Course Overview:

巨量資料分析技術與應用

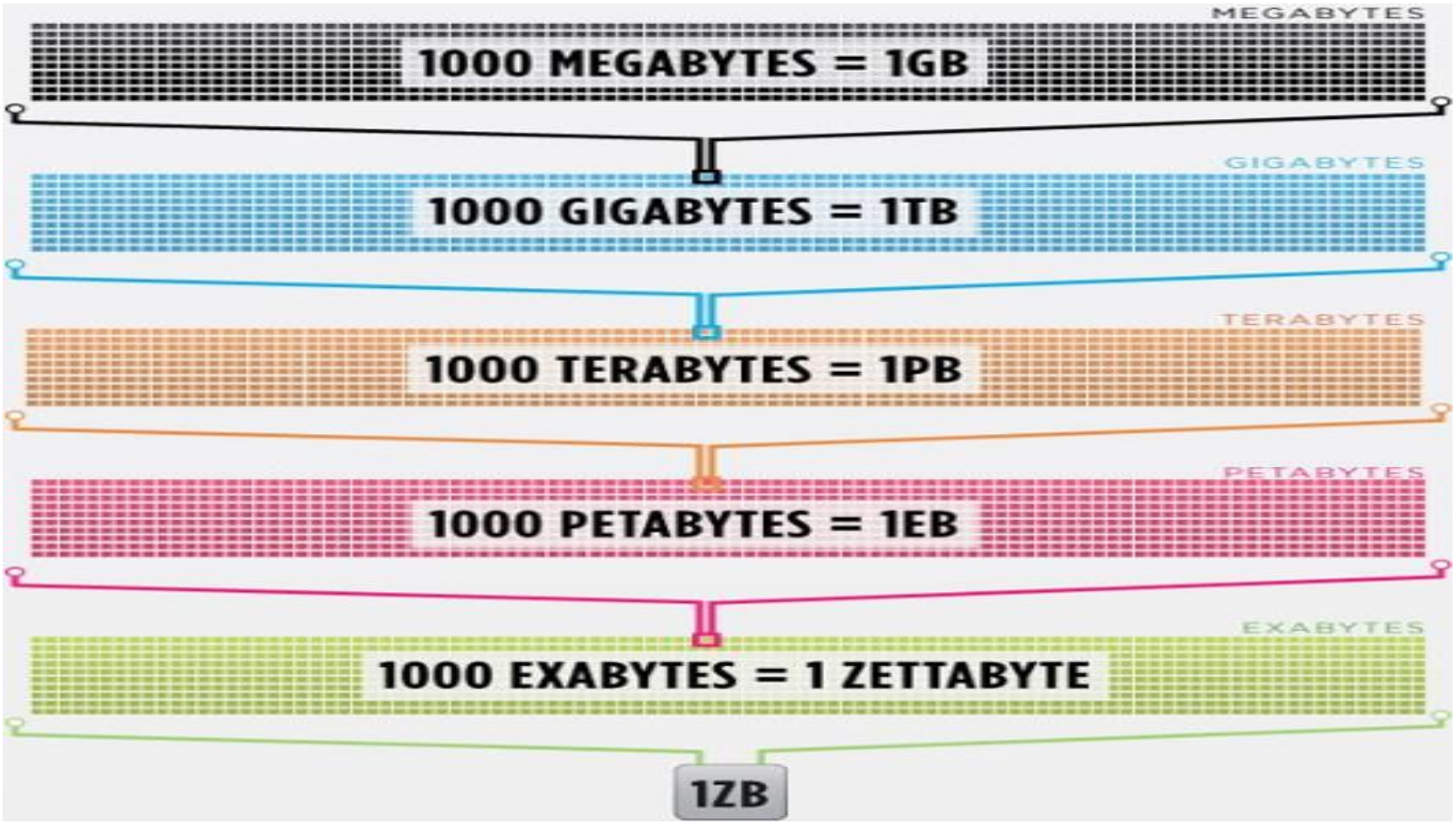
Big Data Analytics Techniques and Applications

