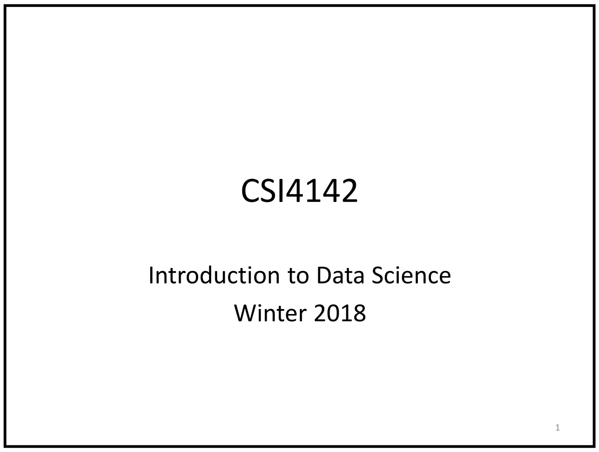
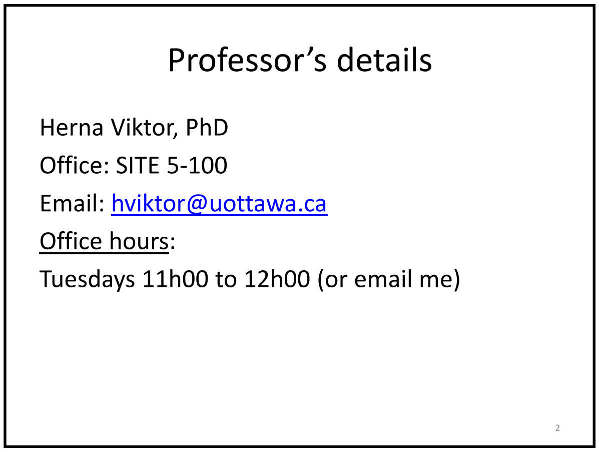


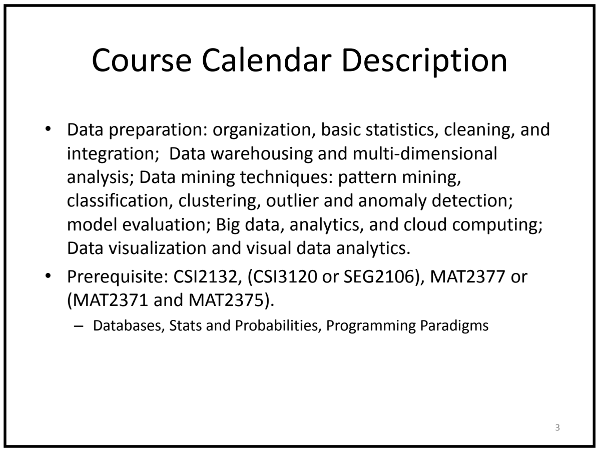
CSI4142

Introduction to Data Science
Winter 2018



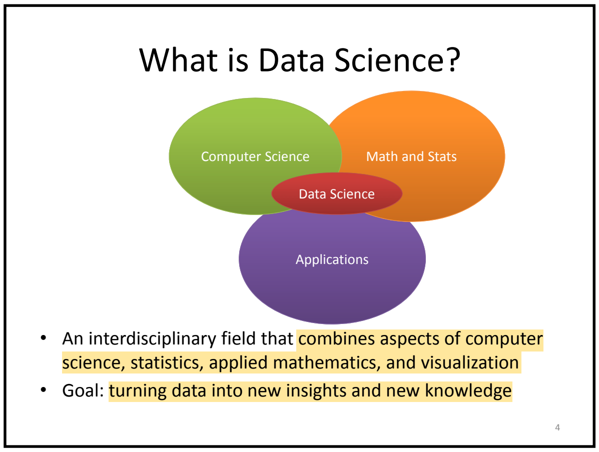
Professor’s details

Herna Viktor, PhD
Office: SITE 5-100
Email: hviktor@uottawa.ca
Office hours:
Tuesdays 11h00 to 12h00 (or email me)



