DSP505: Programming Lab for Data Science and Artificial Intelligence

TPL616: Advanced Programming for DSAI

(Introduction)

Vishwesh Jatala

Assistant Professor

Department of CSE

Indian Institute of Technology Bhilai

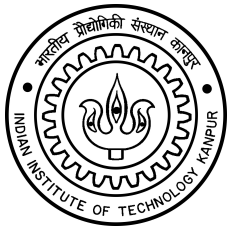vishwesh@iitbhilai.ac.in

My Background

Assistant Professor
[Aug 2020- Present]

Postdoctoral Fellow
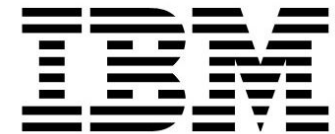[May 2018- June 2020]

Ph.D, IIT Kanpur
[2011- 18]

B.Tech, VNIT Nagpur [2005-09]

Member of Technical Staff [2009-11]

Research Intern [2013-13]

Graphics Processing Units (GPUs), High-Performance Computing, Graph Neural Networks
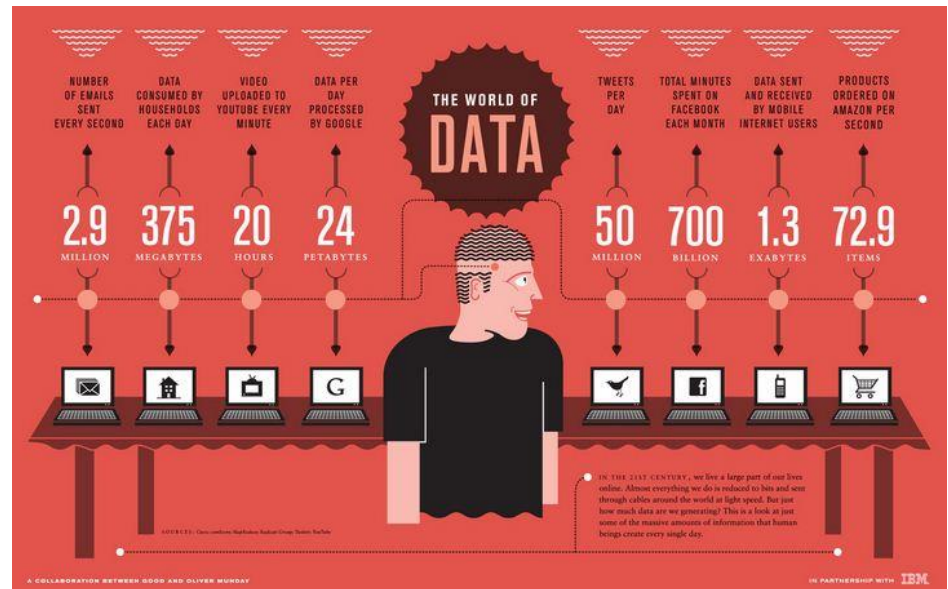
This Lecture

- Why?
- What?
- How?

