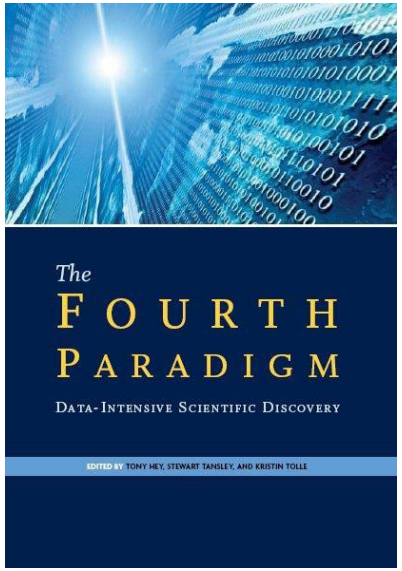
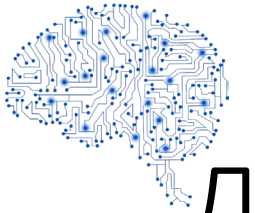
Лекція 1. Сучасна парадигма наукових досліджень

Професор, д.е.н. Ставицький А.В.

Четверта парадигма наукових досліджень

- Теорія, гіпотеза та логічні міркування, спостереження або експеримент, наприклад:
  - Ньютон спостерігав, як падають яблука, щоб створити його теорію механіки
  - Галілео Галілей робив експерименти з падаючими предметами з пізанської вежі
- Моделювання теорії або моделі
  - Цифрове моделювання може довести теорію або модель
- Наукове відкриття на основі даних (відоме також як Data Science)
  - Більше даних перевершує гіпотетичну теорію
  - Е-наука (обчислювальна техніка та інформаційні технології) розширили науку
- Комп’ютерно-людська наука?
  - Машина відкриває нові закономірності та формулює гіпотези в одному або декількох просторах знань
  - Вчений перевіряє та розробляє додаткові тести чи експерименти





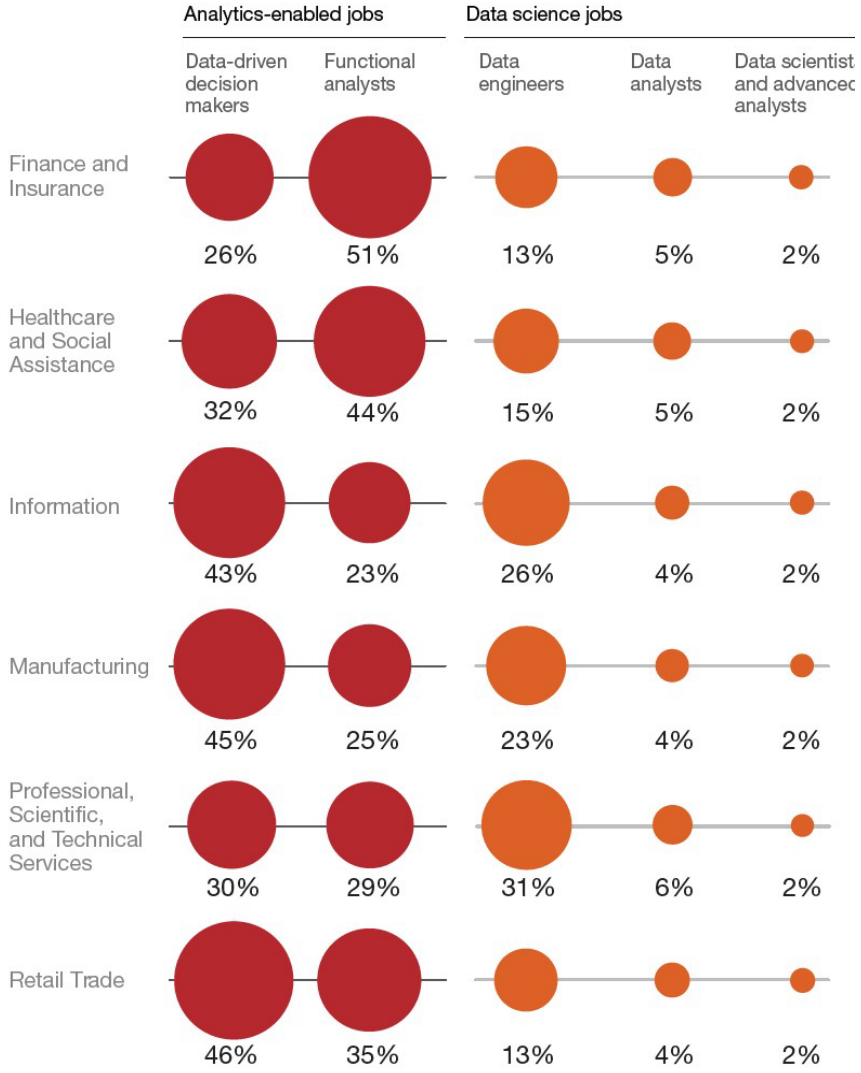
Дані як нова парадигма

- У XXI столітті спостерігається експоненційне зростання обсягів даних ("Великі дані").
- Класичний науковий метод доповнюється аналізом величезних масивів інформації.
- Нові методи, такі як машинне навчання, дозволяють відкривати закономірності без формулювання чітких гіпотез на початку дослідження.

