Course: Data Science Tools and Techniques

Dr. Safdar Ali

Number of Lectures	Topics	Material
3	**Introduction** Data Science Life Cycle, Motivation, Market Value, Issues, Challenges and Opportunities	Reference Textbook, Online references
6	**Data Preprocessing** Clean and filter the data, convert the data from one format to another	Reference Textbook, Online references
3	**Data Visualization** Visualizing the data in different ways	Reference Textbook
6	**Probabilistic View of Data** Basics of probability and statistics, Bayes rule, text Modelling	Reference Textbook
8	**Data Modelling** Machine Learning, classification, regression, clustering	Online Reference
2	**Model Evaluation & Performance Metrics** Train/val/test splits, accuracy, precision-recall, F-1, etc.	Reference Textbook
8	**Big Data Processing Tools** Hadoop and its Ecosystem, Spark	Online references, Research Papers
6	**Diverse Topics in Data Science** Various recent Trends in Data Science, Ethical Issues, Research Opportunities	Research Papers

Dictionary

Definitions from Oxford Languages · Learn more

data

/ˈdeɪtə/

noun

facts and statistics collected together for reference or analysis.
"there is very little data available"

Similar: facts figures statistics details particulars specifics features

- the quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.
- PHILOSOPHY
  things known or assumed as facts, making the basis of reasoning or calculation.

Data

Data are a collection of discrete or continuous values that convey information, describing the quantity, quality, fact, statistics, other basic units of meaning, or simply sequences of symbols that may be further interpreted formally. A datum is an individual value in a collection of data. Wikipedia ›

Dictionary

Definitions from Oxford Languages · Learn more

science

/ˈsʌɪəns/

noun

1. the systematic study of the structure and behaviour of the physical and natural world through observation, experimentation, and the testing of theories against the evidence obtained.
   "the world of science and technology"

   Similar: branch of knowledge · area of study · discipline · field
2. ARCHAIC
   knowledge of any kind.
   "his rare science and his practical skill"

Science

Discipline

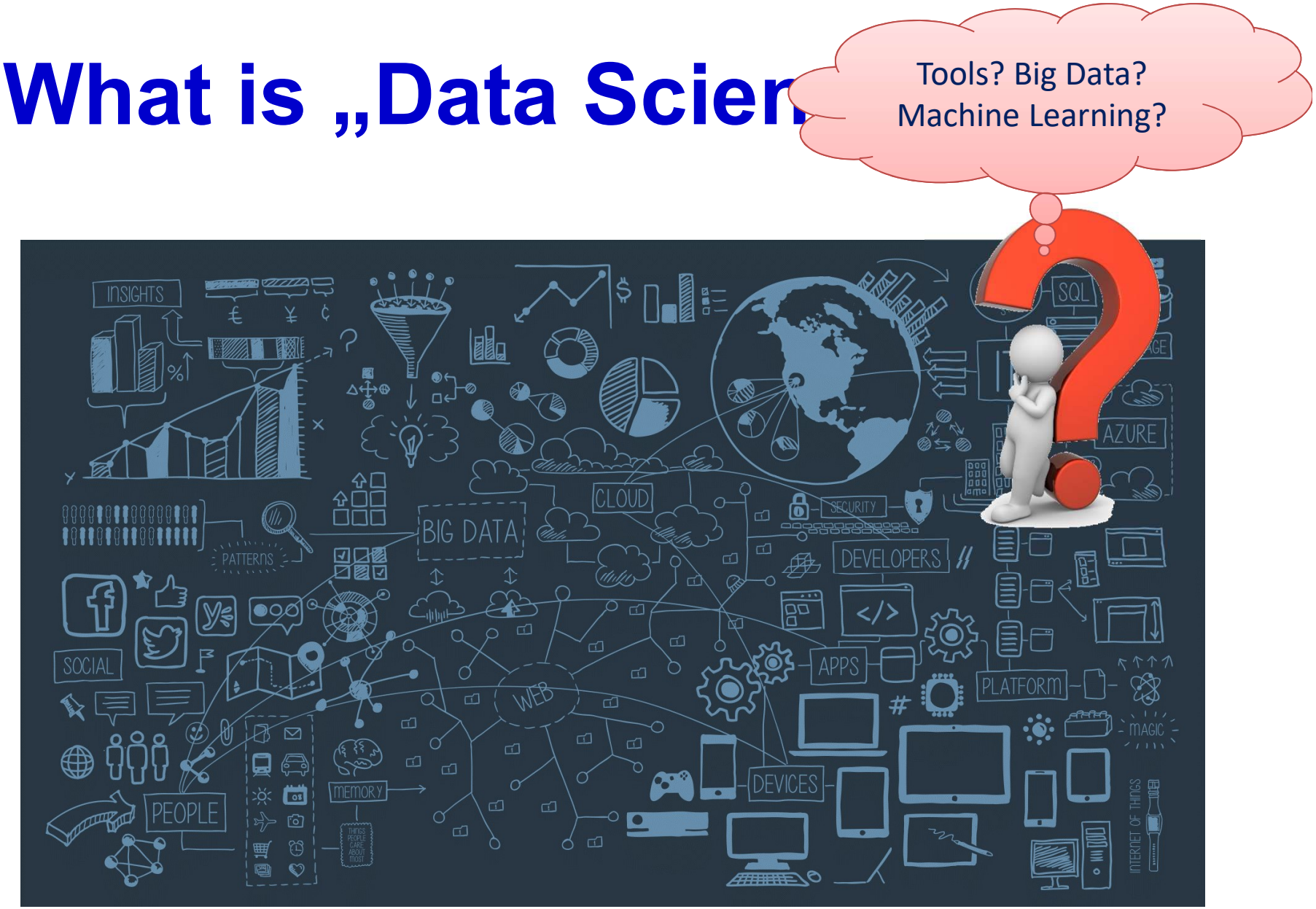
Science is a systematic discipline that builds and organises knowledge in the form of testable hypotheses and predictions about the universe. Wikipedia

Science, any system of knowledge that is concerned with the physical world and its phenomena and that entails unbiased observations and systematic experimentation. In general, a science involves a pursuit of knowledge covering general truths or the operations of fundamental laws.

Defining Data Science

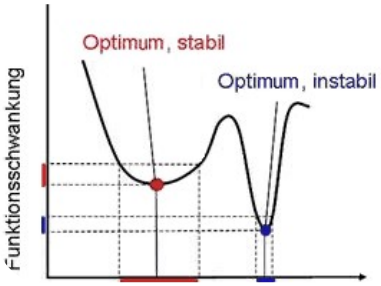
- Unfortunately, there is no clear definition (yet?)
- Goal is the extraction of knowledge from data
- Combination of techniques from different disciplines
- Scientific principles guide the data anaylsis

What is „Data Scien



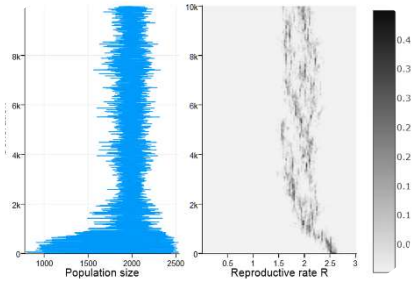
Mathematical Aspects

Computational Geometry



Optimization

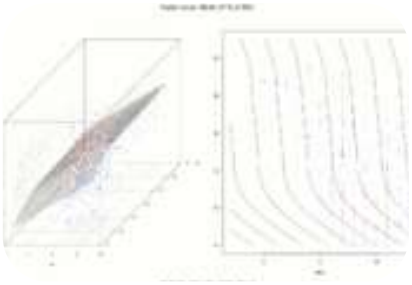
Stochastics



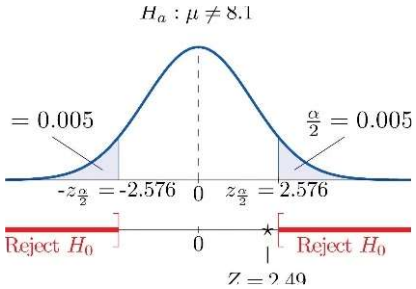
Scientific Computing

Machine Learning

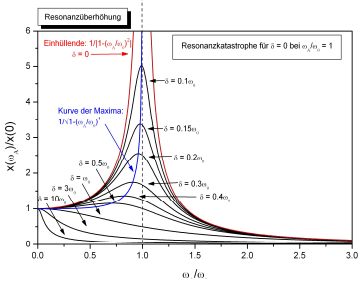
Statistical Aspects



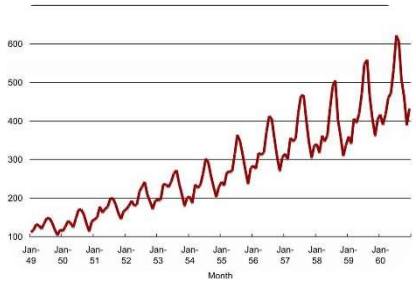
Linear Models



Statistical Tests



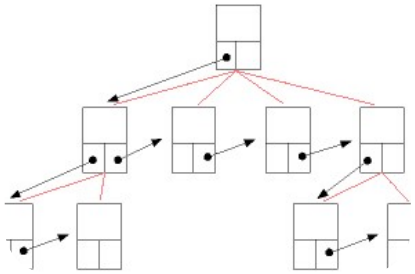
Inference



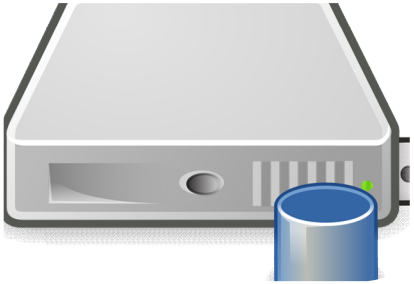
Time Series Analysis

Machine Learning

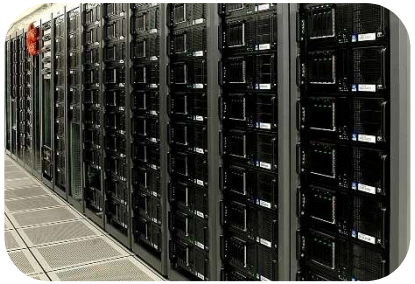
Computer Science Aspects



Data Structures and Algorithms

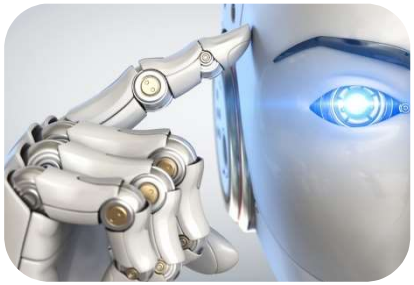


Databases



Distributed Computing

Software Engineering



Artificial Intelligence

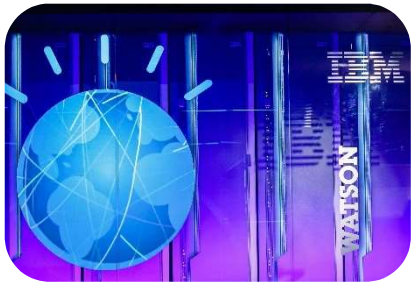
Machine Learning

Applications

Intelligent Systems

Robotics

Marketing



Medicine

Autonomous Driving

Social Networks

Data Science

- A multidisciplinary field

combining
- Math/statistical methods
- computer science
- domain expertise and
- data visualization

to extract insights and knowledge from structured and unstructured data.

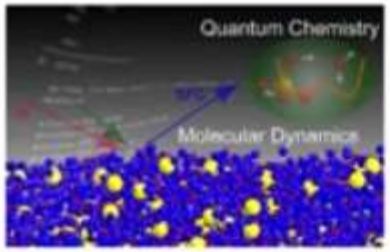
- It powers decisions in various industries such as healthcare, finance, retail, manufacturing, and technology.

