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

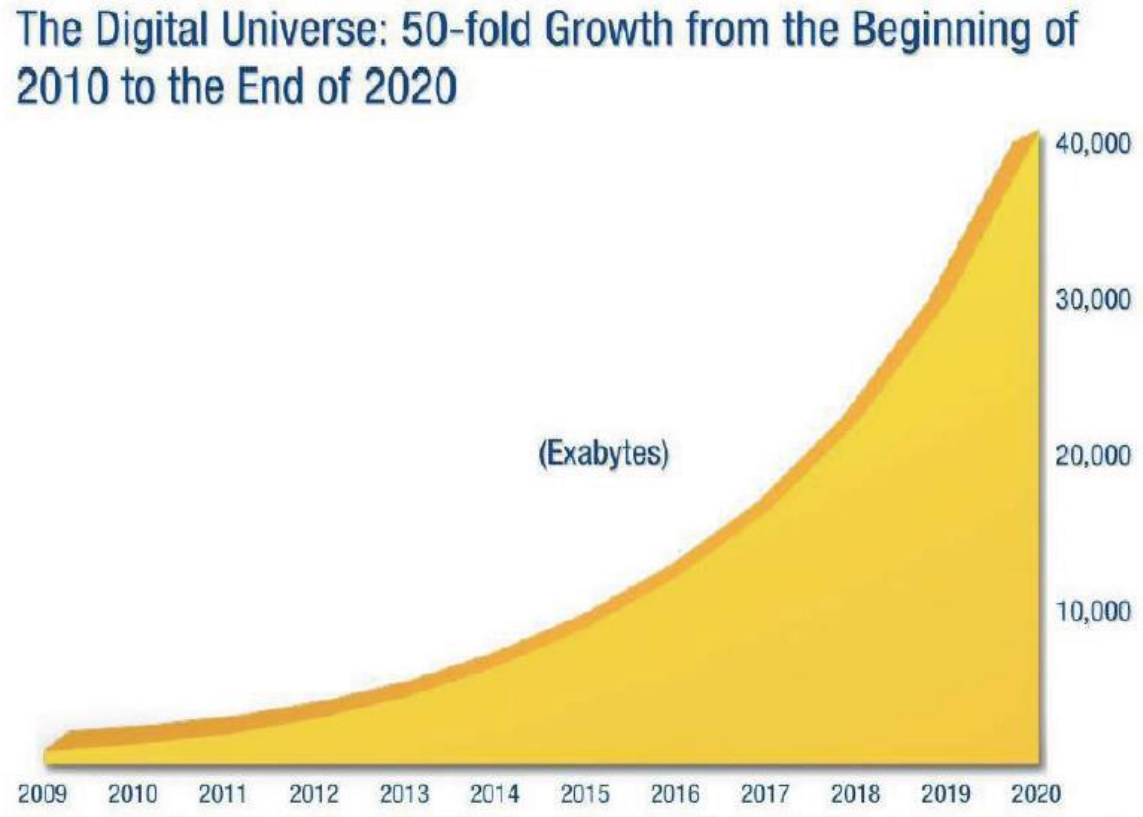
Тема 6.
Інструментальні засоби аналізу даних

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



За інформацією IBM 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 слабко 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)