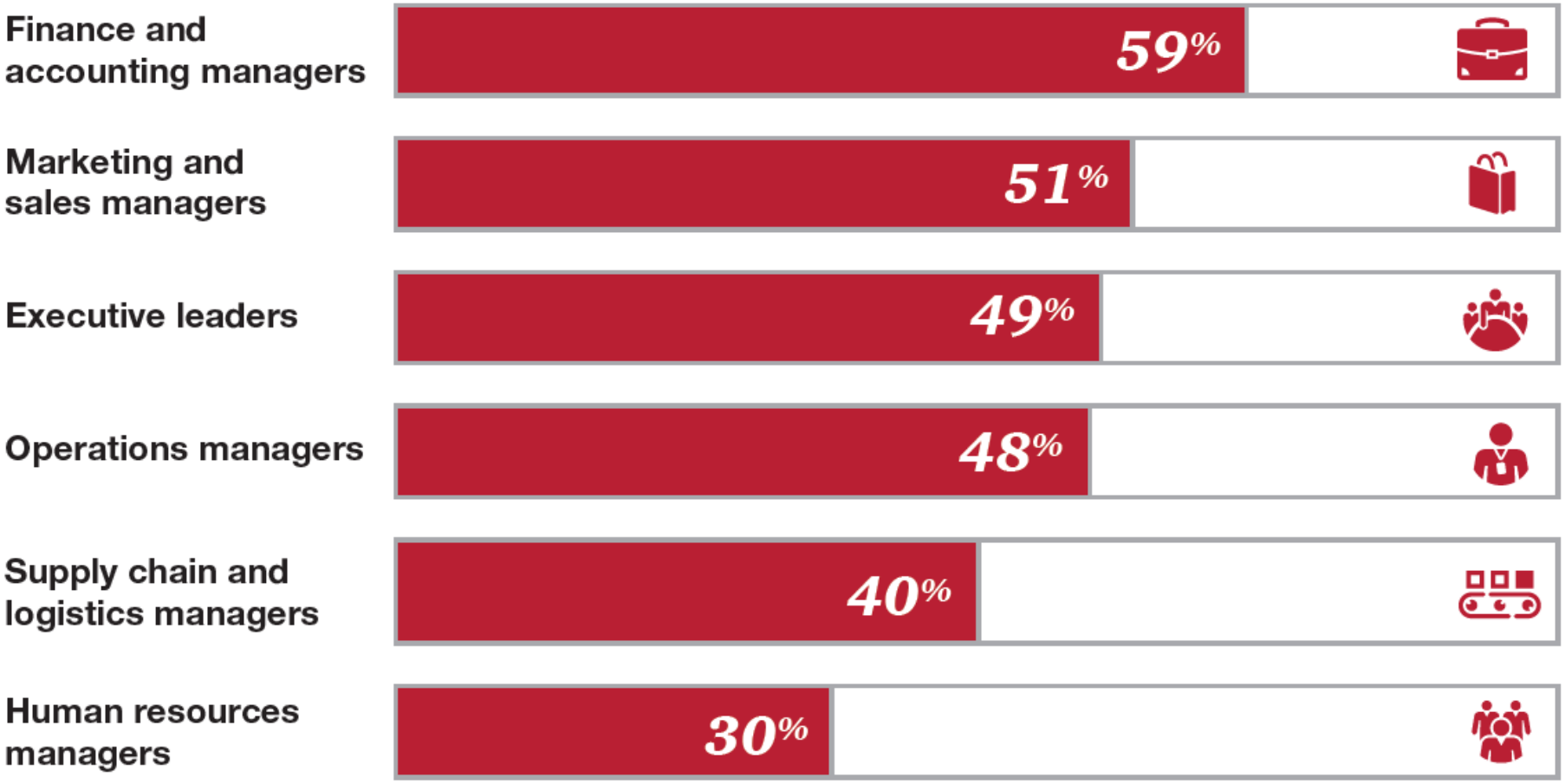
Попит на робочі місця з підтримкою DSA

- Попит на ділових людей з навичками аналітики, не тільки науковців
- З 2,35 мільйона оголошень про роботу в США
  - 23% науковий співробітник
  - 67% завдань із підтримкою DSA

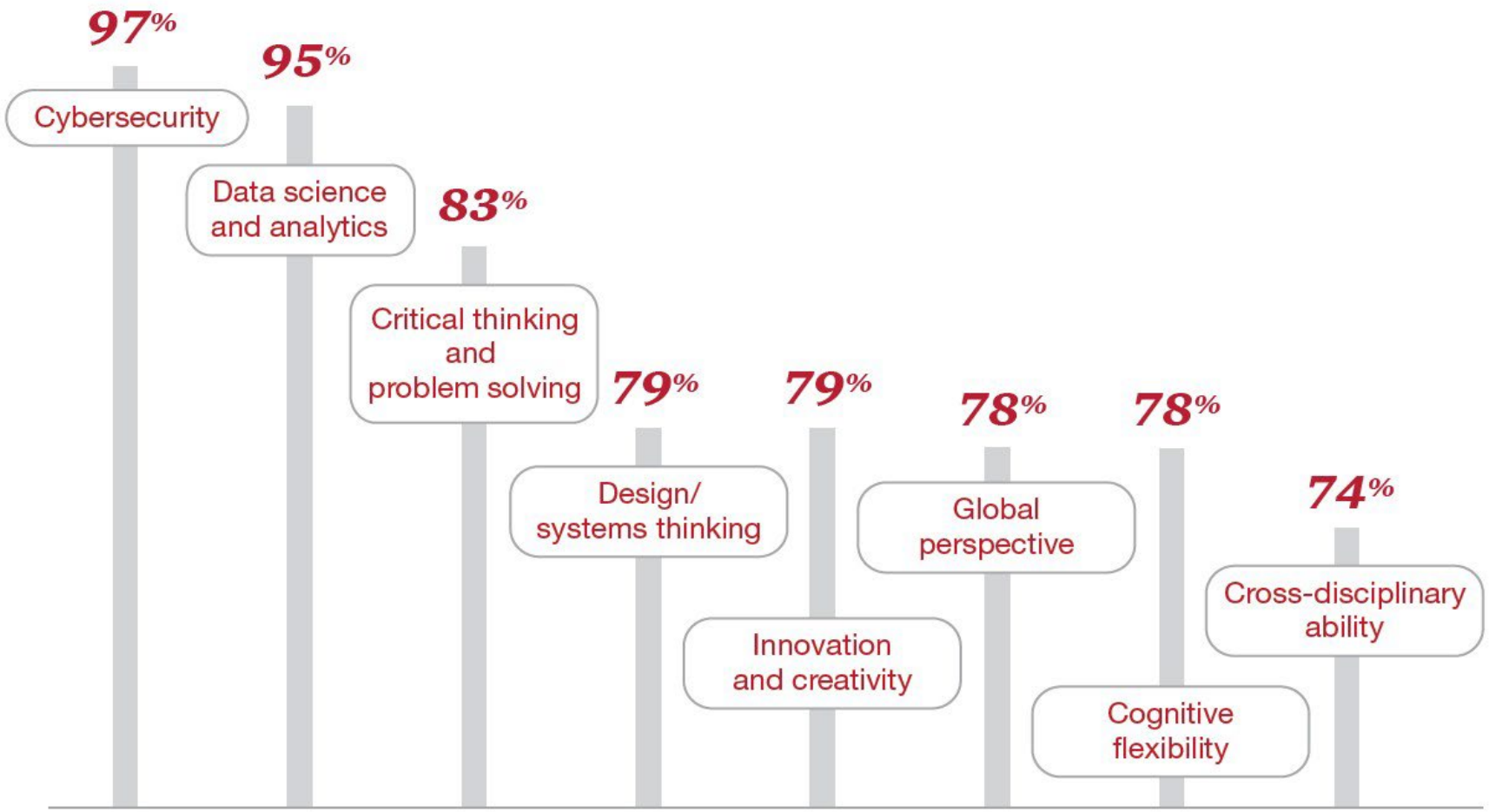
- Великий попит на менеджерів та осіб, які приймають рішення, що володіють навичками / розумінням даних науки (аналітика даних)
  - Завдання забезпечити діючі знання та компетентності менеджерам рівня генерального директора



Відсоток роботодавців, які заявляють про потрібні навички щобо обробки даних та аналітики (від усіх менеджерів у 2020 році)



Навички, які важко знайти



Source: Business Roundtable (2017).

