BGGN 213

Foundations of Bioinformatics

Barry Grant

UC San Diego

http://thegrantlab.org/bggn213

HELLO
my name is

BARRY

bjgrant@ucsd.edu

HELLO
HER ~~my~~ name is

DANIELA

dsamanie@ucsd.edu

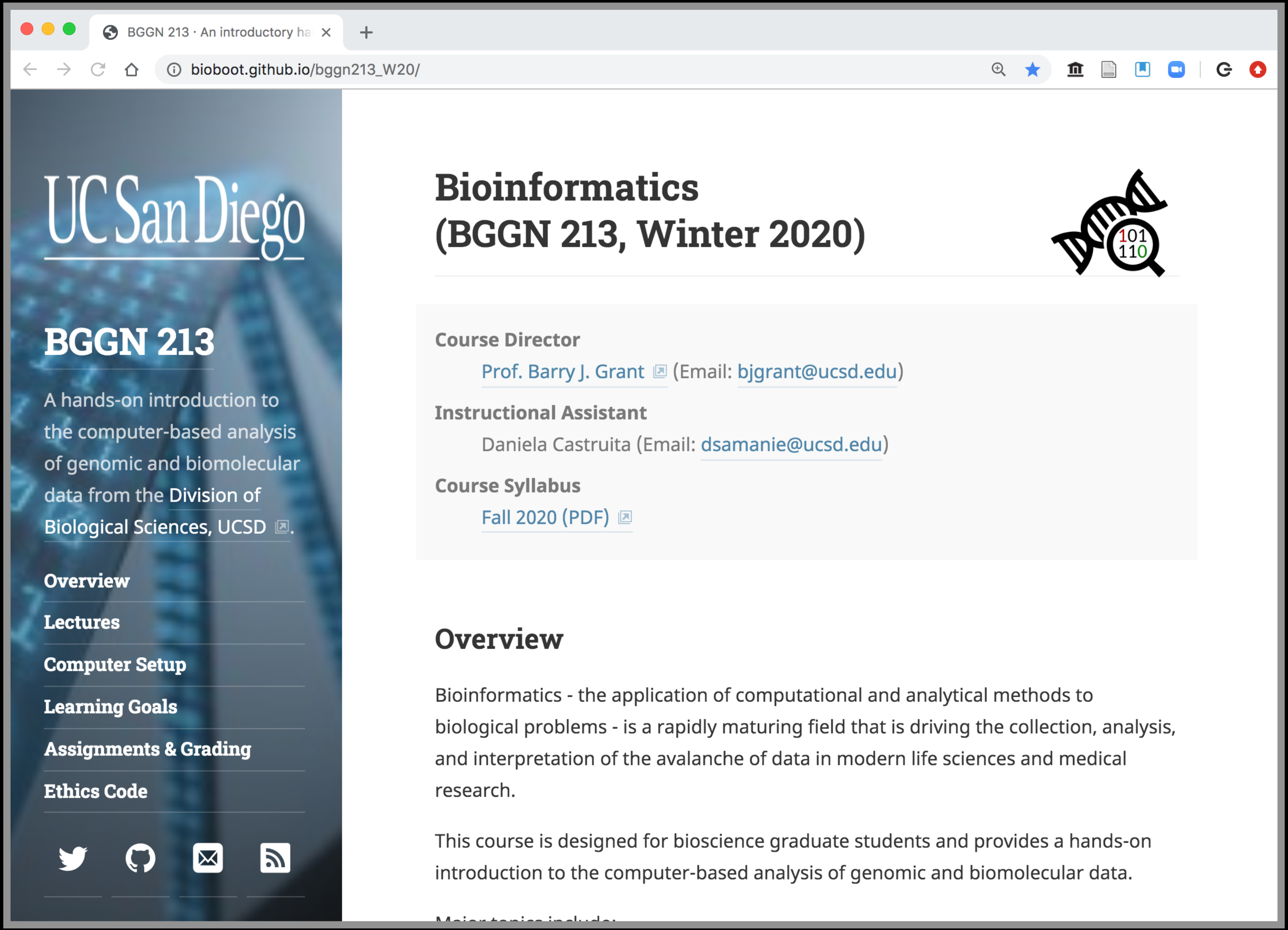
Introduce Yourself!

Your preferred name,
Place you identify with,
Major area of study/research,
Favorite joke (optional)!

Today's Menu

Course Logistics	Website, screencasts, survey, ethics, assessment and grading.
Learning Objectives	What you need to learn to succeed in this course.
Course Structure	Major lecture topics and specific leaning goals.
Introduction to Bioinformatis	**Introducing the *what*, *why* and *how* of bioinformatics?**
Bioinformatics Database	**Hands-on** exploration of several major databases and their associated tools.

http://thegrantlab.org/bggn213/



BGGN 213

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

- Overview
- Lectures
- Computer Setup
- Learning Goals
- Assignments & Grading
- Ethics Code

Bioinformatics (BGGN 213, Winter 2020)

Course Director

Prof. Barry J. Grant (Email: bjgrant@ucsd.edu)

Instructional Assistant

Daniela Castruita (Email: dsamanie@ucsd.edu)

Course Syllabus

Fall 2020 (PDF)

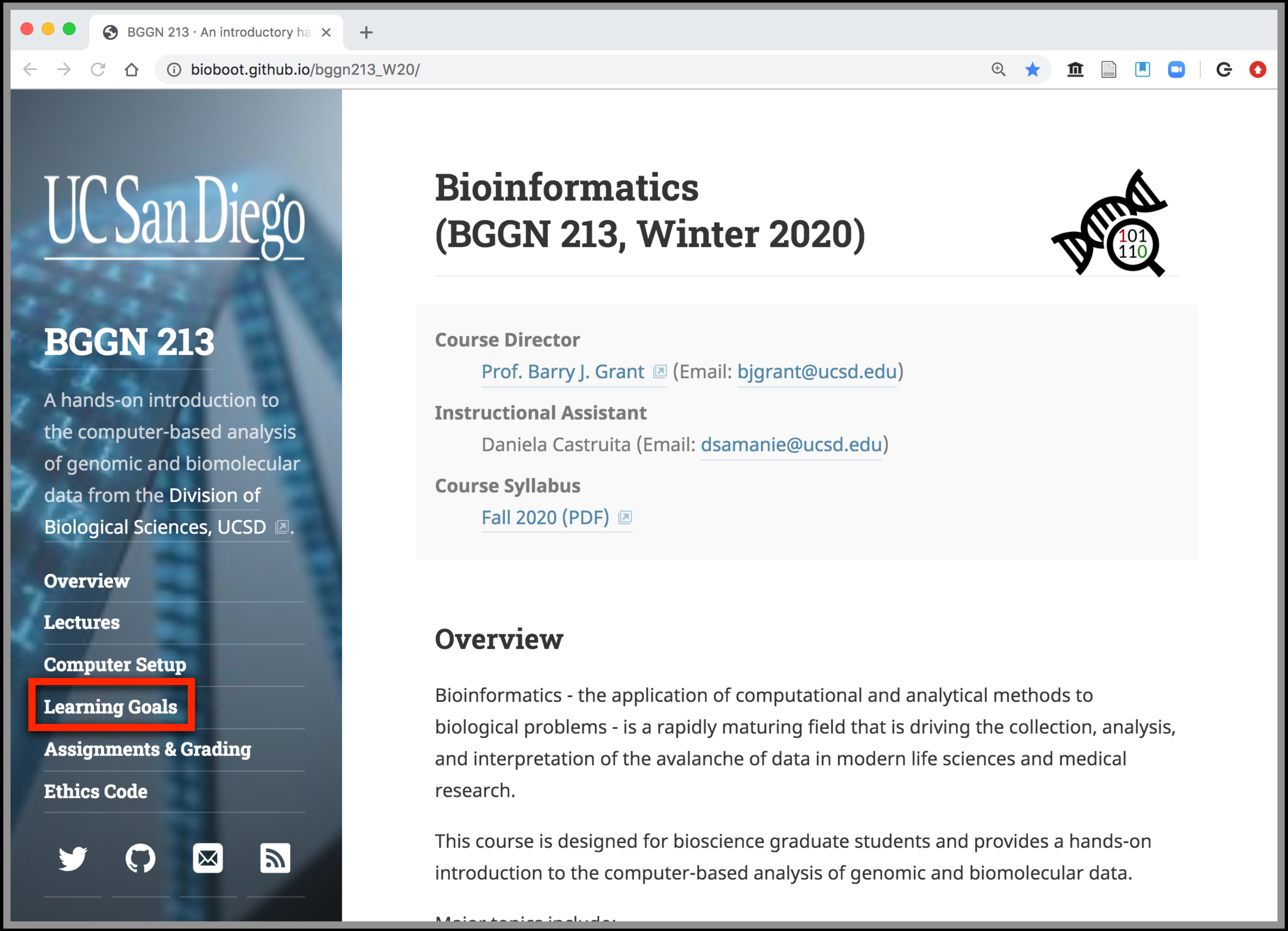
Overview

Bioinformatics - the application of computational and analytical methods to biological problems - is a rapidly maturing field that is driving the collection, analysis, and interpretation of the avalanche of data in modern life sciences and medical research.

This course is designed for bioscience graduate students and provides a hands-on introduction to the computer-based analysis of genomic and biomolecular data.

Major topics include:

http://thegrantlab.org/bggn213/



BGGN 213

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

Bioinformatics (BGGN 213, Winter 2020)

Course Director

Prof. Barry J. Grant (Email: bjgrant@ucsd.edu)

Instructional Assistant

Daniela Castruita (Email: dsamanie@ucsd.edu)

Course Syllabus

Fall 2020 (PDF)

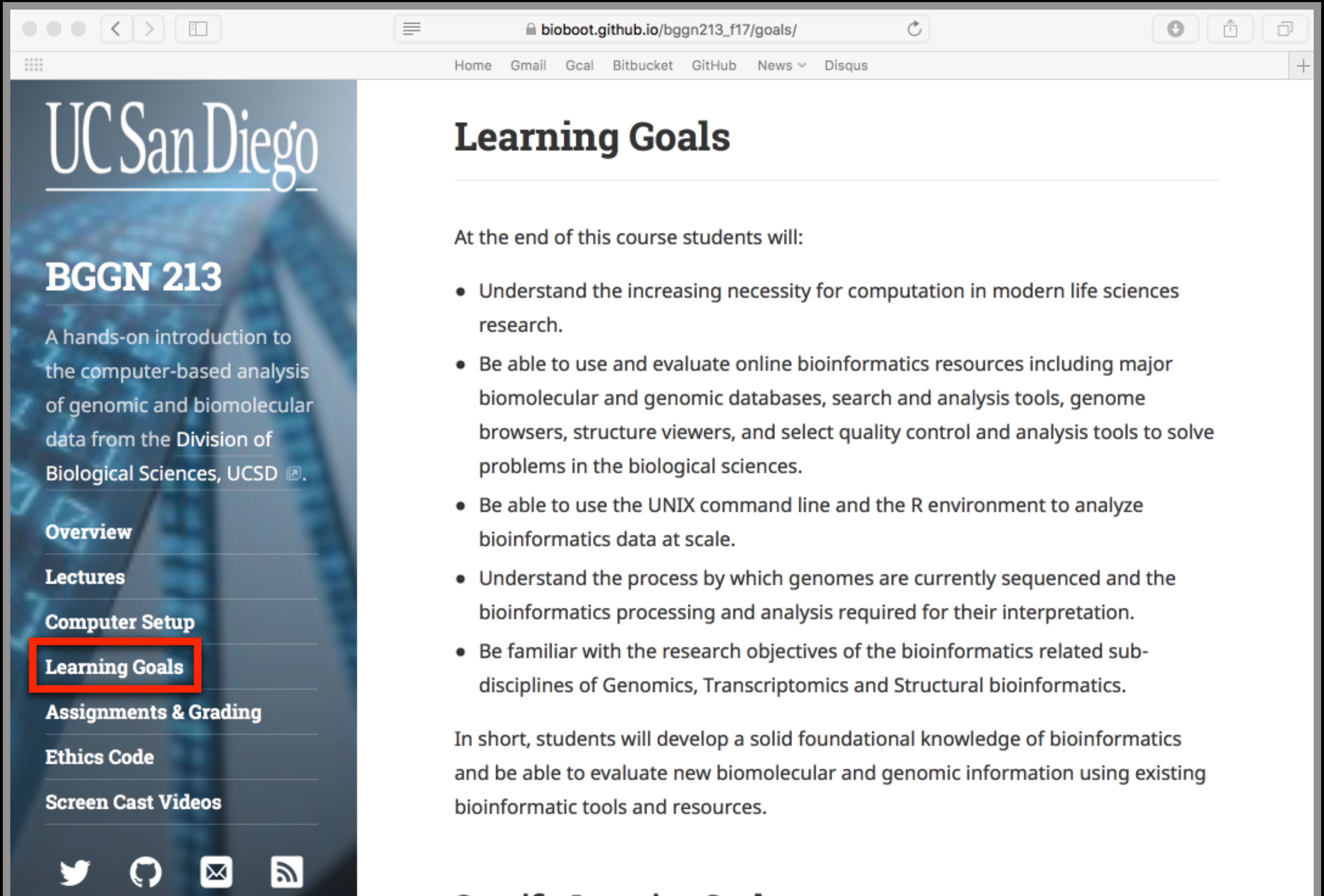
Overview

Bioinformatics - the application of computational and analytical methods to biological problems - is a rapidly maturing field that is driving the collection, analysis, and interpretation of the avalanche of data in modern life sciences and medical research.

This course is designed for bioscience graduate students and provides a hands-on introduction to the computer-based analysis of genomic and biomolecular data.

Major topics include:

What essential concepts and skills should YOU attain from this course?



BGGN 213

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

Learning Goals

At the end of this course students will:

- Understand the increasing necessity for computation in modern life sciences research.
- Be able to use and evaluate online bioinformatics resources including major biomolecular and genomic databases, search and analysis tools, genome browsers, structure viewers, and select quality control and analysis tools to solve problems in the biological sciences.
- Be able to use the UNIX command line and the R environment to analyze bioinformatics data at scale.
- Understand the process by which genomes are currently sequenced and the bioinformatics processing and analysis required for their interpretation.
- Be familiar with the research objectives of the bioinformatics related sub-disciplines of Genomics, Transcriptomics and Structural bioinformatics.

In short, students will develop a solid foundational knowledge of bioinformatics and be able to evaluate new biomolecular and genomic information using existing bioinformatic tools and resources.

At the end of this course students will:

- Understand the increasing necessity for computation in modern life sciences research.

- Be able to use and evaluate online bioinformatics resources and analysis tools to solve problems in the biological sciences.

