Machine Learning, Methods and Applications

Yapay Öğrenme Metotlar, Uygulamalar

A. Taylan Cemgil, Bilgisayar Müh.

20.10.2016, NMK, Şirince

http://www.cmpe.boun.edu.tr/pilab

- Suzan Üsküdarlı
  - Onur Güngör, Ahmet Yıldırım
- Arzucan Özgür
  - Çağıl Uluşahin
- İlker Birbil (Sabancı)
- Figen Öztoprak (Bilgi)
- Almila Akdağ (UvA)
- ATC
  - Umut Şimşekli, Beyza Ermiş, Caner Türkmen
  - Yener Ülker, Can Kavaklıoğlu

Konu Başlıkları

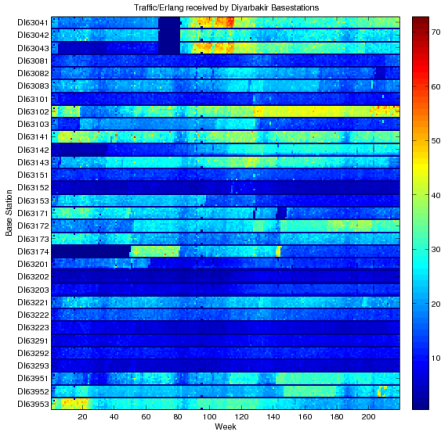
- Büyük Veri - Yapay Öğrenme
- Kullanım örnekleri (Use Cases)
- Güdümlü Öğrenme (Supervised Learning)
  - Sınıflandırma (Classification)
- Güdümsüz Öğrenme (Unsupervised Learning)
  - Öbekleme (Clustering)
  - Boyut indirgeme (Dimensionality Reduction)
- Büyük verilere uygulama (Scaling)
  - Mimariler/Büyük veri araçları
- Referanslar
- Sonuç

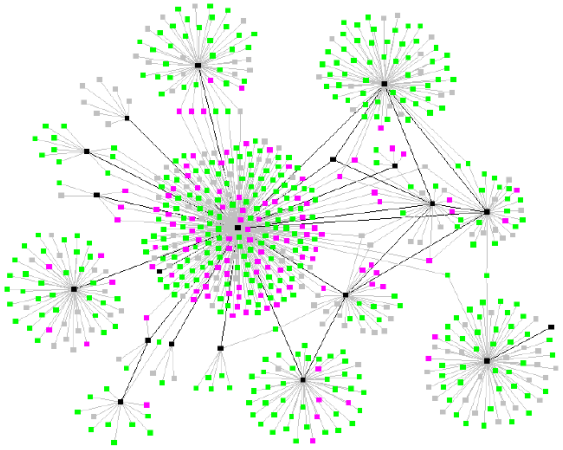
Büyük Veri

- Yazılım/Donanım Altyapısı (SW/HW Infrastructure)
- Görselleştirme/Etkilşim (Visualization/Interaction)
- Data Analytics ←Yapay Öğrenme

Yapay Öğrenme

- Hesaplama tabanlı metotlar topluluğu
  - Gizli örüntü (pattern) yakalama
  - Öngörüler üretmek
  - Belirsizlik altında karar desteği
  - Ham veriyi faydalı bilgiye dönüştürmek





Yapay Öğrenme, Veri Madenciliği, İstatistik

Yapay öğrenme ve Büyük veriler. Gerçekten yeni mi?

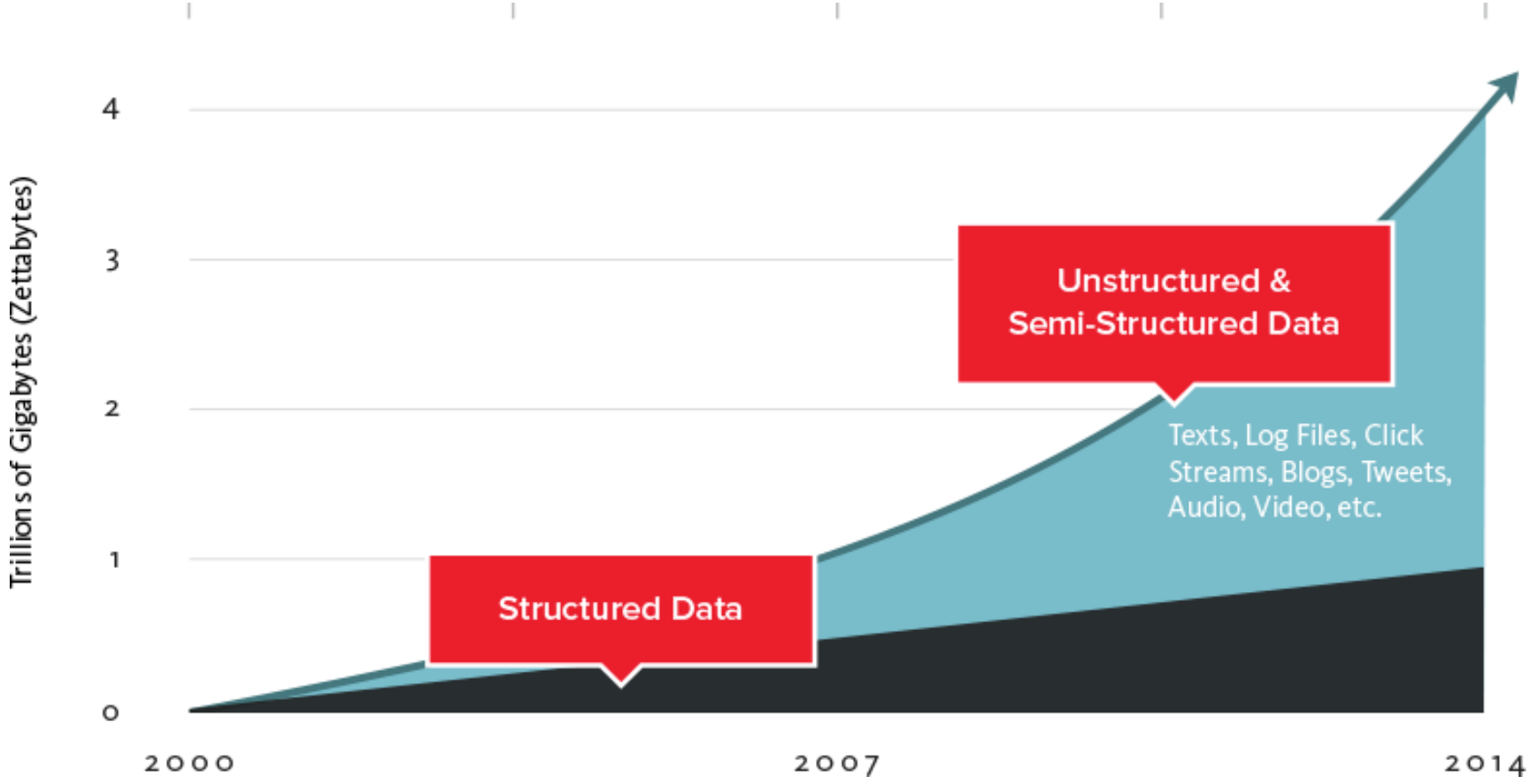
- Eski metotlara yeni bir bakış
- … ve yenilerinin geliştirilmesi
- Büyük boyutun Laneti/Nimeti ‘Curse/Blessing of Dimensionality’
- Ucuzlayan Altyapı
  - Bulut Bilişim
  - Sensör Ağları, Nesnelerin İnterneti (“yeni veri”)
  - Hız (“gerçek zaman”)

Büyük Veri?

- Pragmatik Bakış Açısı
  - Küçük Veri: Naif algoritmalar uygulanabilir
  - Ortaboy Veri: Bir bilgisayar sisteminde işlenebilen miktarda (Feasibly processed on one machine)
  - **Büyük Veri: Bir makinaya sığmayan miktarda**
- Karmaşık ilişkisel veri
  - ikili, üçlü veya daha üst seviyeden etkileşimler
- Hız, akan veri
- Yapılandırılmamış veri (blog metinleri/video/fotoğraf)
- 3V: Volume-Variety-Velocity

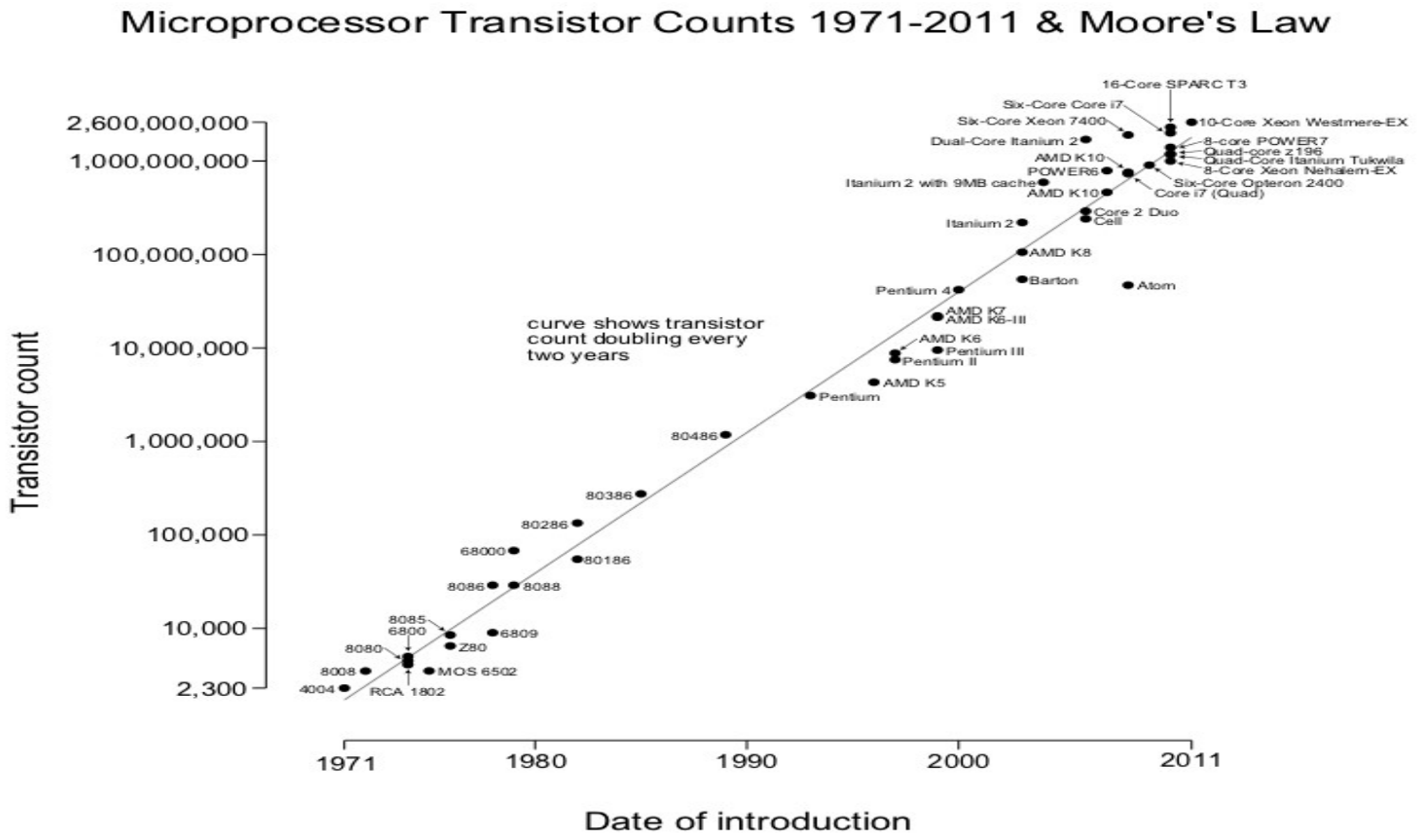
Büyük Veri?

- http://www.couchbase.com/nosql-resources/what-is-no-sql



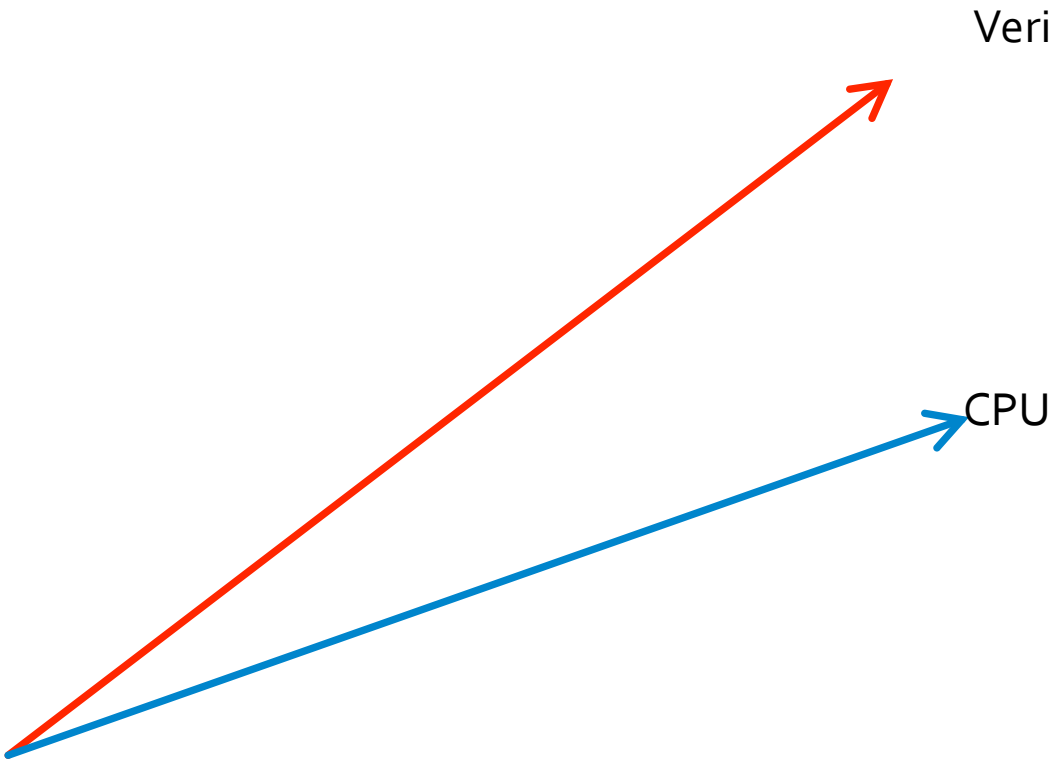
Moore Kanunu günü kurtarır mı?

- "Transistor Count and Moore's Law - 2011" by Wgsimon - Own work. Licensed under CC BY-SA 3.0 via Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Transistor_Count_and_Moore%27s_Law_-_2011.svg#/media/File:Transistor_Count_and_Moore%27s_Law_-_2011.svg



Moore Kanunu günü kurtarır mı?

- Veri patlaması Moore Kanunundan hızlı
- Fiziksel Enerji bariyeri



Parkinson'un 1. Kanunu

- 'Data Expands to fill the space available for storage'
- Veri, kaplar her yeri

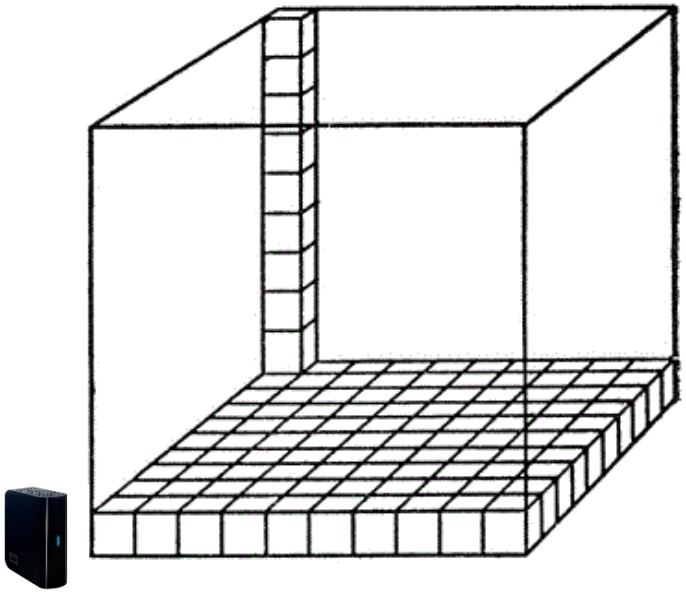
Hafıza Büyüklükleri

kilobyte (kB)	10^3	2^{10}
megabyte (MB)	10^6	2^{20}
gigabyte (GB)	10^9	2^{30}
terabyte (TB)	10^{12}	2^{40}
petabyte (PB)	10^{15}	2^{50}
exabyte (EB)	10^{18}	2^{60}
zettabyte (ZB)	10^{21}	2^{70}
yottabyte (YB)	10^{24}	2^{80}

Hafıza Büyüklükleri

= 1TB = 1 000 000 000 000 Byte

=1 Trilyon Byte



= 1PB

= 1 000 000 000 000 000B

=1 Katrilyon Bytes

Bazı Sayılar

- CERN: Büyük Hadron Çarpıştırıcısı: Yılda 15 petabyte (2013) Günde 1 petabyte (2015)


- Google: Günde 24 petabyte (2013) 100 petabyte (2014)


×24 000


Bazı Sayılar

- Facebook *Hadoop Dağıtık Dosya Sistemi* (*HDFS*) büyüklüğü 100 PB (2012) 300 PB (2014)

×100 000

- Aylık Küresel İnternet Trafiği (2011) yaklaşık 27500 PB (Kaynak: Cisco)

×27 500 000

Bazı Gözlemler (kaynak:eSpatial)

- Google's Eric Schmidt: "every two days now we create as much information as we did from the dawn of civilization up until 2003"
- According to McKinsey – a retailer using big data to the full could increase its operating margin by more than 60%.
- Bad data or poor data quality costs US businesses $600 billion annually.