Зарплати аналітиків (2016-2017)

DSA Framework Category	Top Industries (by Demand Volume)	Average Time to Fill (Days)	Average Annual Salary
Data-Driven Decision Makers	Professional Services	50	$96,845
	Finance & Insurance	37	$98,131
	Manufacturing	43	$93,641
Functional Analysts	Finance & Insurance	35	$71,937
	Professional Services	48	$69,135
	Manufacturing	39	$72,571
Data Systems Developers	Professional Services	51	$82,447
	Finance & Insurance	35	$87,039
	Manufacturing	43	$81,138
Data Analysts	Professional Services	47	$74,917
	Finance & Insurance	31	$83,209
	Manufacturing	41	$72,742
Data Scientists & Advanced Analysts	Professional Services	51	$97,457
	Finance & Insurance	43	$106,610
	Manufacturing	45	$92,543
Analytics Managers	Finance & Insurance	38	$113,754
	Professional Services	53	$107,185
	Manufacturing	40	$106,926

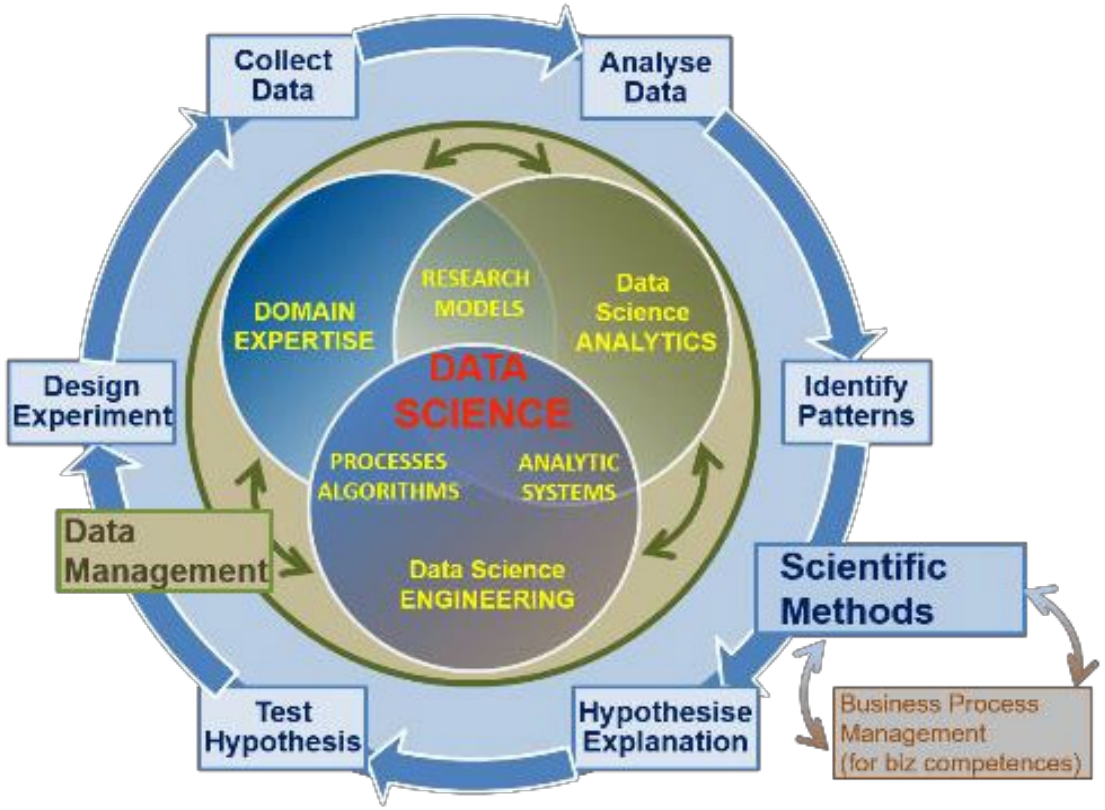
Зарплати 2022

- Директор з розробки програмного забезпечення та надійності сайту (Каліфорнія): $389 000
- Технічний менеджер: $182 000 - 260 000
- Інженер з питань конфіденційності (Каліфорнія): $120 000–184 000
- Вчений-дослідник: $132 000 - 268 000
- Фахівець за даними: $105 000–200 000
- Інженер-проектувальник: $93 000-153 000
- Аналітик даних: 84 000-162 000
- Інженер з апаратного забезпечення (Каліфорнія): $ 110 000 - 243 000
- Менеджер з продукту: $114 650 - 217 416
- Глобальний продакт-лід: $124 000–200 000
- Менеджер з продукції групи (Каліфорнія): $250 000
- Менеджер з маркетингу товару: $118 000–234 000
- Менеджер проекту (Каліфорнія): $153 000–204 000
- Аналітик продукту (Каліфорнія): $85 571–163 000
- Бізнес-аналітик: $93 000-169 000
- Операційний аналітик: $114 000-137 000
- Фінансовий аналітик: $100 000-175 000
- Фінансовий менеджер: $184 350 - 187 000
- Бухгалтер: $106000 - 146 000

