Introduction to Machine-Learning Methods for Patient-Reported Outcomes Data

Yuelin Li

October 19, 2022

ISOQOL Conference, Prague

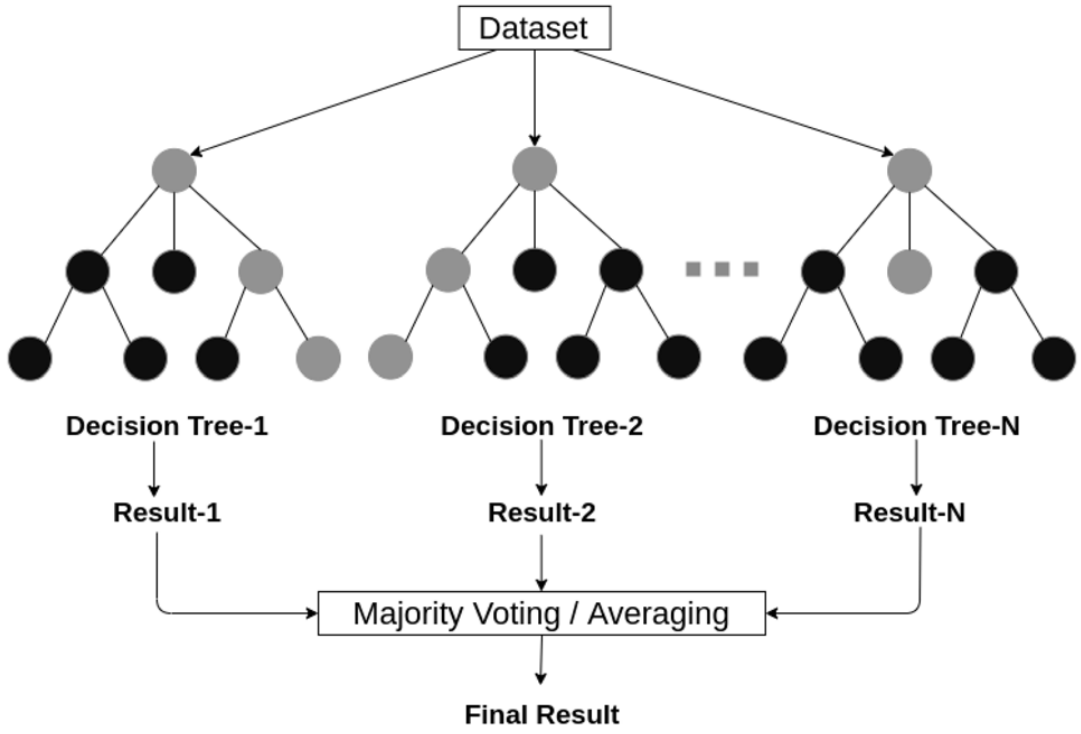
Machine Learning Methods for PRO Data

My goal is to provide a general overview of ML methods

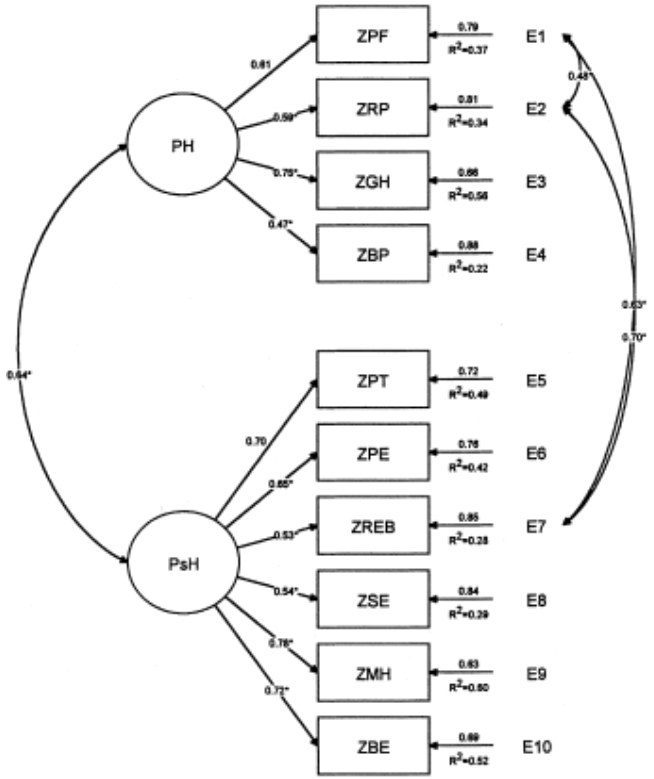
1. Why use ML in PRO data?
2. What is a typical and basic ML workflow?
3. Examples to illustrate the ML workflow
   - Tree-based method
   - Text classification by BERT, a Natural Language Processing (NLP) method
4. How to avoid common pitfalls?

