Week 0_Introduction and Course Information

This 10-week self-paced course is part of the Data Science MicroMasters program and will introduce you to a collection of powerful, open-source, tools needed to analyze data and to conduct data science. Specifically, you'll learn how to use:

- Python
- Jupyter notebooks
- pandas
- NumPy
- Matplotlib
- git
- scikit-learn
- NLTK
- and many other tools!

You will learn these tools all within the context of solving compelling data science problems.

After completing this course, you'll be able to find answers within large datasets by using python tools to import data, explore it, analyze it, learn from it, visualize it, and ultimately generate easily sharable reports. You'll also be introduced to Machine Learning techniques and Natural Language Processing tools to expand your data analysis abilities (e.g., being able to analyze twitter data for user sentiments).

By learning these skills, you'll also become a member of a world-wide community which seeks to build data science tools, explore public datasets, and discuss evidence-based findings. Last but not least, this course will provide you with the foundation you need to succeed in later courses in the Data Science MicroMasters program.

Your Instructors

Ilkay Altintas is the chief data science officer at the San Diego Supercomputer Center (SDSC), UC San Diego, where she is also the founder and director for the Workflows for Data Science Center of Excellence. She received her Ph.D. degree from the University of Amsterdam in the Netherlands with an emphasis on provenance of workflow-driven collaborative science and she is currently an assistant research scientist at UCSD.

Leo Porter is an Assistant Teaching Professor at the Department of Computer Science and Engineering at the University of California, San Diego. He received his Ph.D. in Computer Science, specifically computer architecture, from UC San Diego in 2011.

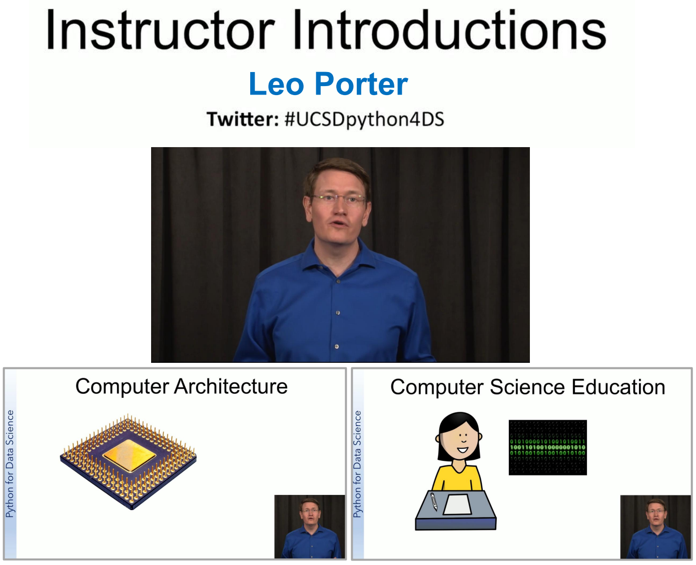
总结:

0.1 Course Introduction

... you probably didn't get interested in data sciencejust because you want to learn a bunch of data analysis tools. You probably became interested in data science because you wanted to better understand some aspect of the world, and you want to use data to do so. That's why we won't be learning these Python tools in isolation, like you might find in other courses or textbooks. For most topics, we'll have you learning these tools as part of exploring and analyzing a real-world data setting.

- Yep, that means you'll be learning data science by doing data science. And doing data science is what we are all excited about.

