CSI5180. Machine Learning for Bioinformatics Applications

Course **overview**

by

Marcel Turcotte

Version **March 1, 2021**

Preamble

Preamble

Course overview

Machine Learning for Bioinformatics Applications is about the analysis of complex biological data using modern machine learning methods. No prior machine learning knowledge is assumed. However, a basic understanding of probability and statistics is needed, as well as, calculus and linear algebra. Also, I am expecting that you can write programs in Python. Now, **what about biology?** Biology is important as bioinformatics strives to solve "real-world" problems. There will be at least two lectures introducing essential concepts of the molecular biology of the cell. Inevitably, we will revisit these concepts each time that a new problem will be introduced. At the very least, I am expecting a desire to learn more about biology.

General objective :

- **Summarize** the learning objectives and the expectations for this course

Learning objectives

- **Clarify** the proposition
- **Summarize** what bioinformatics is about
- **Give** an overview of the instructor's background
- **Discuss** the syllabus
- **Articulate** the expectations

Reading:

Chunming Xu and Scott A Jackson.
Machine learning and complex biological data.
Genome Biol, 20(1):76, 04 2019.

Plan

1. Preamble
2. Proposition
3. About the course
4. About me
5. What is Bioinformatics?
6. Syllabus
7. What is Machine Learning?
8. Prologue

Proposition

AI detects mutations behind autism

- "Using **artificial intelligence**, a Princeton University-led team has decoded the functional impact of such **mutations** in people with **autism**."
- Zhou et al. Nat Genet, 51(6):973980, June 2019.
- https://bit.ly/2QtnmxS

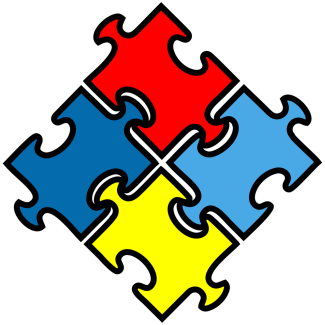
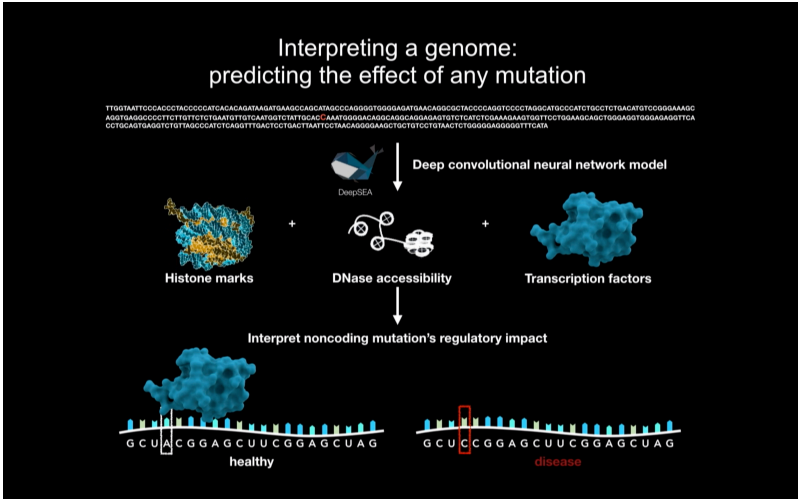


Image: Autism Daily Newscast

Olga Troyanskaya/Princeton at AI NY 2019



https://oreilly.com/go/ainy19

AI detects mutations behind autism

- "We address the challenge of **detecting the contribution of noncoding mutations to disease** with **a deep-learning-based framework** that predicts the specific regulatory effects and the deleterious impact of genetic variants."
- "Our predictive genomics framework **illuminates the role of noncoding mutations** in ASD [autism spectrum disorder] and **prioritizes** mutations with high impact for **further study**, and is broadly applicable to complex human diseases."
- Zhou et al. Nat Genet, 51(6):973980, June 2019.

"Together, the HMP1 and HMP2 phases have produced a total of **42 terabytes** of multi-omic data."

Integrative HMP (iHMP) Research Network Consortium.
The Integrative Human Microbiome Project.

Nature 569, 641648 (2019).

