



Data Science Course

Introduction

IFT6758, Fall 2020



Introduction: the era of big data



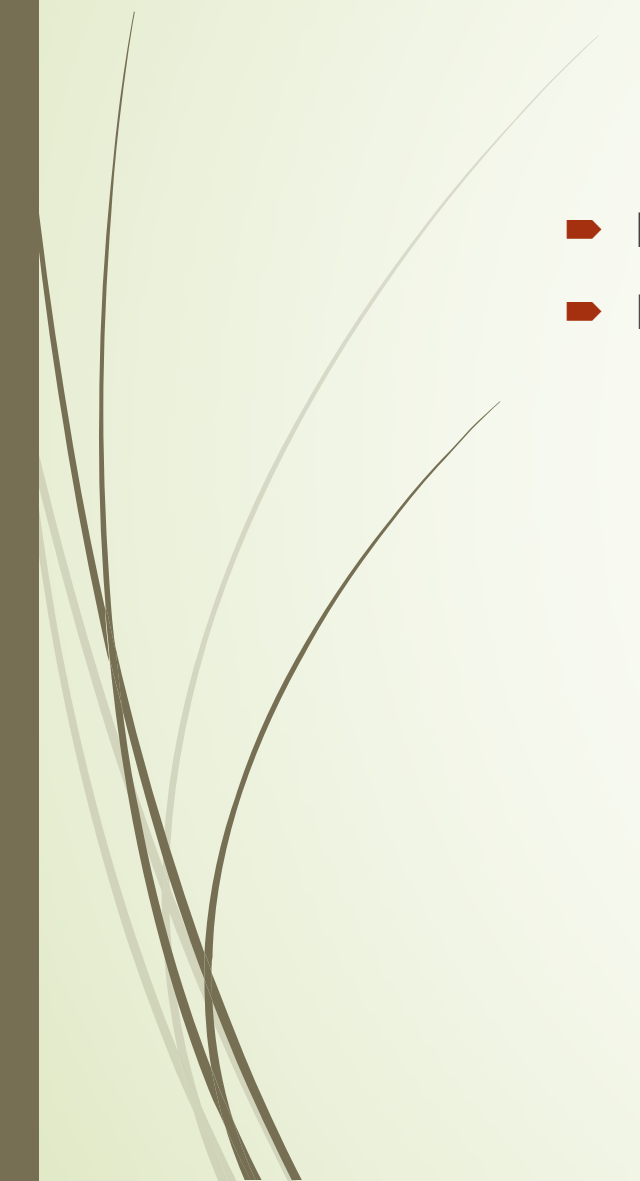


Introduction: the era of big data

- **Billions of sensors** have been collecting data for decades
- 

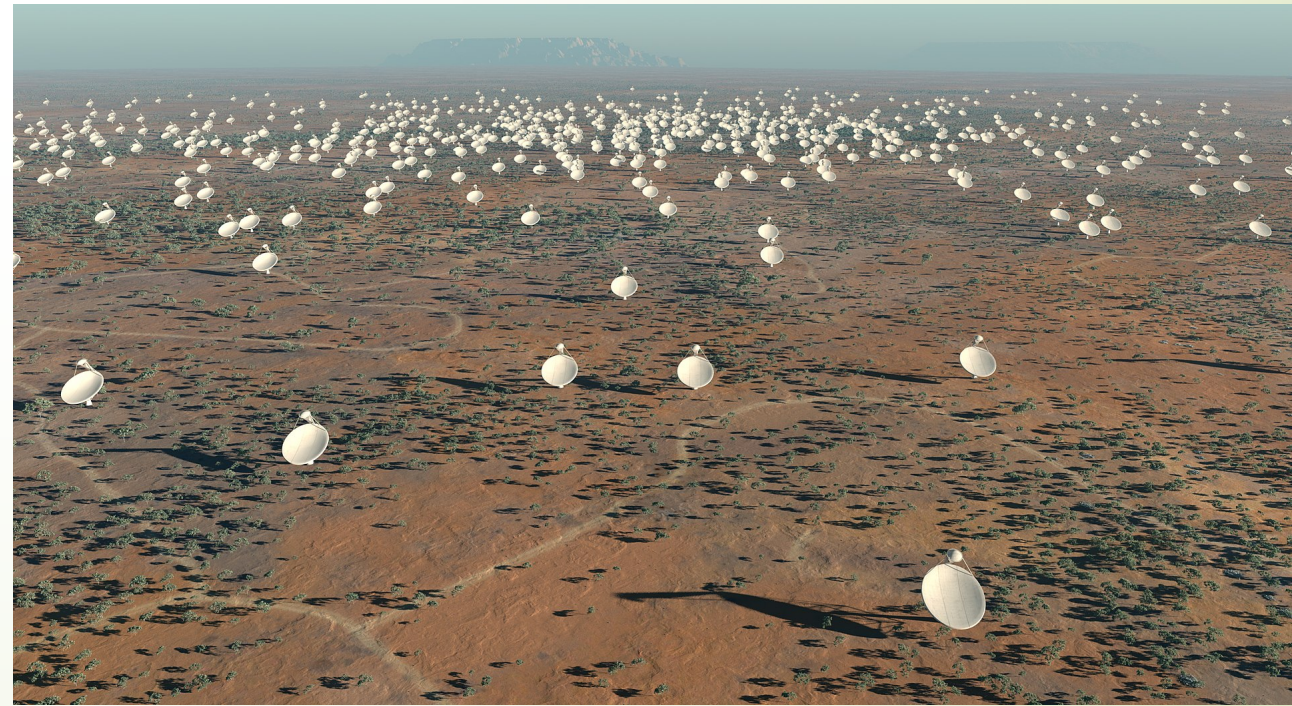


Introduction: the era of big data