Data Science

- **Data Science** is the science of analyzing **raw data** using statistics and machine learning techniques with the purpose of drawing conclusions about that information
- **Data Science** is used in many industries to allow them to make better business decisions, and in the sciences to test models or theories
- This requires a process of inspecting, cleaning, transforming, modeling, analyzing, and interpreting raw data

Data is everywhere

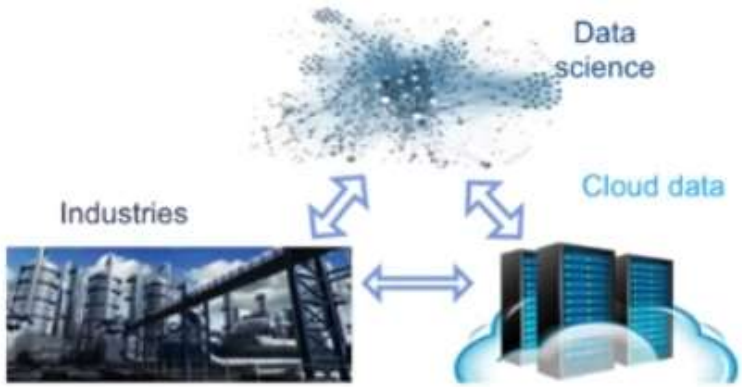
- Enormous amounts of data are being collected and stored
  - Web data, e-commerce
  - Point-of-sale at stores
  - Bank transactions
  - Social Network
  - Real-time plant data
  - Materials Formulation & Design
  - Molecular Simulations



Internet of Things (IOT) + Data Science

- IOT
  - Information from anywhere
  - Information from anything
  - Accessible, query-able and organized
  - Uncover interconnections and patterns that were not possible to derive before
- Data science
  - Algorithms for deriving useful information about state of the equipment and the process
  - All-weather algorithms for all kinds of problems
  - Self-learning

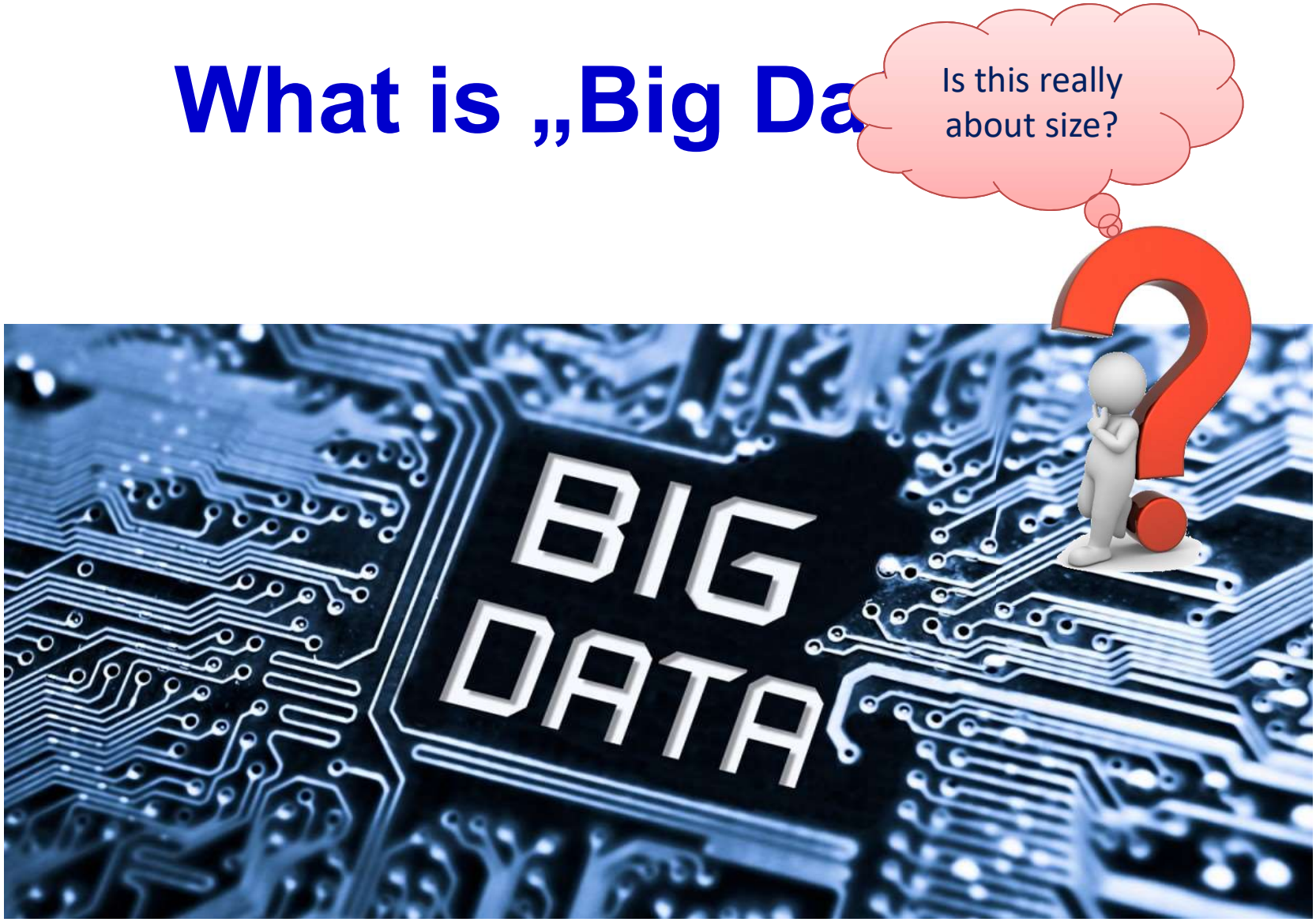
- For engineering processes, impact in
  - Productivity
  - Profitability
  - Safety



Data Science - Essentials

- Collecting does not mean discovering
  - Data should lead to value propositions
  - The science (and art) of converting "collected data" to useful knowledge is called Data Science
- Data Science
  - Domain knowledge
  - Statistics and Machine Learning
  - Software

What is „Big Da



Naive Definition

- Naive definition:
  - Big data only depends on the data size
  - **1 Gigabyte? 1 Terabyte? 1 Petabyte?**

- Naive interpretation misses important aspects
  - **Time:**
    - Analyzing 1 Gigabyte of data per day is different from analyzing 1 Gigabyte of data per second
  - **Diversity:**
    - Analyzing spread sheets with numeric data is different from analyzing Web pages that contain a mixture of text and images
  - **Distribution:**
    - Analyzing data from a single source is different from analyzing data from multiple sources

Definition of Big Data

- Following Gartner‘s IT Glossary:
  - Big data is high-**volume**, high-**velocity** and/or high-**variety** information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

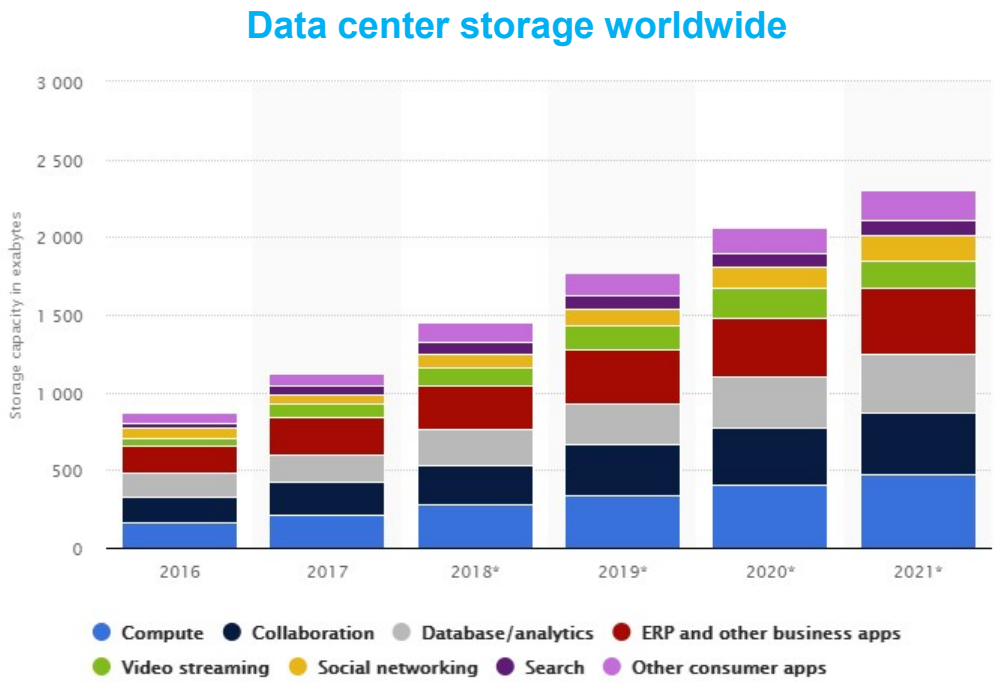
- The three Vs
  - **Volume**
  - **Velocity**
  - **Variety**

Some people actually use 10 Vs to define big data!

- Variability
- Veracity
- Validity
- Vulnerability
- Volatility
- Visualization
- Value

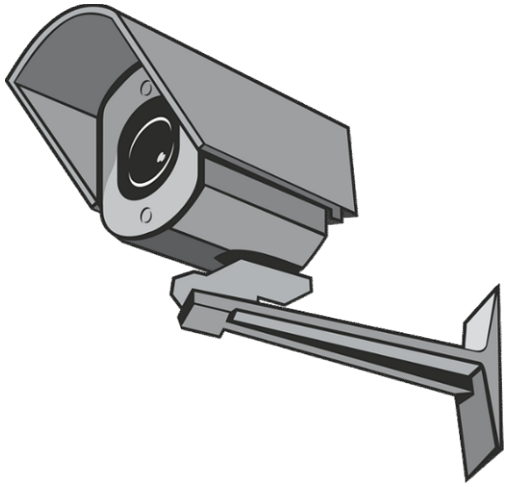
The 3 Vs: Volume

- Scale of the data must be „big“
  - No clear definition
  - „that demand […] innovative forms of information processing“ (Gartner)



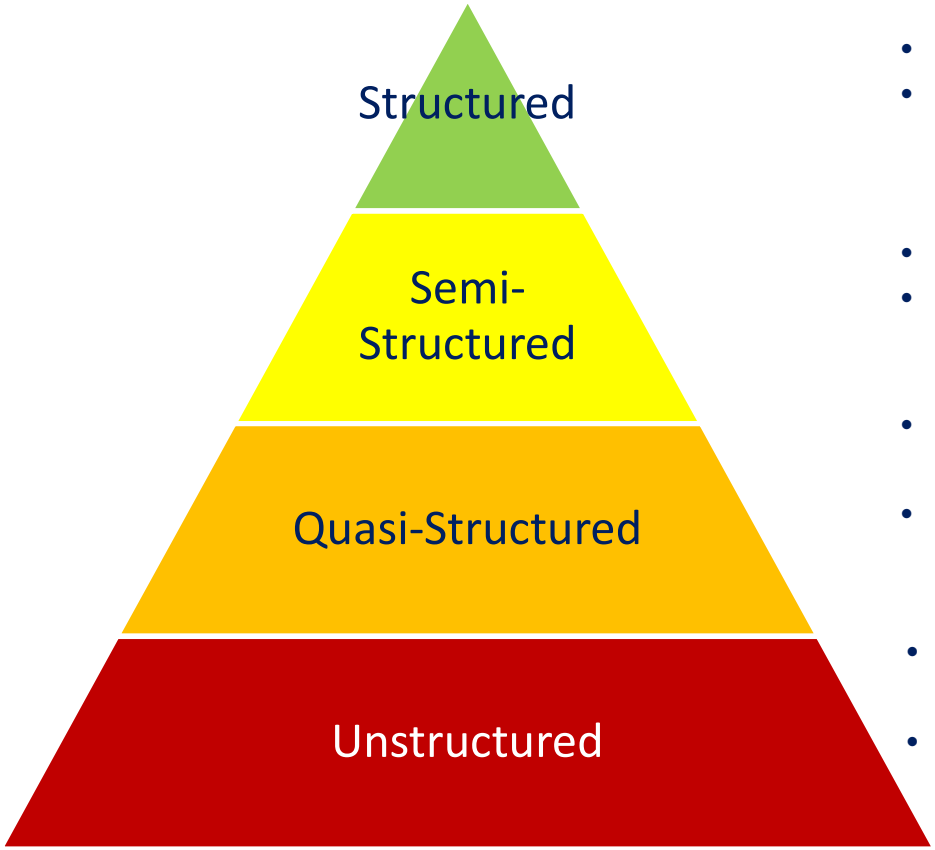
The 3 Vs: Velocity

- Speed at which new data is created
- Speed at which data must be processed and analyzed
  - Often close to real-time



