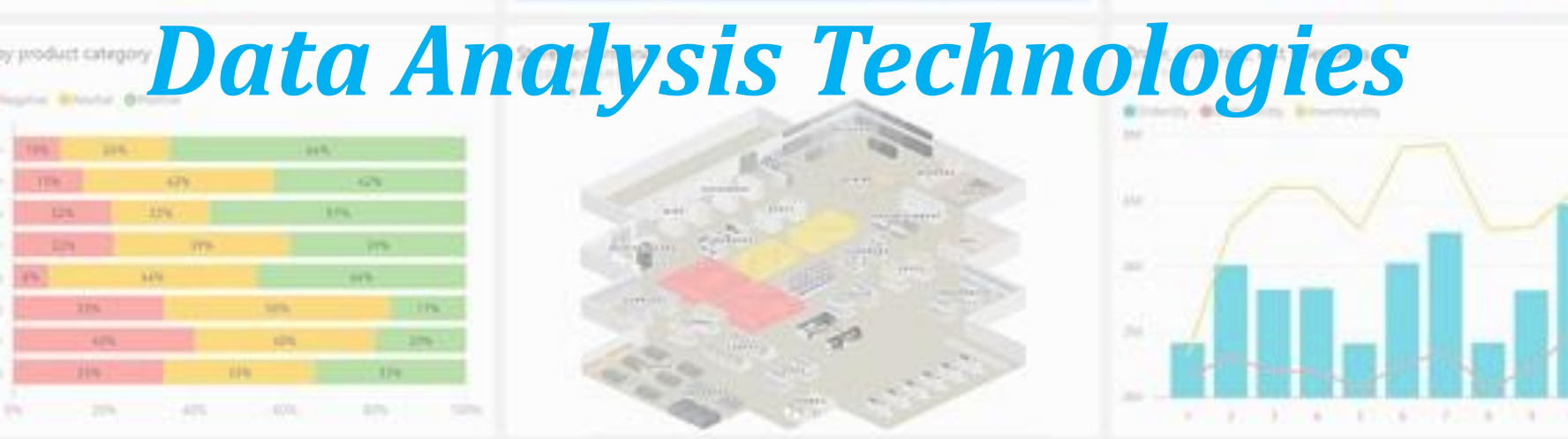


Технології аналізу даних



Data Analysis Technologies

Мета, завдання і предмет дисципліни

Метою вивчення дисципліни «Технології аналізу даних» є надання фундаментальних теоретичних знань і набуття практичних навичок з питань формування, дослідження та всебічного аналізу даних у різних галузях сферах людської діяльності.

Завданням вивчення дисципліни «Технології аналізу даних» є надання студентам ґрунтовних знань в області аналітичних досліджень інформаційного простору, вивчення методів створення, добування, консолідації, переробки, трансформації та аналізу даних.

Предметом вивчення дисципліни є основні положення й методи аналізу даних та їх комп'ютерна реалізація з використанням аналітичних платформ та спеціалізованих мов програмування.

