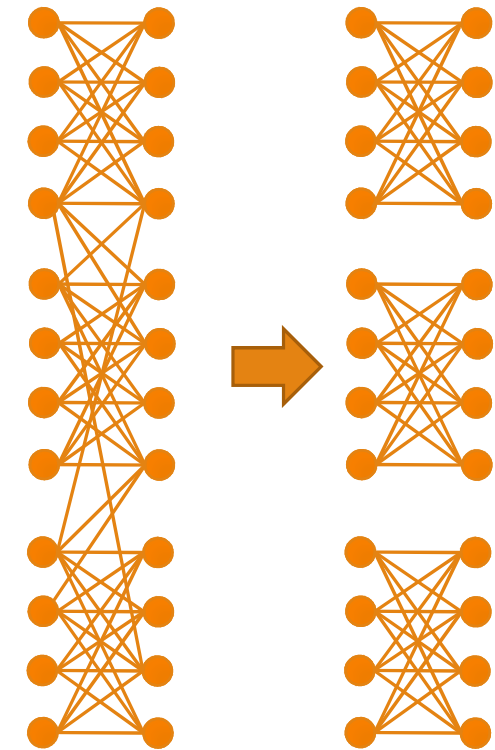
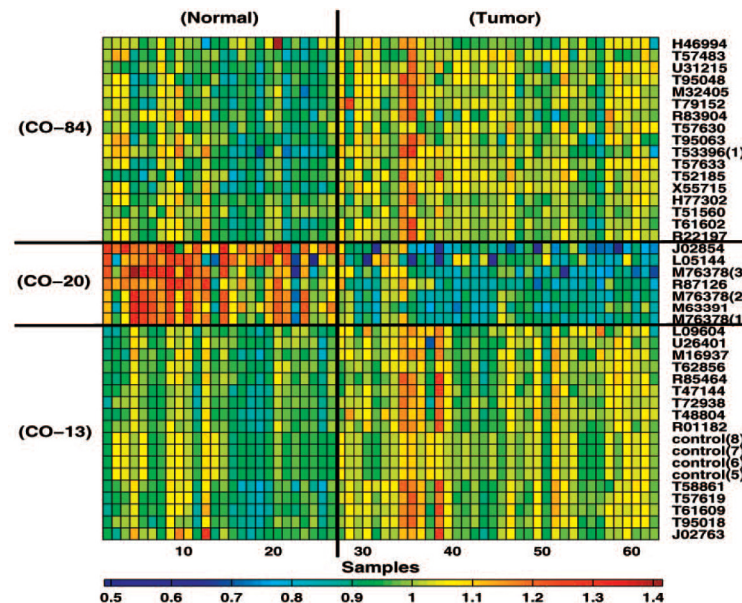
- Divide the data in meaningful partitions
- Need a goodness measure
- Tool: [Weka](#), Matlab