The 3 Vs: Variety

- Diversity in data types and data sources



- Data with defined types and structure
- Example: comma separated values

- Textual data with parseable pattern
- Example: XML files with schema

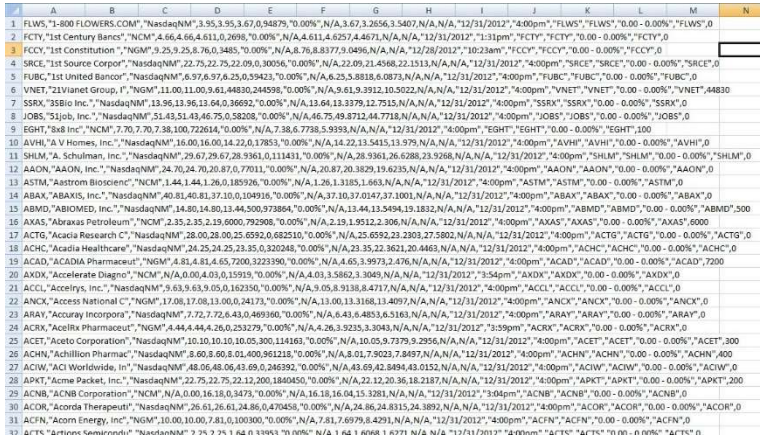
For example: Syntax: *<xs:element name = "phone_number" type = "xs:int" />*

- Textual data with erratic formats that can be formated with effort
- Example: Clickstream data

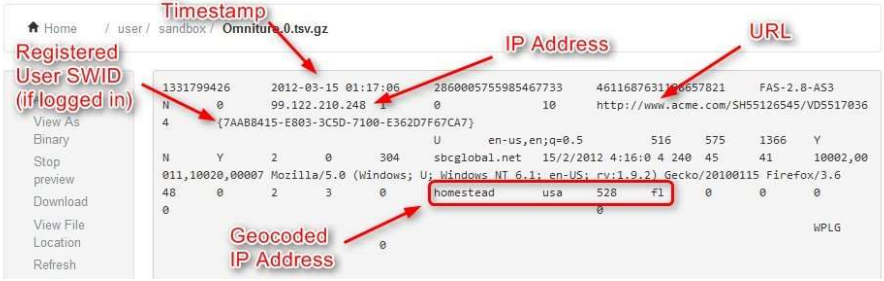
- Data that has no inherent structure, often with multiple formats
- Example: Web site, videos

Examples for data types

Structured



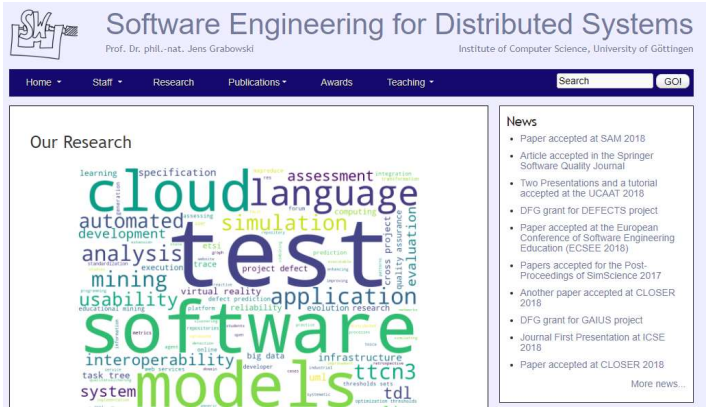
Quasi-Structured



Semi-Structured

```
<?xml version="1.0" encoding="iso-8859-8" standalone="yes" ?>
<CURRENCIES>
  <LAST_UPDATE>2004-07-29</LAST_UPDATE>
  <CURRENCY>
    <NAME>dollar</NAME>
    <UNIT>1</UNIT>
    <CURRENCYCODE>USD</CURRENCYCODE>
    <COUNTRY>USA</COUNTRY>
    <RATE>4.527</RATE>
    <CHANGE>0.044</CHANGE>
  </CURRENCY>
  <CURRENCY>
    <NAME>euro</NAME>
    <UNIT>1</UNIT>
    <CURRENCYCODE>EUR</CURRENCYCODE>
    <COUNTRY>European Monetary Union</COUNTRY>
    <RATE>5.4417</RATE>
    <CHANGE>-0.013</CHANGE>
  </CURRENCY>
</CURRENCIES>
```

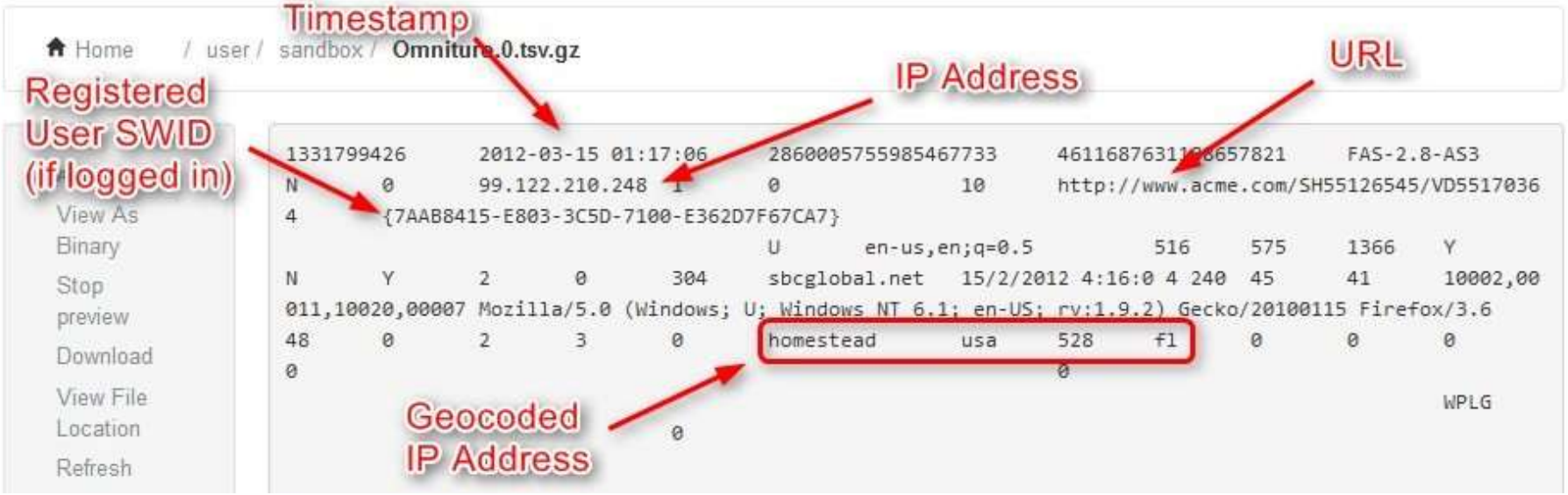
Unstructured



Structured

	A	J	K	L	M	N
1	FLWS,"1-800 FLOWERS.COM","NasdaqNM",3.95,3.95,3.67,0,94879,"0.00%",N/A,3.67,3.2656,3.5407,N/A,N/A,"12/31/2012","4:00pm","FLWS","FLWS","0.00 - 0.00%","FLWS",0					
2	FCTY,"1st Century Bancs","NCM",4.66,4.66,4.611,0,2698,"0.00%",N/A,4.611,4.6257,4.4671,N/A,N/A,"12/31/2012","1:31pm","FCTY","FCTY","0.00 - 0.00%","FCTY",0					
3	FCCY,"1st Constitution ","NGM",9.25,9.25,8.76,0,3485,"0.00%",N/A,8.76,8.8377,9.0496,N/A,N/A,"12/28/2012","10:23am","FCCY","FCCY","0.00 - 0.00%","FCCY",0					
4	SRCE,"1st Source Corpor","NasdaqNM",22.75,22.75,22.09,0,30056,"0.00%",N/A,22.09,21.4568,22.1513,N/A,N/A,"12/31/2012","4:00pm","SRCE","SRCE","0.00 - 0.00%","SRCE",0					
5	FUBC,"1st United Bancor","NasdaqNM",6.97,6.97,6.25,0,59423,"0.00%",N/A,6.25,5.8818,6.0873,N/A,N/A,"12/31/2012","4:00pm","FUBC","FUBC","0.00 - 0.00%","FUBC",0					
6	VNET,"21Vianet Group, I","NGM",11.00,11.00,9.61,44830,244598,"0.00%",N/A,9.61,9.3912,10.5022,N/A,N/A,"12/31/2012","4:00pm","VNET","VNET","0.00 - 0.00%","VNET",44830					
7	SSRX,"3SBio Inc.","NasdaqNM",13.96,13.96,13.64,0,36692,"0.00%",N/A,13.64,13.3379,12.7515,N/A,N/A,"12/31/2012","4:00pm","SSRX","SSRX","0.00 - 0.00%","SSRX",0					
8	JOBS,"51job, Inc.","NasdaqNM",51.43,51.43,46.75,0,58208,"0.00%",N/A,46.75,49.8712,44.7718,N/A,N/A,"12/31/2012","4:00pm","JOBS","JOBS","0.00 - 0.00%","JOBS",0					
9	EGHT,"8x8 Inc","NCM",7.70,7.70,7.38,100,722614,"0.00%",N/A,7.38,6.7738,5.9393,N/A,N/A,"12/31/2012","4:00pm","EGHT","EGHT","0.00 - 0.00%","EGHT",100					
10	AVHI,"A V Homes, Inc.","NasdaqNM",16.00,16.00,14.22,0,17853,"0.00%",N/A,14.22,13.5415,13.979,N/A,N/A,"12/31/2012","4:00pm","AVHI","AVHI","0.00 - 0.00%","AVHI",0					
11	SHLM,"A. Schulman, Inc.","NasdaqNM",29.67,29.67,28.9361,0,111431,"0.00%",N/A,28.9361,26.6288,23.9268,N/A,N/A,"12/31/2012","4:00pm","SHLM","SHLM","0.00 - 0.00%","SHLM",0					
12	AAON,"AAON, Inc.","NasdaqNM",24.70,24.70,20.87,0,77011,"0.00%",N/A,20.87,20.3829,19.6235,N/A,N/A,"12/31/2012","4:00pm","AAON","AAON","0.00 - 0.00%","AAON",0					
13	ASTM,"Aastrom Bioscienc","NCM",1.44,1.44,1.26,0,185926,"0.00%",N/A,1.26,1.3185,1.663,N/A,N/A,"12/31/2012","4:00pm","ASTM","ASTM","0.00 - 0.00%","ASTM",0					
14	ABAX,"ABAXIS, Inc.","NasdaqNM",40.81,40.81,37.10,0,104916,"0.00%",N/A,37.10,37.0147,37.1001,N/A,N/A,"12/31/2012","4:00pm","ABAX","ABAX","0.00 - 0.00%","ABAX",0					
15	ABMD,"ABIOMED, Inc.","NasdaqNM",14.80,14.80,13.44,500,973864,"0.00%",N/A,13.44,13.5494,19.1832,N/A,N/A,"12/31/2012","4:00pm","ABMD","ABMD","0.00 - 0.00%","ABMD",500					
16	AXAS,"Abraxas Petroleum","NCM",2.35,2.35,2.19,6000,792908,"0.00%",N/A,2.19,1.9512,2.306,N/A,N/A,"12/31/2012","4:00pm","AXAS","AXAS","0.00 - 0.00%","AXAS",6000					
17	ACTG,"Acacia Research C","NasdaqNM",28.00,28.00,25.6592,0,682510,"0.00%",N/A,25.6592,23.2303,27.5802,N/A,N/A,"12/31/2012","4:00pm","ACTG","ACTG","0.00 - 0.00%","ACTG",0					
18	ACHC,"Acadia Healthcare","NasdaqNM",24.25,24.25,23.35,0,320248,"0.00%",N/A,23.35,22.3621,20.4463,N/A,N/A,"12/31/2012","4:00pm","ACHC","ACHC","0.00 - 0.00%","ACHC",0					
19	ACAD,"ACADIA Pharmaceut","NGM",4.81,4.81,4.65,7200,3223390,"0.00%",N/A,4.65,3.9973,2.476,N/A,N/A,"12/31/2012","4:00pm","ACAD","ACAD","0.00 - 0.00%","ACAD",7200					
20	AXDX,"Accelerate Diagno","NCM",N/A,0.00,4.03,0,15919,"0.00%",N/A,4.03,3.5862,3.3049,N/A,N/A,"12/31/2012","3:54pm","AXDX","AXDX","0.00 - 0.00%","AXDX",0					
21	ACCL,"Accelrys, Inc.","NasdaqNM",9.63,9.63,9.05,0,162350,"0.00%",N/A,9.05,8.9138,8.4717,N/A,N/A,"12/31/2012","4:00pm","ACCL","ACCL","0.00 - 0.00%","ACCL",0					
22	ANCX,"Access National C","NGM",17.08,17.08,13.00,0,24173,"0.00%",N/A,13.00,13.3168,13.4097,N/A,N/A,"12/31/2012","4:00pm","ANCX","ANCX","0.00 - 0.00%","ANCX",0					
23	ARAY,"Accuray Incorpora","NasdaqNM",7.72,7.72,6.43,0,469360,"0.00%",N/A,6.43,6.4853,6.5163,N/A,N/A,"12/31/2012","4:00pm","ARAY","ARAY","0.00 - 0.00%","ARAY",0					
24	ACRX,"AcelRx Pharmaceut","NGM",4.44,4.44,4.26,0,253279,"0.00%",N/A,4.26,3.9235,3.3043,N/A,N/A,"12/31/2012","3:59pm","ACRX","ACRX","0.00 - 0.00%","ACRX",0					
25	ACET,"Aceto Corporation","NasdaqNM",10.10,10.10,10.05,300,114163,"0.00%",N/A,10.05,9.7379,9.2956,N/A,N/A,"12/31/2012","4:00pm","ACET","ACET","0.00 - 0.00%","ACET",300					
26	ACHN,"Achillion Pharmac","NasdaqNM",8.60,8.60,8.01,400,961218,"0.00%",N/A,8.01,7.9023,7.8497,N/A,N/A,"12/31/2012","4:00pm","ACHN","ACHN","0.00 - 0.00%","ACHN",400					
27	ACIW,"ACI Worldwide, In","NasdaqNM",48.06,48.06,43.69,0,246392,"0.00%",N/A,43.69,42.8494,43.0152,N/A,N/A,"12/31/2012","4:00pm","ACIW","ACIW","0.00 - 0.00%","ACIW",0					
28	APKT,"Acme Packet, Inc.","NasdaqNM",22.75,22.75,22.12,200,1840450,"0.00%",N/A,22.12,20.36,18.2187,N/A,N/A,"12/31/2012","4:00pm","APKT","APKT","0.00 - 0.00%","APKT",200					
29	ACNB,"ACNB Corporation","NCM",N/A,0.00,16.18,0,3473,"0.00%",N/A,16.18,16.04,15.3281,N/A,N/A,"12/31/2012","3:04pm","ACNB","ACNB","0.00 - 0.00%","ACNB",0					
30	ACOR,"Acorda Therapeuti","NasdaqNM",26.61,26.61,24.86,0,470458,"0.00%",N/A,24.86,24.8315,24.3892,N/A,N/A,"12/31/2012","4:00pm","ACOR","ACOR","0.00 - 0.00%","ACOR",0					
31	ACFN,"Acorn Energy, Inc","NGM",10.00,10.00,7.81,0,100300,"0.00%",N/A,7.81,7.6979,8.4291,N/A,N/A,"12/31/2012","4:00pm","ACFN","ACFN","0.00 - 0.00%","ACFN",0					
32	ACTS,"Actions Semicondu","NasdaqNM",2.25,2.25,1.64,0,33953,"0.00%",N/A,1.64,1.6068,1.6271,N/A,N/A,"12/31/2012","4:00pm","ACTS","ACTS","0.00 - 0.00%","ACTS",0					