Course Goal

- **To receive entry knowledge and skills on big data analytics**
  - Concepts
  - Relevant platforms and techniques
  - The way to apply into real applications
  - Hands-on exercises and practical project implementation
  - Linking to further advanced studies
- **Pre-requisite**
  - No strict pre-requisite courses
  - Fundamental programming capability is a MUST
  - Background on database/data mining/statistics is plus

Briefing on Big Data

The Big Data Era

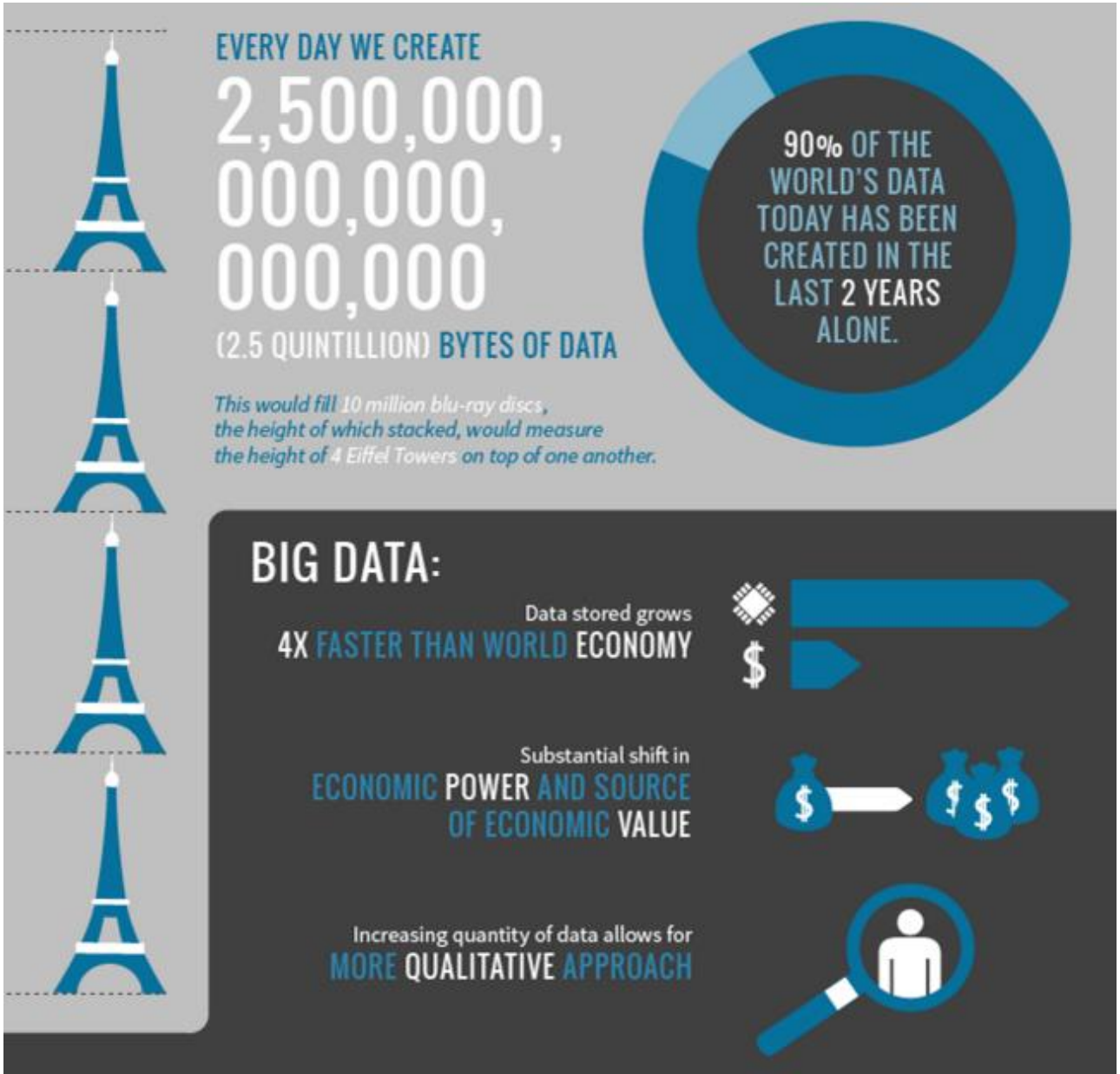


* http://www.engadget.com/2011/06/29/visualized-a-zettabyte/

The Big Data Era (cont.)