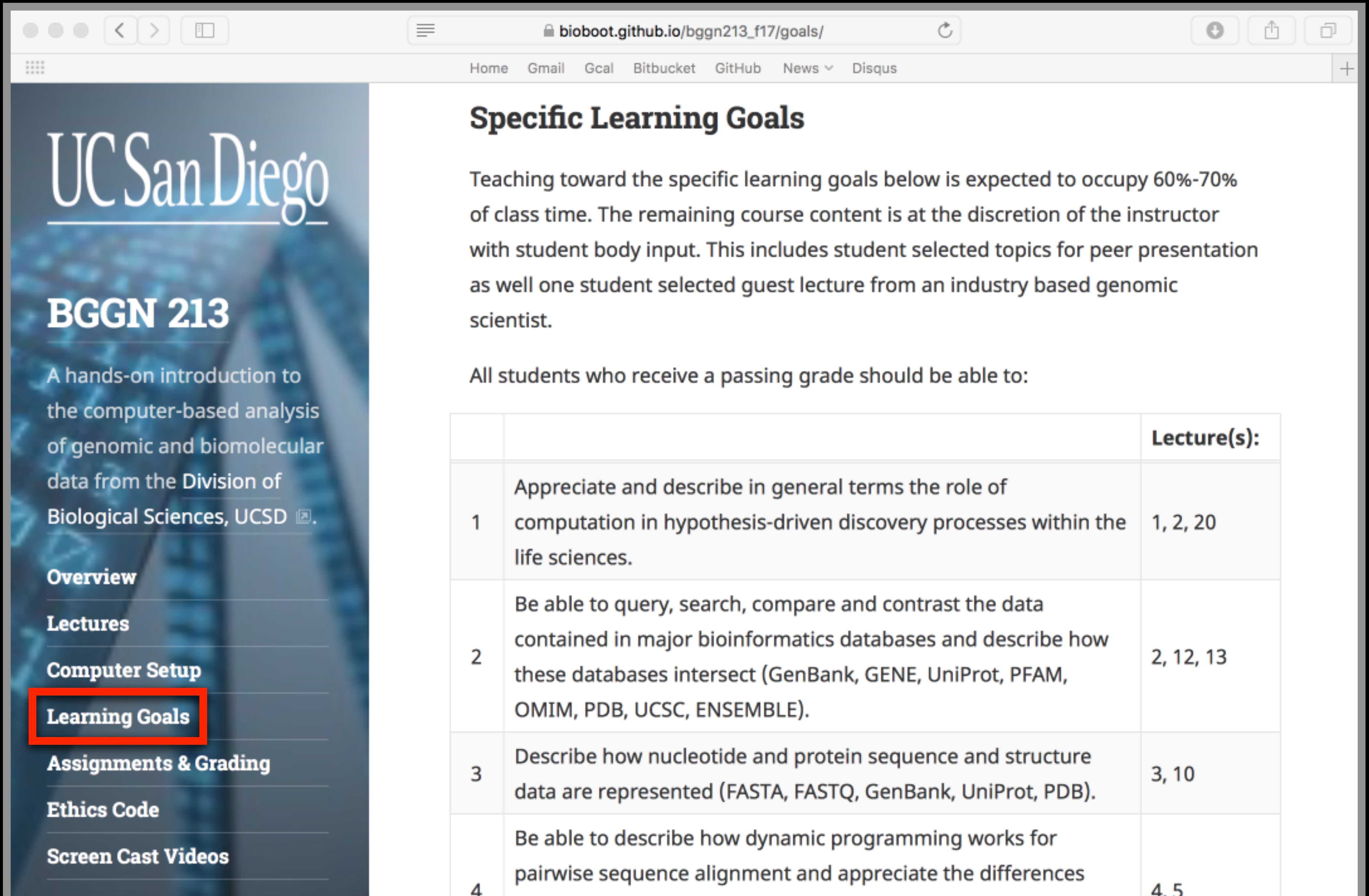
- Be able to use the R environment to analyze bioinformatics data at scale.

- Be familiar with the research objectives of the bioinformatics related sub-disciplines of Genome informatics, Transcriptomics and Structural informatics.

In short, you will develop a solid foundational knowledge of **bioinformatics** and be able to evaluate new biomolecular and genomic information using **existing bioinformatic tools and resources**.

Specific Learning Goals….

What I want you to know by course end!



bioboot.github.io/bggn213_f17/goals/
Home Gmail Gcal Bitbucket GitHub News Disqus
UC San Diego

BGGN 213

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

- Overview
- Lectures
- Computer Setup
- Learning Goals
- Assignments & Grading
- Ethics Code
- Screen Cast Videos

Specific Learning Goals

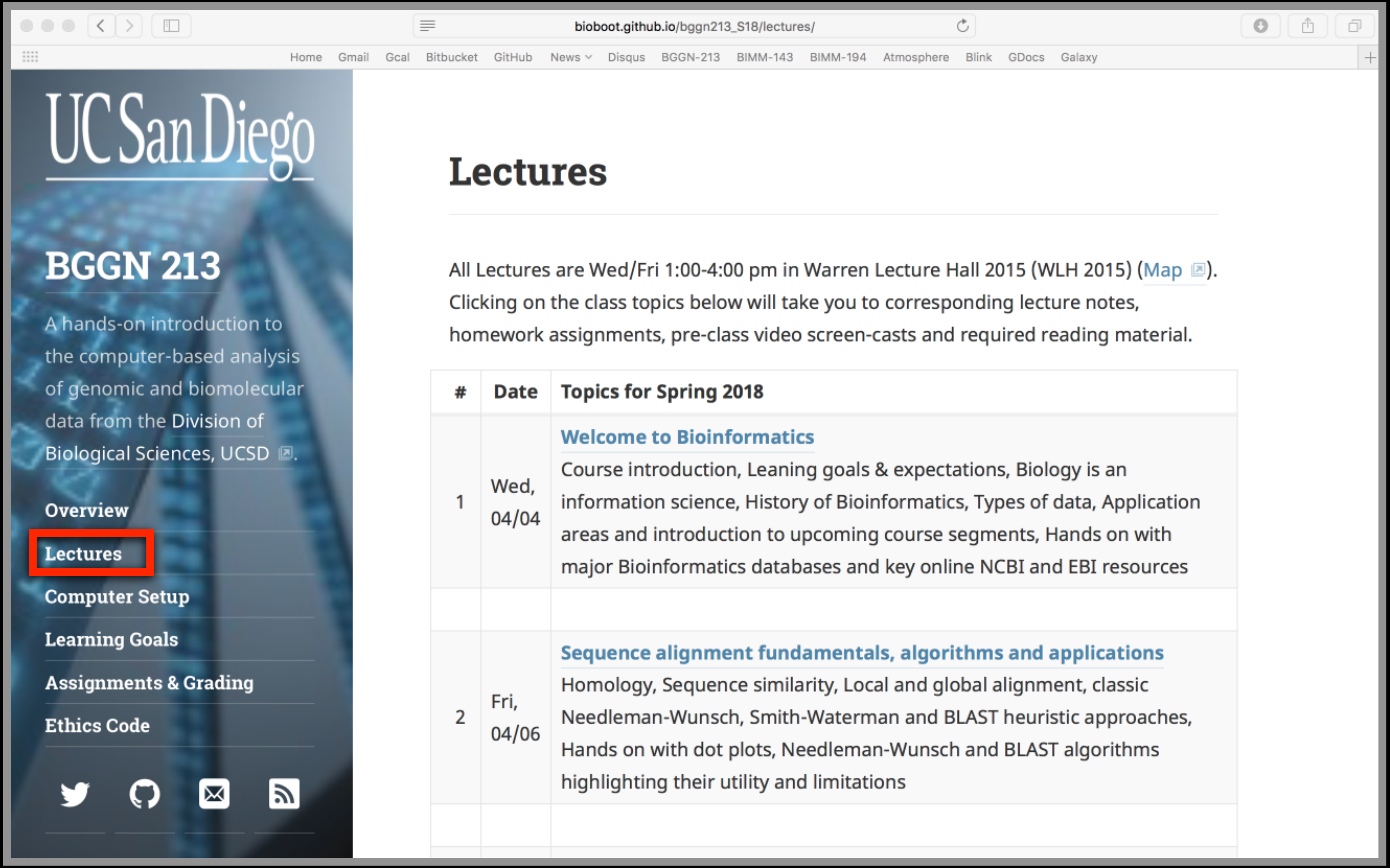
Teaching toward the specific learning goals below is expected to occupy 60%-70% of class time. The remaining course content is at the discretion of the instructor with student body input. This includes student selected topics for peer presentation as well one student selected guest lecture from an industry based genomic scientist.

All students who receive a passing grade should be able to:

		Lecture(s):
1	Appreciate and describe in general terms the role of computation in hypothesis-driven discovery processes within the life sciences.	1, 2, 20
2	Be able to query, search, compare and contrast the data contained in major bioinformatics databases and describe how these databases intersect (GenBank, GENE, UniProt, PFAM, OMIM, PDB, UCSC, ENSEMBLE).	2, 12, 13
3	Describe how nucleotide and protein sequence and structure data are represented (FASTA, FASTQ, GenBank, UniProt, PDB).	3, 10
4	Be able to describe how dynamic programming works for pairwise sequence alignment and appreciate the differences	4, 5

Course Structure

Derived from specific learning goals



bioboot.github.io/bggn213_S18/lectures/
Home Gmail Gcal Bitbucket GitHub News Disqus BGGN-213 BIMM-143 BIMM-194 Atmosphere Blink GDocs Galaxy
UC San Diego

BGGN 213

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

- Overview
- Lectures
- Computer Setup
- Learning Goals
- Assignments & Grading
- Ethics Code

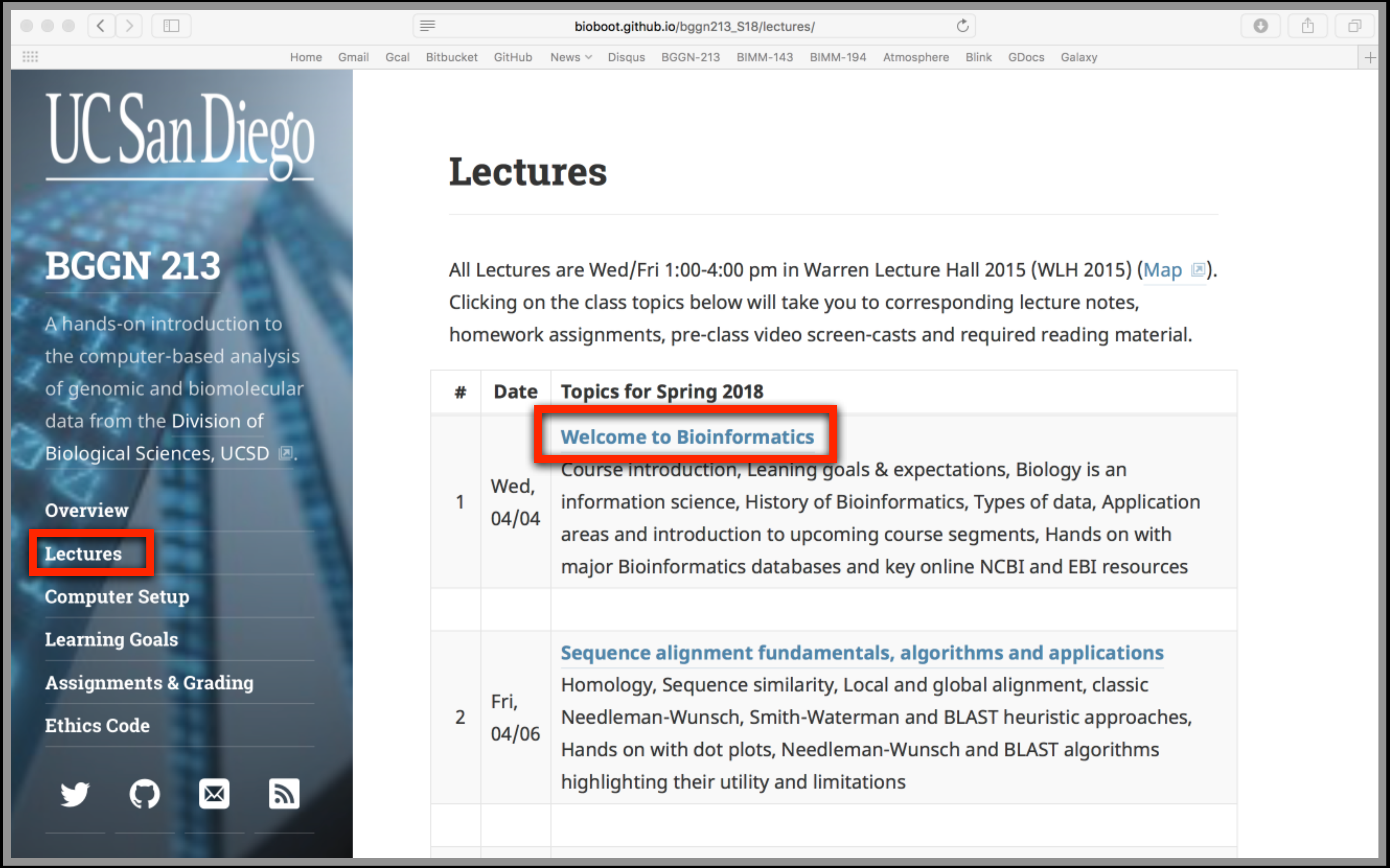
Lectures

All Lectures are Wed/Fri 1:00-4:00 pm in Warren Lecture Hall 2015 (WLH 2015) (Map). Clicking on the class topics below will take you to corresponding lecture notes, homework assignments, pre-class video screen-casts and required reading material.