Why use ML in PRO Data?

ML: *person-oriented* approach [1]



Traditional, *variable-oriented* approach



1: Bergman & Magnusson (1997) *Development and Psychopathology* https://doi.org/10.1017/S095457949700206X

Machine Learning: What is it?

- **Unsupervised Learning**
- **Supervised Learning**

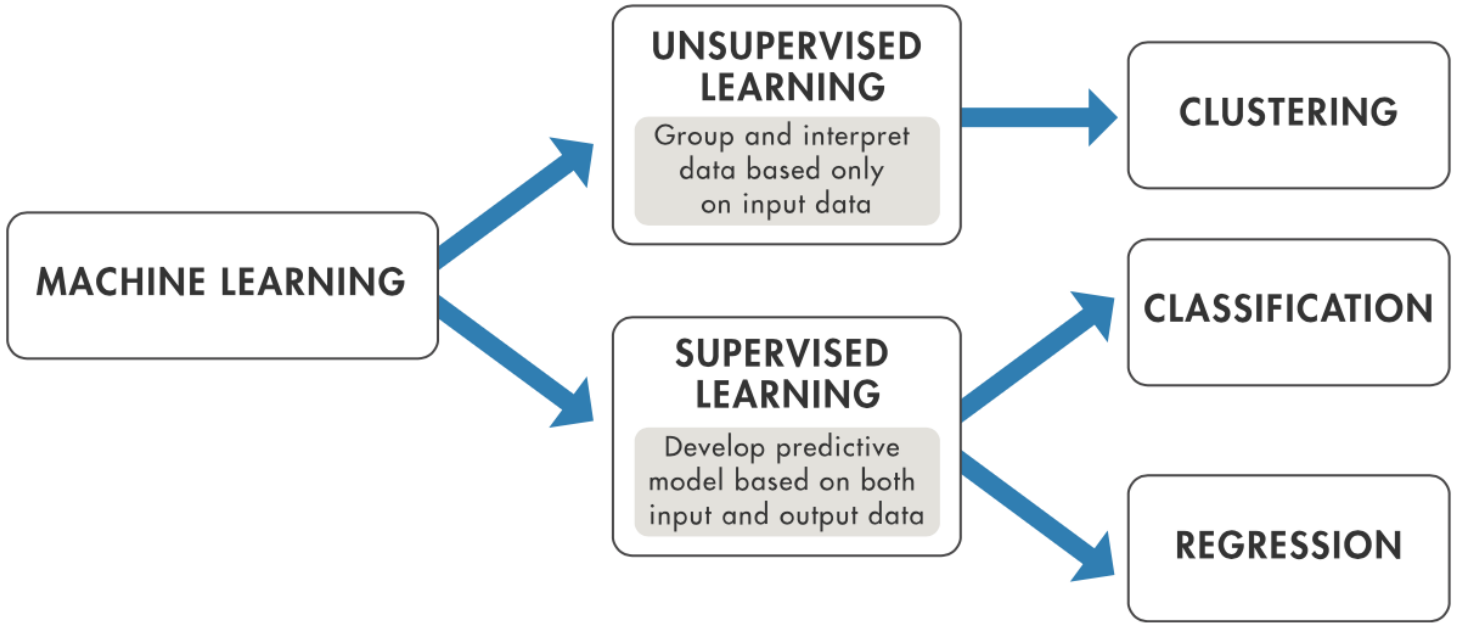


Figure 1. Machine learning techniques include both unsupervised and supervised learning.

https://www.mathworks.com/discovery/machine-learning.html

Sarker I. SN Computer Science (2021) 2:160 https://doi.org/10.1007/s42979-021-00592-x

Unsupervised Learning

- Unsupervised learning finds hidden patterns or intrinsic structures in input data
- Clustering
  - K-means
  - Finite mixture modeling
  - Bayesian nonparametric methods [1]
  - NLP (e.g., Latent Dirichlet Allocation [2])
  - Psychometric network models [4]