Veri =Enformasyon ≠Bilgi

- Data =Information ≠Knowledge

Enformasyon içinde boğulurken bilgiye açlık çekiyoruz
We are drowning in data and starving for knowledge
– J. Naisbitt

(Machine Learning, a probabilistic perspective, KP Murphy)

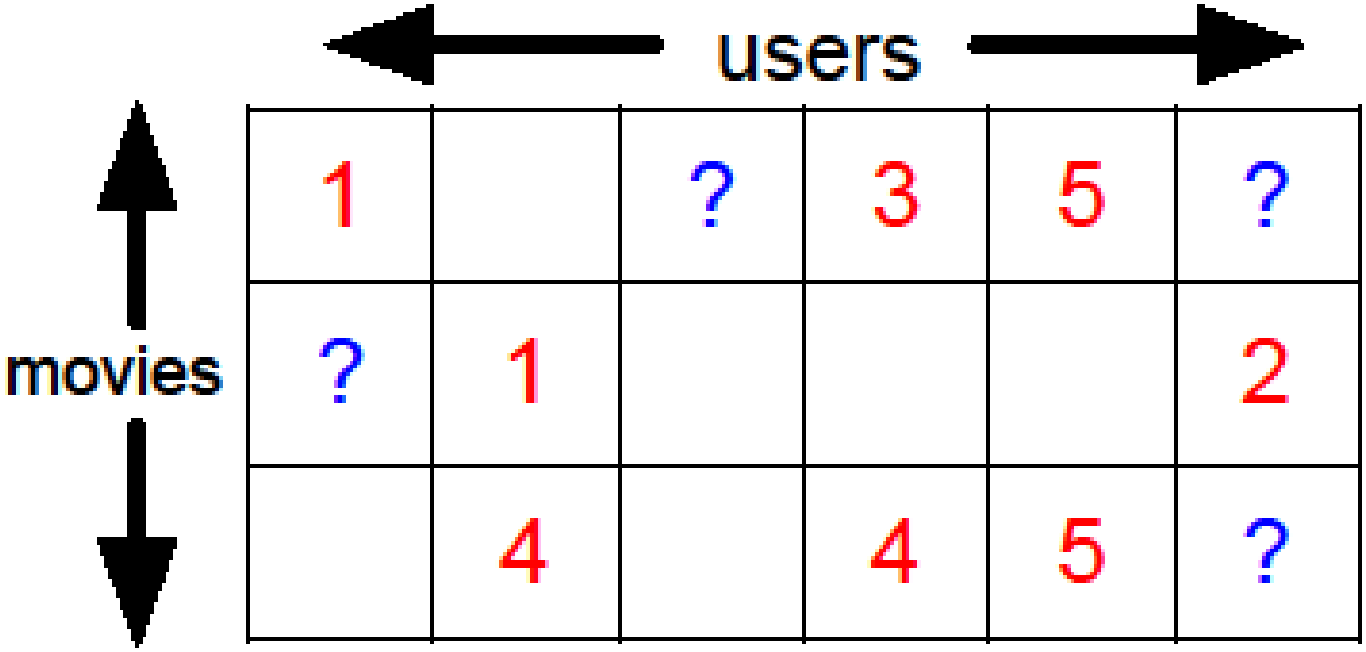
Kullanım Senaryoları: Perakende/ Tüketim

- Ürün Tavsiye Sistemleri
- Sepet Analizi (Market Basket Analysis)
- Olay/Aktivite/Davranış Analizi (Event/Activity/ Behavior Analysis)
- Kampanya yönetimi ve eniyilemesi
- Tedarik zinciri yönetimi
- Pazar ve Tüketici ayrıştırması

Kullanım Senaryoları: Tavsiye Sistemleri

Kullanım Senaryoları: Tavsiye Sistemleri

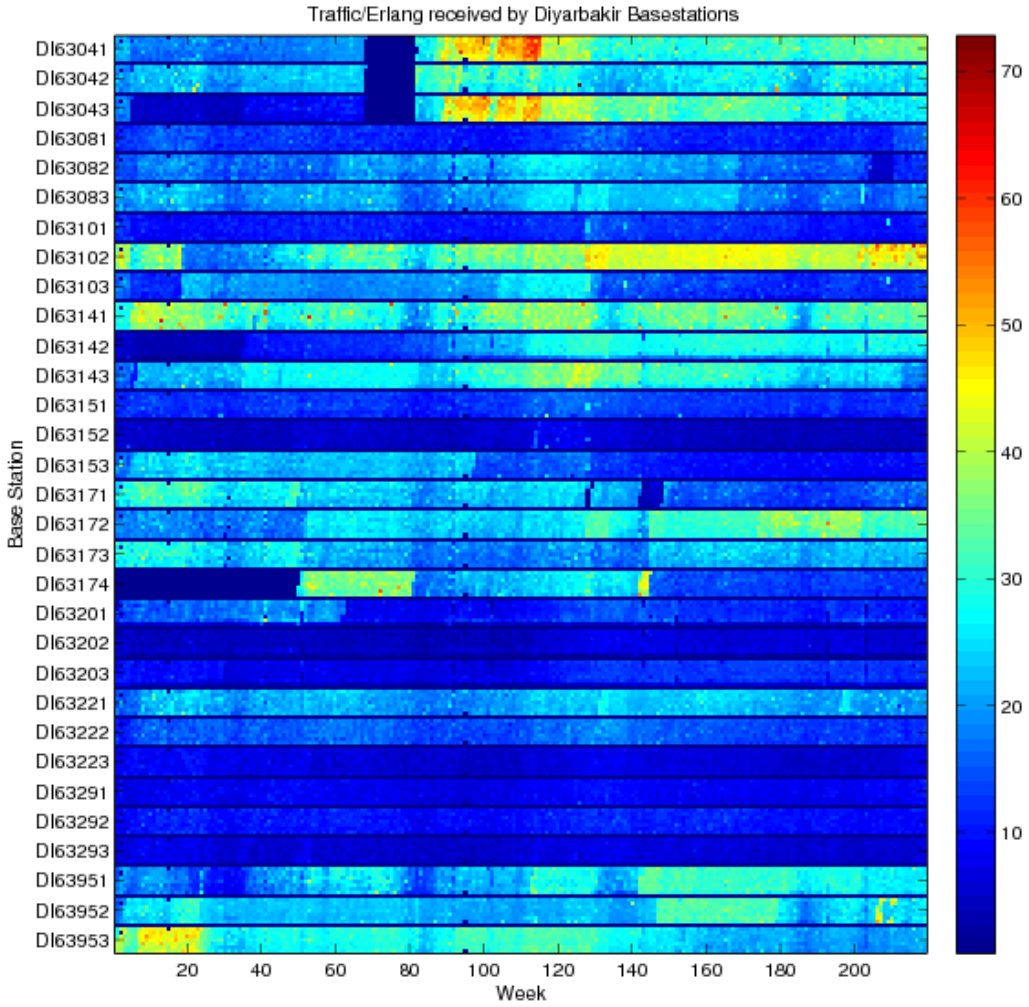
- Netflix: 18K film × 500K kullanıcı %99 seyrek

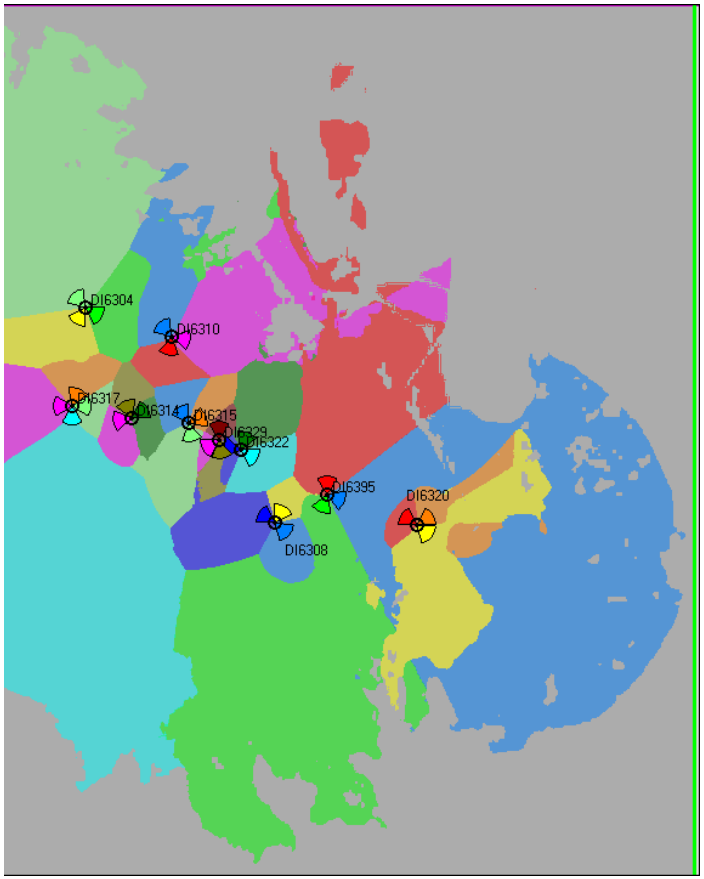


Kullanım Senaryoları: Haberleşme

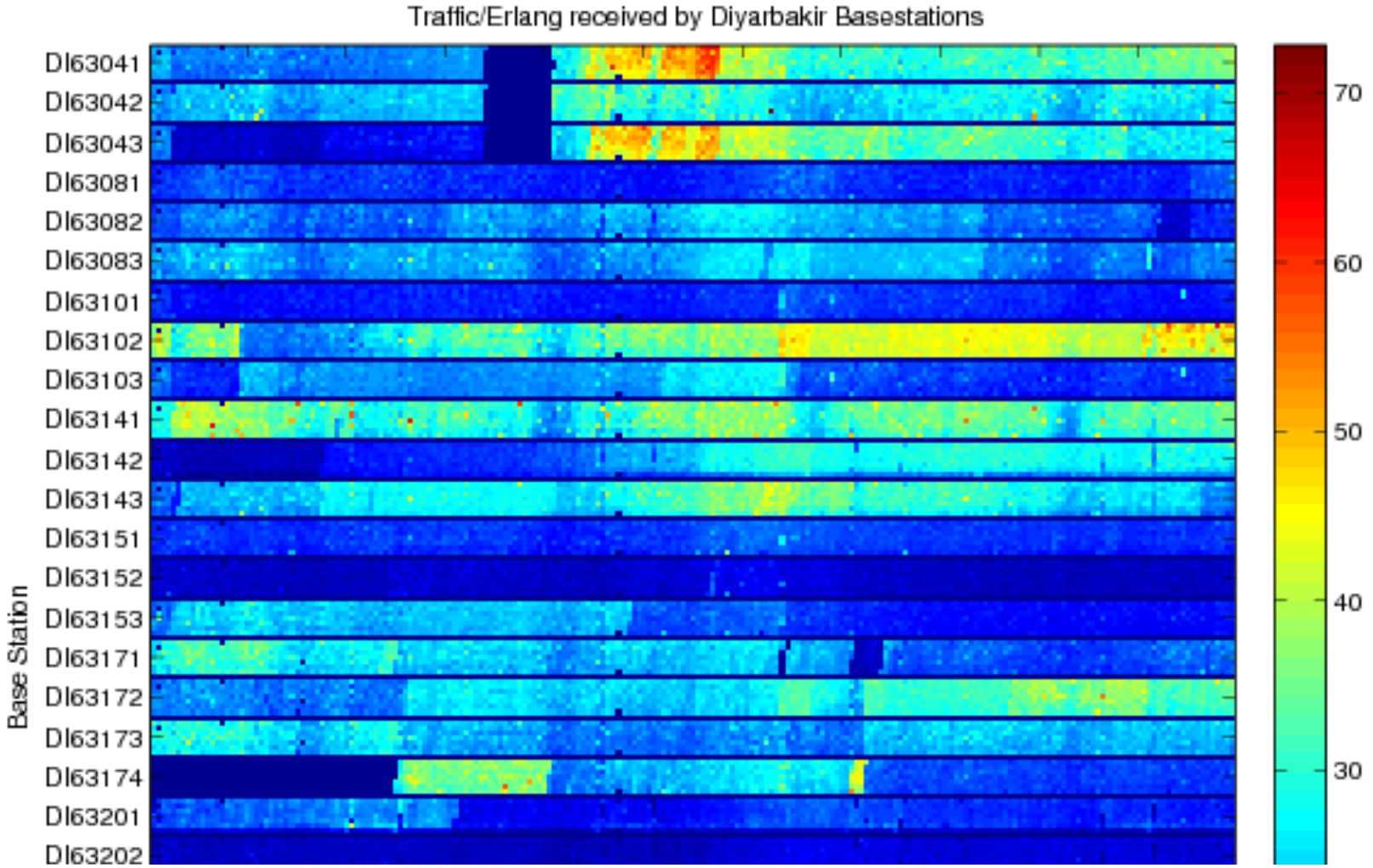
- Ağ izleme ve performans eniyileme
- Fiyatlandırma
- Müşteri ayrılma (Churn) tahmini
- Call Detail Record (CDR) Analizi
- (Mobile) Kullanıcı Davranış Analizi
- Siber güvenlik, DDOS saldırılarının tespiti ve önlenmesi
- Altyapı Planlaması