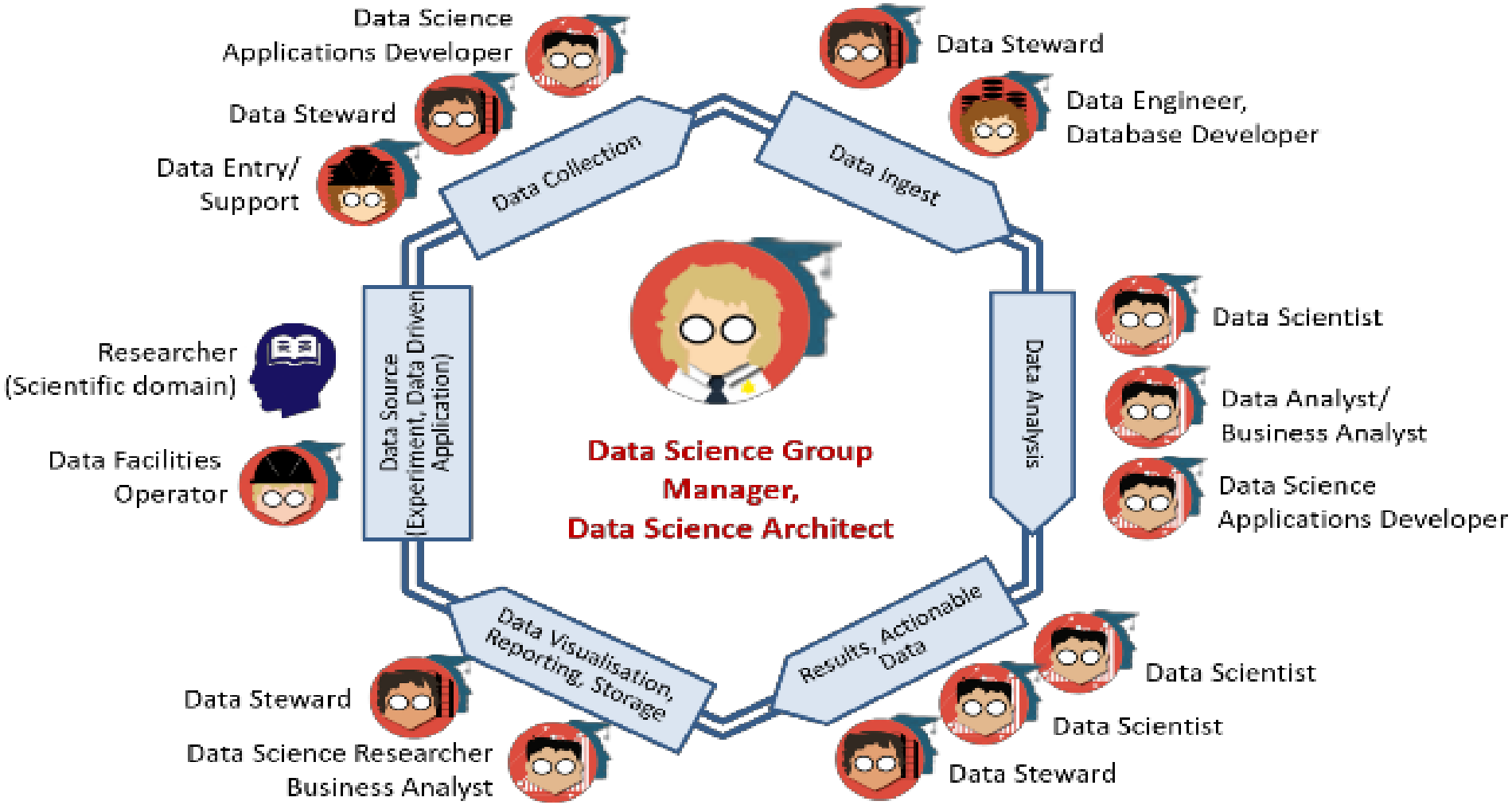
Індустрія 4.0 та попит на нові навички

- Четверта промислова революція, що включає розробки в раніше нерозвинутих областях, таких як штучний інтелект та машинне навчання, робототехніка, нанотехнології, тривимірний друк, генетика та біотехнології, спричинить широкі зміни не тільки бізнес-моделей, але й ринків праці у наступні п’ять років
- Про це свідчить новий звіт "Майбутнє робочих місць", опублікований Світовим економічним форумом.

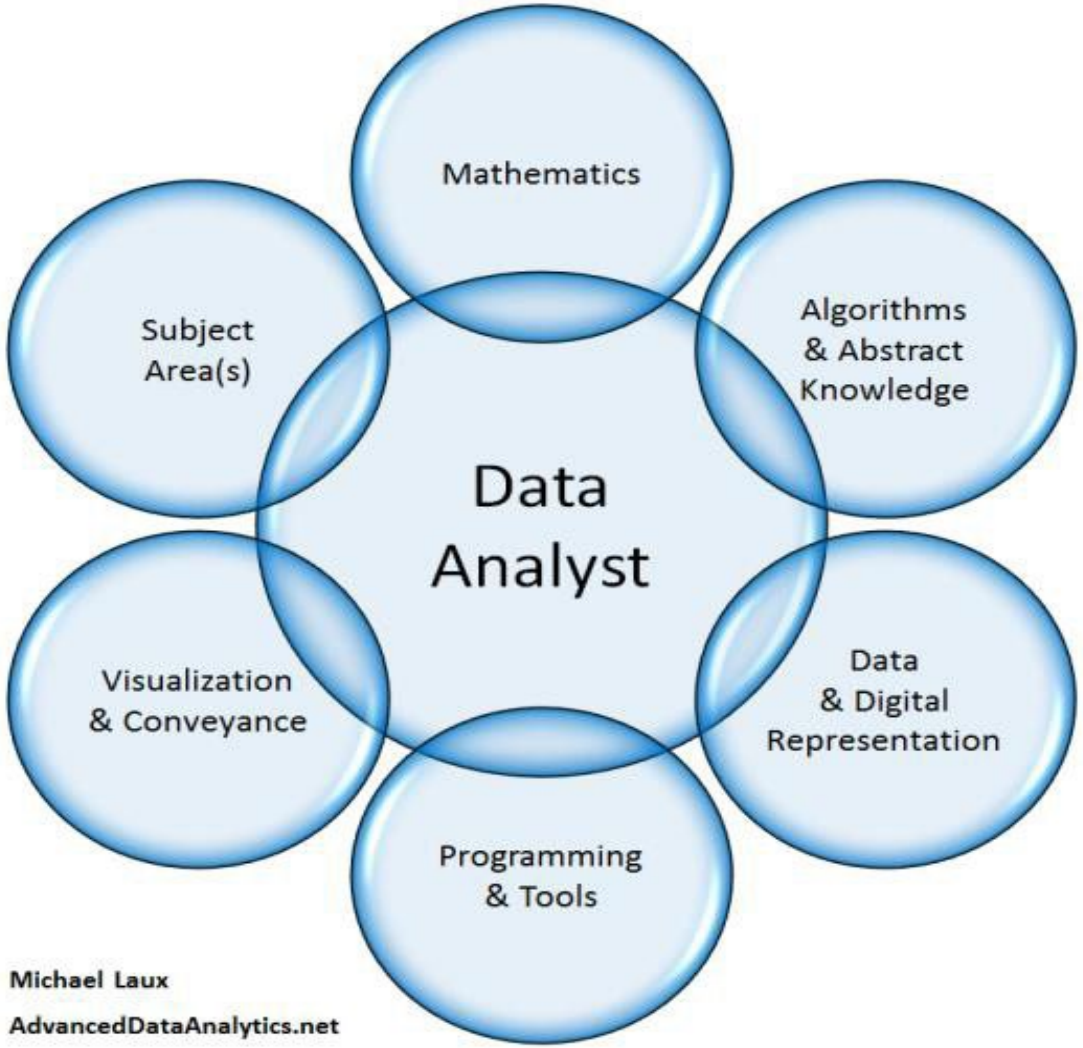
Процес аналізу даних



Професії


Data Science Applications Developer
Data Steward
Data Entry/ Support
Data Collection
Data Steward
Data Engineer, Database Developer
Data Ingest
Data Scientist
Data Analyst/ Business Analyst
Data Science Applications Developer
Data Analysis
Data Science Group Manager, Data Science Architect
Researcher (Scientific domain)
Data Facilities Operator
Data Source (Experiment, Data Driven Application)
Data Steward
Data Science Researcher Business Analyst
Data Visualisation, Reporting, Storage
Results, Actionable Data
Data Scientist
Data Scientist
Data Steward

Аналітик


Mathematics
Subject Area(s)
Algorithms & Abstract Knowledge
Data Analyst
Visualization & Conveyance
Data & Digital Representation
Programming & Tools
Michael Laux
AdvancedDataAnalytics.net