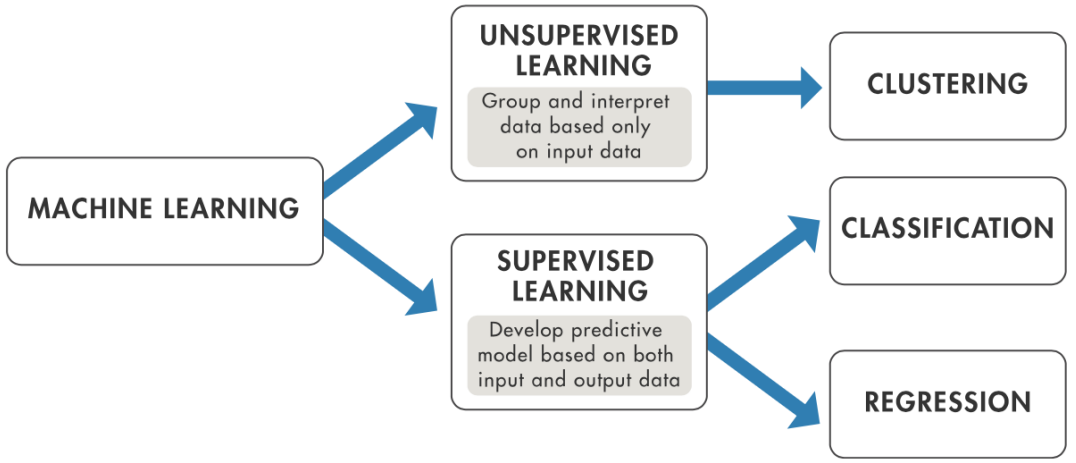


Figure 1. Machine learning techniques include both unsupervised and supervised learning.

1: Li et al. (2019, PMID: 31217637); 2. Li et al. (2019, PMID: 30798421); 3. (2022, JMSS); 4. Sacha Epskamp (2020) *Psychometrika* https://doi.org/10.1007/s11336-020-09697-3

Supervised Learning

- Supervised machine learning builds a model that makes predictions based on evidence in the presence of uncertainty
  - Classification methods
    - Classification and Regression Trees (CART)[1]
    - SVM (Support Vector Machine)
    - Random Forest
    - XGBoost (eXtreme Gradient Boosting)
  - Artificial neural networks [2]
  - NLP (e.g., BERT for Text Classification [3])
  - Regression and its cousins