Kullanım Senaryoları, Haberleşme



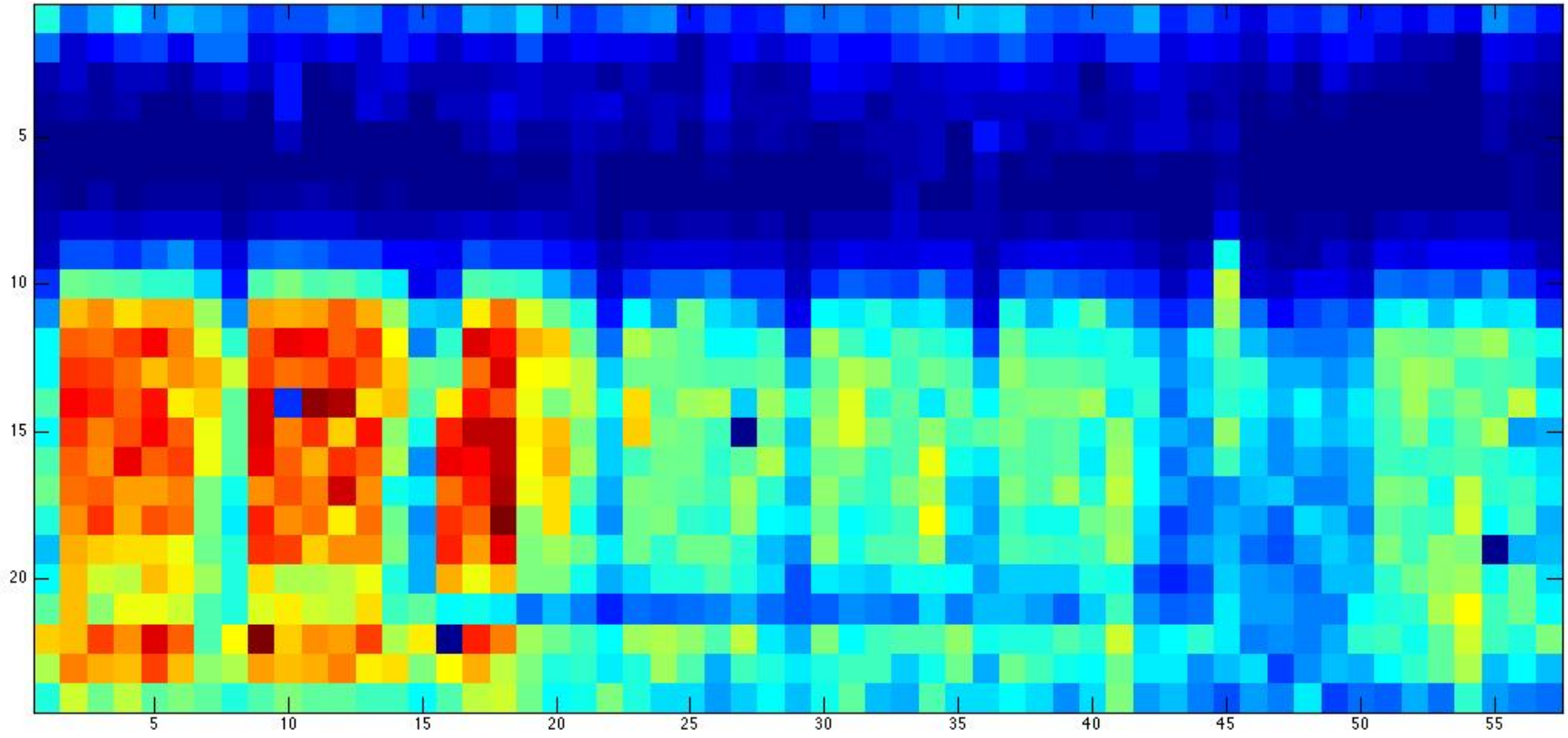


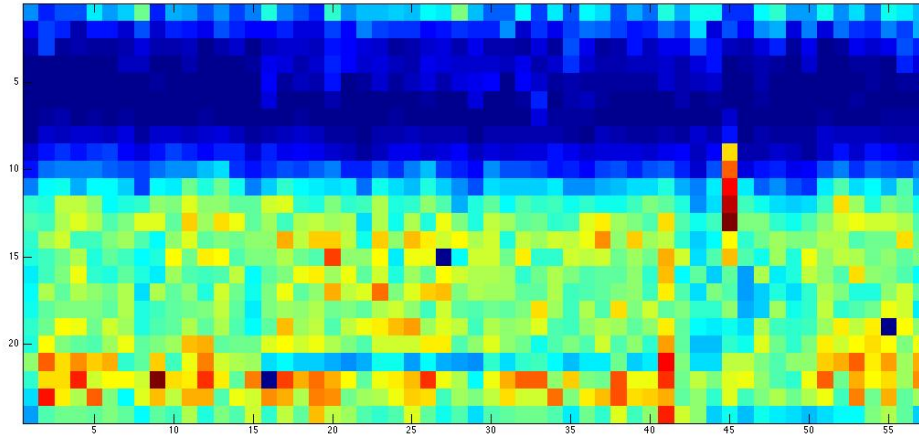
Kullanım Senaryoları

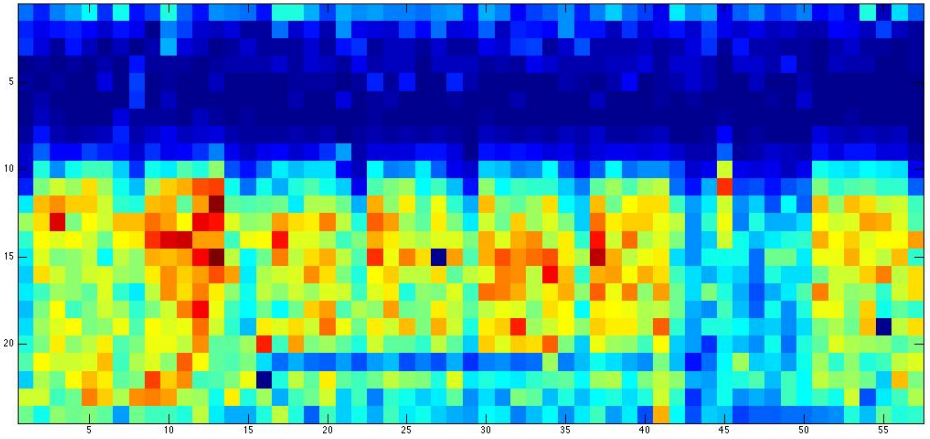


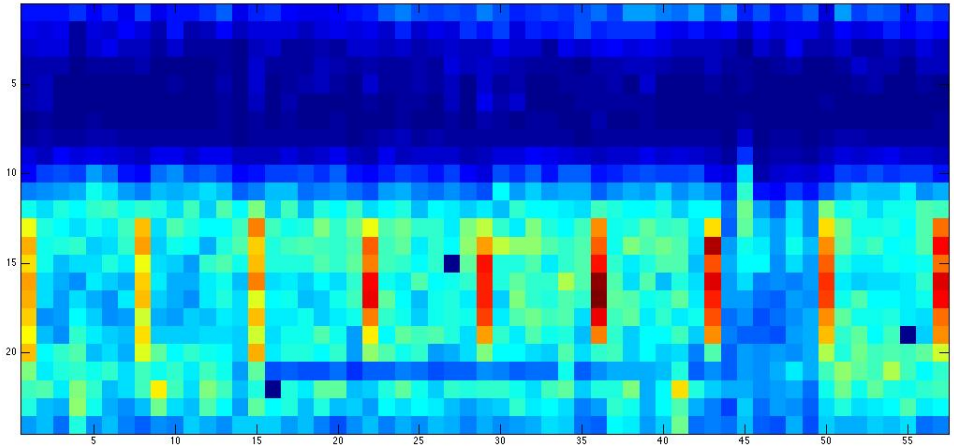
Kullanım Senaryoları, Haberleşme

- X: gün, Y: saat, renk: Kullanım Miktarı




5
10
15
20
5
10
15
20
25
30
35
40
45
50
55

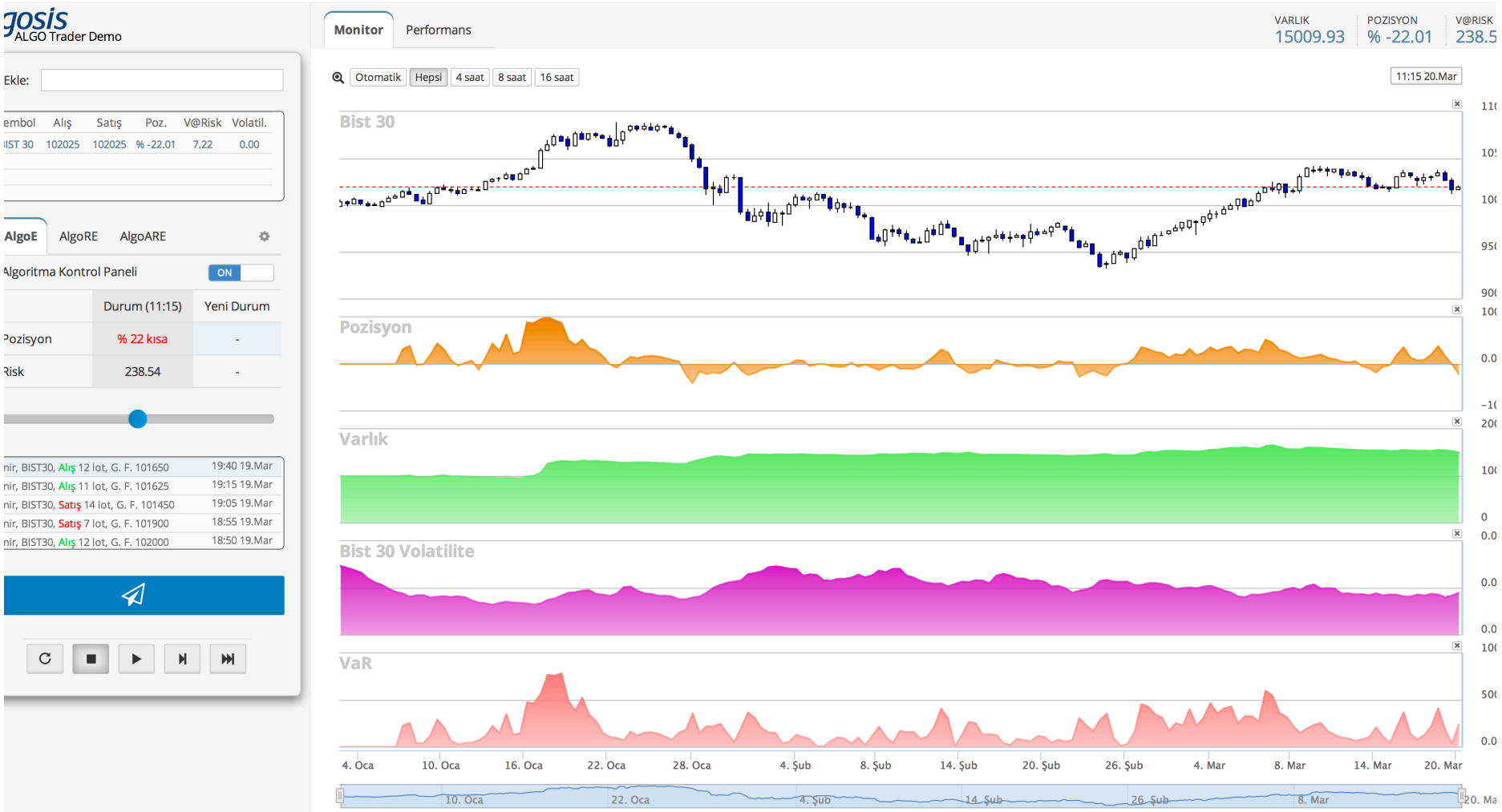

5
10
15
20
5
10
15
20
25
30
35
40
45
50
55


5
10
15
20
5
10
15
20
25
30
35
40
45
50
55

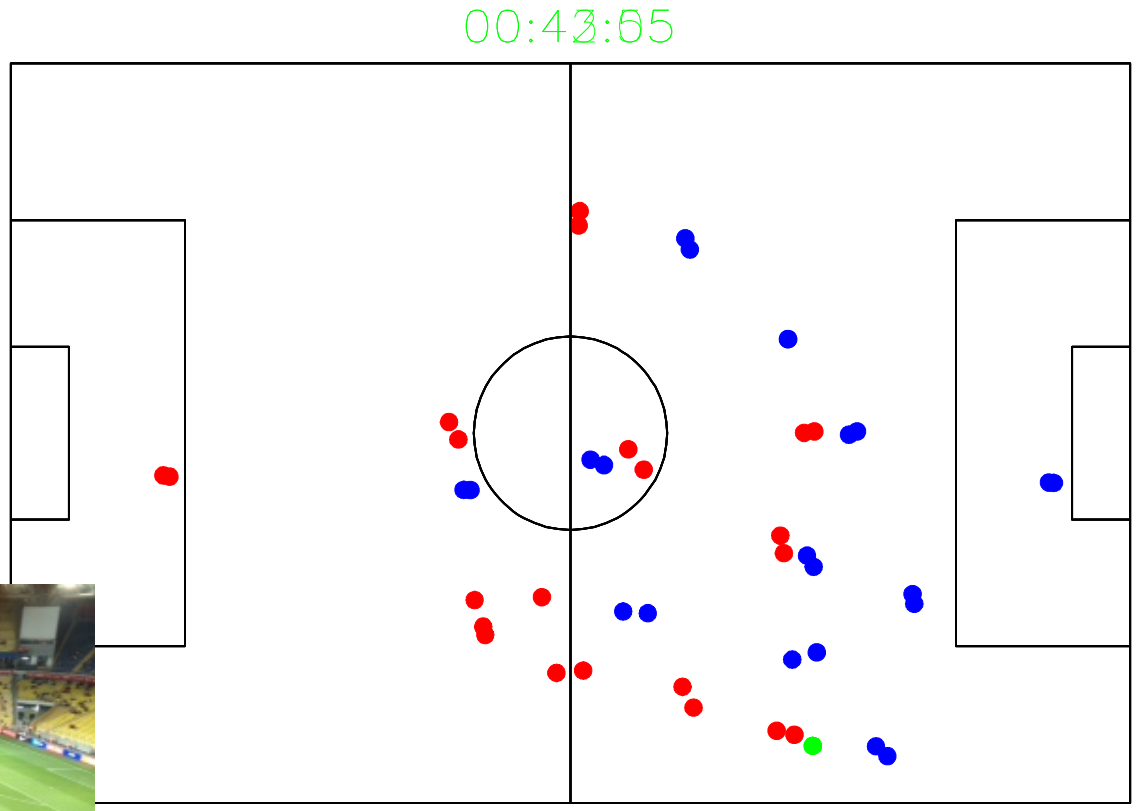
Kullanım Senaryoları: Finans/Ticaret/ Bankacılık

- Yolsuzluk (Fraud) Tespiti/Risk Kestirimi
- Yüksek hızda Trading
- Anomalite/Değişim noktası tanıma

Finansal Veri Analizi (Algosis)



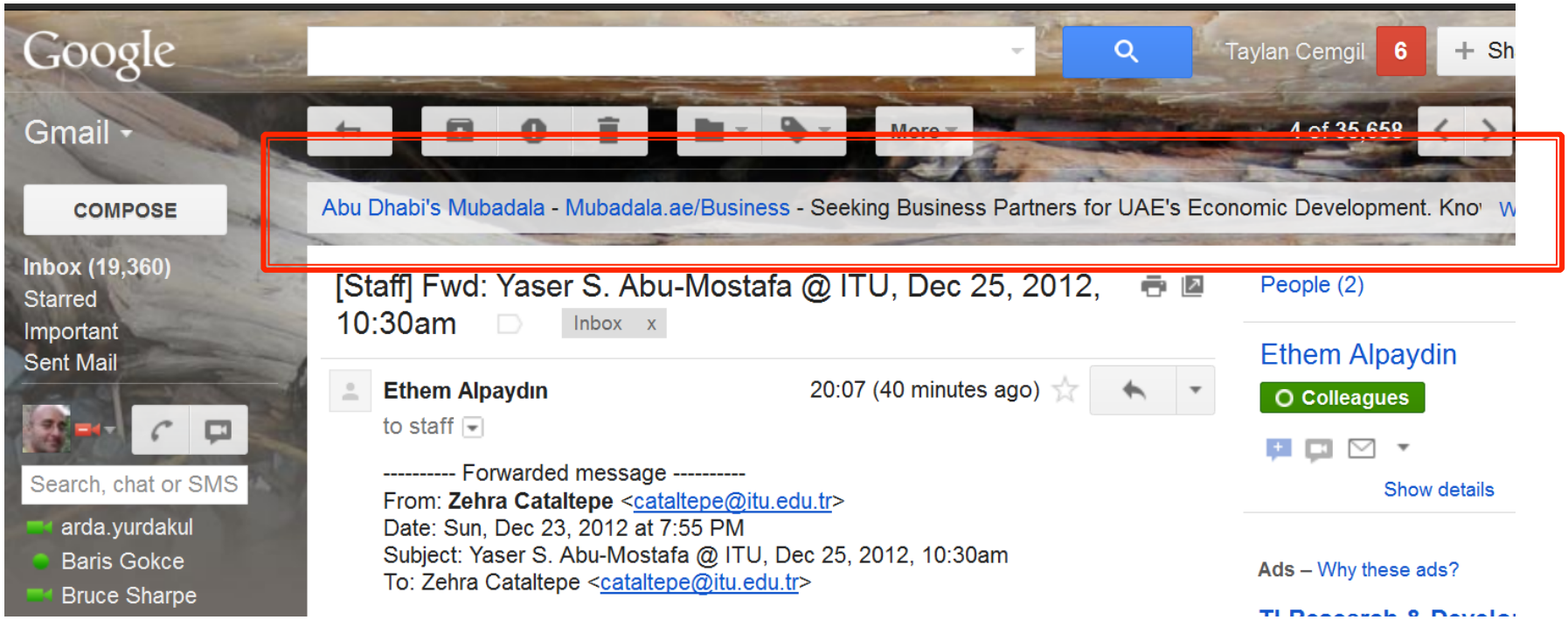
Spor Analitiği (Exatech)



- Oyuncu Takibi

Kullanım Senaryosu örnekleri

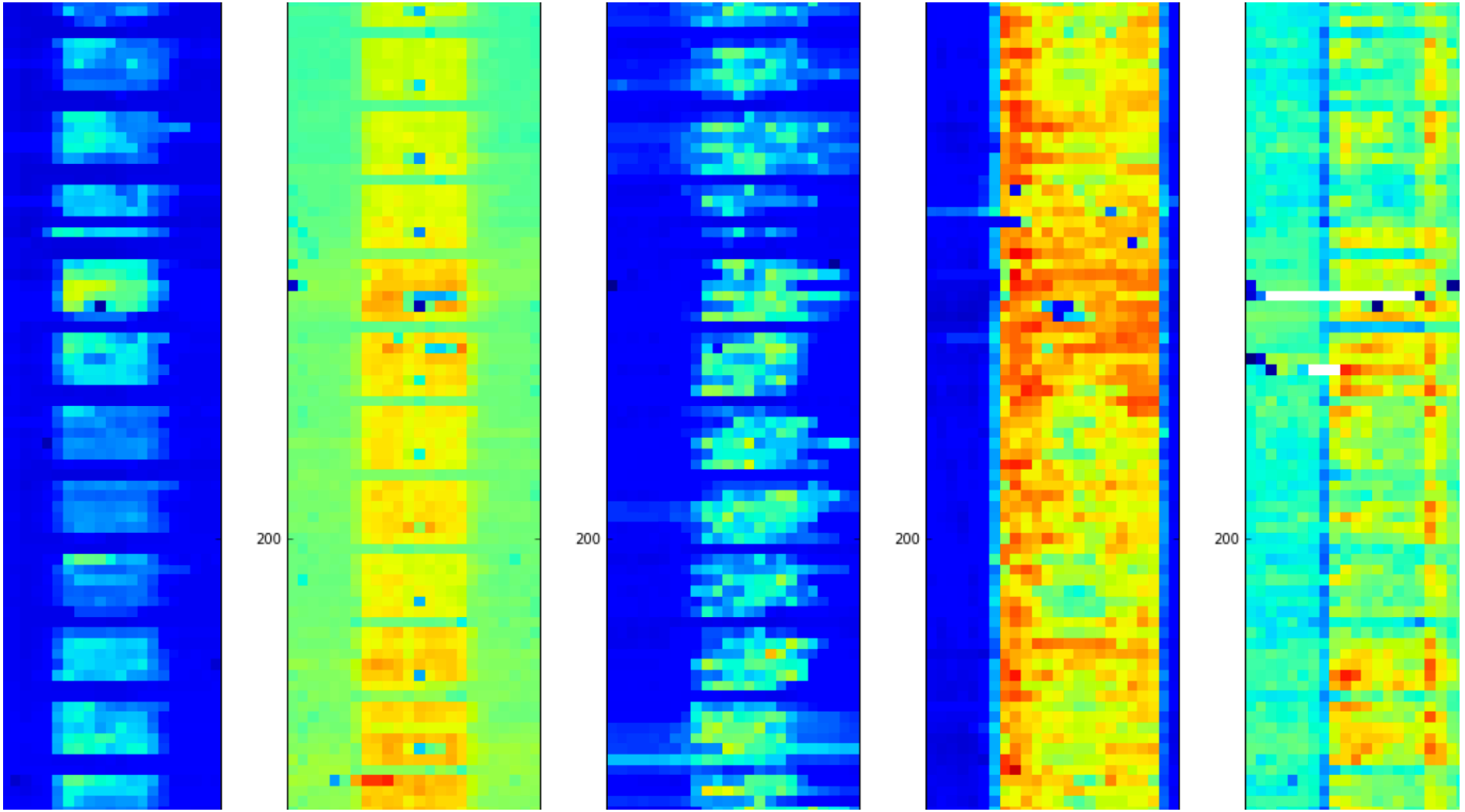
- Reklam Kişiselleştirme
  - Google ve Yahoo'nun temel gelir kaynağı