Improving fitness and health

- "**MyExome**, a new **DNA test** designed by Toronto entrepreneur Zaid Shahatit, claims to be able to provide a little insight into our personal quirks by testing 57 different genes that could determine our ability to **metabolize certain things**, **sleep patterns** and **physical performance**."
- **Can a DNA test improve your fitness and health?** by Christine Sismondo, The Star. July 31, 2019.

Image: myexome.com

“A Brief History of Tomorrow”

- Yuval Noah **Harari** argues that **artificial intelligence** and **genetic engineering** will play a central role shaping the future of society.

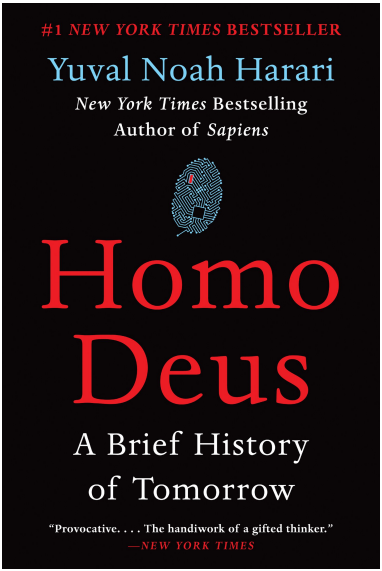


Image: Amazon.ca

About the course

What this course is not

Although the following are of **paramount importance**, this **is not** what this course is about:

- Computational Learning **Theory:**
  - **Probably approximately correct learning** (PAC Learning)
    proposed by Leslie Valiant;
  - **VC theory**
    proposed by Vladimir Vapnik and Alexey Chervonenkis;
  - **Bayesian inference**
    influenced by Judea Pearl;
  - **Algorithmic learning theory**
    from E. Mark Gold;
  - **Online machine learning**
    from Nick Littlestone.
- Compression bounds and learnability in general.

What is course is

- **Practical applications** of machine learning to **biological sequence data**, **gene expression**, **genomics** and **proteomics**.

Aurélien Géron.
Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow.
O'Reilly Media, 2nd edition, 2019.

Andriy Burkov.
The Hundred-Page Machine Learning Book.
Andriy Burkov, 2019.

What I would like the course to be. . .

- In **future editions** of this course:
  - Extensive set of examples
  - **Practical Machine Learning Applications in Bioinformatics** (textoobk)
  - Hackathon, hackfest, codefest, and (friendly) **competitive challenges**;
  - Participation to **international competitions**:
    - `https://dream.recomb2019.org`.
- Activity in the **bioGARAGE**;
- **Guests** lectures.

Cellular Molecular Biology Problems

- Predicting protein stability changes upon mutation, intrinsically disordered protein region
- Protein secondary and tertiary structure prediction
- Prediction of anti-hypertensive peptides
- Genome assembly, gene prediction, genome annotation
- Identifying DNA landmark sites: methylation, splice site, promotors, protein binding sites, etc.
- Prediction and prioritization of gene functional annotations.
- Clustering and classification of non-coding RNA genes
- Subtypes cancer classification
- Toxicity, carcinogenicity, structure activity relationships
- Predicting disease associations, identify robust prognostic gene signatures
- Sub-cellular localization

Machine Learning Concepts

- Feature Engineering, Data Imputation, Dimensionality Reduction
- Unsupervised Learning
- Linear and Logistic Regression
- Decision Trees, Random Forests and eXtreme Gradient Boosting, Ensemble
- Hidden Markov Models
- Kernel Methods, Support Vector Machines
- Deep Learning: Fundamentals, Embeddings, Architectures
- Concept and Rule-based
- Learning Graphs
- Semi-supervised Learning
- Automated Scientific Discovery

Learning objectives

- **Encode** and **clean** biological data for machine learning applications
- **Apply** modern machine learning methods to solve bioinformatics problems
- **Find** optimal values for the hyperparameters a given machine learning algorithm and data set
- **Use** a sound methodology for your machine learning projects
- **Critically review** scientific publications in this field
- **Locate** and **critically evaluate** scientific information
- **Present** scientific content to a small technical audience

About me

Professional experience

- 1989, **Honours project**, implementation of a graphical user interface for a protein folding/unfolding system

Professional experience

- 1989, **Honours project**, implementation of a graphical user interface for a protein folding/unfolding system
- 1989–95, **Université de Montréal**, graduate studies under the direction of Guy Lapalme (IRO), Robert Cedergren (Biochemistry), work on methods for building nucleic acids' 3-D structures