Тематичний план дисципліни

Тема 1.
Передобробка
даних

Тема 2.
Асоціація даних

Тема 3.
Кластеризація
даних

Тема 4.
Класифікація та
регресія даних

Тема 5.
Технології
інтелектуальної
обробки даних

Тема 6.
Інструментальні
засоби аналізу
даних

Тема 7.
Створення моделі
даних

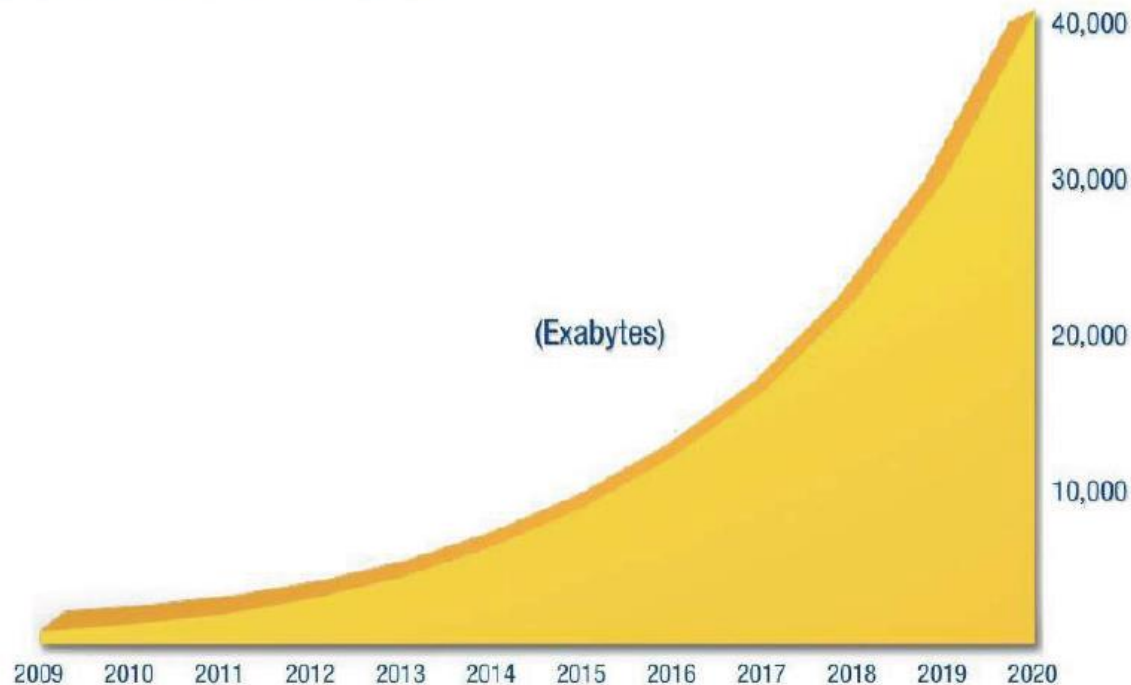
Тема 8.
Побудова
аналітичних звітів



Кількість даних у світі

90 % даних, набутих людством упродовж свого існування, отримано за останні два роки

The Digital Universe: 50-fold Growth from the Beginning of 2010 to the End of 2020



За інформацією ІВМ у наступні два роки (до 2022 року) кількість даних подвоється і досягне величини 40 000 ексабайтів (1 exabyte= 2^{60} bytes).

Фундаментальні терміни аналізу даних

Особливе місце в дисципліні займають два фундаментальні терміни, на яких власне і тримається *Data Science* – **штучний інтелект** та **бізнес-аналітика**

Artificial intelligence, AI (штучний інтелект) – розділ комп'ютерної лінгвістики та інформатики, що опікується формалізацією проблем та завдань, які імітують поведінку людини.

Artificial Intelligence



Артур Семюель (1901-1990)

Піонер у галузі штучного інтелекту, машинного навчання та комп'ютерних ігор. Програма *Samuel Checkers-playing* – одна з перших програм у світі, що самостійно навчаються, є ранньою демонстрацією фундаментального поняття штучного інтелекту.

«Машинне навчання – це галузь досліджень, яка дає комп'ютерам здатність навчатися без того, щоб їх явно програмували»

Артур Семюель, 1959р.

Business Intelligence



Ханс Петер Лун (1896-1964)

Уперше запропонував термін Business Intelligence (BI). Народження BI датується 1958 р., коли Ханс Петер Лун опублікував в IBM System Journal статтю «A Business Intelligence System».

Business Intelligence розуміють як «інструменти для аналізу даних, побудови звітів і запитів, що можуть допомогти бізнес-користувачам подолати море даних для того, щоб синтезувати з них значущу інформацію».

Технології *Data Science* дають можливість перетворювати дані в інформацію, а потім інформацію в знання.

Дані – це реальність, яку комп'ютер записує, зберігає і обробляє – це «сирі дані».

Інформація – це дані, яка людина в стані зрозуміти.

Знання – це інформація, що використовується в бізнесі для прийняття рішень.

Big Data

До категорії «великі дані» (*Big Data*) відноситься інформація, яку навряд чи можна обробляти традиційними способами, в тому числі слабо структуровані дані, медіа і випадкові об'єкти

Big Data: найбільш просте визначення

Термін «великі дані» відноситься до управління та аналізу великих обсягів даних, розмір яких перевищує можливості типових баз даних із занесення, зберігання, управління та аналізу інформації.

Big Data: більш складне визначення

Проблема не в тому, що організації створюють величезні обсяги даних, а в тому, що більша їх частина представлена в форматі, що не відповідає традиційному структурованому формату бази даних, – це веб-журнали, відеозаписи, текстові документи, машинний код або, наприклад, геопросторові дані.

Big Data: найкраще визначення

Поняття великих даних має на увазі роботу з інформацією величезного обсягу і різноманітного складу, яка часто оновлюється і знаходиться в різних джерелах з метою збільшення ефективності роботи, створення нових продуктів і підвищення конкурентоспроможності.

Таким чином *Big Data* – це дані, які великі не стільки за обсягом, скільки за складністю.

Методологія аналізу даних

*Ключова відмінність Data Mining
від інших методів аналізу даних*

Традиційні методи аналізу даних орієнтовані на перевірку заздалегідь сформульованих гіпотез, що становить основу оперативної аналітичної обробки

Одне з основних положень Data Mining – пошук неочевидних закономірностей і здатність самостійно будувати гіпотези про взаємозв'язки

Методи аналізу даних

Статистичні методи

- дескриптивний аналіз і опис вихідних даних.
- аналіз зв'язків (кореляційний і регресійний аналіз, факторний аналіз, дисперсійний аналіз).
- багатомірний статистичний аналіз (компонентний аналіз, дискримінантний аналіз, багатовимірний регресійний аналіз, канонічні кореляції).
- аналіз часових рядів (динамічні моделі і прогнозування).

Кібернетичні методи

- штучні нейронні мережі (розпізнавання, кластеризація, прогноз);
- еволюційне програмування;
- генетичні алгоритми;
- асоціативна пам'ять (пошук аналогів, прототипів);
- нечітка логіка;
- дерева рішень;
- системи обробки експертних знань.