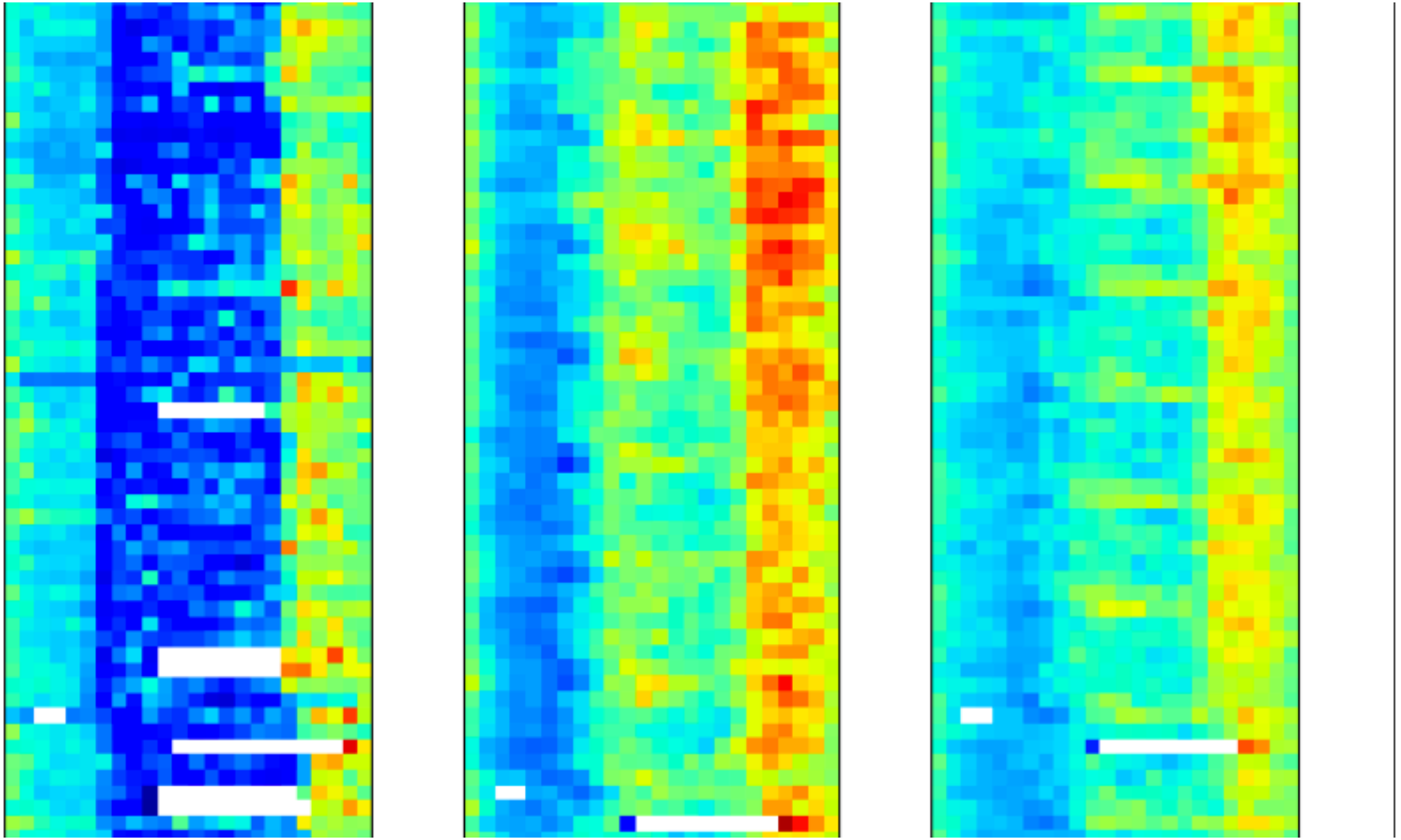


Kullanım Senaryosu: Kamu Yönetimi

- Trafik Yönetimi (Urban Traffic Management)
- Enerji Dağıtım şebekesi yönetimi/eniyilemesi (Energy Grid Management/Optimization)
- Power Generation Management
- Çevre gözlemleme (Environment Monitoring)

Enerji

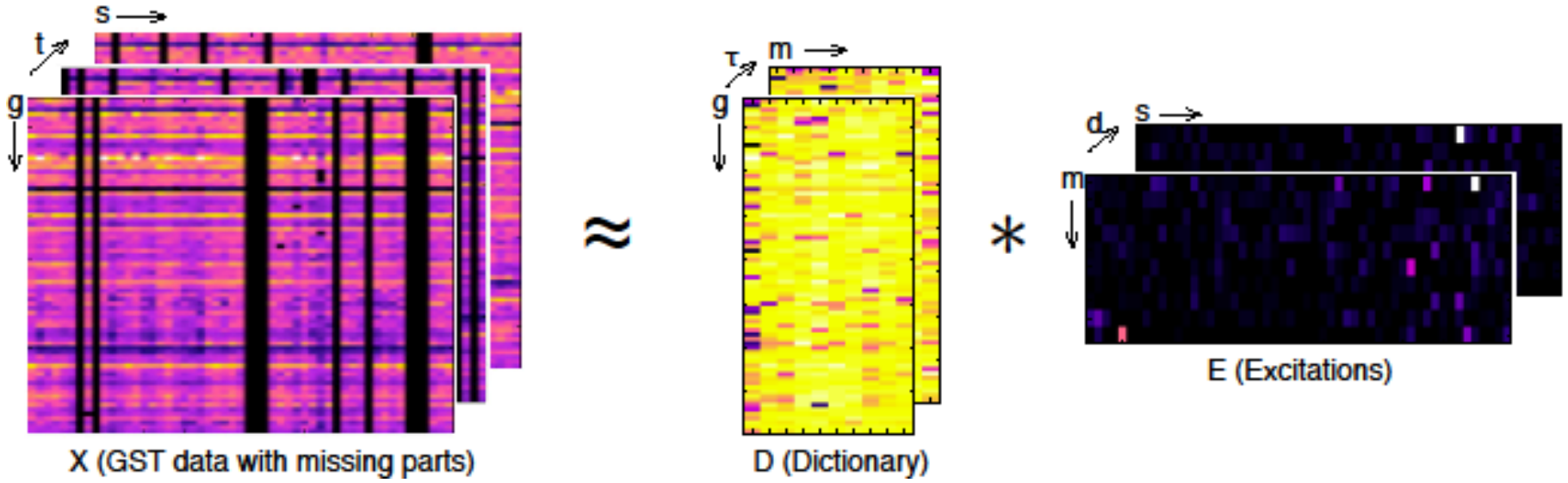




Sağlık/Yaşam Bilimleri ve Biyoloji

- Diagnosis and Medical Expert systems
- Health Insurance fraud detection
- Patient care quality and program analysis
- Drug discovery
- Remote Monitoring

3-boyutlu Microarray verisi analizi



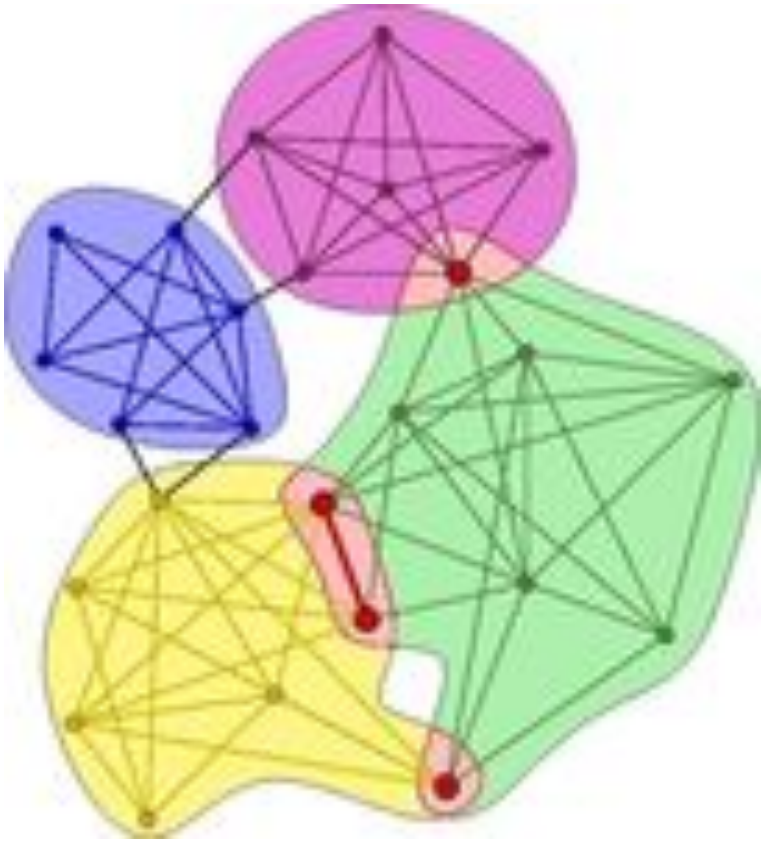
- $X(gen, örnek, zaman)$

Kullanım Senaryoları: Web

- Klik akısı öbekleme ve analizi (Clickstream Segmentation and Analysis)
- İlan hedefleme/seçim/tahmin/eniyileme (Ad Targeting/Selection)
- Klik Yolsuzluğu/Engelleme (Click Fraud Detection/Prevention)
- Sosyal Ağ Analizi
- Müşteri Bölütlemesi
- Newsgroup/Blog/Sosyal Medya gündem takibi

Çizge/Ağ Analizi

- Sosyal Ağlarda gruplanmalar (source: matlab exchange)

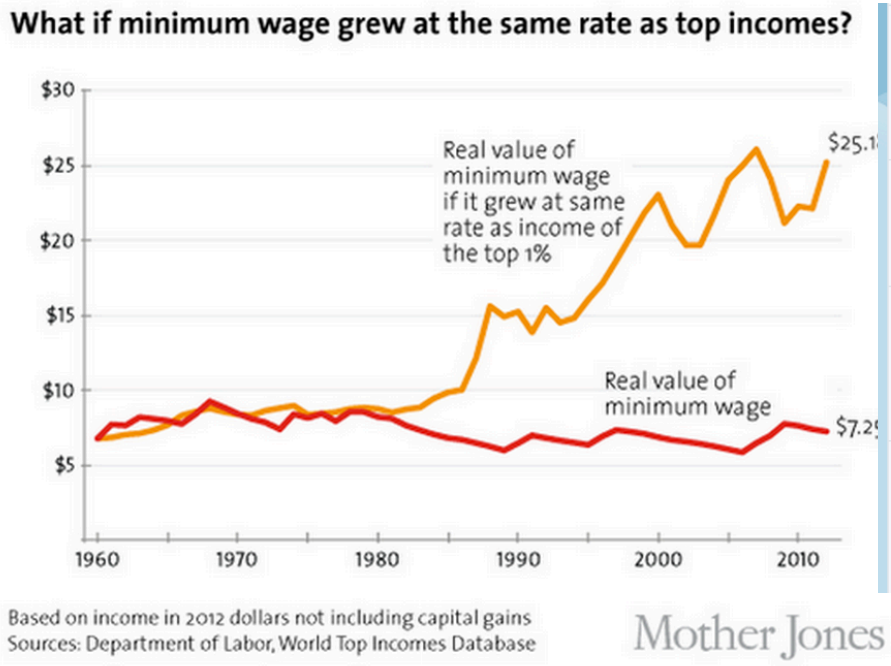


Duygu, İlgi, Eğilim Tahmini

Conrad Hackett @conradhackett · 5h

Minimum wage would be $25/hr if it grew like income of the top 1%

motherjones.com/politics/2013/…



RETWEETS 160 FAVORITES 82

Pivot @jeffposter1211 · 5h

@conradhackett somebody ought to paint an elaborate picture of what a beautiful world we would have if we applied that math

Emmanuel Acain @CitizenCainII · 4h

@conradhackett a good way to share that surplus value.

NikFromNYC @NikFromNYC · 4h

@conradhackett But would the top income, redistributed, at all cover the $25 minimum wage? No! So your implication is moot.

N647 @night647 · 4h

@conradhackett is interesting you can see the trend picking up around 1985 wonder the conditions that produced this situation.

Doğal Dil İşleme

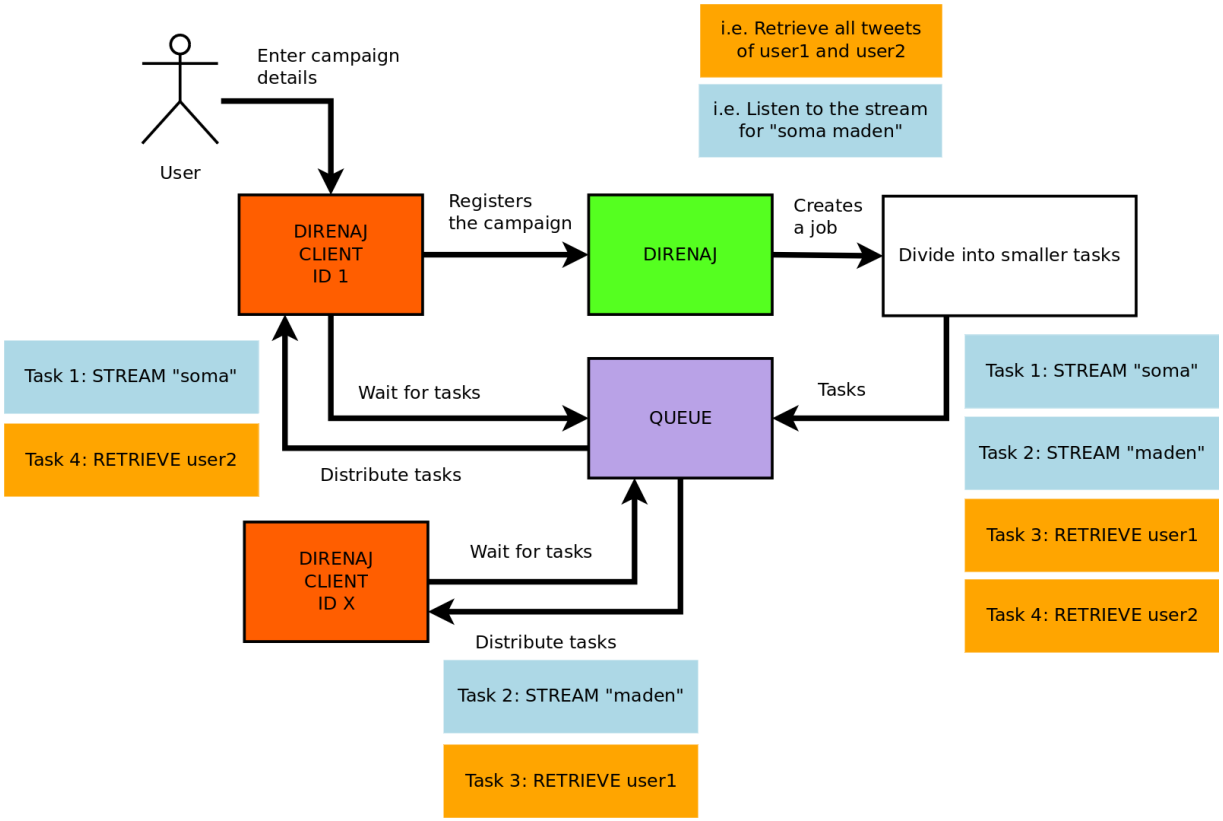
Its a btf nite, lukin for smth fun to do,
I think I wanna be w ma frnds.

Its a beautiful night, looking for something fun to do,
I think I want to be with my friends.

Dun aksam okulun sıtesı gene cokmustü.

Social Text Normalization, Çağıl Uluşahin ve Arzucan Özgür

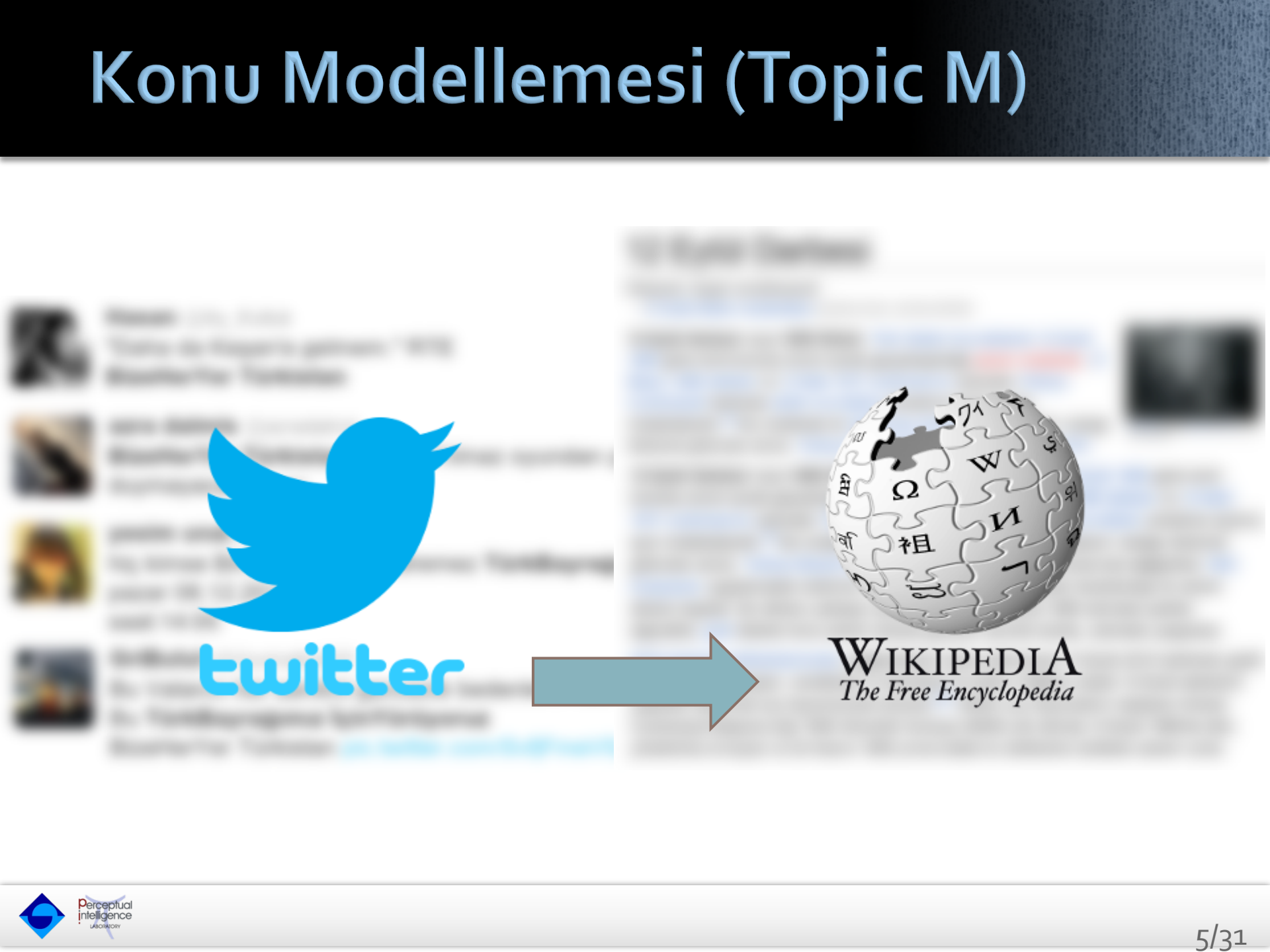
Veri Toplama, Hesaplama ve Yazılım altyapıları