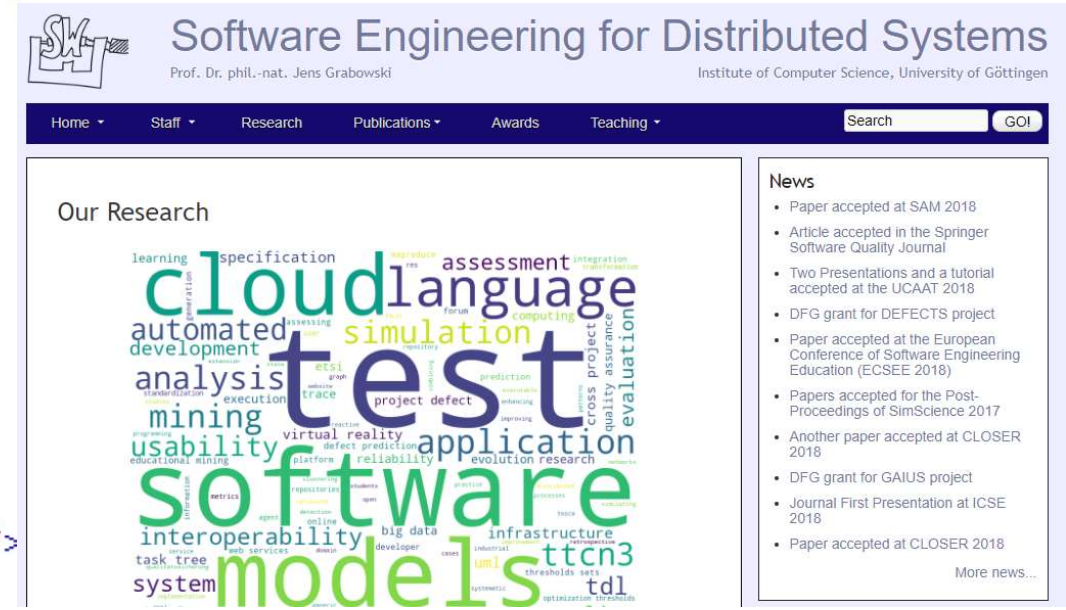
Quasi-Structured



Semi-Structured

```
<?xml version="1.0" encoding="iso-8859-8" standalone="yes" ?>
<CURRENCIES>
  <LAST_UPDATE>2004-07-29</LAST_UPDATE>
  <CURRENCY>
    <NAME>dollar</NAME>
    <UNIT>1</UNIT>
    <CURRENCYCODE>USD</CURRENCYCODE>
    <COUNTRY>USA</COUNTRY>
    <RATE>4.527</RATE>
    <CHANGE>0.044</CHANGE>
  </CURRENCY>
  <CURRENCY>
    <NAME>euro</NAME>
    <UNIT>1</UNIT>
    <CURRENCYCODE>EUR</CURRENCYCODE>
    <COUNTRY>European Monetary Union</COUNTRY>
    <RATE>5.4417</RATE>
    <CHANGE>-0.013</CHANGE>
  </CURRENCY>
</CURRENCIES>
```

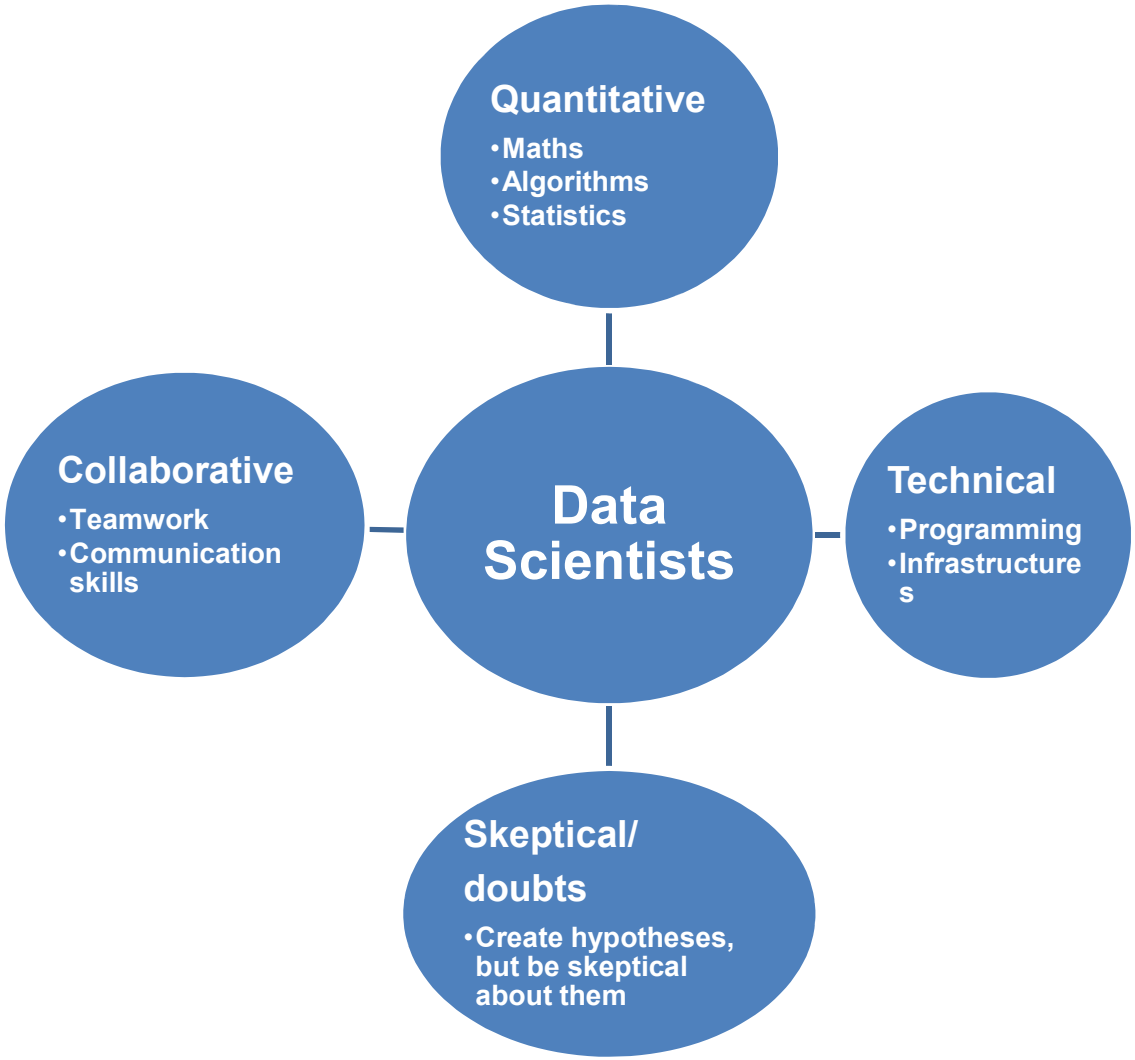
Unstructured



What are Data Scientists?