- **Data is growing explosively!**
  - **For every 30 minutes that a Boeing jet engine runs, the system creates 10 terabytes of operations information.**
  - **For every second that the Large Hadron Collider at CERN runs an experiment, the instrument can generate 40 terabytes of data.**
  - **In 2011 alone, 1.8 zettabytes (1E21 bytes) of data were created.**
    - **If all 7 billion people on Earth joined Twitter and continually tweeted for one century, they would generate one zettabyte of data (Hadhazy, 2010).**
- **IDC estimates that the amount of data available is doubling every two years.**

The Big Data Era (cont.)



What is Big Data?

"Datasets (usually unstructured, distributed, and noisy) that are beyond the ability of typical database/tools to capture, store, manage, and analyze in terms of volume, variety and velocity of coming.", *McKinsey Global Institute*

Volume

Massive scale and growth of unstructured data

- 80%~90% of total data
- Growing 10x~50x faster than structured (relational) data
- 10x~100x of traditional data warehousing

Variety

Heterogeneity and variable nature of Big Data

- Many different forms (text, document, image, video, ...)
- No schema or weak schema
- Inconsistent syntax and semantics

Velocity

Realtime rather than batch-style analysis

- Data streamed in, tortured, and discarded
- Making impact on the spot rather than after-the-fact