Houston, Ethnic Distribution

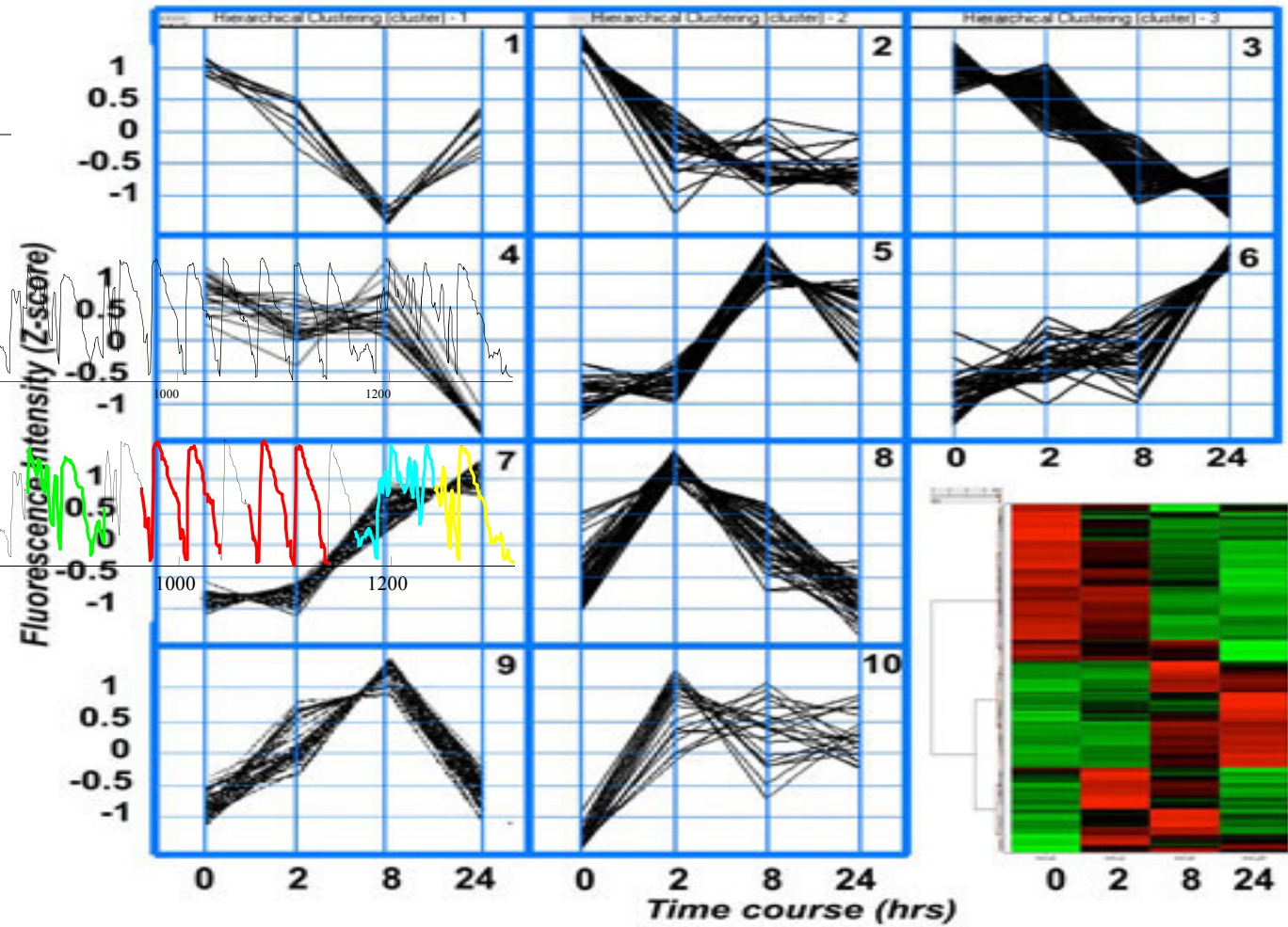
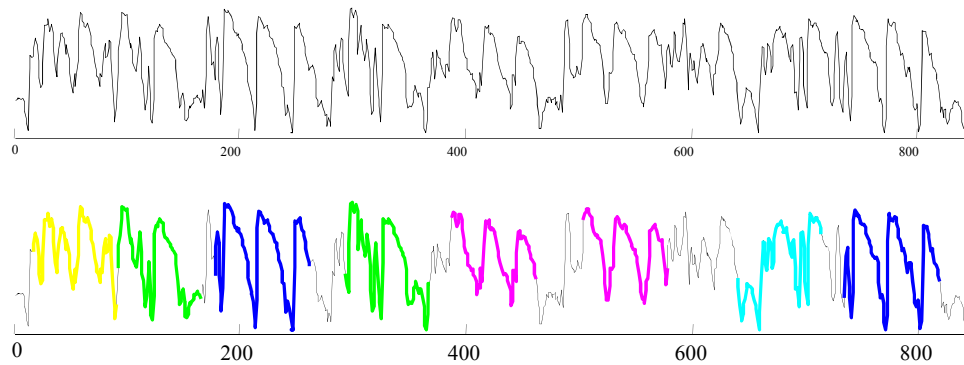
# Graph Clustering

- Neighborhood based similarity
- Co-Clustering is a way to find the heavily connected components of a bipartite graph.
- Tool: [cocluster](#)



