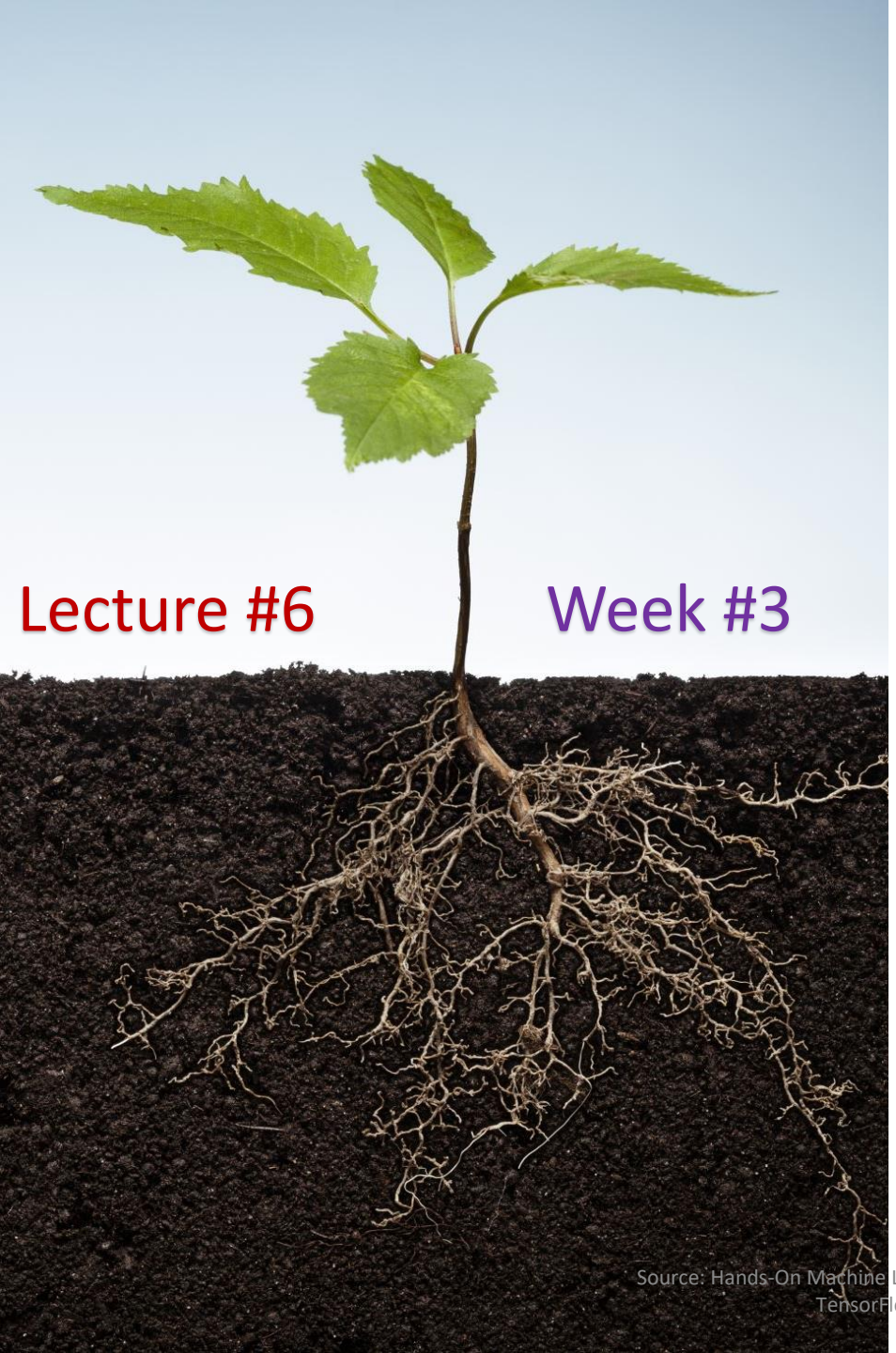

Source: Hands-On Machine Learning with Scikit-Learn and TensorFlow (book)

Pattern Recognition
ECSE 4410/6410 CAPA
Spring 2021

Machine Learning / Pattern Recognition

Problem Checklist

Course Instructor - Thirimachos Bourlai

January to May 2021

Please Check the 2021 Syllabus

Machine Learning Problems

Checklist – Main Steps

1. Problem
2. Data
3. Solution
4. Outcomes

Source: Hands-On Machine Learning with Scikit-Learn and TensorFlow (book)

Machine Learning Problems

Checklist – Details on the Main Steps

1. Frame the problem and look at the big picture.
2. Get the data.
3. Explore the data to gain insights.
4. Prepare the data to better expose the underlying data patterns to Machine Learning algorithms.
5. Explore many different models and short-list the best ones.
6. Fine-tune your models and combine them into a great solution.
7. Present your solution.
8. Launch, monitor, and maintain your system.

