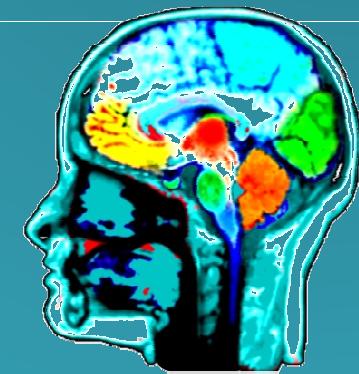




Introduction To Data Mining

Isfahan University of Technology (IUT)
Bahman 1401



Introduction

Dr. Hamidreza Hakim
hamid.hakim.u@gmail.com

بِسْمِ اللّٰهِ الرَّحْمٰنِ الرَّحِيْمِ

Content

References

Grading

Data

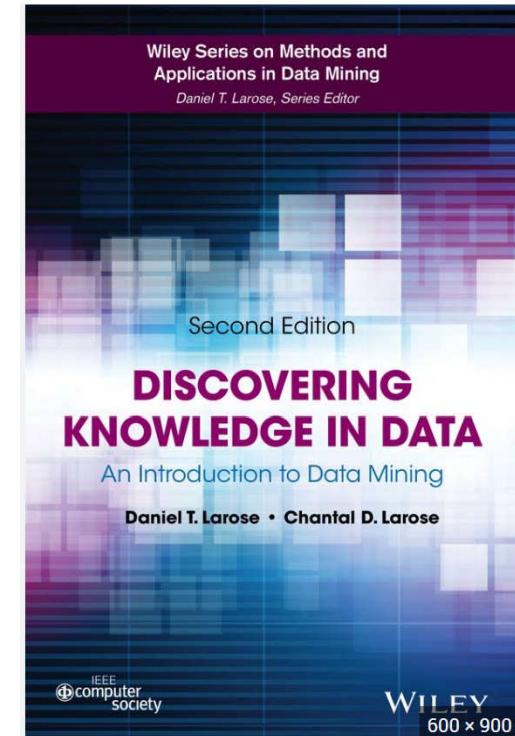
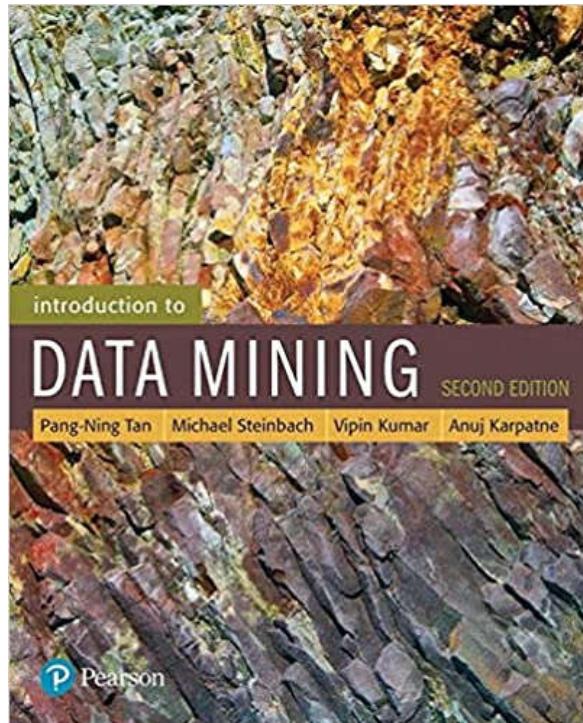
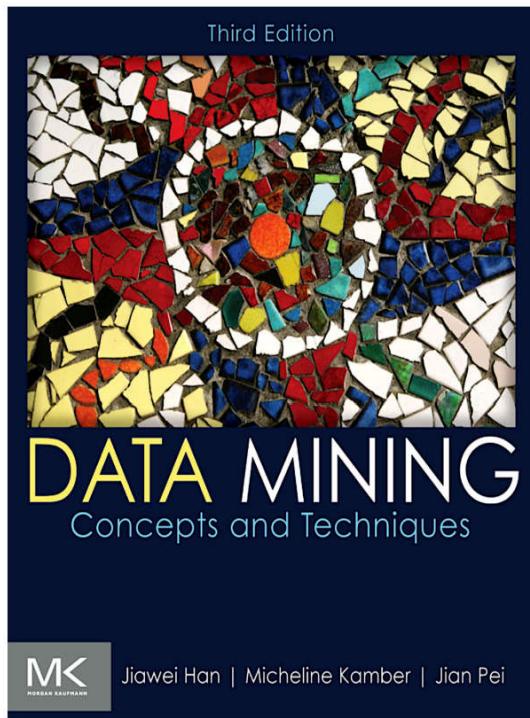
What is Data mining?

Why Data mining

Multi-Dimensional View of Data Mining

Some Examples

References



Grades

- Project and Presentation: 2 points
- Exercises: 5 points
- Exams and Quizzes: 13 points

Large-scale Data is Everywhere!

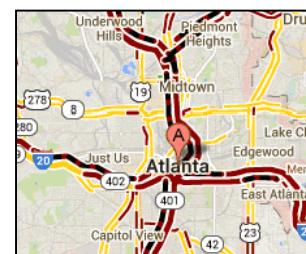
- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies
- New mantra
 - Gather whatever data you can whenever and wherever possible.
- Expectations
 - Gathered data will have value either for the purpose collected or for a purpose not envisioned.



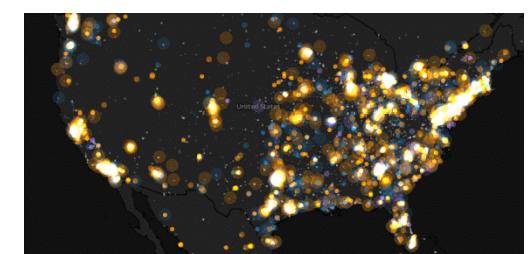
Cyber Security



E-Commerce



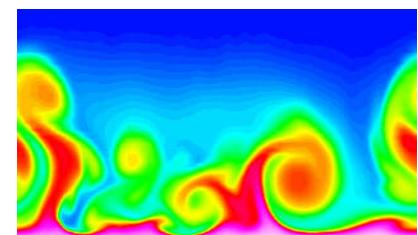
Traffic Patterns



Social Networking: Twitter



Sensor Networks



Computational Simulations

- : Large-scale Data

دیتا همه جا هست

توی سال های اخیر به واسطه شرایطی که پیش اومده شرکت ها به این نتیجه رسیده اند که دیتا شده یک تکنولوژی جدید و می گن دیتا نفت است و داده کاو ها کسانی هستند که توانایی استخراج دانش از این داده ها رو دارند

Why Data Mining?

چرا Data Mining مهم شده؟

- 1- حجم دیتا خیلی بیشتر شده
- 2- تکنولوژی ذخیره سازی رشد کرده
- 3- بستر برای اشتراک دیتا فراهم شده

Why Data Mining? Commercial Viewpoint

- Lots of data is being collected and warehoused
 - Web data
 - ◆ Google has Peta Bytes of web data
 - ◆ Facebook has billions of active users
 - purchases at department/grocery stores, e-commerce
 - ◆ Amazon handles millions of visits/day
 - Bank/Credit Card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
 - Provide better, customized services for an edge (e.g. in Customer Relationship Management)



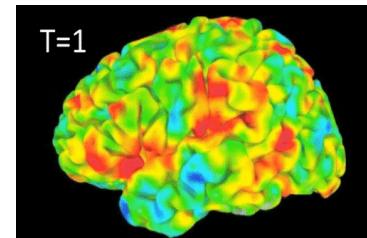
-
حجم دیتا خیلی زیاد شده

پردازش ها از دو بعد تحول داشته:
یکی الگوریتم ها توسعه پیدا کردند
دوم سخت افزارها پیشرفته تر شدن

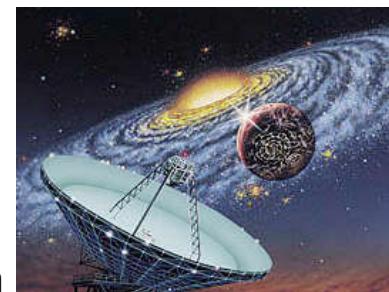
فشار رقابتی است --> این فشار باعث میشه که همه سعی میکنند برن سمت Data Mining چون احساس می کنند دیگه با ساز و کار سنتی نمیشه رفتارها رو شناخت و نمیشه سرویس ها رو بررسی کرد و مشتری ها رو بهتر شناخت و شرکت ها دارن برای افزایش بهره وری دارن روی این سرمایه گذاری می کنند --> همین بحث Data Mining رو خیلی جدی تر کرده

Why Data Mining? Scientific Viewpoint

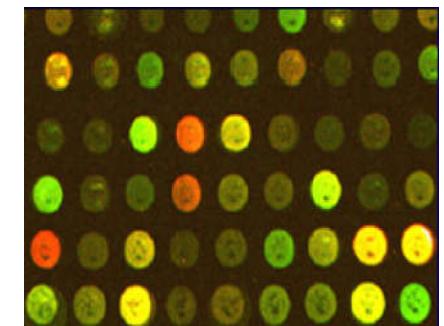
- Data collected and stored at enormous speeds
 - Remote sensors on a satellite
 - ◆ NASA EOSDIS archives over petabytes of earth science data/year
 - Telescopes scanning the skies
 - ◆ Sky survey data
 - High-throughput biological data
 - Scientific simulations
 - ◆ terabytes of data generated in a few hours



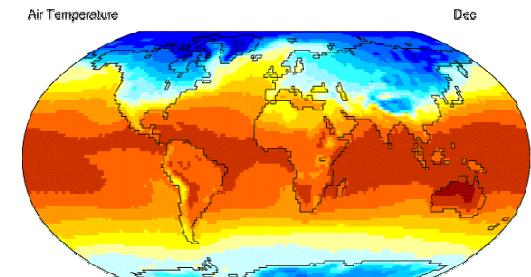
fMRI Data from Brain



Sky Survey Data



Gene Expression Data



Surface Temperature of Earth

نگاه از دیدگاه علمی:

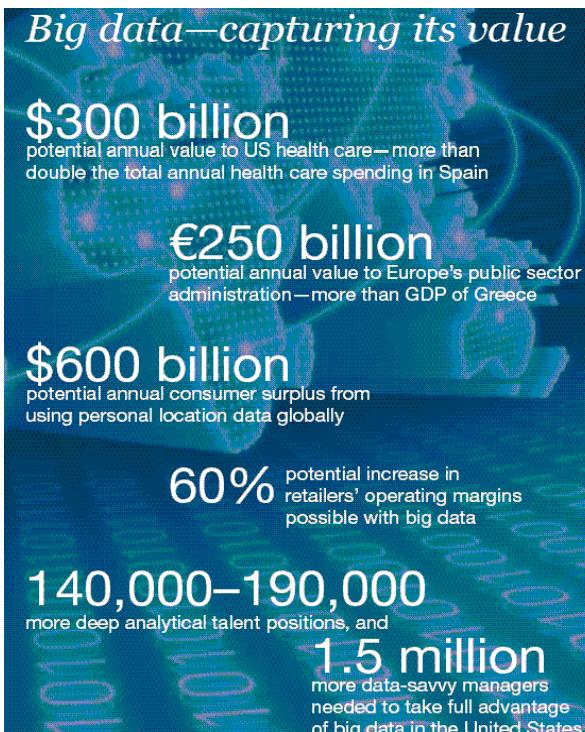
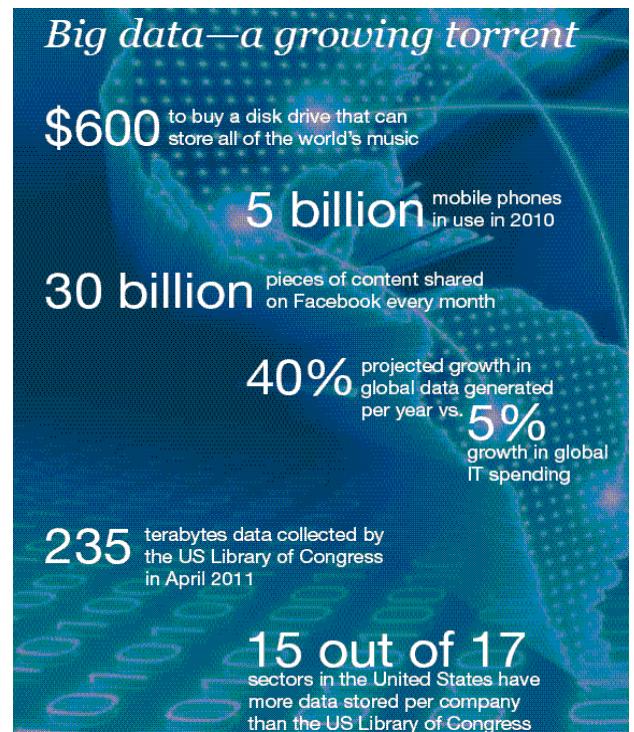
یه جوری همه جا تبدیل شده به مکان جمع اوری دینا و همه دارن تلاش میکنن دینا جمع اوری میکنن
مثلًا توی بحث مغز

....

Great opportunities to improve productivity in all walks of life

McKinsey Global Institute

Big data: The next frontier for innovation, competition, and productivity



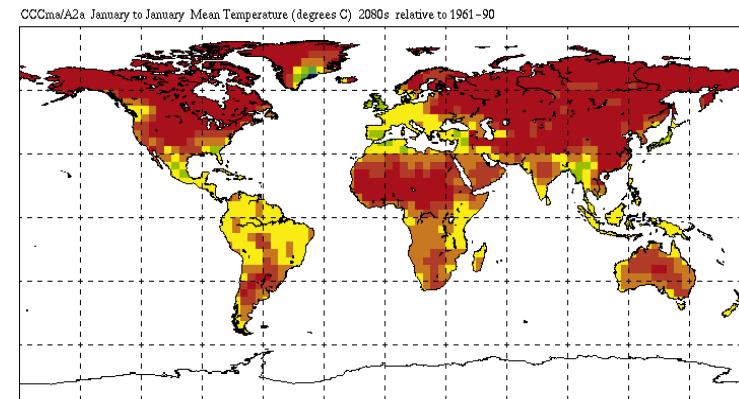
Great Opportunities to Solve Society's Major Problems



Improving health care and reducing costs



Finding alternative/ green energy sources



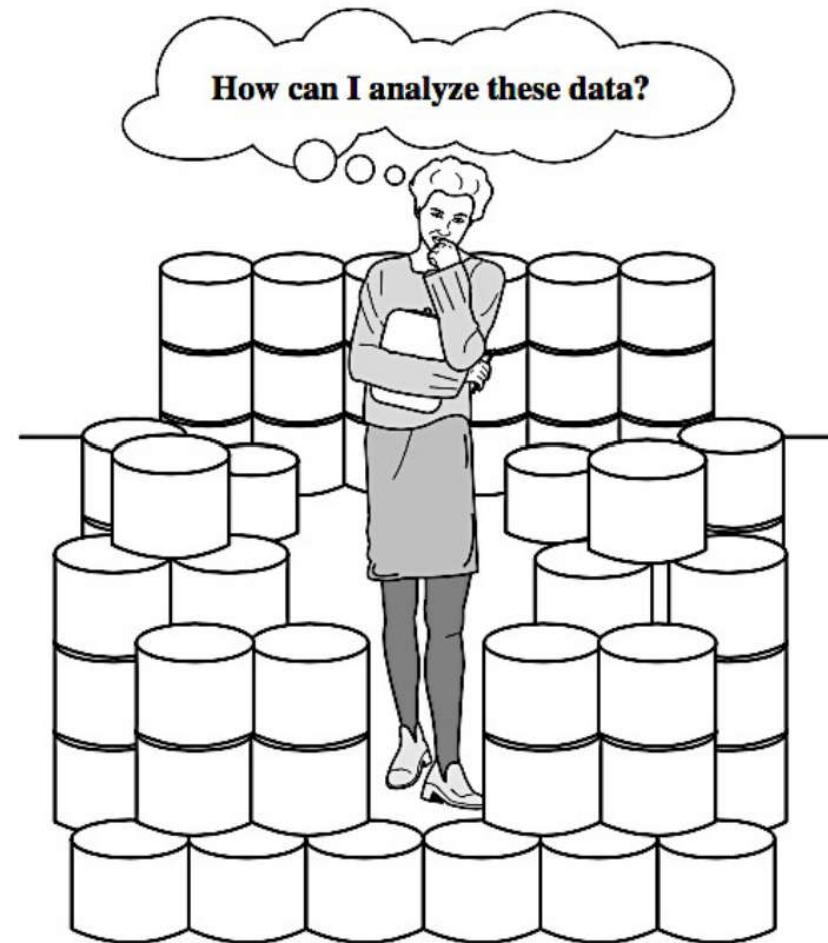
Predicting the impact of climate change



Reducing hunger and poverty by increasing agriculture production

فرصت های بزرگ برای حل مشکلات عمدۀ جامعه

**We Are Drowning In Data,
But Starving For Knowledge!**



The world is data rich but information poor.

-
دنیا از داده ها غنی است ولی از اطلاعات خالی است
ما توانی دنیای زندگی میکنیم که حجم زیادی از داده ها وجود داره ولی اطلاعات از اون ها به خوبی
استخراج نمیکنیم چون ما هنوز نمی دونیم با خیلی از اطلاعات چطوری باید برخورد بکنیم و اینکه
توی پردازش این اطلاعات ما انسان ها دخیل هستیم یعنی نیاز به مهارت ما انسان ها داره

What is Data Mining?

- Many Definitions

- Non-trivial **extraction** of implicit, previously unknown and potentially useful **information** from data
- **Exploration & analysis, by automatic or semi-automatic means**, of large quantities of data in order to discover meaningful patterns.

(Knowledge or Data) Mining!!!



Data mining—searching for knowledge (interesting patterns) in data.

-
تعريف:

علم استخراج اطلاعات از داده هاست یا بحث کشف کردن و انالیز کردن با یکسری تکنیک های اتوماتیک و شبیه اتوماتیک است ولی چیزی که وجود داره ما توانی داده کاوی به دنبال دانش هستیم و شاید بهتره بگیم ما می خوایم Knowledge mining بکنیم

داده کاوی یک فرایند کشف الگوهای مخفی و اطلاعات با ارزش از یک حجم زیادی از داده هاست

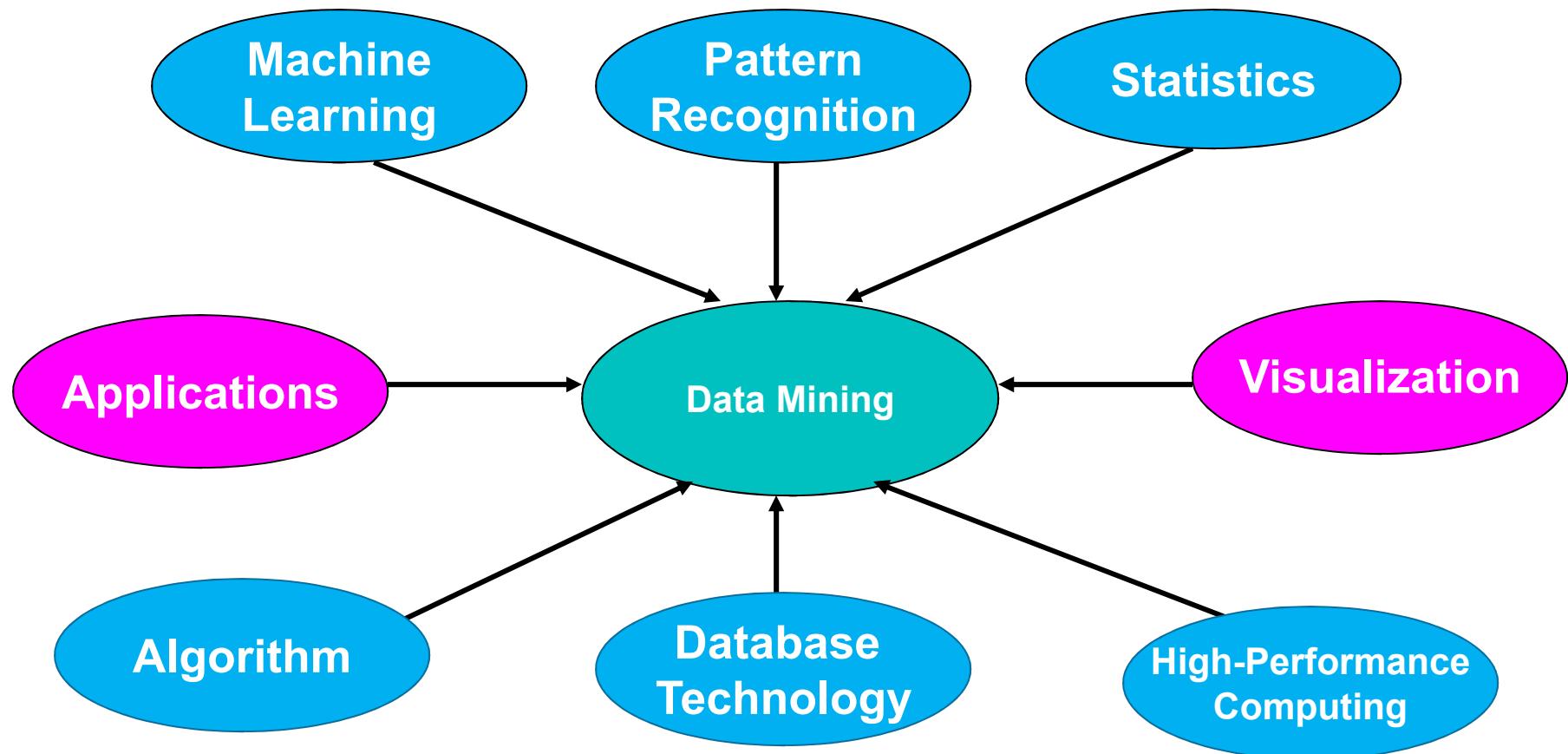
What is Data Mining?

