Lecture 11:
Assignment and Project Discussion

COSC 526: Introduction to Data Mining
Spring 2020

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE
BIG ORANGE. BIG IDEAS.®

Assignment 11

Assignment 11

- Last assignment, we performed some early stages of data downscaling using a common and powerful Python package for data analysis: Pandas.
- This week, we execute a fully developed workflow to reproduce scientific experiments using powerful machine learning methods (i.e., KNN, Surrogate-based Model, HYPPO, and Random Forest).

Setup of the SOMOSPIE workflow

This process consists of two major steps:

- creating a new virtual machine in Jetstream from a custom VM image
- cloning the SOMOSPIE GitHub repository on your new VM

Setup of the SOMOSPIE workflow

This process consists of two major steps:

- creating a new virtual machine in Jetstream from a custom VM image
- cloning the SOMOSPIE GitHub repository on your new VM

Setup of the SOMOSPIE workflow

Your assignment is to:

- Setup and run the full SOMOSPIE workflow
- Use SOMOSPIE to partially reproduce a published study on one region with multiple machine learning methods
- Use SOMOSPIE to downscale soil moisture using one method across multiple regions
- Use Pandas to analyze the reported accuracies ($R^2$and RMSE) of the predictions across multiple regions and months

You will report the results of your experiments within this Notebook.

Do not push a copy of the SOMOSPIE repo in your class repo!

SOMOSPIE Setup

Step 1: Create a new VM

- In Jetstream (https://use.jetstream-cloud.org), launch an m1.medium Virtual Machine (VM) using the image titled "Ubuntu 18.04 for SOMOSPIE" (version 1.1). Consult the JetstreamGuide.ipynb for instructions on launching a VM.

SOMOSPIE Setup

Step 2: Clone the repo

- **Inside your new virtual machine**, clone the SOMOSPIE repository:

```
git clone –recursive https://github.com/TauferLab/SOMOSPIE
```

SOMOSPIE Setup

Step 3: Set up environment

- Before opening the SOMOSPIE notebook, we need to update the .bashrc:

```
cd SOMOSPIE
make bash
source ~/.bashrc
```

SOMOSPIE Setup

Step 4: Load data and run simple test

- Now open the notebook SOMOSPIE.ipynb and in the *Cell* menu select *Run All*

IMPORTANT: The first time you run it may take 45 minutes. The reason is because you are loading the data from public datasets.

Project: Status

Project

Motivation Describe the motivation of your work. To build the motivation, you can answer these questions:

- What is the problem you are tackling?
- How is the problem solved today?

Contributions List between 2 and 4 contributions of your work. Contributions are bullet points that define your solution. E.g.,

- We build a system that ....
- We validate the system accuracy by ....
- We measure the performance of the system by ...
- Write a section of 150 - 200 words

Project

Tests List the type of tests (measurements) you will perform. E.g.,

- What are your metrics of success?
- Where do you run your tests?
- What tests do you perform?
- How many times do you run each test?
- What do you measure?
- Write a section of 250 - 350 words.

Milestones

- March 26: **Define your project (DONE)**
  - Describe the motivation of your work
  - List between 2 and 4 contributions of your work
  - List the type of tests (measurements) you will perform
- April 2: **No lecture (DONE)**

Milestones (TODAY)

- April 9: **Create a new notebook with your solution**
  - Write down the steps of your solution in distinct text cells; add one or multiple cells (as needed) to hold your code for each step. You can leave these software cells empty for the moment. Expand the text cells describing your solution.
  - Add visualization cells that allow you to visualize results. You can leave these software cells empty for the moment.
  - Add software to the code cells that upload data from source and pre-process data.
  - Push your notebook into your GitHub repository as frequently as needed

Milestones

- April 16: **Finalize software and complete the test run within your notebook**
- April 23: **Create your poster and get feedback, submit draft**
- **April 30: Submit your final notebook and poster in GitHub**
- **May 7: Submit your 2-page abstract in GitHub and video in Youtube**

How to create a story for a poster (or a paper)

How to create a research story

- Build a set of ppt slides (use template provided) that summarize your work; use text slides to tell the story of your project and figures with the key results of your work.
- Make sure your slides include: motivation and problem definition, related work and background, your methodology (e.g., with flowcharts and code sections), your results, summary, and conclusions.
- Copy and paste your slides into the poster template.
- Shuffle as needed, extend and fill gaps, embellish fonts and text, enlarge text and figures to make them readable.

Content

- Goal: present research results
- Think of the audience!
  - Knows foundation of e.g., bioinformatics
  - Does not know specific research
- Focus on concepts, not details!
- Point out advantages/disadvantages

Organization of a Talk

Title

Title of the Talk
Name
Affiliation
etc.