#	Date	Topics for Spring 2018
1	Wed, 04/04	**Welcome to Bioinformatics** Course introduction, Leaning goals & expectations, Biology is an information science, History of Bioinformatics, Types of data, Application areas and introduction to upcoming course segments, Hands on with major Bioinformatics databases and key online NCBI and EBI resources
2	Fri, 04/06	**Sequence alignment fundamentals, algorithms and applications** Homology, Sequence similarity, Local and global alignment, classic Needleman-Wunsch, Smith-Waterman and BLAST heuristic approaches, Hands on with dot plots, Needleman-Wunsch and BLAST algorithms highlighting their utility and limitations

Course Structure

Derived from specific learning goals



bioboot.github.io/bggn213_S18/lectures/
Home Gmail Gcal Bitbucket GitHub News Disqus BGGN-213 BIMM-143 BIMM-194 Atmosphere Blink GDocs Galaxy
UC San Diego

BGGN 213

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

- Overview
- Lectures
- Computer Setup
- Learning Goals
- Assignments & Grading
- Ethics Code

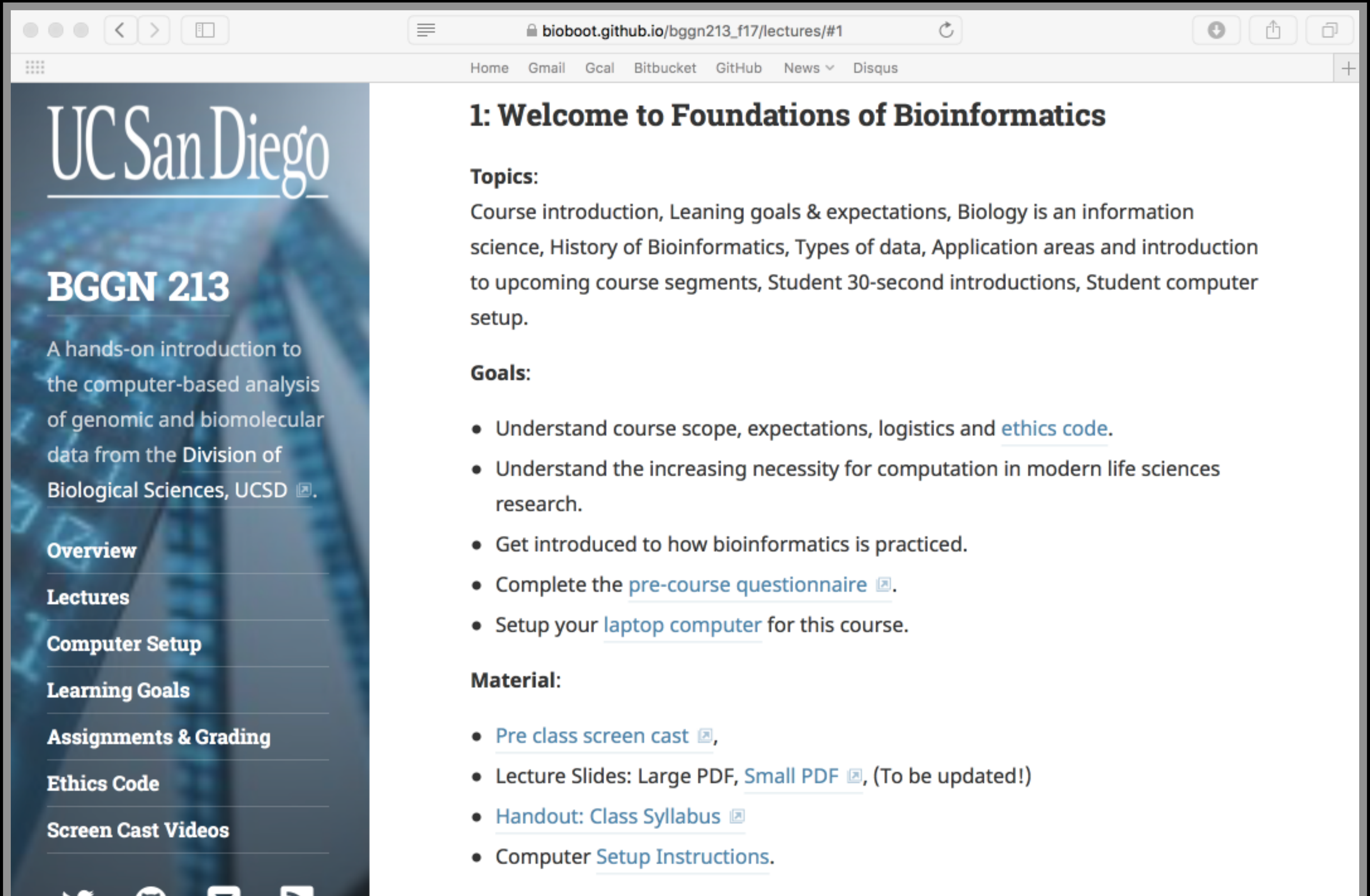
Lectures

All Lectures are Wed/Fri 1:00-4:00 pm in Warren Lecture Hall 2015 (WLH 2015) (Map). Clicking on the class topics below will take you to corresponding lecture notes, homework assignments, pre-class video screen-casts and required reading material.

#	Date	Topics for Spring 2018
1	Wed, 04/04	**Welcome to Bioinformatics** Course introduction, Leaning goals & expectations, Biology is an information science, History of Bioinformatics, Types of data, Application areas and introduction to upcoming course segments, Hands on with major Bioinformatics databases and key online NCBI and EBI resources
2	Fri, 04/06	**Sequence alignment fundamentals, algorithms and applications** Homology, Sequence similarity, Local and global alignment, classic Needleman-Wunsch, Smith-Waterman and BLAST heuristic approaches, Hands on with dot plots, Needleman-Wunsch and BLAST algorithms highlighting their utility and limitations

Class Details

Goals, Class material, Screencasts & **Homework**



BGGN 213

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

1: Welcome to Foundations of Bioinformatics

Topics:
Course introduction, Leaning goals & expectations, Biology is an information science, History of Bioinformatics, Types of data, Application areas and introduction to upcoming course segments, Student 30-second introductions, Student computer setup.

Goals:

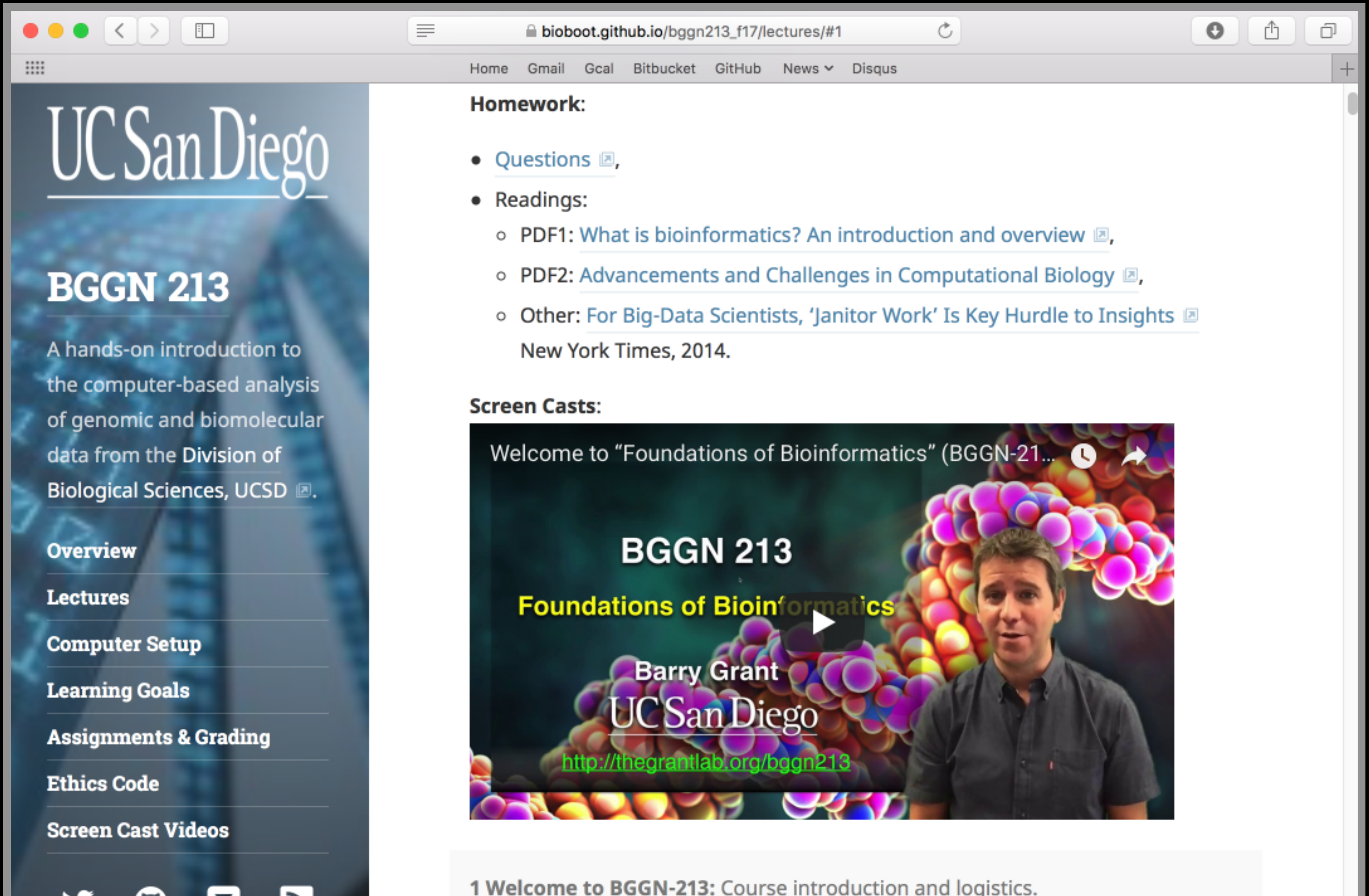
- Understand course scope, expectations, logistics and ethics code.
- Understand the increasing necessity for computation in modern life sciences research.
- Get introduced to how bioinformatics is practiced.
- Complete the pre-course questionnaire.
- Setup your laptop computer for this course.

Material:

- Pre class screen cast,
- Lecture Slides: Large PDF, Small PDF, (To be updated!)
- Handout: Class Syllabus
- Computer Setup Instructions.

Homework

Goals, Class material, Screencasts & **Homework**



UC San Diego

BGGN 213

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

- Overview
- Lectures
- Computer Setup
- Learning Goals
- Assignments & Grading
- Ethics Code
- Screen Cast Videos

Homework:

- Questions,
- Readings:
  - PDF1: What is bioinformatics? An introduction and overview,
  - PDF2: Advancements and Challenges in Computational Biology,
  - Other: For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights New York Times, 2014.

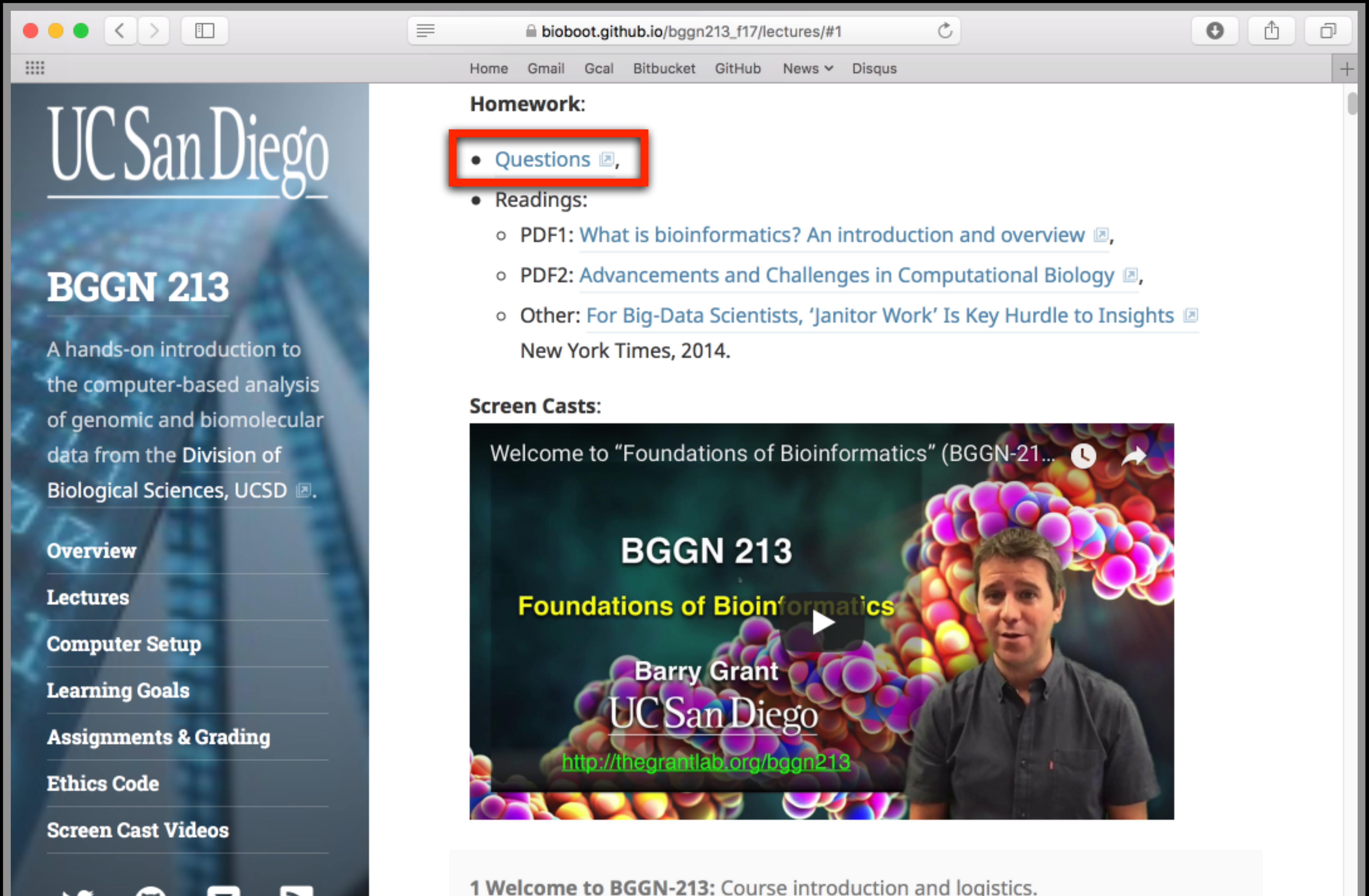
Screen Casts:

Welcome to "Foundations of Bioinformatics" (BGGN-21...
BGGN 213
Foundations of Bioinformatics
Barry Grant
UC San Diego
http://thegrantlab.org/bggn213

1 Welcome to BGGN-213: Course introduction and logistics.

Homework

Goals, Class material, Screencasts & **Homework**



BGGN 213

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

Homework:

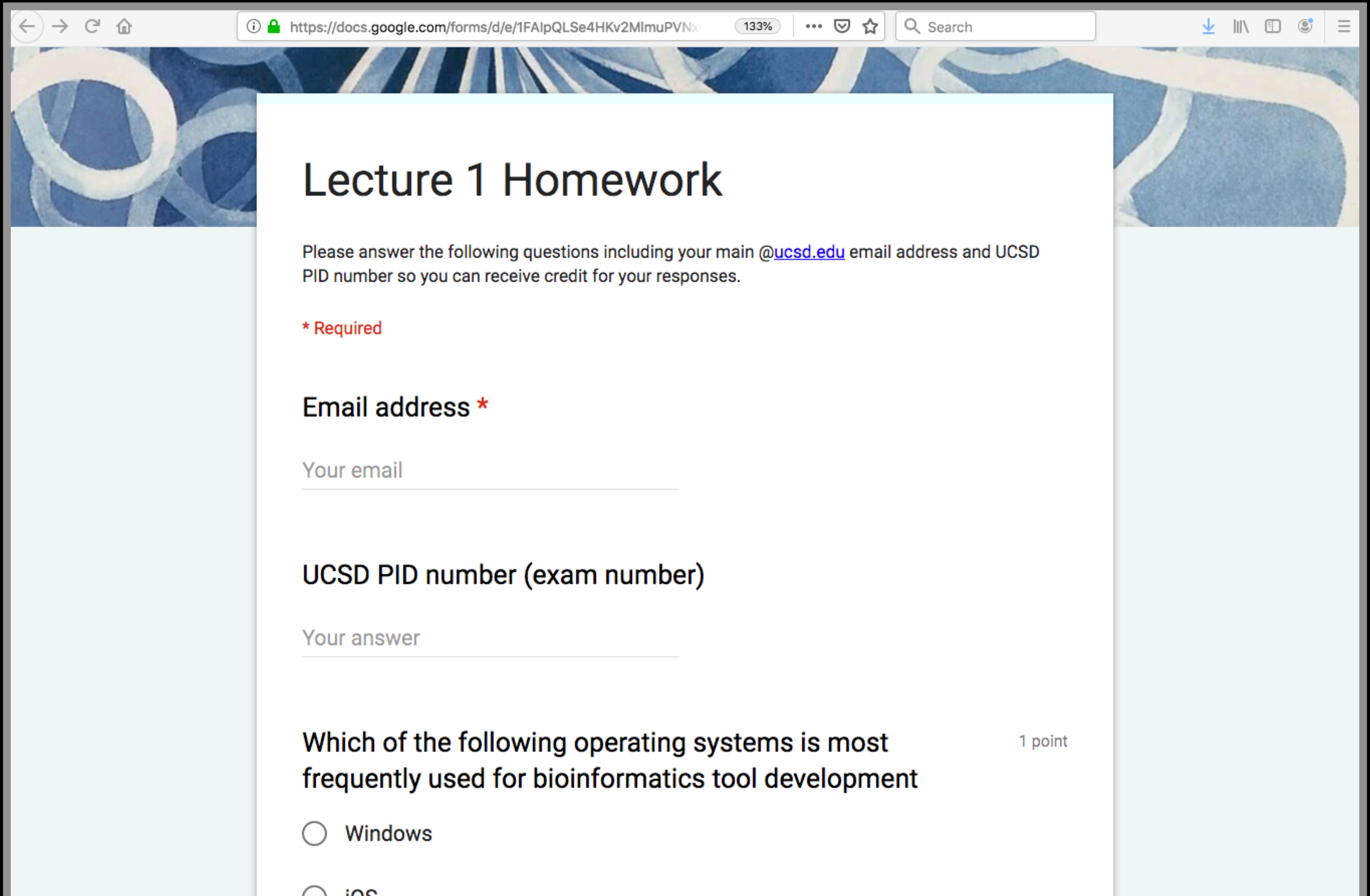
- Questions,
- Readings:
  - PDF1: What is bioinformatics? An introduction and overview,
  - PDF2: Advancements and Challenges in Computational Biology,
  - Other: For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights New York Times, 2014.

Screen Casts:

1 Welcome to BGGN-213: Course introduction and logistics.

Homework

Goals, Class material, Screencasts & **Homework**



Lecture 1 Homework

Please answer the following questions including your main @ucsd.edu email address and UCSD PID number so you can receive credit for your responses.

* Required

Email address *

Your email

UCSD PID number (exam number)

Your answer

Which of the following operating systems is most frequently used for bioinformatics tool development

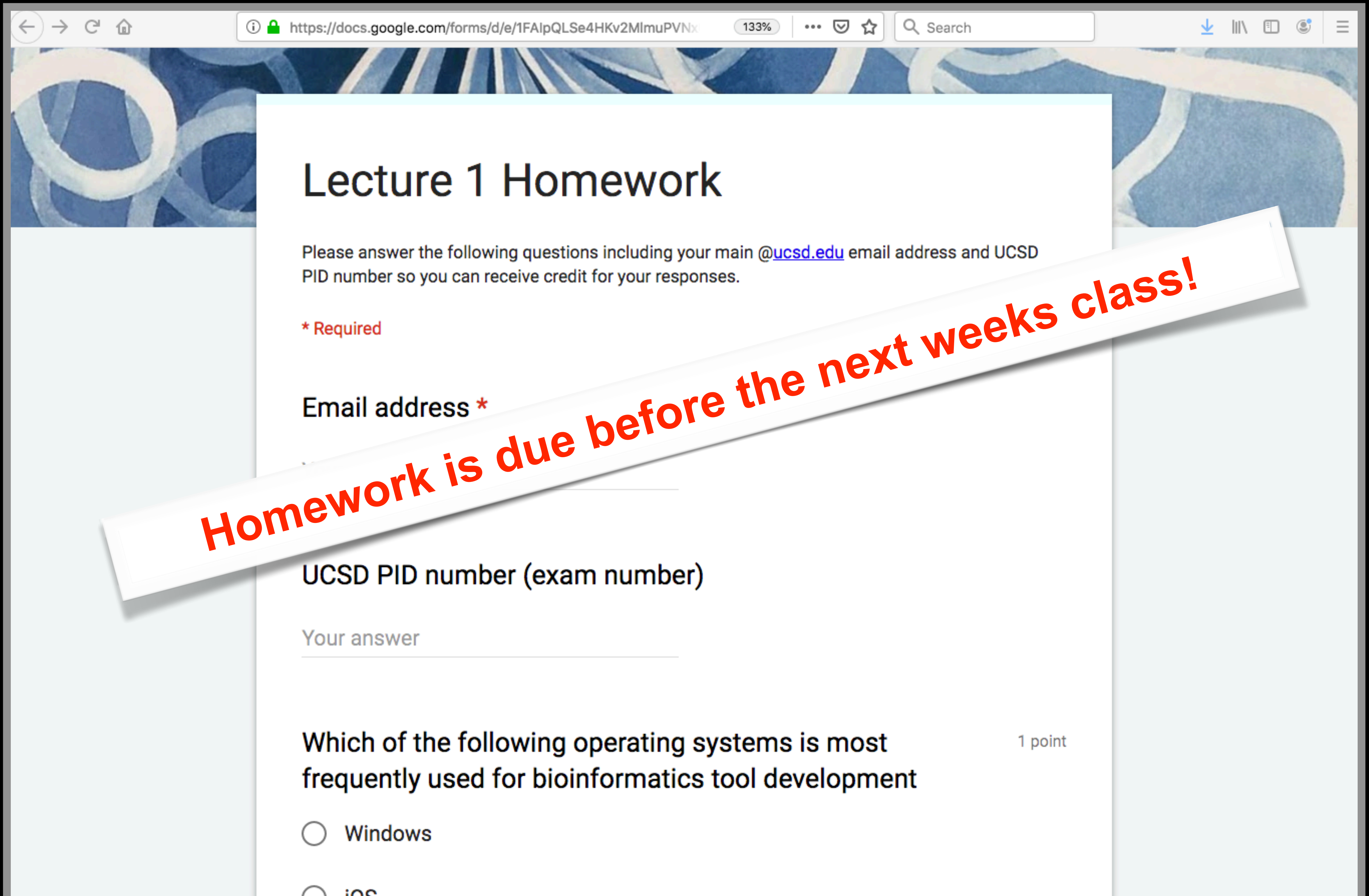
1 point

○ Windows

○ iOS

Homework

Goals, Class material, Screencasts & **Homework**



Lecture 1 Homework

Please answer the following questions including your main @ucsd.edu email address and UCSD PID number so you can receive credit for your responses.

* Required

Email address *

UCSD PID number (exam number)

Your answer

Which of the following operating systems is most frequently used for bioinformatics tool development

1 point

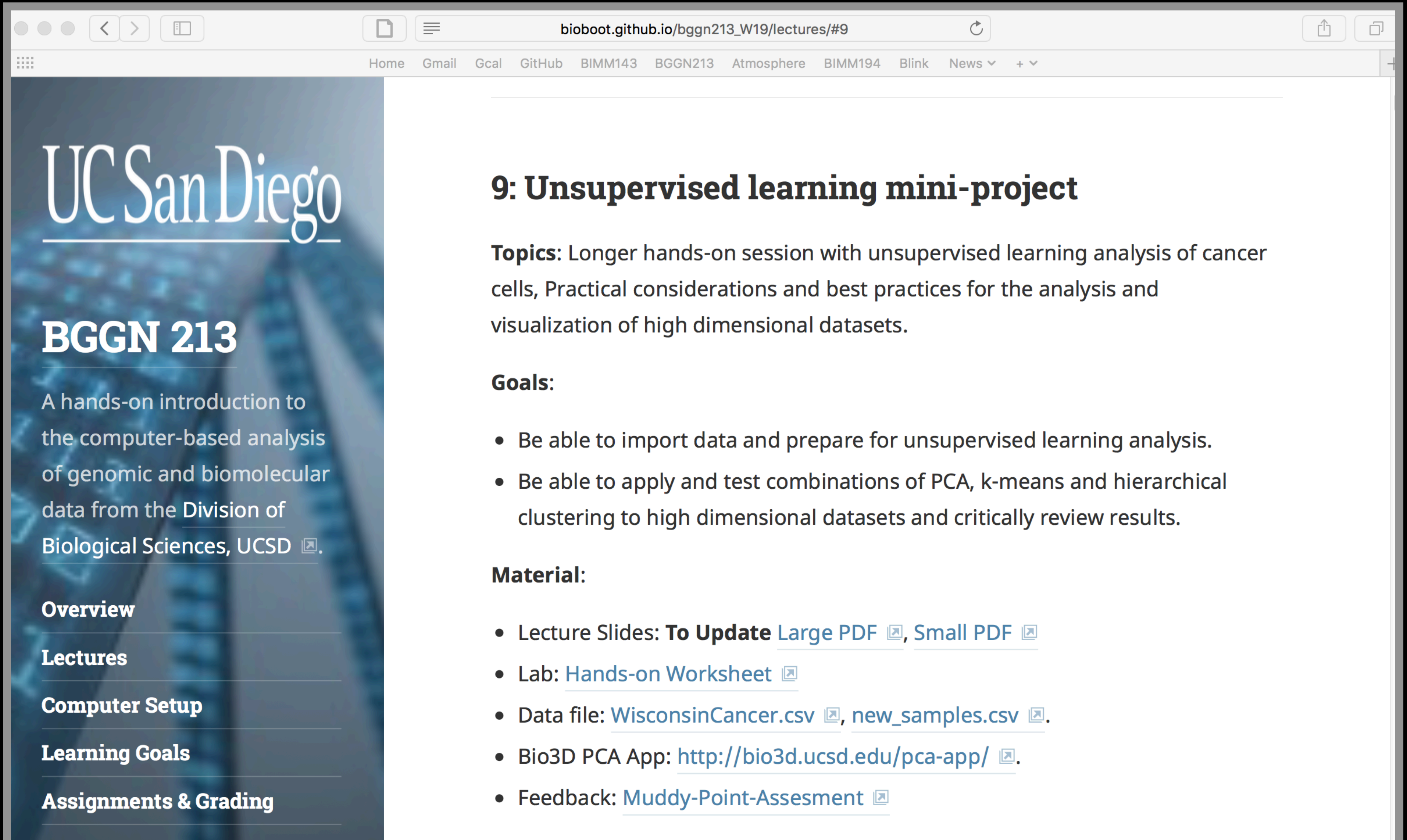
○ Windows

○ iOS

Homework is due before the next weeks class!

Projects

Week long **mini-projects** (x2), and 1 five week main project


bioboot.github.io/bggn213_W19/lectures/#9
Home Gmail Gcal GitHub BIMM143 BGGN213 Atmosphere BIMM194 Blink News +
UC San Diego

BGGN 213

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

Overview

Lectures

Computer Setup

Learning Goals

Assignments & Grading

9: Unsupervised learning mini-project

Topics: Longer hands-on session with unsupervised learning analysis of cancer cells, Practical considerations and best practices for the analysis and visualization of high dimensional datasets.

Goals:

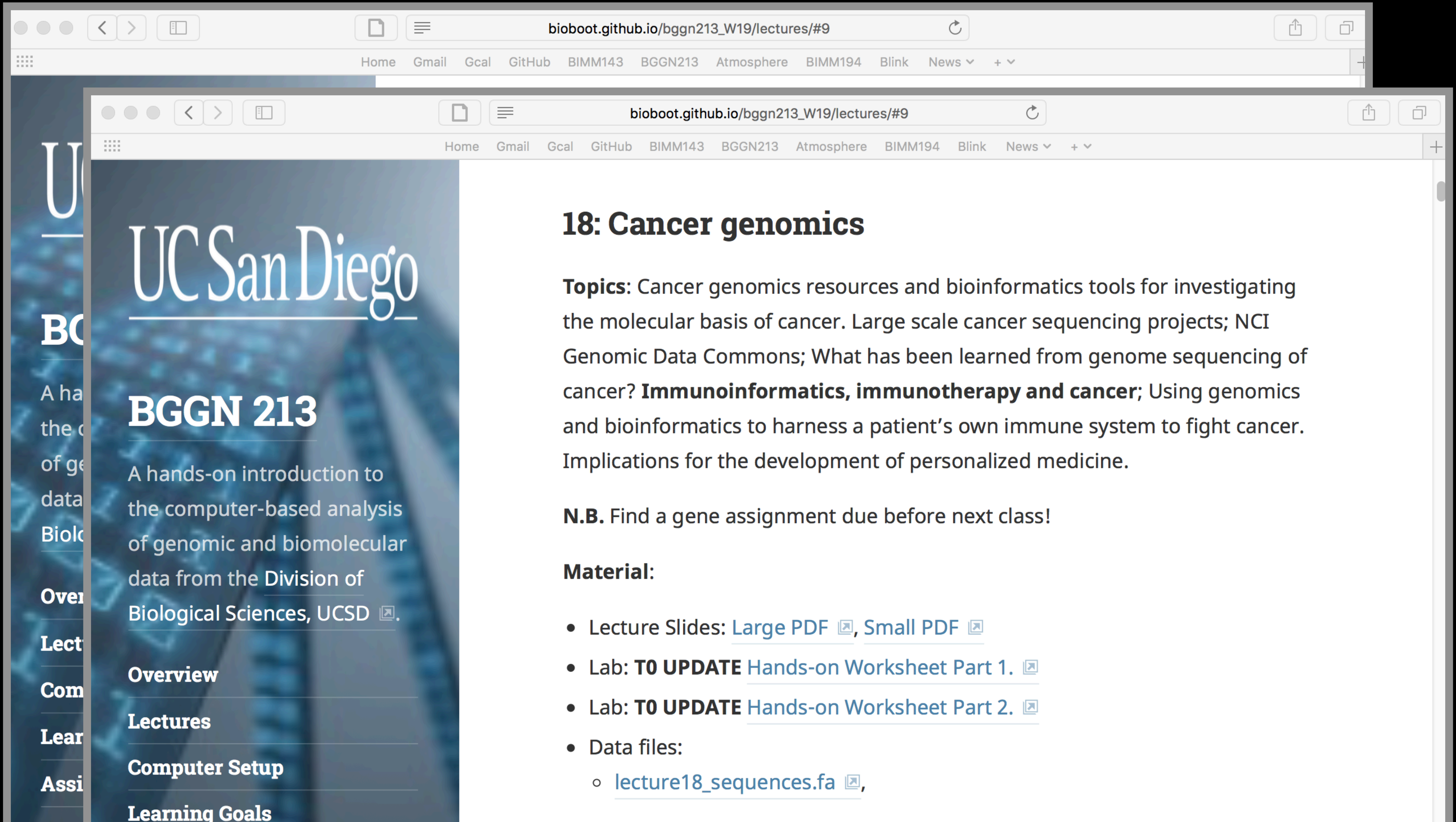
- Be able to import data and prepare for unsupervised learning analysis.
- Be able to apply and test combinations of PCA, k-means and hierarchical clustering to high dimensional datasets and critically review results.