Порівняльна характеристика методів *Data Mining*

Алгоритм	Точність	Масштабованість	Інтерпретація	Придатність до використання	Трудомісткість	Різносторонність	Швидкість	Популярність
Класичні методи (лінійна регресія)	Середня	Висока	Середня	Висока	Середня	Середня	Висока	Низька
Нейронні мережі	Висока	Низька	Низька	Низька	Середня	Низька	Дуже низька	Низька
Методи візуалізації	Висока	Дуже низька	Висока	Висока	Дуже висока	Низька	Надзвичайно низька	Висока
Дерева рішень	Низька	Висока	Висока	Середня	Висока	Висока	Середня	Висока
Поліноміальні нейронні мережі	Висока	Середня	Низька	Середня	Середня	Середня	Середня	Середня

Основні методи аналізу даних

- ✓ Класифікація (Classification)
- ✓ Кластеризація (Clustering)
- ✓ Асоціація (Association)
- ✓ Прогнозування (Forecasting)
- ✓ Візуалізація (Visualization)

Найбільш авторитетний веб-сайт з інформацією про роботу *Glassdoor*, який щорічно публікує 50 найбільш актуальних професій у США, чотири роки поспіль віддає перевагу фахівцям в області аналізу даних.



Top 10 Best Jobs in America in 2019

Rank	Job Title	Median Base Salary	Job Satisfaction	Job Openings
1	Data Scientist	\$108,000	4.3	6,510
2	Nursing Manager	\$83,000	4.0	13,931
3	Marketing Manager	\$82,000	4.2	7,395
4	Occupational Therapist	\$74,000	4.0	17,701
5	Product Manager	\$115,000	3.8	11,884
6	Devops Engineer	\$106,000	4.1	4,657
7	Program Manager	\$87,000	3.9	14,753
8	Data Engineer	\$100,000	3.9	4,739
9	HR Manager	\$85,000	4.2	3,908
10	Software Engineer	\$104,000	3.6	49,007

Source: Glassdoor Economic Research ([Glassdoor.com/research](https://www.glassdoor.com/research))

Гарвардський бізнес-огляд визначив професіонала в галузі аналізу даних як «найпривабливішу роботу 21-го століття»

Data Scientist:

The Sexiest Job of the 21st Century

**Meet the people who
can coax treasure out of
messy, unstructured data.**

*by Thomas H. Davenport
and D.J. Patil*

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants



PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

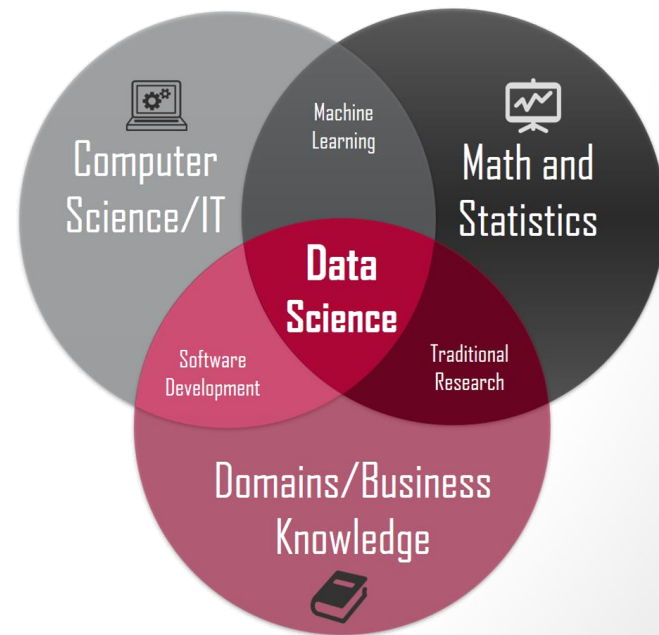
DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Ffare, D3.js, Tableau

Data Scientist – це не програміст. Це фахівець з чудовими кросдисциплінарними знаннями математики, штучного інтелекту, інформаційних технологій та бізнесу і суперздатностями до аналізу.