Splash

Fancy Picture
or
Fancy Quote

Summary

- important problem
- our solution is good
- techniques/methodology

most important

Overview

1. the problem
2. related work
3. notations
4. our solution
5. experiments
6. conclusion

Problem

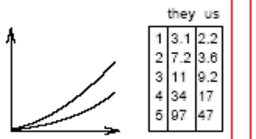
State-of-the-Art

Preliminaries

Solution

important

Tables



important

Conclusion

- summary
- outlook

most important

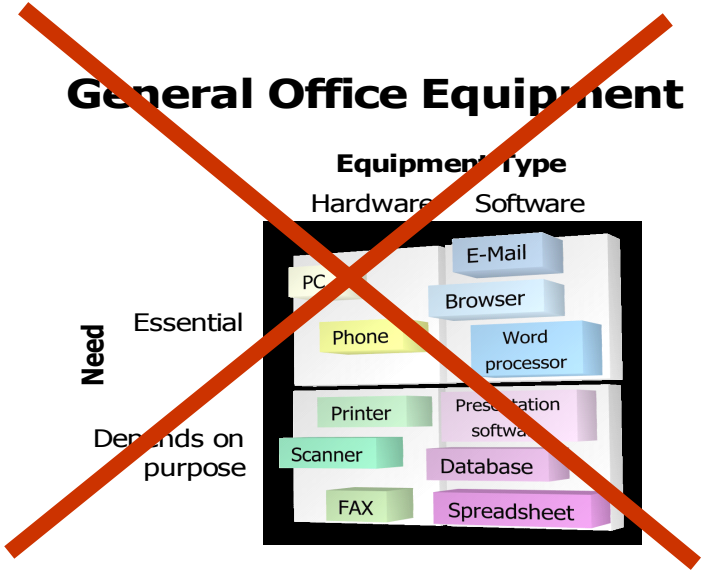
Giving the Talk

- Allow at least two minutes per slide
  - Not enough time? Cut it!
- Keep introduction and motivation short
- Use less than a minute on outline
- Use time for complex topics instead
  - Repeat if really important
  - Use examples

Slide Layout

- Keep it simple!
- Use Diagrams
- Try to keep it in one line
  - Many topics cannot be explained fully with a few bullets. In such situations, one is frequently tempted to write a full definition of the problem, usually in a very small font, which is read aloud while looking at the slide when giving the talk. Instead of this, a diagram together with a few key words would always be more suited to showing the context, which can then be elaborated on in more detail orally, in the talk.

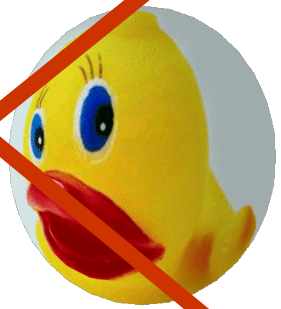
Diagrams



TOO FANCY

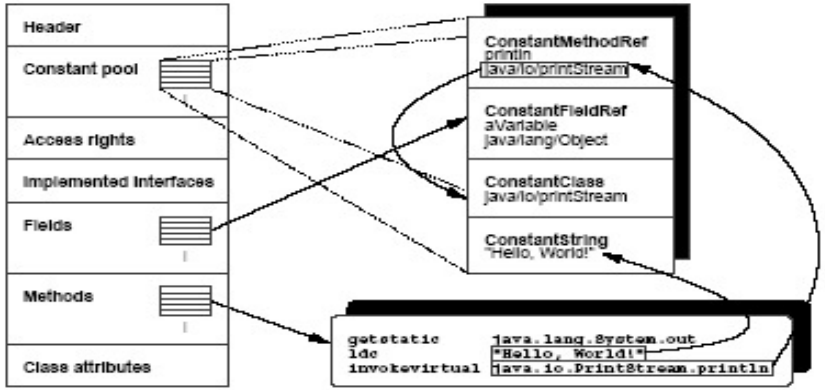
Ugly Duckling Syndrome

- Feathers
  - Stubby
  - Brown
- Told to leave town
- Low self esteem
- Does as is told