Co-clustering

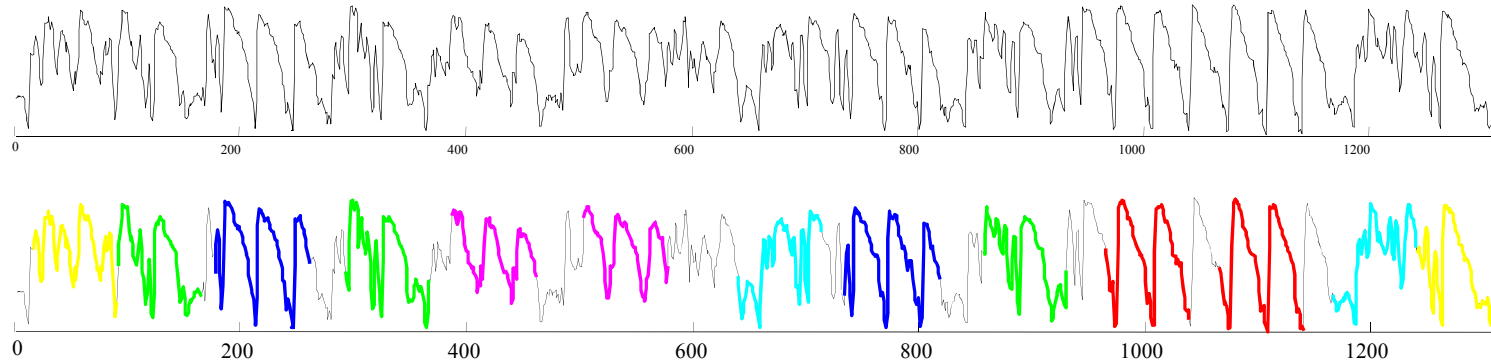
# Signal Clustering



[Link](#)



# Signal Clustering



- Clusters the subsequences of the signal
- Ignores unnecessary segments
- Tool: [Epenthesis](#)

== Poem (original order)==

In a sort of Runic rhyme,  
To the throbbing of the bells--  
Of the bells, bells, bells,  
To the sobbing of the bells;  
Keeping time, time, time,  
As he knells, knells, knells,  
In a happy Runic rhyme,  
To the rolling of the bells,--  
Of the bells, bells, bells--  
To the tolling of the bells,  
Of the bells, bells, bells, bells,  
Bells, bells, bells,--  
To the moaning and the  
groaning of the bells.

==Poem (grouped by clusters)==

bells, bells, bells,  
Bells, bells, bells,  
Of the bells, bells, bells,  
Of the bells, bells, bells--  
To the throbbing of the bells--  
To the sobbing of the bells;  
To the tolling of the bells,  
To the rolling of the bells,--  
To the moaning and the groan-  
time, time, time,  
knells, knells, knells,  
sort of Runic rhyme,  
groaning of the bells.