Professional experience

- 1989, **Honours project**, implementation of a graphical user interface for a protein folding/unfolding system
- 1989–95, **Université de Montréal**, graduate studies under the direction of Guy Lapalme (IRO), Robert Cedergren (Biochemistry), work on methods for building nucleic acids' 3-D structures
- 1995–97, **University of Florida**, work with Steven A. Benner (Chemistry) on evolutionary-based approaches to predict protein secondary structure

Professional experience

- 1989, **Honours project**, implementation of a graphical user interface for a protein folding/unfolding system
- 1989–95, **Université de Montréal**, graduate studies under the direction of Guy Lapalme (IRO), Robert Cedergren (Biochemistry), work on methods for building nucleic acids' 3-D structures
- 1995–97, **University of Florida**, work with Steven A. Benner (Chemistry) on evolutionary-based approaches to predict protein secondary structure
- 1997–00, **Imperial Cancer Research Fund** (London/UK), work with Michael J.E. Sternberg and Stephen H. Muggleton (York) on the application of Inductive Logic Programming to discover automatically protein folding rules

Professional experience

- 1989, **Honours project**, implementation of a graphical user interface for a protein folding/unfolding system
- 1989–95, **Université de Montréal**, graduate studies under the direction of Guy Lapalme (IRO), Robert Cedergren (Biochemistry), work on methods for building nucleic acids' 3-D structures
- 1995–97, **University of Florida**, work with Steven A. Benner (Chemistry) on evolutionary-based approaches to predict protein secondary structure
- 1997–00, **Imperial Cancer Research Fund** (London/UK), work with Michael J.E. Sternberg and Stephen H. Muggleton (York) on the application of Inductive Logic Programming to discover automatically protein folding rules
- 2000–, **University of Ottawa**, work on nucleic acids secondary structure determination, motifs inference and pattern matching

Learning protein structure principles (1/3)

M. Turcotte, S.H. Muggleton, and M.J.E. Sternberg.
Application of inductive logic programming to discover rules governing the three-dimensional topology of protein structure.
In C.D. Page, editor, *Proc. of the 8th International Workshop on Inductive Logic Programming (ILP-98)*, LNAI 1446, pages 53–64, Berlin, 1998. Springer-Verlag.

M. J. E. Sternberg, P. A. Bates, L. A. Kelley, R. M. MacCallum, A. Müller, S. Muggleton, and M. Turcotte.
Exploiting protein structure in the post-genome era.
In *Intelligent Systems for Molecular Biology 1999*, 1999.
Oral Presentation.

Learning protein structure principles (2/3)

M. Turcotte, S.H. Muggleton, and M.J.E. Sternberg.
Learning protein structure principles.
In *The 17th Machine Intelligence Workshop*, Suffolk, UK, July 19-21 2000.
Oral Presentation.

M. Turcotte, S.H. Muggleton, and M.J.E. Sternberg.
Generating protein three-dimensional folds signatures using inductive logic programming.
In *2000 Convention of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour*, Birmingham, UK, April 17-20 2000.
Oral Presentation.

Learning protein structure principles (3/3)

Marcel Turcotte, Stephen H. Muggleton, and Michael J. E. Sternberg.
Automated discovery of structural signatures of protein fold and function.
Journal of Molecular Biology, 306(3):591–605, February 2001.

Marcel Turcotte, Stephen H. Muggleton, and Michael J. E. Sternberg.
Generating protein three-dimensional fold signatures using inductive logic programming.
Computers & Chemistry, 26(1):57–64, December 2001.

Marcel Turcotte, Stephen H. Muggleton, and Michael J. E. Sternberg.
The effect of relational background knowledge on learning of protein three-dimensional fold signatures.
Machine Learning, 43(1-2):81-95, 2001.

Annotation concept synthesis

Mikhail Jiline, Stan Matwin, and Marcel Turcotte.
Annotation Concept Synthesis and Enrichment Analysis.
Canadian AI 2010: Advances in Artificial Intelligence, 304–308, 2010.

Mikhail Jiline, Stan Matwin, and Marcel Turcotte.
Annotation Concept Synthesis and Enrichment Analysis: a Logic-Based Approach to the Interpretation of High-Throughput Experiments.
Bioinformatics (Oxford, England), 27(17):2391-2398, September 2011.

