Управление и анализ на данни

доц. д-р Тодорка Димитрова
катедра ПКТ, Факултет КСУ

каб. 2526, тел. 9653453
dimitrova@tu-sofia.bg

СЪДЪРЖАНИЕ

Извадка от учебния план	4
Учебна програма	5
Учебни ресурси	7
Контрол	8
Въведение в тематиката на дисциплината	9

Извадка от учебния план

ДИСЦИПЛИНА

No MCT03

Управление и анализ на данни

Седмичен хорариум

Л	2
СУ	0
ЛУ	2
Самоподготовка	6
Общо	10

Контрол И

Кредити по ЕСТК 6

Лекции

Управление и обработка на данни

- Информационни процеси
- Събиране и подбор на данни
- Складове за данни (*Data Warehouse*)

Моделиране на данните за анализ и откриване на информация

- Методи и алгоритми за анализ
- Специализирани приложения на методи и алгоритми за анализ

Визуализация и разпространение на информацията

Програмни средства за анализ на данни

Упражнения

- Запознаване и работа с MS SQL Server и Business Intelligence Development Studio
- Експериментално изследване на методи и алгоритми за откриване на информация
  - Програмиране на модели за анализ
  - Използване на моделите в програмна система

Учебни ресурси

- Moodle: **pct.tu-sofia.bg**
- Internet

тема	адрес
Data Management	http://www.tech-faq.com/data-management.html
What is Data Analysis?	http://www.wisegeek.com/what-is-data-analysis.htm
Business Intelligence communications platform for professionals	http://businessintelligence.com/
Data Mining Community's Top Resource	http://www.kdnuggets.com/
Data Analysis, Statistics, and Probability	http://www.learner.org/courses/learningmath/data/
HTML, XML, Server Scripting, etc.	http://www.w3schools.com/

Контрол

- Текущ
  - Теоретични задачи
  - Лабораторни задачи
- Изпит
  - Тест върху теорията
  - Задача за програмиране

Въведение в тематиката на курса

Въведение

- Необходимост от управление на данните
- Значение на управлението на данни
- Информационни процеси и анализ на данни
  - Идентифициране, измерване, събиране и съхраняване на данните
  - Управление на потоци от данни (*Data Flow*)
  - Обработка на данни (*Data Processing*)
  - Извличане на информация (*Information Retrieval*)
  - Откриване на информация (*Data Mining*)

Необходимост от управление на данните

- Приложни области
  - бизнес
  - наука
  - технологии
  - административно управление
  - и др.

Необходимост от управление на данните

- Необходимост за бизнеса
  - лавинно нарастване на количествата събрани данни и необходимостта от съхранявенето им
    - от web приложения
    - от електронна търговия
    - от финансовите институции, които използват on-line транзакции
  - усилване на конкуренцията
    - връзка с икономическата криза
    - изисквания на потребителите
  - разширени възможности на потребителите за генериране и обработка на данни
    - нови технологии
    - разширяване на достъпността и приложимостта на технологиите
    - ниска цена на хардуера
    - широк кръг приложения

Необходимост от управление на данните

- Необходимост за науката
  - Ускорено събиране и съхраняване на данни
    - отдалечено събиране на данни чрез сателити
    - генно инженерство и биоинформатика
    - симулационно моделиране
- Традиционните техники за обработка са непригодни за анализ на неструктурирани и хетерогенни данни

Данни, информация и знание

- Необходимо е знание
- Знанието се изгражда на основата на натрупана информация
- Информацията често е скрита в данните
  - разкриването е сложен процес
  - разбирането на откритата информация е трудно
  - много данни изобщо не се използват

Значение на управлението на данни

- Оптимизира се съхраняването на данните
- Извлича се информация от данните
- Подпомага се изграждането на знания
- Улеснява се вземането на решения

Цел

- Подпомагане на извличането на потенциално полезна информация и изграждането на нови знания от натрупаните данни посредством
  - автоматизиран анализ на данните и
  - представяне във форма, достъпна за потребителите

