Introduction to Bioinformatics II

BCB 5250 Introduction to Bioinformatics II

Spring 2020

Tae-Hyuk (Ted) Ahn
Department of Computer Science
Program of Bioinformatics and Computational Biology
Saint Louis University

SAINT LOUIS
UNIVERSITY
EST. 1818

Welcome

- Welcome to **Introduction to Bioinformatics II**

- Tell me about you!
  - What is your name, major, and year?

Survey & Syllabus

- Fill our the survey form and let's go over the syllabus.

Topical Outline

- Introduce new trends of tools and environments for Bioinformatics
- Genome assembly and genome/gene annotation
- Metagenomics
- RNA-Seq Analysis
- Biological Modeling and Simulation

- Any additional topics you want to know?

Student Learning Outcomes

After successfully complete this course, students are expected to:

- Know fundamental concepts of bioinformatics
- Understand underlying basic bioinformatics algorithms
- Run bioinformatics applications and tools to study diverse and complex omics data
- Recognize how to apply different bioinformatics tools
- Understand cutting edge bioinformatics research topics
- Write pipeline scripts to automate existing applications
- Increase the ability to propose new algorithms and implement software tools
- Able to evaluate peer's research works and understand the importance of peer review process
- Study the knowledge including substantive findings, as well as theoretical and methodological contributions to a particular topic in a literature
- Practice and improve presentation skills including logical format of contents, ordered in clear manner, effective information, and so on
- Conduct a research as a project to answer or analysis of a biological problem as a group for obtaining a successful, high-quality, collaborative experience

Course Textbook and Resources

- No textbook is required, but check the optional books as references.
  - Bioinformatics and Functional Genomics 2nd or 3rd Edition (Jonathan Pevsner)
  - Bioinformatics Programming Using Python: Practical Programming for Biological Data (Mitchell Model)
  - Essential Bioinformatics 1st Edition (Jin Xiong)
  - An Introduction to Bioinformatics Algorithms (Neil C. Jones and Pavel A. Pevzner 2004)
  - Bioinformatics Algorithms: an Active Learning Approach 3rd Edition (Pavel A. Pevzner and Phillip Compeau)
  - Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology (Dan Gusfield 1997)
  - Genome Scale Algorithm Design (Mäkinen, Belazzougui, Cunial, Tomescu 2015)
  - Biological Sequence Analysis (Durbin, Eddy, Krogh, Mitchinson 1998)

Rosalind

- Students can enroll in the class with this link: http://rosalind.info/classes/enroll/c023b4d3bf/. It is free to register it. We will use the Rosalind for solving some programming assignments.
- Solve the first problem (20 mins)
  - Any language is OK (but I will mainly use Python3)

Terminal, Linux, and Shell

- How many of you are familiar to Terminal, Linux, and Shell Programming?
- After I hear more about your background, I will decide what resources will be provided and used in our class.

Grading

- **Project** (40%)
  - Midterm Proposal Presentation and Midterm Report (15%)
    - Contribution
  - Final Project Presentation and Final Report (25%)
    - Report should include below sections:
      - Abstract, Introduction, Methods, Results, Discussion, References, Contribution
- **Lab and Homework Assignments** (40%)
  - There will be labs and assignments in a variety of format such as (but not exclusively) in-class labs and take-home assignments.
  - I usually give you a week time frame to submit the homework. Check the due in the online system.
- **Literature Review and Presentation** (15%)
  - Submit literature review report (one paragraph or less than 1 page) for each reference paper with your critiques.
  - Present assigned reference paper and Q&A.
- **Attendance and Class Activities** (5%)

Literature Review

- Example 1:

Strategies to improve reference database for soil microbiomes

This paper is about the development of reference database for soil metagenomics study. It emphasises about why a reference database specific to soil is necessary and how RefSoil is the first step towards it. RefSeq is developed from curating the genomic data obtained from the cultured representative originating from soil. Obligate host-associated pathogens and extremphiles are not included. 11 Refeq associated phyla whose genetic information was available was excluded because of the difficulty in their culture in soil or their absence in soil.

The paper shows that RefSeq is not representative of the soil genome and provides suggestions regarding how the reference database for soil can be further developed. Samples that are represented unfairly in RefSeq can be identified by comparing with other soil resources like those from Earth Microbiome project for further curation of the database. By targeting the genomes that are abundant in EMP but not represented in RefSeq, culture and other techniques can be developed for functional annotation. Also by incorporating data from the single cell genomics, the RefSoil can be further closer to representing microbial soil diversity.