- Data mining,
 - is the **process** of **uncovering patterns and other valuable information from large data sets.**

- Alternative names

Knowledge discovery (mining) in databases (KDD),
knowledge extraction, data/pattern analysis, data
archeology, data dredging, information harvesting,
business intelligence, etc.

Data Mining: Confluence of Multiple Disciplines



-
توی داده کاوی چندین مسئله داریم و به چندین حوزه باید توجه بکنیم:
که توی شکل نوشته شده

Human role in data mining?

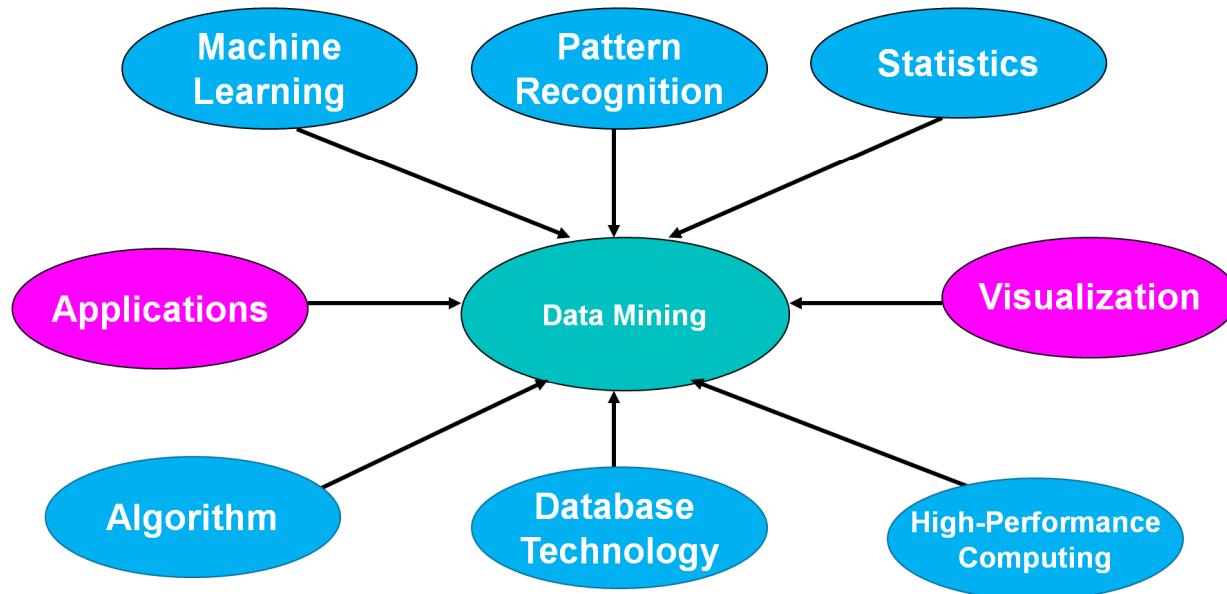
- Berry and Linoff, in their 1997 book gave the following definition for data mining:
 - “Data mining is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules”.

Human role in data mining?

- Berry and Linoff, in their 1997 book gave the following definition for data mining:
 - “Data mining is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules”.
- **Three years** later, in their Mastering Data Mining book, they mentioned that,
 - “If there is anything we regret, it is the phrase ‘by automatic or semiautomatic means’ . . . because we feel there has come to be too much focus on the automatic techniques and not enough on the exploration and analysis. This has misled many people into believing that **data mining is a product that can be bought** rather than **a discipline that must be mastered**.”

Human need to be actively involve in every phase of data mining.

Human role in data mining?



Human need to be actively involve in every phase of data mining.

انسان باید به طور فعال در هر مرحله از داده کاوی مشارکت داشته باشد.

Question?

Your Name, ID, Major

Q1: What do you think Data Mining is?

Q2: What project have you done so far that you think is most relevant to Data Mining?

Not necessarily research project; can be your course project or any hackathon event you participated in.

Q3: What do you expect to learn from this course?

Multi-Dimensional View of Data Mining

- Data to be mined
 - Database data (extended-relational, object-oriented, heterogeneous), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks

اولین بعد اينه که ما با چه ديتايی سرکار داريم ينی با چه ديتايی می خوايم عملیات کاوش رو انجام بدیم
 نوع ديتا ينی ما قراره با تایم سری کار بکnim یا با ديتابیس ...

یک چارچوبی داريم توی داده و اون اينه که داده ها نرمالايز باشن و جدول ها تا اونجايی که میشه فشرده باشه و اطلاعات توش ذخیره بشه
 ديتا مارت يا **data mart** ينی چی؟ ما نياز داريم یک سری اطلاعات رو حتی با وجود افزونگی که داريم کnar هم بذاريM و یک جدول بزرگی درست بکnim که توش اطلاعات رو از جدول های مختلف گذاشته و حتی ممکنه يکسری ستون هاش تكراري باشه

Multi-Dimensional View of Data Mining

- **Data to be mined**
 - Database data (extended-relational, object-oriented, heterogeneous), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- **Knowledge to be mined (or: Data mining functions)**
 - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, ...
 - Descriptive vs. predictive data mining
 - Multiple/integrated functions and mining at multiple levels

- دانشی که دنبالش هستیم که استخراج بشه:

مثلاً دیجی کالا میگه فروش ما کمه و این دیتای ما و الان باید چی کار بکنیم و ما باید ببینیم الان چه رویکردی براساس این دیتا پیاده سازی بکنیم که فروش این رو افزایش بدیم

Multi-Dimensional View of Data Mining

- **Data to be mined**
 - Database data (extended-relational, object-oriented, heterogeneous), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- **Knowledge to be mined (or: Data mining functions)**
 - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, ...
 - Descriptive vs. predictive data mining
 - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
 - Data warehousing (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance computing, etc.

-
تکنیک هایی که قراره استفاده بکنیم:
ینی قراره ما چه تکنیک هایی توی کار داشته باشیم

Multi-Dimensional View of Data Mining

- **Data to be mined**
 - Database data (extended-relational, object-oriented, heterogeneous), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- **Knowledge to be mined (or: Data mining functions)**
 - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, ...
 - Descriptive vs. predictive data mining
 - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
 - Data warehousing (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance computing, etc.
- **Applications adapted**
 - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

- کاربردی که انتظار داریم:

خود کسب کاره یک کمک هایی میکنه توی داده کاوی که باید اونارو بفهمیم ینی از دید کسب و کار به داده کاوی نگاه بکنیم ینی کسب و کار رو بفهمیم ینی این داده ها از کجا او مده و ...

Data Mining Tasks

- **Prediction Methods**
 - Use some variables to predict unknown or future values of other variables.

- **Description Methods**
 - Find human-interpretable patterns that describe the data.

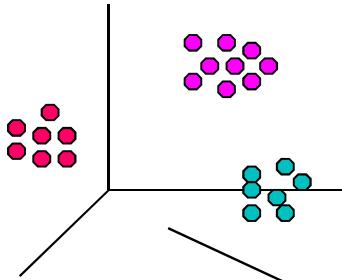
From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

- دسته بندی های مختلفی از لحاظ کارهایی که توانی داده کاوی انجام میدیم وجود دارد:
: Prediction

Description : معمولاً مسئله هایی که با داده داریم از Description شروع می شه بینی اول برای من خود داده کاو باید یک تعریف خیلی خوبی از این داده ها چیه و چه اتفاقی توش می افته حاصل بشه که بعد بتونیم راجع به Prediction حرف بزنیم یه جاهایی همین Description جواب میده

مثالاً یک پروژه ای هست و یک دیتایی به ما داده و بعد بیاین این دیتا رو شفاف بکن پس یک کار Description اینجا کردیم و الان که وضعیت الان رو خوب فهمیدیم بعد میایم میگیم الان میخوایم پیش بینی بکنم که بعداً چه اتفاقی می افته که این میشه Prediction

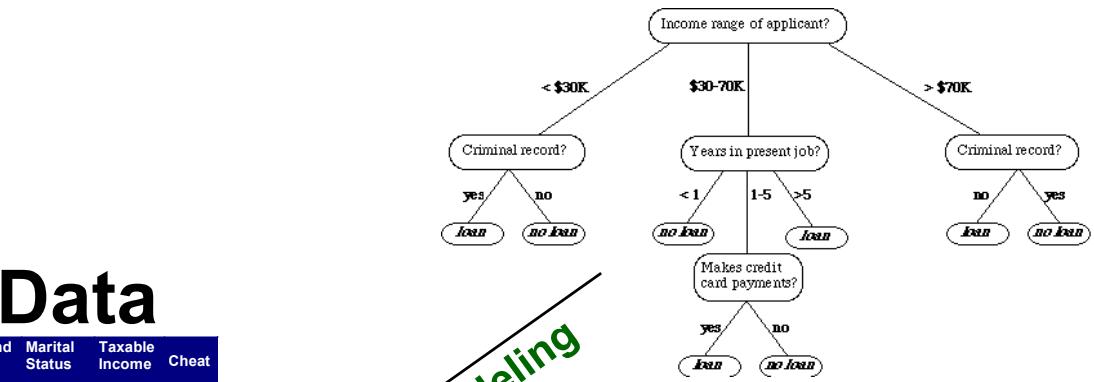
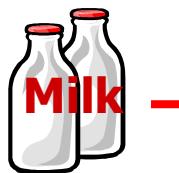
Data Mining Tasks ...