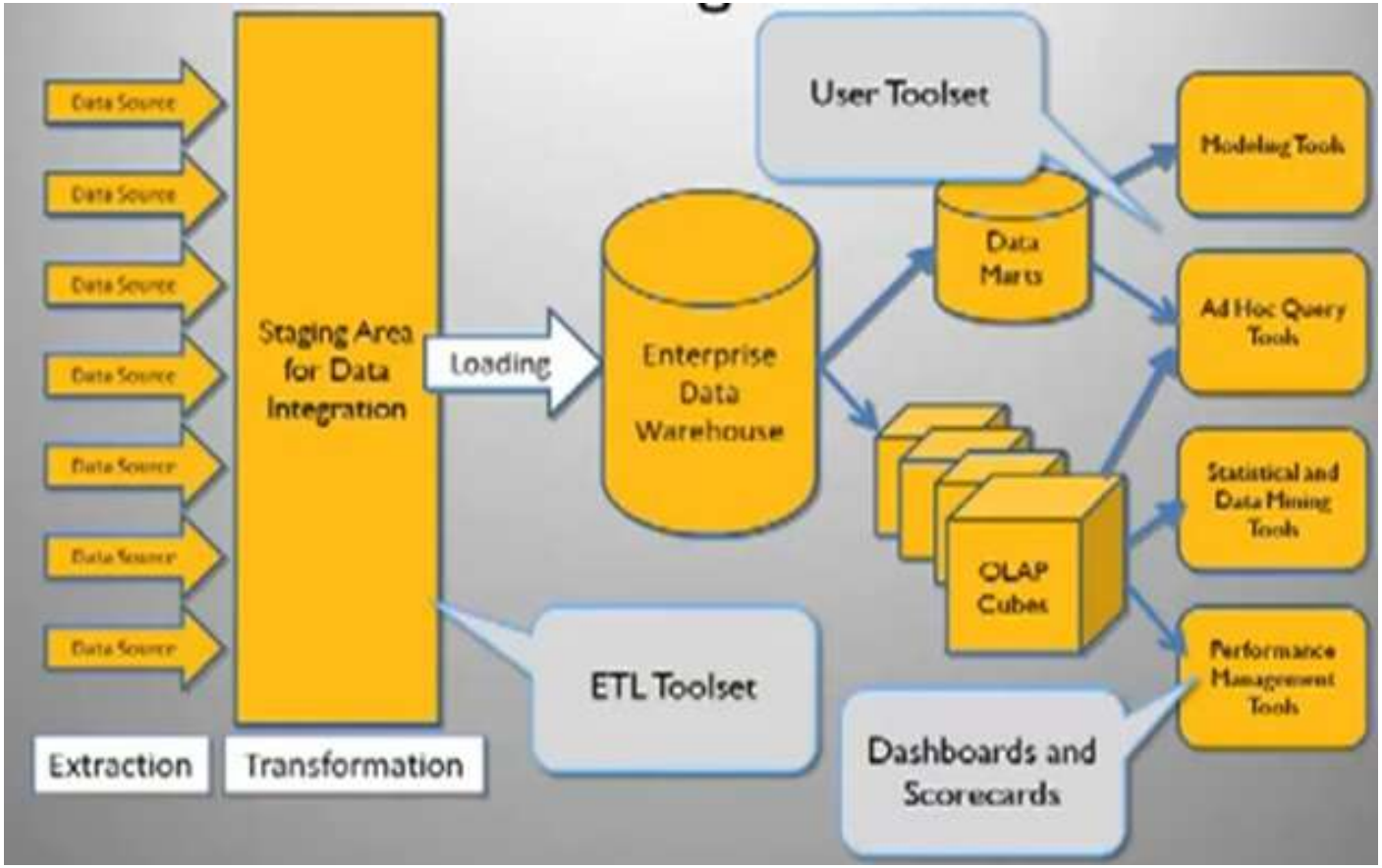
Процеси

- Разбиране на бизнеса
- Разбиране на данните
- Подготовка на данните
- Създаване на модели за анализ
- Апробиране и оценка на моделите
- Анализ на данните
- Прилагане на резултатите от анализа