World Economic Forum: перспективи Data Science



COMMITTED TO
IMPROVING THE STATE
OF THE WORLD

Insight Report

The Future of Jobs Report 2018

Centre for the New Economy and Society



COMMITTED TO
IMPROVING THE STATE
OF THE WORLD

Insight Report

Data Science in the New Economy

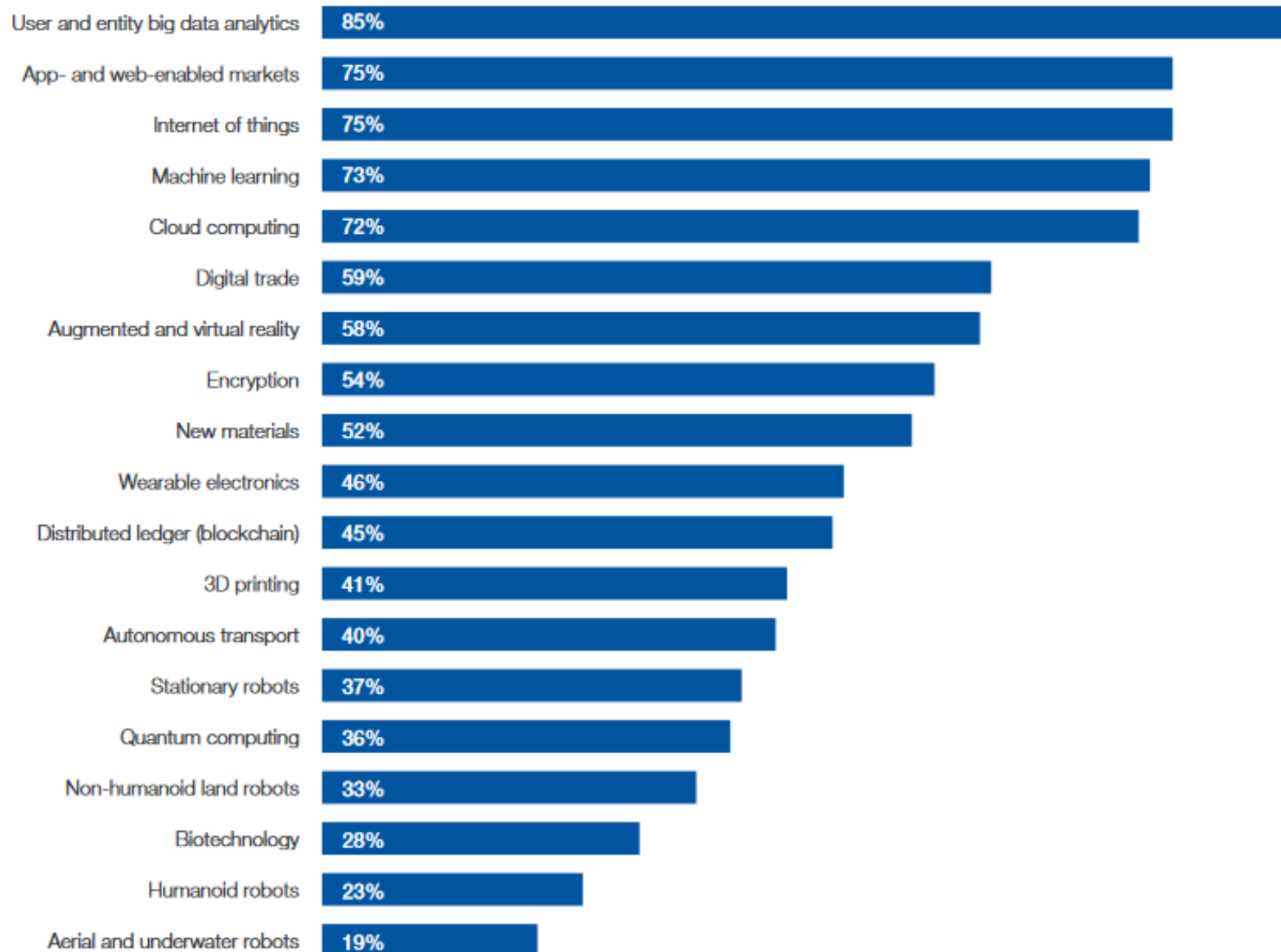
A new race for talent in the
Fourth Industrial Revolution

Centre for the New Economy and Society

July 2019



Figure 2: Technologies by proportion of companies likely to adopt them by 2022 (projected)



Source: Future of Jobs Survey 2018, World Economic Forum.

Table 3: Examples of stable, new and redundant roles, all industries




Stable Roles	New Roles	Redundant Roles
Managing Directors and Chief Executives General and Operations Managers* Software and Applications Developers and Analysts* Data Analysts and Scientists* Sales and Marketing Professionals* Sales Representatives, Wholesale and Manufacturing, Technical and Scientific Products Human Resources Specialists Financial and Investment Advisers Database and Network Professionals Supply Chain and Logistics Specialists Risk Management Specialists Information Security Analysts* Management and Organization Analysts Electrotechnology Engineers Organizational Development Specialists* Chemical Processing Plant Operators University and Higher Education Teachers Compliance Officers Energy and Petroleum Engineers Robotics Specialists and Engineers Petroleum and Natural Gas Refining Plant Operators	Data Analysts and Scientists* AI and Machine Learning Specialists General and Operations Managers* Big Data Specialists Digital Transformation Specialists Sales and Marketing Professionals* New Technology Specialists Organizational Development Specialists* Software and Applications Developers and Analysts* Information Technology Services Process Automation Specialists Innovation Professionals Information Security Analysts* Ecommerce and Social Media Specialists User Experience and Human-Machine Interaction Designers Training and Development Specialists Robotics Specialists and Engineers People and Culture Specialists Client Information and Customer Service Workers* Service and Solutions Designers Digital Marketing and Strategy Specialists	Data Entry Clerks Accounting, Bookkeeping and Payroll Clerks Administrative and Executive Secretaries Assembly and Factory Workers Client Information and Customer Service Workers* Business Services and Administration Managers Accountants and Auditors Material-Recording and Stock-Keeping Clerks General and Operations Managers* Postal Service Clerks Financial Analysts Cashiers and Ticket Clerks Mechanics and Machinery Repairers Telemarketers Electronics and Telecommunications Installers and Repairers Bank Tellers and Related Clerks Car, Van and Motorcycle Drivers Sales and Purchasing Agents and Brokers Door-To-Door Sales Workers, News and Street Vendors, and Related Workers Statistical, Finance and Insurance Clerks Lawyers

Source: Future of Jobs Survey 2018, World Economic Forum.