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

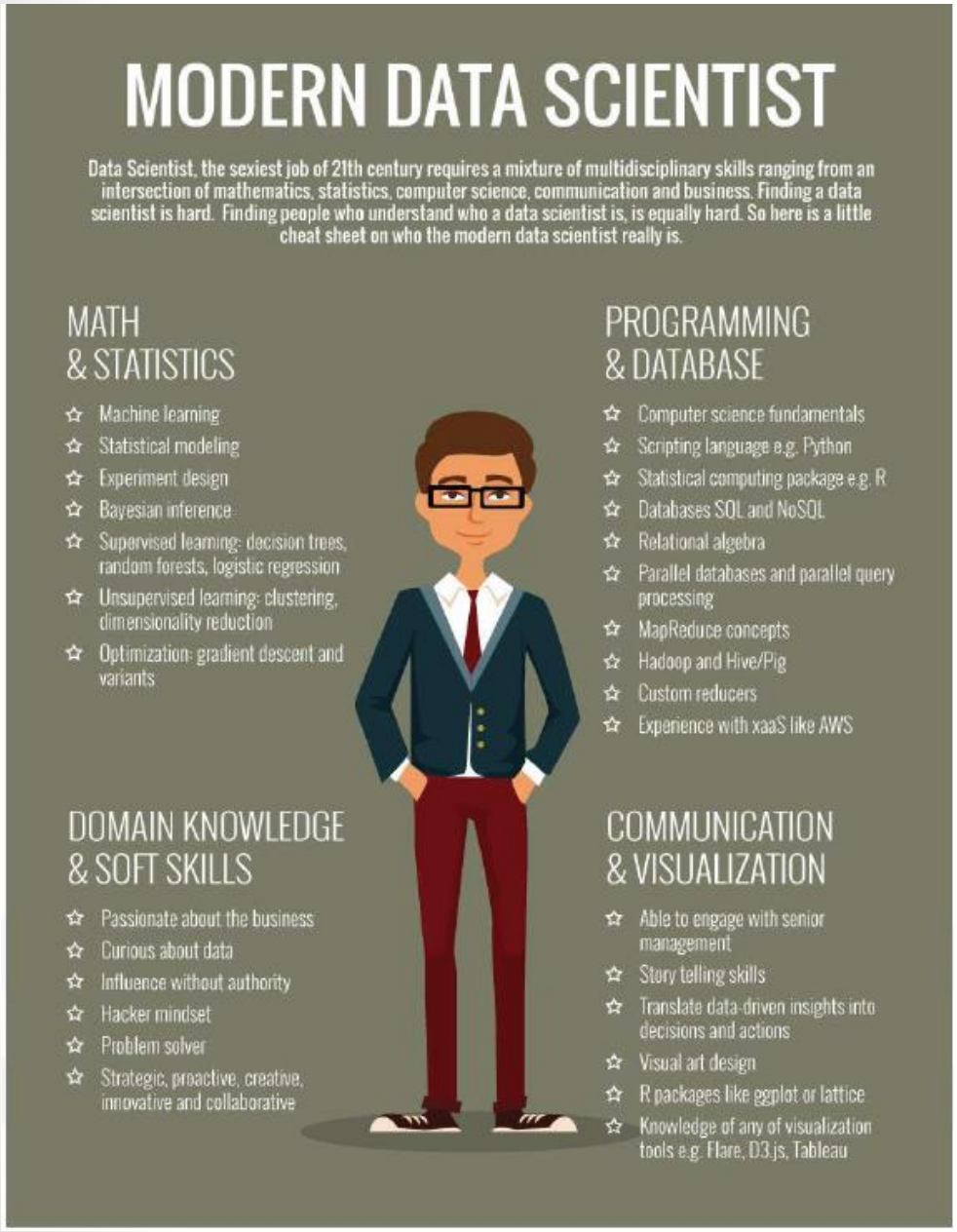
Data Scientist:

The Sexiest Job of the 21st Century

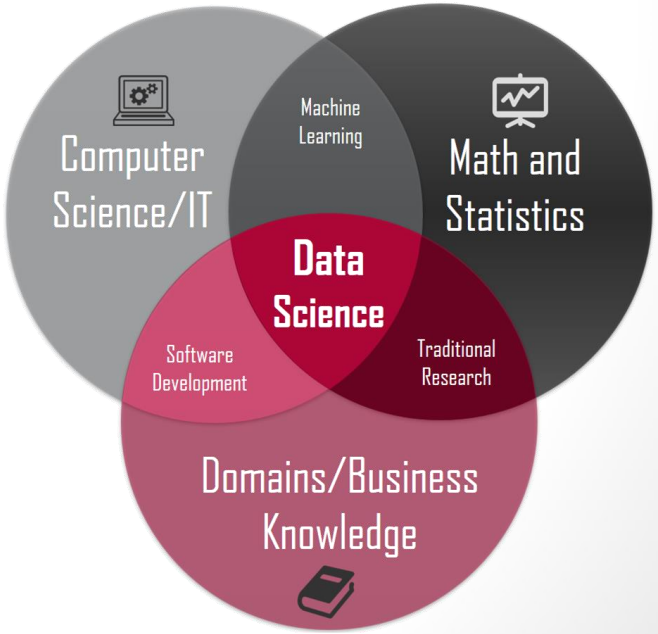
Meet the people who can coax treasure out of messy, unstructured data.

