CISC 372
Advanced Data Analytics

https://l1nna.com/course/cisc372/

Terminology:

Data mining

= [Advanced] Data analytics

= Knowledge discovery

= Inductive modelling of systems from data

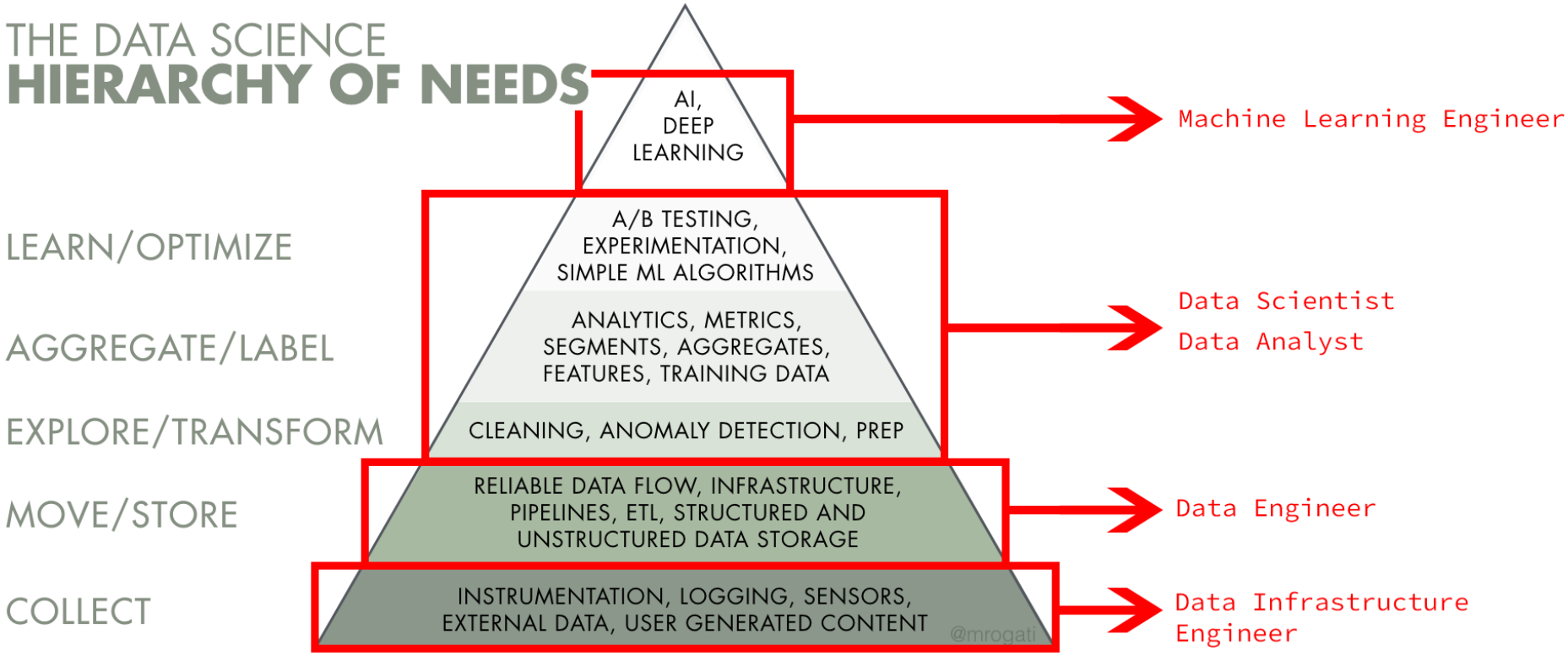
= Machine learning

= **Data Science**

>_ whomai

Dr. Steven Ding

- AI, Machine Learning, Data Mining, and **Cybersecurity**
  - PhD, McGill University (2019)
  - Assistant professor, Queen's (2019–)
  - Director, L1NNA Research Lab, l1nna.com
    - AI for security, and security for AI
  - Created **Kam1n0**
- The father of a child


THE DATA SCIENCE
HIERARCHY OF NEEDS
AI, DEEP LEARNING
A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS
ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA
CLEANING, ANOMALY DETECTION, PREP
RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE
INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT
@mrogati
LEARN/OPTIMIZE
AGGREGATE/LABEL
EXPLORE/TRANSFORM
MOVE/STORE
COLLECT
Machine Learning Engineer
Data Scientist
Data Analyst
Data Engineer
Data Infrastructure Engineer

Evolution of Sciences: New Data Science Era

- Before 1600: **Empirical science**
- 1600-1950s: **Theoretical science**
  - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s: **Computational science**
  - Most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
  - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now: **Data science**
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes
  - Data mining/analytics is a major new challenge!
- Jim Gray and Alex Szalay, *The World Wide Telescope: An Archetype for Online Science*, Comm. ACM, 45(11): 50-54, Nov. 2002

Learning outcomes:

After successful completion of this course you will:

- Have a new view of how to understand complex systems
- Be able to design and execute a process for modelling complex systems that is appropriate, effective, and revealing
- Be able to assess the validity of models for their intended purpose
- Have a deep understanding of the issues and tradeoffs of data collection, data quality, and privacy
- Be able to earn a LOT more money

Topics

- Data analytics as an epistemology, review of optimization-based prediction and clustering, ethical issues in data analytics (1 week)
- Assessing model quality (accuracy, F1 score, precision, recall, ROC, AUC) (1 week)
- Predictors based on counting: decision trees, rule systems (2 weeks)
- Bias, variance, ensemble techniques (random forests, xgboost, bagging, boosting) (2 weeks)
- Visualization (1 week)
- Data analytics for graph data (Internet search, recommender systems) (2 weeks)
- Social network analysis (1 week)
- Natural language analytics (1 week)
- Introduction to deep learning (1 week)

Workload and Grading

- Be prepared to spend adequate time and effort on this course.