- **Not computer scientists**
  - But should know about databases, data structures, algorithms, etc.
- **Not mathematicians**
  - But should know about optimization, stochastics, etc.
- **Not statisticians**
  - But should know about regression, statistical tests, etc.
- **Not domain experts**
  - But must work together with them

Skills of Data Scientists



A bit of everything

… but actually as much as possible of everything

- **Rare qualities**

Data science takes unstructured data, then finds order, meaning and value

- **High demand**

Data science provides insight and competitive advantage

Different types of Data Scientists

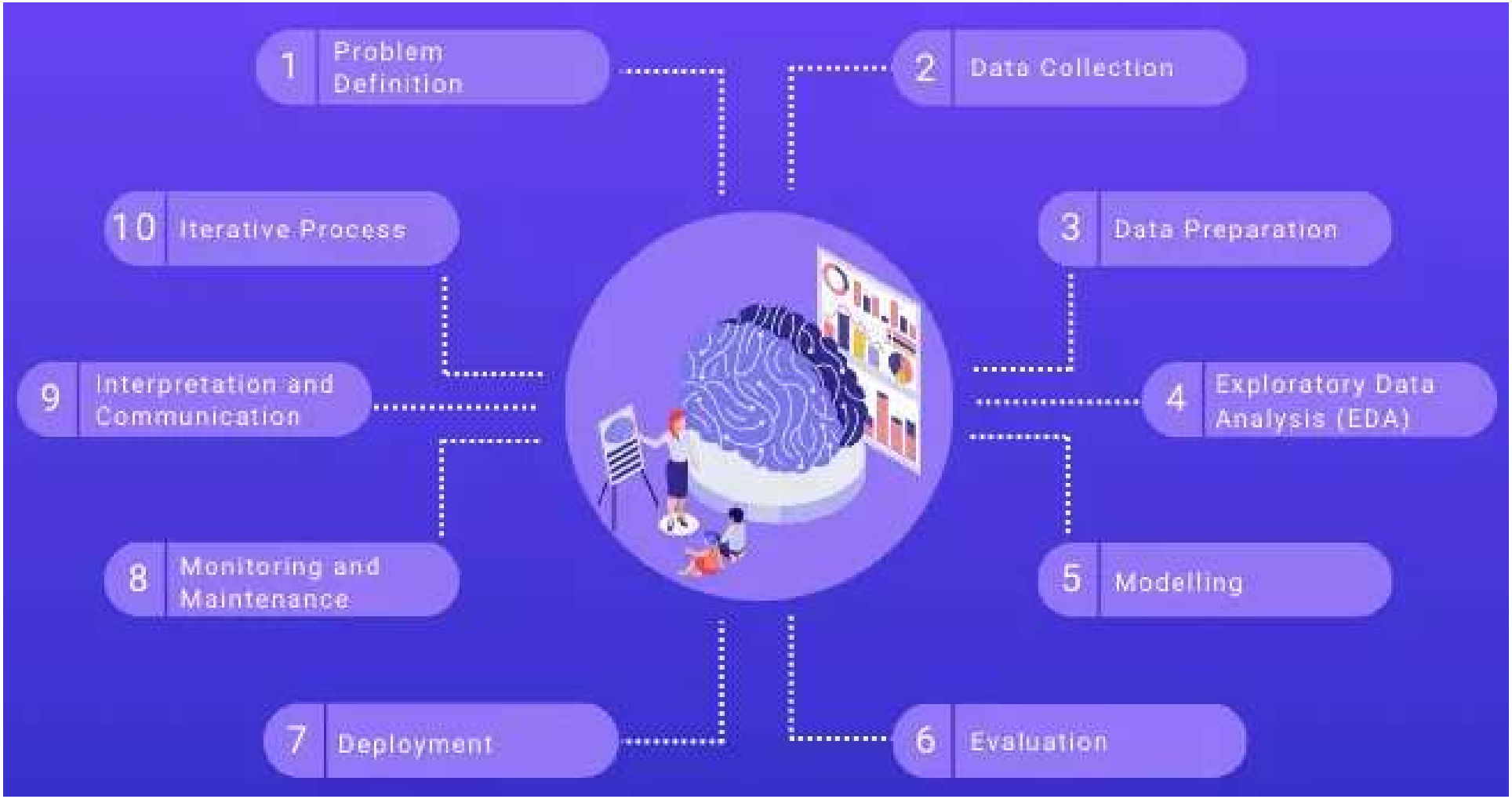
- **According to Microsoft Research:**
  - **Polymath**
    - **„Do it all“**
  - **Data Evangelist**
    - **Data analysis, disseminating and acting on insights**
  - **Data Preparer**
    - **Querying existing data, preparing data for analysis**
  - **Data Shapers**
    - **Analyzing and preparing data**
  - **Data Analyzer**
    - **Analyzing data**
  - **Platform Builder**
    - **Collect data and create infrastructures**
  - **Moonlighters (50%/20%)**
    - **„Spare time“ data scientists**
  - **Insight Actors**
    - **Use the outcome and act on insights.**

Miyung Kim, Thomas Zimmermann, Robert DeLine, Andrew Begel: Data Scientists in Software Teams: State of the Art and Challenges, IEEE Transactions on Software Engineering (Online First)

Data Science Life Cycle

- **It represents the structured process of solving data-driven problems.**
- **It provides a systematic approach for collecting, processing, analyzing, and leveraging data to generate insights and drive decision-making.**

Data Science Life Cycle -Stages



Stage 1: Problem Definition

Aim: Identify the problem and define clear goals.

Major Actions:

- Understand the problem domain.
- Define project objectives and success criteria.
- Determine the scope and limitations.

Questions to Ask:

- What is the problem we aim to solve?
- What decisions will be made based on the analysis?

Stage 2: Data Collection

Aim: Gather relevant data from various sources.

Major Actions:

- Identify data requirements (structured, unstructured, or semi-structured).
- Collect data from databases, APIs (*Application Programming Interface*), web scraping, surveys, or sensors.
- Ensure data relevance and sufficiency.

Common Tools:

- SQL, Python (requests, BeautifulSoup), web APIs.

Stage 3: Data Preparation

Aim: Clean and preprocess data for analysis.

Major Actions:

- Handle missing values, outliers, and duplicates.
- Transform and encode data (normalization, standardization, encoding).
- Split data into training, validation, and test sets.

Challenges:

- Poor-quality data may lead to inaccurate results.

Common Tools:

- Python (Pandas, NumPy), R.

Stage 4: Exploratory Data Analysis (EDA)

Aim: Explore and understand data patterns and trends.

Major Actions:

- Visualize distributions, correlations, and relationships.
- Generate descriptive statistics and summaries.
- Detect anomalies or biases in the data.

•Common Tools:

- Python (Matplotlib, Seaborn), Tableau, Power BI.

Stage 5: Modeling and Algorithm Development

Aim: Build predictive or descriptive models using machine learning or statistical methods.

Major Actions:

- Select appropriate algorithms based on the problem type (classification, regression, clustering, etc.).
- Train models on the data and fine-tune hyperparameters.
- Use techniques like cross-validation to prevent overfitting.

Common Algorithms:

- Linear regression, decision trees, SVM, random forests, k-means, neural networks.

Common Tools:

- Python (Scikit-learn, TensorFlow, PyTorch), R.

Stage 6: Model Evaluation

Aim: Assess the model's performance and reliability.

Major Actions:

- Use metrics like accuracy, precision, recall, Sn, Sp, F1-score, and ROC-AUC.
- Compare multiple models to select the best-performing one.
- Validate results using test data.

Challenges:

- Balancing bias and variance.

Stage 7: Deployment

Aim: Implement the model into production for practical use.

Major Actions:

- Create APIs or dashboards for real-time predictions.
- Integrate models into existing business systems or workflows.
- Ensure scalability and reliability.

Common Tools:

- Flask, Docker, Kubernetes, cloud platforms (AWS, Azure, GCP).

Amazon Web Service, Microsoft Azure: Cloud Computing Services , Google Cloud Platform

Stage 8: Monitoring and Maintenance

Aim: Track the model's performance over time and adapt as needed.

Major Actions:

- Monitor metrics to detect performance degradation.
- Update or retrain models to incorporate new data.
- Handle concept drift (when data patterns change over time).

Common Tools:

- Prometheus, Grafana, MLflow.

Challenges in the Data Science Life Cycle

Data Quality Issues: Missing or inconsistent data can compromise analysis.

Ethical Concerns: Ensuring data privacy and addressing biases.

Scalability: Handling large-scale data efficiently.

Collaboration: Bridging gaps between data scientists, engineers, and business teams.

Summary

- Data science is so much technical… but creative.
- Reality.

- Use tools from coding, statistics, and math to work creatively with data
- Goal is insight.
- Because everything signifies

Best Practices

1. Start with a well-defined problem statement.
2. Maintain clear documentation at every stage.
3. Prioritize data quality and ethical considerations.
4. Regularly validate and update models to ensure relevance.
5. Use version control for data and code.

References

Textbooks

1."Data Science for Business" by Foster Provost and Tom Fawcett
2."An Introduction to Statistical Learning" by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani
3."Python for Data Analysis" by Wes McKinney
4."Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" by Aurélien Géron

Online References

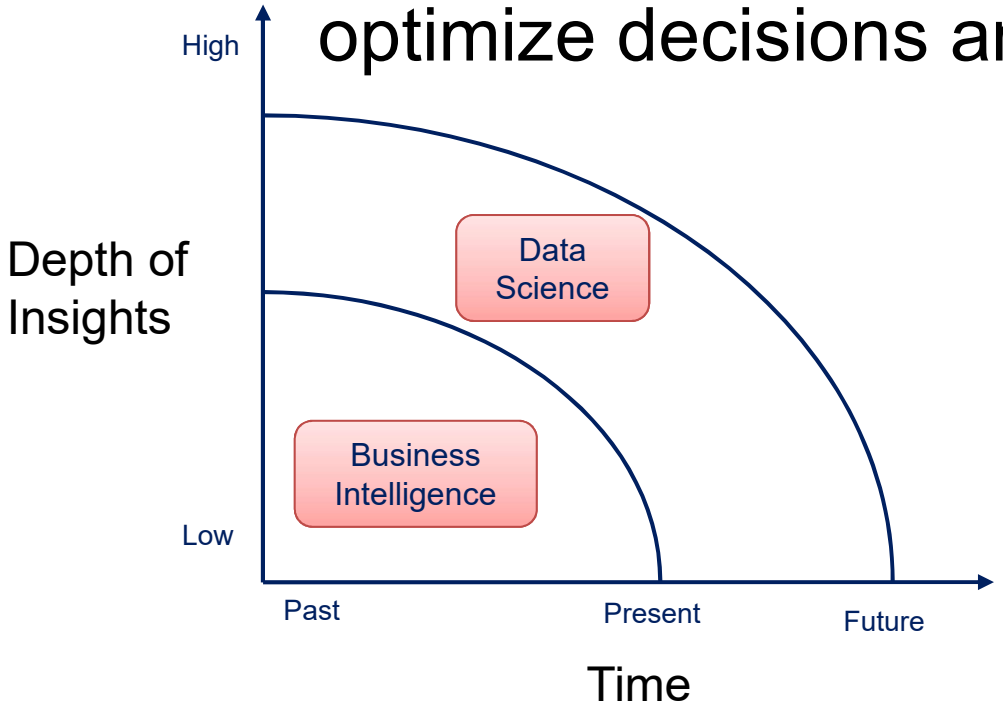
1.Scikit-learn Documentation - Comprehensive guide for machine learning in Python (https://scikit-learn.org)
2.DataCamp - Interactive learning platform for data science (https://www.datacamp.com)
3.UCI Machine Learning Repository - Open dataset repository (https://archive.ics.uci.edu/ml/index.php)

Data Repositories

1.Kaggle Datasets - A large collection of datasets for practice (https://www.kaggle.com/datasets)
2.Google Dataset Search - A search engine for datasets (https://datasetsearch.research.google.com)
3.Open Data Portal (data.gov) - U.S. government's open data (https://www.data.gov)
4.GitHub Datasets - Public datasets hosted on GitHub (https://github.com/awesomedata/awesome-public-datasets)
5.World Bank Open Data - Economic and development data (https://data.worldbank.org)
6.AWS Public Datasets - Large-scale datasets hosted on AWS (https://registry.opendata.aws)

Data Science vs. Business Intelligence

- Business Intelligence (Gartner IT Glossary)
  - […] best practices that enable access to and analysis of information to improve and optimize decisions and performance.



	Business Intelligence	Data Science
Techniques	Dashboards, alerts, queries	Optimization, predictive modelling, forecasting
Data Types	Structured, data warehouses	Any kind, often unstructured
Common questions	What happened…? How much did…? When did…?	What if…? What will…? How can we…?

More Data → More Opportunities

1990's
Relational Databases
& Data Warehouses

2000's
Content Management

2010's
Key-Value Storages
& Unstructured Data

one Exabyte (EB) = 1,000 Petabytes or one billion gigabytes (GB)

Motivation

Actionable Insights: Data Science enables organizations to uncover actionable insights for strategic decision-making and problem-solving.

Competitive Advantage: Organizations that adopt data-driven strategies often outperform their competitors and businesses can predict market trends, identify opportunities, and optimize operations.

Automation and Efficiency: Improving operational efficiency through predictive analytics, artificial intelligence, machine learning, and reducing manual effort.