Clustering

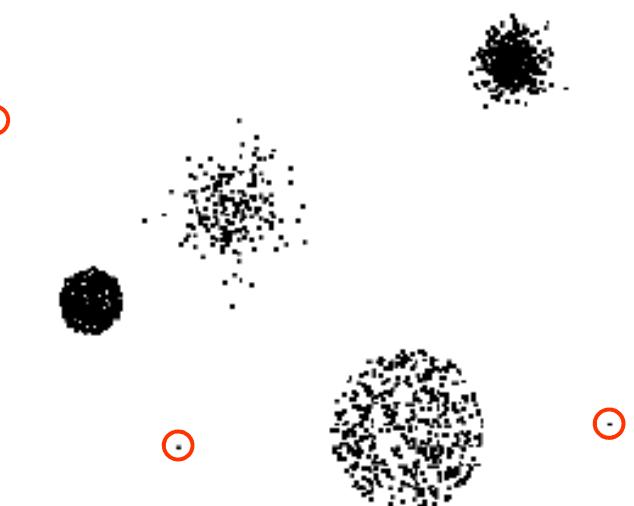
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Association
Rules



Predictive Modeling

Anomaly
Detection



دسته بندی بعدی:

این 4 محور است که توی شکل نوشته

Association Rule Discovery: Definition

Given a set of records each of which contain some number of items from a given collection

Produce dependency rules

which will

predict occurrence of an item based on occurrences of other items.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

$\{\text{Milk}\} \rightarrow \{\text{Coke}\}$

$\{\text{Diaper}, \text{Milk}\} \rightarrow \{\text{Beer}\}$

: Association Rule Discovery

بعضی وقت‌ها اون ابجکت‌هایی که می‌ایم بررسی می‌کنیم یک سری ویژگی‌های مشترکی دارن و اون ویژگی مشترک برای ما مهم می‌شده است. مثلاً مشتری‌ها چه کالایی رو با هم می‌خرند؟ چون می‌خوایم اون کالایی که با هم می‌خرن رو کنار هم بذاریم --> دوست داریم ببینیم که چه فیچرهایی با هم داره اتفاق می‌افته.

مثال:

جدوله اطلاعات ادم‌های مختلفی است که 5 تا ابجکت است:

وقتی که می‌ایم کالاهای پرتکرارش رو استخراج می‌کنیم یک قانونی از توش درمی‌آید: که اگه هر کی شیر بخره می‌رمه کیک هم می‌خره.

Association Rule Discovery: Definition

Given a set of records each of which contain some number of items from a given collection

Produce dependency rules

which will

predict occurrence of an item based on occurrences of other items.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

$\{\text{Milk}\} \rightarrow \{\text{Coke}\}$

$\{\text{Diaper}, \text{Milk}\} \rightarrow \{\text{Beer}\}$

Association Analysis: Applications

- Market-basket analysis

- Rules are used **for sales promotion**, shelf management, and inventory management

مسئله ای که توی صفحه قبلی گفتیم مسئله Market-basket analysis بود ینی انالیز سبد خرید که ما میایم تشخیص میدیم که چه کالایی با هم خریده میشه

Association Analysis: Applications

- Market-basket analysis
 - Rules are used **for sales promotion**, shelf management, and inventory management
- Telecommunication alarm diagnosis
 - Rules are used to find combination of alarms that occur together frequently in the same time period

یک کاربرد دیگش بحث Telecommunication alarm diagnosis است ینی ما میخوایم بینیم یک سری اتفاق ها با هم دیگه می افته و انالیز می کنیم و ممکنه یک سری خطاهایی توی سیستم پیچیده ای ایجاد بشه و بعد ما لاگ اینارو داریم و ما می خوایم بفهمیم اگر این خطا پیش اومد بعدش چه خطایی پیش میاد و اگر اینارو کنار هم بذاریم می بینیم یک زنجیره خطا داره اتفاق می افته

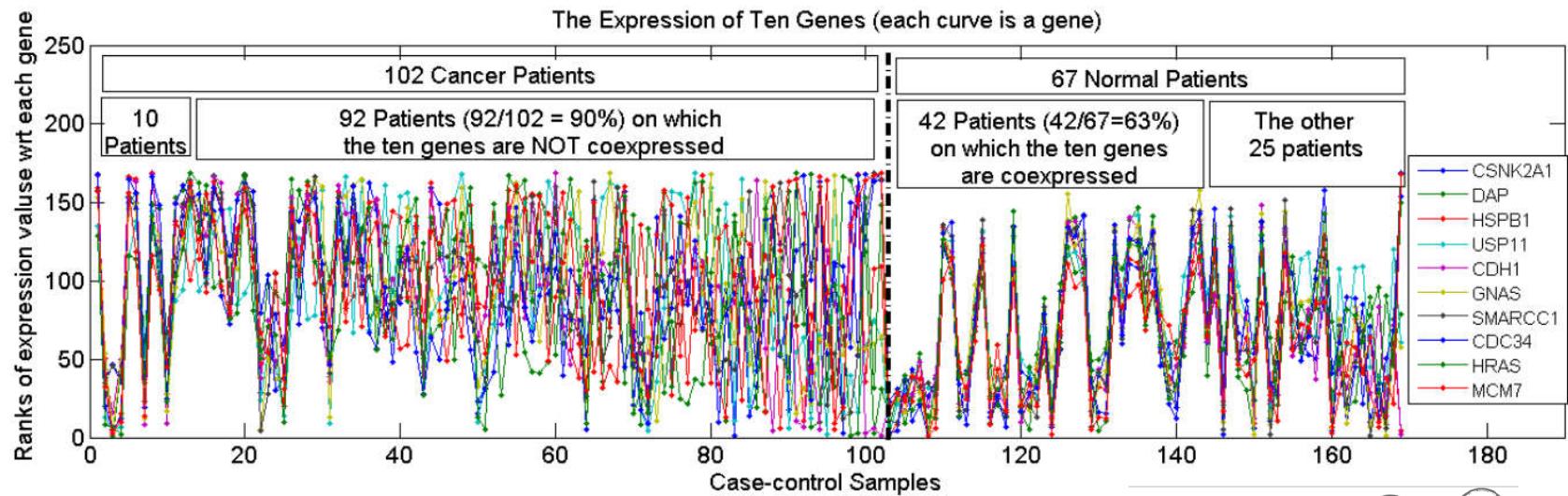
Association Analysis: Applications

- Market-basket analysis
 - Rules are used **for sales promotion**, shelf management, and inventory management
- Telecommunication alarm diagnosis
 - Rules are used to find combination of alarms that occur together frequently in the same time period
- Medical Informatics
 - Rules are used to find combination of patient symptoms and test results associated with certain diseases
- Some Other Example
 - Associate Events

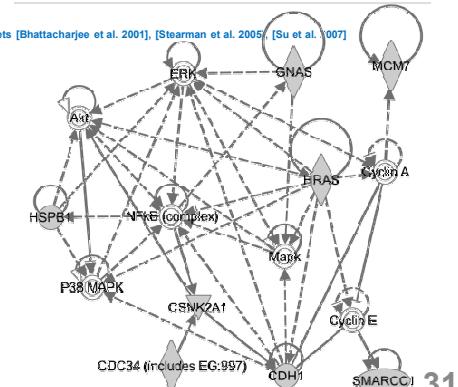
توی بحث پزشکی هم خیلی استفاده میشه مثلًا توی بحث ژن درمانی

Association Analysis: Applications

- An Example Subspace Differential Coexpression Pattern from lung cancer dataset



Enriched with the TNF/NFB signaling pathway
which is well-known to be related to lung cancer
P-value: 1.4×10^{-5} (6/10 overlap with the pathway)



[Fang et al PSB 2010]

09/09/2020

Introduction to Data Mining, 2nd Edition
Tan, Steinbach

مثلا: برای تشخیص ادم سالم از سرطان استفاده شده ینی یک سری بازنمایی داریم از ژن ها که او نا رو او مدن گزارش کردن

محور های افقی سابجکت های ما هستن و تیکه سمت راست ادم های سالم اند و تیکه سمت چپ ادم هایی هستن که سرطان دارن و هر کدام این ها میزان بازنمایی اون ژن است که جوری اندازه گیری شده

ما می بینیم ادم های سالم بازنمایی ژن هاشون مثل هم است ولی توی بیمار سرطانی نه اینطوری نیست و اینجا ما با یک تصمیم می تونیم تشخیص بدیم این فرد سرطان داره یا نه

Classification Example

predicting (credit worthiness)

categorical categorical quantitative class				
Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
...

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
...

Some Other Examples

- Predict Job by action
- Predict which Key press by sound OR Monitor
- Best Action in Trade

: Classification

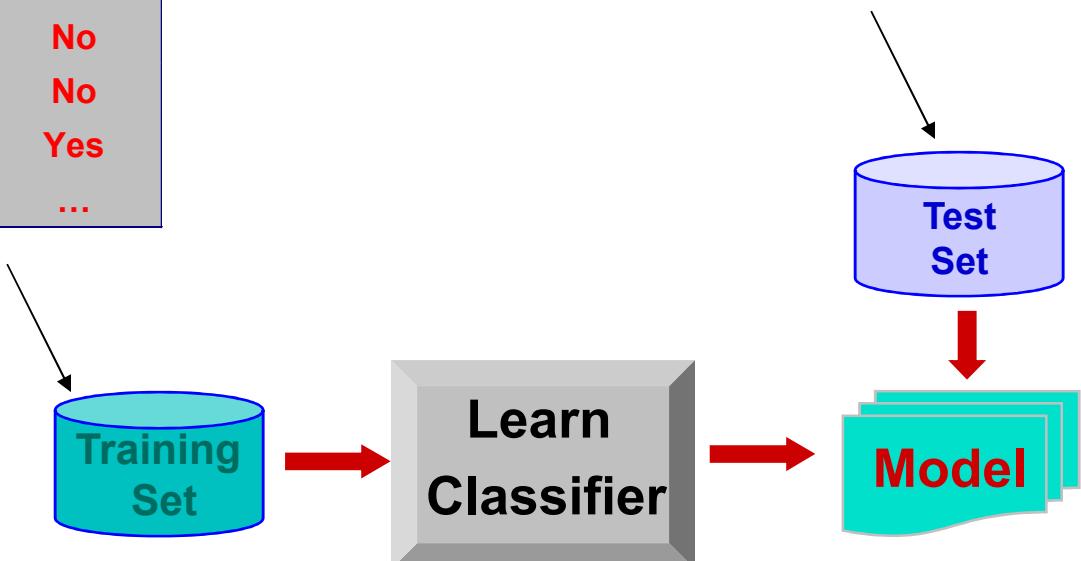
مثال: فرض میکنیم که اطلاعات یک بانک رو داریم و این اطلاعات مشتریان یک بانک است و برای هر مشتری مشخص شده این ادم استخدام شده یا نشده و ستون دوم سطح تحصیلات مشتری رو میگه و ستون سوم مدت زمانی است که توی اوم مکان حضور داشته و ستون اخر مگه وام دادیم بهش ینی ارزش داشته که بهش وام دادیم و این یک سری اطلاعاتی است که بانک از قبلا جمع اوری کرده ینی گفته ما به این مشتری ها وام دادیم و این ها وام هاشون رو برگردانند حالا یکسری مشتری های جدید داریم و میخوایم ببینیم به این ها وام بدیم یا ندیم: مثلا اگر سطح تحصیلاتش بالا بود می تونیم بگیم وام رو برمنی گردونه --> این که الان به این نتیجه رسیدیم بخاراطر اون 4 تا سمپل است و اگر سمپل های بیشتری داشتیم مطمئن تر راجع بهش حرف می زدیم

Classification Example

predicting credit worthiness

predicting credit worthiness				
Tid	categorical		quantitative	
	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
...