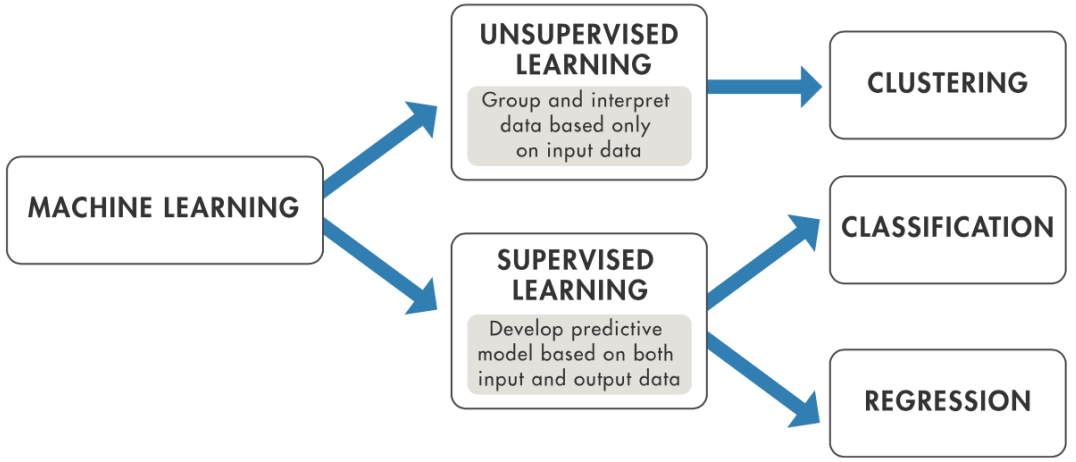
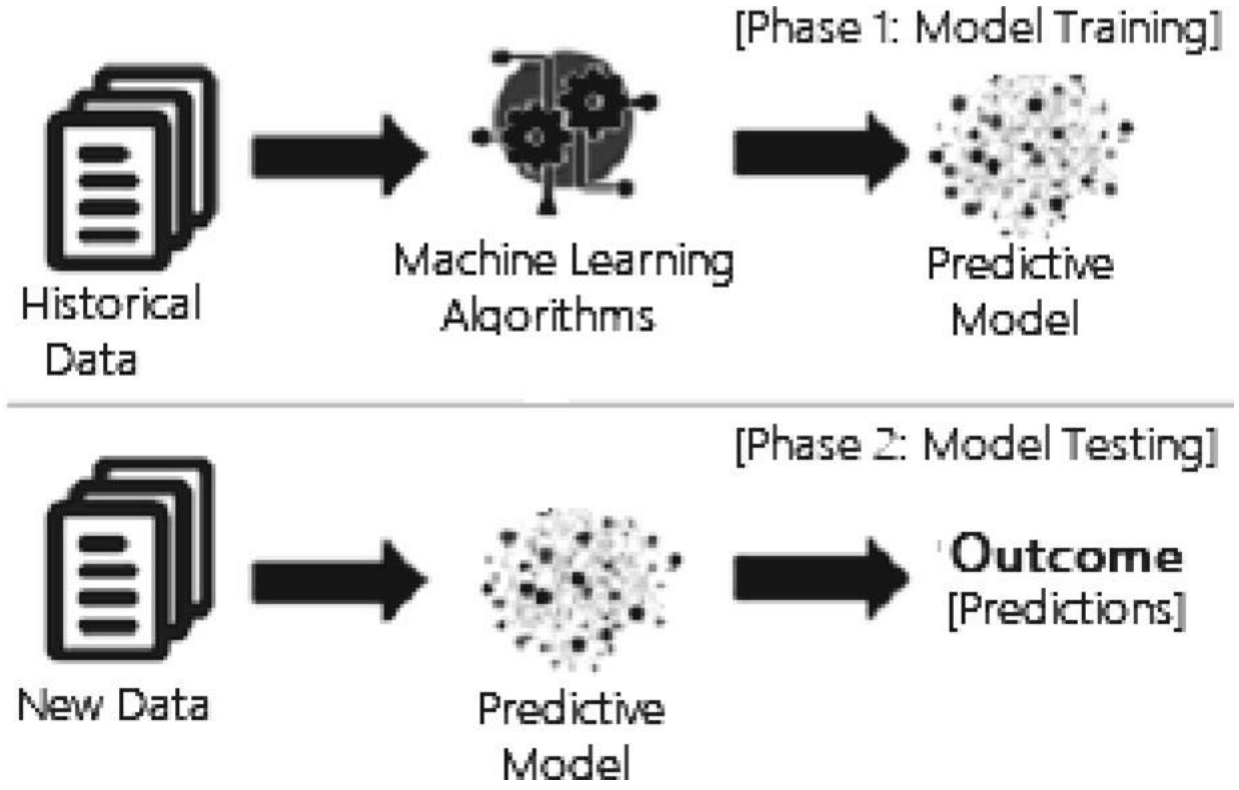


Figure 1. Machine learning techniques include both unsupervised and supervised learning.

1: Li & Rapkin (2009, PMID: 19595576); 2. Pfob et al. (2021, PMID: 33914464); 3. Schwartz et al. (2022, JMSS)

What is a typical ML workflow?

- Model training: feed historical data (often large) to ML algorithms to build a predictive model (e.g., logistic regression)
- Model testing: new data (never seen before by predictive model) fed into the model to generate predictions
- Predictive accuracy in new data (e.g., sensitivity/specificity, ROC curves, mean squared error)
- Often, as part of phase 1 model training, you do *cross-validation* to select the desired model



Sarker (2021)

Typical ML workflow

- Training sample, validation sample, and test sample
- Training sample to set the parameters of the predictive models ("train the models" in ML parlance)
- Validation to choose the best model among different models trained on the training sample
- How choose? By selecting (say) model that has the lowest empirical risk on the validation sample
- E.g., choice between a simple vs. a complex model using same training sample. We choose the model that has the lowest mean squared prediction error on the validation sample
- Test sample yields unbiased estimate of risk of model selected in the validation step
- Why test step? Because model selection can also make mistakes
- Sometimes tested by someone else on blinded data (ML competitions)
- Workflow is iterative (e.g., Train-Validation-Test1-Test2)