0.2 Instructor Introductions



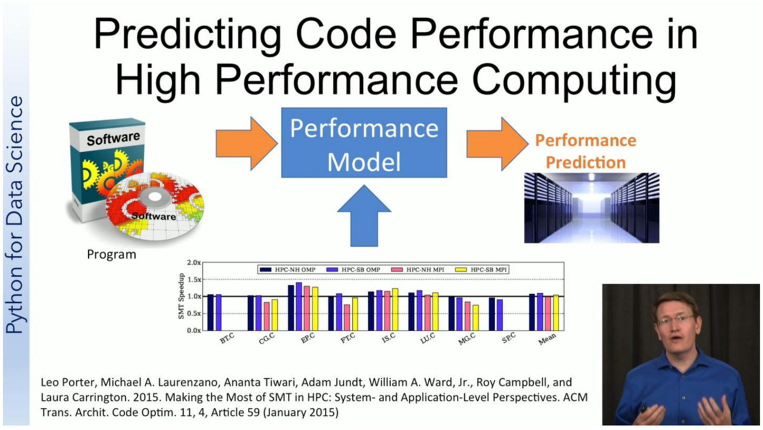
Instructor Introductions

Leo Porter

Twitter: #UCSDpython4DS

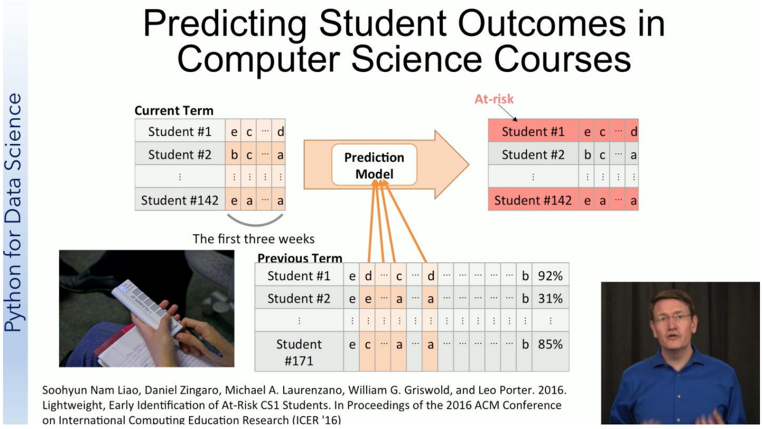
Computer Architecture

Computer Science Education



Predicting Code Performance in High Performance Computing

Leo Porter, Michael A. Laurenzano, Ananta Tiwari, Adam Jundt, William A. Ward, Jr., Roy Campbell, and Laura Carrington. 2015. Making the Most of SMT in HPC: System- and Application-Level Perspectives. ACM Trans. Archit. Code Optim. 11, 4, Article 59 (January 2015)



Predicting Student Outcomes in Computer Science Courses

Soohyun Nam Liao, Daniel Zingaro, Michael A. Laurenzano, William G. Griswold, and Leo Porter. 2016. Lightweight, Early Identification of At-Risk CS1 Students. In Proceedings of the 2016 ACM Conference on International Computing Education Research (ICER '16)

Thank you to my co-authors!

Leo Porter, Michael A. Laurenzano, Ananta Tiwari, Adam Jundt, William A. Ward, Jr., Roy Campbell, and Laura Carrington. 2015. Making the Most of SMT in HPC: System- and Application-Level Perspectives. ACM Trans. Archit. Code Optim. 11, 4, Article 59

Soohyun Nam Liao, Daniel Zingaro, Michael A. Laurenzano, William G. Griswold, and Leo Porter. 2016. Lightweight, Early Identification of At-Risk CS1 Students. In Proceedings of the 2016 ACM Conference on International Computing Education Research (ICER '16)

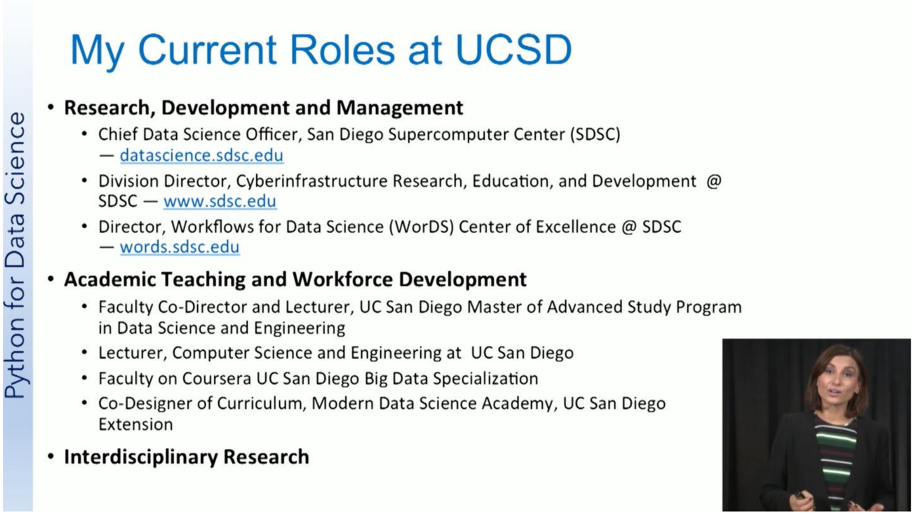
- Hi, I'm Leo Porter.And I just wanted to take this opportunity to introduce myself a bit more. I already told you that I'm an assistant teaching professor in the Computer Science and Engineering Department at UC San Diego. But I haven't told you the kind of work that I do. My background is in Computer Architecture, which is an area of computer science focused on the high-level designer processors. There, my research is aimed at making processors faster, or more power efficient, or both. These days I spend most of my time doing research on how students learn Computer Science and how to improve their learning. Most my work is focused on improving student outcomes through the adoption of an evidence-based teaching practice known as Peer Instruction. But I've also worked to identify common student misconceptions and to produce meaningful assessments of student learning. In both of these fields, I've used skills from the areas of Data Science and Machine Learning, to inform and conduct research. Let me give you two examples. One example from Computer Architecture is we wanted to be able to predict how a program might run on a high performance computing cluster given certain architectural features of the cluster. Specifically, we want to know, if it'd be better to run the software using or not using a vary specific hardware optimization. To figure this out, we took statistics from a whole bunch of programs and used Machine Learning to build a model of performance. We could then feed statistics about a single program into the model and then the model would then produce a prediction of how that program might perform on that system. It turned out that our predictions were quite accurate for the problem we were trying to solve. Similarly, in Computer Science Education, we wanted to see if we could use data to predict when students were in jeopardy of failing a course. In many of our classes at UC San Diego, we use clickers like the one you see here. Students answer questions during the term using clickers, and we wondered, could those responses tell uswho might be at risk of failing. So what we'd have is a whole term worth of clicker responses per student and how well they did on the final exam. Similar to the architecture example, we just feed that data into a machine learning model designed to predict student outcomes. Then we'd use clicker data responses from the first few weeks of a new class, feed them into the Prediction Model, and bingo, we'd see which students were in trouble. Again, the Prediction Model was quite accurate, enabling instructors to potentially intervene to help students at risks, at risk. For this work, and for the prior, I want to reinforce, the fact that Computer Science, research is often a highly collaborative effort. And I want to thank my collaborators on these projects. Also, thank you, for letting me introduce you to the kind of work that I do.