- **Billions of sensors** have been collecting data for decades
 - Data is routinely collected and employed for
- 

Introduction: the era of big data

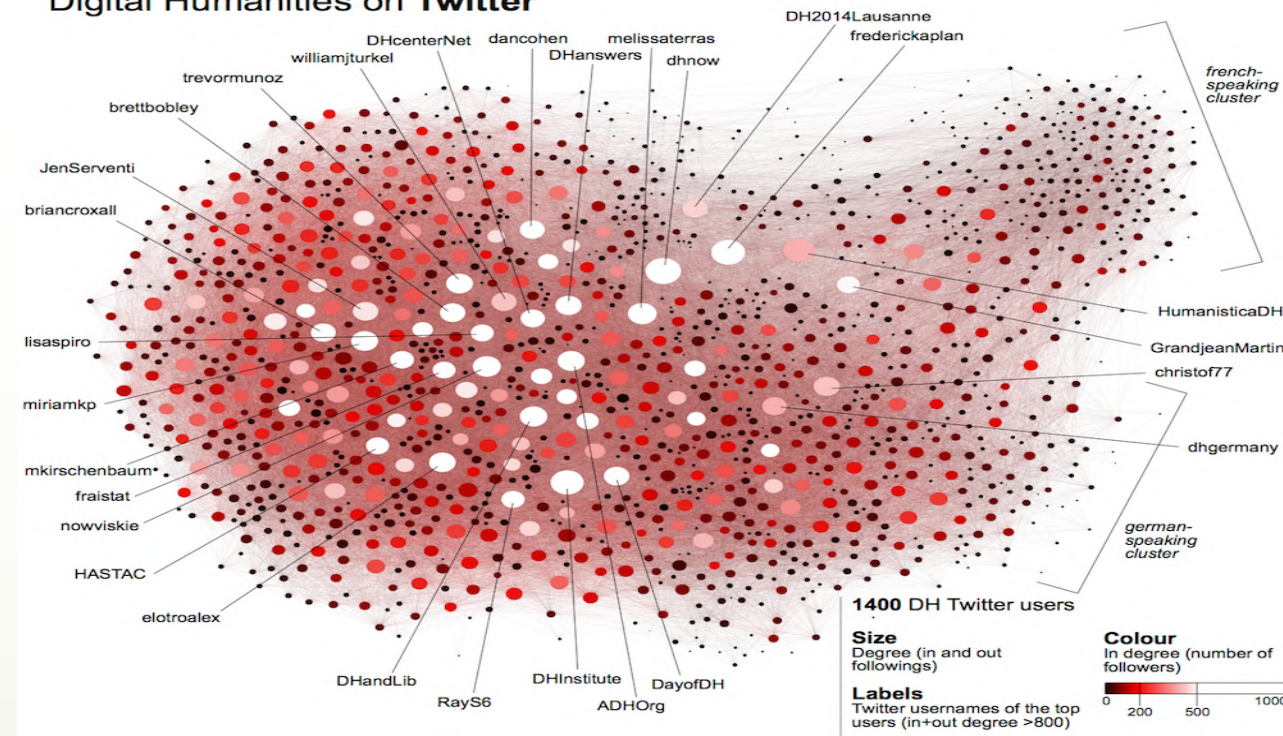
- **Billions of sensors** have been collecting data for decades
- Data is routinely collected and employed for
 - Scientific discovery