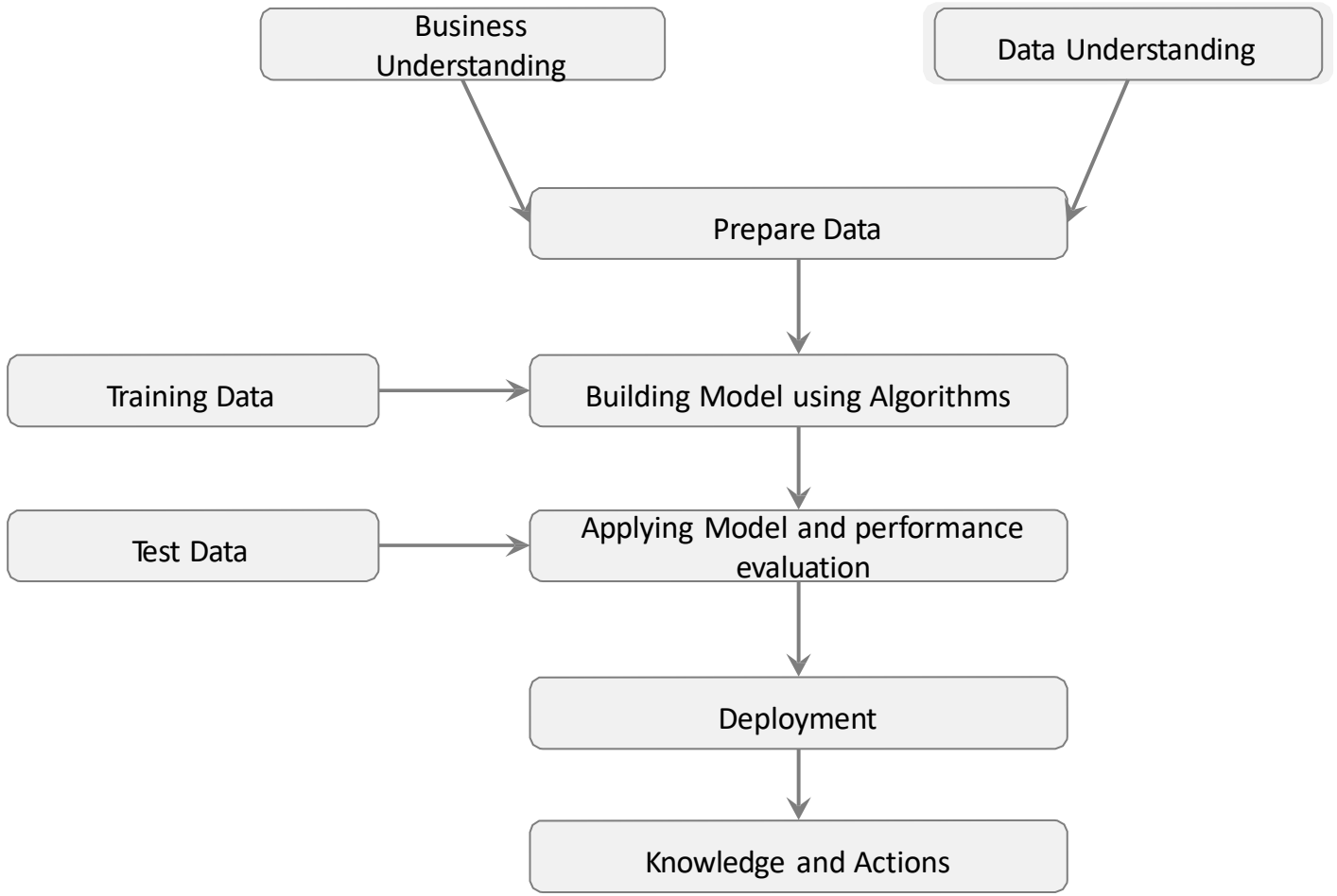
Data Scientist

- Data Scientist - це практик, який має достатні знання бізнес-потреб, знання доменів, аналітичні навички, а також знання програмування та системної інженерії для управління наскрізним процесом наукових методів на кожному етапі життєвого циклу великих даних до надання очікуваної наукової та ділової цінності організації чи проекту.

Професійні навички в галузі науки про дані

- Визнавайте цінність даних, працювати з необробленими даними, показувати хорошу інтуїцію в даних, використовувати відкриті дані
- Будьте готові до ітеративного розвитку
- Розуміючи метрики, розуміючи важливість перевірки результатів, ніколи не переставайте дивитись на окремі приклади
- Ставте правильні запитання
- Поважайте знання предмету у галузі науки про дані
- Вирішуйте проблеми на основі даних, застосовуйте розумовий підхід, заснований на даних
- Будьте в курсі потужності та обмежень основного машинного навчання та інструментів аналітики
- Зрозумійте, що більшість алгоритмів аналізу даних ґрунтуються на статистиці та ймовірності, тому будь-яка відповідь чи рішення має певний ступінь вірогідності та представляє оптимальне рішення для ряду змінних та факторів

Процес аналізу даних



1. Prior Knowledge

2. Preparation

3. Modeling

4. Application

5. Knowledge

Що таке цифрова трансформація сьогодні (O`Reilly)

- Data-Resistant–Супротив даним - Ми так завжди робили - Прагнення рухатися «накатаним» шляхом
- Data-Curious–Інтерес до даних - Що говорять дані? - Усвідомлення потенційної цінності зібраних даних, хоча сама цінність ще не ясна
- Data-Aware–Поінформованість про дані - Що відбувається? - Ухвалення тактичних рішень на основі даних. Використання різних інструментів аналітики, щоб вибрати найбільш підходящі для себе.
- Data-Savvy–Розуміння даних - Чому це відбувається? - Розуміння того, що цінність даних не тактична, дані —це стратегічний актив. Прагнення отримати інсайт -глибинний сенс, який несуть у собі дані.
- Data-Driven–Управлінняна основі даних - Що робити далі? - Розуміння, що сталося, чому це сталося і що робити далі. Прийняття рішення зводиться до формулювання чистих і коротких тверджень, що ґрунтуються на даних і вказують на дію, яку потрібно зробити

- DataDrivenCompany – це підприємство кероване даними, де бізнес-процеси та оргструктура побудована на основі наскрізної інтеграції інформаційних потоків та їх безперервної предиктивної (прогнозної) аналітики. У такому підприємстві впроваджено та широко застосовуються інструменти Data Science, Machine Learning, BigData.

Інформаційні системи

- Інформаційні системи і технології дають можливість оптимізувати і раціоналізувати управлінські функції за рахунок застосування сучасних засобів отримання, опрацювання та передавання інформації.

Основні компоненти ІС

- функціональні компоненти;
- компоненти системи опрацювання даних;
- організаційні компоненти.

Інформаційні системи та великі дані

- Інформаційні системи (ІС) традиційно використовувалися для збереження та обробки структурованих даних.
- З появою великих даних (Big Data) вони стали основою для аналітики, прогнозування та автоматизованого ухвалення рішень.
- Основні виклики: обробка неструктурованих даних, швидкість обробки, інтеграція різних джерел інформації.

Роль інформаційних систем у Big Data

- Збір даних
  - сенсори
  - мобільні пристрої
  - соціальні мережі
  - IoT
- Зберігання
  - розподілені бази даних (Hadoop, NoSQL, Cloud Storage).
- Аналіз
  - машинне навчання
  - нейромережі
  - обчислювальні моделі.
- Візуалізація
  - BI-системи
  - інтерактивні дашборди.