总结:

Meet Your Instructor:

Dr. Ilkay Altintas

Twitter: #UCSDpython4DS

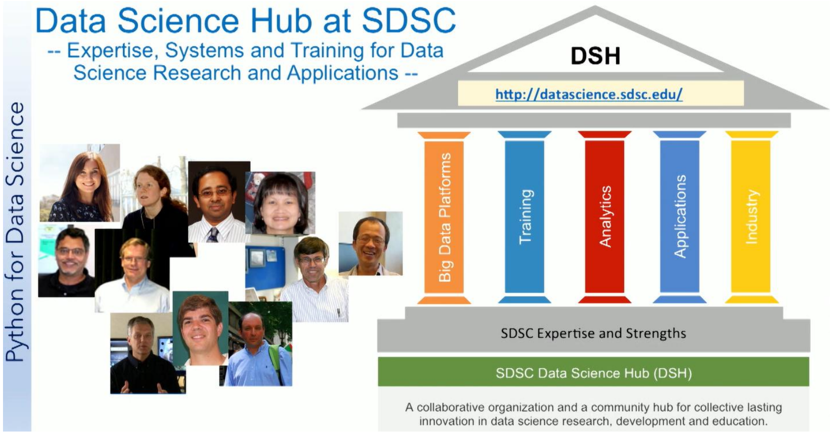


Python for Data Science

My Current Roles at UCSD

- **Research, Development and Management**
  - Chief Data Science Officer, San Diego Supercomputer Center (SDSC) — datascience.sdsc.edu
  - Division Director, Cyberinfrastructure Research, Education, and Development @ SDSC — www.sdsc.edu
  - Director, Workflows for Data Science (WorDS) Center of Excellence @ SDSC — words.sdsc.edu
- **Academic Teaching and Workforce Development**
  - Faculty Co-Director and Lecturer, UC San Diego Master of Advanced Study Program in Data Science and Engineering
  - Lecturer, Computer Science and Engineering at UC San Diego
  - Faculty on Coursera UC San Diego Big Data Specialization
  - Co-Designer of Curriculum, Modern Data Science Academy, UC San Diego Extension
- **Interdisciplinary Research**

I am Ilkay Altinas. Just like Leo, I'd like to tell you a little bit about my work here at the San Diego Supercomputer Center and UC San Diego. You might even think, what does a supercomputer center have to do with data science? Let me tell you. As you might have heard, there are many different, exciting applications that are being enabled by data science in the big data era. I am the Chief Data Science Officer at SDSC, leading our collaborative data science hub activities. At the same time, I lead our research development and education division, where I oversee many exciting research programs. And a role that I am passionate about is my research center for data science workflows, which I built over time as an area since the year 2001, which has been a joint SDSC at UC San Diego. In terms of some of my education activities, I'm the Faculty Co-Director for the Master of Advanced Studies Program on Data Science and Engineering here at UC San Diego, in which I teach a capstone project course. I also work as a lecturer for the Computer Science and Engineering Department and have done big data courses as part of other online and offline course programs. What is common to all these roles is that I work on interdisciplinary research. As a part of my core research development and teaching activities, I work on building methodologies and tools to make big data, data science, and computational science useful to dynamic, data-driven scientific applications. I work with many UC San Diego centers in these areas.