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."

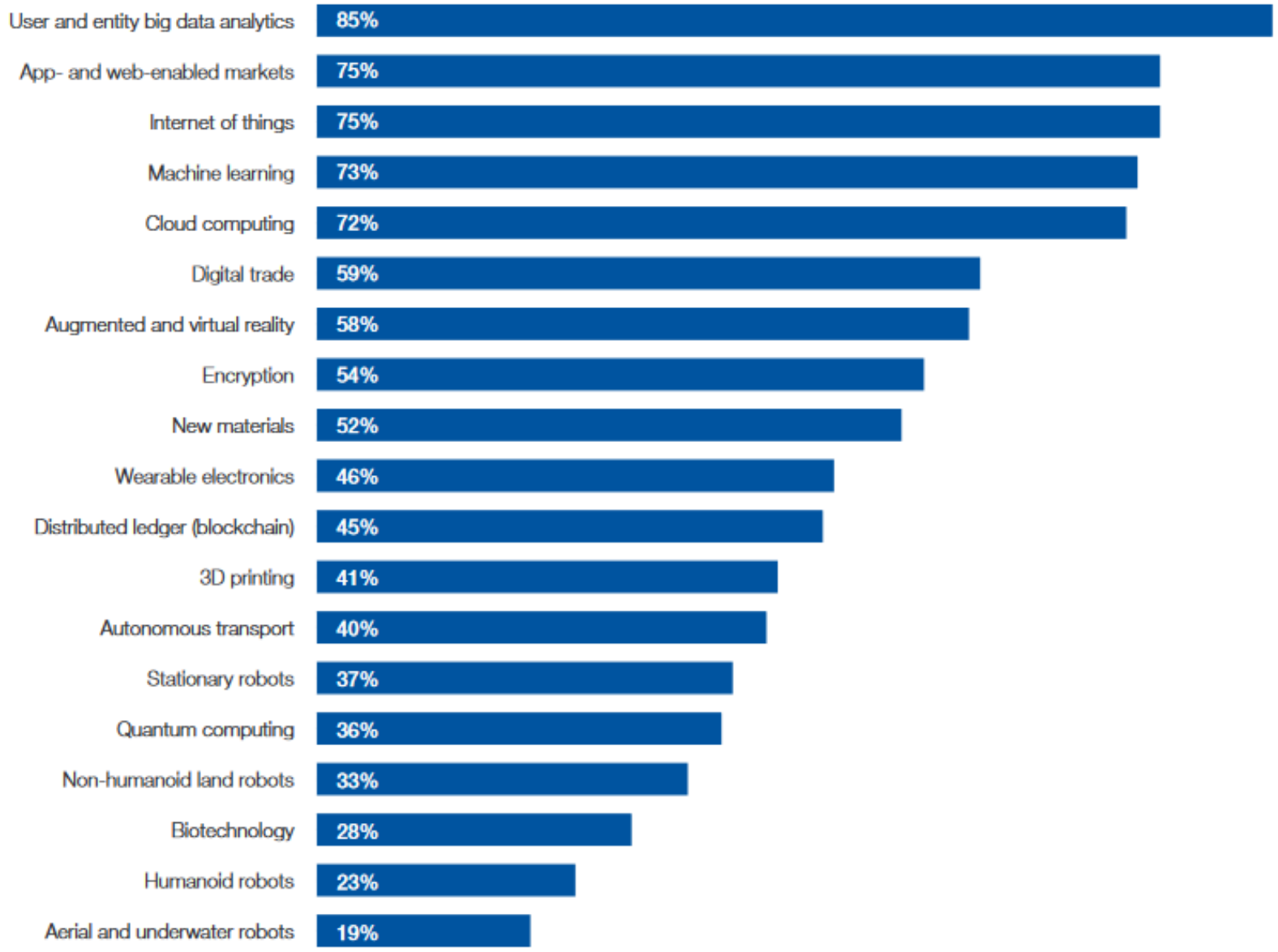


Data Scientist – це не програміст. Це фахівець з чудовими кросдисциплінарними знаннями математики, штучного інтелекту, інформаційних технологій та