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
...



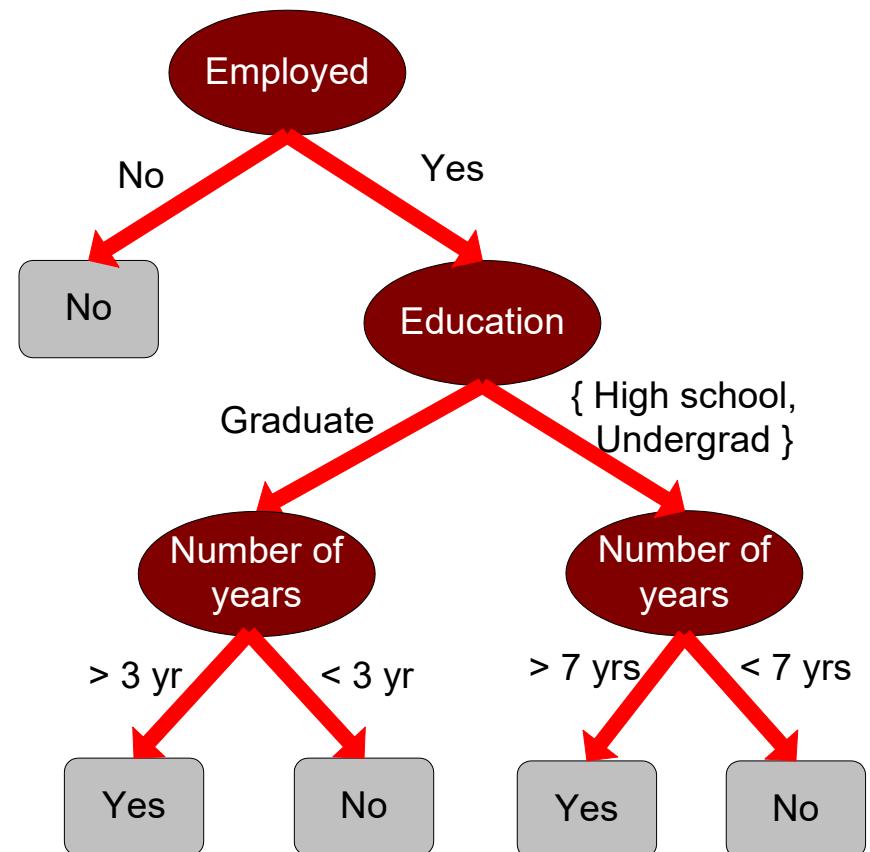
رویه ای که توی Classification داریم اینه که ما یک ترین ست داریم و توی این ترین ست ما می دونیم برچسب هر چیزی چیه و یک classifier است که learn میشه و این تبدیل به یک سری مدل میشه و اون داده های تست میاد کنار هم می شینه توی این مدل و میگه برچسب چی هست

Predictive Modeling: Classification

- Find a model for class attribute as a function of the values of other attributes

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
...

Model for predicting credit worthiness



-

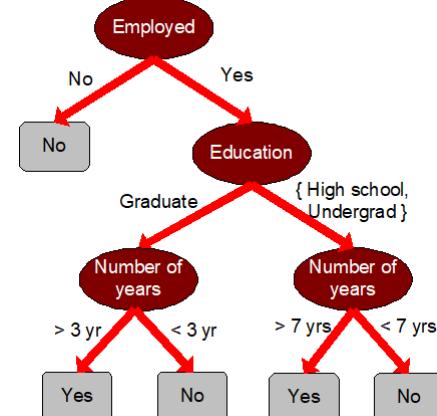
توی این عملیات Classification یک Predictive Modeling ایجاد میشه
چیه؟ یک چکیده ای است از داده ها --> چکیده ای که داره اون مسئله
رو برای ما حل میکنه و با کمک اون چکیده می تونیم داده های جدید رو ارزیابی بکنیم
اینجا ما یک Predictive Modeling داریم از نوع درخت تصمیم --> اگر این درخت تصمیم رو
داشته باشیم انگار کل این دیتاهای رو داریم

Classification Example

predicting credit worthiness

categorical categorical quantitative class				
Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
...

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
...



هر کدوم از این ستون ها یکسری تایپ داره که جلوتر میگیم اینا رو

Classification: Application 1

Fraud Detection

- **Goal:** Predict fraudulent cases in credit card transactions.
- **Approach:**



مسئله : Fraud Detection

میخوایم تشخیص تقلب بدیم --> ینی اطلاعات کارت های بانکی به ما داده شده و بهمون میگن این کارت جدید کارت بانکی شخص متقلب هست یا نه

رویکرد:

اینجا باید الگوها را یک حالت پویا بکنیم

پس ما باید یک مجموعه ای از تراکنش های بانکی رو جمع اوری بکنیم و اول کار به یکی بگیم بهمون بگه کدامشون تقلبی هستن یا نیستن ینی بهش بگیم یک برچسبی بهمون بده و بعد ما میایم یک مدلی روی این ها استخراج میکنیم که اون رفتار تقلبشو در بیاره و اونو روی نمونه های جدید تستش میکنیم

Classification: Application 1

Fraud Detection

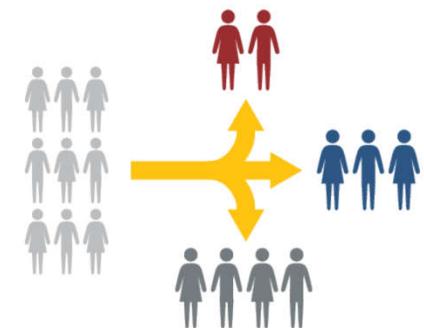
- **Goal:** Predict fraudulent cases in credit card transactions.
- **Approach:**
 - ◆ Use credit card transactions and the information on its account-holder as attributes.
 - When does a customer buy, what does he buy, how often he pays on time, etc
 - ◆ Label past transactions as fraud or fair transactions. This forms the class attribute.
 - ◆ Learn a model for the class of the transactions.
 - ◆ Use this model to detect fraud by observing credit card transactions on an account.



Classification: Application 2

Churn prediction for telephone customers

- **Goal:** To predict whether a customer is likely to be lost to a competitor.
- **Approach:**
 - ◆ Use detailed record of transactions with each of the past and present customers, to find attributes.
 - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
 - ◆ Label the customers as loyal or disloyal.
 - ◆ Find a model for loyalty.



From [Berry & Linoff] Data Mining Techniques, 1997

- مسئله: دسته بندی مشتری:

از ما میخوان که مشتری ها رو دسته بندی بکنیم مثلا ما یک سرویس دهنده تلفنی هستیم و نگران این هستیم که این مشتری دیگه کارتشو شارژ نکنه ینی می خوایم یک کاری بکنیم مشتری ها بیشتر بشن

رویکرد:

اینجا برچسبی وجود نداره و خودمون باید برچسب بزنیم و یکسری قواعد در بیاریم و با اون قواعد بتونیم برچسب بزنیم

مثلا بگیم تعریف ما از این که این مشتری خوبی است یا .. چی است
مثلا اگر 6 ماه شارژ نکرد ینی مشتری است که از دستش دادیم

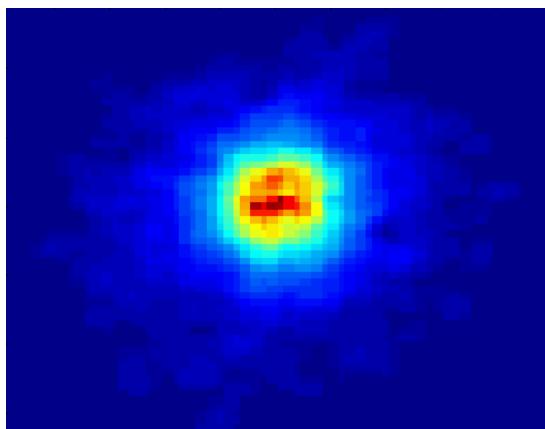
پس یک چارچوب اینطوری تعریف میکنیم و برچسب می زنیم به اون ها و با خود کار داده کاوی یک برچسبی تولید میکنیم و برچسب می زنیم و دیگه میشه مسئله Classification و بعدش سعی میکنیم الگو پیدا بکنیم

پس برچسب می زنیم به مشتری ها وفادار و بی وفا و .. و بعد سعی میکنیم این مدل رو ایجاد بکنیم

Classifying Galaxies

Courtesy: <http://aps.umn.edu>

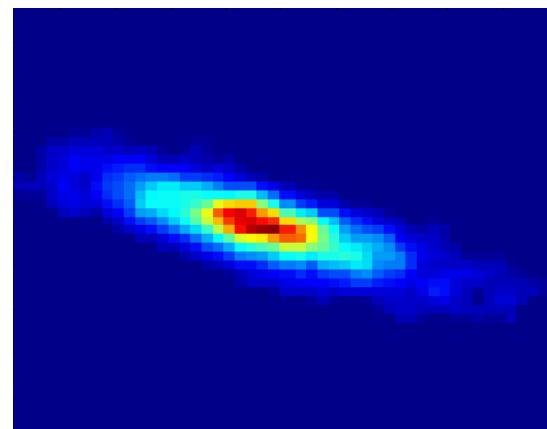
Early



Class:

- Stages of Formation

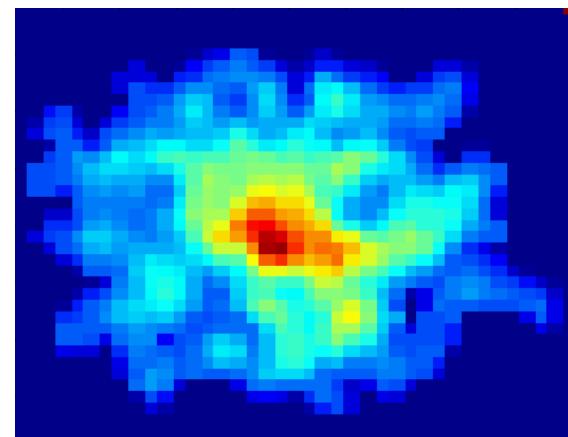
Intermediate



Attributes:

- Image features,
- Characteristics of light waves received, etc.