Note: Roles marked with * appear across multiple columns. This reflects the fact that they might be seeing stable or declining demand across one industry but be in demand in another.

Data Scientist Salaries

– San Francisco, CA Area

Salaries in \$ (USD)	Average	Salaries in \$ (USD)	Average
 Data Scientist Facebook 46 salaries See 2,149 salaries from all locations	\$138,713 per year	 Data Scientist IBM 2 salaries	About \$97k - \$170k
 Data Scientist Airbnb 19 salaries See 2,149 salaries from all locations	\$126,159 per year	 Data Scientist Tesla Motors 3 salaries See 2,149 salaries from all locations	\$103,625 per year
 Data Scientist Twitter 15 salaries See 2,149 salaries from all locations	\$134,861 per year	 Data Scientist PayPal 10 salaries	\$132,909 per year

Вакансії та заробітна плата фахівців з аналізу даних від філій лідерів світових компаній у м. Сан-Франциско (США)

Microsoft Power BI

Power BI



Microsoft




```

leisch@galadriel:~/work/tnp
R> n <- 5
R> g <- gl(n, 100, n*100)
R> x <- rnorm(n*100) + sqrt(codes(g))
R> boxplot(split(x,g), col="lavender", notch=TRUE)
R> title(main="Notched Boxplots", xlab="Group", font.main=4, font.lab=1)
R>
R> ctl <- c(4.17,5.58,5.18,6.11,4.50,4.61,5.17,4.53,5.33,5.14)
R> trt <- c(4.81,4.17,4.41,3.59,5.87,3.83,6.03,4.89,4.32,4.69)
R> group <- gl(2,10,20,labels=c("Ctl","Trt"))
R> weight <- c(ctl,trt)
R> anova(lm.D9 <- lm(weight~group))

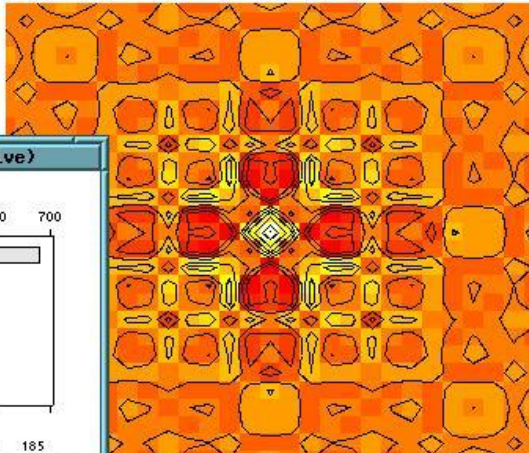
Analysis of Variance Table
Response: weight
          Df Sum Sq Mean Sq    F Pr(>F)
group      1  0.6882   0.6882  1.419  0.249
Residual  18  8.7293   0.4850

R>
R>

```

R Graphics: Device 2 (inactive)

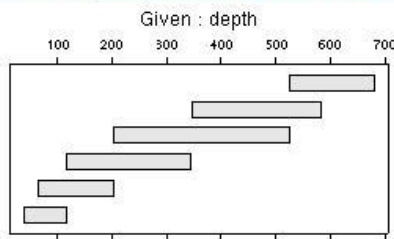
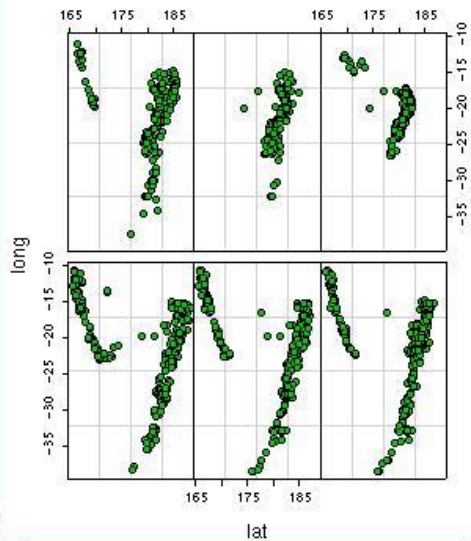
Math can be beautiful ...



$\cos(r^2)e^{-r^{16}}$

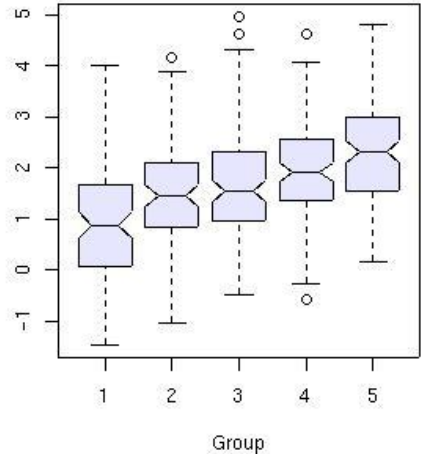
R Graphics: Device 3 (inactive)