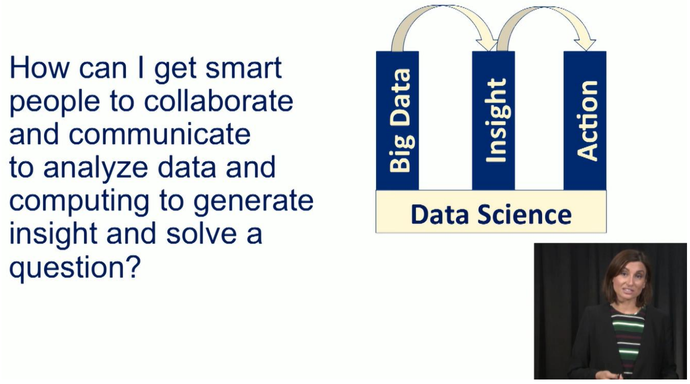
Python for Data Science

Data Science Hub at SDSC

-- Expertise, Systems and Training for Data Science Research and Applications --

DSH
http://datascience.sdsc.edu/
Big Data Platforms
Training
Analytics
Applications
Industry
SDSC Expertise and Strengths
SDSC Data Science Hub (DSH)

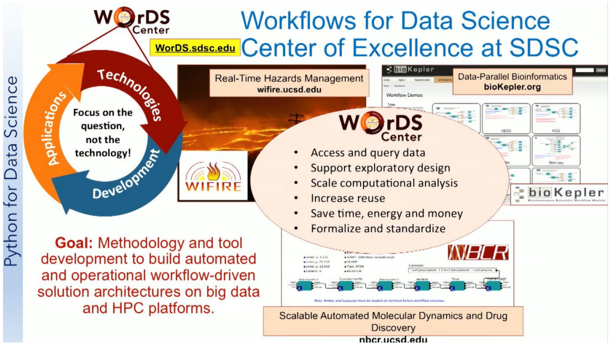
A collaborative organization and a community hub for collective lasting innovation in data science research, development and education.



How can I get smart people to collaborate and communicate to analyze data and computing to generate insight and solve a question?

My colleagues and I work on many Grand Challengedata science applications in all areas of science and engineering, including genomics, geoinformatics, metro data science or smart cities, energy management, biomedicine, personalized health, and many others as part of our data science hub. What is common to all these applications is their unique way of bringing together new modes of data and computing research. So it's all about collaboration and what the supercomputer center brings to it as the computing and data management expertise.

So what do I do on a regular day,is to think about how I can get a group of very smart people collaborate to solve data science challenges. I worry about questions like, how do we close the loop on data science methods, tools, computing and data systems, and domain experts? These questions also get translated to research challenges and toolboxes for data science by my research group. Our main goal is methodology and tool developmentto build automated and operational workflow-driven solution architectures on big data and high performance computing platforms. This applies to many scientific disciplines. Let me tell you a project we work on as a part of many of these applications. It is for wildfire analytics,which breaks up into two components: prediction and emergency response.



Python for Data Science

Workflows for Data Science Center of Excellence at SDSC

WorDS Center
WorDS.sdsc.edu
Technologies
Applications
Development
Focus on the question, not the technology!
Real-Time Hazards Management
wifire.ucsd.edu
WIFIRE
Data-Parallel Bioinformatics
bioKepler.org
bioKepler
NBCR
Scalable Automated Molecular Dynamics and Drug Discovery
nbcr.ucsd.edu

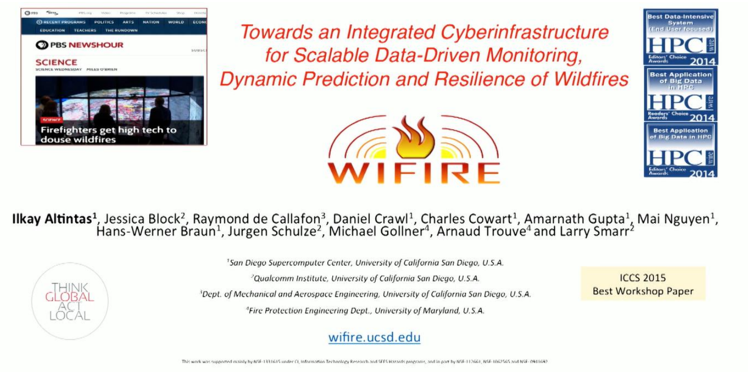
WorDS Center

- Access and query data
- Support exploratory design
- Scale computational analysis
- Increase reuse
- Save time, energy and money
- Formalize and standardize

Goal: Methodology and tool development to build automated and operational workflow-driven solution architectures on big data and HPC platforms.