- **20% Assignments - 4 assignments**
- **30% Quizzes – 4 Quizzes**.
- **50% Project**
  - (proposal 5% + presentation 10% + final report 35%)

E-mail Policy

- When you send e-mail to me, put **"CISC 372"** in the subject area, so that it can pass the spam filter.

- Visit me during my office hour if questions required extensive explanation.

- **Course email list is a must-read**.

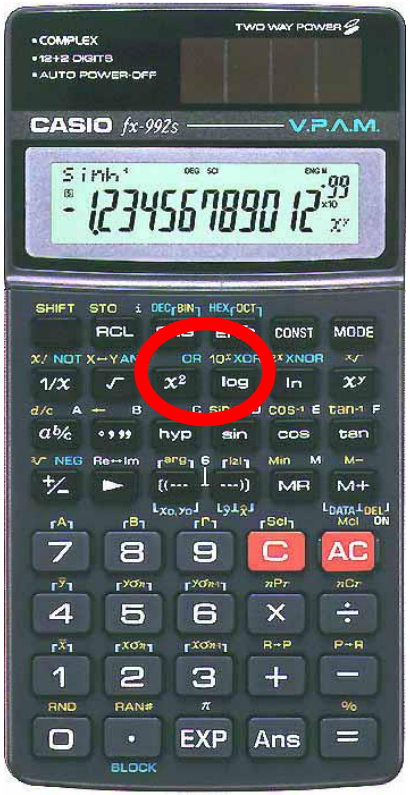
Letter Grades

- Follow the grading scheme specified by the FAS.

- Timeline (see course website)

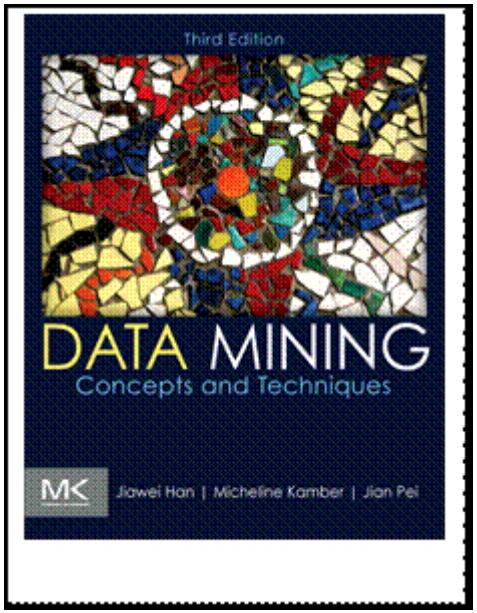
Calculator with “log” function

- The following models are recommended.
- CASIO
  - fx-100MS, fx-115MS,
  - fx-260, fx-570MS,
  - fx-991MS, fx-992S
- SHARP
  - EL-510, EL-520,
  - EL-531, EL-546
  - Models extensions are acceptable.



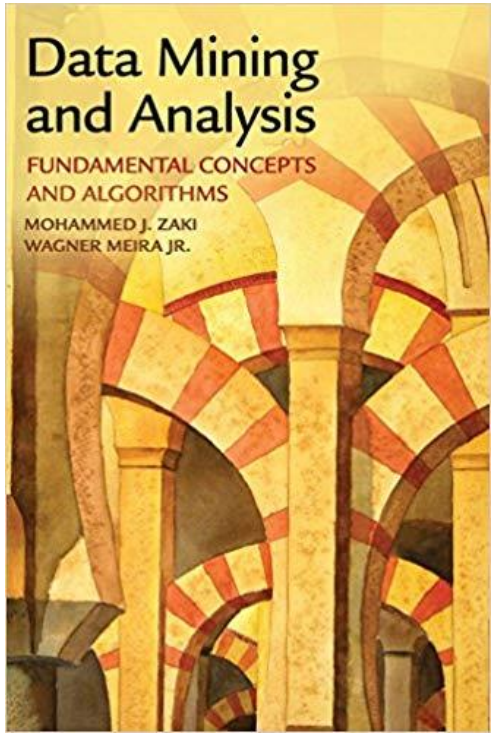
(Optional) Textbook

- ***Data Mining: Concepts and Techniques***, 3rd edition, Jiawei Han, Micheline Kamber, and Jian Pei, 2012.



(Optional) Textbook

- ***Data Mining and Analysis: Fundamental Concepts and Algorithms.*** Zaki and Meira. Cambridge University Press.



Where to Find References? DBLP, CiteSeer, Google Scholar

- Data mining and KDD (SIGKDD: CDROM)
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)
  - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
  - Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.
- AI & Machine Learning
  - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
  - Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.
- Web and IR
  - Conferences: SIGIR, WWW, CIKM, etc.
  - Journals: WWW: Internet and Web Information Systems,
- Statistics
  - Conferences: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization
  - Conference proceedings: CHI, ACM-SIGGraph, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.

Feedback and Suggestions

- Your feedback and suggestions are most welcome!

- Two anonymous course evaluations:
  - Mid-course evaluation
    - Unofficial
    - Gathering feedback, so I can improve in the rest of this semester.
  - Official course evaluation
    - Official -- for administrative purpose
    - For improving the next course offering

All schedules and related resources

- https://l1nna.com/course/cisc372/