Catch Tweet streams
Crawl and Track Friend-Follower Graphs

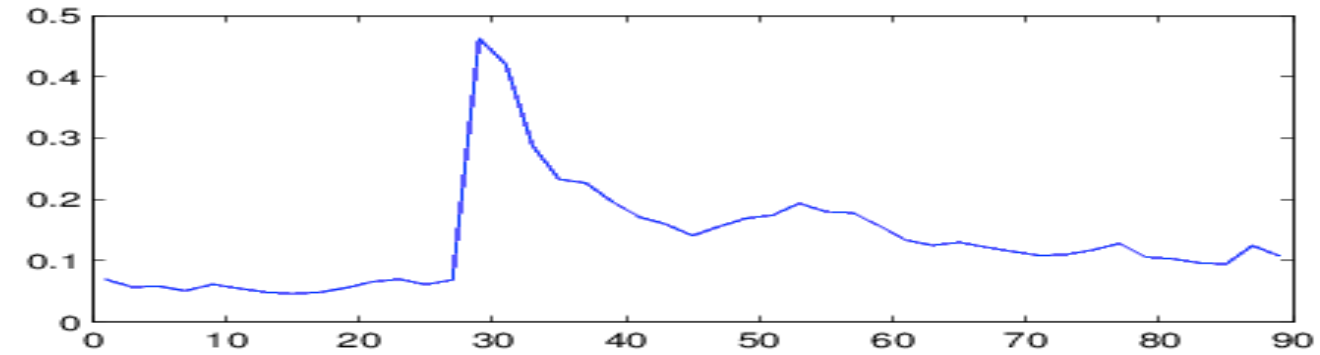
Görselleştirme

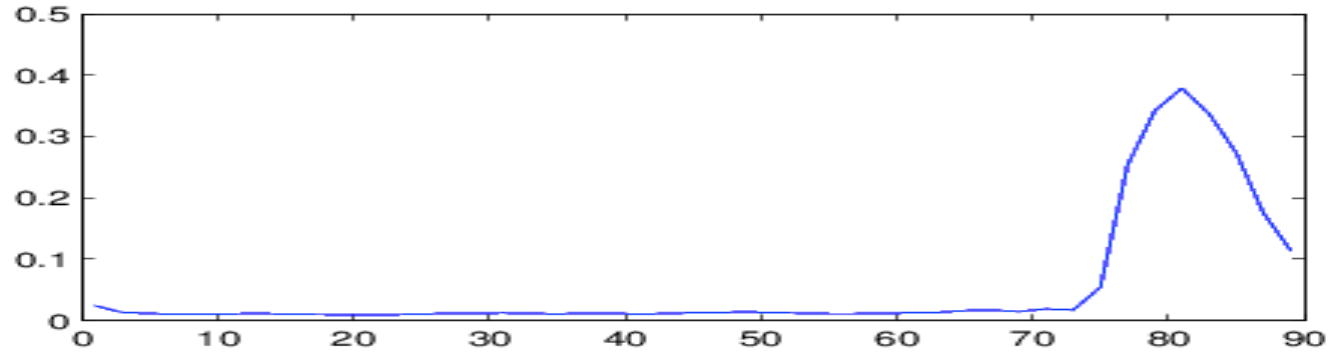
Konu Modellemesi (Topic M)

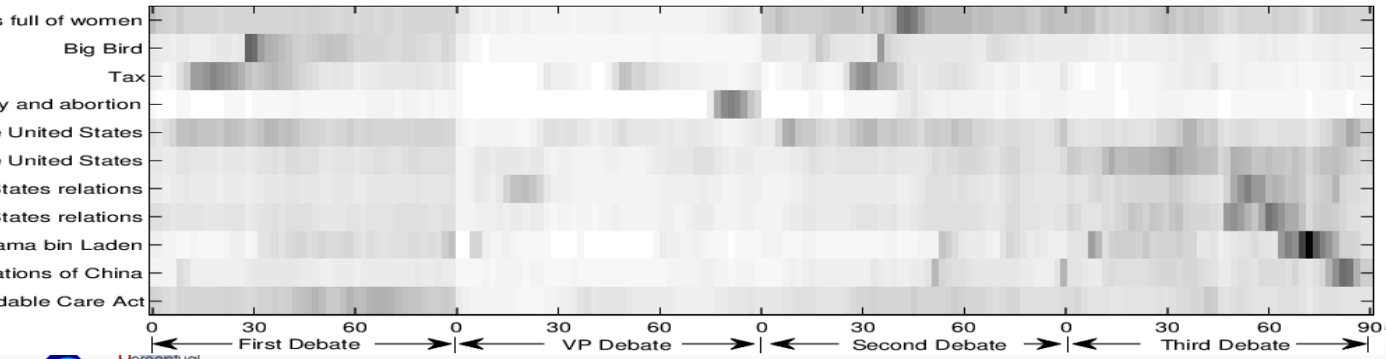


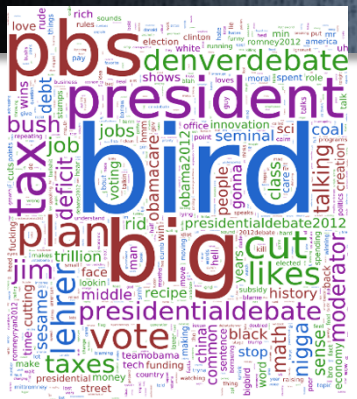
Dinamik Gündem Takibi

Yıldırım ve Üsküdarlı



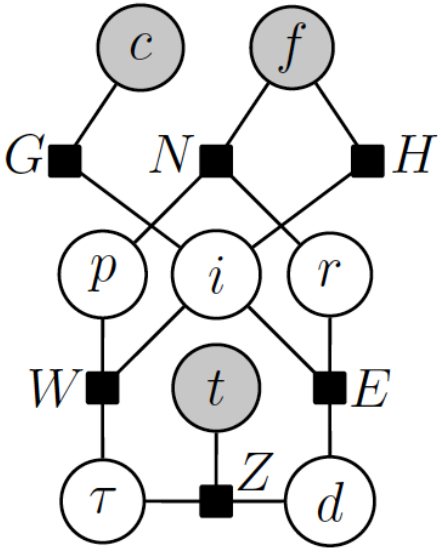






Modern Yapay Öğrenme

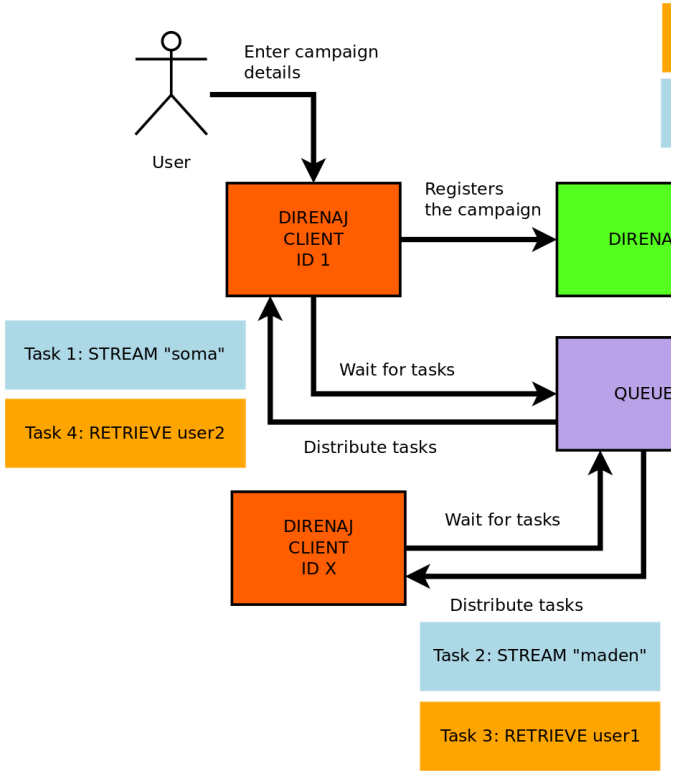
- Modeller – Algoritmalar -- Sistemler



Algorithm 1: quad

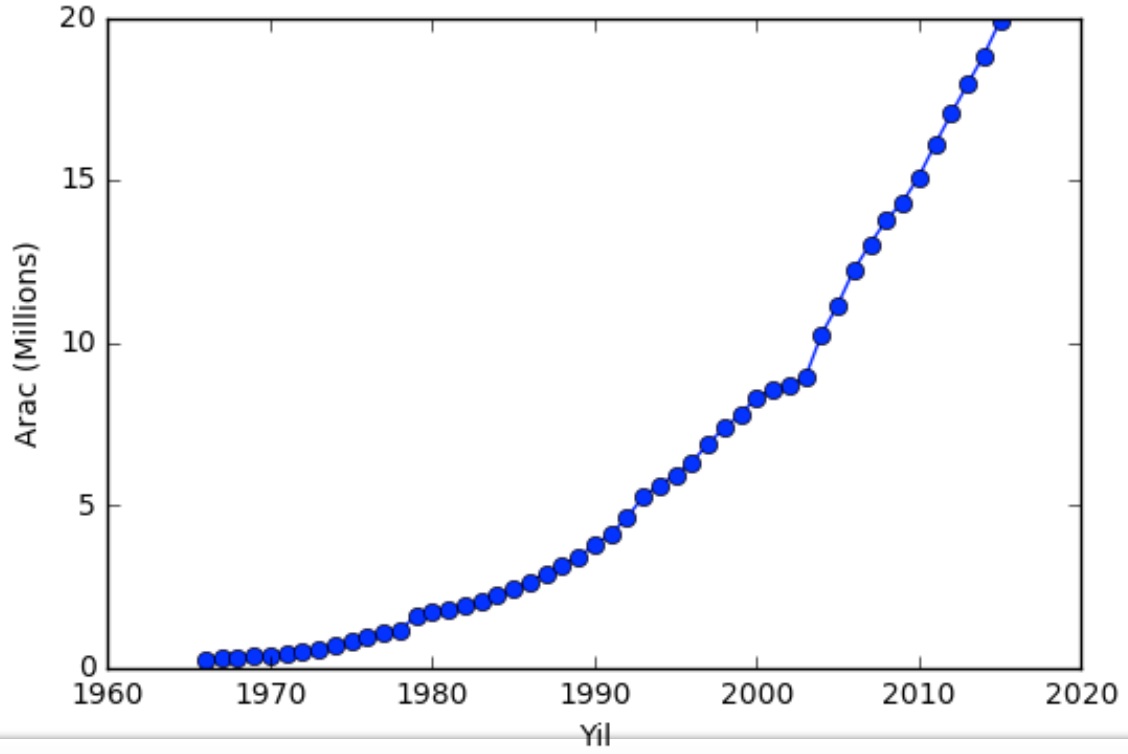
input: z_0, β_1

for $t = 0, 1, 2, \cdots$ **do**
  $x_1 = z_t$
  Compute H_t
  for $k = 1, 2, \cdots,$
    Choose a subs
    Compute ∇_{S_k}
    $x_{k+1} = \arg$ mi
  end
  $z_{t+1} = x_{c+1}$
  Set $\beta_{t+1} \leq \beta_t$
end



Güdümlü Öğrenme: Regresyon

i	Araç Sayısı (y)	Yıl (x)
1	231977	1966
…		
49	19882069	2015



Güdümlü Öğrenme: Regresyon

i	Araç Sayısı (y)	Yıl (x)
1	231977	1966
…		
49	19882069	2015

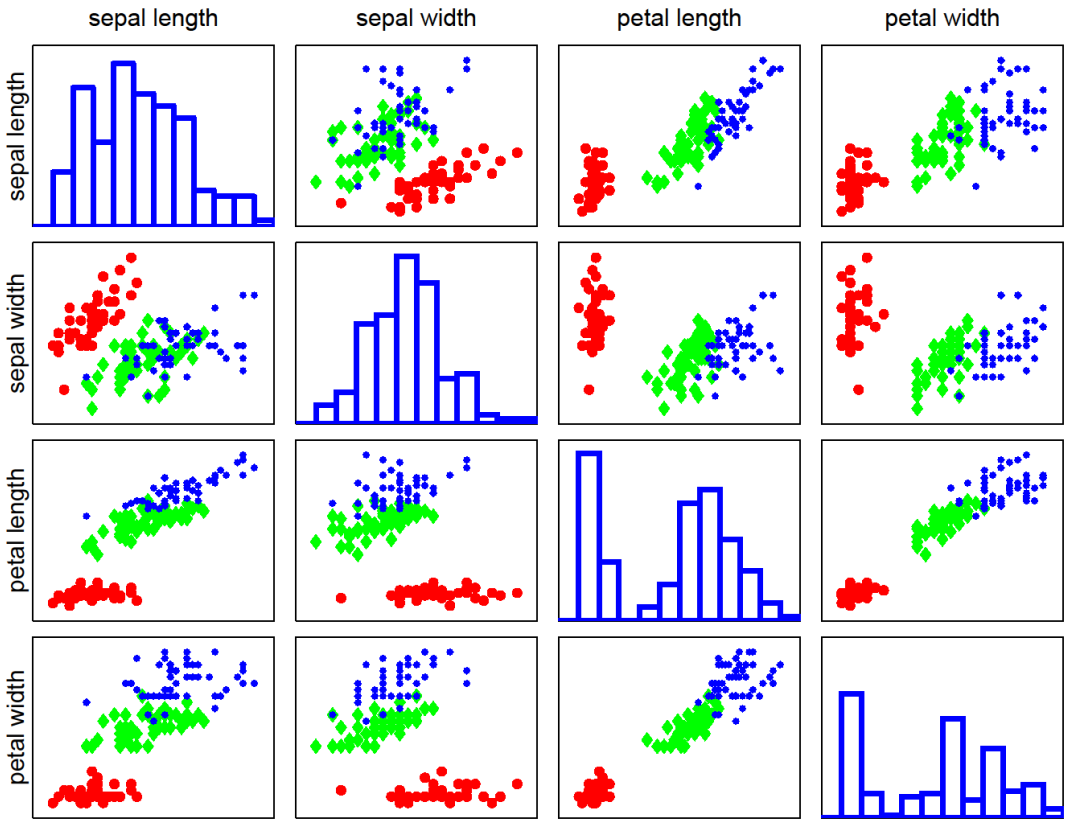
$$y \approx w_1 x + w_0$$

$$y \approx w_2 x^2 + w_1 x + w_0$$

$$y \approx f(x; w)$$

Güdümlü Öğrenme (Supervised Learning)

- Sınıflandırma

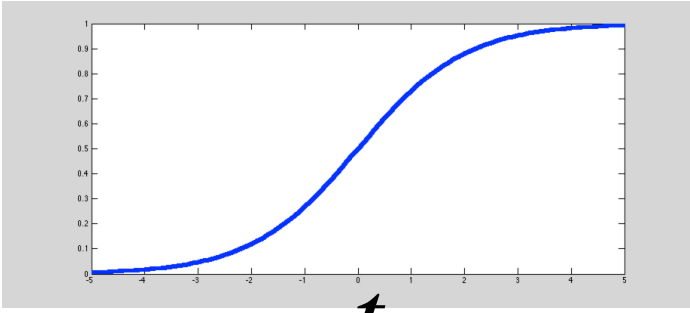


Sınıflandırma: Lojistik Regresyon

Öznitelik 1	Öznitelik 2	Öznitelik 3	Öznitelik 4	Sınıf
5.1	4.3	2.1	0.3	0
5.7	3.5	3.2	0.8	0
3.4	5.2	0.4	0.6	1
X_1	X_2	X_3	X_4	y

$$y \approx \sigma(x_1 w_1 + x_2 w_2 + x_3 w_3 + x_4 w_4)$$

$$\sigma(t)$$

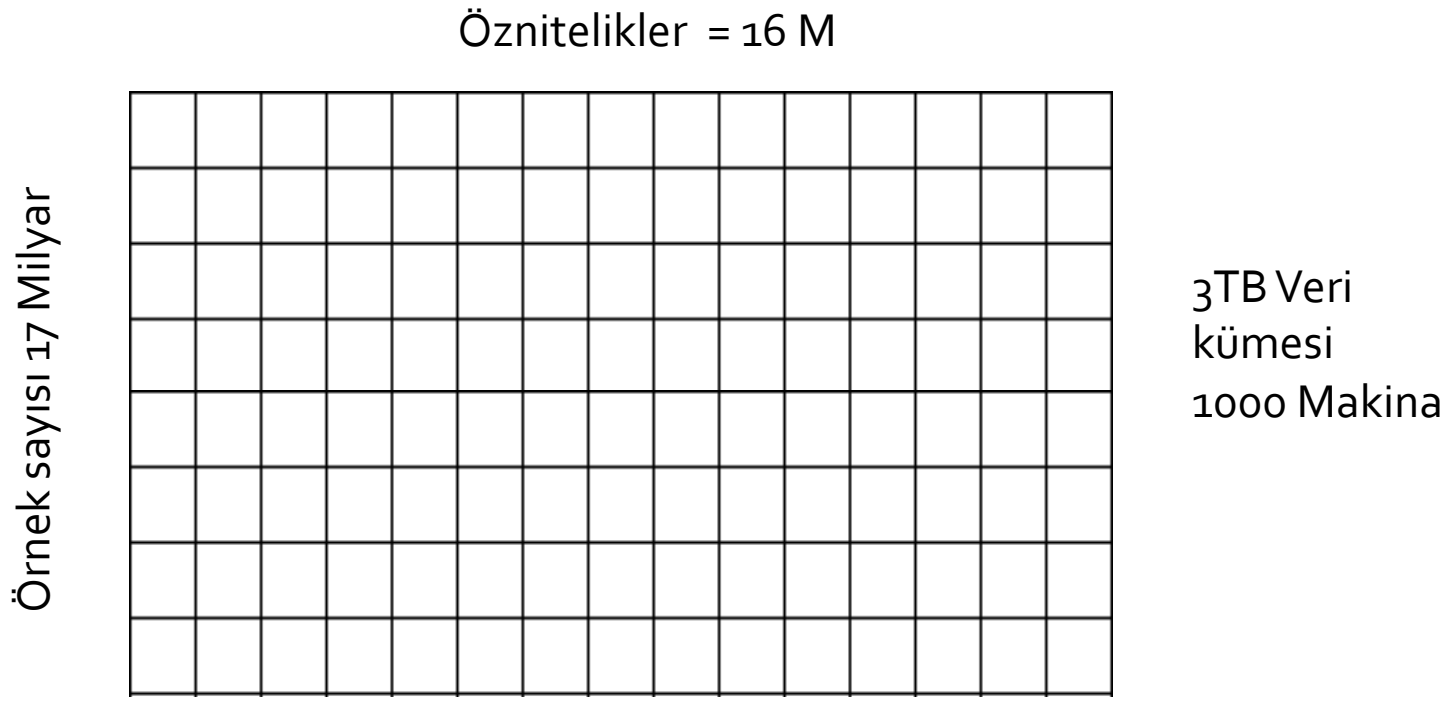


Öznitelik Mühendisliği (Feature Engineering)

- Belirli bir problem için uygun sayısal bir gösterim bulmak
- Örn: e-mail Spam/NoSpam filtering
  - X1 = [email metni 'Rolex' içeriyor mu?]
  - X2 = [email metni 'http://' içeriyor mu?]
  - X3 = email metindeki büyük/küçük harf sayıları oranı
  - …
  - X100000 = [Gönderen kişi adres defterinde var mı?]