Бизнес интелигентност

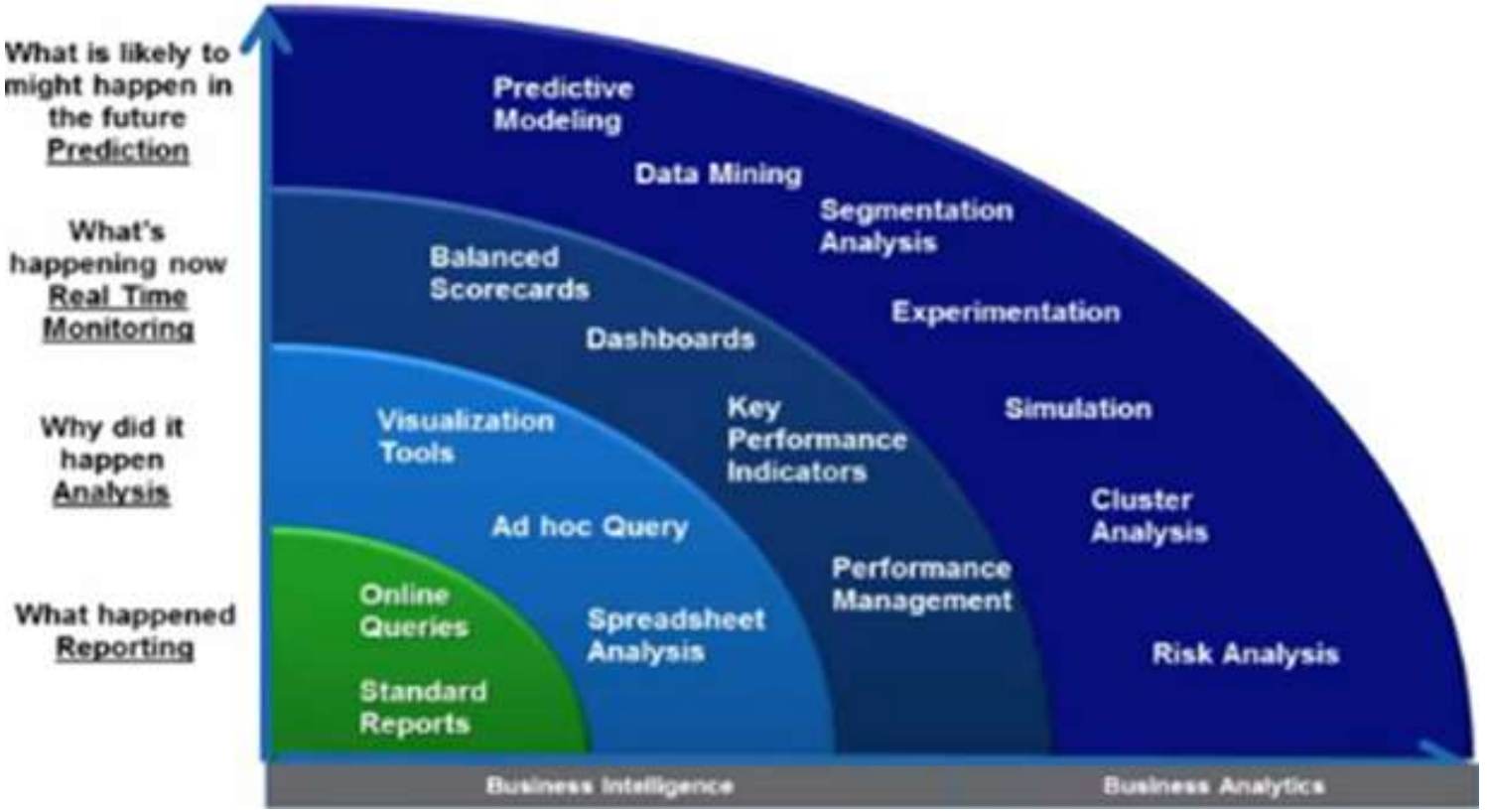
- Едно определение
  - технологии – методи, средства, процеси - за преобразуване на данните от бизнеса в информация, на информацията в знание и на знанието в планиране на бизнеса, водещо до увеличаване на ефективността, ефикасността и печалбите