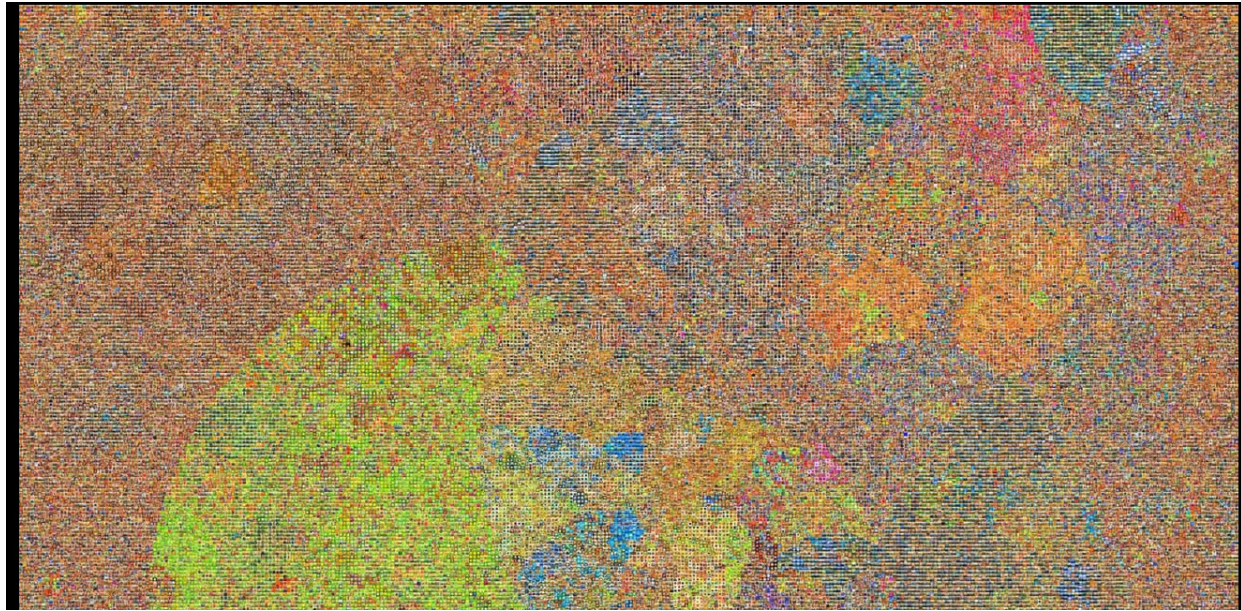


# Image Clustering

- Clustering based on color, texture, background etc.
- Ranges from small scale to web scale.

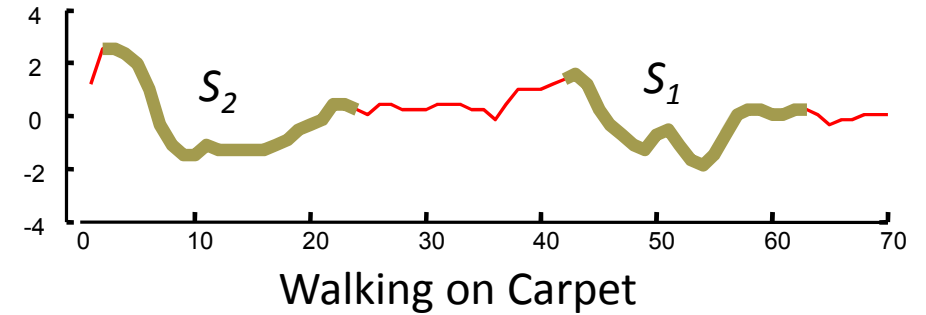
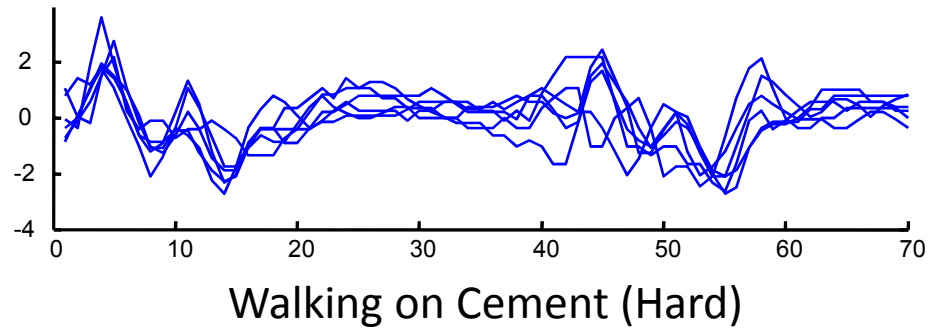
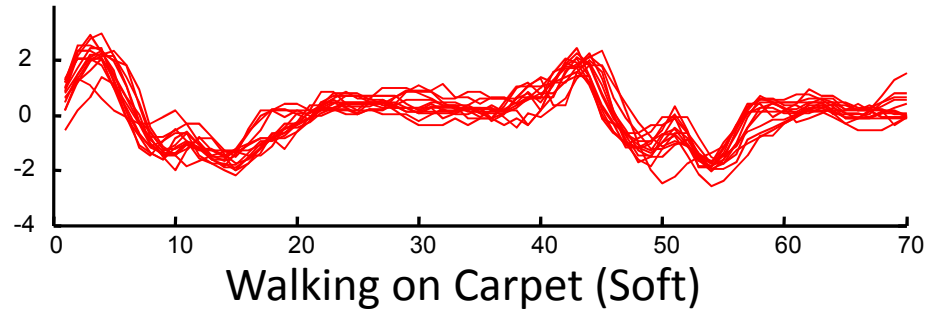