Проєктування ІС

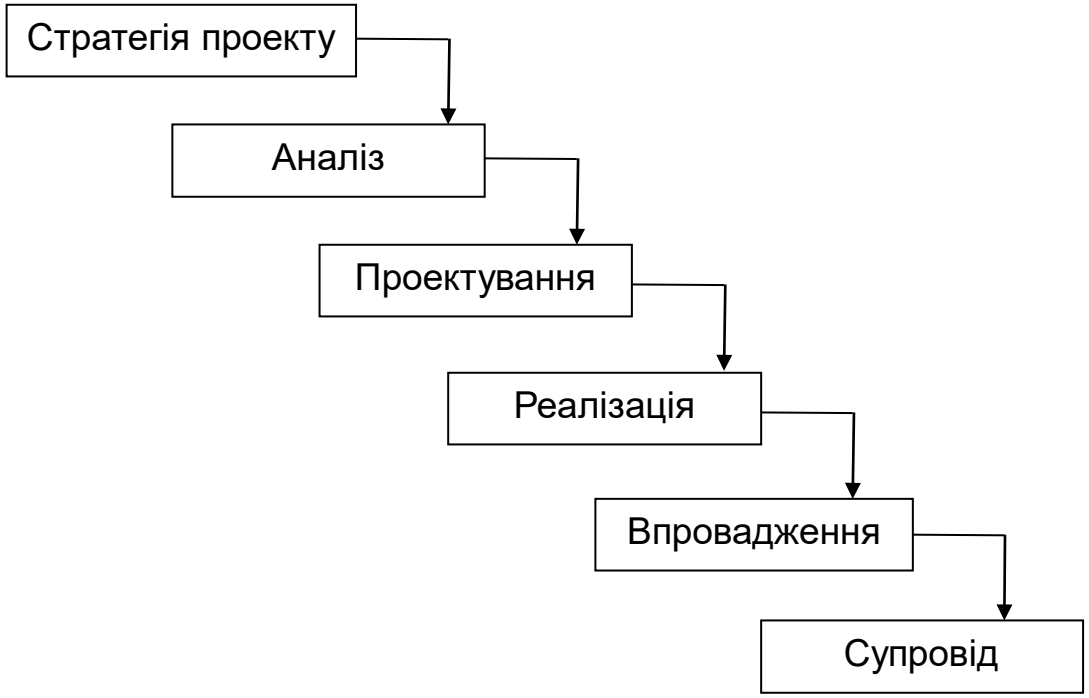
- Життєвий цикл ПЗ – це безперервний процес, який починається з моменту ухвалення рішення про необхідність його створення і закінчується у момент його повного вилучення з експлуатації.
- Структура ЖЦ ПЗ базується на трьох групах процесів:
  - основні процеси ЖЦ ПЗ (придбання, постачання, розробка, експлуатація, супровід);
  - допоміжні процеси, що забезпечують виконання основних процесів (документування, управління конфігурацією, забезпечення якості, верифікація, атестація, оцінка, аудит, рішення проблем);
  - організаційні процеси (управління проектами, створення інфраструктури проекту, визначення, оцінка і поліпшення самого ЖЦ, навчання).

Основні моделі ЖЦ

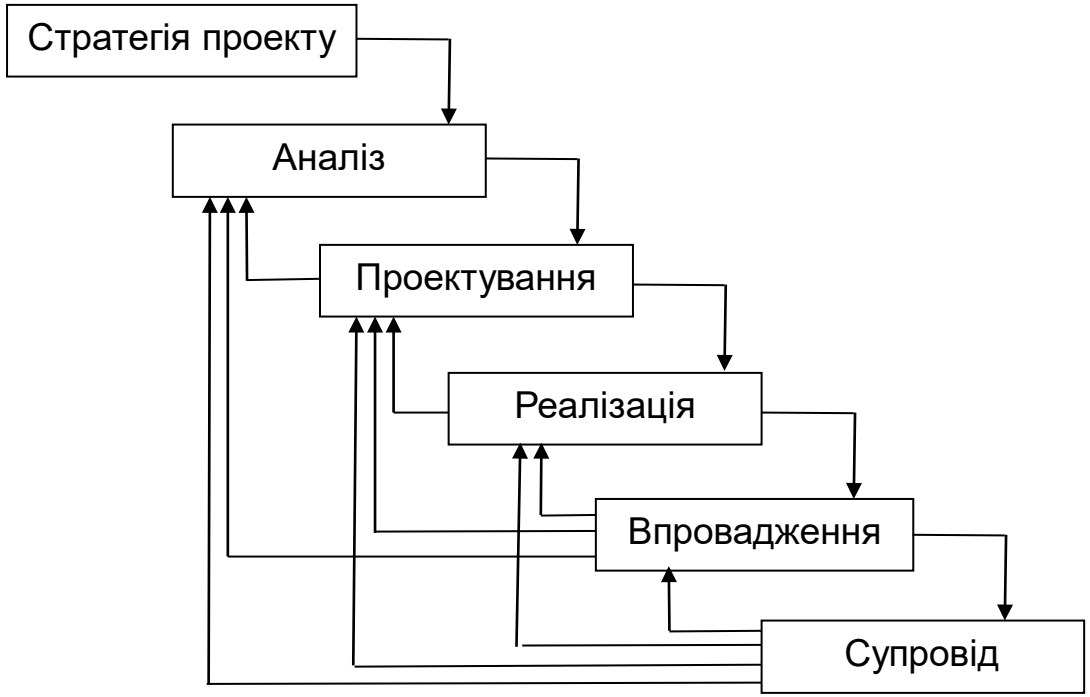
- каскадна модель;
- каскадна схема розробки ІС з проміжним контролем
- спіральна модель.

- Вибір моделі життєвого циклу проектування ІС є багатокритеріальною задачею, якість вирішення якої визначається досвідом та інтуїцією керівників проекту.