Material:

- Lecture Slides: **To Update** Large PDF, Small PDF
- Lab: Hands-on Worksheet
- Data file: WisconsinCancer.csv, new_samples.csv.
- Bio3D PCA App: http://bio3d.ucsd.edu/pca-app/.
- Feedback: Muddy-Point-Assesment

Projects

Week long **mini-projects** (x2),
and 1 five week main project



UC San Diego

BGGN 213

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

Overview

Lectures

Computer Setup

Learning Goals

18: Cancer genomics

Topics: Cancer genomics resources and bioinformatics tools for investigating the molecular basis of cancer. Large scale cancer sequencing projects; NCI Genomic Data Commons; What has been learned from genome sequencing of cancer? **Immunoinformatics, immunotherapy and cancer**; Using genomics and bioinformatics to harness a patient's own immune system to fight cancer. Implications for the development of personalized medicine.

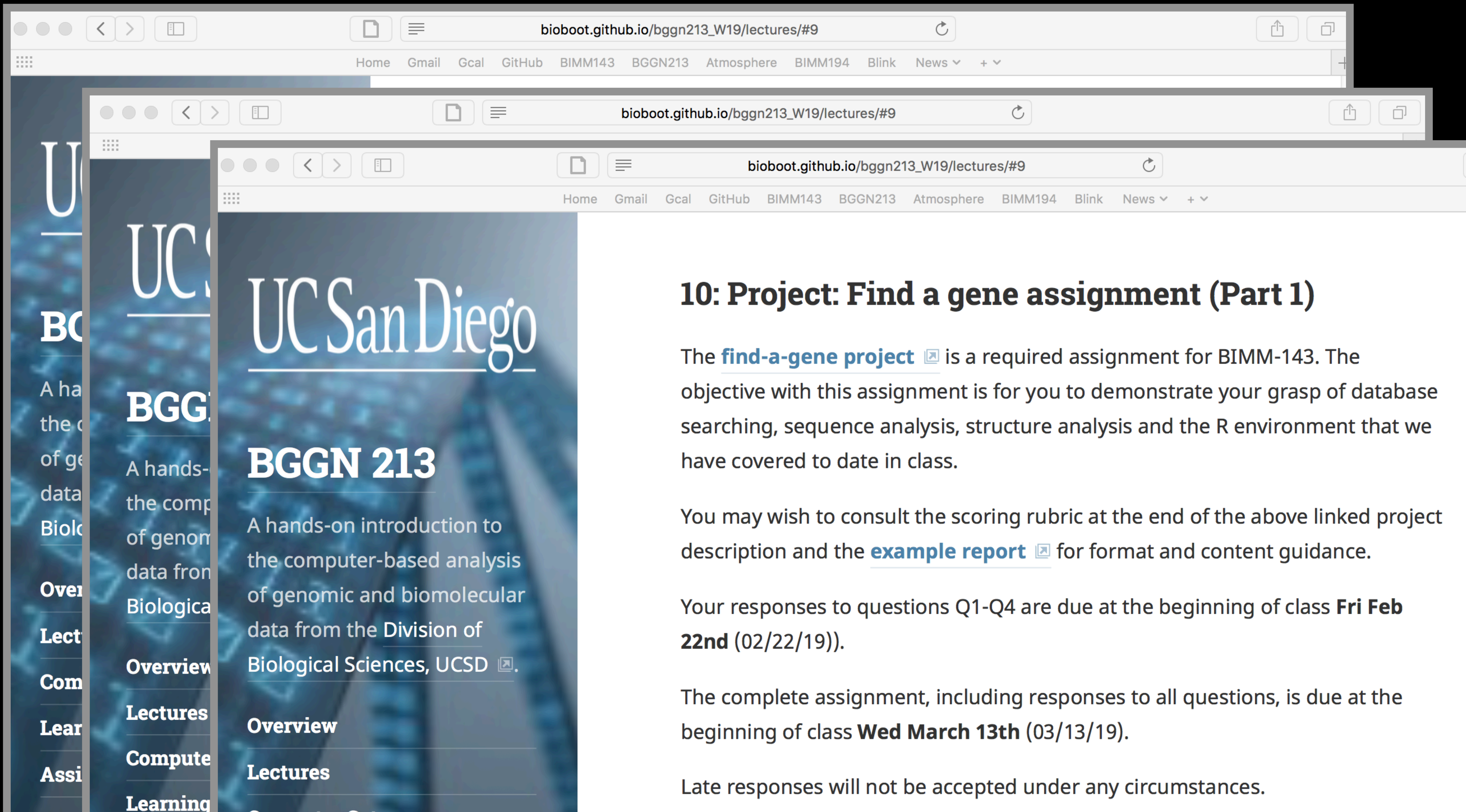
N.B. Find a gene assignment due before next class!

Material:

- Lecture Slides: Large PDF, Small PDF
- Lab: **TO UPDATE** Hands-on Worksheet Part 1.
- Lab: **TO UPDATE** Hands-on Worksheet Part 2.
- Data files:
  - lecture18_sequences.fa,

Projects

Week long mini-projects (x2), and 1 five week **main project**



bioboot.github.io/bggn213_W19/lectures/#9
Home Gmail Gcal GitHub BIMM143 BGGN213 Atmosphere BIMM194 Blink News +
UC San Diego
BGGN 213
A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.
Overview
Lectures

10: Project: Find a gene assignment (Part 1)

The **find-a-gene project** is a required assignment for BIMM-143. The objective with this assignment is for you to demonstrate your grasp of database searching, sequence analysis, structure analysis and the R environment that we have covered to date in class.

You may wish to consult the scoring rubric at the end of the above linked project description and the **example report** for format and content guidance.

Your responses to questions Q1-Q4 are due at the beginning of class **Fri Feb 22nd** (02/22/19)).

The complete assignment, including responses to all questions, is due at the beginning of class **Wed March 13th** (03/13/19).

Late responses will not be accepted under any circumstances.

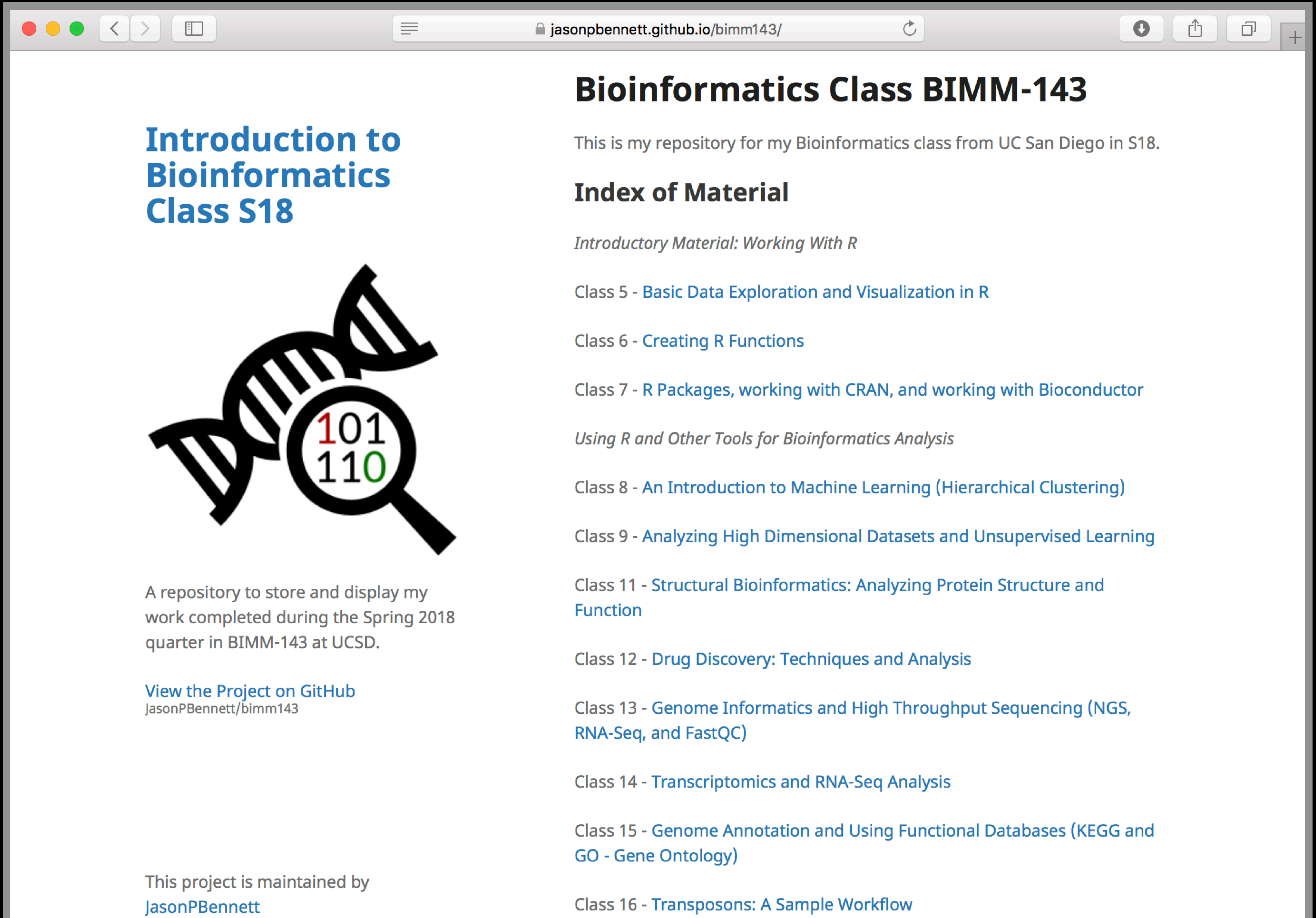
Why Projects?

- Projects allow you to practice your new Bioinformatics skills in a less guided environment.

- In Projects, we provide datasets and ask you questions about them; just like a research project.

- Projects help build a personal portfolio and showcase your new skills, as well as help put what we have learned into practice.

Online portfolio of **your** bioinformatics work!



Introduction to Bioinformatics Class S18

A repository to store and display my work completed during the Spring 2018 quarter in BIMM-143 at UCSD.

View the Project on GitHub
JasonPBennett/bimm143

This project is maintained by JasonPBennett

Bioinformatics Class BIMM-143

This is my repository for my Bioinformatics class from UC San Diego in S18.

Index of Material

Introductory Material: Working With R

Class 5 - Basic Data Exploration and Visualization in R

Class 6 - Creating R Functions

Class 7 - R Packages, working with CRAN, and working with Bioconductor

Using R and Other Tools for Bioinformatics Analysis

Class 8 - An Introduction to Machine Learning (Hierarchical Clustering)

Class 9 - Analyzing High Dimensional Datasets and Unsupervised Learning

Class 11 - Structural Bioinformatics: Analyzing Protein Structure and Function

Class 12 - Drug Discovery: Techniques and Analysis

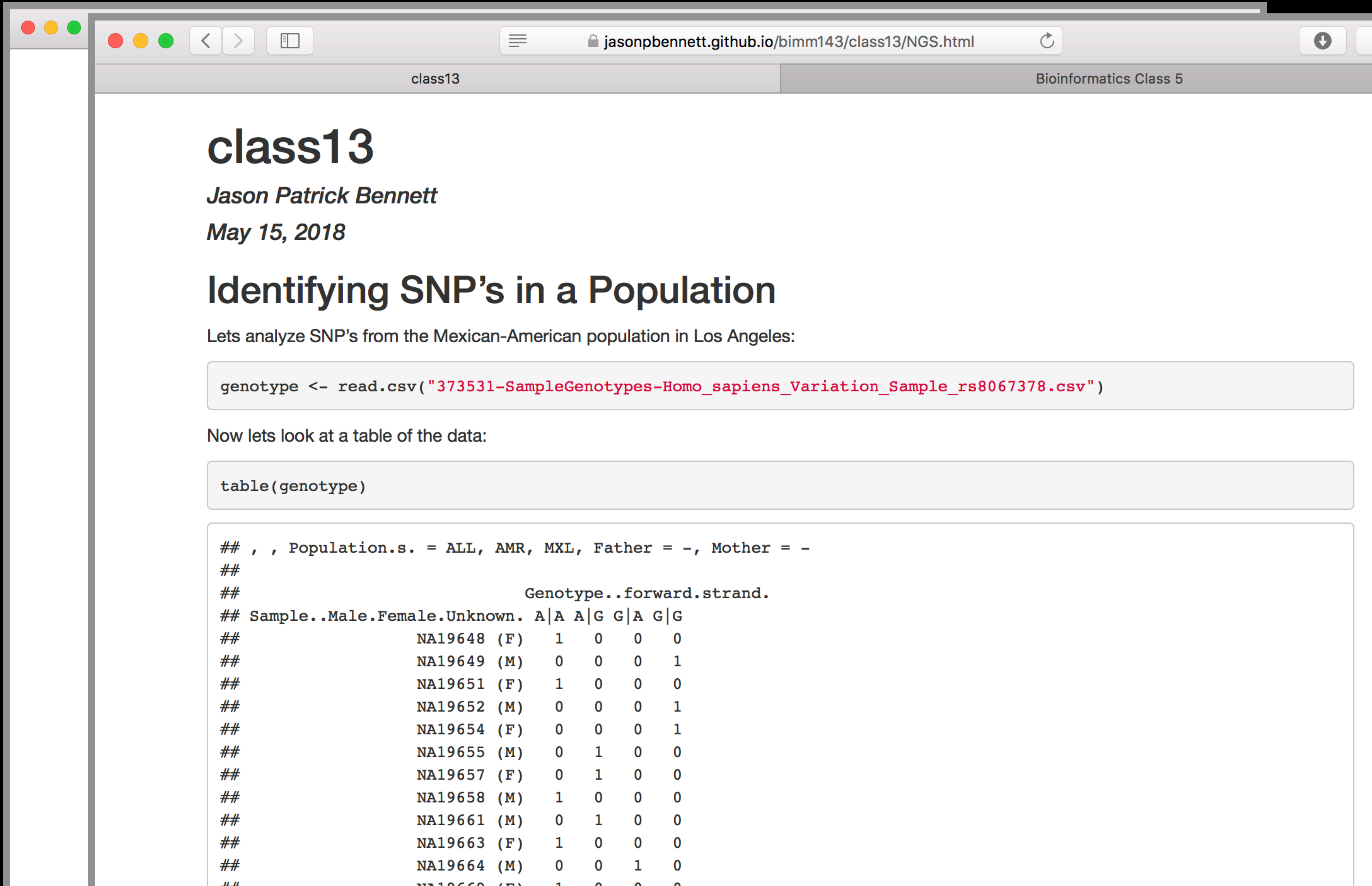
Class 13 - Genome Informatics and High Throughput Sequencing (NGS, RNA-Seq, and FastQC)