Introduction: the era of big data

- **Billions of sensors** have been collecting data for decades
- Data is routinely collected and employed for
 - Scientific discovery
 - Commercial purposes

Digital Humanities on Twitter



Introduction: the era of big data

- **Billions of sensors** have been collecting data for decades
- Data is routinely collected and employed for
 - Scientific discovery
 - Commercial purposes
 - Data driven society





Introduction: the era of big data

- **Billions of sensors** have been collecting data for decades
- Data is routinely collected and employed for
 - Scientific discovery
 - Commercial purposes
 - Data driven society
- Huge resource of public datasets



Introduction: the era of big data

- **Billions of sensors** have been collecting data for decades
- Data is routinely collected and employed for
 - Scientific discovery
 - Commercial purposes
 - Data driven society
- Huge resource of public datasets
 - <https://github.com/awesomedata/awesome-public-datasets>



Introduction: the era of big data

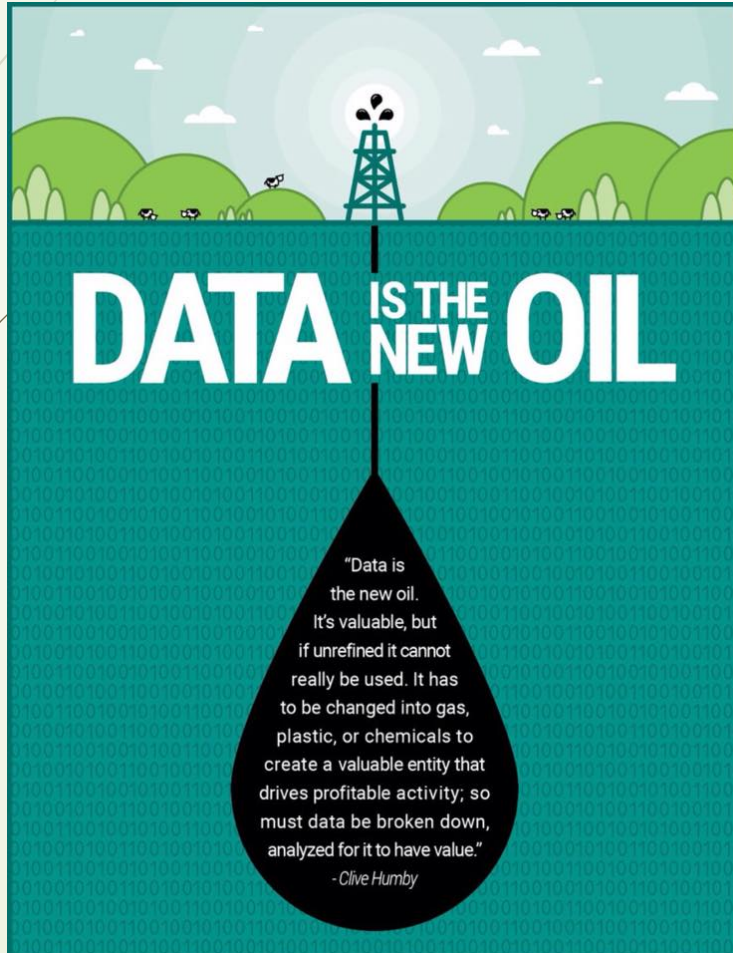
- **Billions of sensors** have been collecting data for decades
 - Data is routinely collected and employed for
 - Scientific discovery
 - Commercial purposes
 - Data driven society
 - Huge resource of public datasets
 - <https://github.com/awesomedata/awesome-public-datasets>
 - It's important that data scientists work **responsibly** and for **greater good**
- 



