ARMY CYBER INSTITUTE AT WEST POINT

Rapid ML Prototyping

MAJ Iain Cruickshank

iain.cruickshank@westpoint.edu

30 NOV 22

Agenda

- Basis for This Course
- Principles of Rapid ML Prototyping
- Worked Examples
- Practical Exercise
- Conclusion

About Me

- Military Background
  - Functional Area 49 (ORSA)
  - Base branch of Military Intelligence
  - Assignments at 101st, 780th, and AI2C
  - Currently a senior research scientist at the ACI
- Academic Background
  - BS in Mathematics from USMA (2010)
  - MS in OR from U of Edinburgh (2011)
  - Ph.D. in Societal Computing from CMU (2020)
- Professional Background
  - Frequent data science competitor
  - Instruct at the undergraduate and graduate levels in applied machine learning, research, network science, and data science

Democratization of Machine Learning

- Machine Learning, and the data powering it, is becoming increasingly more available, more accessible, and more applicable
- The same frameworks underlie research and production
- Widely available compute, especially cloud compute

Scope of This Course

- This course is not…
  - Going to make you a data scientist
  - Going to teach data literacy or coding
  - Going to cover the math and theory behind ML
- This course will…
  - Augment you with new tools to use in your work
  - Teach you how to look at data problems
  - Teach you how to get to a working solution in a matter of hours.

Scope of this Course

The overall objective for this course is to give you instruction on machine learning such that you, in the course of the work in your field, can quickly apply machine learning tools to your problems.

What is Machine Learning?

- Machine Learning consists of four things:
  - The Model ~ $h_\theta(.)$
  - Examples ~ (x_i, y_i)
  - A Loss Function ~ $L(h_\theta(x_i), y_i)$
  - An Optimization Procedure ~ $\min_\theta L(h_\theta(x_i), y_i)$

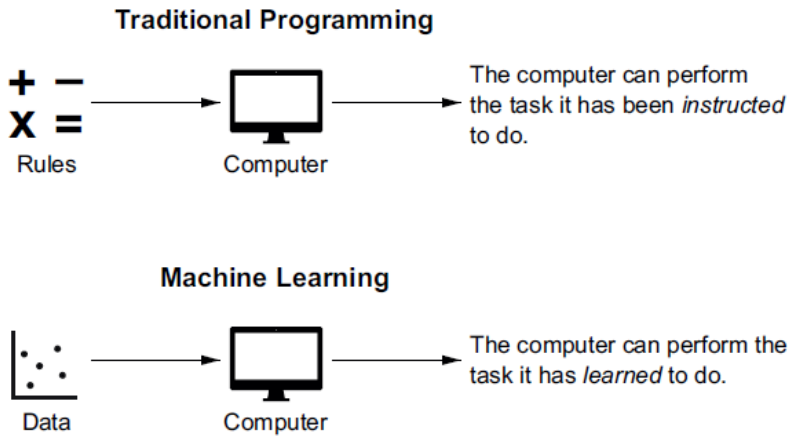


Figure 1.1 The difference between the traditional programming approach and machine learning: the first relies on precise rules and instructions, the latter on data and learning.

Mauro and Valigi, *Zero to AI* (2020)

What is Machine Learning?

- Three Main Types of Machine Learning
  - Supervised
  - Unsupervised
  - Reinforcement
- Several variations and combinations on these themes exist (i.e. self-supervised, semi-supervised, etc.)

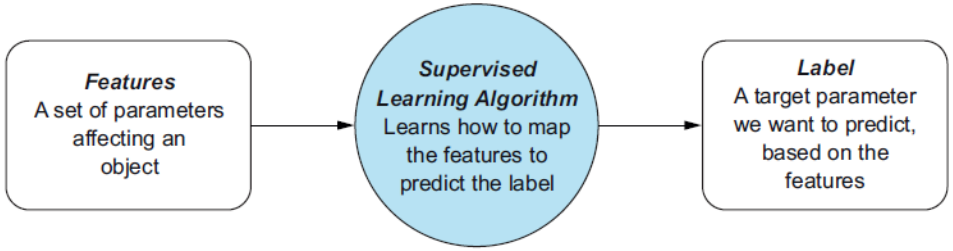


Figure 2.4 The core concept of supervised learning: finding a mapping between a set of features and a label

Mauro and Valigi, *Zero to AI* (2020)

Starting Development on an ML Capability

When considering development of an ML-based system, there are some key considerations

- Is there data available?
  - Does the data already have labels or a proxy for the label that you want
  - What kind of data is this?
- Does the problem break down into task(s) that resemble a type of ML?
- Is this something worth my time?