Class 14 - Transcriptomics and RNA-Seq Analysis

Class 15 - Genome Annotation and Using Functional Databases (KEGG and GO - Gene Ontology)

Class 16 - Transposons: A Sample Workflow

Online portfolio of **your** bioinformatics work!



class13

Jason Patrick Bennett

May 15, 2018

Identifying SNP's in a Population

Lets analyze SNP's from the Mexican-American population in Los Angeles:

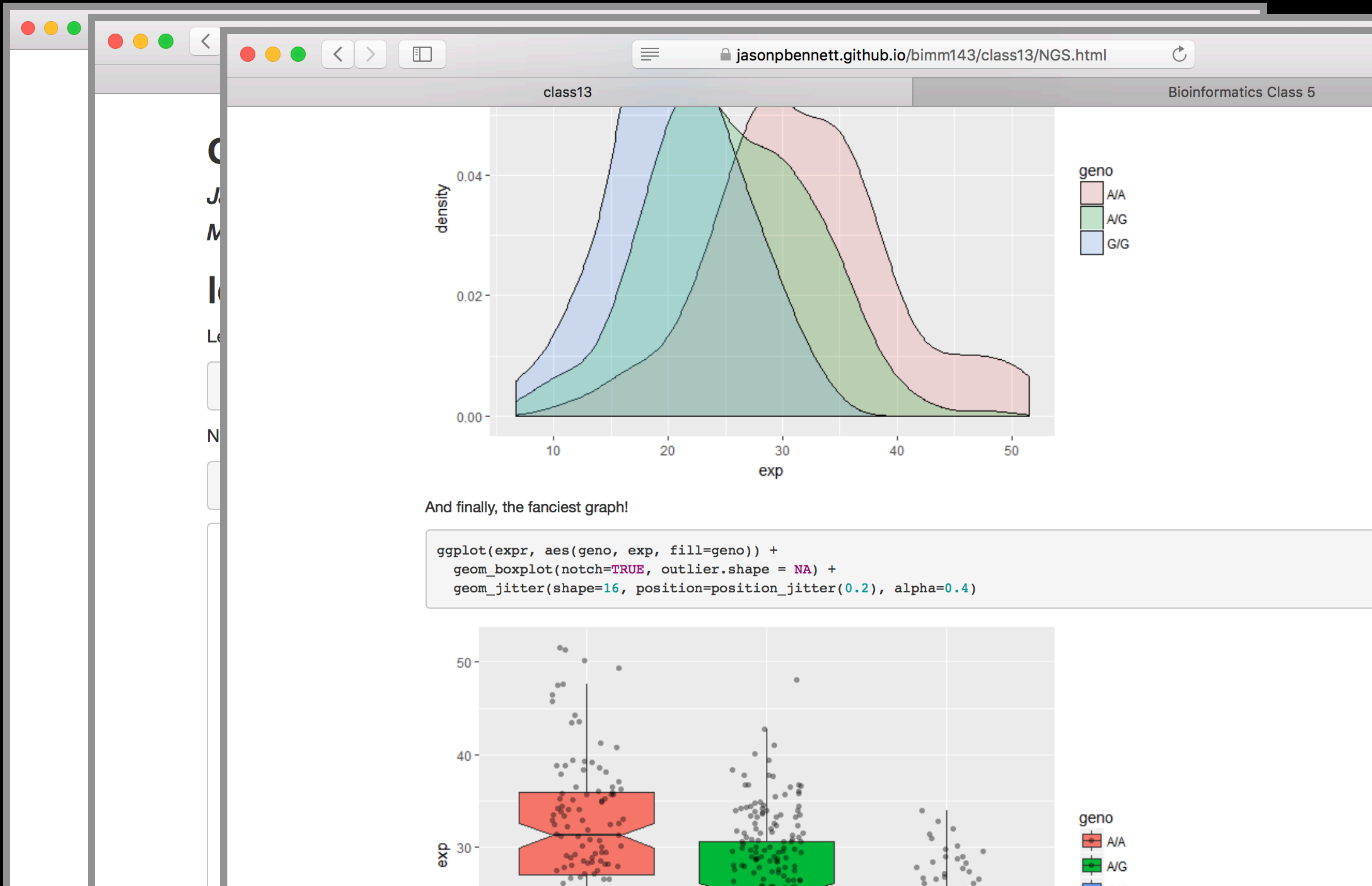
```
genotype <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")
```

Now lets look at a table of the data:

```
table(genotype)
```

```
## , , Population.s. = ALL, AMR, MXL, Father = -, Mother = -
##
##                             Genotype..forward.strand.
## Sample..Male.Female.Unknown. A|A A|G G|A G|G
##                  NA19648 (F)   1   0   0   0
##                  NA19649 (M)   0   0   0   1
##                  NA19651 (F)   1   0   0   0
##                  NA19652 (M)   0   0   0   1
##                  NA19654 (F)   0   0   0   1
##                  NA19655 (M)   0   1   0   0
##                  NA19657 (F)   0   1   0   0
##                  NA19658 (M)   1   0   0   0
##                  NA19661 (M)   0   1   0   0
##                  NA19663 (F)   1   0   0   0
##                  NA19664 (M)   0   0   1   0
```

Online portfolio of **your** bioinformatics work!

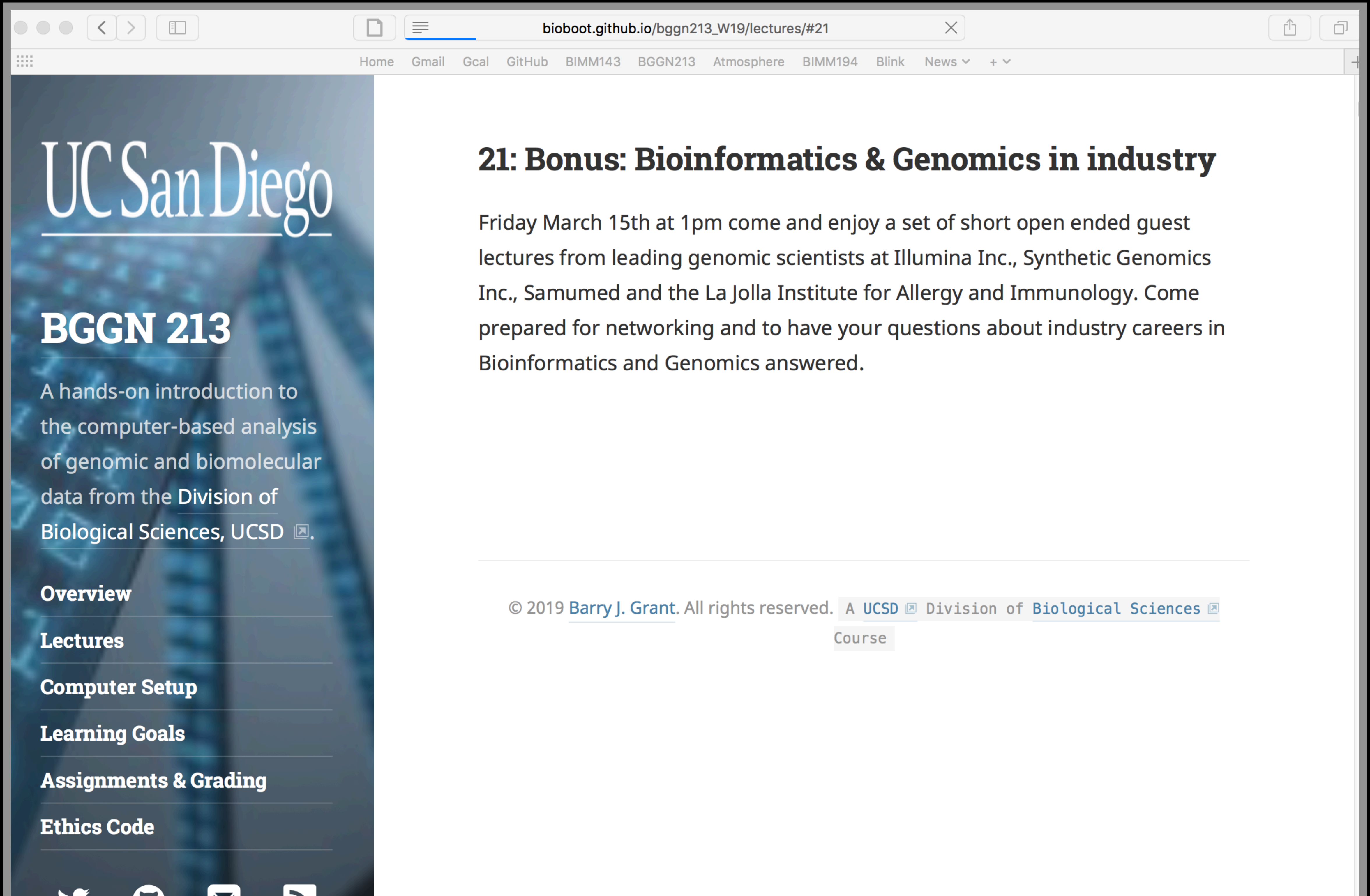


And finally, the fanciest graph!

```
ggplot(expr, aes(geno, exp, fill=geno)) +
  geom_boxplot(notch=TRUE, outlier.shape = NA) +
  geom_jitter(shape=16, position=position_jitter(0.2), alpha=0.4)
```

Bonus:

Bioinformatics & Genomics in industry



BGGN 213

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

Overview

Lectures

Computer Setup

Learning Goals

Assignments & Grading

Ethics Code

21: Bonus: Bioinformatics & Genomics in industry

Friday March 15th at 1pm come and enjoy a set of short open ended guest lectures from leading genomic scientists at Illumina Inc., Synthetic Genomics Inc., Samumed and the La Jolla Institute for Allergy and Immunology. Come prepared for networking and to have your questions about industry careers in Bioinformatics and Genomics answered.

© 2019 Barry J. Grant. All rights reserved. `A UCSD Division of Biological Sciences Course`

Side Note: **Why stick with this course?**

Provides a hands-on practical introduction to major bioinformatics concepts and resources.

Covers modern hot topics and the intimate coupling of informatics with biology - highlighting the impact of computing advances and ‘big data’ on biology!

Designed for graduates in the biosciences with no programing experience or high level math skills.

Provides a hook for increasing computational and data science competencies in the biosciences - valuable high demand translational skills!

Side Note: **Why stick with this course?**

Provides a hands-on practical introduction to major bioinformatics concepts and resources.

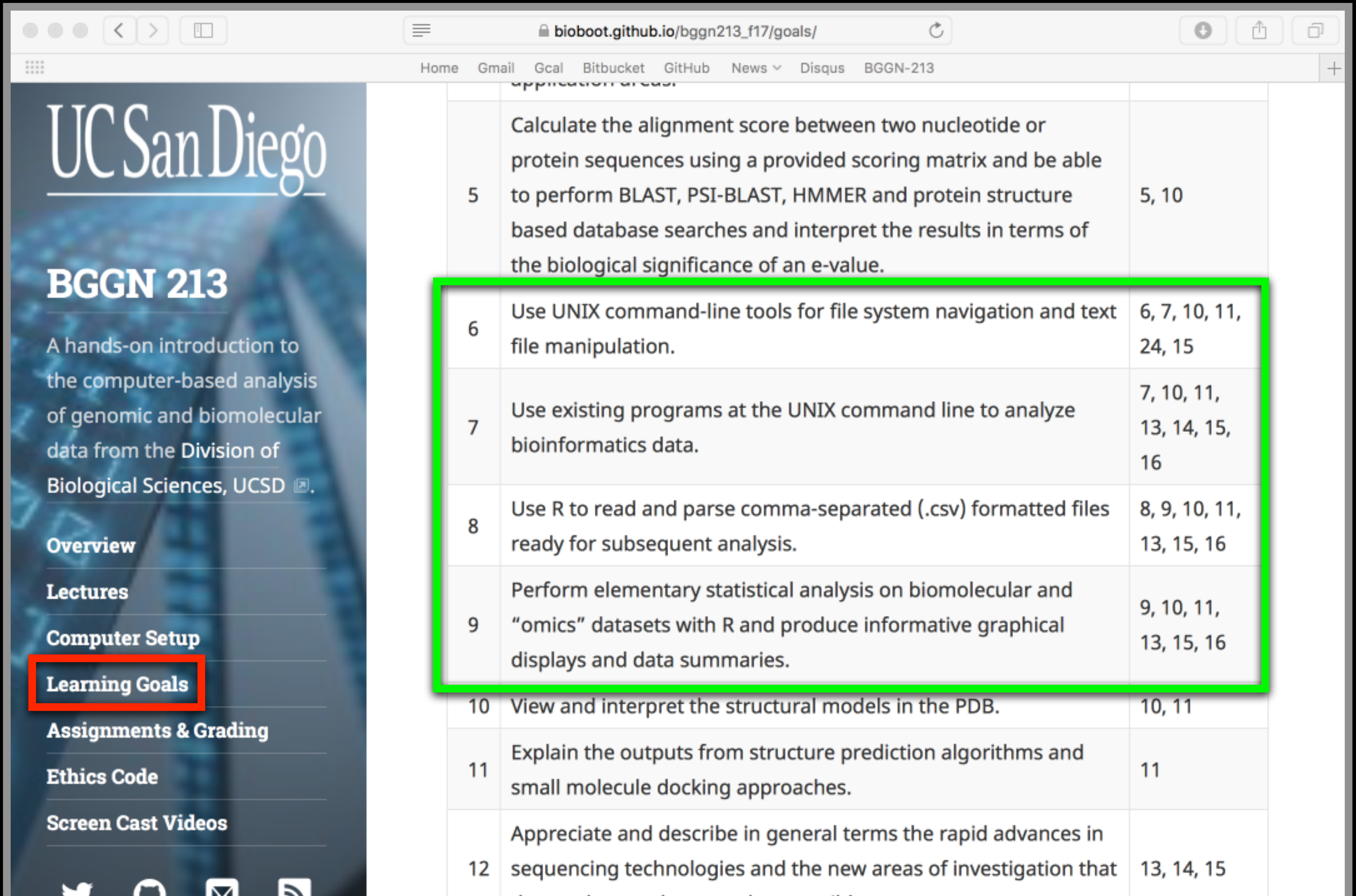
Covers modern hot topics and the intimate coupling of informatics with biology - highlighting the impact of computing advances and 'big data' on biology!

