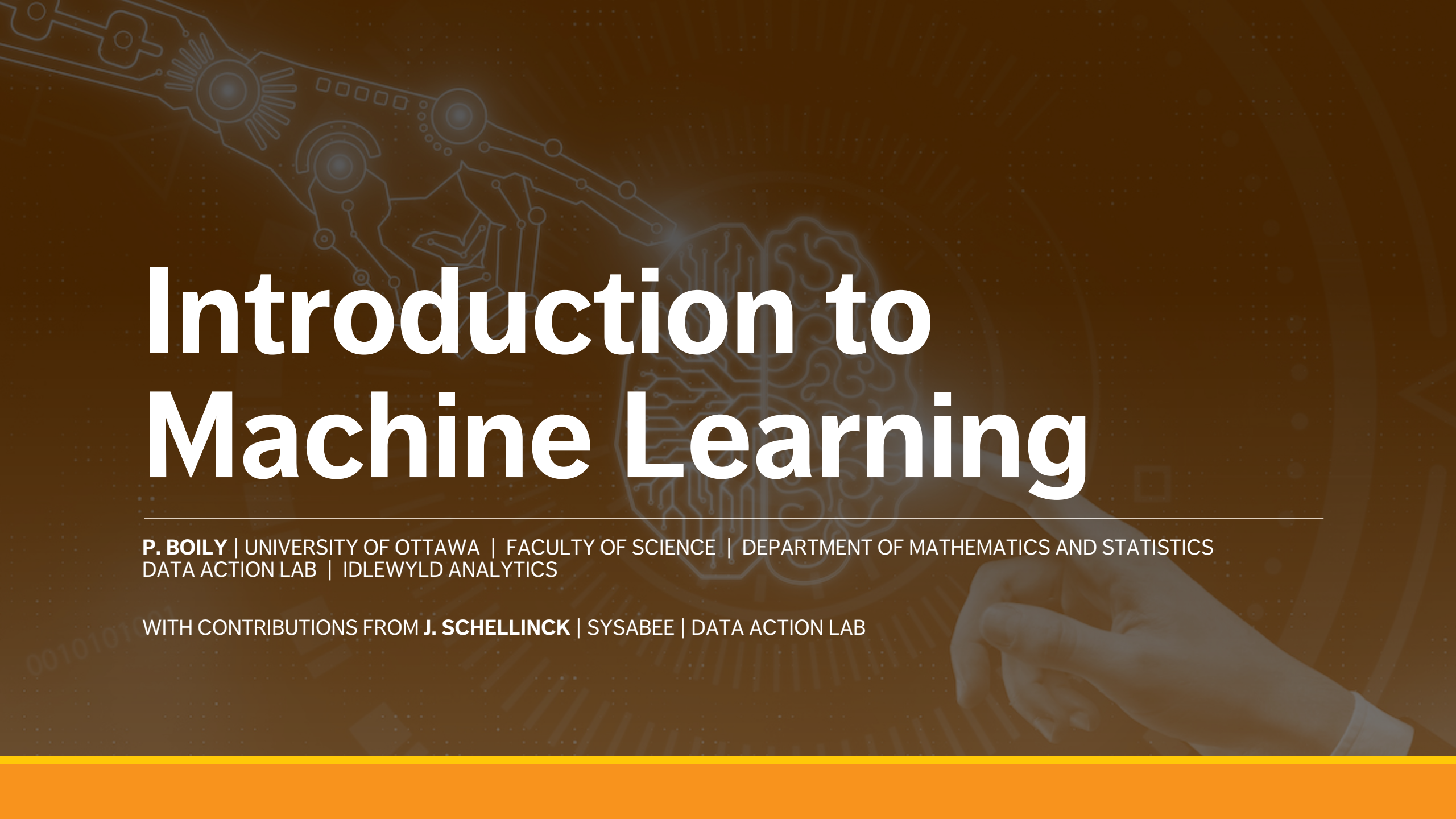


Introduction to Machine Learning

P. BOILY | UNIVERSITY OF OTTAWA | FACULTY OF SCIENCE | DEPARTMENT OF MATHEMATICS AND STATISTICS
DATA ACTION LAB | IDLEWYLD ANALYTICS