Given: depth

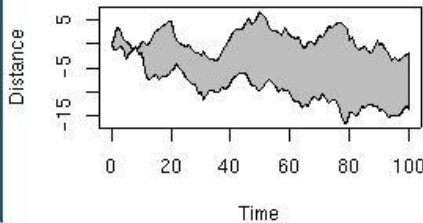
R Graphics: Device 4 (ACTIVE)

Notched Boxplots

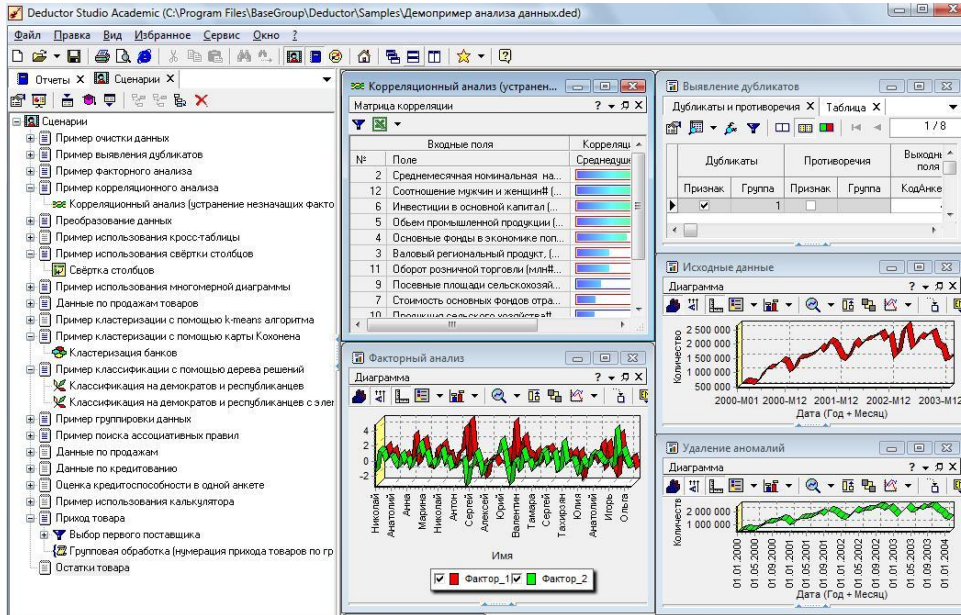


R Graphics: Device 5 (inactive)

Distance Between Brownian Motions



Deductor Studio




Корреляционный анализ (устранение...)
Матрица корреляции

№	Поле	Корреляция	Средний
2	Среднемесячная номинальная на...		
12	Соотношение мужчин и женщин# (...)		
6	Инвестиции в основную капитал (...)		
5	Объем промышленной продукции (...)		
4	Основные фонды в экономике поп...		
3	Валовый региональный продукт, (...)		
11	Оборот розничной торговли (млн...		
9	Посевные площади сельскохозяй...		
7	Стоимость основных фондов обра...		
10	Производство сельскохозяйствен...		

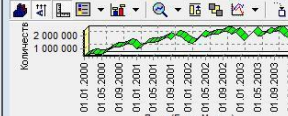
Выявление дубликатов
Дубликаты и противоречия X Таблица X

Дубликаты	Противоречия	Выходные поля		
Признак	Группа	Признак	Группа	КодАнкет
1				

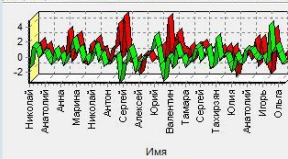
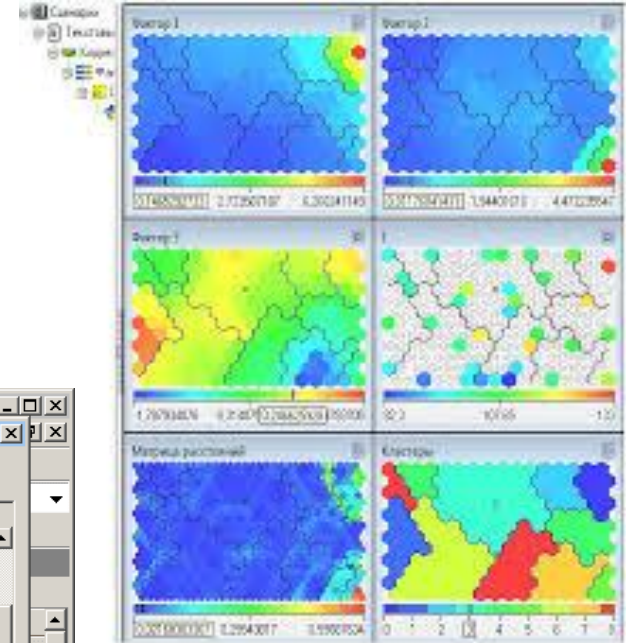
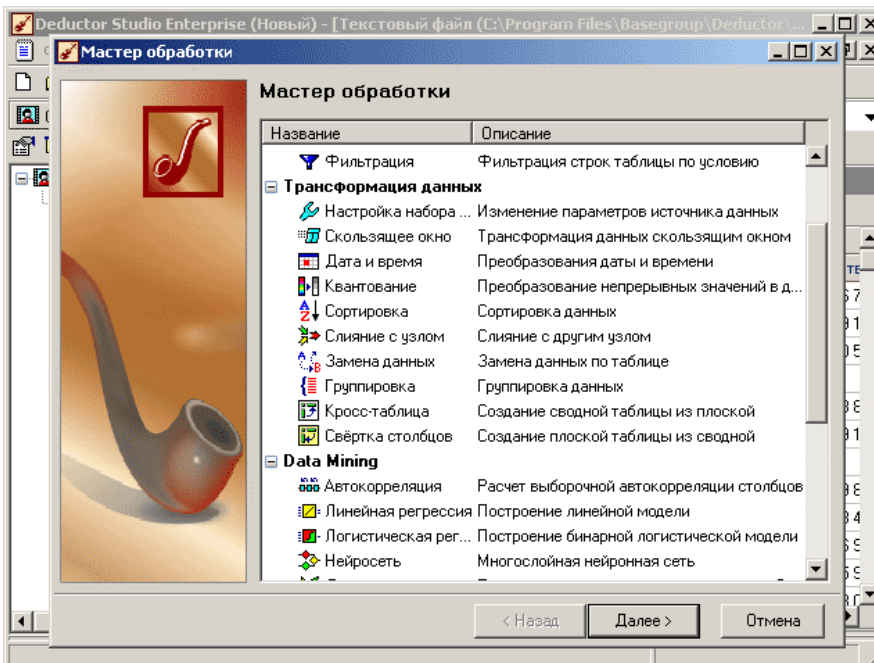
Исходные данные
Диаграмма