Some of the issues I found with the paper are explained below.

1. The database formed from subset of NCBI database of cultured organism is very under representative of all the possible genome in soil. It may be less crucial that RefSeq does not contain information about uncultured organism but it completely ignores the eukaryota domain of life which includes fungi and protists which are abundantly available in soil and of whose genomic information is easily accessible. It does not include any information on the viral genome as well. These aspects are not discussed in paper as a strategy to improve the soil reference database.

Literature Review

- Example 2:

Review: Choi, J., et al. (2017). Strategies to improve reference databases for soil microbiomes. *ISME Journal*, *11*(4), 829–834. https://doi.org/10.1038/ismej.2016.168

I found this paper to be a straight forward read, the authors set out to develop a reference data base for genomes derived from soil microbes (i.e. *RefSoil*). The authors present compelling examples of the use of the data base to current applications (e.g. the Earth Microbiome Project; (Thompson et al., 2017), allowing for OTUs generated from 16s amplicon sequencing to be identified to a reference genome. From the reference researchers can probe the genes that allow these microbes to inhabit the environments they are found, such as those in extreme environments. The utilization of single cell genomics is a perfect fit for this database with this technology it is possible avoid the time extensive and difficult process of cultivating each microbe before sequencing it genome. The drawback to this design is that when a microbe's culture requirements are identified it can allow for study's aimed at understanding how the microbe reacts to environmental stimuli, such as raising temperature as expected from climate change. I found the last point made, integration/compare of *RefSoil* to other large microbe genome databases (e.g. Human Microbiome Project and RefSeq), by the authors in the conclusion section to be weak. I would think that many of the properties (e.g. fixing nitrogen, methanogenesis, production of siderophores, solubilizing inorganic phosphates, etc) that are of the highest interest to humanity as a whole in soil microbes to not be shared with microbes that inhabit the gut and surfaces of humans.

Of particular interest to me in the paper was the identification of OTUs from the Earth Microbiome Project to sequences currently in the *RefSoil* database. This allowed the authors to generate a list of microbes with high abundances and that are common to many soils types but currently have no reference genome (at least not one with <97% similarity in the 16s rRNA region). These are the low hanging fruit, the authors had chosen the moniker of "most wanted" for these microbe genomes. Although they are keen to use the Earth Microbiome Project and identify the microbes found in many samples. The authors also utilized the metadata from the Earth Microbiome Project to understand how the soil properties effected the distribution of microbes. I felt the analysis aimed at understanding soil environmental patterns (e.g. pH, sand/silt/clay content, etc.) only scratched the surface. For instance consider microbes in high abundance in only high or low pH soils which are ignored because there maybe few soil samples in the Earth Microbiome Project with these soil properties. Thus although it will be important to understand microbes ubiquitous to many soil types, we have to understand that many extremophile microbes that are vital to the function of extreme environments will be missed. It is these microbes that might process unique functions and gene that we could utilize for bioremediations of human altered landscapes, such as mine drainages.

Another paper that is taking aim at understanding microbe and environment patterns global is the recent work by Delgado-baquerizo et al. (2018). I feel this project is taking a better approach to building a most wanted list, similar to the paper above they are looking for microbes of high abundance but are taking into account the environmental preference of the microbes. Thus from their work they can better understand the environmental preferences that govern the microbial communities in a given environment. These environmental preference will also be key in developing techniques to culture the microbes for genomic and future studies.

Literature cited

Delgado-baquerizo, M., Oliverio, A. M., Brewer, T. E., Benavent-gonzález, A., Eldridge, D. J., Bardgett, R. D., … Fierer, N. (2018). A global atlas of the dominant bacteria found in soil. *Science*, *359*(6373), 320–325. https://doi.org/10.1126/science.aap9516

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., … Zhao, H. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*. https://doi.org/10.1038/nature24621

Literature Review Presentation

- Mostly every week Thursday from 1/24/2019.
- I will post the review paper on Tuesday a week ahead.
- Try to understand most of contents, but don't get too much stress to fully understand everything. Authors prepare 1-3 years to write a manuscript, and it is very natural to not understand everything. However, try to catch important concepts, methods, results, and some points that should be discussed together.
- Literature review presentation rubric will be posted soon.
- Now, it's time to decide the order of the presentations. How?