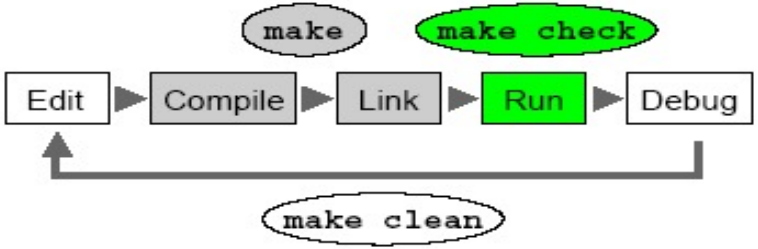
TOO ANNOYING

Diagrams

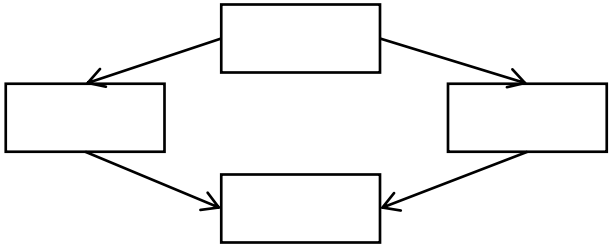


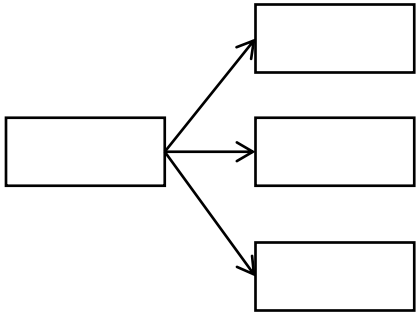

```
private static Method helloifyMethod(Method m) {
    /* Create instruction list to be
       inserted at method start.
     */
    String mesg = "Hello, World!";
    InstructionList patch  = new InstructionList();
    patch.append(new GETSTATIC(out));
    patch.append(new PUSH(cp, mesg));
    patch.append(new INVOKEVIRTUAL(println));
    MethodGen          mg  =
        new MethodGen(m, class_name, cp);
    InstructionList    il  =
        mg.getInstructionList();
    InstructionHandle[] ihs =
        il.getInstructionHandles();
    il.insert(ihs[0], patch);
    m = mg.getMethod();

    return m;
}
```

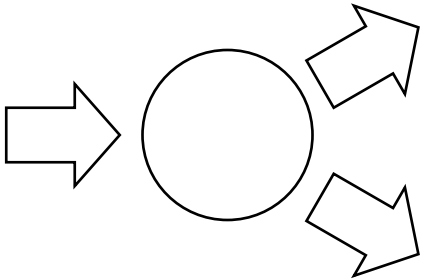
Think about the people in the back row!

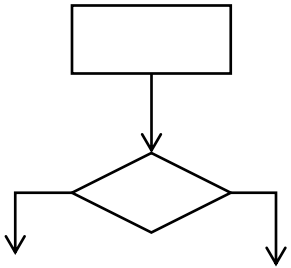


Structural Figures









Better than long text

Formulas

- Use only when needed
- Use descriptive variable names

$$velocity = \frac{distance}{time}$$

- Powerpoint equation editor is bad
  - Consider Latex + Acrobat Reader
  - Consider Texpoint

Colors

- Avoid too many colors
- LCD ≠ Projector
- Background colors:

  Yellow, orange, green, light blue

- Text colors:

  Blue, magenta, red, dark green

- Avoid distracting bitmap backgrounds

Project progress

Project Titles

- Levente Dojcsak: Predicting the Development of CKD and Identifying Preventative Treatments
- Konstantinos Georgiou: Analyzing and predicting bottlenecks on the distribution of COVID-19 Vaccines
- Michael Wermert: Stock Predictor: Using Machine Learning to Predict Stock Market Behavior

Project Titles

- Georgia Channing: Traces in the Noise: Identifying Invalid Vessel Paths
- Anuj Gautam: Add a symbol and try again: A comprehensive study of password policies
- Mirka Mandich & Jake Maeker: Machine Learning Applied to HIT-SI Spheromak Data
- Carter White: Impact of Champion Selection on League of Legends Rank

Project Titles

- Xinlan Jia & Candice Chen: US Airbnb Price Prediction
- Zhixiu Lu: Using Microbiomes to Predict Environmental Factors via Machine Learning Approaches
- Azarang Asadi: Motor control quantification using lower-limb body kinematics
- Tommy Wong: Applications of Variational Autoencoders in Analyzing Ferroelectric Domains

Project Titles

- Jerome Kovoor and Shree Neupane: Analyzing the effect of climate change on global food production using K-means clustering or DBSCAN
- Fabian Fallas Moya: How many are good enough? Finding the best number of annotated images into a self-training algorithm
- Gerald Jones: Subgroups and Factors of ESRD

Live Chat

Live Chat

- GTC 2015 Keynote with Dr. Andrew Ng, Baidu
  https://video.ibm.com/recorded/60113824