Büyük Ölçekte sınıflandırma

- Reklam Tahmini, YouTube video sıralaması
- Bir kullanıcının bir reklamı/videoyu klikleme ihtimali nedir?
- A Reliable Effective Terascale Linear Learning System, Agarwal et.al. 2012



Algoritma

1. Her düğümde sıralı öğrenme kullanarak bir parametre bul
2. AllReduce kullanarak ortalama hesapla
3. Her düğümde, gradyan toplamlarını hesapla
4. AllReduce kullanarak gradyanları topla.
5. L-BFGS kullanarak parametreleri güncelle ve 3. adıma dön

Paralel işleme Platformları (BBL2011)

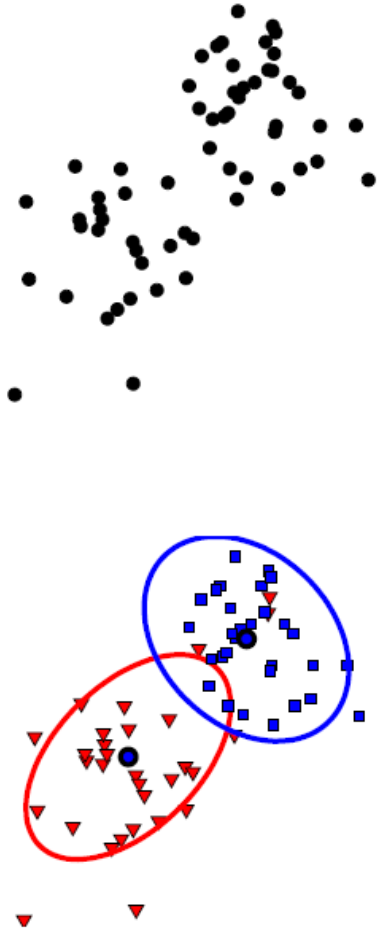
Platform	Communication Scheme	Data size
Peer-to-Peer	TCP/IP	Petabytes
Virtual Clusters	MapReduce / MPI	Terabytes
HPC Clusters	MPI / MapReduce	Terabytes
Multicore	Multithreading	Gigabytes
GPU	CUDA	Gigabytes
FPGA	HDL	Gigabytes

Slide from ICML 2011 tutorial Langford et. al.

Güdümsüz Öğrenme

- Öbekleme (Clustering)
- Boyut Düşürme (Dimensionality Reduction)
- Göreselleştirme (Visualization)

Öbekleme



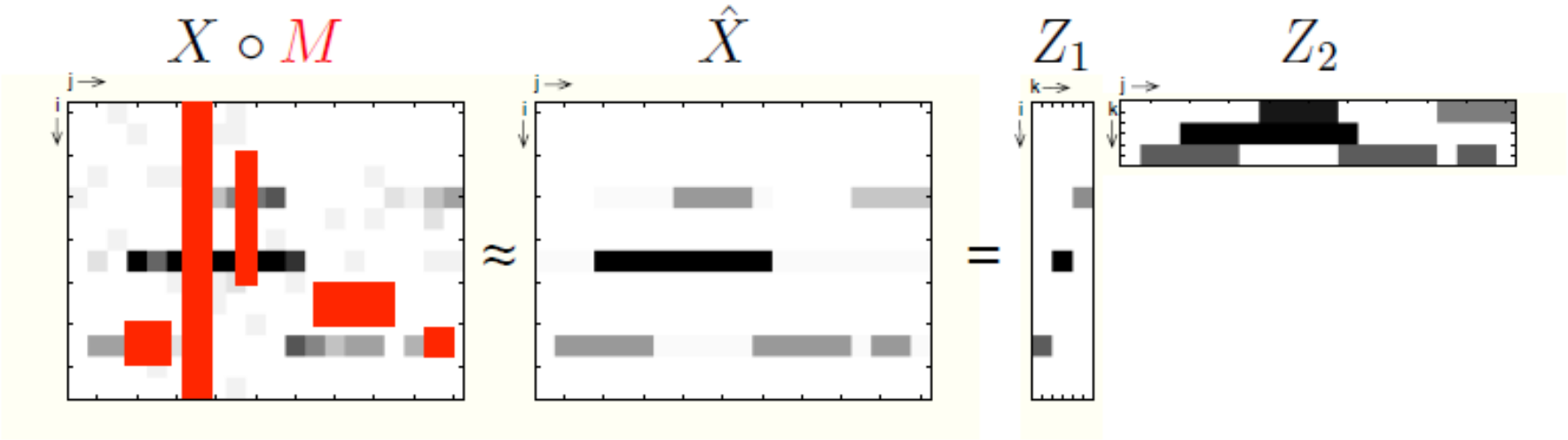
Boyut Düşürme

- Kelime-Kitap

	j.caesar	hamlet	othello	macbeth	rom&jul	sonnets
caesar	270	2	1	1	0	0
brutus	379	1	0	0	0	0
malcolm	0	0	0	60	0	0
muse	0	0	1	1	0	16
⋮						
love	34	68	80	19	150	195
friend	23	14	18	5	13	16
the	610	1148	759	733	682	446
traitor	1	0	0	5	1	0
traitors	9	0	1	3	0	0
⋮						
napkin	0	1	3	0	0	0
sword	15	16	10	14	8	1
laptop	0	0	0	0	0	0

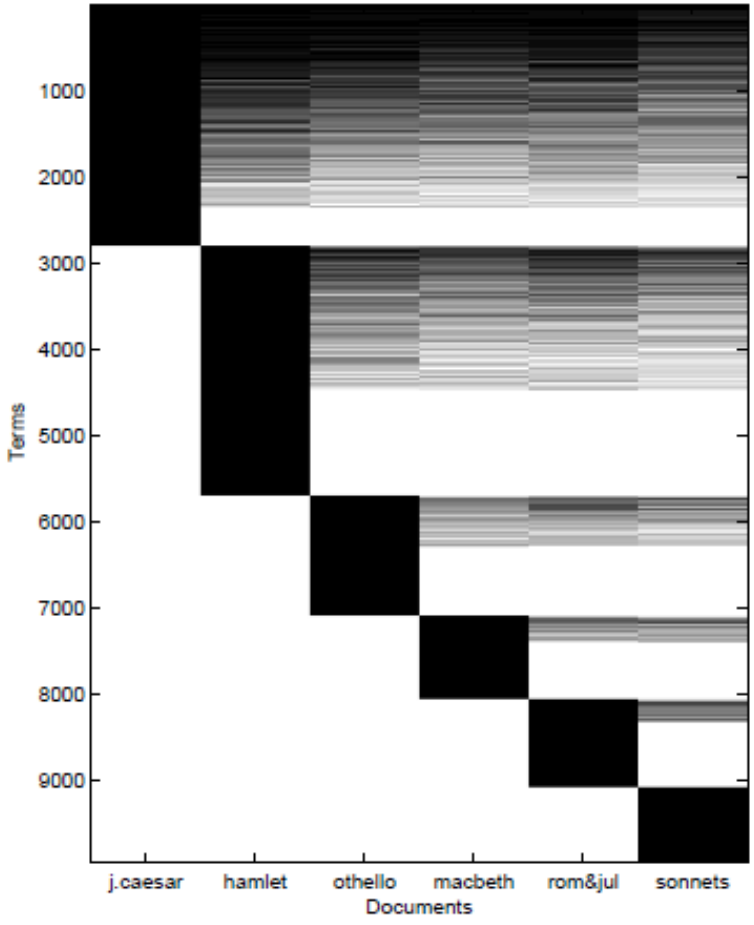
Matrix Factorizations

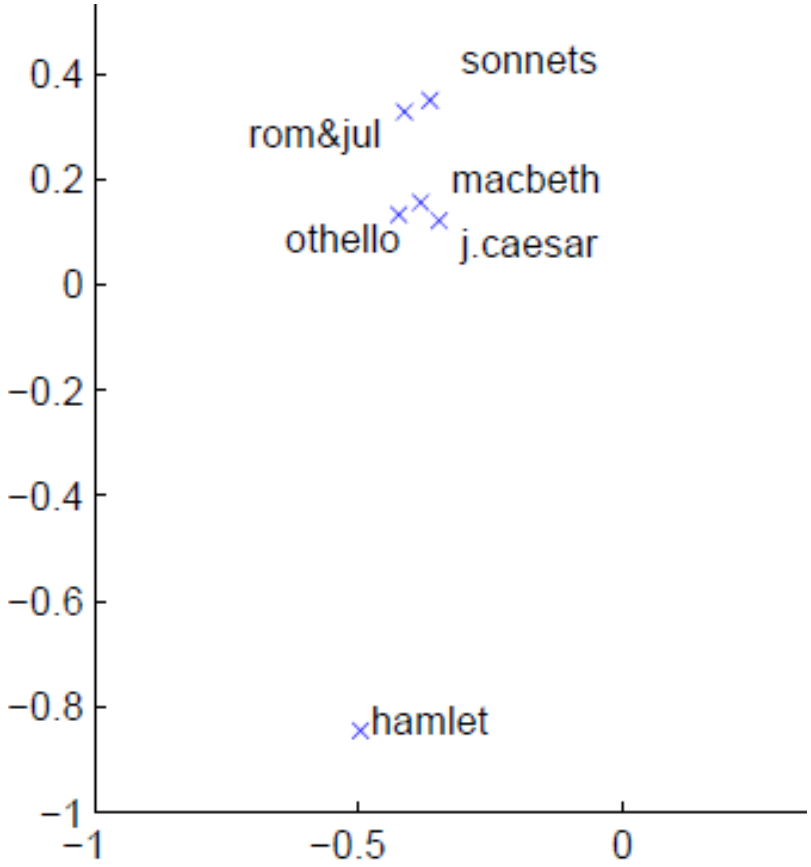
$$X(i,j) \approx \sum_k Z_1(i,k) Z_2(k,j)$$



$$(Z_1, Z_2)^* = \arg\min_{Z_1, Z_2} D(X||Z_1 Z_2) + \lambda R(Z_1, Z_2)$$

Kelime Dokuman Matrisi

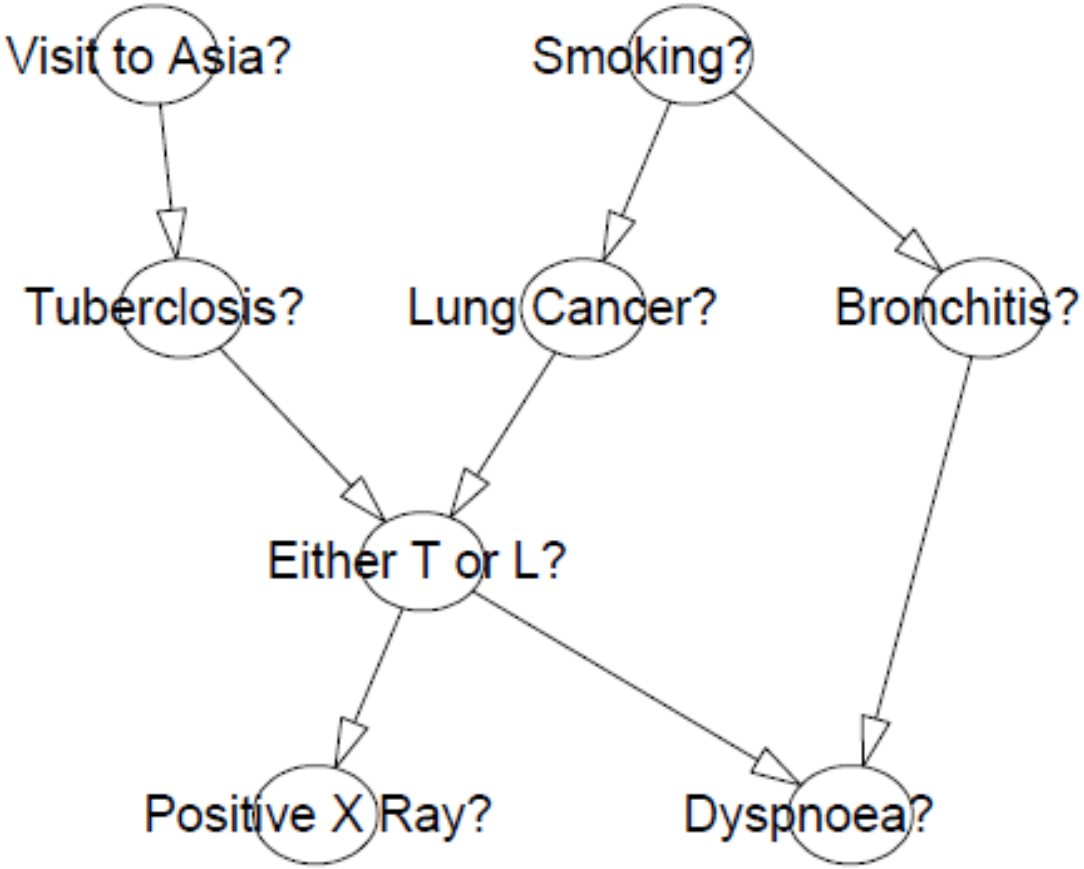




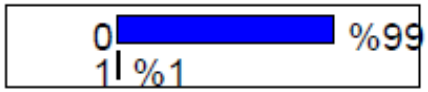
Yapay Öğrenme ve Olasılık Teorisi

- Olasılık Teorisi
  - Olasılık teorisi, sağduyunun hesaplamaya dökülmüş halinden başka bir şey değildir. – P. Laplace
  - Probability theory is nothing but common sense reduced to calculation – P. Laplace
- Grafik Modeller, Olasılıksal Uzman Sistemler
- Zaman Serileri
  - Örnek: 'Network flow' sınıflandırma

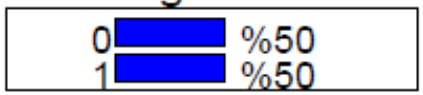
Örnek: Tıbbi uzman sistemler



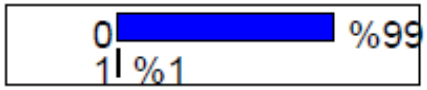
Visit to Asia?



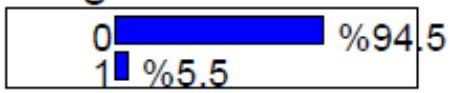
Smoking?



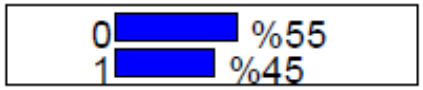
Tuberclosis?



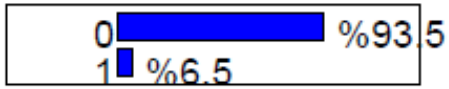
Lung Cancer?



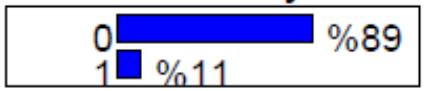
Bronchitis?



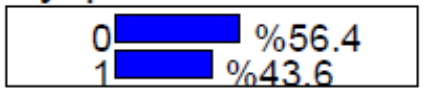
Either T or L?

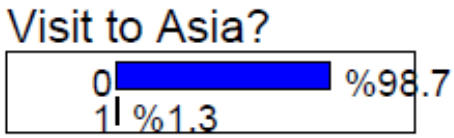


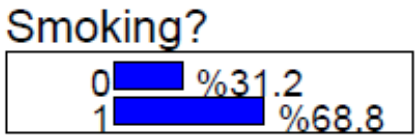
Positive X Ray?

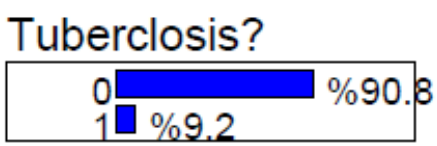


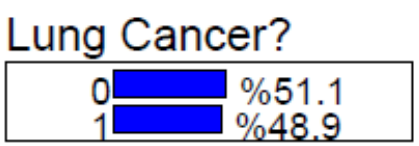
Dyspnoea?