Data All Around

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Financial transactions, bank/credit tran
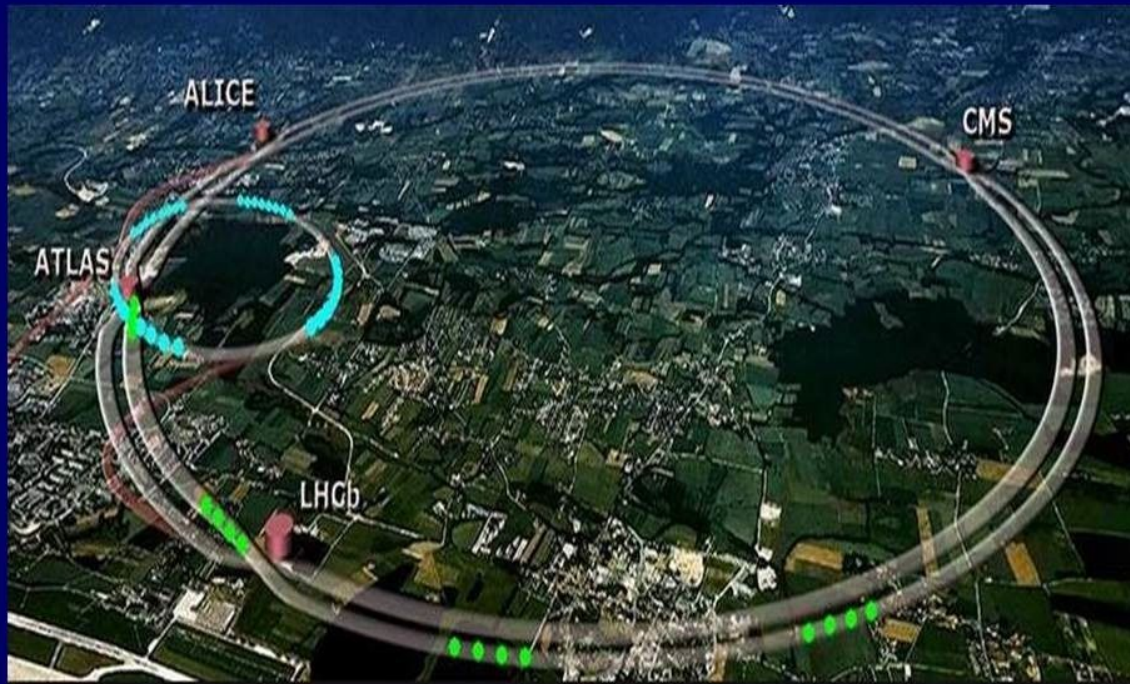 - Online trading and purchasing
  - Social Network

How much data we have?

- Google processes 20 PB a day (2008)
- Facebook has 60 TB of daily logs
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)

CERN

Aerial View of the Large Hadron Collider in Geneva

Type of Data

- Relational Data (Tables/Transaction/Legacy Data)

- Text Data (Web)

- Semi-structured Data (XML)

- Graph Data

- Social Network, Semantic Web (RDF), …

- Streaming Data

- ….

What to do with this data?

- **Aggregation and Statistics**
  - ❑ Data warehousing
- **Indexing, Searching, and Querying**
  - ❑ Keyword based search
  - ❑ Pattern matching (XML/RDF)
- **Knowledge discovery**
  - ❑ Data Mining
  - ❑ Statistical Modeling

What is Data Science?

- An area that manages, manipulates, extracts, and interprets knowledge from tremendous amount of data

- Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data

- Data science principles apply to all data – big and small

What is Data Science?