Innovation: It drives innovation by enabling the creation of new data-driven products, services, and solutions. Examples include personalized recommendation engines, autonomous vehicles, and smart healthcare systems.

Improved Decision-Making: Data Science provides leaders with quantitative evidence to make informed decisions, reducing reliance on intuition and guesswork.

Motivation

Cost Reduction: By identifying inefficiencies, forecasting demand, and detecting fraud, Data Science can significantly reduce costs across industries.

Social Impact: Beyond businesses, Data Science is used to solve global challenges such as disease outbreaks, climate change, and poverty. For instance, predictive models can help allocate resources during natural disasters.

Empowering Small Businesses: With accessible tools and platforms, even small and medium enterprises (SMEs) can harness the power of Data Science to optimize their operations and enhance customer experiences.

Adaptability in a Data-Driven World: In a world increasingly reliant on data, proficiency in Data Science ensures adaptability to changing technologies and market demands.

Market Value

High Demand:

- Companies require data scientists to manage and interpret large volumes of data.
- Rapid growth in data from IoT, social media, and e-commerce.

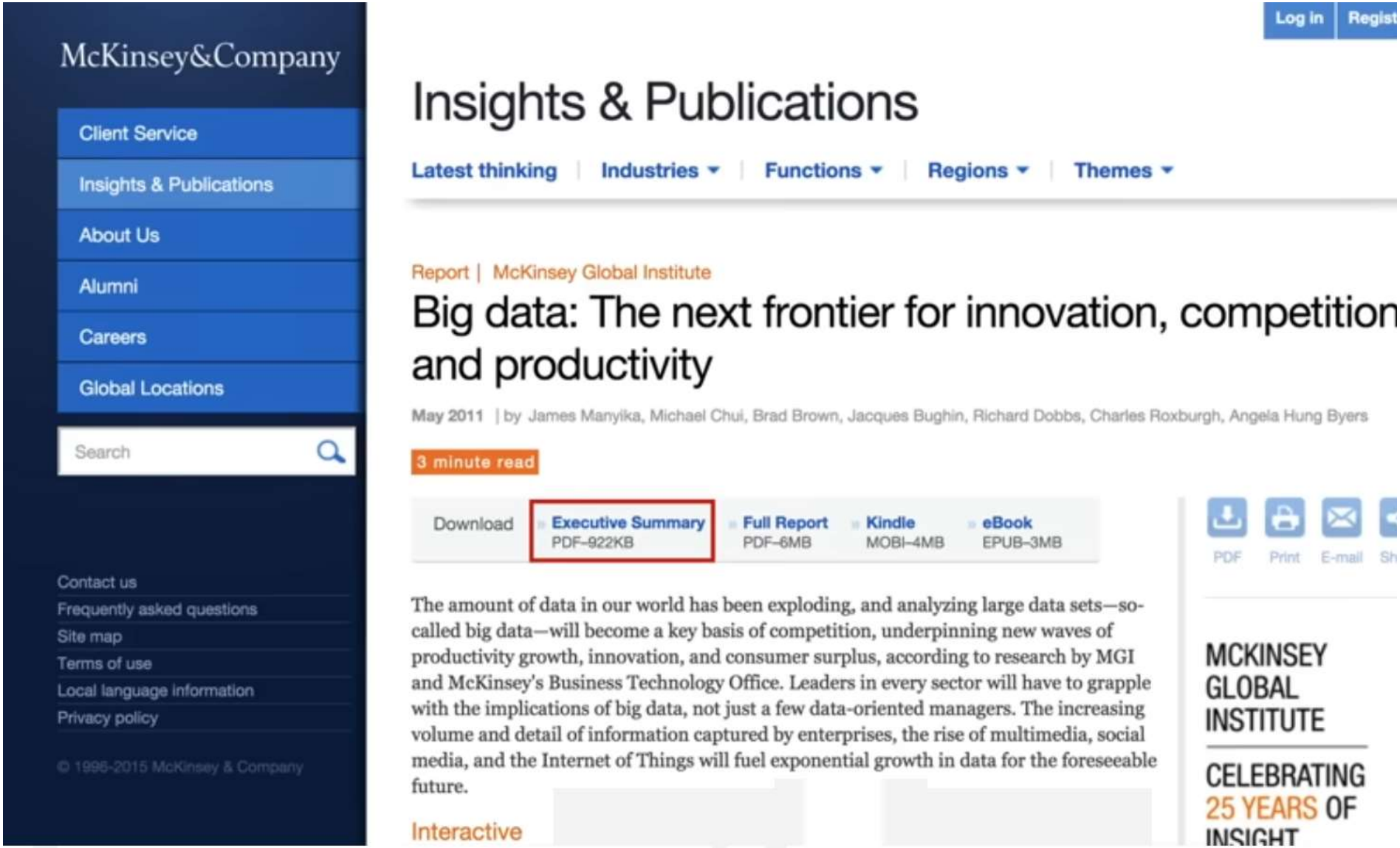
Salary Trends:

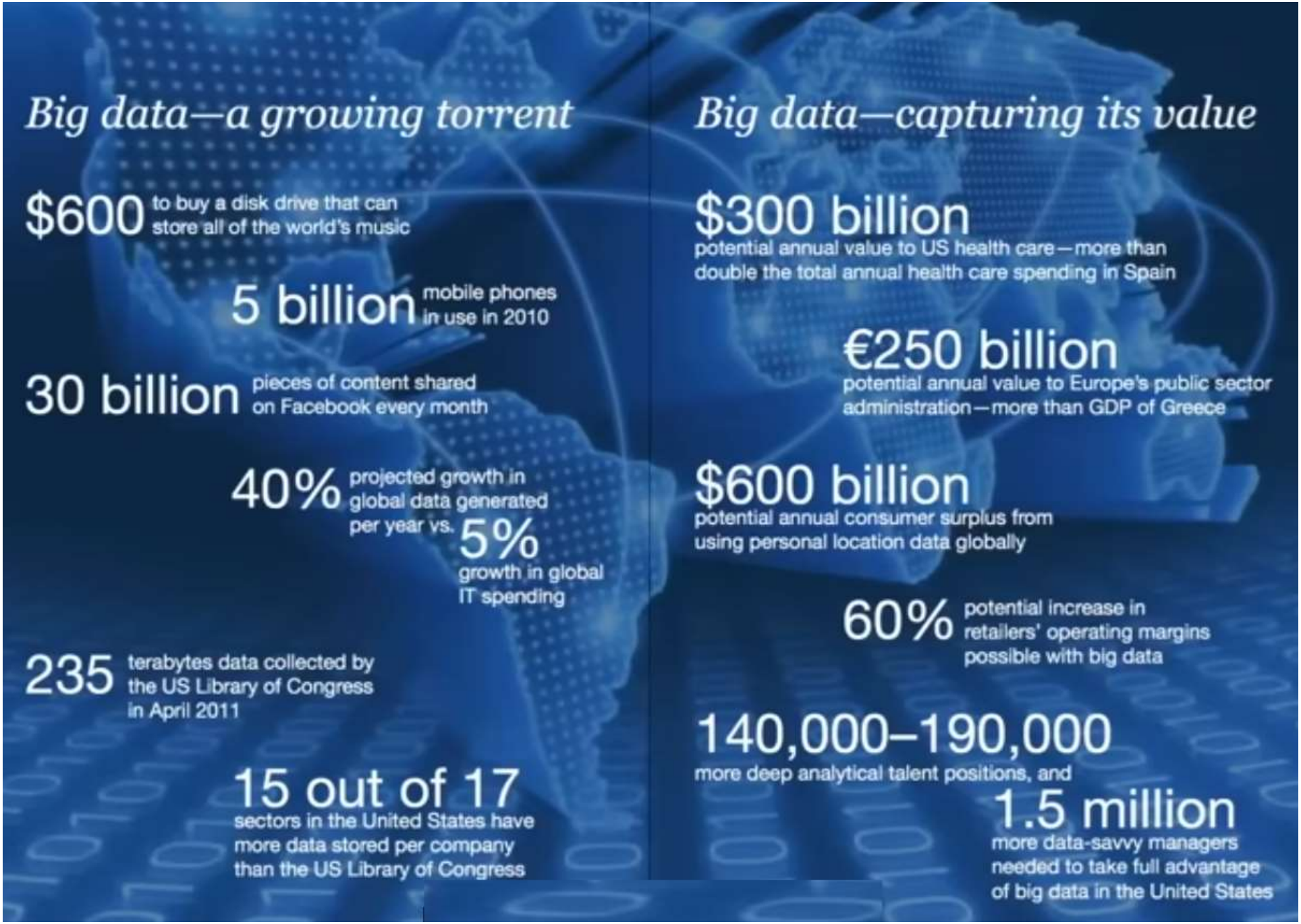
- Data Science roles are among the highest-paying positions globally.
- Salaries vary based on expertise, location, and industry.

Industry Applications:

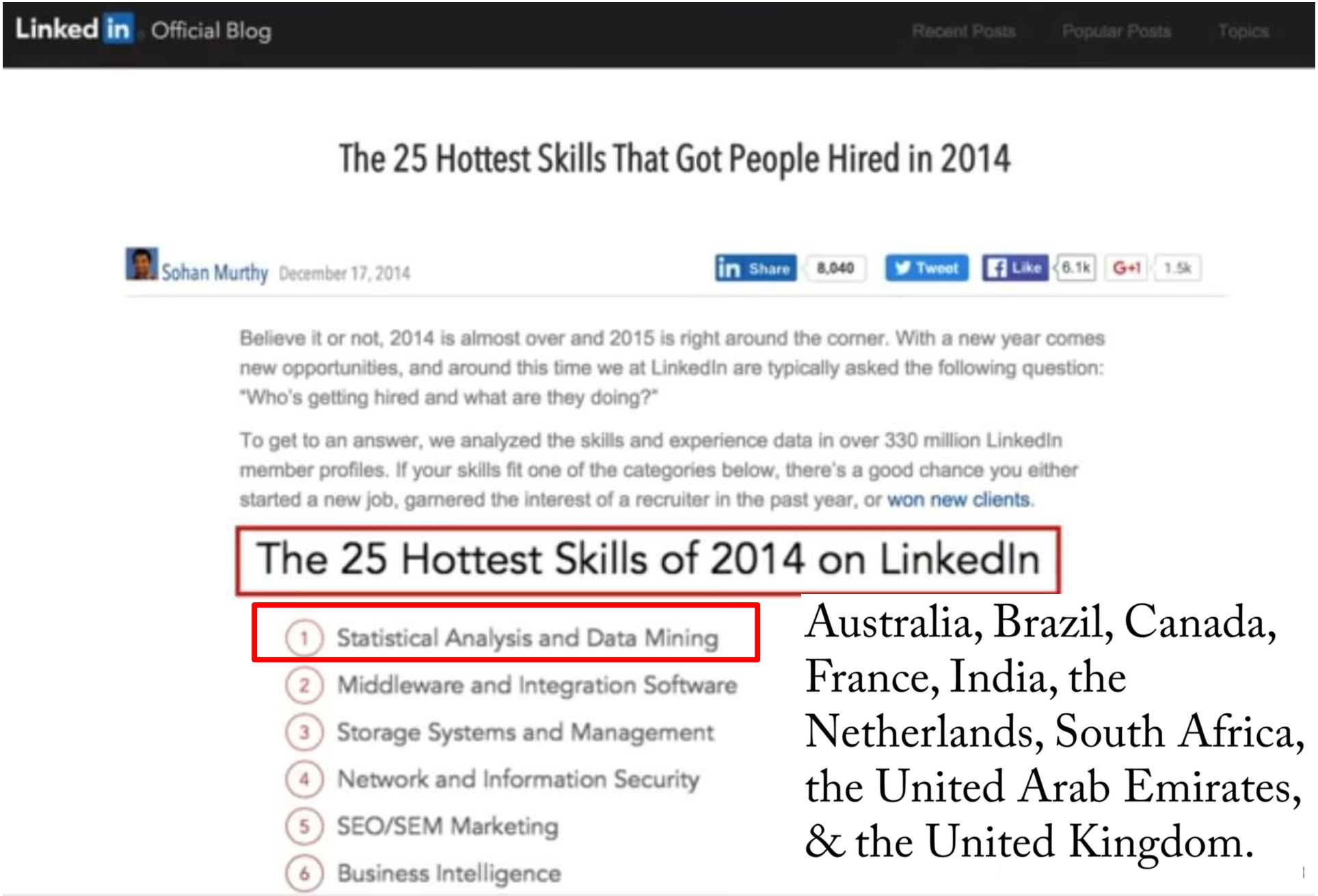
- Healthcare: Predictive models for diseases.
- Finance: Fraud detection and algorithmic trading.
- Retail: Personalized marketing and inventory optimization.