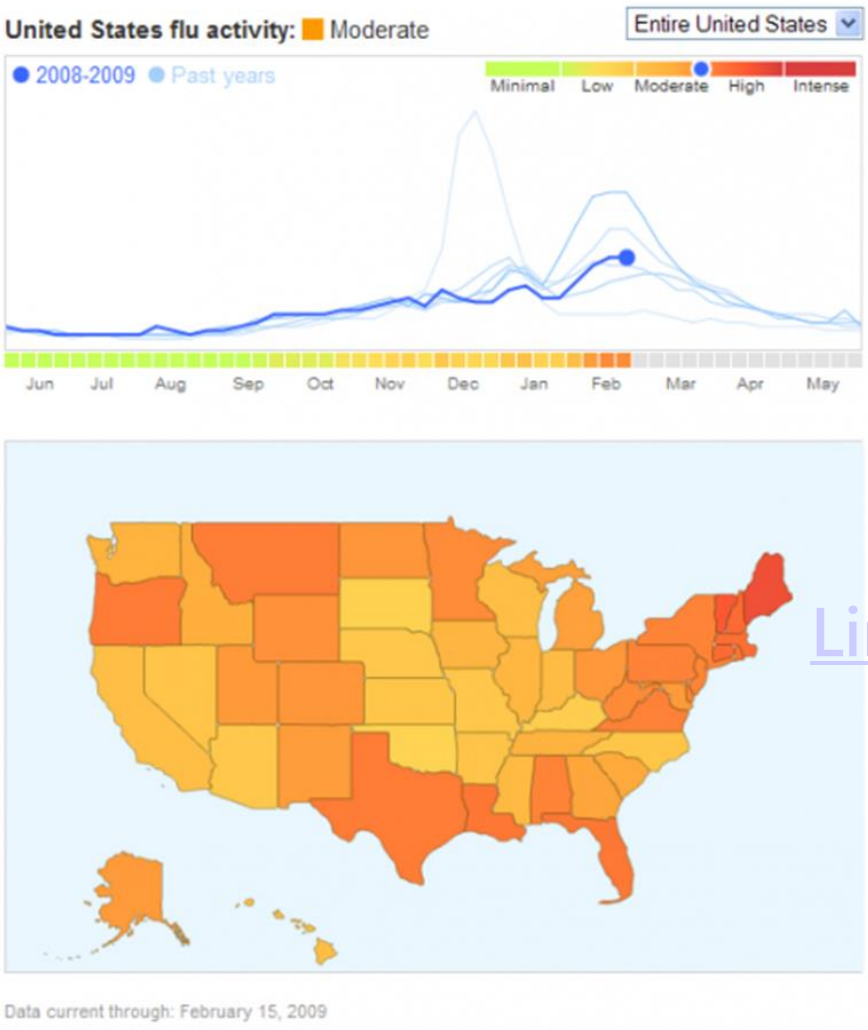
Value

Predictive analytics for future trends and patterns

- Deep, complex analysis (machine learning, statistic modeling, graph algorithms, …), versus
- Traditional business intelligence (querying, reporting, …)

Big Data Applications

Google Flu Trends



J. Ginsberg, *et al.*,

Detecting influenza epidemics using search engine query data,

***Nature*, February 2009**

Link:- www.google.com/flutrends

IBM Watson in Jeopardy!

- IBM 於2011 年2 月，將最新研發成果華生電腦（Watson）推上全美知名益智節目「Jeopardy!」挑戰並擊敗兩位歷史紀錄保持人
- Watson具備高速巨量資料運算力、記憶力、反應力及語言能力等，牽涉龐雜的語言處理、邏輯解答的人工智慧及語音合成技術
- 目前已開始被應用在

醫療照護、媒體分析、智慧家庭等服務中

Application in Movie Industry

- 電影【復仇者聯盟】：成本兩億美金
- 如何知道觀眾之興趣反應?
- 如何訂定最佳之行銷策略?

Application in Movie Industry (cont.)

- 利用Big Data Analytics 監測分析社交媒體對電影預告片之反應:
  - 11億條 Tweets/min
  - 570萬篇Blogs/min
  - 350萬條 Messages/min
  - 擷取關鍵訊息, 分析主題, 判斷網友意向 → 歸結出網友對電影預告片之看法與評價
  - 電影公司針對分析結果進行行銷策略之調整
- 【復仇者聯盟】票房:
  - 2012年5月上片後, 美國本土首周票房達兩億美金(成本)，寫下全美影史最高首周票房紀錄
  - 2012年總票房達15億美金, 成為世界電影史票房排名第三名, 僅次於"阿凡達"、"鐵達尼號"

Covered Topics

- Overview of Big Data
- Overview of Big Data Techniques, Applications and Challenges
- Big Data Technical Platform
- Big Data Analytics Workflow and Techniques
- Statistical Methods
- Machine Learning Methods
- Big Data Analytics Algorithms
  - Association/Sequential-Patterns mining in big data
  - Clustering analysis in big data
  - Classification modeling in big data
  - Stream data analysis in big data
  - Big Graph and Network Analysis
- Big Data Visual Analytics
- Applications Study
- Future Trends and Advanced Topics

Course Schedule

週次	課程進度、內容、主題
1	課程大綱與巨量資料簡介
2	巨量資料技術、應用與挑戰
3	資料探勘技術 - I
4	資料探勘技術 - II
5	巨量資料分析工具與平台 - I
6	巨量資料分析工具與平台 - II
7	巨量資料分析演算法 - I
8	巨量資料分析演算法 - II
9	巨量資料分析演算法 - III
10	專題提案報告
11	機器學習演算法 - I
12	機器學習演算法 - II
13	基本統計方法
14	巨量資料圖型與網路分析
15	巨量資料之資料視覺化
16	應用實例介紹
17	專題成果報告 - I
18	專題成果報告 - II

Course Design

- Lectures
  - Lectures with slides by the instructor
  - Reference books and websites
  - Invited guest talks
- Lab Assignments
  - 4 homework assignments (individually done)
- Term Project
  - Around 4 persons as a team
  - Implementation, Report and Presentation/Demo
- Participatory Discussions

Grading

- **Lab: 60% (3-4 assignments)**
- **Term Project: 30%**
- **Participation: 10%**