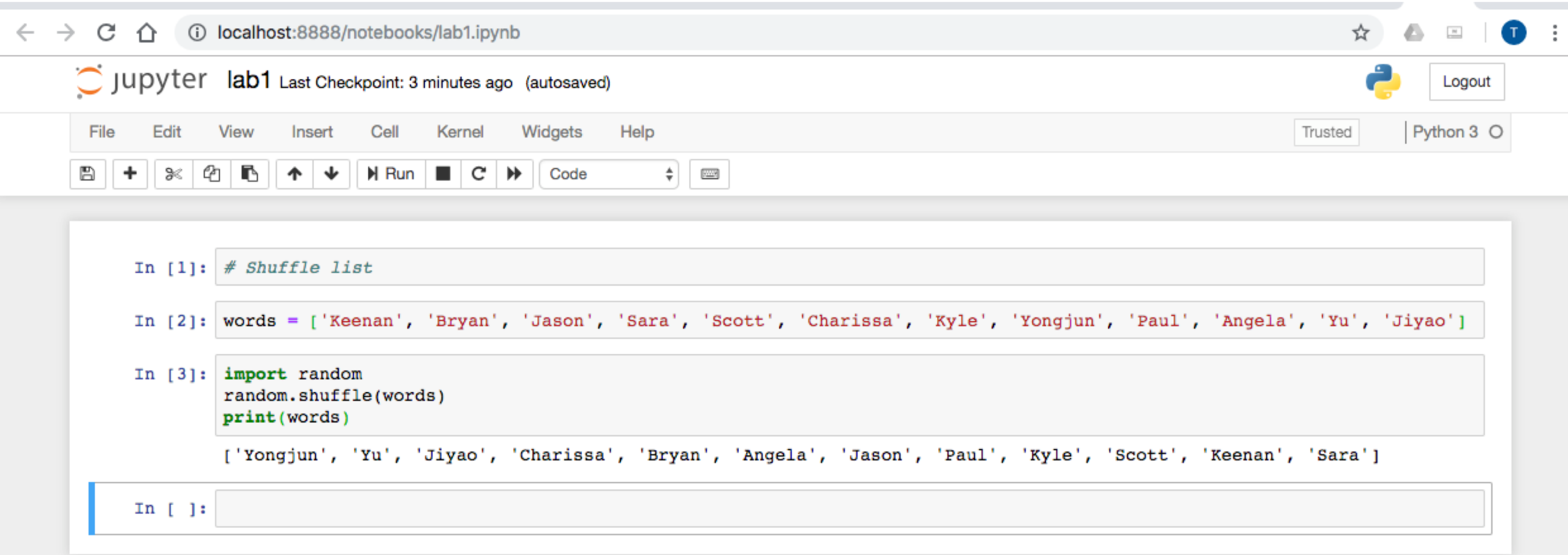
Lab1: Make a Python Program

- Get the list of the names of the students.
- Shuffle a list of strings order to get the presentation list!

Student Name
Ahrens, David P.
Debruin, David E.
Gardner, Cory
Hadfield, Christina M.
Li, Huan
Liu, Wanxiang
Ma, Feiya
Mreyoud, Yassin S.
Peasari, John Reddy
Tahan, Stephen M.
Tran, Peter
Zhang, Yujing

HW1: Run the Lab1 using Jupyter Notebook

- Run the lab1.py using Jupyter Notebook.
- Save the file as lab1.ipynb



Jupyter Notebook

- I strongly recommend you to install Jupyter Notebook in your laptop or your desktop.
- https://jupyter.org/
- If you cannot get started, watch the YouTube video “Jupyter Notebook Tutorial: Introduction, Setup, and Walkthrough” (https://www.youtube.com/watch?v=HW29067qVWk)

1. What is the Jupyter Notebook?

In this page briefly introduce the main components of the *Jupyter Notebook* environment. For a more complete overview see References.

Contents

1.1. Notebook document

Notebook documents (or “notebooks”, all lower case) are documents produced by the Jupyter Notebook App, which contain both computer code (e.g. python) and rich text elements (paragraph, equations, figures, links, etc...). Notebook documents are both human-readable documents containing the analysis description and the results (figures, tables, etc..) as well as executable documents which can be run to perform data analysis.

References: Notebook documents in the project homepage and in the official docs.

1.2. Jupyter Notebook App

The *Jupyter Notebook App* is a server-client application that allows editing and running notebook documents via a web browser. The *Jupyter Notebook App* can be executed on a local desktop requiring no internet access (as described in this document) or can be installed on a remote server and accessed through the internet.