Learning relationships between motifs

Oksana Korol and Marcel Turcotte
Learning relationships between over-represented motifs in a set of DNA sequences. *2012 IEEE Symposium on Computational Intelligence and Computational Biology, CIBCB 2012*, 2012.

Frequent Subgraph Mining (FSM)

Alexander R. Gawronski and Marcel Turcotte.
RiboFSM: Frequent subgraph mining for the discovery of RNA structures and interactions.
BMC bioinformatics, 15(S2), 2014.

Smart controls

Aseel Awdeh, Marcel Turcotte, and Theodore J. Perkins.
WACS: Improving peak calling by optimally weighting controls.
In Great Lakes Bioinformatics Conference, GLBIO 2019, May 1922 2019.

Aseel Awdeh, Marcel Turcotte, and Theodore J. Perkins.
WACS: Improving Peak Calling by Optimally Weighting Controls.
biorxiv.org.

What is Bioinformatics?

Beginnings

"Computers and specialized software have become an essential part of the biologists toolkit. Either for routine DNA or protein sequence analysis or to parse meaningful information in massive gigabyte-sized biological data sets, virtually all modern research projects in biology require, to some extent, the use of computers. (...) **the very beginnings of bioinformatics occurred more than 50 years ago**, when desktop computers were still a hypothesis and DNA could not yet be sequenced."

Gauthier, J., Vincent, A. T., Charette, S. J. & Derome, N. A brief history of bioinformatics. *Brief Bioinform* 79, 137 (2018).

A. Isaev

"Broadly speaking, bioinformatics can be defined as a collection of **mathematical, statistical and computational methods for analyzing biological sequences**, that is, DNA, RNA and amino acid (protein) sequences."

In *Introduction to Mathematical Methods in Bioinformatics*,
A. Isaev, Springer, p. i, 2006.

Lacroix and Critchlow

"Bioinformatics is the design and development of computer-based technology that supports life sciences. Using this definition bioinformatics tools and systems perform a diverse range of functions including: **data collection**, **data mining**, **data analysis**, **data management**, **data integration**, **simulation**, **statistics**, and **visualization**. *Computer-aided technology directly supporting medical applications is excluded from this definition and is referred to as medical informatics*."

In *Bioinformatics: Managing Scientific Data*, **Zoé Lacroix and T. Critchlow** Editors, Morgan Kaufmann, p. 3, 2003.

Jones N.C. and Pevzner P. A.

"Biologists that reduce bioinformatics to **simply the application of computers in biology** sometimes fail to recognize the rich intellectual content of bioinformatics. Bioinformatics has become a part of modern biology and often dictates new fashions, **enables new approaches**, and **drives further biological developments**"'
In *An Introduction to Bioinformatics Algorithms*, **Jones N.C. and Pevzner P. A.**, MIT Press, p. 77, 2004.

J.J. Ramsden

"In bioinformatics, so much is to be done, the raw material to hand is already so vast and vastly increasing, and the problems to be solved are so important (perhaps the most important of any science at present) **we may be entering an era comparable to the great flowering of quantum mechanics in the first three decades of the twentieth century** (...)"

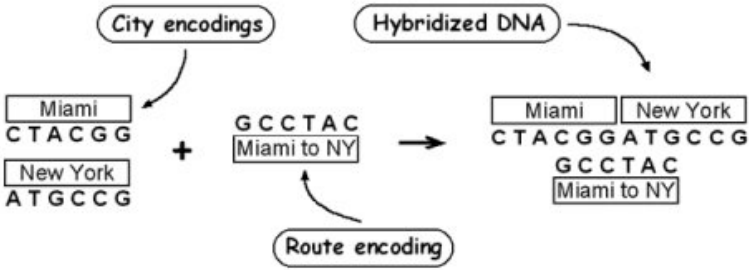
In *Bioinformatics: An introduction*, **J.J Ramsden**, Kluwer, p. xiii, 2004.

SIB - Swiss Institute of Bioinformatics

https://youtu.be/182AzhLiwxc

What it's not!

Leonard **Adleman** (*Science*, December 1994) solved a particular instance of the Hamiltonian Path problem using DNA molecules!