- Theories and techniques from many fields to analyze a large amount of data to help decision makers in many industries such as science, engineering, economics, politics, finance, and education
  - Computer Science
    - Pattern recognition, visualization, data warehousing, High performance computing, Databases, AI
  - Mathematics
    - Mathematical Modeling
  - Statistics
    - Statistical and Stochastic modeling, Probability.
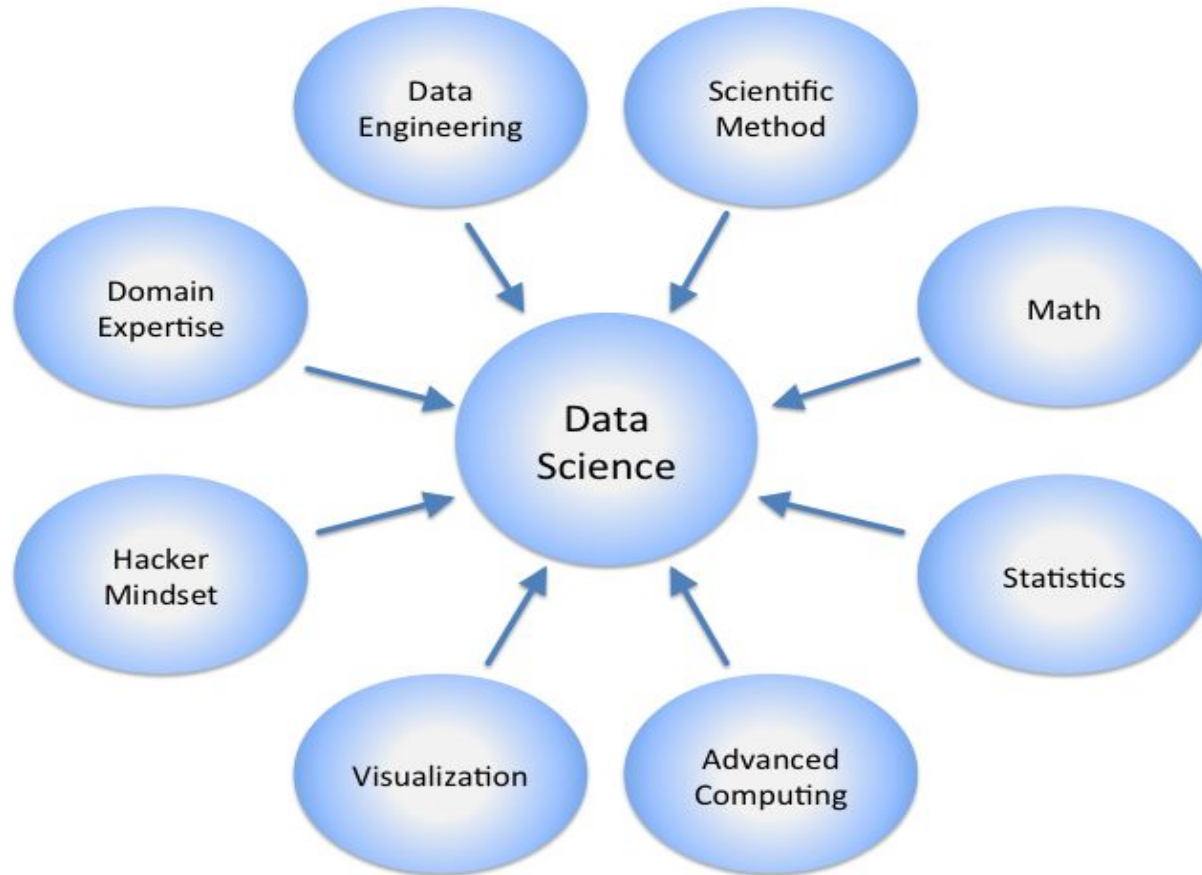
Data Science