Source: Hands-On Machine Learning with Scikit-Learn and TensorFlow (book)

Machine Learning Problems

Adapt the steps to your needs

1. **Understand**
   - Business Objective
   - Application of your solution
2. **Know**
   - Other/existing solutions
   - Comparable problems to base your solution
   - Potential manual solution
3. **Define the type of ML model**
4. **Determine your performance evaluation process**
   - Performance measure alignment with the business objective
   - Define acceptable performance (business/class etc.)
5. **Determine availability of experts**
6. **Make necessary assumptions**
   - List the assumptions you (or others) have made so far.
   - Verify assumptions if possible.

Source: Hands-On Machine Learning with Scikit-Learn and TensorFlow (book)

Machine Learning Problems

Automation of processes & Preparation

- **Data**
  - Where?
  - How much?
  - Should you document?
  - Are there any legal obligations? Do I need to get authorization? Get it. Delete sensitive information – deanonymize.
  - Check the size and type of data (images, audio, geospatial, videos, etc.).
  - Convert to an easy-to-use format.
- **Workspace**
  - Check your space and how much is needed on your HD
  - Need access / authorization

Training and Testing

- Leave a test set sample aside
  - Do not check at these dataset…

Source: Hands-On Machine Learning with Scikit-Learn and TensorFlow (book)

Machine Learning Problems

More Check Points …

- **Create a copy of the data**
    - Explore it… is it at a manageable size? What do you do if it is not?
- **Create a record of your data copy and exploration processes**
- **Data attributes and its characteristics**
    1. Name
    2. Type (categorical, int/float, bounded/unbounded, text, structured, etc.)
    3. % of missing values
    4. Noisiness and type of noise (stochastic, outliers, rounding errors, etc.)
    5. Possibly useful for the task?
    6. Type of distribution (Gaussian, uniform, logarithmic, etc.)
- **Supervised Learning**
    - Identify the target Attribute(s)/Features/Characteristics

Source: Hands-On Machine Learning with Scikit-Learn and TensorFlow (book)

Machine Learning Problems

More Check Points …

- Data visualization can be important
- Check features their correlation
- Can you solve the problem manually first?
- Are the data you have sufficient or do you need extra data
  - If you need extra data that would be useful find MORE relevant / good data.
- Perform Data cleaning in your “to be used” dataset:
  - Fix or remove outliers (if needed).
  - Missing values:
    - Fill with zero, mean, median, or
    - Drop their rows (or columns).

Source: Hands-On Machine Learning with Scikit-Learn and TensorFlow (book)

Machine Learning Problems

Features

- Feature selection:
  - Drop unnecessary features.

- Other feature processes:
  - Discretize continuous features.
  - Decompose features (e.g., categorical, date/time, etc.).
  - Add promising transformations of features (e.g., log(x), sqrt(x), x2, etc.).
  - Aggregate features into new features.

- Feature scaling.
  - Normalization/scaling is VERY important
  - Start with sample, smaller, training sets so you can train many different models in a reasonable time (be aware that this penalizes complex models such as large neural nets or Random Forests).

Source: Hands-On Machine Learning with Scikit-Learn and TensorFlow (book)

Machine Learning Problems
Algorithm Selection Process

- Start with simple models using standard parameters.
- Compute and compare their performance.
- For each model:
  - Use N-fold cross-validation
  - Compute/Visualize the mean and standard deviation of the performance measure on the N folds.
- Analyze the most significant variables for each algorithm.
- Analyze the types of errors the models make.
- Discuss solution to avoid these errors
- Maybe feature selection is needed (ML) or use DL, or both
- Determine the top models (3-5)
  - Select base on the different types of errors, NOT blindly

Source: Hands-On Machine Learning with Scikit-Learn and TensorFlow (book)

Machine Learning Problems
Communicate the findings

- Documentation
- Presentation of solution/results
- Highlight the big picture first, then go to the details
- Explain why your solution achieves the business/research objective
- Present interesting points you noticed along the way
  - There are limitations (what works or not) to present and discuss
  - List the original hypothesis vs. solution vs. limitations
  - Nobody cares only about YOUR best solution – there are ALWAYS limitations
- Communicate key findings -- use visualizations; table and easy-to-remember statements (e.g., “the median income is the number-one predictor of housing prices”).
- Get your solution ready for demo

Source: Hands-On Machine Learning with Scikit-Learn and TensorFlow (book)

Machine Learning Problems

Final Points

- Beware of slow degradation as you add more data
- In larger projects with a lot of data → measuring performance may require more human resources and support e.g, crowdsourcing
- Monitor your inputs' quality
  - Going back to data quality and cleaning (e.g. random values sent by a sensor)
  - Do not trust the input you get 100%
- In online learning systems; real time; or your own research project as it evolves.
  - Retrain your models on a regular basis
  - Find new data
  - Automate / improve

Source: Hands-On Machine Learning with Scikit-Learn and TensorFlow (book)

Questions?