Технологии



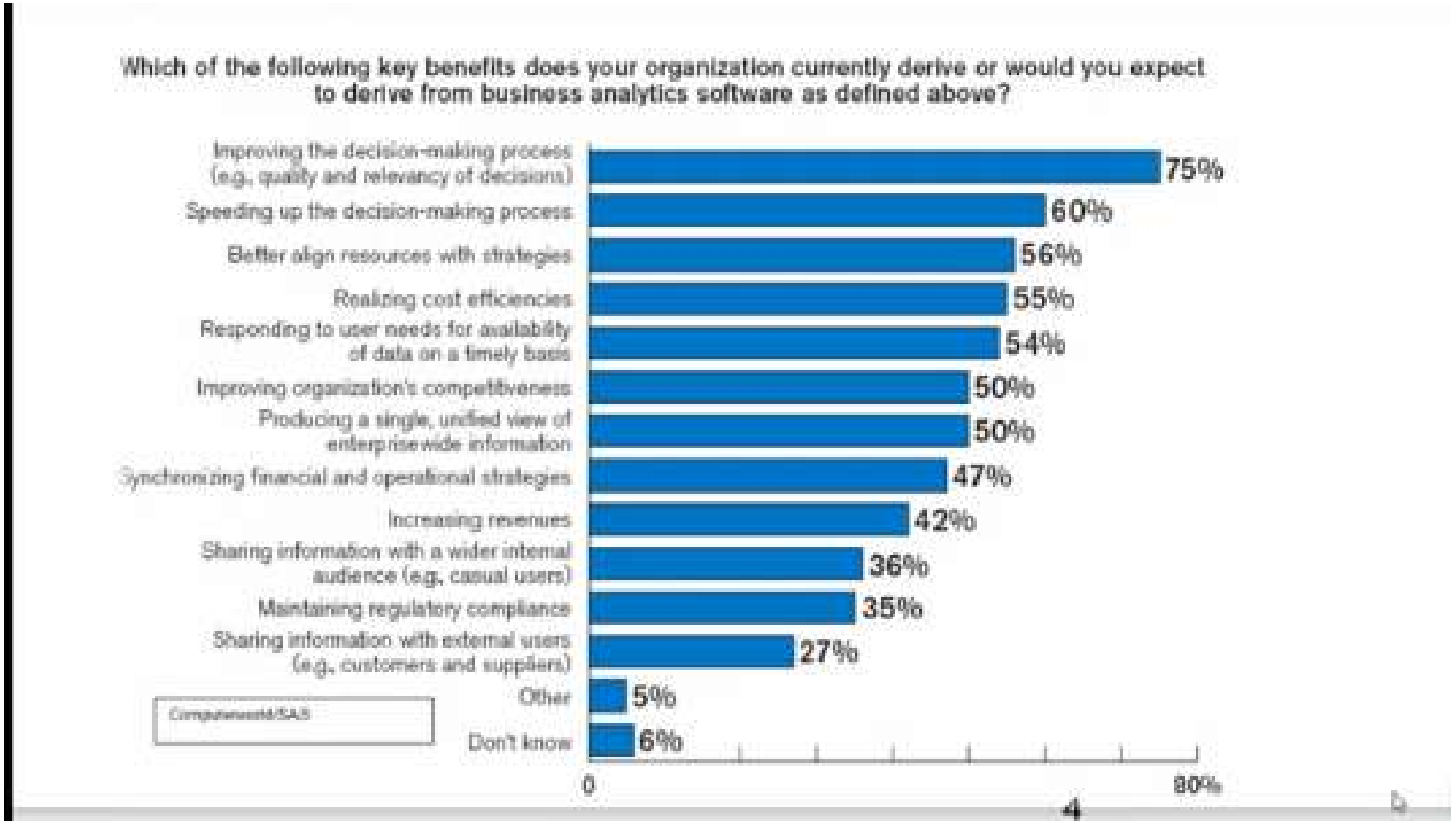
tribridge.com

Бизнес интелигентност и анализ



tribridge.com

Значение за бизнеса



tribridge.com

Данни

разбиране на данните

Събиране на данни

- Идентифициране, измерване, събиране и съхраняване на данните
  - данни – регистрирани стойности на атрибути на обекти и процеси от реалния свят
  - атрибути – свойства или характеристики на обект
    - примери: размери, цвят, националност и др.
    - множество атрибути описват обекта
    - правилата за изменение на стойностите на атрибутите описват поведението на обекта
  - стойности на атрибутите
    - числа или символи, определени по избрани скали
    - един атрибут – разклични стойности за различни обекти
  - измерителни скали
    - оценяват стойността на атрибут
    - определят допустимите граници на различните стойности

Типове атрибути и скали

- Номинални
  - код – ЕГН, цвят на очите, име
- Ординални
  - ранжиране – оценка от изпит, образователна степен, курс, размер на дрехи и обувки
- Интервални
  - равни интервали между стойностите - дати, температури
- Пропорционални
  - получени от преобразуване на други атрибути - брой, продължителност, разстояние и други

Типове атрибути и свойства

- Характеристични свойства на атрибутите
  - различимост $= \neq$
  - подредба $< >$
  - събиране $+ -$
  - умножение $* /$

- Притежание на свойствата
  - номинални различимост
  - ординални различимост, подредба
  - интервални различимост, подредба , събиране
  - пропорционални всички

Типове атрибути и операции

тип	значение на стойностите	примери	операции
Номинални	информация за идентификация (=, ≠)	пощенски код, факултетн номер, числа, цвят на очите, пол: {*male, female*}	мода, ентропия, корелация, χ^2 тест
Ординални	информация за подреждане на стойности (<, >)	качество {*good, better, best*}, степен, пореден номер	медиана, персентили, рангова корелация и др.
Интервални	съществени са разликата между стойностите и измервателната скала (+, -)	дати от календара, температура по Celsius или Fahrenheit	средно аритметично, стандартно отклонение, Pearson корелация, *t* и *F* тест