Удаление аномалий
Диаграмма



Факторный анализ
Диаграмма

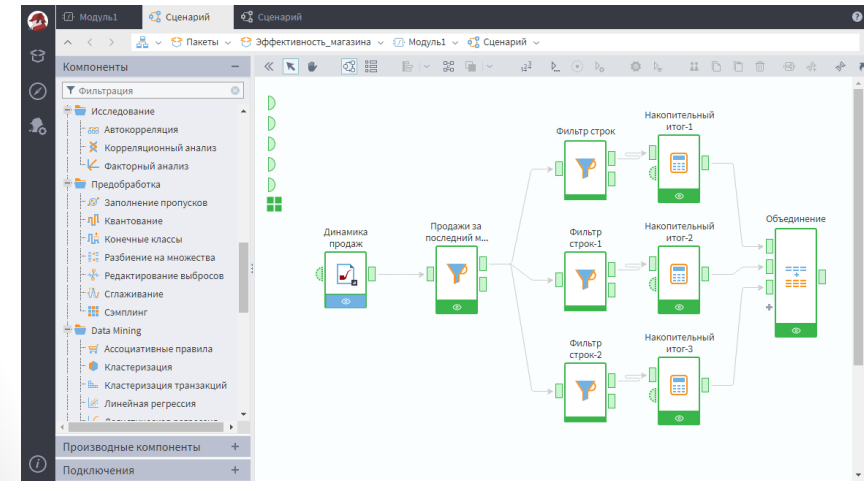
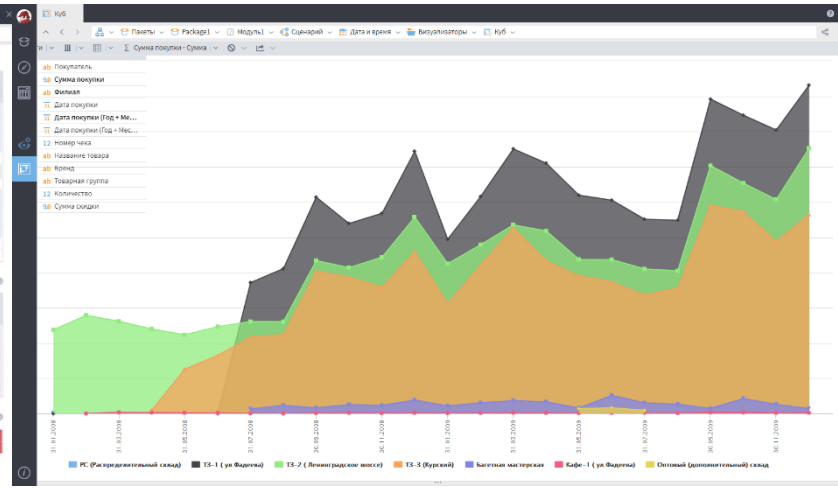
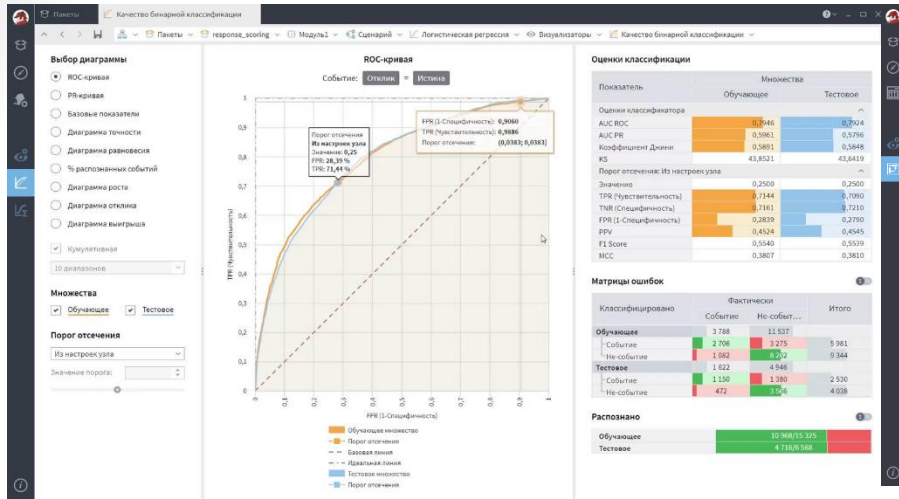