Course Calendar Description

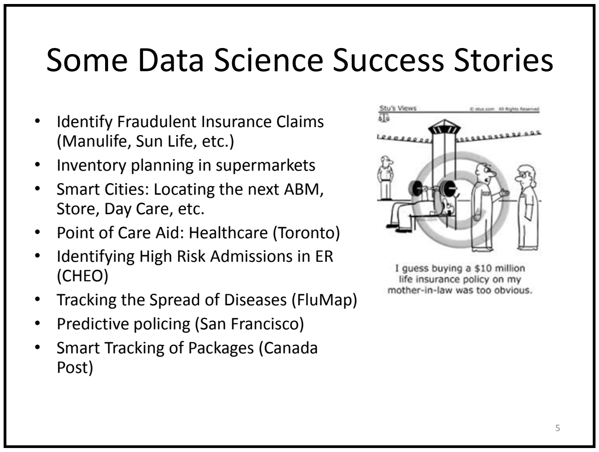
- Data preparation: organization, basic statistics, cleaning, and integration; Data warehousing and multi-dimensional analysis; Data mining techniques: pattern mining, classification, clustering, outlier and anomaly detection; model evaluation; Big data, analytics, and cloud computing; Data visualization and visual data analytics.
- Prerequisite: CSI2132, (CSI3120 or SEG2106), MAT2377 or (MAT2371 and MAT2375).
  - Databases, Stats and Probabilities, Programming Paradigms



What is Data Science?

Computer Science
Math and Stats
Data Science
Applications

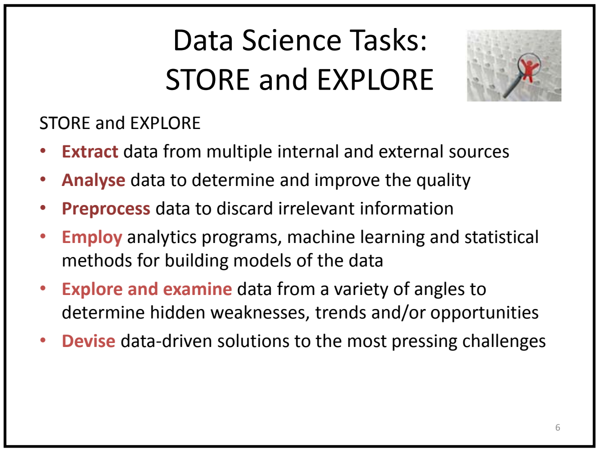
- An interdisciplinary field that combines aspects of computer science, statistics, applied mathematics, and visualization
- Goal: turning data into new insights and new knowledge



Some Data Science Success Stories

- Identify Fraudulent Insurance Claims (Manulife, Sun Life, etc.)
- Inventory planning in supermarkets
- Smart Cities: Locating the next ABM, Store, Day Care, etc.
- Point of Care Aid: Healthcare (Toronto)
- Identifying High Risk Admissions in ER (CHEO)
- Tracking the Spread of Diseases (FluMap)
- Predictive policing (San Francisco)
- Smart Tracking of Packages (Canada Post)

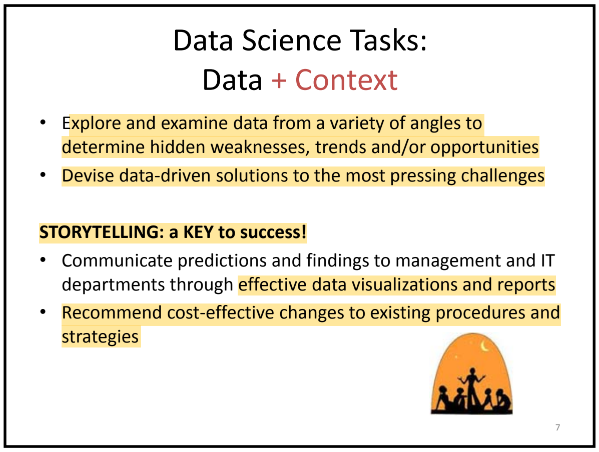
Stu's Views
© stus.com All Rights Reserved
I guess buying a $10 million life insurance policy on my mother-in-law was too obvious.