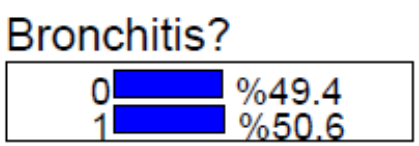


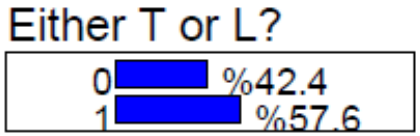

Visit to Asia?
0
%98.7
1
%1.3

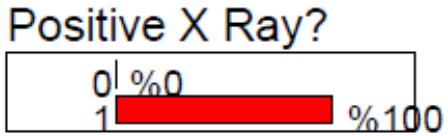

Smoking?
0
%31.2
1
%68.8

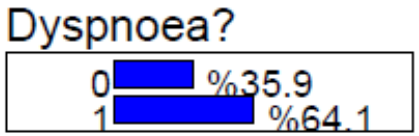

Tuberclosis?
0
%90.8
1
%9.2

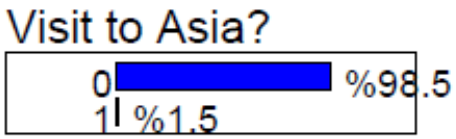

Lung Cancer?
0
%51.1
1
%48.9

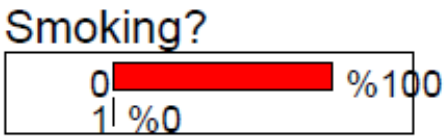

Bronchitis?
0
%49.4
1
%50.6

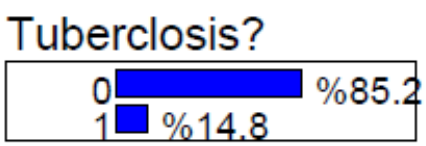

Either T or L?
0
%42.4
1
%57.6

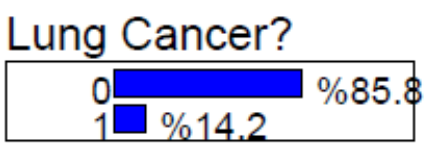

Positive X Ray?
0
%0
1
%100

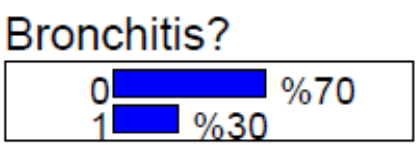

Dyspnoea?
0
%35.9
1
%64.1

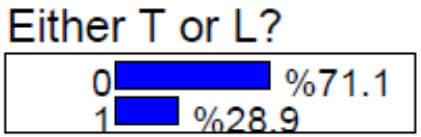

Visit to Asia?
0
1
%98.5
%1.5

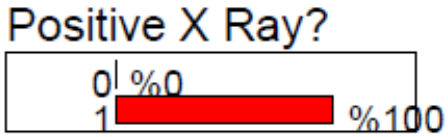

Smoking?
0
1
%100
%0

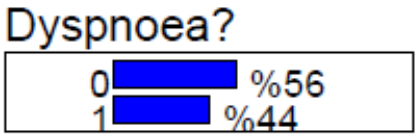

Tuberclosis?
0
1
%85.2
%14.8


Lung Cancer?
0
1
%85.8
%14.2

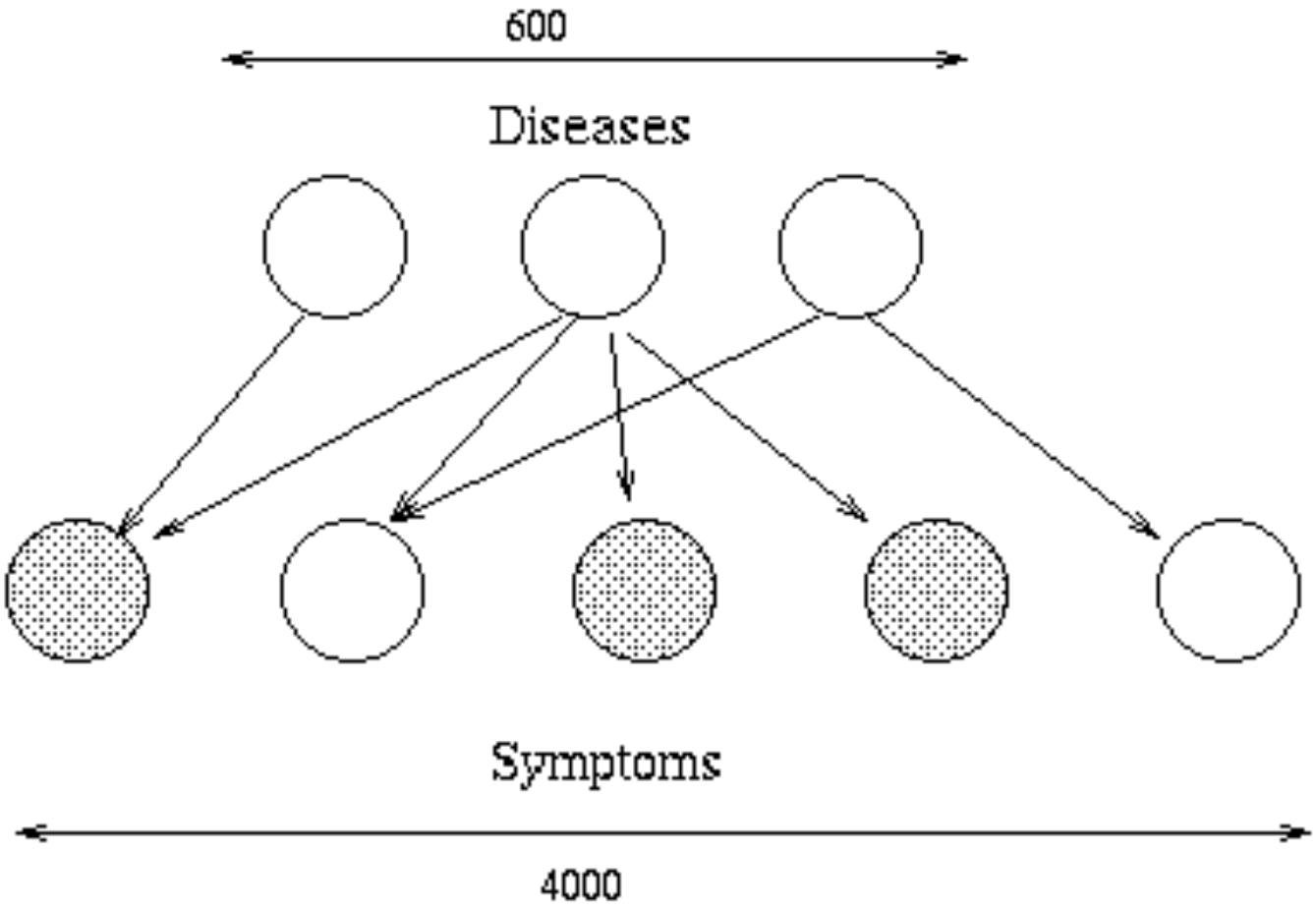

Bronchitis?
0
1
%70
%30


Either T or L?
0
1
%71.1
%28.9

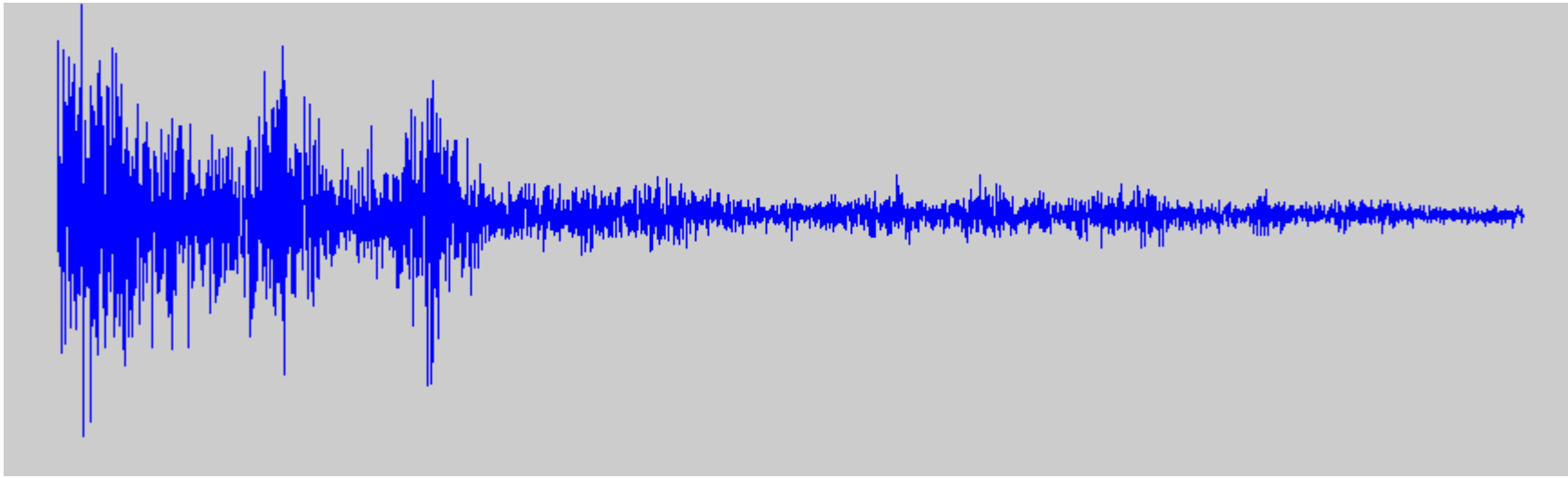

Positive X Ray?
0
1
%0
%100


Dyspnoea?
0
1
%56
%44

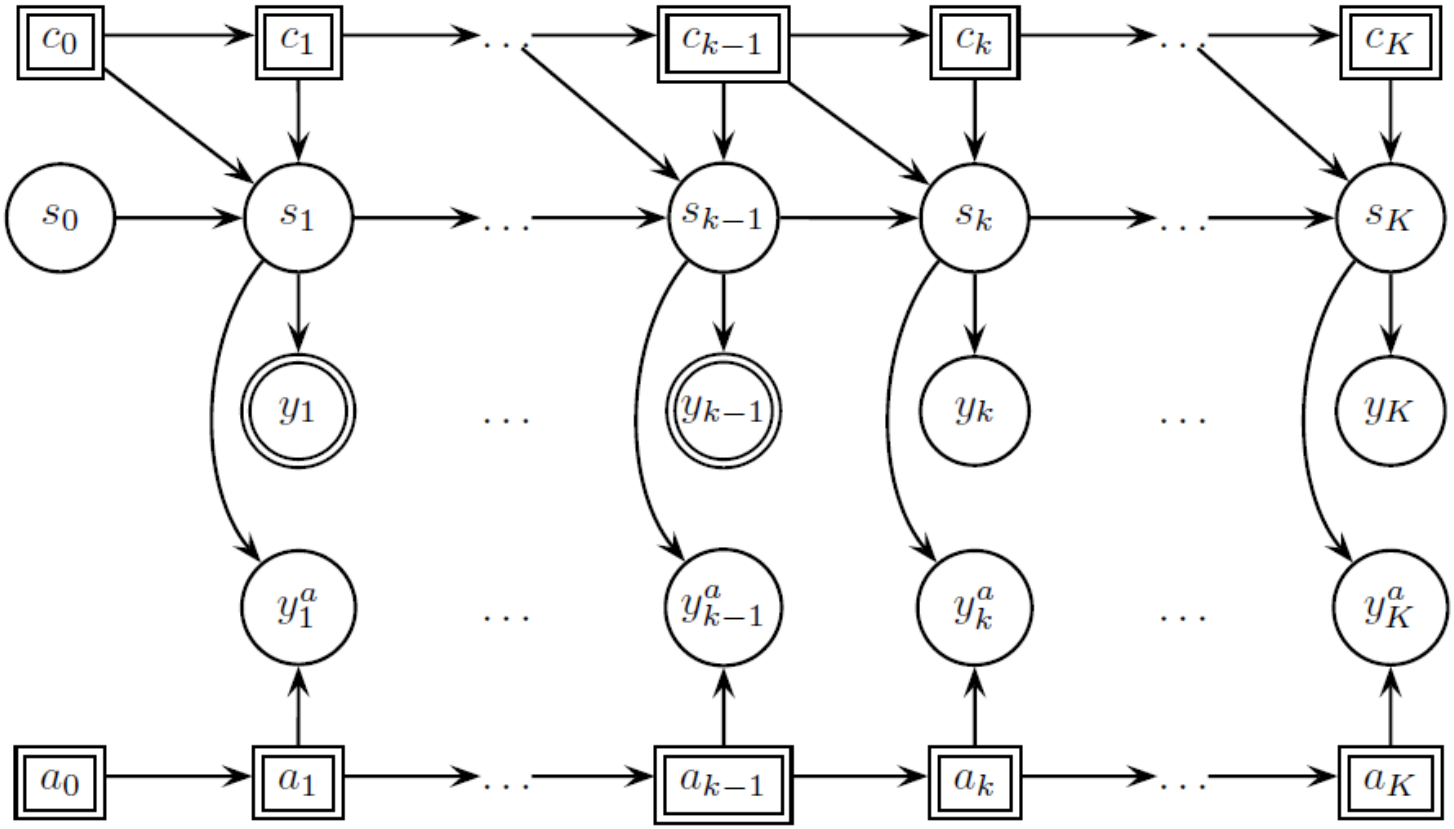
QMR-DT



Zaman Serileri

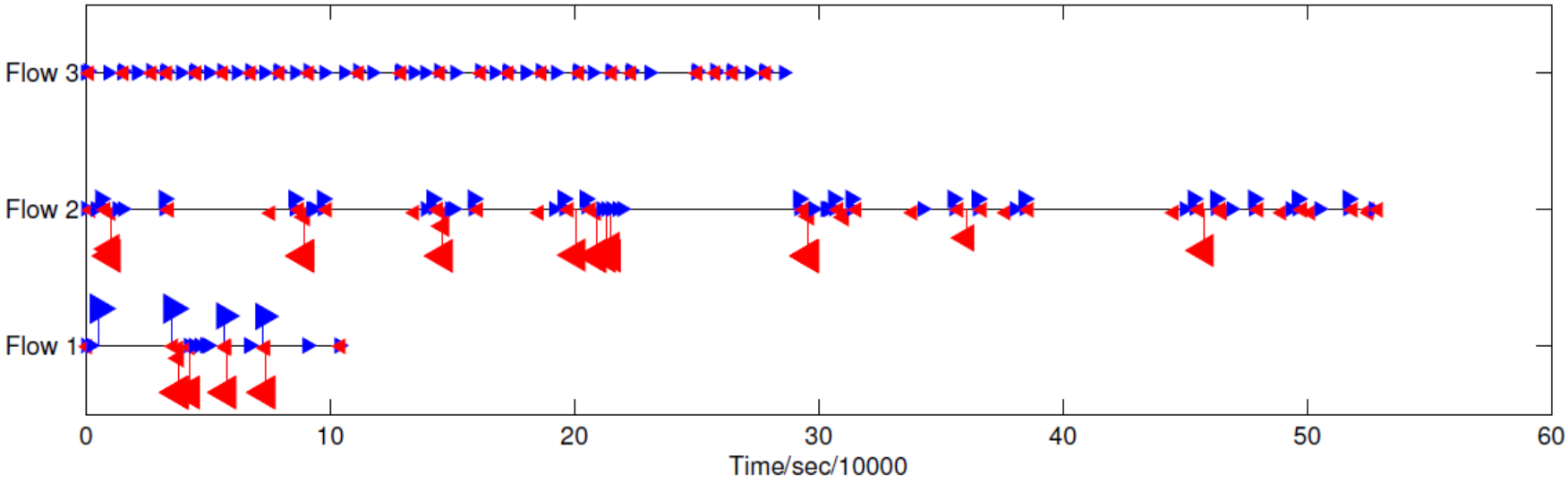


Zaman Serileri, Saklı Markov Modelleri



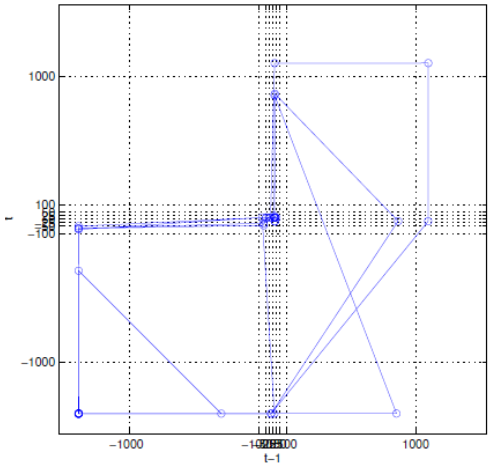
Graphical Model Through Time

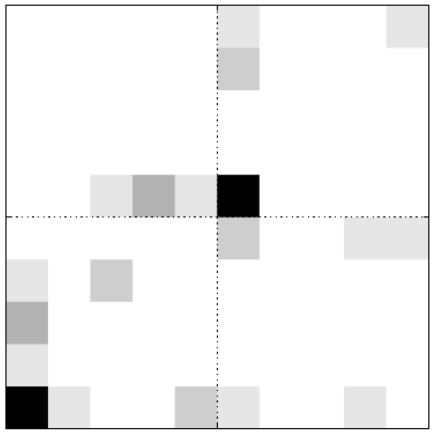
Zaman Serisi Sınıflandırma

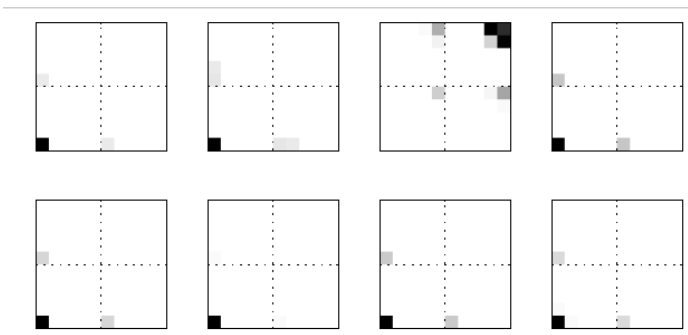


Mobil 3G kullanım örüntüleri, Uygulamalar
Deep Packet Inspection (DPI) kullanmadan
8 Saatlik Kayıt, **Anonimize edilmiş, Payload olmadan 1TB**
Kurt, Mungan, Saygun with Ericsson/Avae FP7 Mevico