бізнесу і суперздатностями до аналізу.



World Economic Forum: перспективи Data Science

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



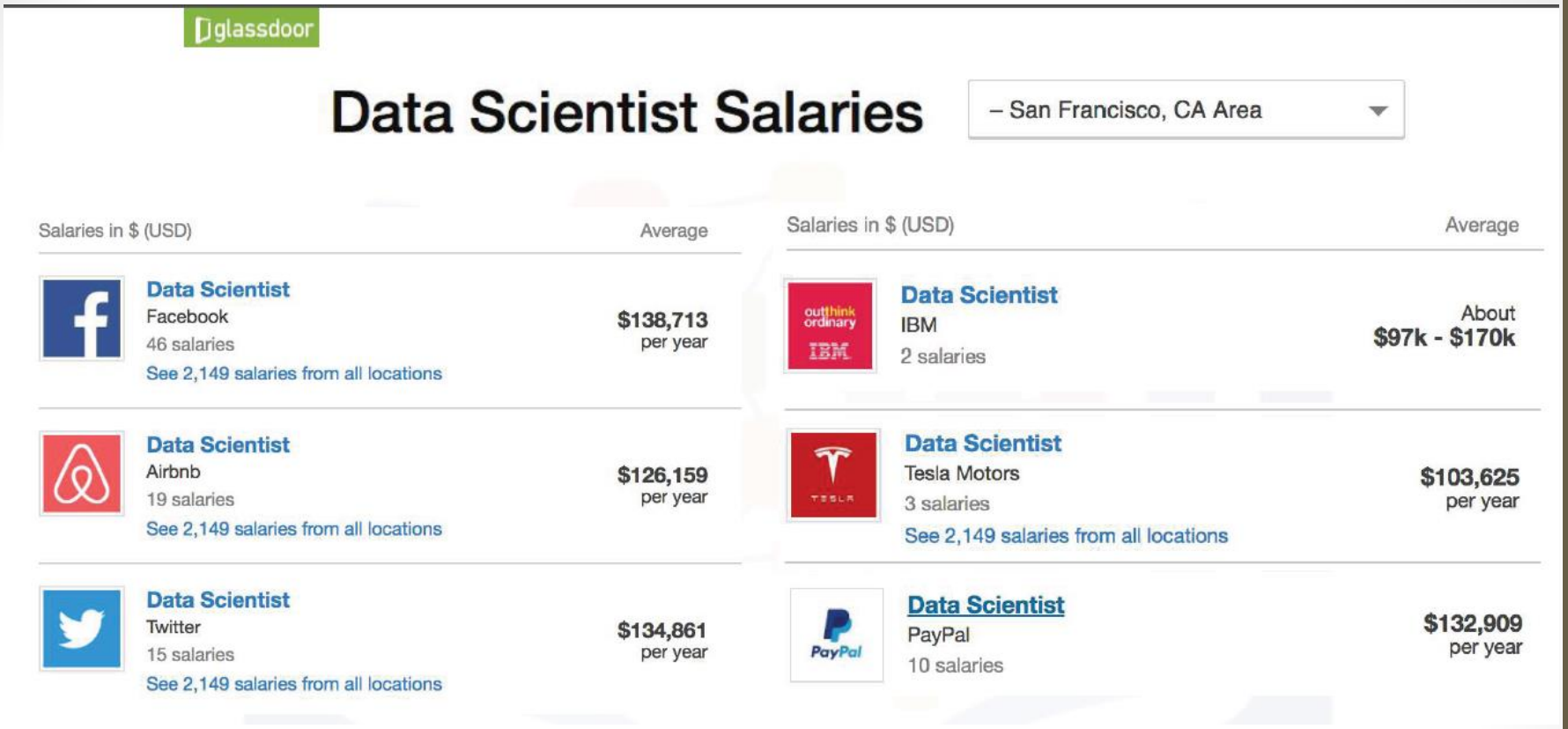
Source: Future of Jobs Survey 2018, World Economic Forum.