Designed for graduates in the biosciences with no programing experience or high level math skills.

Provides a hook for increasing computational and data science competencies in the biosciences - valuable high demand translational skills!

BGGN-213 Learning Goals….

Advanced UNIX and R based learning goals



BGGN 213

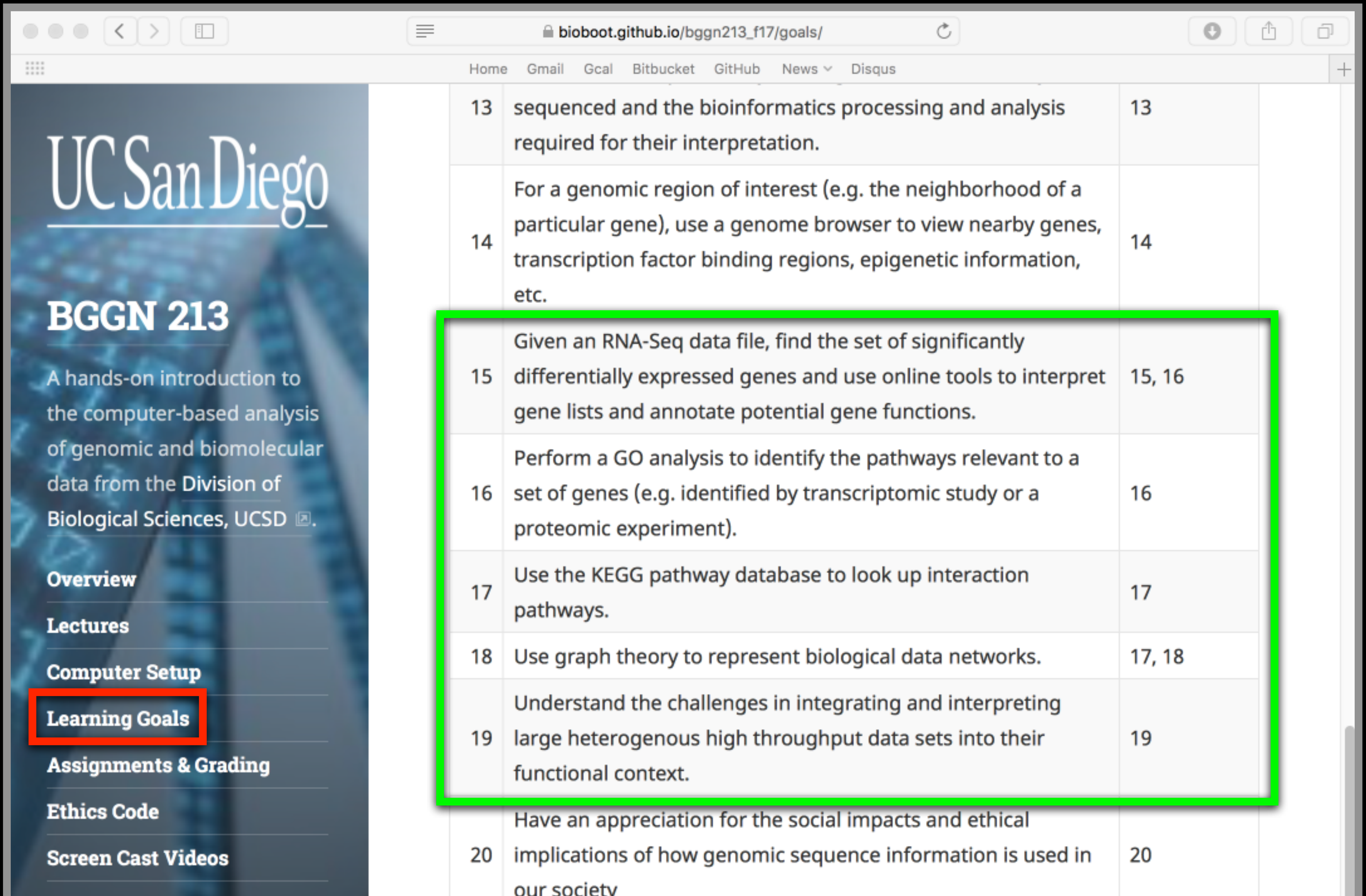
A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

- Overview
- Lectures
- Computer Setup
- Learning Goals
- Assignments & Grading
- Ethics Code
- Screen Cast Videos

	application areas.	
5	Calculate the alignment score between two nucleotide or protein sequences using a provided scoring matrix and be able to perform BLAST, PSI-BLAST, HMMER and protein structure based database searches and interpret the results in terms of the biological significance of an e-value.	5, 10
6	Use UNIX command-line tools for file system navigation and text file manipulation.	6, 7, 10, 11, 24, 15
7	Use existing programs at the UNIX command line to analyze bioinformatics data.	7, 10, 11, 13, 14, 15, 16
8	Use R to read and parse comma-separated (.csv) formatted files ready for subsequent analysis.	8, 9, 10, 11, 13, 15, 16
9	Perform elementary statistical analysis on biomolecular and “omics” datasets with R and produce informative graphical displays and data summaries.	9, 10, 11, 13, 15, 16
10	View and interpret the structural models in the PDB.	10, 11
11	Explain the outputs from structure prediction algorithms and small molecule docking approaches.	11
12	Appreciate and describe in general terms the rapid advances in sequencing technologies and the new areas of investigation that	13, 14, 15

BGGN-213 Learning Goals….

Delve deeper into "real-world" bioinformatics



bioboot.github.io/bggn213_f17/goals/
Home
Gmail
Gcal
Bitbucket
GitHub
News
Disqus
UC San Diego
BGGN 213
A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.
Overview
Lectures
Computer Setup
Learning Goals
Assignments & Grading
Ethics Code
Screen Cast Videos

13	sequenced and the bioinformatics processing and analysis required for their interpretation.	13
14	For a genomic region of interest (e.g. the neighborhood of a particular gene), use a genome browser to view nearby genes, transcription factor binding regions, epigenetic information, etc.	14
15	Given an RNA-Seq data file, find the set of significantly differentially expressed genes and use online tools to interpret gene lists and annotate potential gene functions.	15, 16
16	Perform a GO analysis to identify the pathways relevant to a set of genes (e.g. identified by transcriptomic study or a proteomic experiment).	16
17	Use the KEGG pathway database to look up interaction pathways.	17
18	Use graph theory to represent biological data networks.	17, 18
19	Understand the challenges in integrating and interpreting large heterogenous high throughput data sets into their functional context.	19
20	Have an appreciation for the social impacts and ethical implications of how genomic sequence information is used in our society	20

These support a major learning objective

At the end of this course students will:

- Understand the increasing necessity for computation in modern life sciences research.

- Be able to use and evaluate online bioinformatics resources and analysis tools to solve problems in the biological sciences.

- Be able to use UNIX and the R environment to analyze bioinformatics data at scale.

- Be familiar with the research objectives of the bioinformatics related sub-disciplines of Genome informatics, Transcriptomics and Structural informatics.

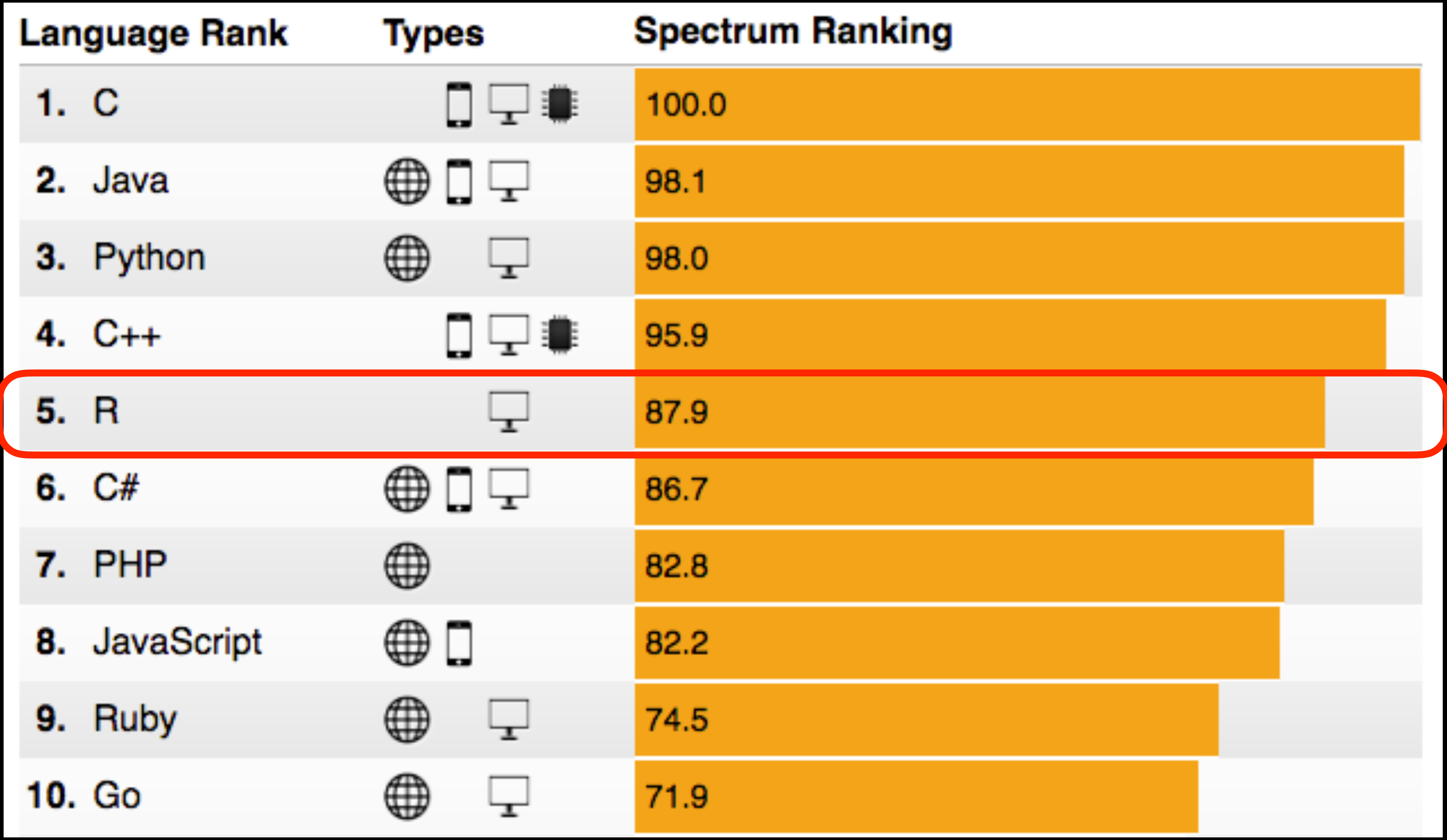
Why use R?

Productivity

Flexibility

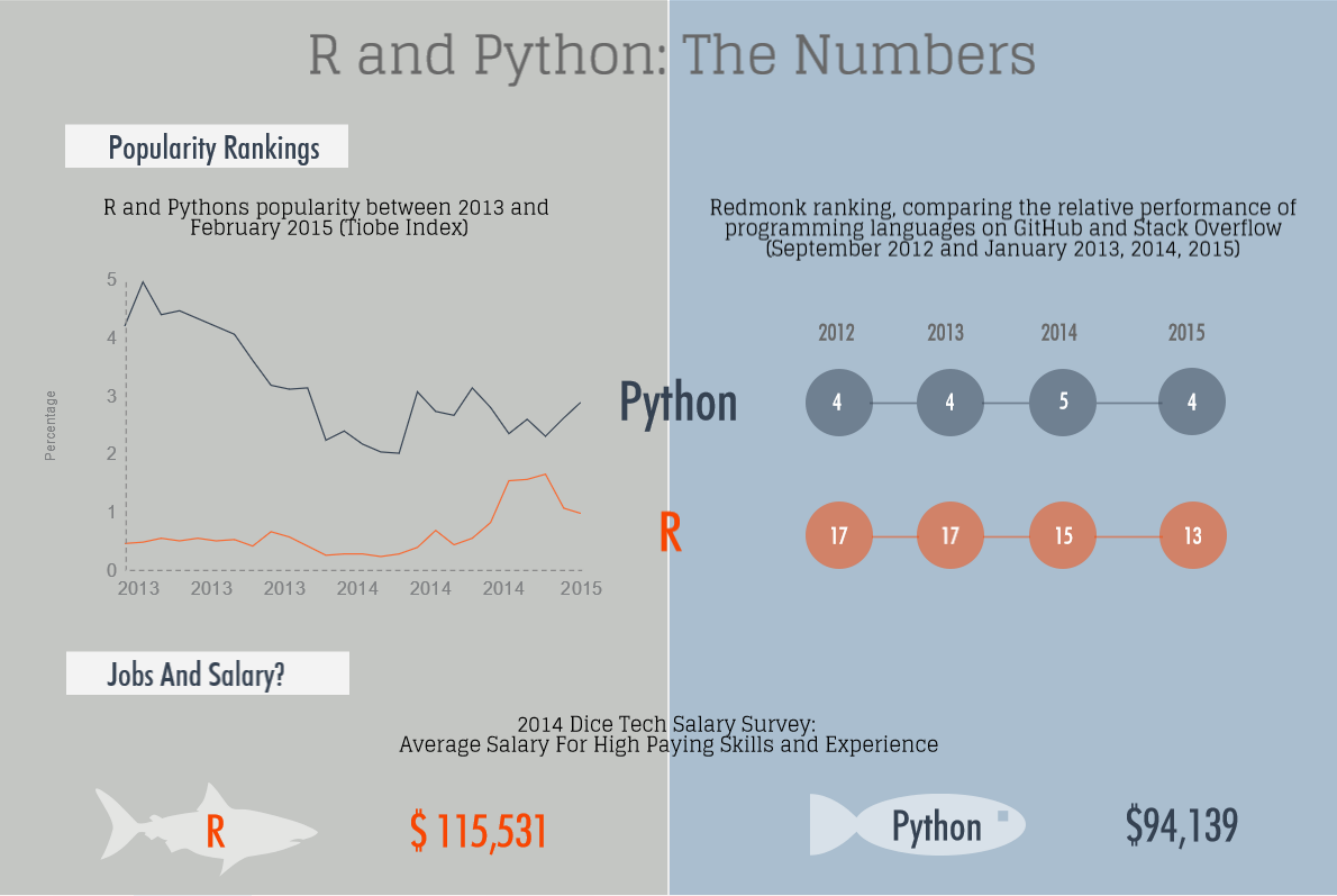
Genomic data analysis

IEEE 2016 Top Programming Languages



Language Rank	Types	Spectrum Ranking
1. C	📱 🖥 ▣	100.0
2. Java	🌐 📱 🖥	98.1
3. Python	🌐 🖥	98.0
4. C++	📱 🖥 ▣	95.9
5. R	🖥	87.9
6. C#	🌐 📱 🖥	86.7
7. PHP	🌐	82.8
8. JavaScript	🌐 📱	82.2
9. Ruby	🌐 🖥	74.5
10. Go	🌐 🖥	71.9

http://spectrum.ieee.org/computing/software/the-2016-top-programming-languages



R and Python: The Numbers

Popularity Rankings

R and Pythons popularity between 2013 and February 2015 (Tiobe Index)

Redmonk ranking, comparing the relative performance of programming languages on GitHub and Stack Overflow (September 2012 and January 2013, 2014, 2015)

	2012	2013	2014	2015
Python	4	4	5	4
R	17	17	15	13

Jobs And Salary?