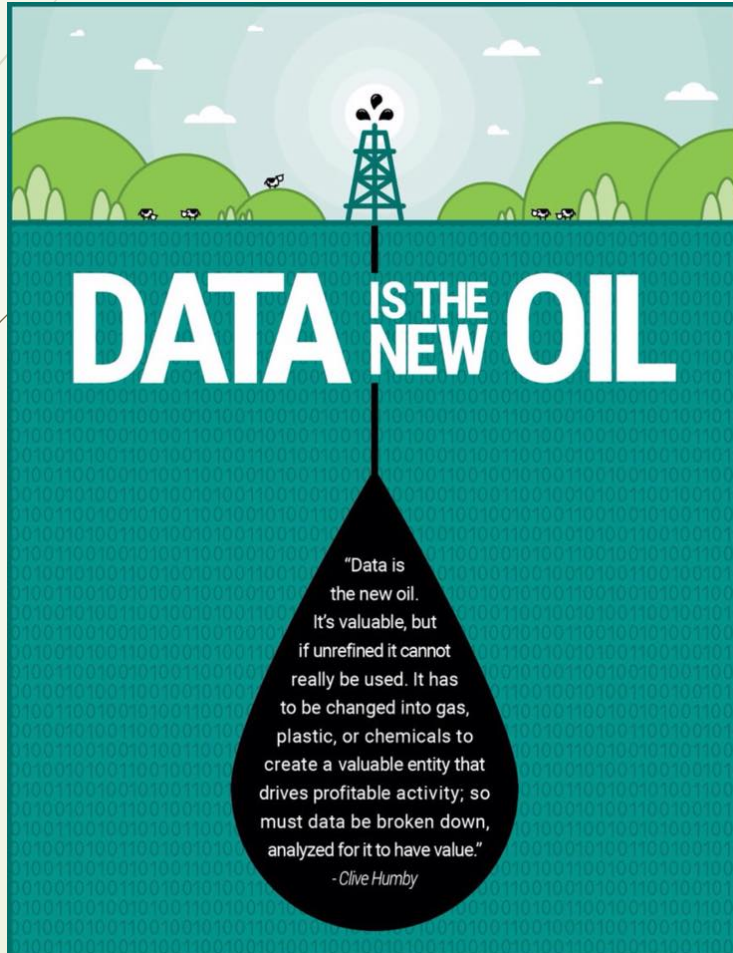
Data is the new oil



Data is the new oil

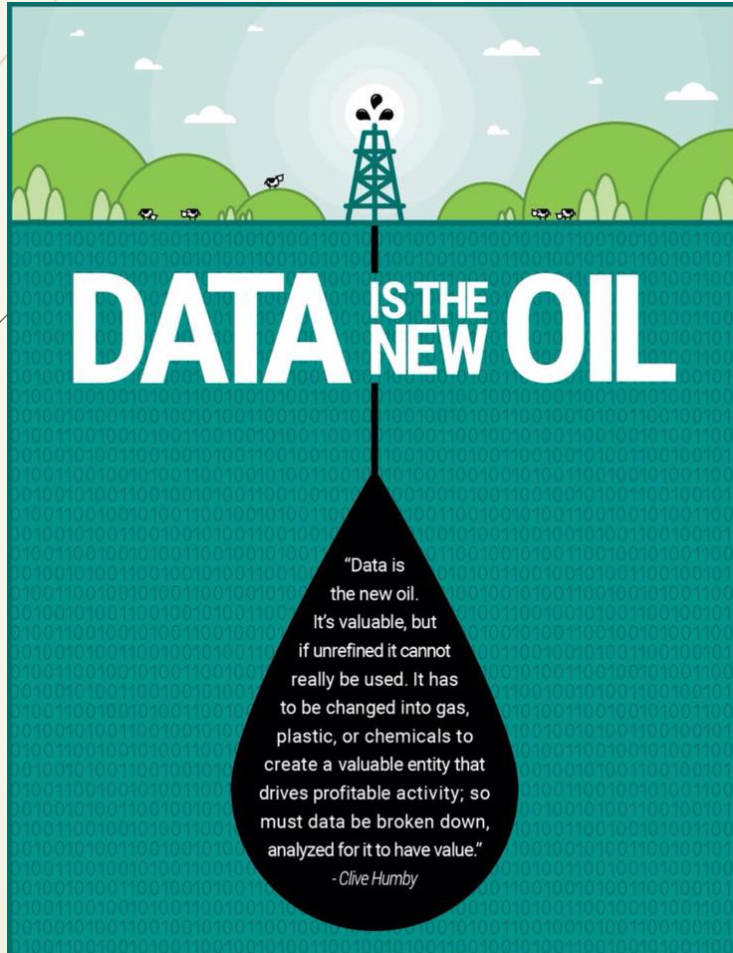


Data is the new oil



Or, is it?