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

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

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.



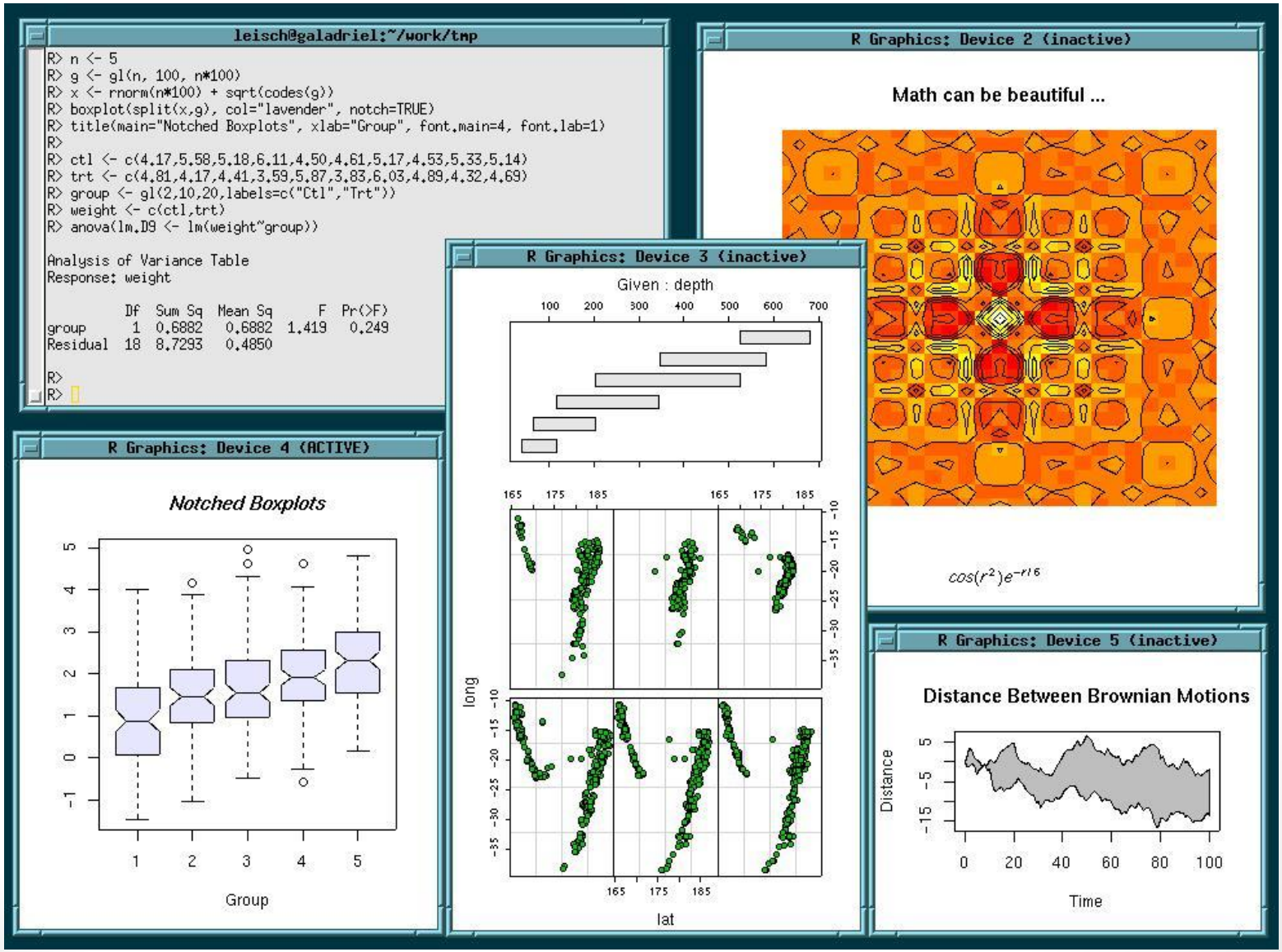
Вакансії та заробітна плата фахівців з аналізу даних від філій лідерів світових компаній у м. Сан-Франциско (США)