Пропорционални	съществени са разликата между стойностите и пропорцията между тях (*, /)	валути, възраст, дължина, електрически величини	средно геометрично, процентно отклонение и др.

Типове атрибути и трансформации

тип	трансформации	пояснение
Номинални	всички пермутации на стойностите	ако всички факултетни номера бъдат подменени, това не променя значението им
Ординални	запазване на функцията на изменение на стойности *new_value = f(old_value)* , където *f* е монотонна функция	{good, better, best} могат да се заменят с {1, 2, 3} или с { 0.5, 1, 10}.
Интервални	*new_value =a * old_value + b* където a и b са константи	температурните скали на Fahrenheit и Celsius по нулата и размера на деленията
Пропорционални	*new_value = a * old_value*	дължината може да се мери в метри или мили

Дискретни и непрекъснати атрибути

- Дискретни атрибути
  - краен брой, ограничено множество от стойности
  - често се представят с цели числа (integer)
  - бинарни атрибути – частен случай – приемат една от две стойности

- Непрекъснати атрибути
  - реални стойности
  - на практика се закръгляват до определен знак след десетичната точка .
  - представят се чрез реално число (floating-point)

Структури от данни

- Записи
  - матрици от данни
  - документи
  - данни за транзакции
- Графи
  - World Wide Web
  - Молекулярни структури
- Подредени множества
  - пространствени данни
  - времеви данни
  - последователности
  - генетични последователности
- Свойства на структурираните данни
  - размерност
  - резолюция – стойностите зависят от скалата
  - непълно множество - не е задължително всички стойности да присъстват

Записи

- Колекция от записи, всеки от които съдържа стойности на набор от атрибути

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

Матрица от данни

- Ако обектите от множеството имат фиксиран брой числови атрибути, могат да се представят в многомерно пространство, в което всеки атрибут задава една размерност - проекция