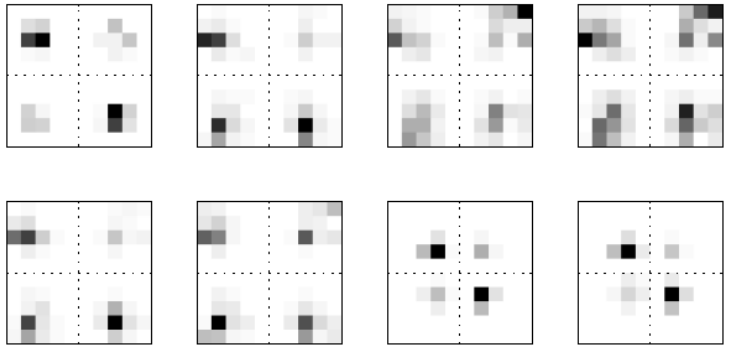
Öznitelik çıkarımı





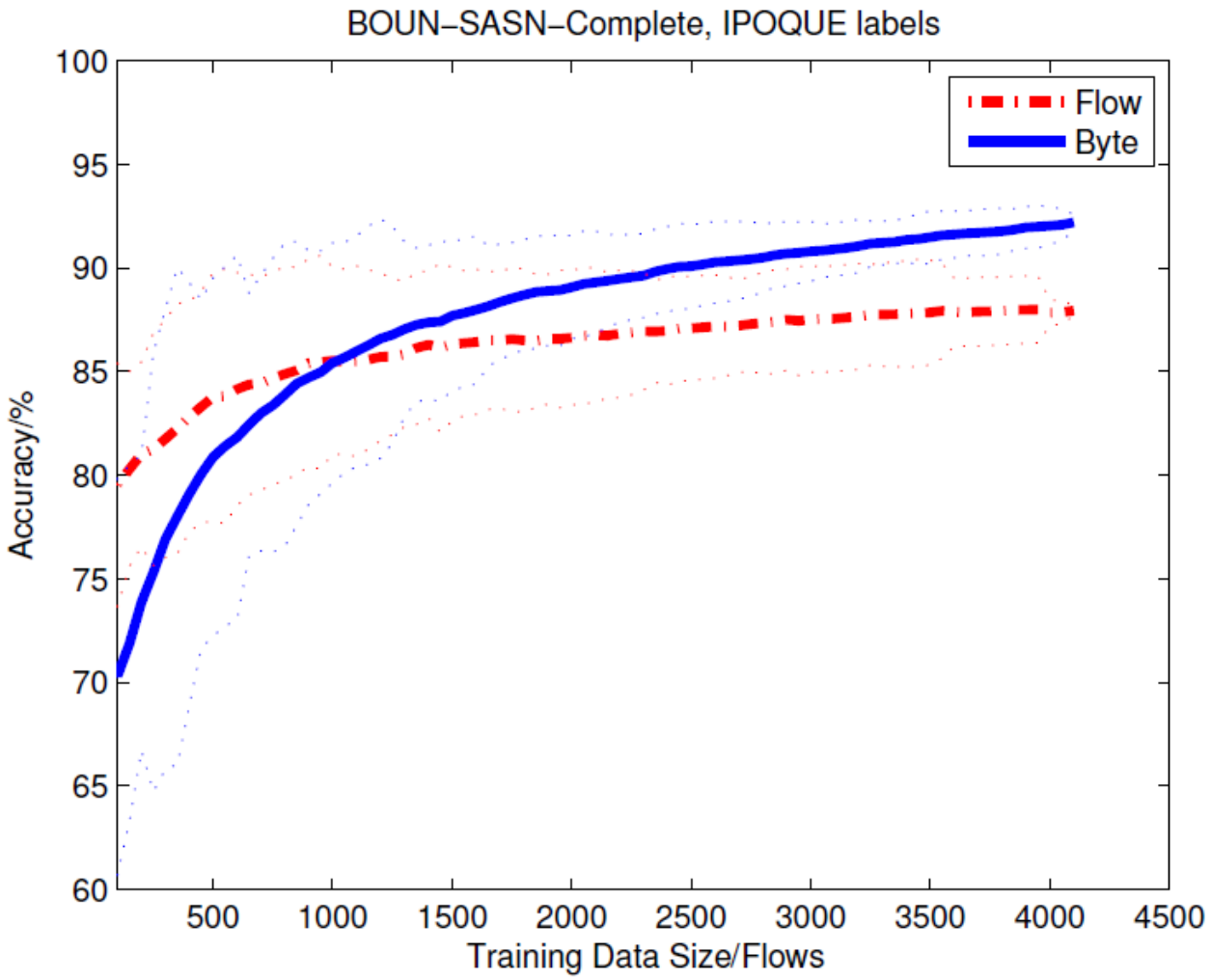


VIDEO



VIDEO2

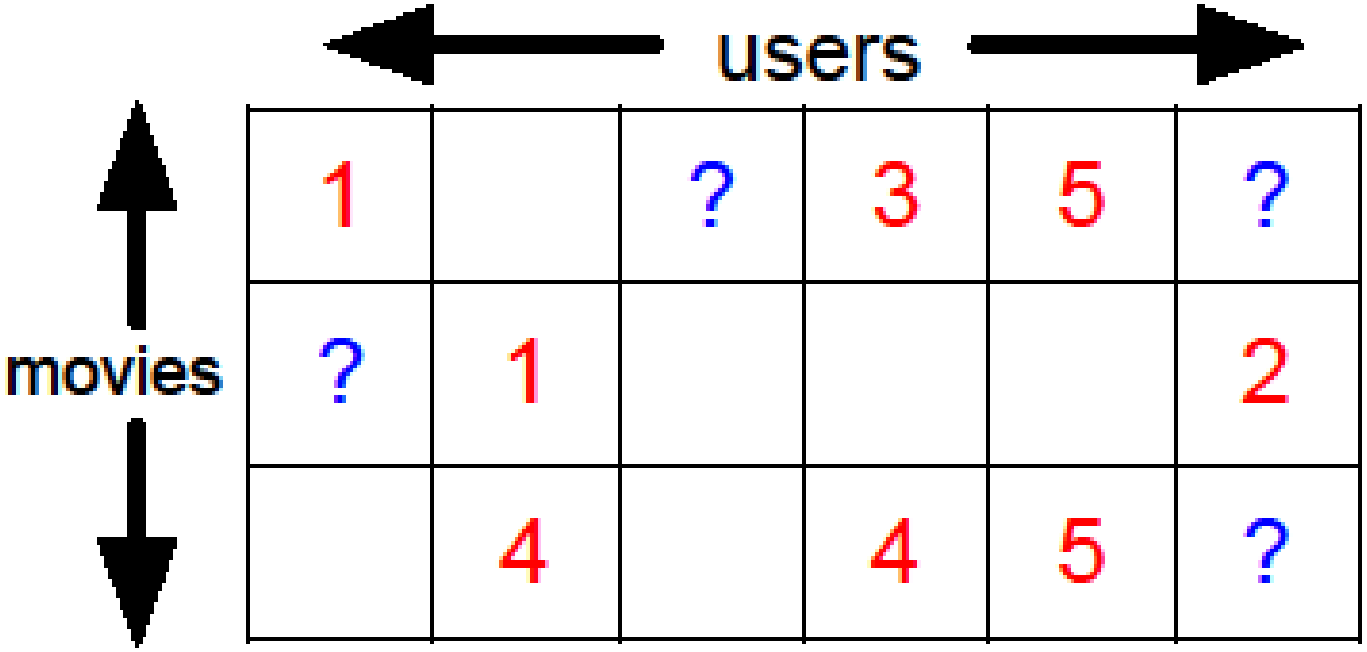
Training Data Size - Accuracy



Tavsiye Sistemleri

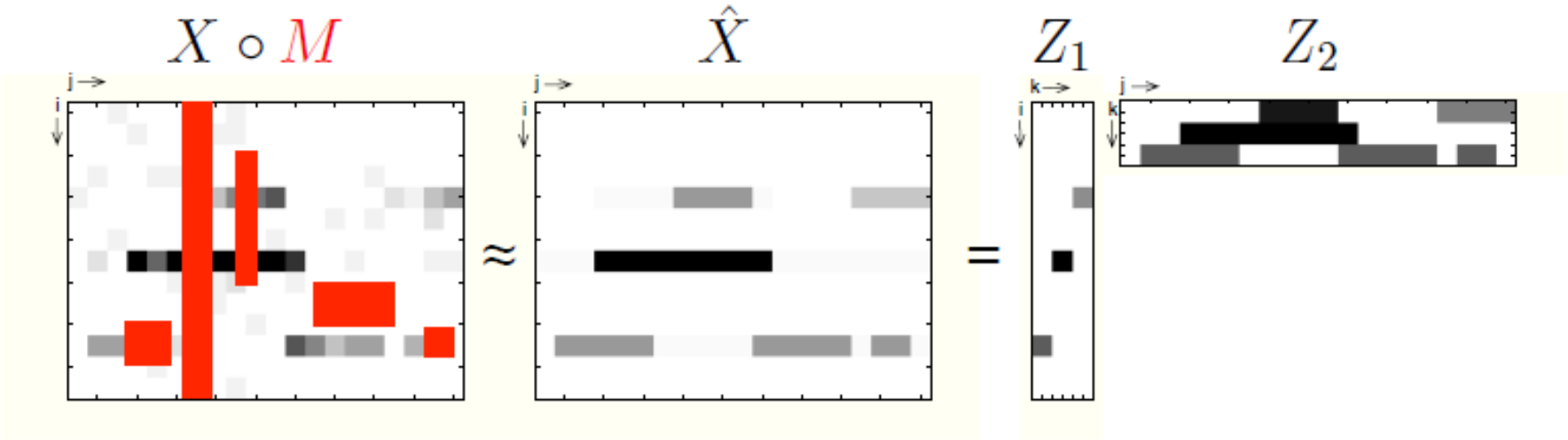
Matris tamamlama

- Netflix: 18K film × 500K kullanıcı %99 seyrek



Matris ve Tensor Ayrıştırma

$$X(i,j) \approx \sum_k Z_1(i,k) Z_2(k,j)$$



$$(Z_1, Z_2)^* = \arg\min_{Z_1,Z_2} D(X||Z_1 Z_2) + \lambda R(Z_1, Z_2)$$

Tavsiye Sistemleri

	1	?	3	4
	2	4	6	8
	1.5	3	?	6.1

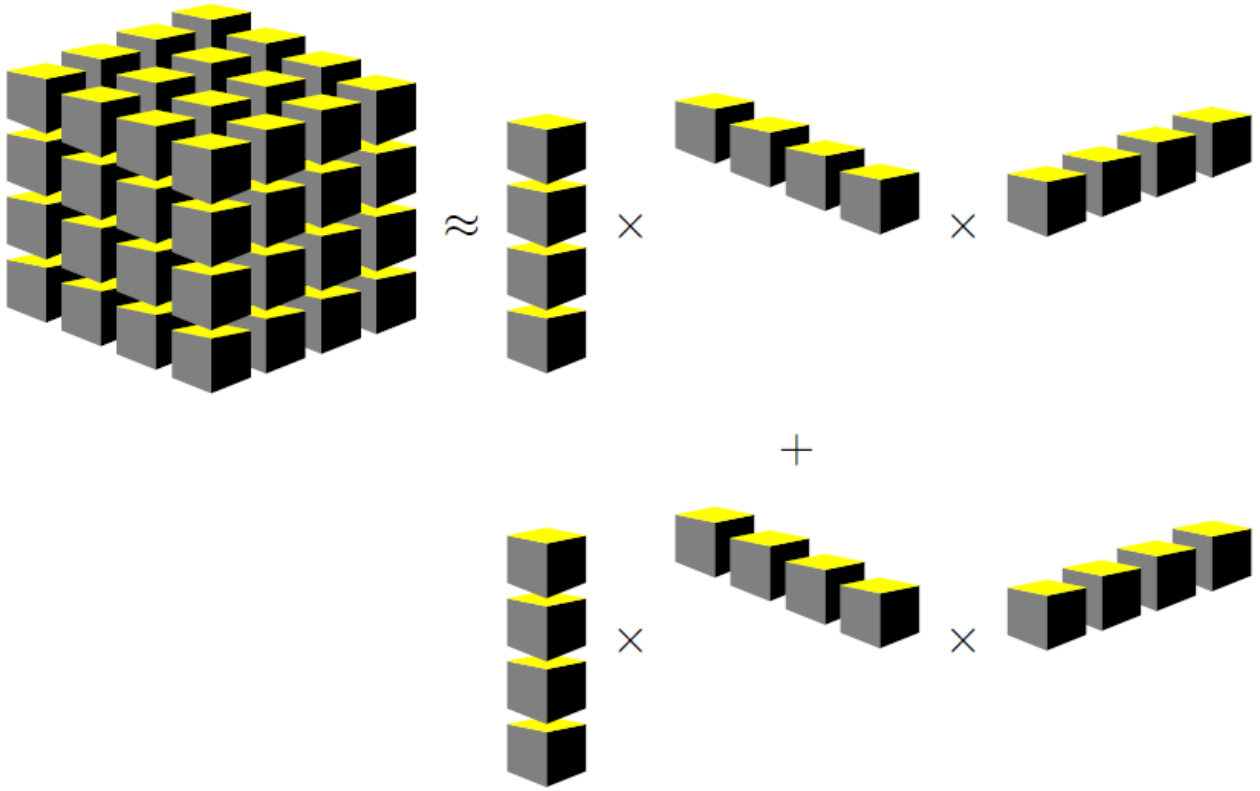
Tavsiye Sistemleri: Öğrenme

	1	2	3	4
1	1	?	3	4
2	2	4	6	8
1.5	1.5	3	?	6.1

Tavsiye Sistemleri

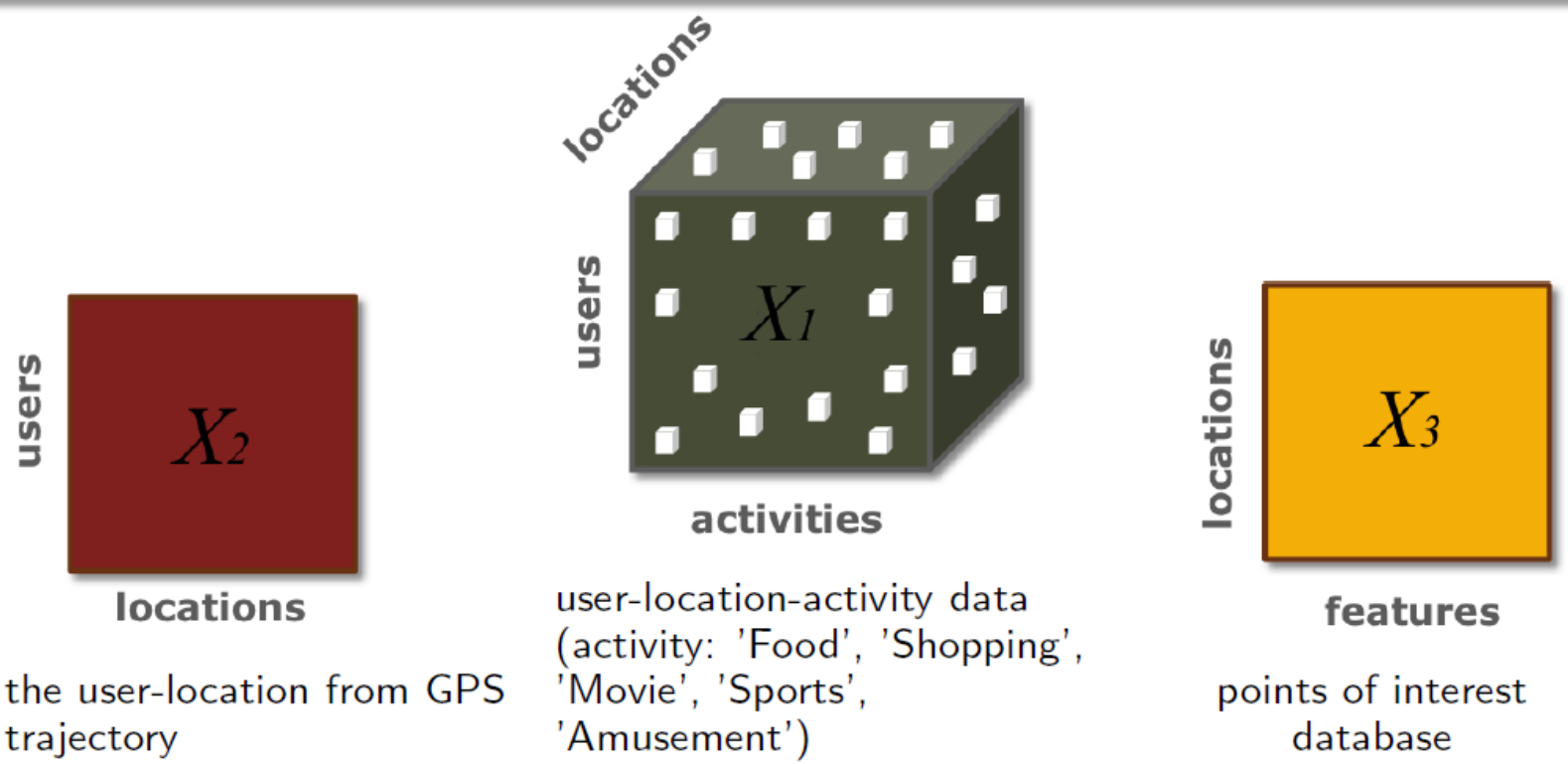
	1	2	3	4
1	1	2	3	4
2	2	4	6	8
1.5	1.5	3	4.5	6.1

Tensor Ayrıştırma



$$X(i,j,k) \approx \sum_r Z_1(i,r) Z_2(j,r) Z_3(k,r)$$

Veri Birleştirme (Sensor Fusion)



the user-location from GPS trajectory

user-location-activity data (activity: 'Food', 'Shopping', 'Movie', 'Sports', 'Amusement')

points of interest database

- Mahremiyet gözeten yapay öğrenme
  - Dağıtık Sistemlerde öğrenme
  - Veri değil parametre paylaşımı

Paralel işleme Platformları (BBL2011)

Platform	Communication Scheme	Data size
Peer-to-Peer	TCP/IP	Petabytes
Virtual Clusters	MapReduce / MPI	Terabytes
HPC Clusters	MPI / MapReduce	Terabytes
Multicore	Multithreading	Gigabytes
GPU	CUDA	Gigabytes
FPGA	HDL	Gigabytes

Slide from ICML 2011 tutorial Langford et. al.

Veri Analitiği Platformları (Açık Kaynak)

- Hadoop/MapReduce

- Apache Spark

- Storm

- …

Özet

- Veri ≠ Bilgi
- Algoritma tasarımında yeni bakış açıları
- Büyük veri: veri noktaları arasındaki ilşkiler ve etkileşimler
- Yapay Öğrenme : bir çok uygulamada olgun teknoloji
- Bilgisayar bilimleri eğitimi:
  - Daha çok Matematik, Fizik, İstatistik ve Sosyal Bilimler etkileşimi
- Büyük veri = Büyük Potansiyel

- http://www.winshuttle.com/big-data-timeline/

Teşekkürler, Sorular