Using Data Science for Wildfire Prediction and Emergency Response

总结：



PBS NEWSHOUR

SCIENCE

Firefighters get high tech to douse wildfires

Towards an Integrated Cyberinfrastructure for Scalable Data-Driven Monitoring, Dynamic Prediction and Resilience of Wildfires

WIFIRE

Ilkay Altintas[1], Jessica Block[2], Raymond de Callafon[3], Daniel Crawl[1], Charles Cowart[1], Amarnath Gupta[1], Mai Nguyen[1], Hans-Werner Braun[1], Jurgen Schulze[2], Michael Gollner[4], Arnaud Trouve[4] and Larry Smarr[2]

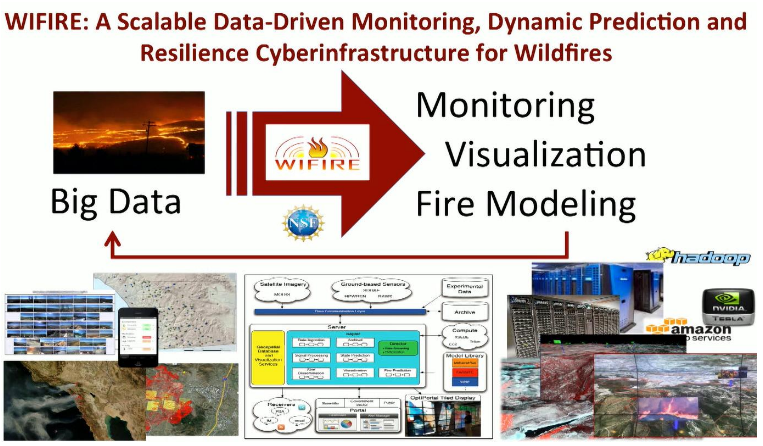
[1]San Diego Supercomputer Center, University of California San Diego, U.S.A.
[2]Qualcomm Institute, University of California San Diego, U.S.A.
[3]Dept. of Mechanical and Aerospace Engineering, University of California San Diego, U.S.A.
[4]Fire Protection Engineering Dept., University of Maryland, U.S.A.

ICCS 2015
Best Workshop Paper

wifire.ucsd.edu

This work was supported mainly by NSF-1331615 under CI, Information Technology Research and SEES Hazards programs, and in part by NSF-112661, NSF-1062565 and NSF-0941692

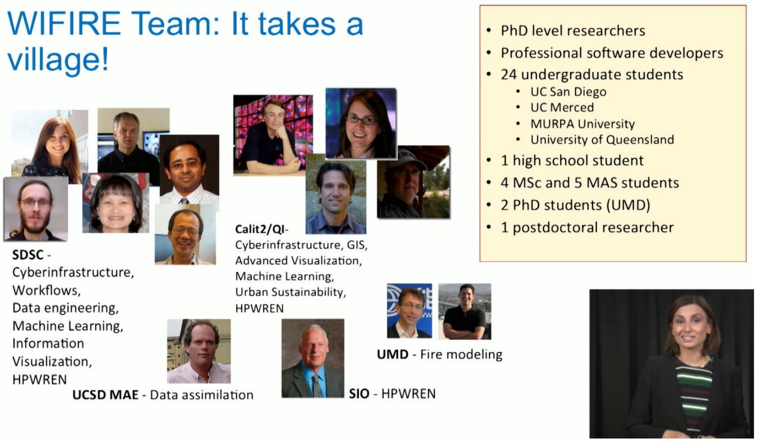
Best Data-Intensive System (End User focused)
HPC wire Editors' Choice Awards 2014
Best Application of Big Data in HPC
HPC wire Readers' Choice Awards 2014
Best Application of Big Data in HPC
HPC wire Editors' Choice Awards 2014
THINK GLOBAL ACT LOCAL



WIFIRE: A Scalable Data-Driven Monitoring, Dynamic Prediction and Resilience Cyberinfrastructure for Wildfires

WIFIRE is a collaborative projectfunded by the National Science Foundation to build a cyberinfrastructure for wildfire monitoring, prediction, and resilience. It's a research project, which lead to very useful insights. In WIFIRE, we built a scalable cyberinfrastructure that can utilize any high-end computing, cloud and big data platform, for dynamic big data fire modeling and prediction. The approach here is to use real-time data to learn about the dynamics of fire behavior and environment, and using data science techniques, assimilate what we have learned into a fire model to adopt to changes in the situation over time. Here, data science methods and workflows were used for system integration and dynamic application scalability. All the data, models, and computing systems being used for data before WIFIRE, but a programmable system integration that can match the application needs we're lacking. Data science enables such computing capabilities to become available to fire response, research, and planning communities. In fact, the system is being used as a special awareness tool by some fire departments already and we are really happy about that. But WIFIRE represents a wide range of applications where real-time big data can be assimilated with modeling and simulation tools for better situational awareness and dynamic decision support.