Late



Data Size:

- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

اینجا یک فضای بسیار بزرگی داریم و کهکشان ها یک چارچوبی دارن و هر کدامشون توی یک تیپی است و ما تصاویری که داریم حجم این تصاویر به شدت زیاد است و فرض میکنیم این تصاویر رو به ما دادن و گفتن این تصاویر رو دسته بندی بکن

رویکرد:

روی یک سمپل کوچک برچسب می سازیم و می زنیم و بعد اون دیتا رو مدل میکنیم و بعد بسطش میکنیم

مثلًا میگیم اونی که کارشناس نجوم است و وارد است بیاد این تعداد نمونه رو توی یک مقایس کوچیک برآمون برچسب بزن و بعد میایم اونو مدلش میکنیم و بسطش میدیم

نکته: یک جاهایی برای برچسب گذاشتن می تونیم از خود انسان ها استفاده بکنیم مثل کپچر گوگل--> مثلا میگه اتوبوس رو پیدا بکن و ما هم با دقیق زیاد اتوبوس رو پیدا میکنیم توی تصاویر ولی در واقعیت این است اون میاد یه تعدادی از این ها رو چک میکنه و یک تعدادی رو گذاشته که ما بیایم برآش برچسب بزنیم پس یک تکنیکی ساخته که این مسئله سمت ما حل بشه و برای خودش هم برچسب جمع اوری بکنه

Classification: Application 3

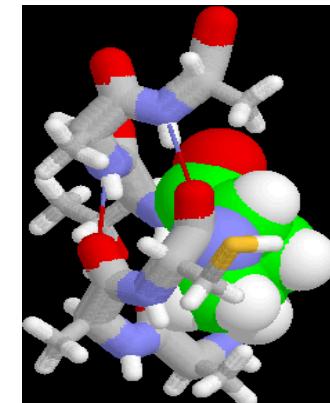
Sky Survey Cataloging

- **Goal:** To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
 - 3000 images with 23,040 x 23,040 pixels per image.
- **Approach:**
 - ◆ Segment the image.
 - ◆ Measure image attributes (features) - 40 of them per object.
 - ◆ Model the class based on these features.
 - ◆ Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Examples of Classification Task

- **Classifying land covers** (water bodies, urban areas, forests, etc.) using satellite data
- **Categorizing news** stories as finance, weather, entertainment, sports, etc
- Identifying intruders in the cyberspace
- Predicting tumor cells as benign or malignant
- **Classifying secondary structures of protein** as alpha-helix, beta-sheet, or random coil



-
مثال های دیگه از Classification :

Regression

- Predict a value of a given continuous valued variable based on the values of other variables, (assuming a linear or nonlinear model of dependency.)
- Extensively studied in statistics, neural network fields.
- Examples:
 - Predicting sales amounts of new product based on advertising expenditure.
 - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
 - Time series prediction of stock market indices.

-
Regression یا تقریب زدن:

توی دسته بندی دنبال این هستیم که بگیم داده های ما یک برچسبی رو بهش نسبت بدیم ولی توی

Regression ما دیگه برچسب 1 یا 2 یا 3.. نداریم بلکه توی Regression ما یک طیف داریم

که باید به یک مقداری از اون طیف نسبت بدیم

Regression برای پیش بینی یک مقدار پیوسته هستش مثلا میزان فروش یک مغازه رو

می خوایم تخمین بزنیم یا مثلا می خوایم میزان باد رو اندازه گیری بکنیم

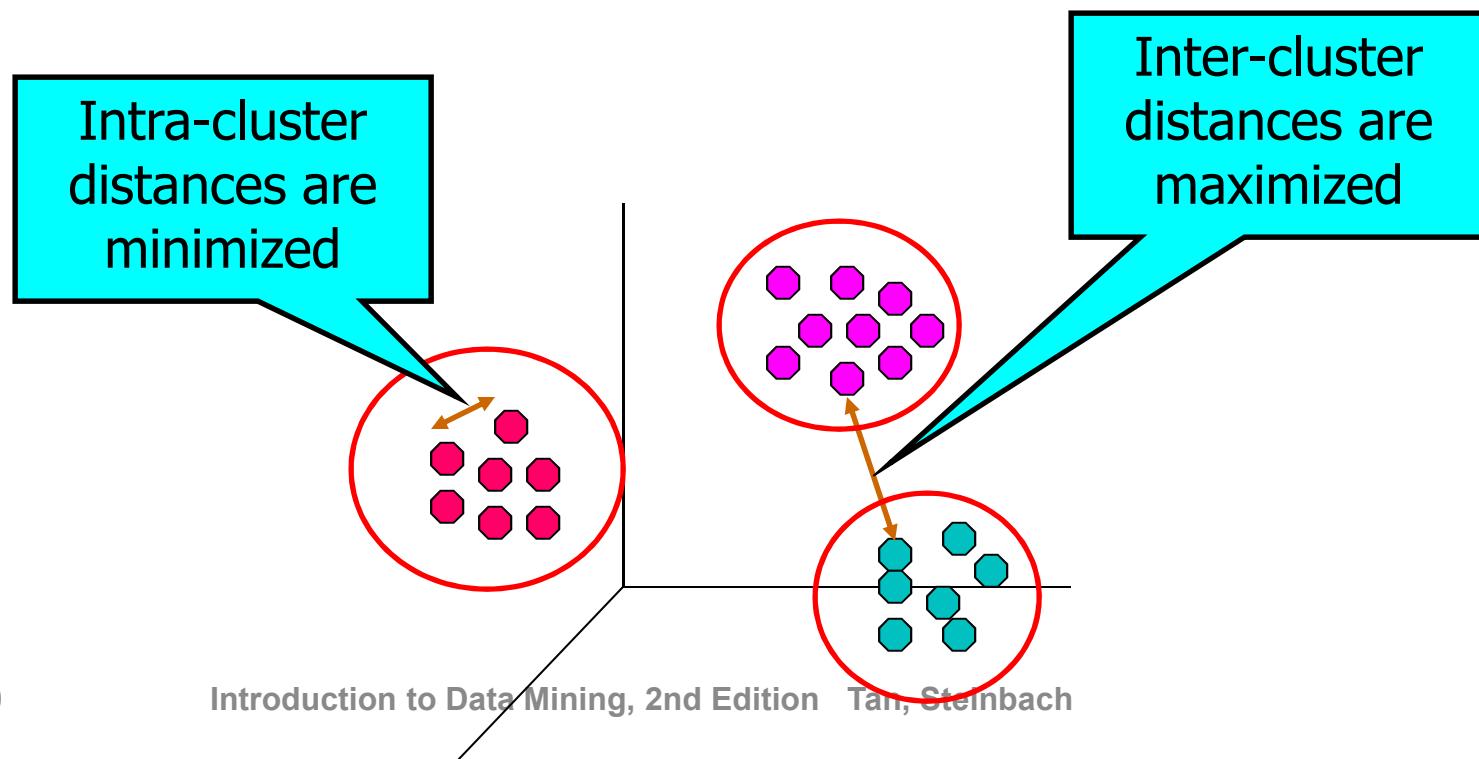
Clustering Task

Finding **groups** of objects

such that the objects in a group

will be **similar (or related)** to one another and

different from (or unrelated to) the objects in **other groups**.



- خوشه بندی:

با یک شرایطی روبه رو هستیم که اون چیزی که دنبال پیش بینیش هستیم مقدار دقیقشو نداریم و قراره خود داده ها رو به یک تعداد دسته تقسیمشون بکنیم

پس اینجا دنبال خوشه خوشه کردن داده ها هستیم که توی این خوشه ها داده هایی که توی این خوشه می افتن ویژگی مشترکی دارن

پس توی خوشه بندی می خوایم ببینیم این ابجکت ها با چه منطق هایی به هم شبیه هستند و اونا رو توی یک خوشه قرارشون بدیم پس باید گروه بندی بشن و خودمون هم نمیدونیم گروه درست چیه نکته: ما توی خوشه بندی به دنبال این هستیم که اشیایی که بهم شبیه هستند رو و با بقیه خیلی متفاوت هستند توی یک گروه قرار بدیم

شکل رو ببین!!

مثلا بر اینکه ببینم خوشه خوشه بندی رو خوب انجام دادیم: خوشه بندی خوبه که فاصله یک ابجکت از هم گروهیاش حداقل باشه و در عین حال نسبت به گروه های دیگه بیشترین فاصله رو داشته باشه

Clustering: Application 1

Market Segmentation:

- **Goal:** subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.

- کاربرد خوشه بندی:

میخوایم یک بازار را تقسیم بکنیم

اطلاعات فروشنده های یک فروشگاه به ما داده شده

و به ما میگن یک پلن فروش ارائه بده و یه کاری بکن فروش بیشتر بشه:

اولین سوال این است که مشتری ها رو بشناسیم و مشتری ها چه ویژگی هایی دارند

معمولًا مشتری ها به صورت گروه گروهی ویژگی مشترکی دارند

بعد از خوشه بندی می ریم اونارو بررسی می کنیم و یک پلن فروش برآشون در میاریم

نکته: از قبلش نمی دونیم چه مشتری توی چه خوشه ای هست باید همه مشتری ها همه اطلاعاتشون

رو بهمون بدن و ما می ریم اینارو بررسی میکنیم و خوشه بندی می کنیم و بعد می ریم روی این

خوشه کار های مشخصی انجام می دیم

برای هر خوشه ما برچسب به دست میاریم <-- نکته

خوشه بندی می تونه برآmon برچسب بسازه

Clustering: Application 1

Market Segmentation:

- **Goal:** subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.

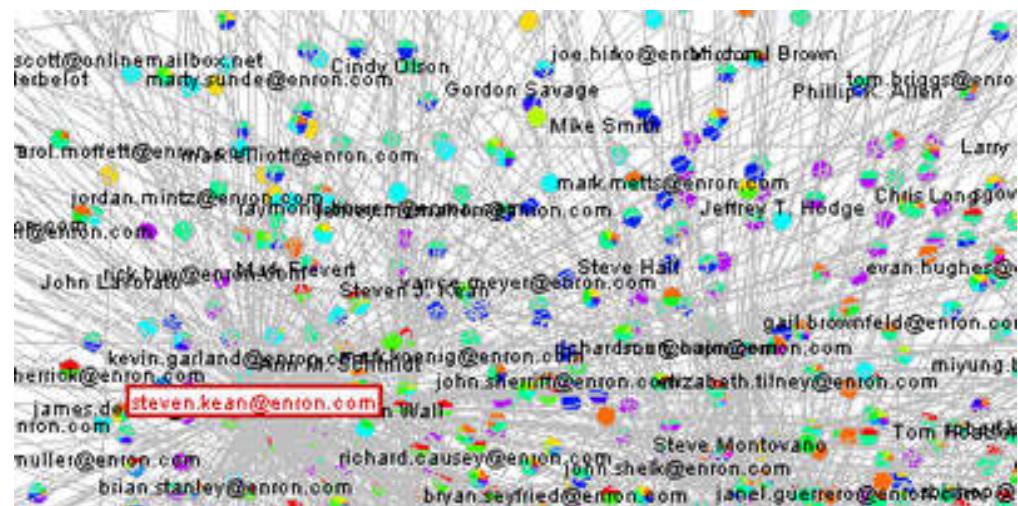
- **Approach:**
 - ◆ Collect different attributes of customers based on their geographical and lifestyle related information.
 - ◆ Find clusters of similar customers.
 - ◆ Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

Document Clustering:

- **Goal:** To find groups of documents that are similar to each other based on the important terms appearing in them.
- **Approach:** To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.