⇒ A Hamiltonian path visits every node of a graph exactly once.

What it's not! (continued)

DNA computing is the theoretical study of the use of DNA molecules to solve challenging problems or as a new architecture (what class of problems can be solved, what are the properties, limits, etc.).

What it's not! (continued)

- **Biotechnology** and **biomedical engineering** apply engineering approaches to problems dealing with biological systems.
- Examples of biomedical engineering include **developing biomedical devices** for human implantation, **drug delivery systems**, **simulation of organs and micro-fluids**, **medical imaging**, and many more.

Bioinformatics courses on campus

- `http://www.bioinformatics.uottawa.ca`
- **CSI 5126.** Algorithms in bioinformatics
- **BNF5106** Bioinformatics[1]
- **BCH5101** Analysis of -omics data

[1] www.bioinformatics.uottawa.ca/stephane/bnf5106.syllabus.pdf

Collaborative programs in bioinformatics

- Starting from January 2008, Carleton University and the University of Ottawa offers a Collaborative Program leading to an **MSc degree with Specialization in Bioinformatics** or **MSc of Computer Science degree with Specialization in Bioinformatics**;
- A proposal for a **Ph.D.** program is under review.

Most cited publications in Science

- Van Noorden, R., Maher, B. & Nuzzo, R. The top 100 papers. *Nature* **514**:550553, 2014.
- Wren, J. D. Bioinformatics programs are 31-fold over-represented among the highest impact scientific papers of the past two decades. *Bioinformatics* **32**(17):2686-91, September 2016.

www.bioinformatics.ca/jobs

Syllabus

Course information

Web sites

- https://www.eecs.uottawa.ca/~turcotte/teaching/csi-5180/
- https://piazza.com/uottawa.ca/fall2019/csi5180
- https://uottawa.brightspace.com

Schedule

- **Lectures**: Tuesday, 13:00 to 14:30, and Thursday, 11:30 to 13:00, MNT 103
- **Office hours**: Tuesday from 14:30 to 16:00 at STE 5-106
- **Official schedule**: www.uottawa.ca/course-timetable

Evaluation

- 30% — assignments (3)
- 10% — presentation (1)
- 20% — project (1)
- 40% — examinations (2)

Deadlines

- Assignments
  - **A1** - October 10, 2019, 18:00
  - **A2** - October 31, 2019, 18:00
  - **A3** - November 21, 2019, 18:00
- Presentation
  - Schedule will be published on September 19, 2019
  - Presentations between October 1, 2019 and December 3, 2019
- Project
  - Outline - October 1, 2019
  - Report - December 3, 2019
- Examinations
  - Midterm - October 24, 2019
  - Final - December 5 to 18, 2019

What is Machine Learning?

"Lets start by telling the truth: **machines dont learn**. (...) just like **artificial intelligence is not intelligence**, **machine learning is not learning**."

The Hundred-Page Machine Learning Book, Andriy Burkov, 2019

1959

- Samuel, A. L. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development* 3, 211229 (1959).
  - "A computer can be programmed so that it will learn to play a better game of checkers than **can be played by the person who wrote the program**."
  - "Programming computers **to learn from experience should eventually eliminate the need for much of this detailed programming effort**."

1951

- Stephen .H. Muggleton. Logic and learning: Turing's legacy. In K. Furukawa, D. Michie, and S.H. Muggleton, editors, **Machine Intelligence** 13, pages 37-56. Oxford University Press, 1994.
  - "Inspired by a radio talk given by **Turing** in **1951**, **Christopher Strachey** went on to implement the worlds first machine learning program."

Mitchell

- Tom M Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.
  - "A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Machine learning in computational biology

- Chicco, D. Ten quick tips for machine learning in computational biology. *BioData Mining* 10, 35 (2017).
  - "A machine learning algorithm is a **computational method based upon statistics**, implemented in software, able to **discover hidden non-obvious patterns** in a dataset, and moreover to make **reliable statistical predictions about similar new data**."
  - "The ability [of machine learning] to automatically identify patterns in data [...] is particularly important when the expert knowledge is incomplete or inaccurate, when the amount of available data is too large to be handled manually, or when there are exceptions to the general cases."