Wildfire Command Centers of the Future



WIFIRE Team: It takes a village!

- PhD level researchers
- Professional software developers
- 24 undergraduate students
  - UC San Diego
  - UC Merced
  - MURPA University
  - University of Queensland
- 1 high school student
- 4 MSc and 5 MAS students
- 2 PhD students (UMD)
- 1 postdoctoral researcher

SDSC - Cyberinfrastructure, Workflows, Data engineering, Machine Learning, Information Visualization, HPWREN

Calit2/QI- Cyberinfrastructure, GIS, Advanced Visualization, Machine Learning, Urban Sustainability, HPWREN

UMD - Fire modeling

UCSD MAE - Data assimilation

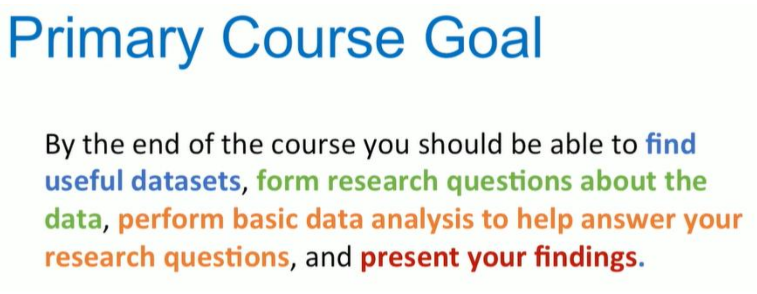
SIO - HPWREN

Just take a moment to imagine how data sciencewill help with firefighting in the future. Many streams of data will come together in 3-D displays that can show all the related information along with weather and fire predictions. That would be a great application of data science with societal impact.

Needless to say, none of this would be possiblewithout the collaboration of many individuals, showing how important interdisciplinary collaboration is to data science. Hope you enjoyed hearing about what I do. And I'm looking forward to sharing some exciting data science use cases with you over the next 10 weeks.

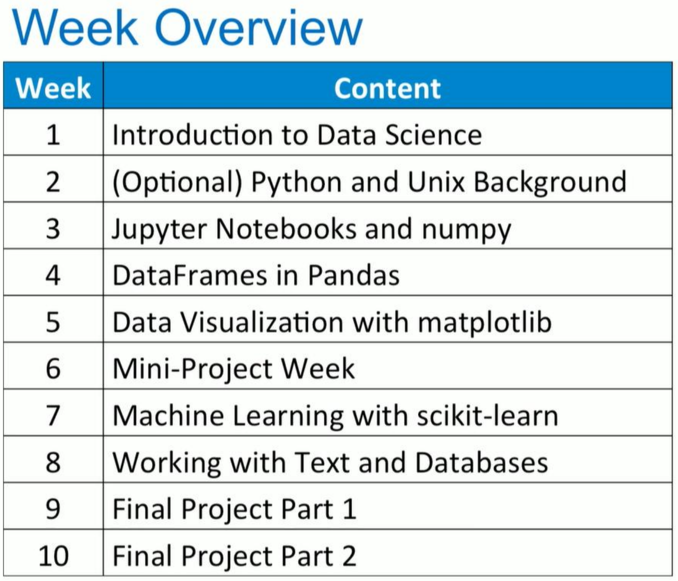
总结：

0.3 Syllabus and Course Information



Primary Course Goal

By the end of the course you should be able to **find useful datasets**, **form research questions about the data**, **perform basic data analysis to help answer your research questions**, and **present your findings**.

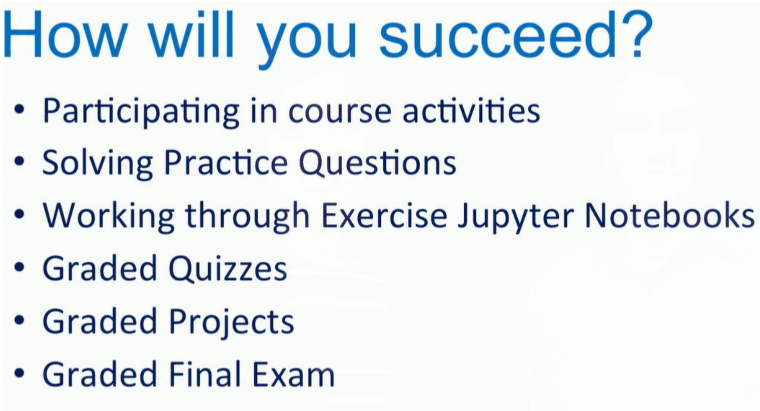


Week Overview

Week	Content
1	Introduction to Data Science
2	(Optional) Python and Unix Background
3	Jupyter Notebooks and numpy
4	DataFrames in Pandas
5	Data Visualization with matplotlib
6	Mini-Project Week
7	Machine Learning with scikit-learn
8	Working with Text and Databases
9	Final Project Part 1
10	Final Project Part 2