Jupyter Lab

jupyterlab.readthedocs.io/en/stable/

JupyterLab

stable

Search docs

GETTING STARTED

- Overview
- Installation
- Starting JupyterLab
- Reporting an issue
- Frequently Asked Questions (FAQ)
- JupyterLab Changelog

USER GUIDE

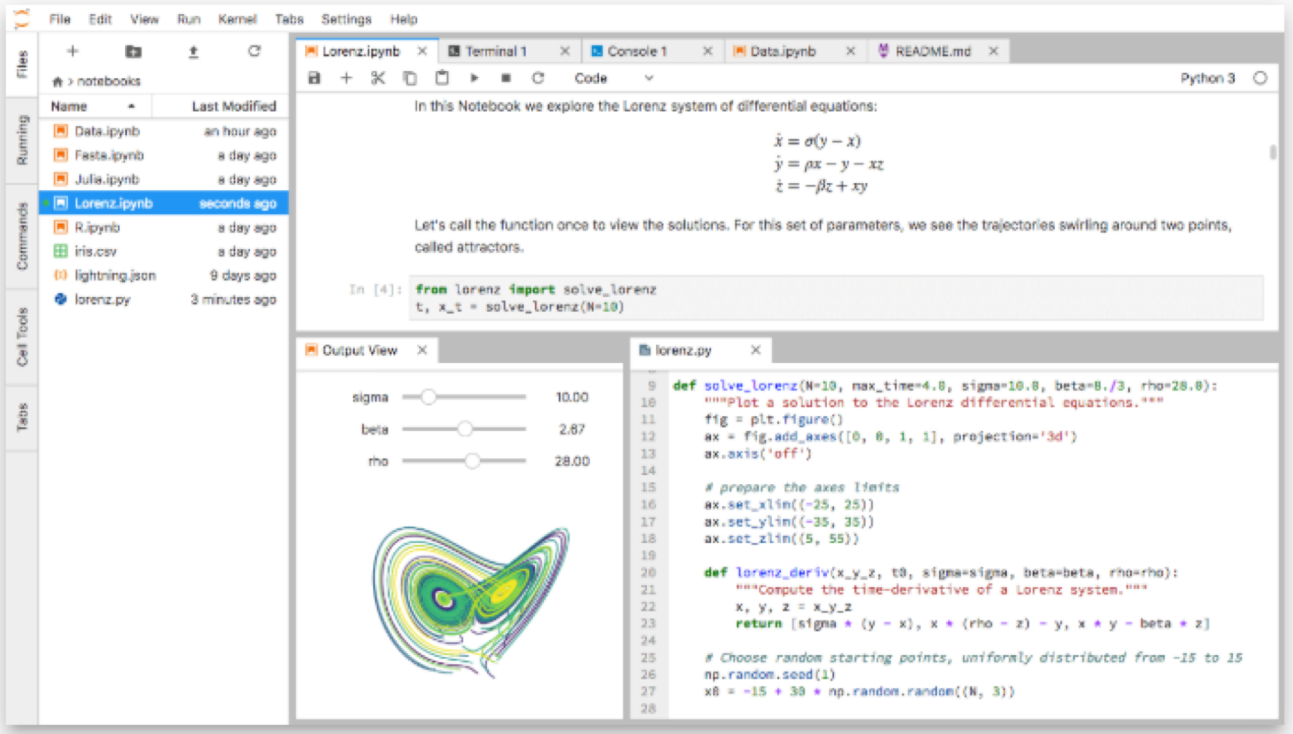
- The JupyterLab Interface
- JupyterLab URLs
- Working with Files
- Text Editor
- Notebooks
- Code Consoles
- Terminals
- Managing Kernels and Terminals
- Command Palette

Docs » JupyterLab Documentation

Jupyter | Edit on GitHub

JupyterLab Documentation

JupyterLab is the next-generation web-based user interface for Project Jupyter. Try it on Binder. JupyterLab follows the Jupyter Community Guides.