Enron email dataset



- کاربرد دیگه خوشه بندی:

یک حجم زیادی از متن ها و خبرها رو به ما دادن و به ما میگن که این هارو خوشه بندی بکن مثلا بگو کدوم خبر ورزشی است کدوم سیاسی و ... از قبل هم ما نمی دونیم و همینطور زبان این خبرها رو هم ما نمی شناسیم چی کار کنیم اینجا برای خوشه بندی؟ ینی چجوری خوشه بندی بکنیم؟ اون خبرهایی که ویژگی مشترک دارن مال یک خبرند مثلا یکسری کلمات یکسان توی اون خبرها هست که می تونیم بگیم ویژگی مشترکی دارند اینجا ویژگی مشترک میشه کلمات مشترک پس اینارو خوشه بندی میکنیم ولی نمی تونیم دقیق بفهمیم که توی این خوشه چی هست ولی می دونیم که این ها راجع به یک متن یا یک موضوع هستش

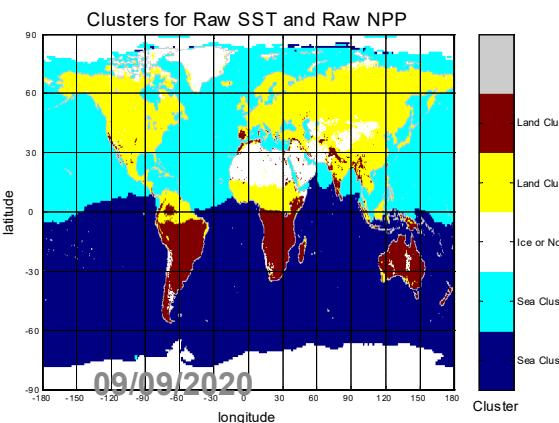
Applications of Cluster Analysis

● Understanding

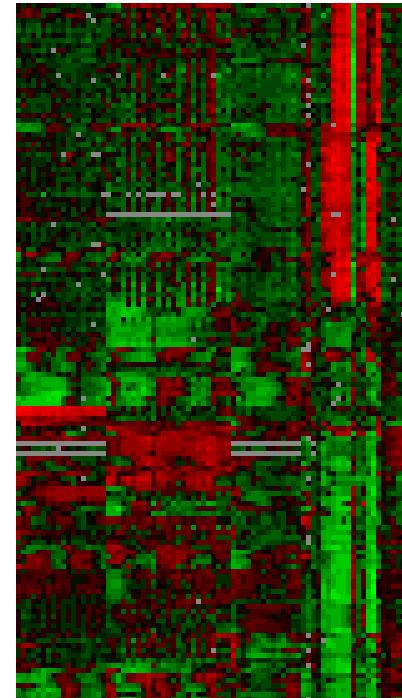
- Custom profiling for targeted marketing
- Group related documents for browsing
- Group genes and proteins that have similar functionality
- Group stocks with similar price fluctuations

● Summarization

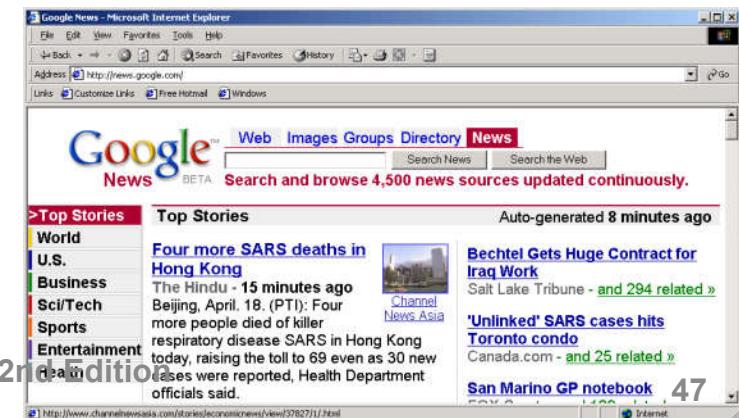
- Reduce the size of large data sets



Use of K-means to partition Sea Surface Temperature (SST) and Net Primary Production (NPP) into clusters that reflect the Northern and Southern Hemispheres.
Introduction to Data Mining, 2nd Edition
Tan, Steinbach



Courtesy: Michael Eisen



- دو تا کاربرد مهم داره خوشه بندی:
1- برای ما برچسب تولید میکنه
بقیش: Understanding است:

میخوایم ببینیم چخبره توی دیتا و چه گروه هایی داریم
چه نمونه هایی نمونه هایی پرتری هست و با بقیه متفاوت است

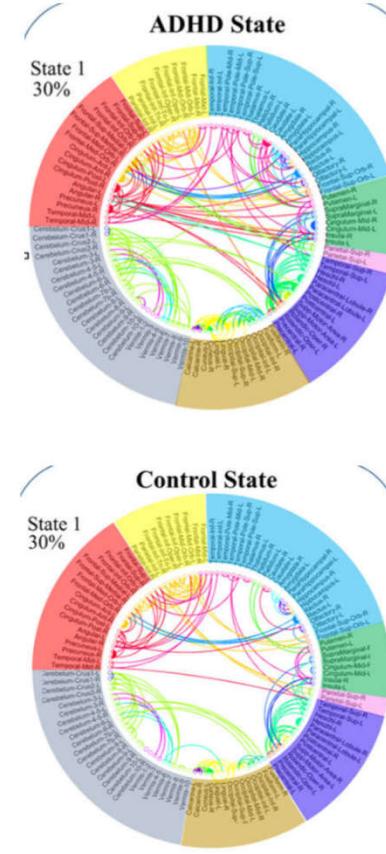
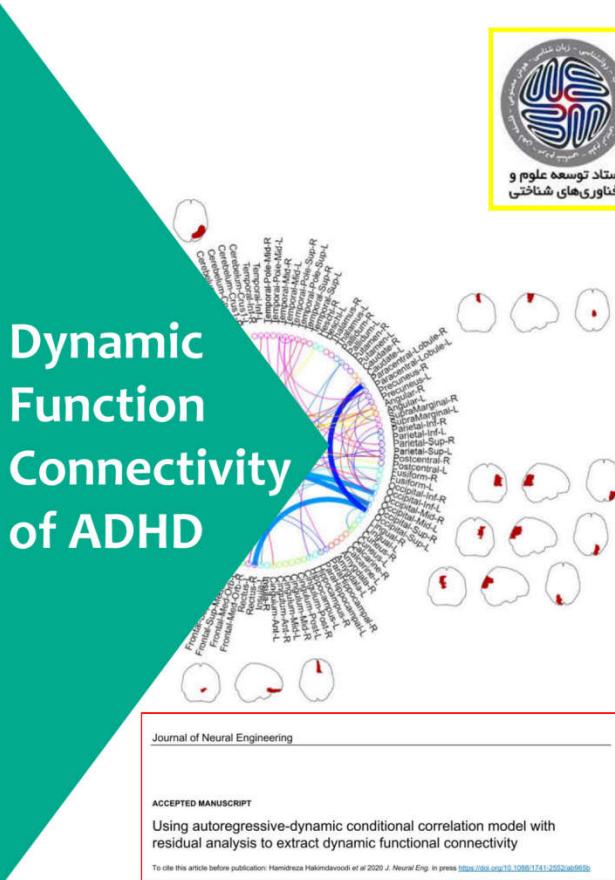
.....

2- بحث Summarization است: ما یک حجم زیادی اطلاعات داریم و ما میخوایم ببینیم چخبره توی این اطلاعات --> با یک خوشه بندی ما می تونیم نماینده ها و مشتری هایی که به عنوان نماینده یک گروه هستن استخراج بکنیم و بعد بباییم فقط با اونها کار بکنیم یعنی حجم زیادی از دیتا رو برای ما فشرده می کنه

پس یه جاهایی وقتی ما میخوایم بفهمیم توی داده ها چخبره میایم سراغ خوشه بندی --> حجم دیتا زیاده و ما نمی تونیم همه این دیتا رو با هم ببینیم پس میایم خوشه بندی می کنیم اینا رو و بعد هر خوشه رو آنالیز میکنیم

Application of Cluster analysis

- Summarization of Brain Network in ADHD



8

-
کاربر خوشه بندی توی تحلیل مسائل پزشکی:
فعالیت هایی که توی مغز ما داریم روی هر شخصی متفاوت است

افرادی که پیش فعالی یا ADHD
و اطلاعات مغزی افراد عادی هم گرفتیم

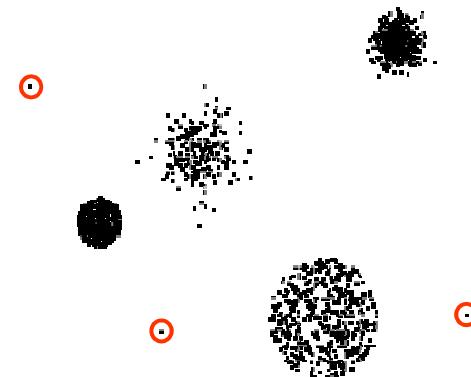
با اینکه ما نمی دونیم کیا ADHD دارد یا کی ندارد ولی تا می دیمش دست الگوریتم های خوشه
بندی یکسری الگوهایی رو استخراج می کنند و می گن اونایی که پیش فعالی دارند مغز پیچیده تری
دارند نسبت به اونایی که ندارند و مغزشون فعال تر است

Deviation/Anomaly/Change Detection

Detect significant deviations from normal behavior

Applications:

- Credit Card Fraud Detection



- Change Detection بحث است:

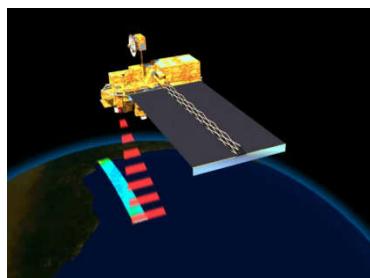
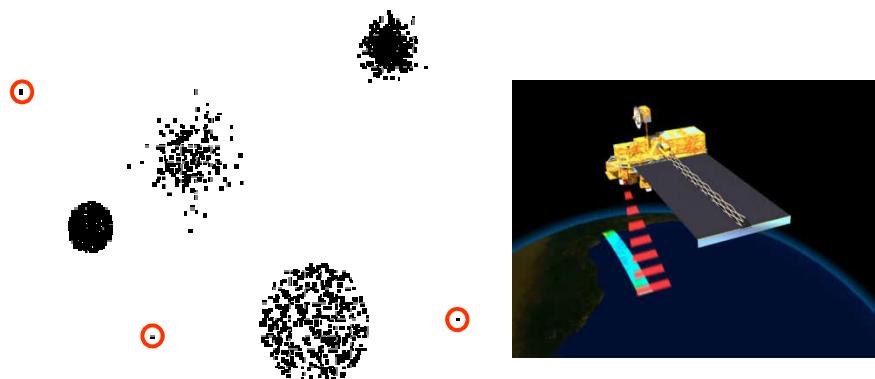
بعضی وقت ها ما دنبال رکوردهایی هستیم که متفاوت با بقیه است و خیلی تعدادشون کمeh ولی می دونیم فقط متفاوت است و نمیدونیم چی داره ولی می دونیم که متفاوت هست فقط مثل تشخیص کارت های تقلبی