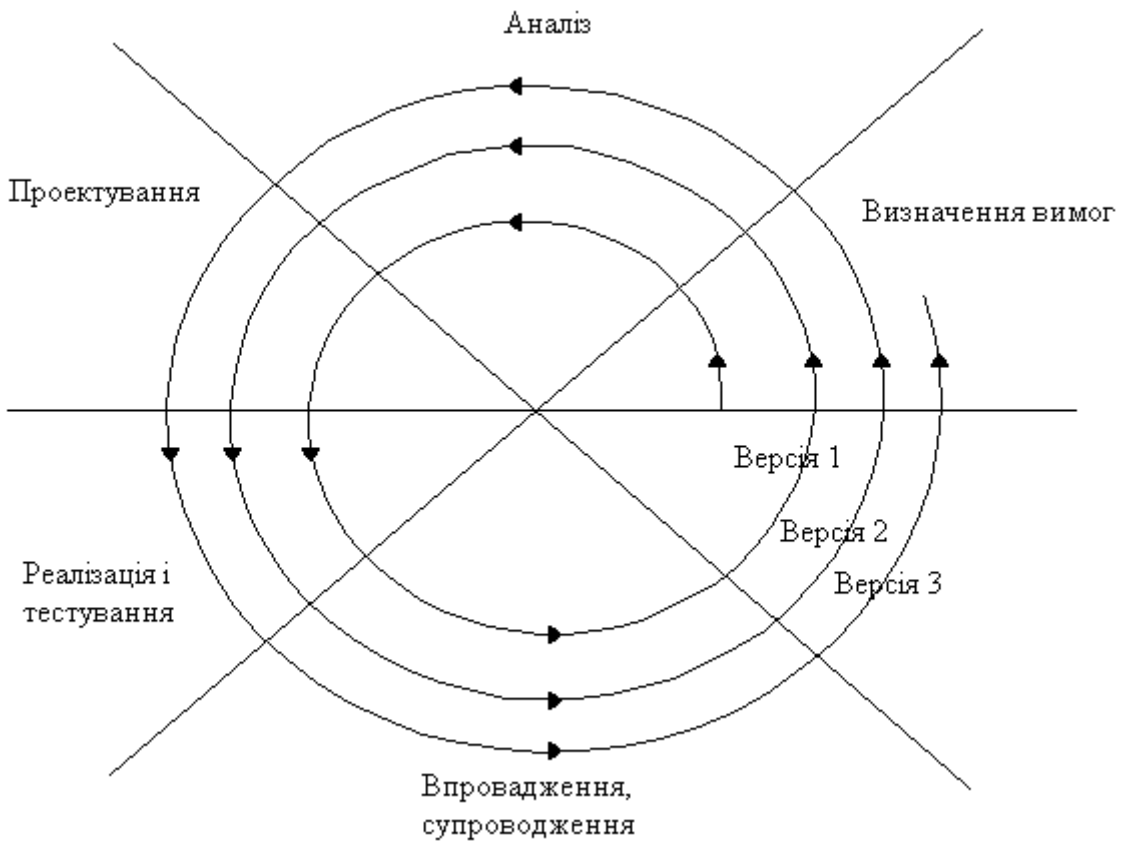
Каскадна схема розробки ІС



Каскадна схема розробки ІС з проміжним контролем



Спіральна модель життєвого циклу ІС



Стадії розробки ІС

1) формування вимог до інформаційної системи (ІС);
2) розробка концепції ІС;
3) технічне завдання;
4) ескізний проєкт;
5) технічний проєкт;
6) робоча документація;
7) введення в експлуатацію;
8) супроводження ІС.

Склад і формування вимог до проєктованої ІС

- Планування робіт включає визначення мети розробки, попередню економічну оцінку проекту, створення плану-графіка виконання робіт, навчання спільної робочої групи;
- Проведення обстеження діяльності об’єкта (організації) автоматизації, у рамках якого здійснюються: попереднє виявлення вимог до майбутньої системи; визначення структури організації; визначення переліку цілей організації; аналіз розподілу функцій за підрозділами і між співробітниками; виявлення функціональних взаємодій між підрозділами, інформаційних потоків усередині підрозділів і між ними, зовнішніх стосовно організації об’єктів і зовнішніх інформаційних взаємодій; аналіз наявних засобів автоматизації діяльності організації;
- Побудову моделей діяльності організації, що передбачає обробку матеріалів обстеження;
- Побудову двох видів моделей:
  - моделі "як є", що відображає наявний на момент обстеження стан справ і допомагає зрозуміти, як саме функціонує певне підприємство, а також виявити вузькі місця і сформулювати пропозиції щодо поліпшення ситуації;
  - моделі "як має бути", що відображає схему про нові технології роботи підприємства. Кожна з моделей містить повну функціональну й інформаційну модель діяльності організації.

Мета створення інформаційних систем

- у гранично короткі терміни створити систему обробки даних, яка має задані споживчі якості. До них належать:
  - функціональна повнота,
  - своєчасність,
  - функціональна надійність,
  - адаптивна надійність,
  - економічна ефективність.

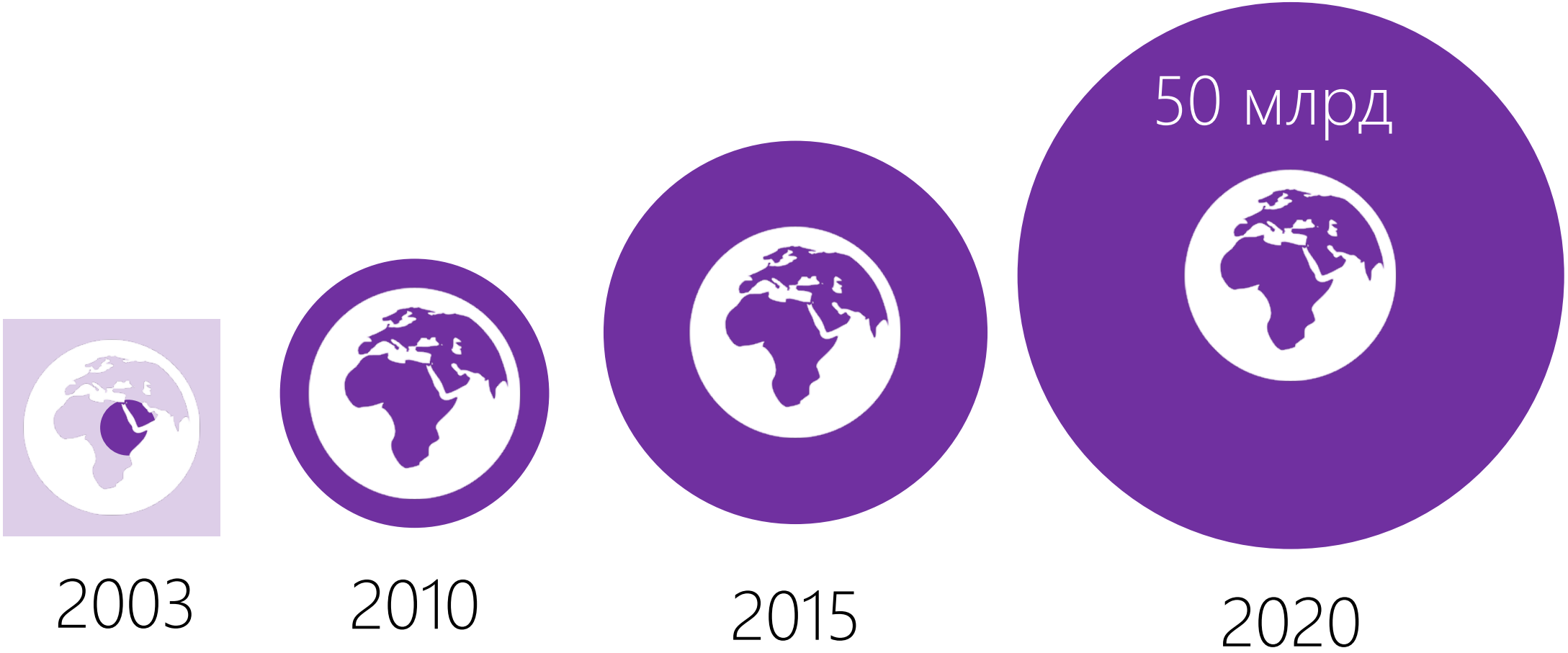
Виклики

- Інфраструктура
  - необхідність потужних серверів та розподілених обчислень
- Безпека
  - ризики витоків даних
  - необхідність шифрування та
  - контролю доступу
- Автоматизація
  - нові підходи до обробки потоків даних у реальному часі
- Етика
  - правильне використання персональних даних
  - уникнення алгоритмічних упереджень