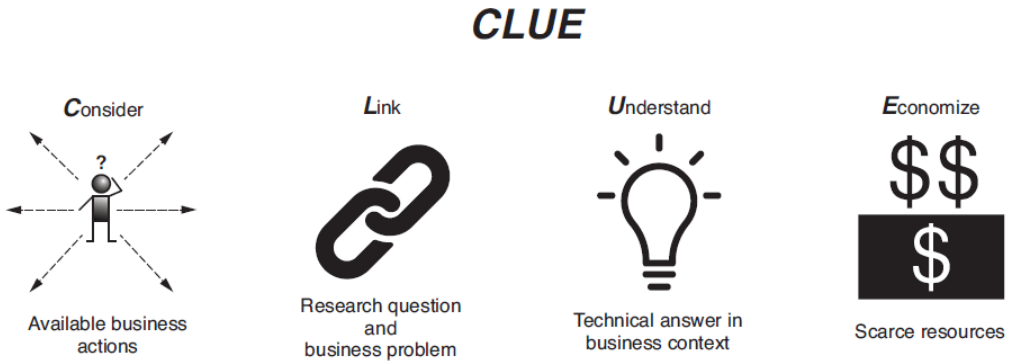


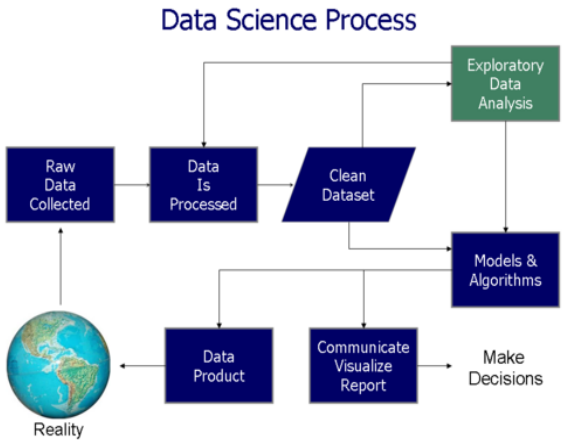
Figure 1.4 Elements of CLUE. AI projects that don't have good answers for all elements of CLUE experience difficulties.

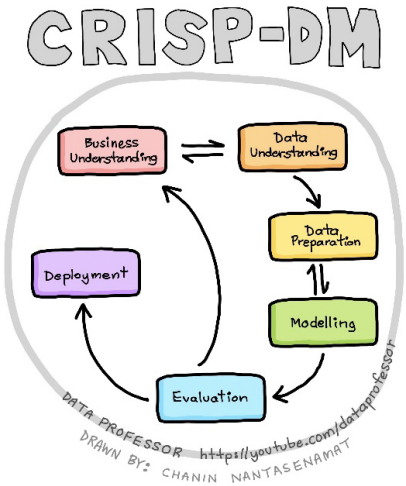
Krunic, *Succeeding with AI* (2020)

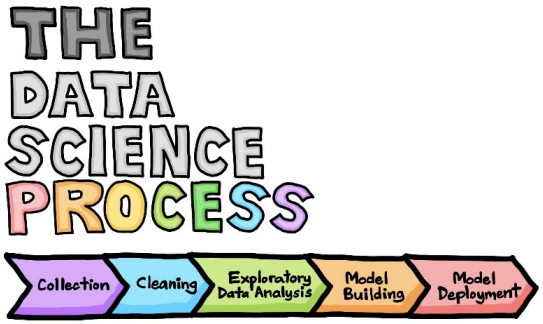
ARMY CYBER INSTITUTE AT WEST POINT

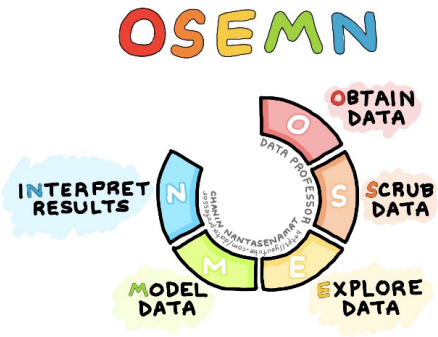
The Data Science Workflow

1. Data Collection/Triage Decisions
2. Exploratory Data Analysis
3. Feature Engineering
4. Select and Optimize a Model
5. Present Results









The Overarching Approach

When seeking to quickly prototype an ML solution

1. Narrowly and specifically design the problem
2. Use existing code/models/workflows/insights whenever possible
3. Start simple and basic, and build more complexity over iterations
4. Stay data-centric in your approach

Practical Tips For Rapid ML Model Development

- When dealing with the data
  - Work data areas with well-established methods
  - When unsure, get the data into a tabular format
- Models to use
  - Use tree-based methods for tabular data (LightGBM, XGB, CatBoost)
  - Use pre-trained deep learning models for text and image
- Where to spend your time
  - Focus on feature engineering, to the exclusion of optimizing for the best model hyperparameters if short on time

Worked Examples

Set Up

- Head to https://colab.research.google.com/
- Go to GitHub option
- Search by username: “ijcruic”
- Connect to repository: “ijcruic/rapid-ml-prototyping”

Practical Exercise

https://www.kaggle.com/t/c572f9a6f9b944079a3a30bf7df2e8f3

Next Steps

- Other techniques to explore for modeling
  - Ensembling
  - Constructing custom model architectures/ training methods
  - Using zero-shot models
  - Time series models
- Other techniques for getting the most out of your data
  - Pseudo-labeling
  - Active learning
  - Low-shot learning

Review

- When seeking to prototype an ML solution
  1. Narrowly and specifically design the problem
  2. Use existing code/models/workflows/insights whenever possible
  3. Start simple and basic, and build more complex over iterations
  4. Keep data-centric in your approach
- Survey feedback
  - https://forms.gle/oJ7zKTNPxDeHnmPg6