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Документи

- Всеки документ се представя като вектор от термини
  - всеки термин е компонент (attribute) на вектора
  - стойността на всеки компонент е число, показващо честотата на срещане на съответния термин в документа

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

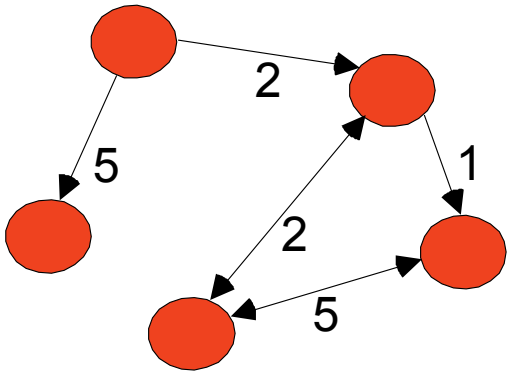
Данни за транзакции

- Специален вид записи
  - записите съдържат списък от елементи
  - Пример

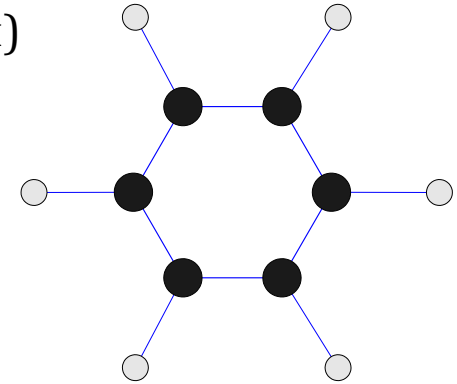
TID	*Items*
1	история, математика, физика
2	литература, история, спорт
3	спорт
4	математика, география
5	математика, физика

Графи

- Пример: web site



- Пример: молекула C_6H_6 (бензин)



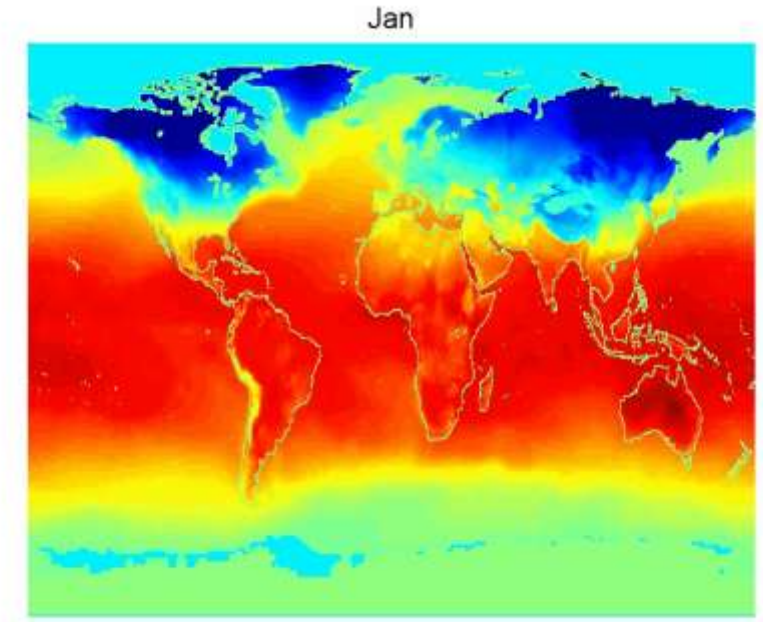
Подредени множества

- Genomic sequence

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Подредени множества

- Пространствено-времеви последователности
  - Пример: средни месечни температури