Світ сьогодні

- $1 \cdot 10^9$ смартфонів
- $1{,}8 \cdot 10^9$ компьютерів
- $4{,}5 \cdot 10^9$ традиційних телефонів
- $8 \cdot 10^9$ жителів Землі

Зростання кількості пристроїв



Світ завтра

- 8,5 – $10 \cdot 10^9$ жителів Землі
- 12 – $30 \cdot 10^9$ клієнтських пристроїв
- 50 –$150 \cdot 10^9$ об'єднані в мережу контролерів
- 50 –$100 \cdot 10^{12}$ ідентифікованих об'єктів
- 200 – $1000 \cdot 10^{21}$ байт хмарного трафіку / рік

1990

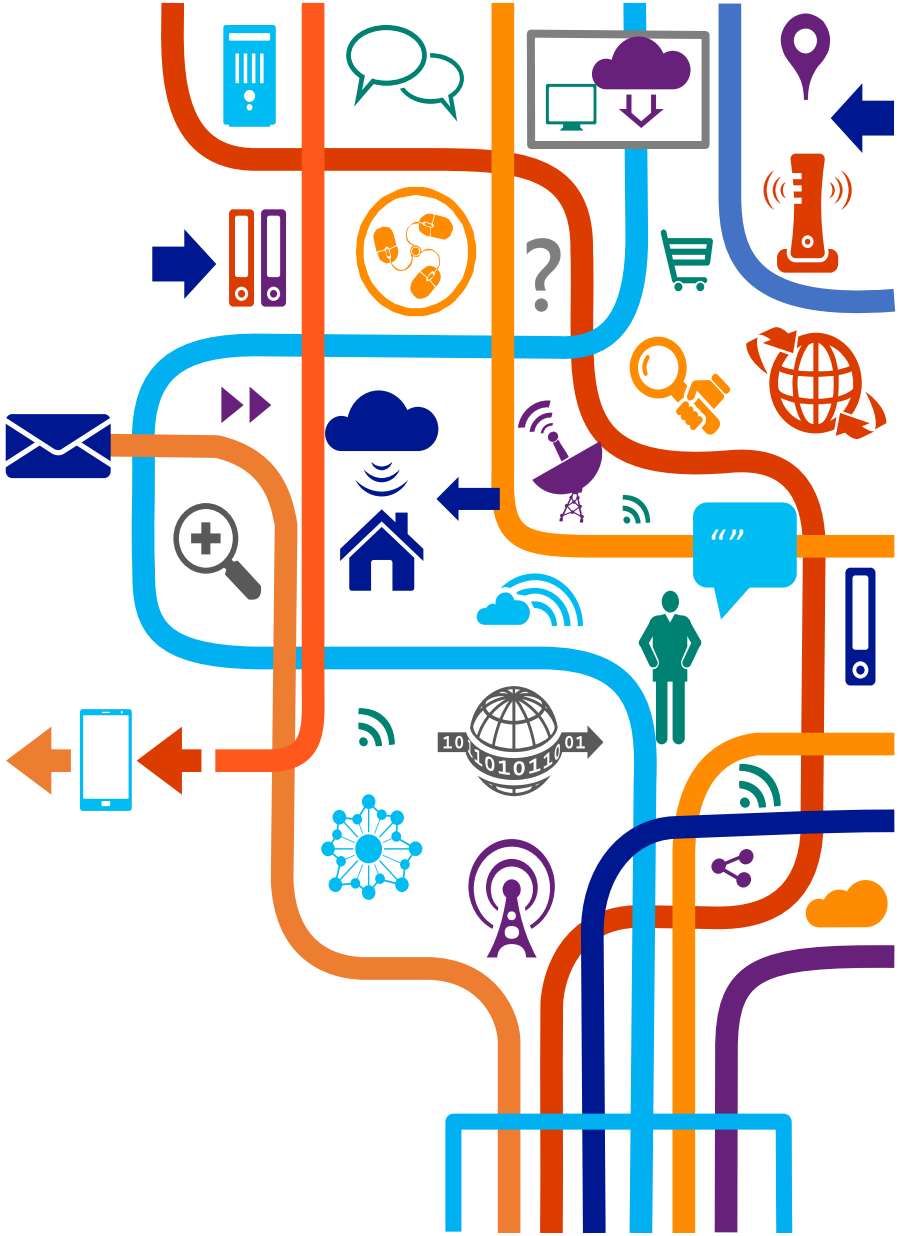
Інтернет

2020

IOT

Що таке інтернет речей?

- Мережа фізичних об'єктів з вбудованою електронікою, здатною підтримувати внутрішній стан, спілкуватися з іншими об'єктами і оточенням.



Питання

Корова + = IoT ?

Два важливих аспекти IoT

Сенсори

Хмара

Клієнтські пристрої

Великі дані

ВЕЛИКІ ДАНІ: дані, отримані з «Інтернету речей» і інших бізнес-систем

БІЛЬШ ЯКІСНІ ідеї: якісні ідеї відкривають можливість перетворення бізнесу та науки

Фундаментальні зміни – 1

- **Volume.** Величезні «обсяги» даних, які організації отримують з бізнес-транзакцій, інтелектуальних (IoT) пристроїв, промислового обладнання, соціальних мереж та інших джерел, потрібно десь зберігати. У минулому це було проблемою, але розвиток систем зберігання інформації полегшило ситуацію і зробило інформацію доступнішою.
- **Velocity.** Найчастіше цей пункт відноситься до швидкості приросту, з якої дані надходять в реальному часі. У більш широкому розумінні характеристика пояснює необхідність високошвидкісної обробки через темпи зміни і сплески активності.
- **Variety.** Різноманітність великих даних проявляється в їх форматах: структуровані цифри з клієнтських баз, неструктуровані текстові, відео- і аудіофайли, а також напівструктурована інформація з кількох джерел. Якщо раніше дані можна було збирати тільки з електронних таблиць, то сьогодні дані надходять в різному вигляді: від електронних листів до голосових повідомлень.

Фундаментальні зміни – 2

Традиційна аналітика	**Big data аналітика**
Поступовий аналіз невеликих пакетів даних	Обробка відразу всього масиву доступних даних
Редакція і сортування даних перед обробкою	Дані обробляються в їх початковому вигляді
Старт з гіпотези і її тестування щодо даних	Пошук кореляцій за всіма даними до отримання шуканої інформації
Дані збираються, обробляються, зберігаються і лише потім аналізуються	Аналіз і обробка великих даних в реальному часі, в міру надходження

Від гіпотез до кореляцій

- Традиційний науковий підхід передбачає формулювання гіпотез та їх перевірку.
- Нові методи, базовані на аналізі великих даних, спершу шукають кореляції, а вже потім формують гіпотези.
- Однак, важливим залишається розуміння: кореляція ≠ причинність. Це одна з найбільших проблем у застосуванні ML у науці.

Аналітика стає доступною для кожного

3-я хвиля
Аналіз кожною особою

2-а хвиля
Самостійний аналіз

1-а хвиля
Технічний аналіз

Всі

Аналітик для кінцевого користувача

IT для кінцевого користувача

Питання?