WITH CONTRIBUTIONS FROM **J. SCHELLINCK** | SYSABEE | DATA ACTION LAB

Instructor – Patrick Boily

Employment

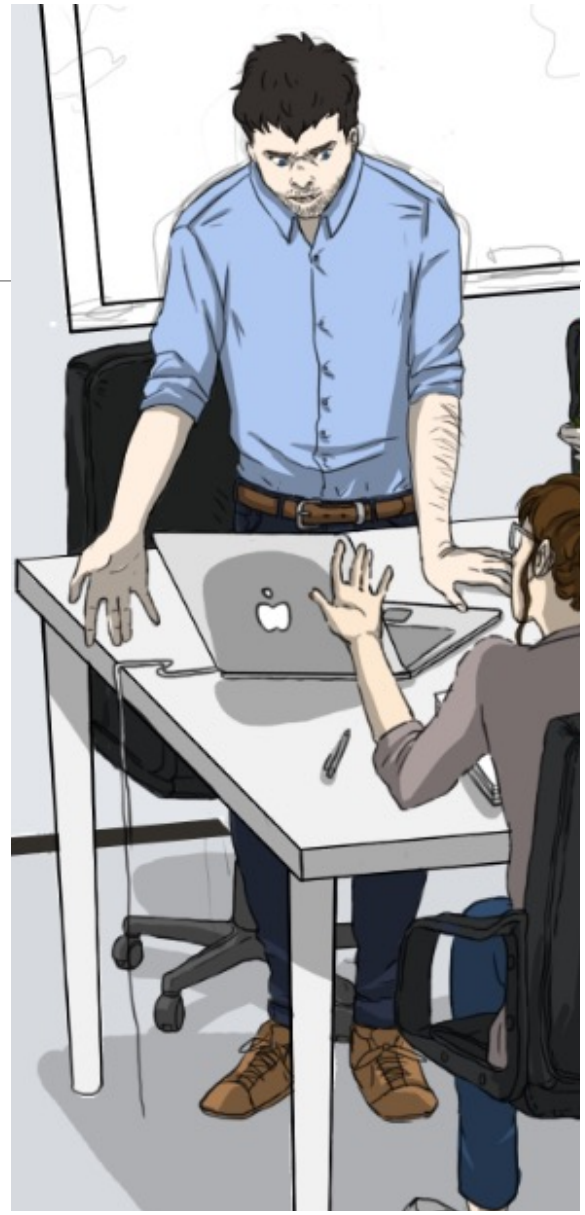
- Professor Math/Stat [‘19 – now, uOttawa]
- President [‘16 – now, Idlewyld Analytics]
- Manager and Senior Consultant [’12 – ‘19, CQADS, Carleton]
- Public Service [’08 – ’12, ASFC | StatCan | TC | TPSGC]
- 60+ uni course; 250+ workshop days

Projects

- GAC; NWMO; CATSA; etc.
- 40+ projects

Specialization

- Data visualization; data cleaning (... unfortunately)
- Application of wide breadth of techniques to all kinds of data
- Mathematical/statistical modeling



Course Material

Course Webpage:

https://data-action-lab.com/103-iml

Course Notes:

https://idlewyldanalytics.com

Contact Info:

pboily@uottawa.ca

Slack Workspace:

https://dspdi.slack.com

Course Description

This course leads the participants to analyze and discuss the general tasks and problems of statistical learning (machine learning), as well as their pitfalls.

In this course, participants will be introduced to simple association rules mining, classification, and clustering algorithms.

Following the course, the participants have the option of working on a guided project, getting feedback from the instructor.

Additional Information

Participants are expected to be familiar with the concepts introduced in the courses *Data Science Essentials* (data preparation, data cleaning), and *Data Visualization and Dashboards* (data exploration), and their pre-requisites.

Familiarity with optimization methods would be beneficial but is not required.

Participants are required to bring a laptop/personal computer on which the current version of R/RStudio (Posit) is installed (for which they may require administrative authorisation to install packages).

Participants doing a guided project must be familiar with R, the tidyverse, and/or Python.

Learning Outcomes

At the end of this course, participants will be able to:

- differentiate between situations which require a supervised learning approach and those which require an unsupervised approach (or some combination of both)
- identify strategies used to overcome common real-world statistical learning issues and challenges
- recognize the variety of machine learning algorithms available to them
- implement simple machine learning algorithms to provide actionable insights
- build a simple data analysis pipeline incorporating machine learning components

Course Outline

Statistical Learning

1. Types of Learning;
   Machine Learning Tasks

Association Rules Discovery

2. Association Rules Overview;
   Case Study: Danish Medical Data
3. Association Rules Concepts

Classification

4. Classification Overview;
   Case Study: Minnesota Tax Audits
5. Decision Trees and Other Algorithms
6. Performance Evaluation

Session 1
Session 2
Session 3
Session 4

Course Outline

Clustering

7. Clustering Overview; *Case Study*: Livehoods
8. k-Means and Other Algorithms
9. Validation and Notes

Issues and Challenges

10. Bad Data and Big Data
11. Underfitting and Overfitting/Transferability
12. Miscellanea

Session 1

Session 2

Session 3

Session 4

Sister Courses

DATA SCIENCE ESSENTIALS

1. Non-Technical Aspects
2. Data Science Basics
3. Data Preparation
4. Data Engineering

DATA VISUALIZATION AND DASHBOARDS

1. Data Viz Concepts
2. Dashboarding
3. Storytelling with Data
4. Data Viz with ggplot2

POWER BI FOR BEGINNERS

1. The Tool
2. Exploration
3. Monitoring
4. Storytelling