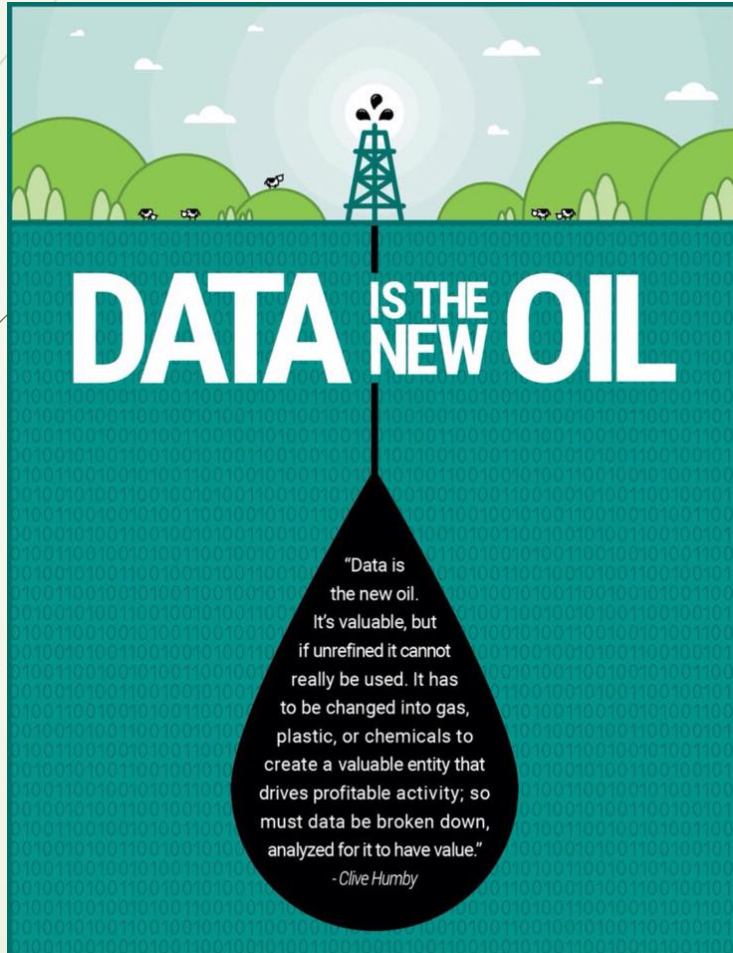
Data is the new oil



Or, is it?

Unlike oil, when dealing with data, it is far from clear how exactly to turn that data into profits.

Data is the new oil



Or, is it?

Unlike oil, when dealing with data, it is far from clear how exactly to turn that data into profits.

Why data is not new oil



Know the data





Know the data

- **Be comfortable with data**
- 



Know the data

- **Be comfortable with data**
 - Build competence working with multimodal data sets
 - Exposure to the full data science workflow



Know the data

- **Be comfortable with data**
 - Build competence working with multimodal data sets
 - Exposure to the full data science workflow
- **Become a data detective**



Know the data

- **Be comfortable with data**

- Build competence working with multimodal data sets
- Exposure to the full data science workflow

- **Become a data detective**

- Learn to ask good questions
- Reason about uncertainty, think critically



Know the data

- **Be comfortable with data**
 - Build competence working with multimodal data sets
 - Exposure to the full data science workflow
- **Become a data detective**
 - Learn to ask good questions
 - Reason about uncertainty, think critically
- **Learn responsible data science**



Know the data

- **Be comfortable with data**
 - Build competence working with multimodal data sets
 - Exposure to the full data science workflow
- **Become a data detective**
 - Learn to ask good questions
 - Reason about uncertainty, think critically
- **Learn responsible data science**
 - Understand risks at all stages of data science workflow