What we'd like to do nextis just give you an overview of the course and its features. More details about things like grading and the honor code will appear in the syllabus readings next. We have one primary goal for you for this course but it's a big goal. We want you to be able to by the end of the course to find an open dataset and explore it using the tools we'll learn in this class, either before the exploration or during, develop a meaningful research question which could be answered by the dataset. Then, further explore the data or analyze it to find an answer to your research question. Based on what you find you should be able to accurately present your results. This might seem daunting now but if you follow through the course, you should be well prepared to conduct this kind of data analysis. In fact, your final project asks you to do exactly this.

- To see how we'll prepare you, let's take a look at what you'll be doing in the course.
Here's our weekly overview.We'll get started in week one with an introduction to the field of data science and big data. Then, we'll have an optional week. If your background is in a language other than Python or you don't have much experience with Unix, this week will give you the background you need to succeed in the course. The next two weeks are all about manipulating and working with data in Jupyter Notebooks. Between Numpy and Pandas you'll be able to read in data, clean and organize the data, and explore the data. Week five is all about data visualization. We'll introduce you to some of the core concepts in the filed of data visualization and show you how to produce visualizations in your Notebooks using matplotlib. By week six you'll already know a lot about how to analyze data and how to present the results. So, the next step is to spend some time doing data analysis yourself in a small Jupyter Notebook project with the dataset we've already shown you. The next two weeks we'll introduce you to more advanced topics in data science. The first focuses on machine learning and the library scikit-learn. In the next week you'll be diving into working with text from the web, databases and some basic natural language processing using the natural language processing toolkit. The last two weeks are gonna bring everything together in that final project that I just mentioned.



How will you succeed?

- Participating in course activities
- Solving Practice Questions
- Working through Exercise Jupyter Notebooks
- Graded Quizzes
- Graded Projects
- Graded Final Exam

We value your feedback

- So, how will you succeed?First, we really want to see you finish the course. We love data science and hope you'll learn to appreciate the field as much as we do. We also know that many of you have competing demands for your time. To help you keep on track and on pace to succeed, there are a variety of activities throughout the course, including videos, discussions and practice questions. We'll give you credit for completing the course for this effort to help motivate you to make progress. We also want to give you practice working in Jupyter Notebooks, so you'll find exercise Notebooks which have built-in tests to help you get feedback on your code. At the end of most weeks, you'll find a graded quiz which will be based on the practice questions and exercise Notebooks. If you've been making progress in the course, you should succeed in these quizzes but because everyone has a bad day sometimes, you get to drop your lowest quiz. Leo has already talked about the projects and those are critical for you to test your new skills. Lastly, once you've finished everything else in the class, you'll have a final exam to gain feedback on your final understanding of the course. If you participate in all the activities in the course, we are confident you can succeed. - One last key point. We want your feedback and we'll ask for it periodically. We both have experience building online courses and we know there's always wrinkles in a course after it first launches. We're committing to remedying any issues you discover and to getting you the help you need if you get stuck, so please don't hesitate to give us feedback. We want to thank you for joining the course and please look through the next readings for more syllabus details.

总结:

Excelling with Integrity

You are in this course because you want to learn and we want to do everything we can to help you learn. In order for that to happen, you need to do your own work and not help other people do work they should be doing.

As we tell our in-person students, focusing on course grades and course credit is short-sighted. We take courses to learn new skills and to learn new ways of thinking about the world. Yes, you get grades for your work in a course. But grades and course credit just helps open doors for you - with these grades comes the expectation you know the course material. When it comes time for you to use your knowledge and skills, you'll want the pride and personal confidence of knowing you did the work yourself.

We care a lot of about this and so should you. Please be sure you read the edX terms of service agreement. If you're unclear about what is permitted and what is not, check out the agreement and if still in doubt, just ask.

Remember, when you are posting in discussion forums we expect you to observe common rules of etiquette (see a visual here). However, here's a few basic categories of etiquette to keep in mind.