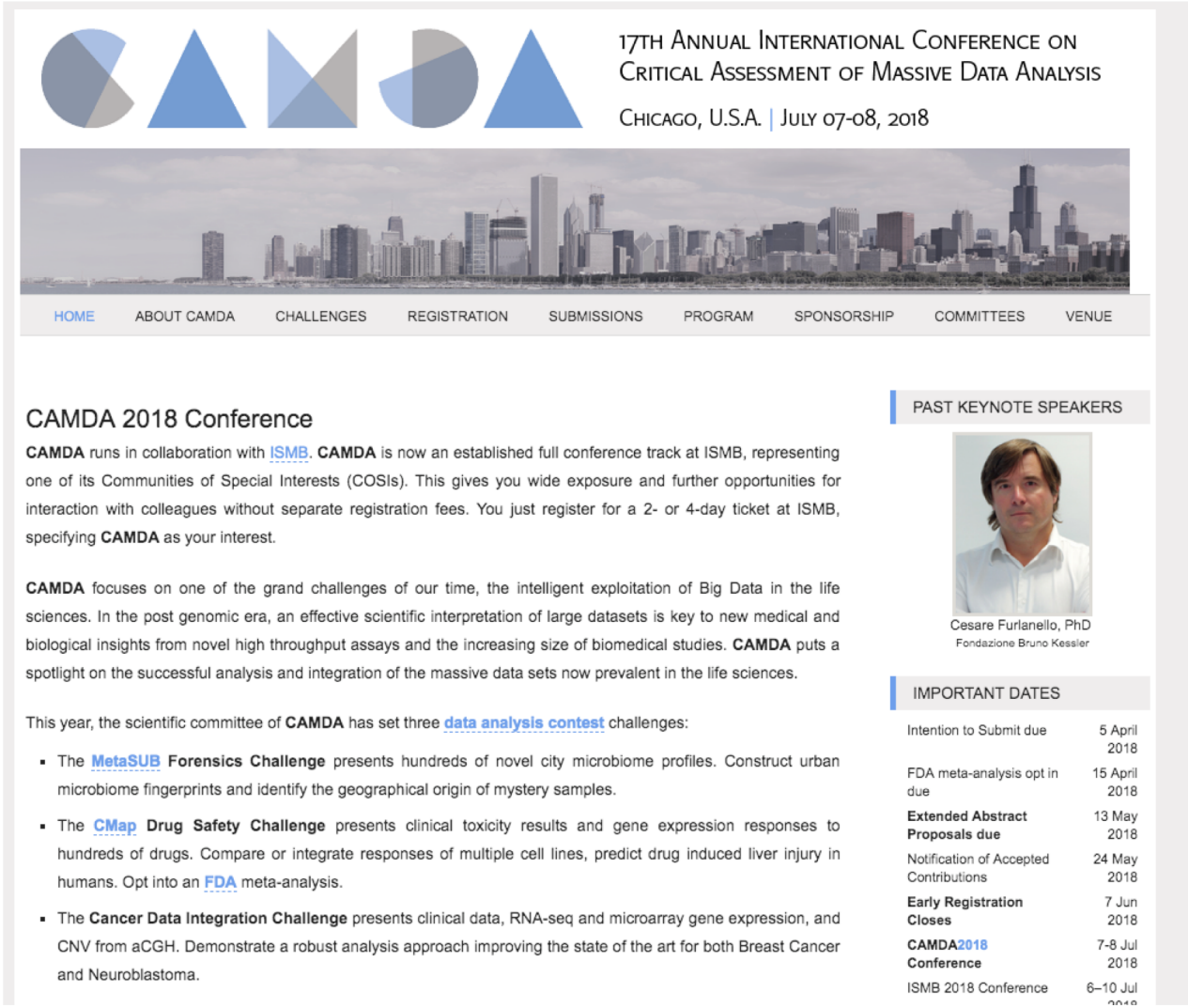


Project and Presentation

- We need 4 check points.
  - End of Jan: Decide group and project idea (no scoring)
  - Close to Midterm: Project - Midterm Presentation
  - Sometime in Apr: Report the summary of progress (no scoring)
  - Close to Final: Project - Final Presentation
- Midterm Proposal Presentation and Midterm Report (15%)
- Final Project Presentation and Final Report (25%)

Term Project Environments

- I will assign your groups (4 groups) very soon
- Last year, we had 4 groups.
  - One groups worked CAMDA 2019 MetaSub challenge project and presented their work in the conferences.
- Let's discuss together. I am planning to do a horseweed project using genome assembly, gene annotation, and RNA-Seq research, and genome browser with Oxford NanoPore MinION.



Internship

- It is time for you to apply your summer internships.
- Check https://www.bioinformatics.org/ and https://www.iscb.org/
- Check local opportunites.
  - Danforth Plant Science Lab, Monsanto (Bayer, now), Nestle-Purina, MOGene, Pfizer, Cofactor Genomics, BioSTL, Sigma-Aldrich, Saint Louis University School of Medicine, Washington University in St. Louis School of Medicine
  - Check the partners at http://bioinformatics.slu.edu/partners.html

Welcome to Bioinformatics and BCB program!

- Welcome!!!
- Lots of opportunities for you!!!