Takeaway from this course





Takeaway from this course

Prepare you to be responsible and competent data scientists



Takeaway from this course

Prepare you to be responsible and competent data scientists

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

COMMUNICATION & VISUALIZATION

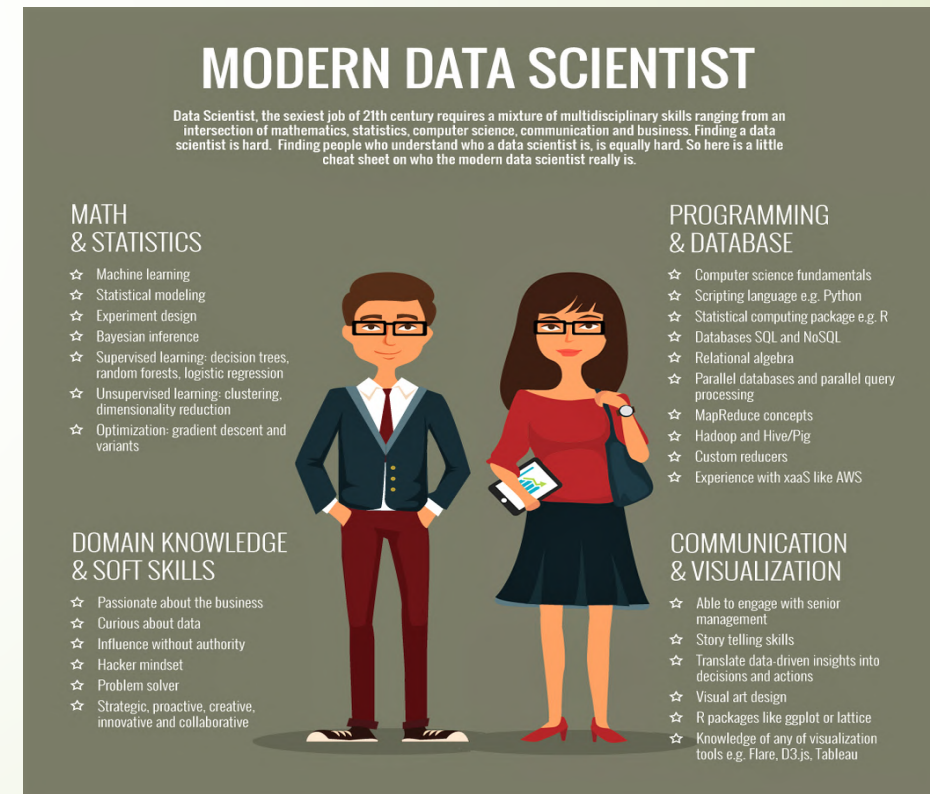
- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau




Takeaway from this course


Prepare you to be responsible and competent data scientists


- Learn how to...
 - Apply data science in your own field
 - Work in industry or research
 - Understand data's role in society





Course outline (subject to minor changes)






Course outline (subject to minor changes)

- **Part 1: Summaries and Inferences**

- Data visualization and transformation
- Supervised and unsupervised learning from data
- Inference and model comparison




Course outline (subject to minor changes)

- **Part 1: Summaries and Inferences**

- Data visualization and transformation
- Supervised and unsupervised learning from data
- Inference and model comparison

- **Part 2: Nontabular data**

- Text and Image
- Graph mining
- Special topics: Reinforcement Learning and Deep Learning



Course outline (subject to minor changes)

- **Part 1: Summaries and Inferences**

- Data visualization and transformation
- Supervised and unsupervised learning from data
- Inference and model comparison

- **Part 2: Nontabular data**

- Text and Image
- Graph mining
- Special topics: Reinforcement Learning and Deep Learning

- **Part 3: Frontiers**

- Advanced inference
- Ensembling
- Privacy and explainability



Logistics



- Course **website**: https://jhelum-ch.github.io/DataScience_IFT6758/
- Fill Survey for access to discussion forum
- Grading: **25%** Final project, **30%** HW, **30%** Midterm, **15%** Kaggle competition
- Instructor & TA Office Hours: **TBA**

- Use **StudiUM Forum** to reach out to the instructor
- In case of emergency:
 - Instructor: Jhelum – chakravj@mila.quebec
 - Head TA: Pravish – pravishsainath@gmail.com
 - TAs: Yutong Yan, Alexander Peplowski, Harmanpreet Singh, Akshay Singh Rana



Resources



- Books: [Statistical Learning](#), [Python Handbook](#), [Introduction to Data Science](#)
- Online courses: [freeCodeCamp](#) , [Harvard CS109](#)
- Harvard's [Data Science Review](#)