Data Science Tasks: STORE and EXPLORE

STORE and EXPLORE

- **Extract** data from multiple internal and external sources
- **Analyse** data to determine and improve the quality
- **Preprocess** data to discard irrelevant information
- **Employ** analytics programs, machine learning and statistical methods for building models of the data
- **Explore and examine** data from a variety of angles to determine hidden weaknesses, trends and/or opportunities
- **Devise** data-driven solutions to the most pressing challenges

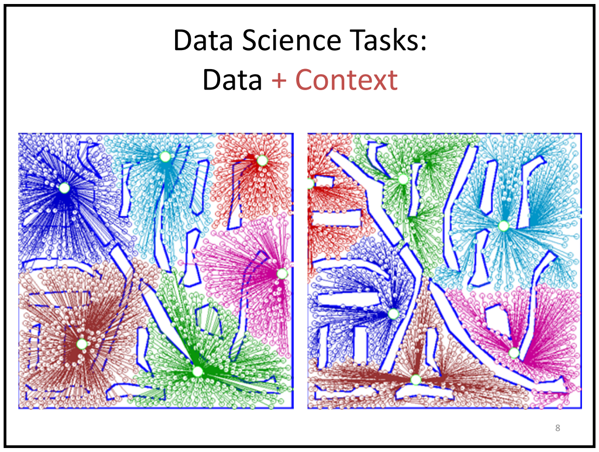


Data Science Tasks: Data + Context

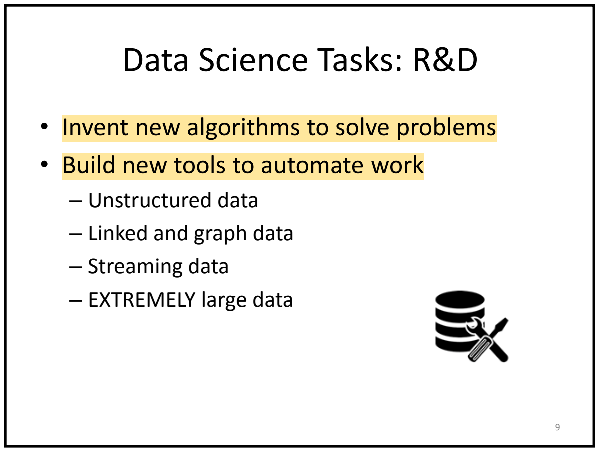
- Explore and examine data from a variety of angles to determine hidden weaknesses, trends and/or opportunities
- Devise data-driven solutions to the most pressing challenges

STORYTELLING: a KEY to success!

- Communicate predictions and findings to management and IT departments through effective data visualizations and reports
- Recommend cost-effective changes to existing procedures and strategies



Data Science Tasks: Data + Context

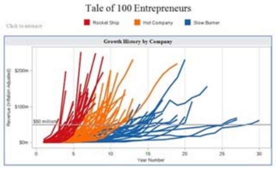


Data Science Tasks: R&D

- Invent new algorithms to solve problems
- Build new tools to automate work
  - Unstructured data
  - Linked and graph data
  - Streaming data
  - EXTREMELY large data

Data Science: CSI Skills

- Database programming (SQL, or a variant such as HIVE or MDX)
- Data preprocessing skills (R or Python, or a dedicated tool)
- Data mining skills (R or Python, or a dedicated tool)
- Data visualization (R or a dedicated tool such as Tableau, QlikView, etc.)



https://www.analyticsvidhya.com/wp-content/uploads/2015/07/entrepreneurs_journey.jpg

Course Outline:

Refer to the Syllabus on the Virtual Campus

About the team project

- The aim of the project is
  - to design and build a data mart (database for decision support)
  - to explore using Online analytical processing (OLAP) and data mining (machine learning)
- Requirements:
  - time/date dimension; data as recorded over time
  - 100,000+ records
- Send me your suggestions by 19 January 2018:
  - Canada Federal and Provincial Open Data
  - World Bank Data
  - https://github.com/caesar0301/awesome-public-datasets
- All teams will use the same data, as determined by majority voting

Last year’s project

- Crowd-sourced prices of basic groceries in 8 different countries
- Collected via mobile phones
- 500,000 records (duplicates)
- Grocery Prices: Rice, Oranges, Apples, Maize, Salt, Tea, etc.
- https://data.worldbank.org/data-catalog/crowd-sourced-price-collection

Next time:
Getting to know our data