- Respect the challenges of written communication (e.g., don't use all CAPS, avoid abbreviations, be careful with jokes, don't say things in writing you wouldn't say in person (e.g., call someone stupid), avoid posting when you are angry)
- Make your post valuable for others (e.g., keep it as short as possible, use the simplest language you can (not everyone speaks English as a first language), stay on topic, be accurate (or state that this is your opinion), before posting a question - check to see if someone else has already asked it, if many people respond to your post - provide a summary).
- Be respectful of others (e.g. respect the opinion of your classmates, be open to others with differing viewpoints, respect the differences in culture and experience among learners in the course).

Optional: Feedback

We care about your learning!

This online course is new for us and for UC San Diego. We believe we are providing you with world-class materials and resources to learn data science or we wouldn't have created the course. But, being experienced educators, we know things often go wrong in new courses. We value your patience as we resolve issues that come up and we promise we'll do our best to resolve them both fairly and quickly.

If you encounter problems, please post on the forums and fill in the surveys which occur periodically in the course. We take your feedback seriously and will use it to improve this, and future iterations, of this course.

This is all to say: Thank you for your patience and feedback!

Self-Paced == Under Construction

Unlike in-person or synchronous classes, not everyone will be at the same place at the same time. With our in-person classes, we just take notes every term on how to make things better for students next time around. Here, we get the benefit of being able to make changes as soon as we get feedback or identify the need. This means you might go back to a prior week and find a whole new practice exercise or video there, and that's okay.

We'll try our best to not let these changes interfere with learners who have already advanced past that point in the course while improving the experience for learners who haven't gotten there yet. Again, thank you for your patience.

总结：

0.4 Succeeding in this Course

Tips

We really want you to succeed and finish this course!

To help with that, we have a lot of experience working with learners online and hope to share some of the best practices we've seen from the data and which have been reported back to us from learners.

Here are some tips for you to help you complete the course:

1. **Find friends or colleagues to work through the course with you.** For our in-person classes, there is a lot of social accountability with being enrolled in a class. Students attend class together and study together. While we know this isn't always possible for online learners, particularly those with already busy schedules, please explore this option if possible.

2. **Set aside time to work on the class.** We've all had the best intentions to start a new hobby, fitness regiment, or learn a new field. One of the best ways to ensure you start, and stick, with the new practice is to set aside time for it on your calendar and hold yourself accountable to work at that time.

3. **Participate in the discussion forums.** We have a number of discussion prompts during the class. We encourage you to participate. This helps you engage with the material in an active way and gain an increased sense of social belonging.

4. **Install the coding platform early.** Learners often stumble at the first point in the course when coding is required. So by the end of this week, be sure to follow our instructions on setting up Jupyter to run on your machine. This leads to....

5. **Play with the code.** Throughout the course, we'll be walking through notebooks we've provided you. Please work along with us and feel free to make changes, try out new ideas, and just generally play around with the code.

6. **Sign up for a certificate.** It's easy -- just click here! This course is new, so we don't have data for it. But for one of our other online courses, learners who sign up for the certificate are over 4x more likely to complete the course learners who do not.* Also remember, this course is the first in our MicroMasters in Data Science. We'll tell you more about it, but you can earn credit toward a real Masters degree!

* Comparing learners who sign up for the certificate against learners who sign up to audit AND complete at least one graded assignment, certificate learners were over 4x more likely to finish the course.

总结：

Certificate

HEY, WAIT! THIS COURSE'S VERIFIED CERTIFICATE IS PRETTY EXPENSIVE!

It's true -- and you'll find that to be true for all courses in edX's MicroMasters programs. Why is this? Well, these courses are selected because they represent Masters level course content at UC San Diego and other edX University Partners. By completing our MicroMasters, not just a single course certificate, you indicate to employers and hiring personnel that you have made a significant investment in completing rigorous, Masters-level courses and content.   So, MicroMasters courses should offer you more.

Also, if you decide you want to do a full Masters degree you can get one faster and less expensively!

UNIVERSITY COURSE CREDIT

edX's MicroMasters Credential program allows MOOC learners, like you, to get credit towards courses for a Masters degree on various university campuses (affiliated with edX).

Current Masters Options

Currently, learners who successfully earn the MicroMasters Credential are eligible to apply for admission to the Master of Predictive Analytics program at Curtin University.

If a learner applies for admission to the Master of Predictive Analytics program at Curtin University, and is accepted, the MicroMasters Credential will count towards 25% (or 100 credits of the 400 credits) of the coursework required for graduation in the Curtin program.

Future Masters Options

At UCSD, we are committed to giving learners everywhere access to the quality, state-of-the-art data science training and preparation we provide our own Masters students in San Diego. We plan to work with other edX partner universities with appropriate Masters programs to encourage them to accept our MicroMasters courses for credit within their own programs. We'll update this page as we do this. These courses are designed directly from our own courses on campus and we believe they serve as an excellent foundation for many Masters programs preparing data scientists.

总结：