Deviation/Anomaly/Change Detection

Detect significant deviations from normal behavior

Applications:

- Credit Card Fraud Detection
- Network Intrusion Detection



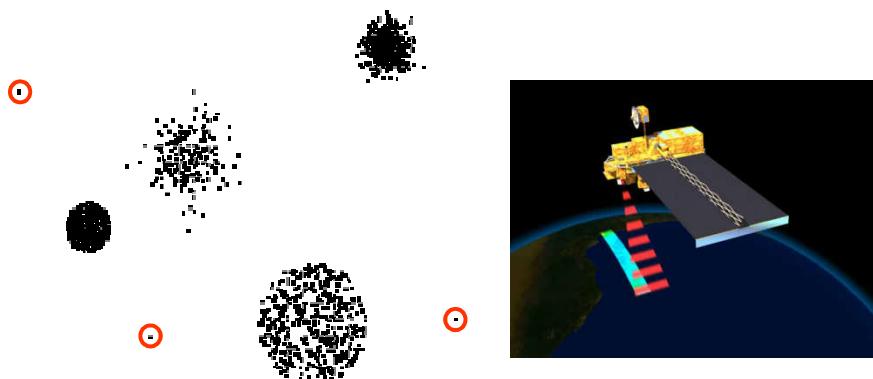
– تشخيص نفوذ شبکه

Deviation/Anomaly/Change Detection

Detect significant deviations from normal behavior

Applications:

- Credit Card Fraud Detection
- Network Intrusion Detection
- Identify anomalous behavior from sensor networks for monitoring and surveillance.



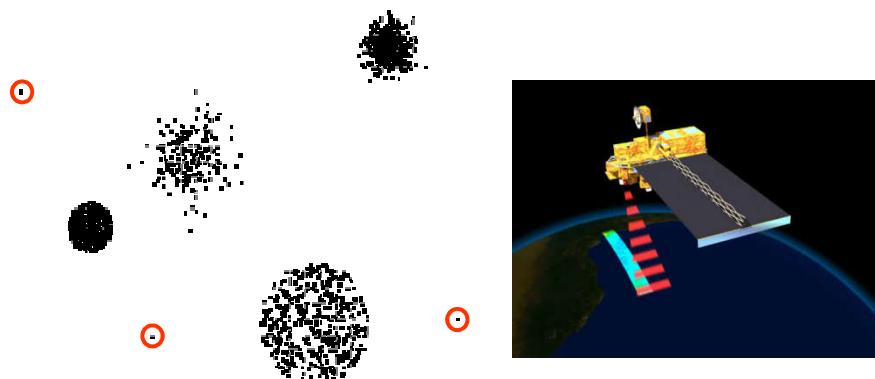
- شناسایی رفتار غیر عادی از شبکه های حسگر برای نظارت و نظارت

Deviation/Anomaly/Change Detection

Detect significant deviations from normal behavior

Applications:

- Credit Card Fraud Detection
- Network Intrusion Detection
- Identify anomalous behavior from sensor networks for monitoring and surveillance.
- Detecting changes in the global forest cover.



- تشخیص تغییرات در پوشش جهانی جنگل.

Major Issues in Data Mining (1)

- Mining Methodology
 - Mining various and new kinds of knowledge (New Question)
 - Mining knowledge in **multi-dimensional** space (Traffic[#+Speed])
 - Data mining: An interdisciplinary effort (bug mining)
 - Boosting the power of discovery in a networked environment (hybrid)
 - **Handling noise, uncertainty, and incompleteness** of data
 - Pattern evaluation and pattern- or constraint-guided mining
(ADHD types details)

چالش هایی که توی دیتا ماینینگ باهاش رو به رو هستیم:

اولین مسئله ای که داریم اینه که تکنولوژی که میخوایم باهاش کار بکنیم به چه صورت است -->ینی روشی که میخوایم باهاش داده هارو استخراج بکنیم چوری است ینی یک بخشی از تحقیقات روی روش مرکز است

یک بحث دیگه ای که داریم اینه که چقدر ما میخوایم بین رشته ای روش مانور بدیم:

ما می تونیم با بحث های داده کاوی فارغ از اینکه اون کسب و کارو بدونیم برخورد بکنیم و یه

جاهایی نه می تونیم یه گام جلوتر بزنیم و از دید کسب و کار هم به مسئله نگا بکنیم مثل bug

--> خیلی مهمه یک نرم افزاری که تولید میشه خطای نداشته باشه یا باگ نداشته باشه -->

برای بحث bug mining چی کار میشه کرد با دیتا ماینینگ؟ میان از دنیای نرم افزار به مسئله نگا

می کنند ینی میگن باگ کی پیدا میشه وقتی پیدا میشه که اون توسعه دهنده حوصله کافی نداره و

کیفیت کدش پایینه و وقتی که کیفیت کد پایینه غیر از باگ داشتن باعث میشه کد ها طولانی تر بشه و

رفتار بهینه ای نداشته باشه

Major Issues in Data Mining (1)

- Mining Methodology
 - Mining various and new kinds of knowledge(New Question)
 - Mining knowledge in **multi-dimensional** space (Traffic[#+Speed])
 - Data mining: An interdisciplinary effort (bug mining)
 - Boosting the power of discovery in a networked environment (hybrid)
 - **Handling noise, uncertainty, and incompleteness** of data
 - Pattern evaluation and pattern- or constraint-guided mining
(ADHD types details)
- User Interaction
 - Interactive mining(Search Engine(Similar))
 - Incorporation of background knowledge(Multi Judge)
 - Presentation and visualization of data mining results
(Better Presentations)

- ارتباط با کاربر:

مثل توی گوگل وقتی که میخوایم یه چیزی رو سرچ بکنیم : نحوه تعاملش به این صورت است که ما براساس تک تک سرچ هایی که داریم میکنیم داریم اطلاعات اون رو بیشتر میکنیم و اون سعی میکنه با توجه به اطلاعات که از ما گرفته پیشنهاد بهتری به ما بده پس یکی بحث Interactive mining است ینی ما چجوری اطلاعات کاربر رو بگیریم و توی داده کاوی ازش استفاده بکنیم: 1: ما یک اطلاعاتی از گذشته داریم و یک اطلاعاتی هم الان به دست میاد 2: ما چطور می خوایم اطلاعات داده کاوی رو به کاربر نشون بدیم؟ توی چه بسترهای میخوایم نشون بدیم ... 3:

Major Issues in Data Mining (2)

- Efficiency and Scalability
 - Efficiency and scalability of data mining algorithms
 - Parallel, distributed, stream, and incremental mining methods(|.|)

-
کارایی و مقیاس پذیری --> الگوریتم و اون تکنیک داده کاوی ما قراره چقدر زمان ببره

بحث توزیع شدگی: ینی تکنیک رو بتونیم توزیع بکنیم روی داده های مختلف کار بکنه

Major Issues in Data Mining (2)

- Efficiency and Scalability
 - (running time of a data mining algorithm must be predictable)
 - Efficiency and scalability of data mining algorithms
 - Parallel, distributed, stream, and incremental mining methods(.)
- Diversity of data types
 - Handling complex types of data (Time Series)
 - Mining dynamic, networked, and global data repositories(Social Network)
- Data mining and society
 - Social impacts of data mining (The social Dilemma!)
 - Privacy-preserving data mining(Fanavard)
 - Invisible data mining(Search Engine)

-
تنوع دیتا تایپ: بعضی وقت ها ما مسئله رو به صورت جدول می گیریم و یک تکنیک داده کاوی روش میدیم ولی بعضی وقت ها هم نه ینی تبدیلش میکنیم به یک فرمی که تکنیک داده کاوی روش عمل بکنه

ابعاد اجتماعی کارهای داده کاوی است: یک تکنیک داده کاوی توسعه میدیم و کمتر به ابعاد اجتماعیش فکر میکنیم

Where to Find References? DBLP, CiteSeer, Google

Data mining and KDD (SIGKDD)

- Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
- Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD"

Database systems (SIGMOD)

- Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
- Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc. "

AI & Machine Learning

- Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
- Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.

Where to Find References? DBLP, CiteSeer, Google

Web and IR

Conferences: SIGIR, WWW, CIKM, etc.

Journals: WWW: Internet and Web Information Systems

..

Statistics

Conferences: Joint Stat. Meeting, etc.

Journals: Annals of statistics, etc.

..

Visualization

Conference proceedings: CHI, ACM-SIGGraph, etc.

Journals: IEEE Trans. visualization and computer graphics, etc.

For next session

- با جستجو (یا کشf) یک کاربرد جذاب و واقعی(عملی) از داده کاوی که در کلاس تاکنون به آن اشاره نشده است است پیدا کنید و (حداقل در یک پاراگراف) آن را توضیح دهید و در سامانه آپلود کنید.

ملاحظات

- در یک فایل word ارسال شود.
- توضیحات شامل
 - ◆ یک تصویر مرتبط باشد.
 - ◆ مرجع مناسب و دقیق باشد.
- به بهترین گروه(طبق رای گیری) نمره فوق العاده مشارکت داده می‌شود.

The social Dilemma

 **The Social Dilemma**
2020 · Documentary/Docudrama · 1h 34m

[Overview](#) [Watch movie](#) [Reviews](#) [Cast](#) [Trailers & clips](#) [Quotes](#)

<https://www.thesocialdilemma.com> ::

The Social Dilemma
From the creators of Chasing Ice and Chasing Coral, **The Social Dilemma** blends documentary investigation and narrative drama to disrupt the disrupters, ...
[The Dilemma](#) · [The Film](#) · [Take a social media reboot](#) · [Take Action](#)

Cast >

Tristan Harris Jaron Lanier Skyler Gisondo Tim Kendall
Ben Kara Hayward Cassandra Sophia Hammons Isla

Watch movie [EDIT SERVICES](#)

 Watch now [Subscription](#)  Already watched  Want to watch

About

7.6/10 [IMDb](#) 85% [Rotten Tomatoes](#) 78% [Metacritic](#)

93% liked this film  
Google users

Tech experts from Silicon Valley sound the alarm on the dangerous impact of social networking, which Big