<http://www.ulrichpaquet.com/current.html>

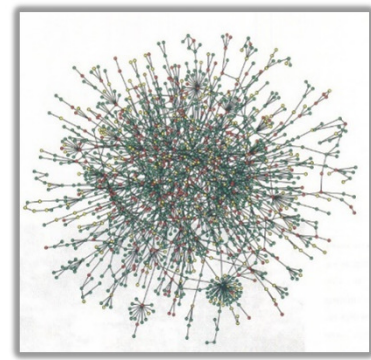


<http://groups.csail.mit.edu/vision/TinyImages/>

# Classification

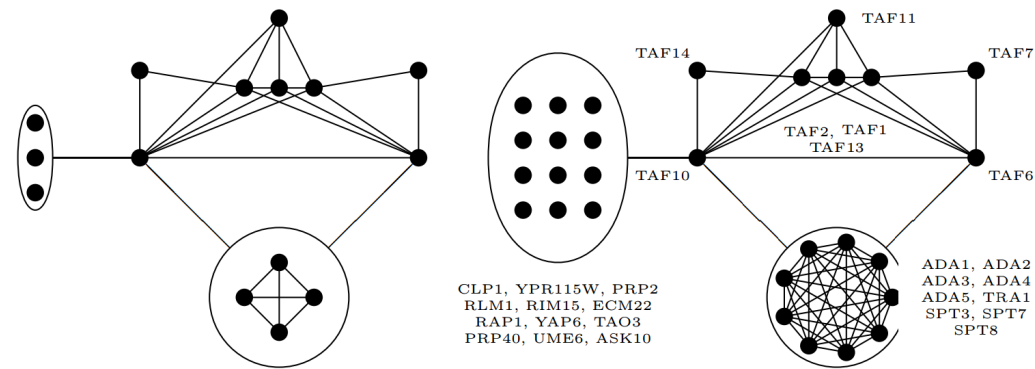


- Intuitive pattern for classification
- Very fast testing
- Tool: [Shapelet](#)



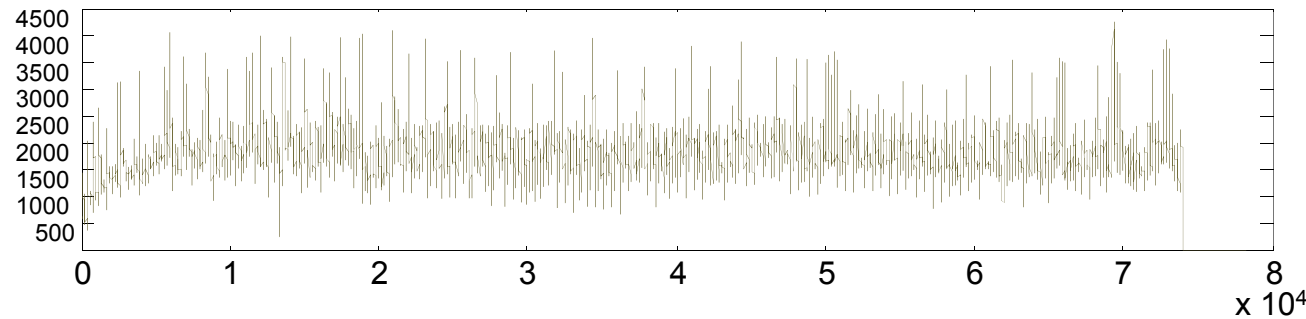
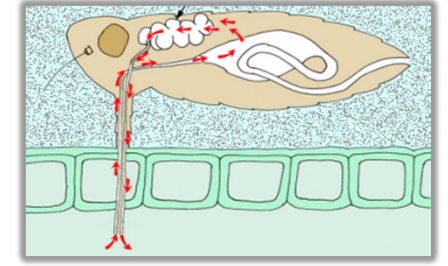
# Repetition Detection: Graph

- Frequent Subgraph Mining
- Various Constraints on the Subgraph
- Tool: [gSpan](#)