Demand

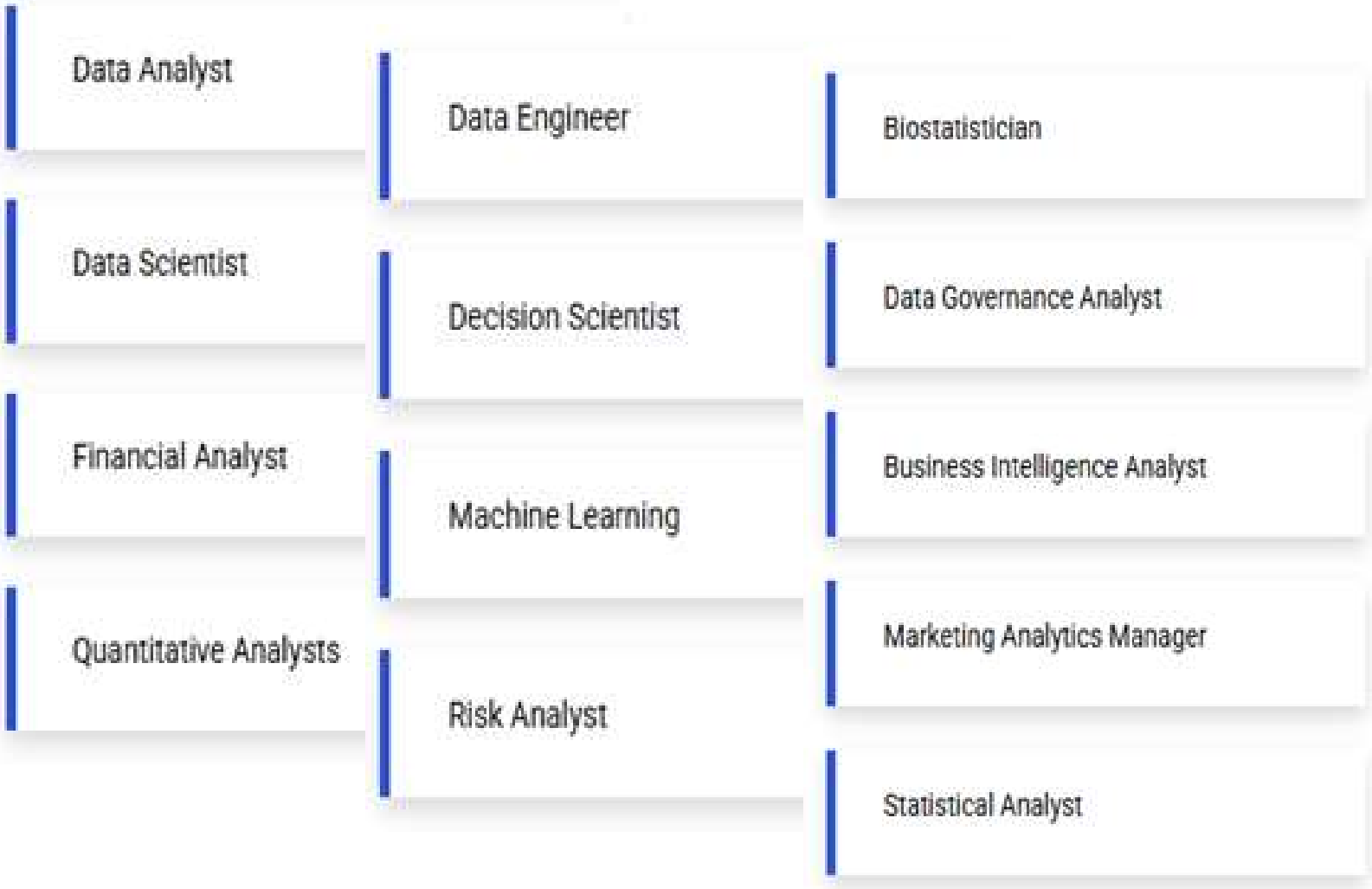

McKinsey&Company
Client Service
Insights & Publications
About Us
Alumni
Careers
Global Locations
Search
Contact us
Frequently asked questions
Site map
Terms of use
Local language information
Privacy policy
© 1996-2015 McKinsey & Company
Log in
Regist
Insights & Publications
Latest thinking
Industries
Functions
Regions
Themes
Report | McKinsey Global Institute
Big data: The next frontier for innovation, competition and productivity
May 2011 | by James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers
3 minute read
Download
Executive Summary PDF–922KB
Full Report PDF–6MB
Kindle MOBI–4MB
eBook EPUB–3MB
PDF
Print
E-mail
The amount of data in our world has been exploding, and analyzing large data sets—so-called big data—will become a key basis of competition, underpinning new waves of productivity growth, innovation, and consumer surplus, according to research by MGI and McKinsey's Business Technology Office. Leaders in every sector will have to grapple with the implications of big data, not just a few data-oriented managers. The increasing volume and detail of information captured by enterprises, the rise of multimedia, social media, and the Internet of Things will fuel exponential growth in data for the foreseeable future.
Interactive
MCKINSEY GLOBAL INSTITUTE
CELEBRATING 25 YEARS OF INSIGHT


Big data—a growing torrent
$600 to buy a disk drive that can store all of the world's music
5 billion mobile phones in use in 2010
30 billion pieces of content shared on Facebook every month
40% projected growth in global data generated per year vs. 5% growth in global IT spending
235 terabytes data collected by the US Library of Congress in April 2011
15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress
Big data—capturing its value
$300 billion potential annual value to US health care—more than double the total annual health care spending in Spain
€250 billion potential annual value to Europe's public sector administration—more than GDP of Greece
$600 billion potential annual consumer surplus from using personal location data globally
60% potential increase in retailers' operating margins possible with big data
140,000–190,000 more deep analytical talent positions, and
1.5 million more data-savvy managers needed to take full advantage of big data in the United States


60% potential increase in retailers' operating margins possible with big data
140,000–190,000
more deep analytical talent positions, and
1.5 million
more data-savvy managers needed to take full advantage of big data in the United States


Linked in Official Blog
Recent Posts
Popular Posts
Topics
The 25 Hottest Skills That Got People Hired in 2014
Sohan Murthy December 17, 2014
Share 8,040
Tweet
Like 6.1k
G+1 1.5k
Believe it or not, 2014 is almost over and 2015 is right around the corner. With a new year comes new opportunities, and around this time we at LinkedIn are typically asked the following question: "Who's getting hired and what are they doing?"
To get to an answer, we analyzed the skills and experience data in over 330 million LinkedIn member profiles. If your skills fit one of the categories below, there's a good chance you either started a new job, garnered the interest of a recruiter in the past year, or won new clients.
The 25 Hottest Skills of 2014 on LinkedIn
1 Statistical Analysis and Data Mining
2 Middleware and Integration Software
3 Storage Systems and Management
4 Network and Information Security
5 SEO/SEM Marketing
6 Business Intelligence
Australia, Brazil, Canada, France, India, the Netherlands, South Africa, the United Arab Emirates, & the United Kingdom.

Career Prospects



Opportunities in Data Science

AI and Machine Learning:

- Develop intelligent systems for automation, natural language processing, and image recognition.

Personalization:

- Enhance user experiences in e-commerce, entertainment, and education.

Healthcare Advancements:

- Precision medicine, early diagnosis, and predictive healthcare models.

Sustainability:

- Optimize resource usage and environmental monitoring with data-driven solutions.

Emerging Technologies:

- Integration with blockchain, IoT, and quantum computing for innovative applications.

Challenges in Data Science

Skill Gap:

- High demand for professionals skilled in both technical and domain expertise.

Scalability:

- Building systems that can process and analyze massive datasets efficiently.

Interpretability:

- Explaining the results of complex models (e.g., deep learning).

Bias in Models:

- Models can perpetuate biases present in the data.

Cost and Resources:

- High costs of computational resources and tools for large-scale analytics.

Real-time Analytics:

- Managing streaming data and delivering real-time insights.

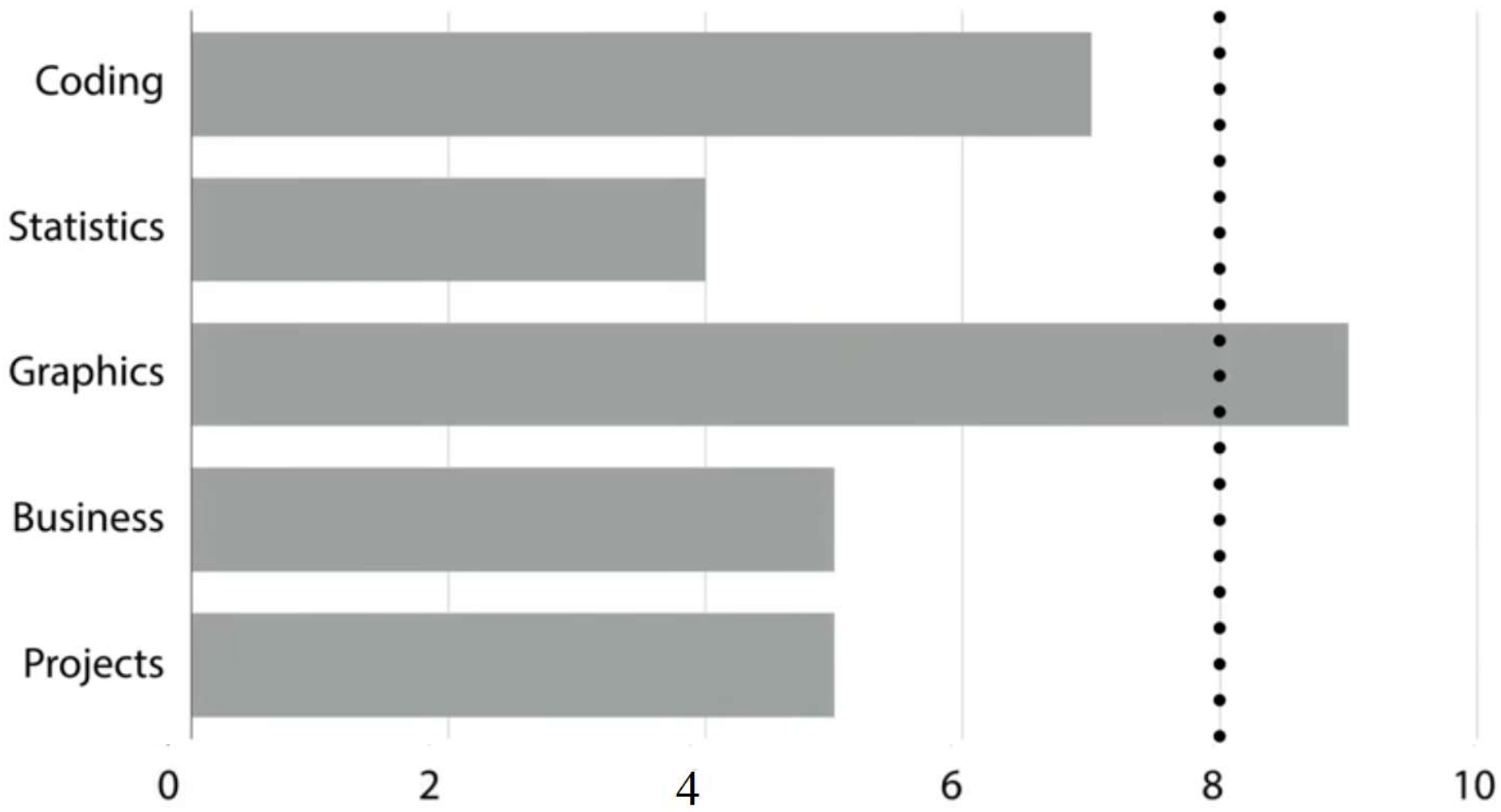
Benefit of working together

- Data science is diverse
- Volume, velocity, and variety
- Different goals and skills
- Different contexts
- Need full skill set
- Coding, statistics, and domain expertise