Prologue

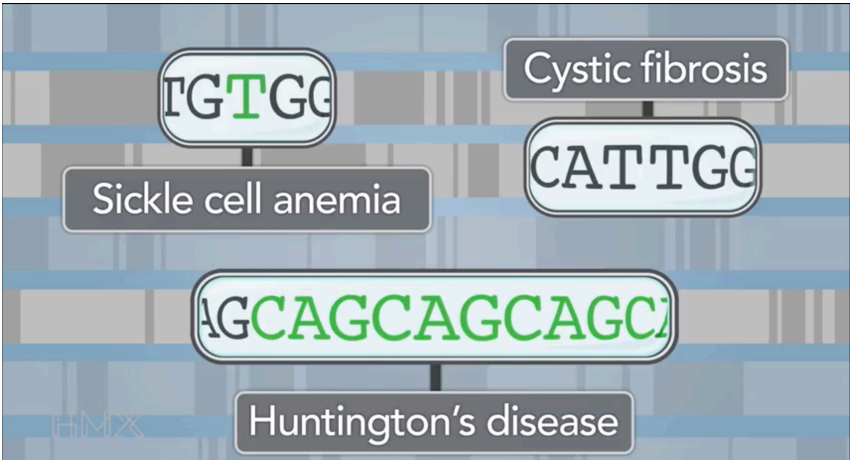
Summary

- A **practical** application of **machine learning** to **biological data**

Summary

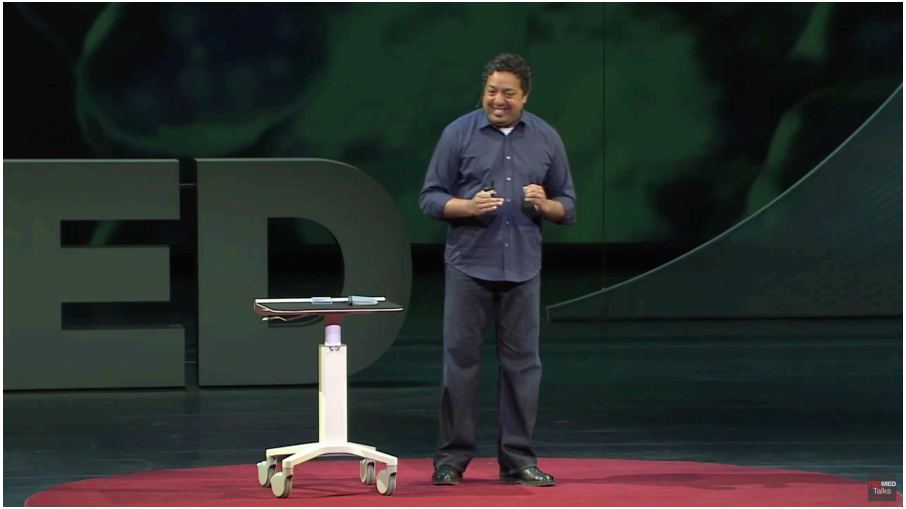
- A **practical** application of **machine learning** to **biological data**
- **Python** programming skills and a **love of biology** are both expected

An Introduction to the Human Genome



https://youtu.be/jEJp7B6u_dY

Atul Butte/Stanford at TEDMED 2012



https://youtu.be/dtNMA46YgX4

Atul Butte/Stanford at TEDx 2017

https://youtu.be/dtNMA46YgX4

Going from CS to bioinformatics

https://www.youtube.com/watch?v=4mMviYCgBSU

https://www.youtube.com/channel/UCG4kmWK8UyzfenZ60xVBapw

Next module

- **Essential** Cell Biology (two lectures)

References

Chunming Xu and Scott A Jackson.
Machine learning and complex biological data.
Genome Biol, 20(1):76, 04 2019.

Jian Zhou, Christopher Y Park, Chandra L Theesfeld, Aaron K Wong, Yuan Yuan, Claudia Scheckel, John J Fak, Julien Funk, Kevin Yao, Yoko Tajima, Alan Packer, Robert B Darnell, and Olga G Troyanskaya.
Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk.
Nat Genet, 51(6):973–980, Jun 2019.

Tom M Mitchell.
Machine Learning.
McGraw-Hill, New York, 1997.

Marcel **Turcotte**

Marcel.Turcotte@uOttawa.ca

School of Electrical Engineering and **Computer Science** (EECS)
University of Ottawa