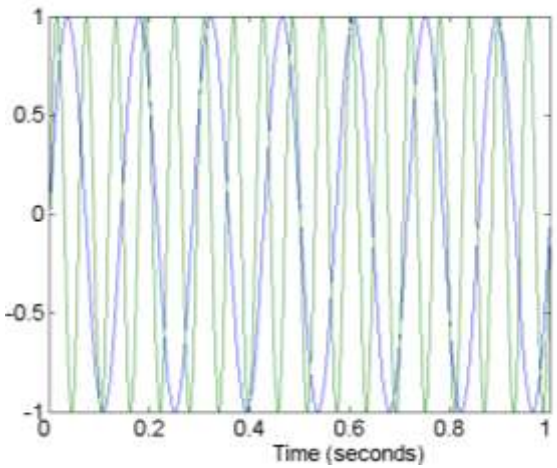


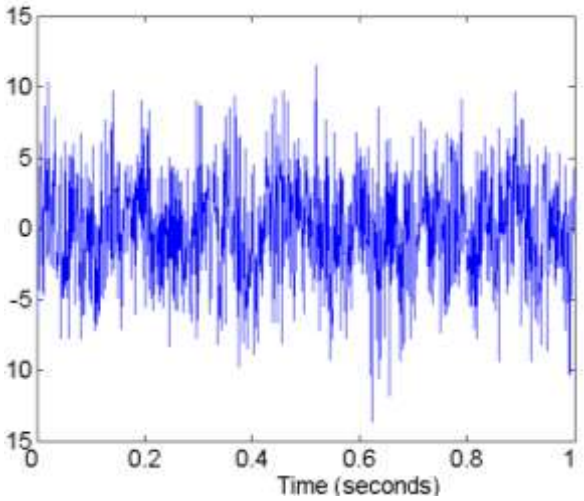
Качество на данните

- Проблеми на качеството
  - какви проблеми
  - как се отстраняват
- Примери за лошо качество
  - шум
  - липсващи стойности
  - дублиране на стойности

Шум

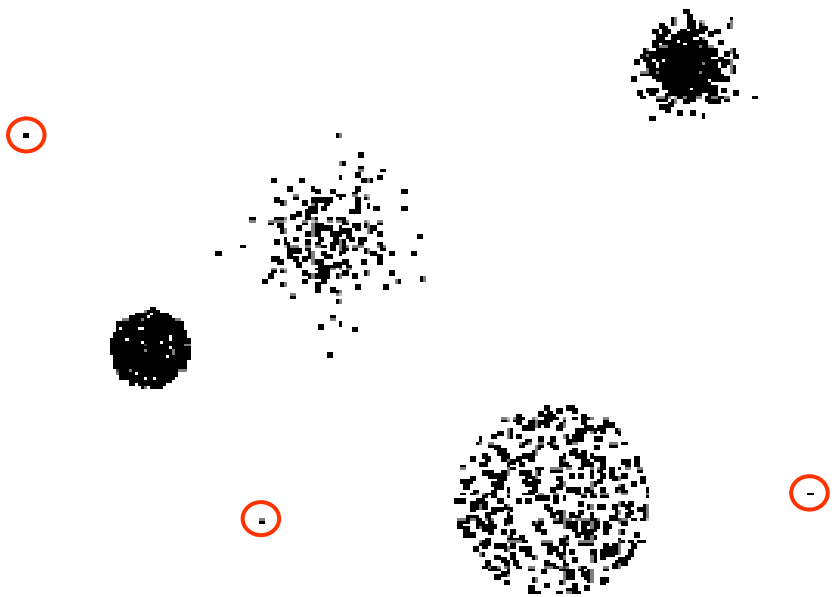
- Изменение на оригиналната стойност под външни въздействия
  - Пример: запис на глас





Шум

- Нехарактерни екземпляри



Липсващи стойности

- Причини
  - липса на отговор при запитване – напр. “За кого ще гласуваш?”
  - атрибутът е неприложим за екземпляра – напр. децата нямат доходи
- Обработване на липсващи стойности
  - елиминиране на обекта
  - игнориране на стойността при анализ
  - апроксимиране на стойност
  - заместване с възможните стойности

Дублиране

- При смесване на данни от различни източници
- Две и повече значения на един атригут за един екземпляр
  - напр. един човек – два адреса
- Изчистване
  - обработка на дублираните стойности