Microsoft Power BI

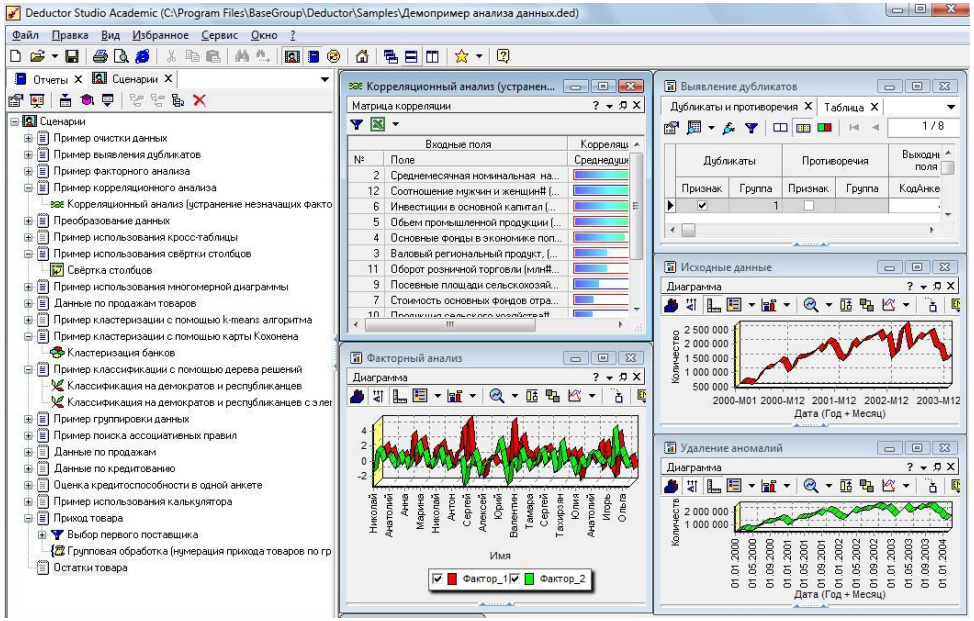
Power BI

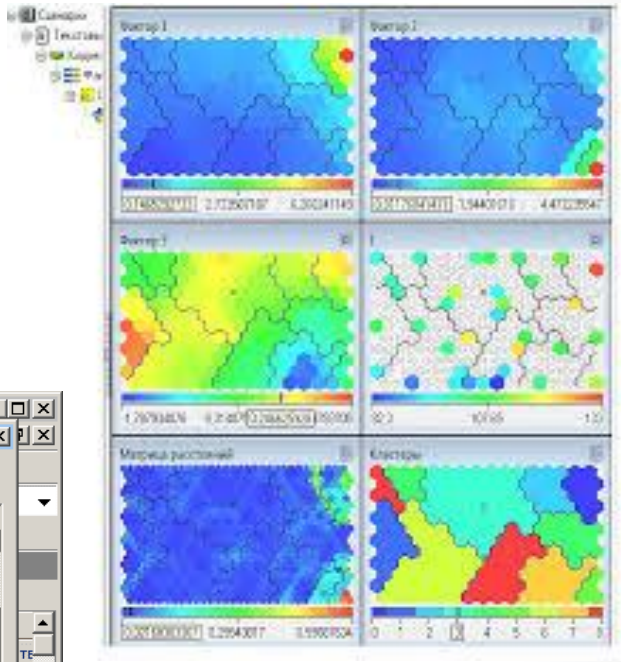
Microsoft

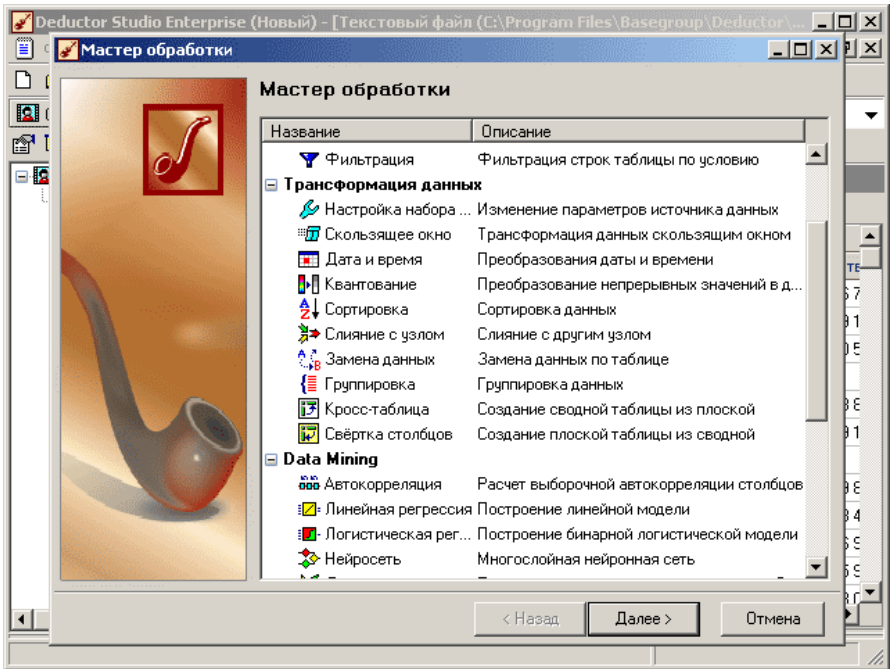
WEKA


leisch@galadriel:~/work/tmp
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 ...
$cos(r^2)e^{-r/6}$
R Graphics: Device 3 (inactive)
Given : depth
long
lat
R Graphics: Device 4 (ACTIVE)
Notched Boxplots
Group
R Graphics: Device 5 (inactive)
Distance Between Brownian Motions
Distance
Time