Cross validation types

- K-fold cross-validation

- Leave-One-Out (LOO) cross validation
  - N-fold cross validation

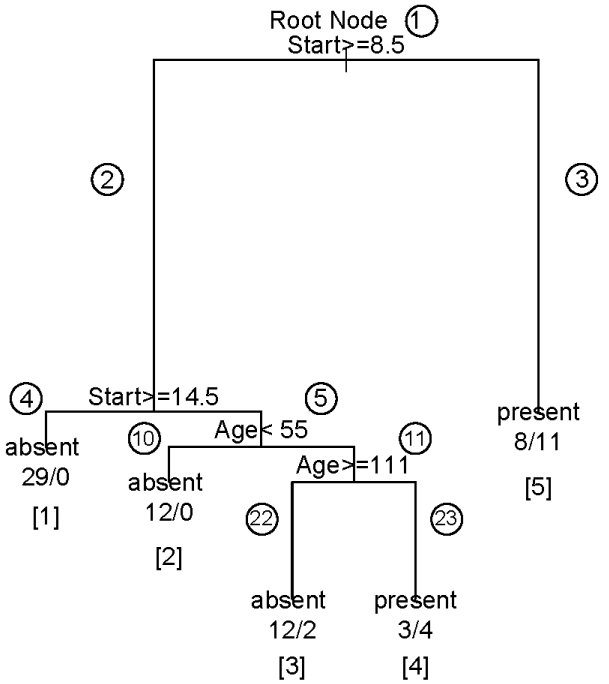
$$\mathrm{MSE} = \frac{1}{N}\sum (y_i - f(x_i))^2$$

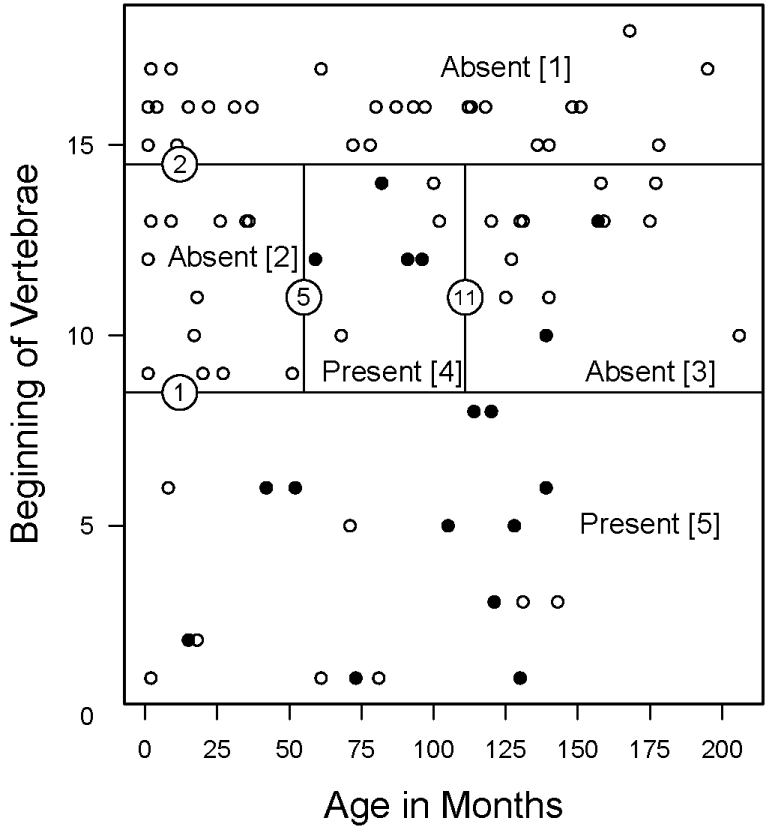
  - N: Total number of observations
  - y_i: Observed outcome value of the ith observation
  - $f(x_i)$: Predicted response value of the ith observation

Cross-Validation Example

Why x-validation?