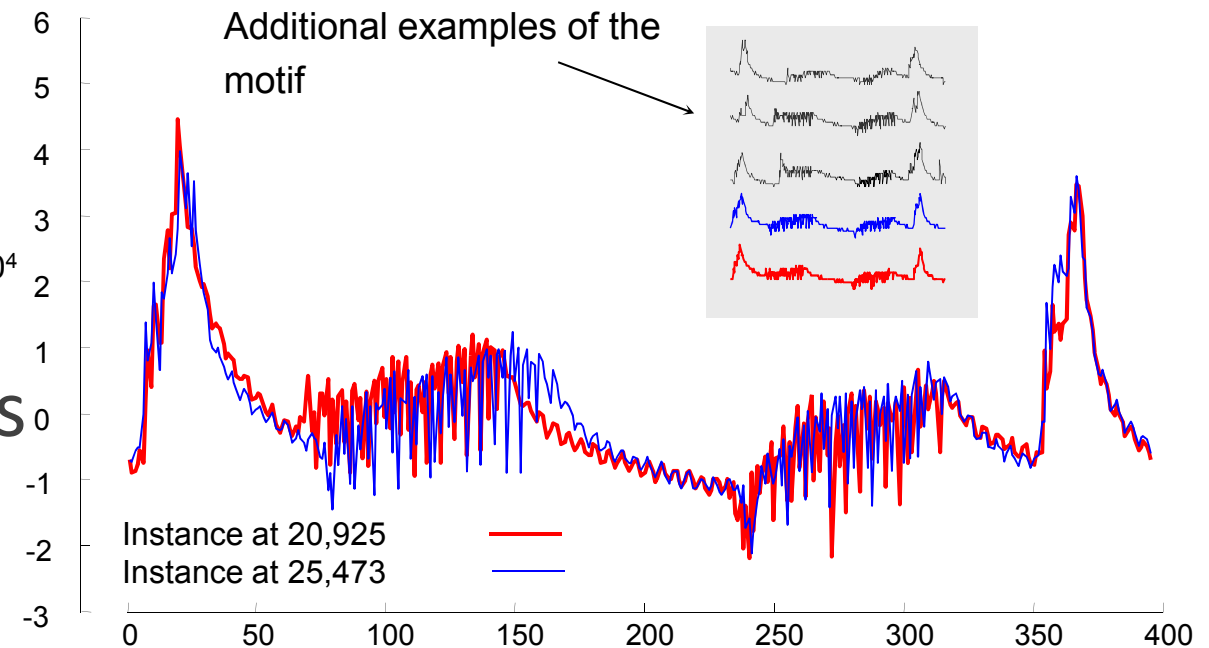


[Reference](#)

# Repetition Detection: Signal



- Motif Discovery in Time Series
- Parameter-free method
- Tool: [MOEN](#)