Deductor Studio

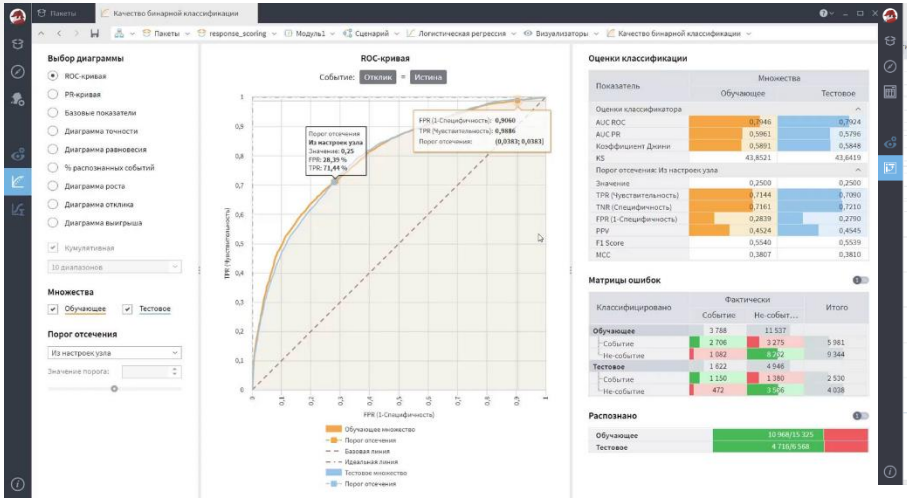


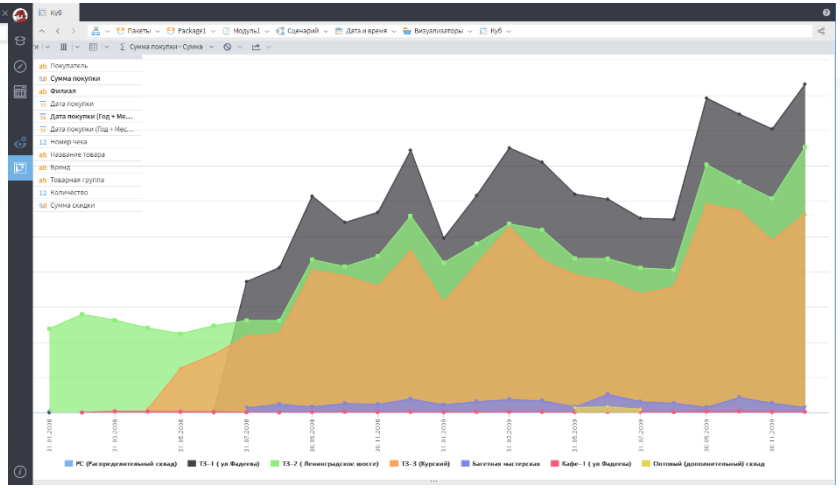


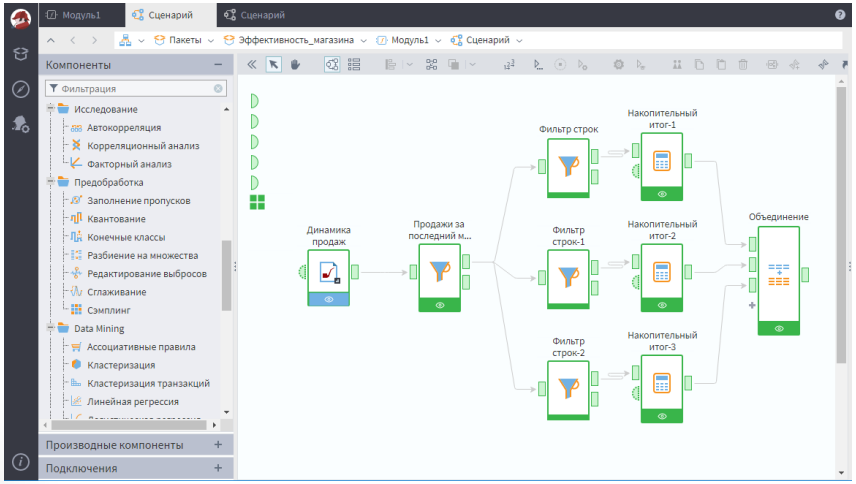


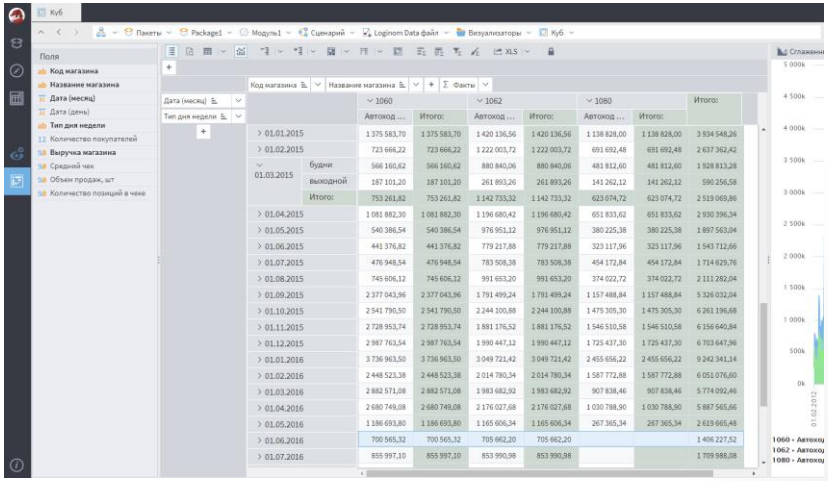
Loginom

Loginom









Технології аналізу даних

Data Analysis Technologies

Кафедра цифрової економіки
та системного аналізу

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