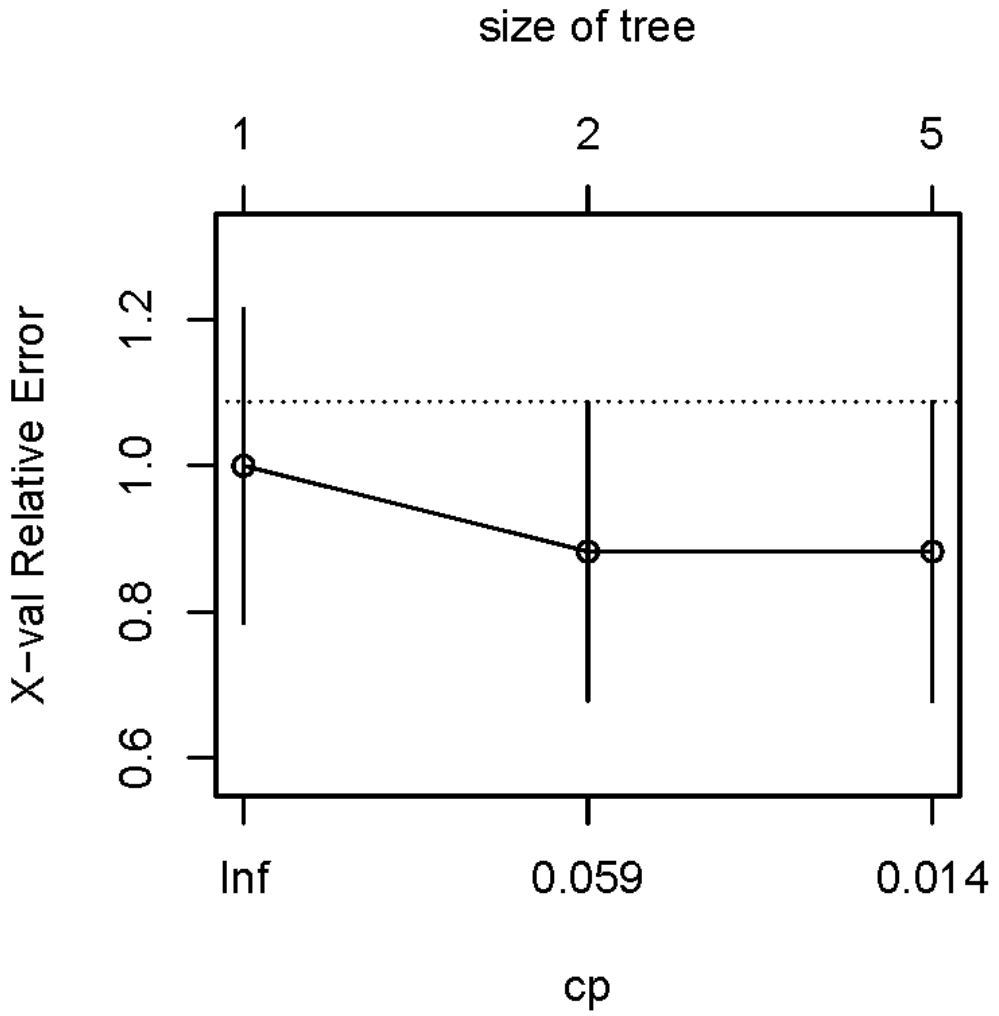
- For example, stock Kyphosis dataset
- Recursive partitioning divides 2D space between `Age` and `Start` into subgroups
- However, individuals falling near the boundaries can be misclassified by chance
- Cross-validation error provides a measure of this problem by running the same model K times
- Now, we will do this in RStudio





Prune rpart tree

- On the previous slide, rpart() grows a tree with 5 terminal nodes
- rpart() does a 10-fold cross-validation under the hood
- The `plotcp()` command plots the size of tree and its corresponding relative error to help prune the tree
- Like a traditional scree plot
- Splitting sample to 2 nodes reduces relative error (good)
- Further into 5 nodes not much worse than 2
- Rationale for a 5-node model



Natural Language Processing

Natural Language Processing by BERT

- BERT stands for ‘Bidirectional Encoder Representations from Transformers’
- Unlike earlier NLP models, BERT’s artificial neural nets take the order of words into account
- E.g., ‘work to live’ and ‘live to work’ are semantic opposites
- In late 2018, Google released BERT [1]
- Other technology companies such as Facebook and OpenAI have also joined forces to pre-train BERT using 3.3 billion words total with 2.5 billion words from Wikipedia and 0.8 billion from BooksCorpus.
- As a result, this “pre-trained” BERT has learned how the English language is structured and how the words are typically arranged in the context of sentences
- Learn more in their blog post [2]
- Carolyn Schwartz and I tested BERT in coding interview transcripts (close to the performance of a human coder) [3]
- BERT is now an integral part of Google search
- You might have noticed that you get better Google search results by complete sentences

1. Devlin, J., et al. (2018). https://arxiv.org/abs/1810.04805; 2. https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html; 3 Schwartz et al. (2022, JMSS)

Text Classification by BERT

- Next, we turn to Python to show you how to do text classification using a “pre-trained” BERT you can download

To Summarize This Part of Workshop

- Countless other points on ML not covered here
- But I hope I have given you a fundamental intuition on ML
- Importance of cross-validation and NLP parameter tuning as an integral part of data analysis in writing a paper
- You must interrogate your models (e.g., cross-validation), not just to accept ML defaults
- Make you feel more confident and efficient in learning more on your own
- Potentially rediscover finer details of ML yourself