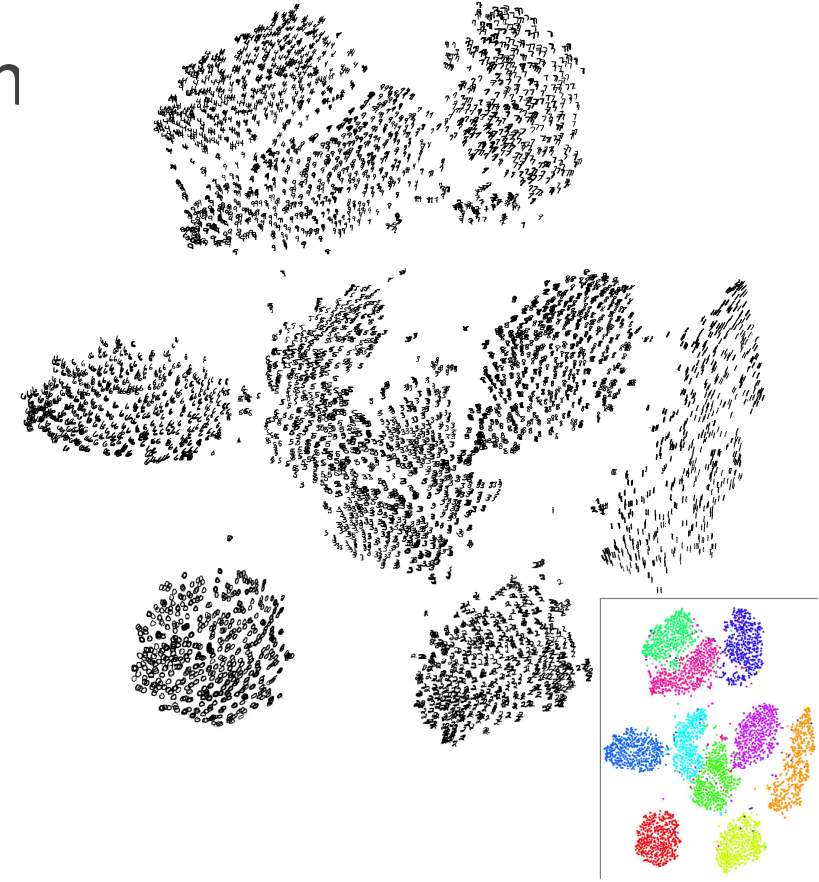




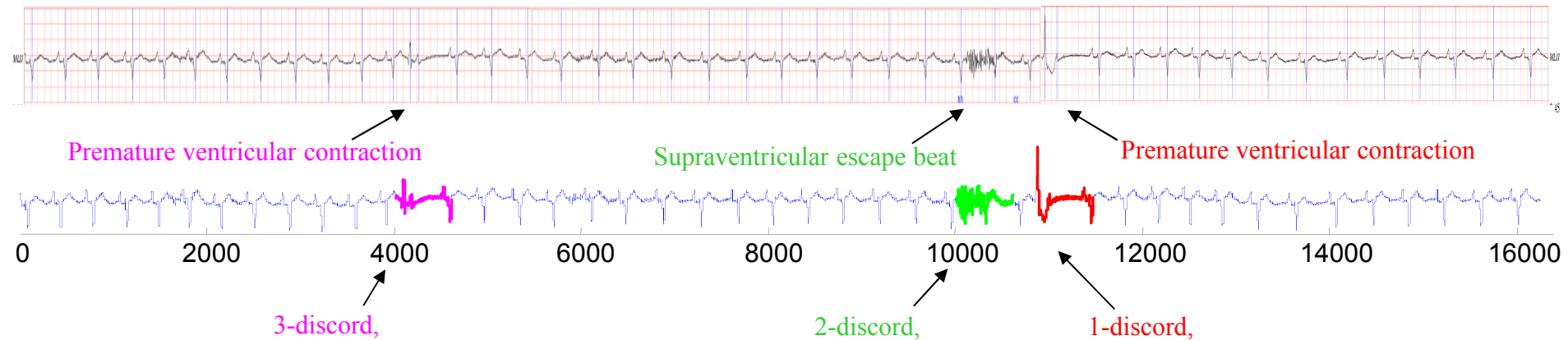
2	6	8	9	3	4	7	5	6
3	4	7	9	5	5	6	7	2
5	8	7	0	9	4	3	5	4
<b>5</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>9</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>

# Visualization

- High Dimensional Data Visualization
- 2D and 3D
- Preserving Neighborhood of the points
- Tool: [t-SNE](#)