Data Science Is Multidisciplinary
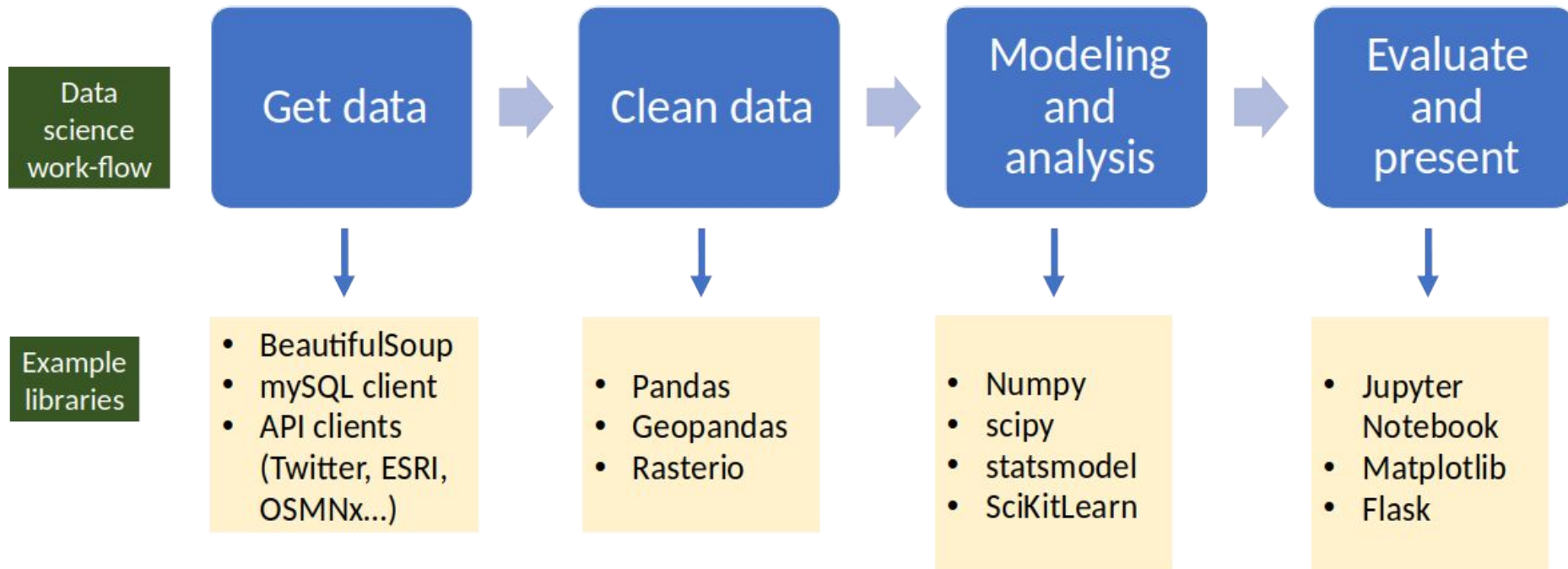
By Brendan Tierney, 2012
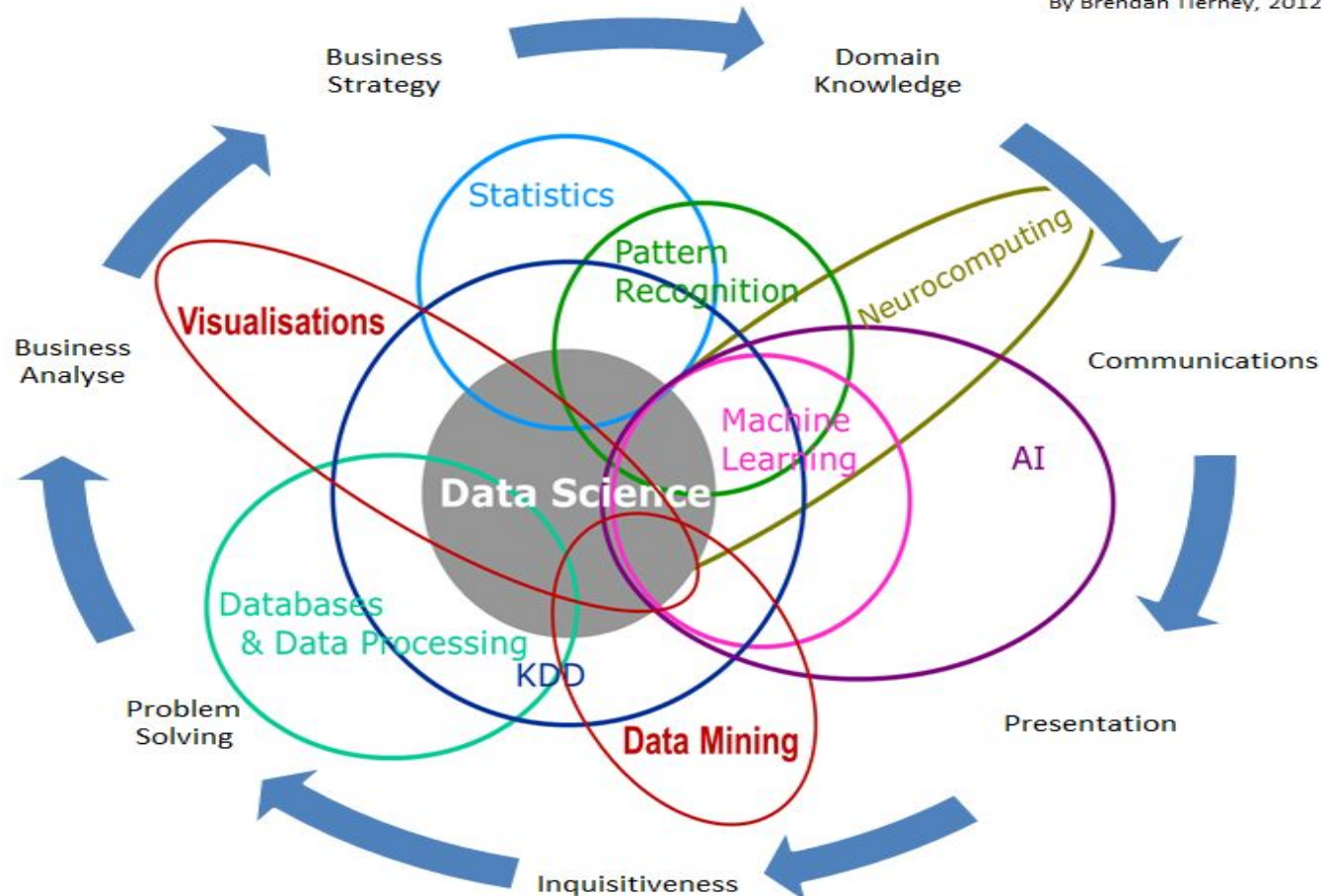
Data Science