Мастер обработки

Название	Описание
Фильтрация	Фильтрация строк таблицы по условию
Трансформация данных	
Настройка набора ...	Изменение параметров источника данных
Скользящее окно	Трансформация данных скользящим окном
Дата и время	Преобразования даты и времени
Квантование	Преобразование непрерывных значений в д...
Сортировка	Сортировка данных
Слияние с узлом	Слияние с другим узлом
Замена данных	Замена данных по таблице
Группировка	Группировка данных
Кросс-таблица	Создание сводной таблицы из плоской
Свёртка столбцов	Создание плоской таблицы из сводной
Data Mining	
Автокорреляция	Расчет выборочной автокорреляции столбцов
Линейная регрессия	Построение линейной модели
Логистическая рег...	Построение бинарной логистической модели
Нейросеть	Многослойная нейронная сеть

< Назад Далее > Отмена

Loginom



Куб

Поло

- Название магазина
- Дата (месяц)
- Дата (день)
- Тип дня недели
- Средний чек
- Объем продаж, шт
- Количество позиций в чеке

Дата (месяц)	1060		1062		1060		Итого:
	Автопод...	Итого:	Автопод...	Итого:	Автопод...	Итого:	
> 01.01.2015	1 375 543,70	1 375 583,70	1 420 136,56	1 420 136,56	1 138 638,08	1 138 628,00	3 934 546,24
> 01.02.2015	723 666,22	723 666,22	1 222 003,72	1 222 003,72	601 692,48	601 692,48	2 637 362,42
> 01.03.2015	586 160,82	586 160,82	880 840,06	880 840,06	481 812,80	481 812,80	1 949 812,80
	187 101,20	187 101,20	261 903,56	261 903,56	341 262,12	341 262,12	590 256,58
	793 563,82	793 563,82	1 342 733,32	1 342 733,32	623 074,72	623 074,72	2 519 069,86
> 01.04.2015	1 081 882,30	1 081 882,30	1 136 869,42	1 136 869,42	603 833,62	603 833,62	2 899 396,34
> 01.05.2015	540 386,54	540 386,54	976 951,12	976 951,12	380 225,38	380 225,38	1 897 563,04
> 01.06.2015	441 376,82	441 376,82	779 217,88	779 217,88	323 117,96	323 117,96	1 543 712,66
> 01.07.2015	476 948,34	476 948,34	783 906,38	783 906,38	404 172,84	404 172,84	1 314 629,76
> 01.08.2015	746 696,12	746 696,12	991 853,20	991 853,20	374 027,72	374 027,72	1 711 282,04
> 01.09.2015	2 377 043,96	2 377 043,96	1 791 499,24	1 791 499,24	1 157 488,84	1 157 488,84	5 328 032,04
> 01.10.2015	2 941 790,50	2 941 790,50	2 244 300,88	2 244 300,88	1 475 305,30	1 475 305,30	6 261 196,68
> 01.11.2015	2 728 953,74	2 728 953,74	1 881 176,52	1 881 176,52	1 546 510,58	1 546 510,58	6 156 640,84
> 01.12.2015	2 987 763,54	2 987 763,54	1 990 447,12	1 990 447,12	1 723 437,30	1 723 437,30	6 793 647,96
> 01.01.2016	3 736 963,50	3 736 963,50	3 049 721,42	3 049 721,42	2 455 656,22	2 455 656,22	9 242 341,14
> 01.02.2016	2 448 523,38	2 448 523,38	2 014 760,34	2 014 760,34	1 587 772,88	1 587 772,88	6 051 076,60
> 01.03.2016	2 882 571,08	2 882 571,08	1 983 682,92	1 983 682,92	907 838,46	907 838,46	5 774 092,46
> 01.04.2016	2 680 749,08	2 680 749,08	2 176 027,68	2 176 027,68	1 030 788,90	1 030 788,90	5 887 565,66
> 01.05.2016	1 186 693,80	1 186 693,80	1 165 606,34	1 165 606,34	267 365,34	267 365,34	2 619 665,48
> 01.06.2016	700 565,32	700 565,32	705 862,20	705 862,20			1 406 227,52
> 01.07.2016	855 997,10	855 997,10	653 990,58	653 990,58			1 709 988,08



Технології аналізу даних

Data Analysis Technologies

**Кафедра цифрової економіки
та системного аналізу**

**Роскладка Андрій Анатолійович,
д.е.н., професор
a.roskladka@knuce.edu.ua**