Example

One of you say Ali have:

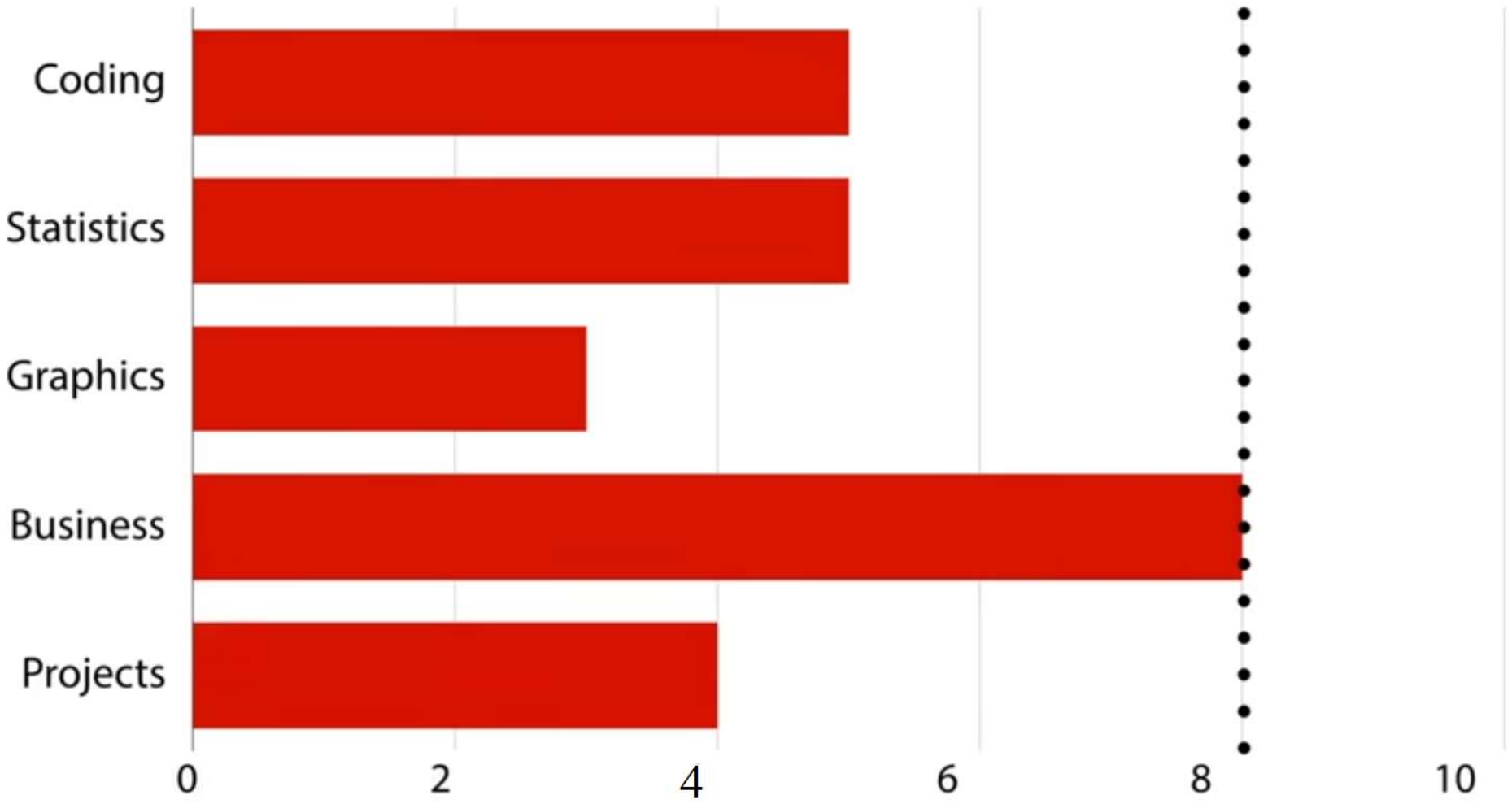
- Strong visualization
- Good coding
- Limited analytics


Coding
Statistics
Graphics
Business
Projects
0
2
4
6
8
10

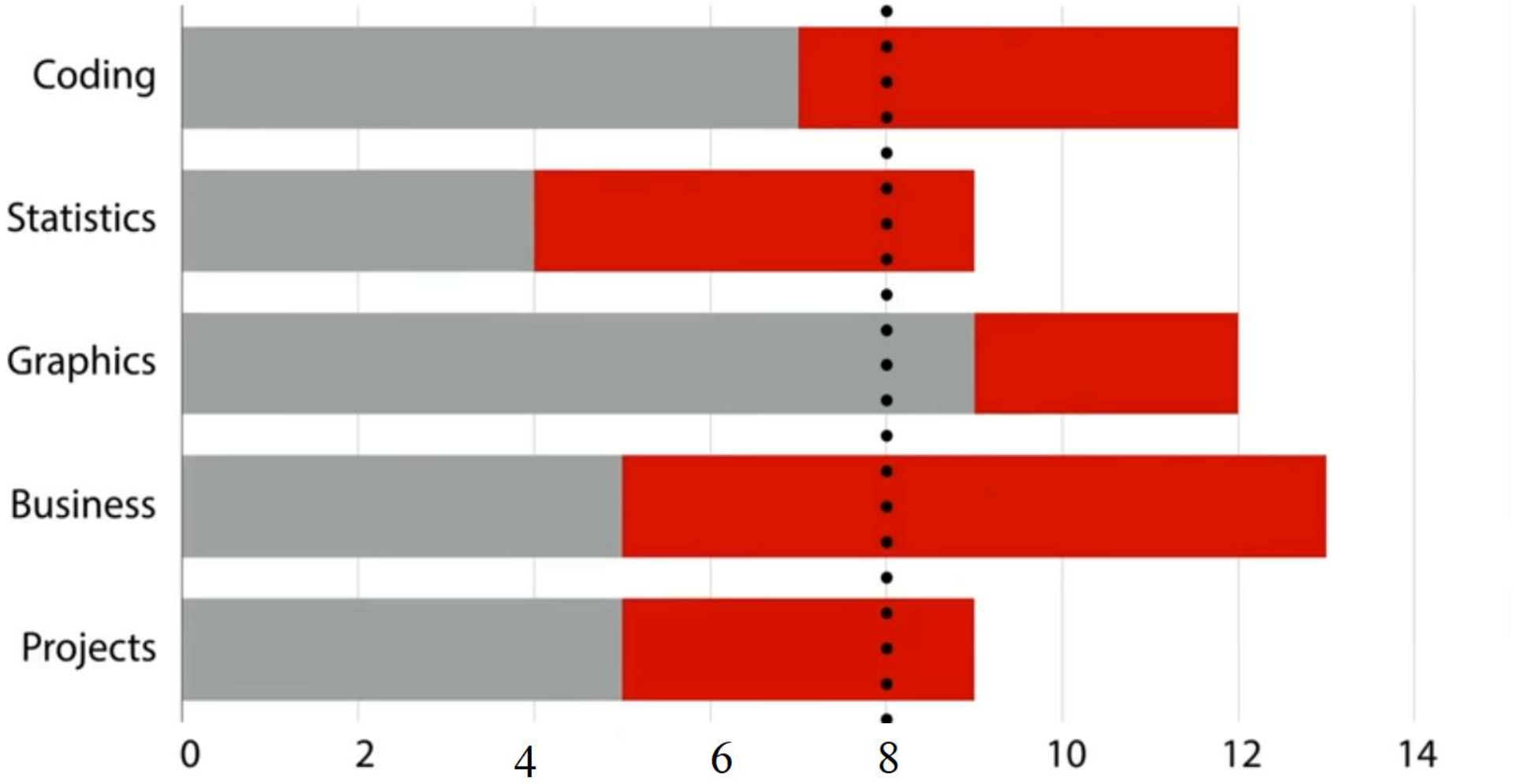
Example

Other is say Aieza Noor have:

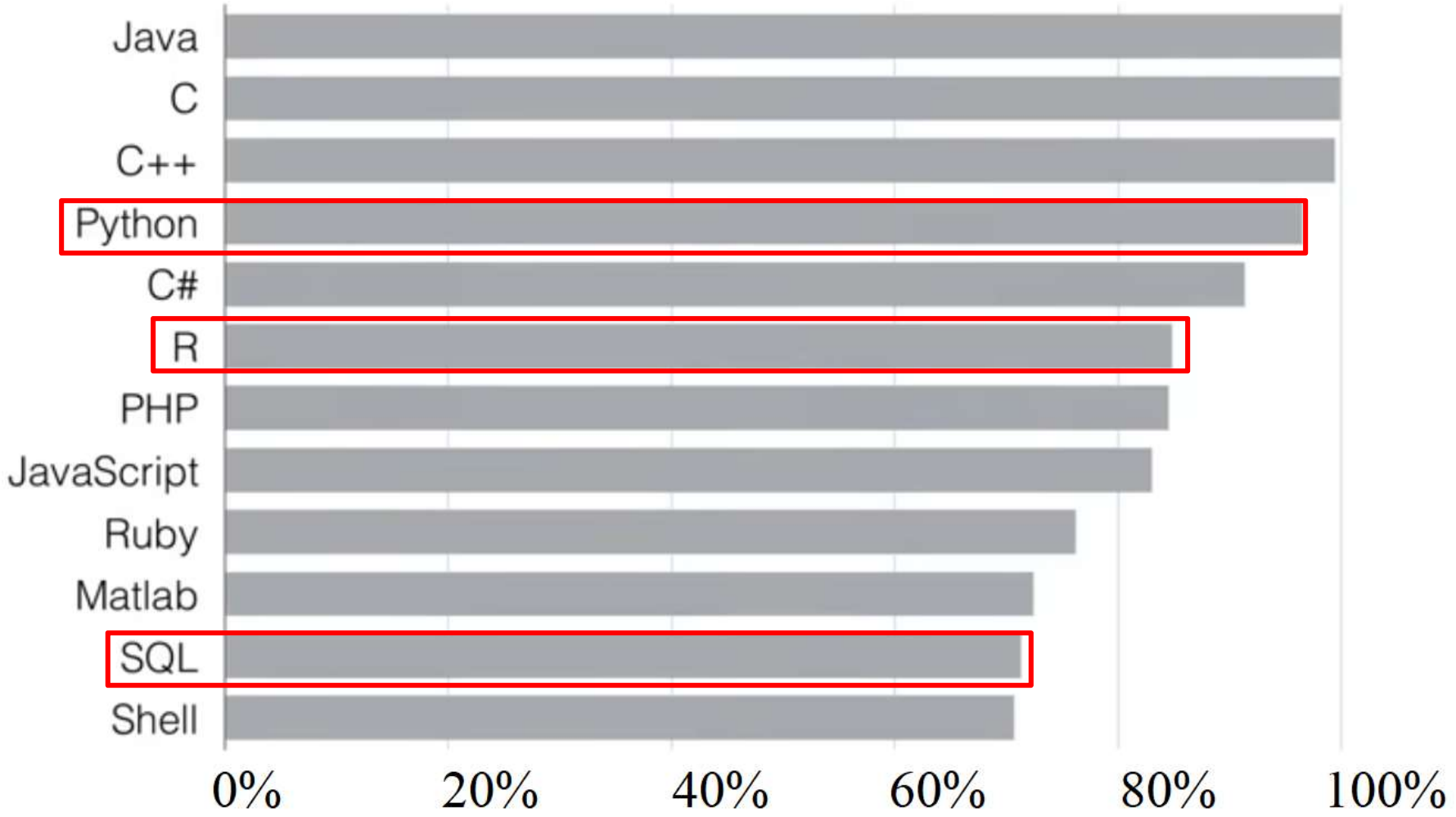
- Strong business
- Good tech skills
- Limited graphics


Coding
Statistics
Graphics
Business
Projects
0
2
4
6
8
10

Now we make a team



Tools for coding.



Tools for data science.