Real Life Examples

- Companies learn your secrets, shopping patterns, and prefernces
    - Next recommended video in youtube
    - Generation of custom news

Data Science Tools

	Data science work-flow			
	Get data →	**Clean data** →	**Modeling and analysis** →	**Evaluate and present**
Example libraries	• BeautifulSoup • mySQL client • API clients (Twitter, ESRI, OSMNx...)	• Pandas • Geopandas • Rasterio	• Numpy • scipy • statsmodel • SciKitLearn	• Jupyter Notebook • Matplotlib • Flask

What?

- Popular Data Science Tools and Techniques
  - ❑ Basic Python Programming
  - ❑ Libraries: Pandas, Numpy
  - ❑ Data visualization Tools
  - ❑ Graph Representations
  - ❑ Dimensionality reduction
  - ❑ Basic ML
  - ❑ Data storage and tools
  - ❑ Data Scraping
  - ❑ Big Data handling
- Hands on experience

How?

Mode of Teaching

- Theory Classes
- Labs: Practice problems

Labs: Practice Sessions

- Practice problems will be provided to solve
- Take it as an opportunity to clear your misconceptions/doubts
- Try to submit your solutions during the lab hours.

Course Logistics

- Lecture schedule: Uploaded

- Course Website: Canvas platform
    - Lecture notes
    - Lab practice problems
    - Any evaluations as per the instructions

Course Logistics

- Evaluation scheme (Tentative):
  - Programming assignments (3-4): ~20%
  - Lab exams: ~40%
  - Project: ~30%
  - Attendance: ~10%

- Attendance
  - 0% - 50%: 0 Marks
  - >50%: Marks will be awarded out of 10 accordingly.
  - Example:
    - Total sessions: 16
    - #sessions attended = 7 (<50%), marks = 0
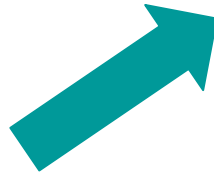   - #sessions attended = 10 (62.5%), marks = 2.5 (2*10/8)

Course Logistics

- Projects:
    - Will be done in a team (logistics will be conveyed)

- Policy:
    - Penalty for late submission: 20% for each day
    - Acknowledge all the sources
    - **Severe penalty for cheating**

References

- Will be uploaded on Canvas
  - ❑ Lecture notes
  - ❑ Reference material

Course Logistics

Office:

ED1, 401A

vishwesh@iitbhilai.ac.in

CANVAS

Thank you