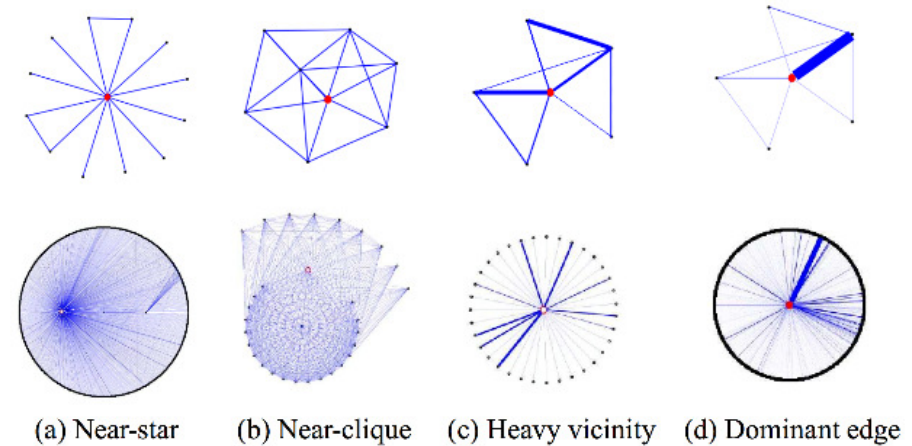
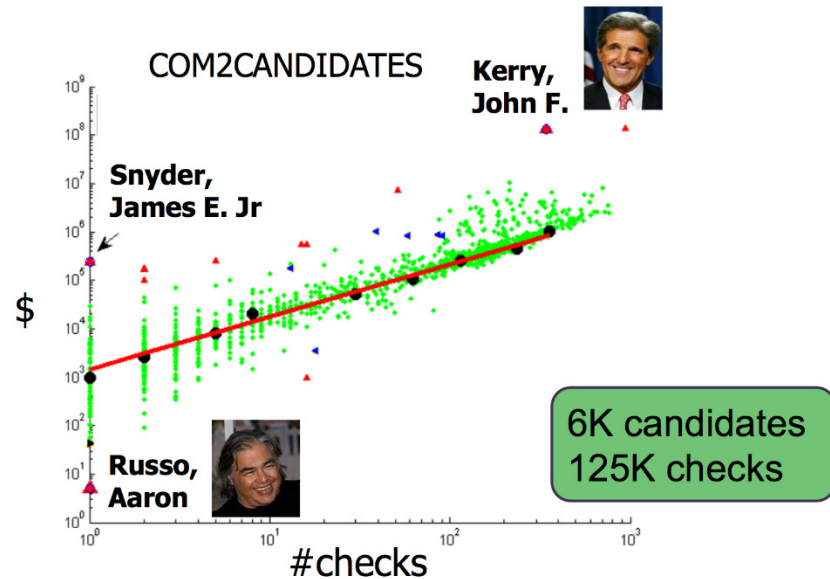
# Anomaly Detection: Signal



- Most unusual pattern in the signal
- Works in two passes
- Tool: [Discord](#)



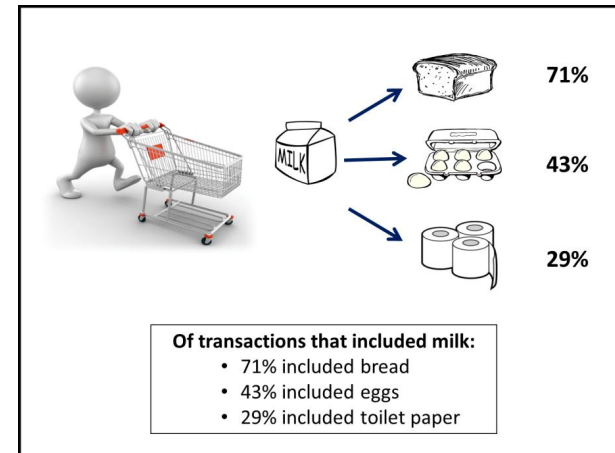
# Anomaly Detection: Graph



- Neighborhood based features
- Finds extremes in both direction
- Tool: [OddBall](#)

# Association Detection

- Finds association among items with high support and confidence
- The algorithms are mostly exponential
- Tool: [SPSS Modeler](#), Weka



No.	Association Rule	Support	Confidence
1	{Vaginal ultrasound; Surgical pathology; Pregnancy test; Hematology; Induced abortion; Penicillin injection} $\implies$ {Legally induced abortion}	173	99.42%
2	{Pulmonary bronchospasm evaluation; Pulmonary vital capacity test; Non-pressurized inhalation treatment for acute airway obstruction; Doctor's office visit } $\implies$ {Asthma}	56	91.80%
3	{Debridement of nails, manual, five or less; Debridement of nails, each additional, five or less; Intestine excision: Enterointerostomy, anastomosis of intestine with or without cutaneous enterostomy; Transurethral surgery (Urethra and bladder)} $\implies$ {Dermatophytosis}	619	91.43%

[Reference](#)