2014 Dice Tech Salary Survey:
Average Salary For High Paying Skills and Experience

$ 115,531

$94,139

http://www.kdnuggets.com/2015/05/r-vs-python-data-science.html?utm_medium=email&utm_source=flipboard

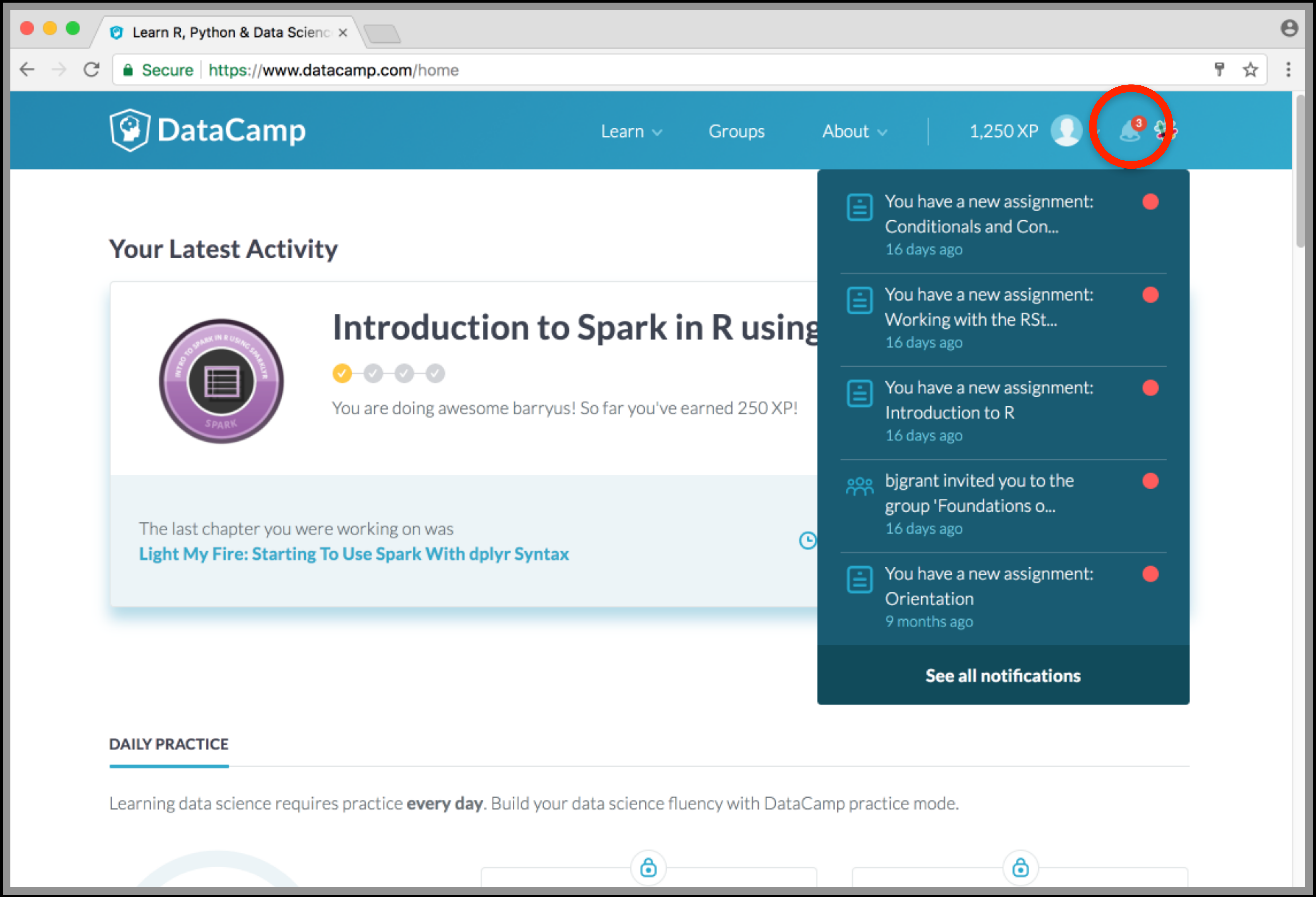
R is designed specifically for data analysis

- Large friendly user and developer community.
  - As of Jan 6th 2019 there are 15,352 add on **R packages** on **CRAN** and 1,823 on **Bioconductor** - much more on these later!

- Virtually every statistical technique is either already built into R, or available as a free package.

- Unparalleled data analysis environment for **high-throughput genomic data**.

< https://www.datacamp.com/ >



Learn R, Python & Data Scienc ×
Secure https://www.datacamp.com/home
DataCamp
Learn
Groups
About
1,250 XP
3

You have a new assignment: Conditionals and Con...
16 days ago

You have a new assignment: Working with the RSt...
16 days ago

You have a new assignment: Introduction to R
16 days ago

bjgrant invited you to the group 'Foundations o...
16 days ago

You have a new assignment: Orientation
9 months ago

See all notifications

Your Latest Activity

Introduction to Spark in R using

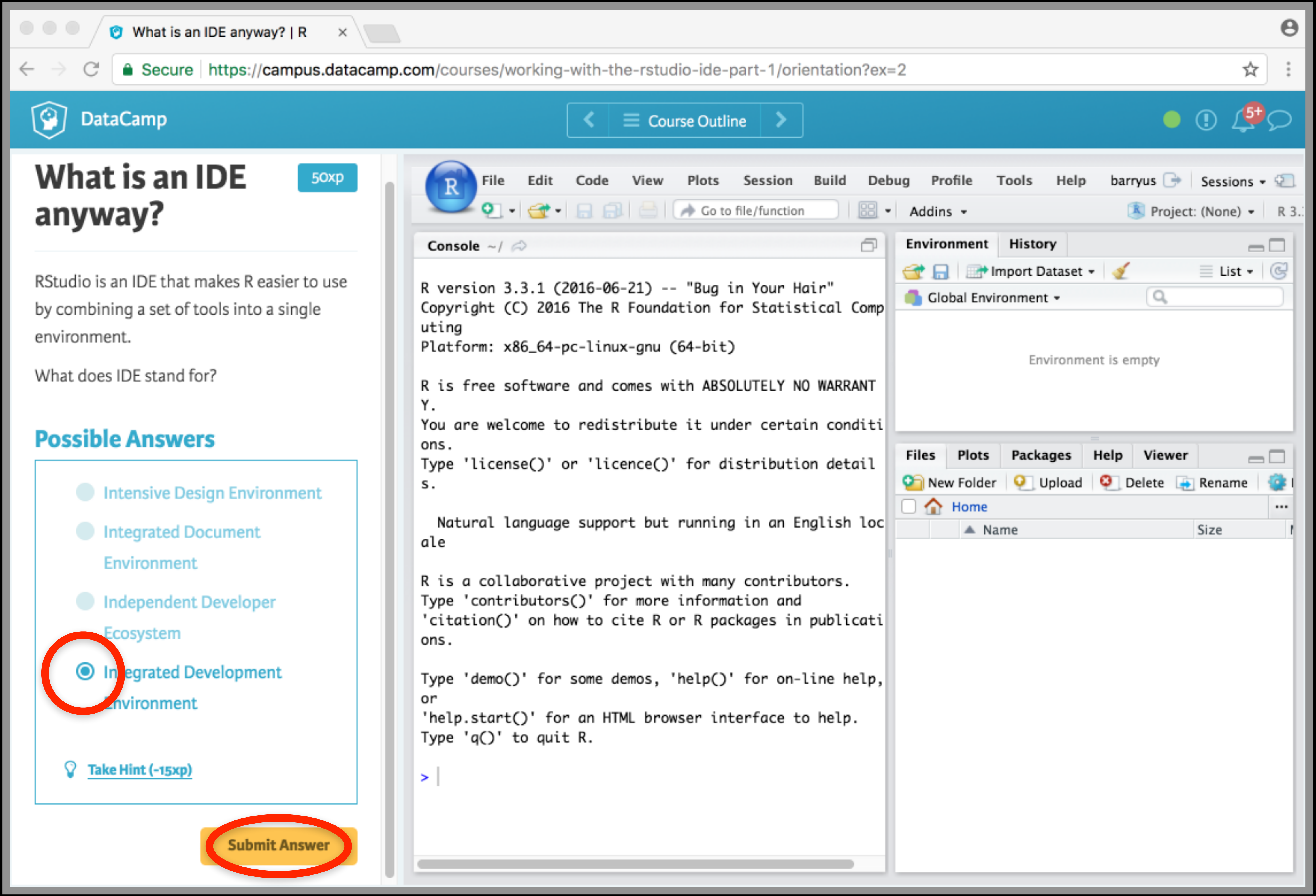
You are doing awesome barryus! So far you've earned 250 XP!

The last chapter you were working on was
Light My Fire: Starting To Use Spark With dplyr Syntax

DAILY PRACTICE

Learning data science requires practice **every day**. Build your data science fluency with DataCamp practice mode.

< https://www.datacamp.com/ >



What is an IDE anyway?

50xp

RStudio is an IDE that makes R easier to use by combining a set of tools into a single environment.

What does IDE stand for?

Possible Answers

○ Intensive Design Environment

○ Integrated Document Environment

○ Independent Developer Ecosystem

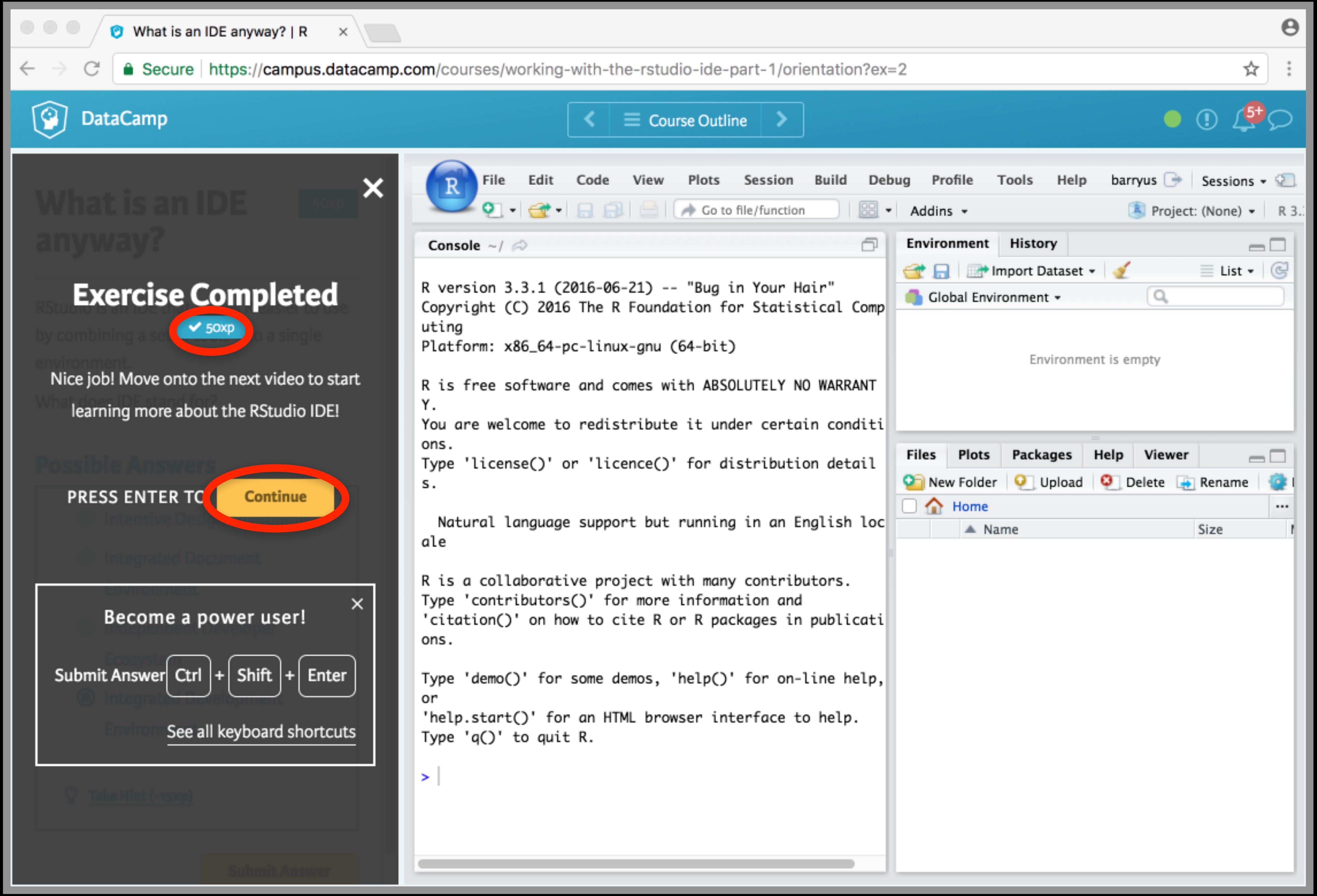
● Integrated Development Environment

Take Hint (-15xp)

Submit Answer

```
R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"
Copyright (C) 2016 The R Foundation for Statistical Comp
uting
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANT
Y.
You are welcome to redistribute it under certain conditi
ons.
Type 'license()' or 'licence()' for distribution detail
s.

  Natural language support but running in an English loc
ale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publicati
ons.

Type 'demo()' for some demos, 'help()' for on-line help,
or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

< https://www.datacamp.com/ >



What is an IDE anyway? | R

Secure | https://campus.datacamp.com/courses/working-with-the-rstudio-ide-part-1/orientation?ex=2

Exercise Completed

✓ 50xp

Nice job! Move onto the next video to start learning more about the RStudio IDE!

PRESS ENTER TC Continue

Become a power user!

Submit Answer Ctrl + Shift + Enter

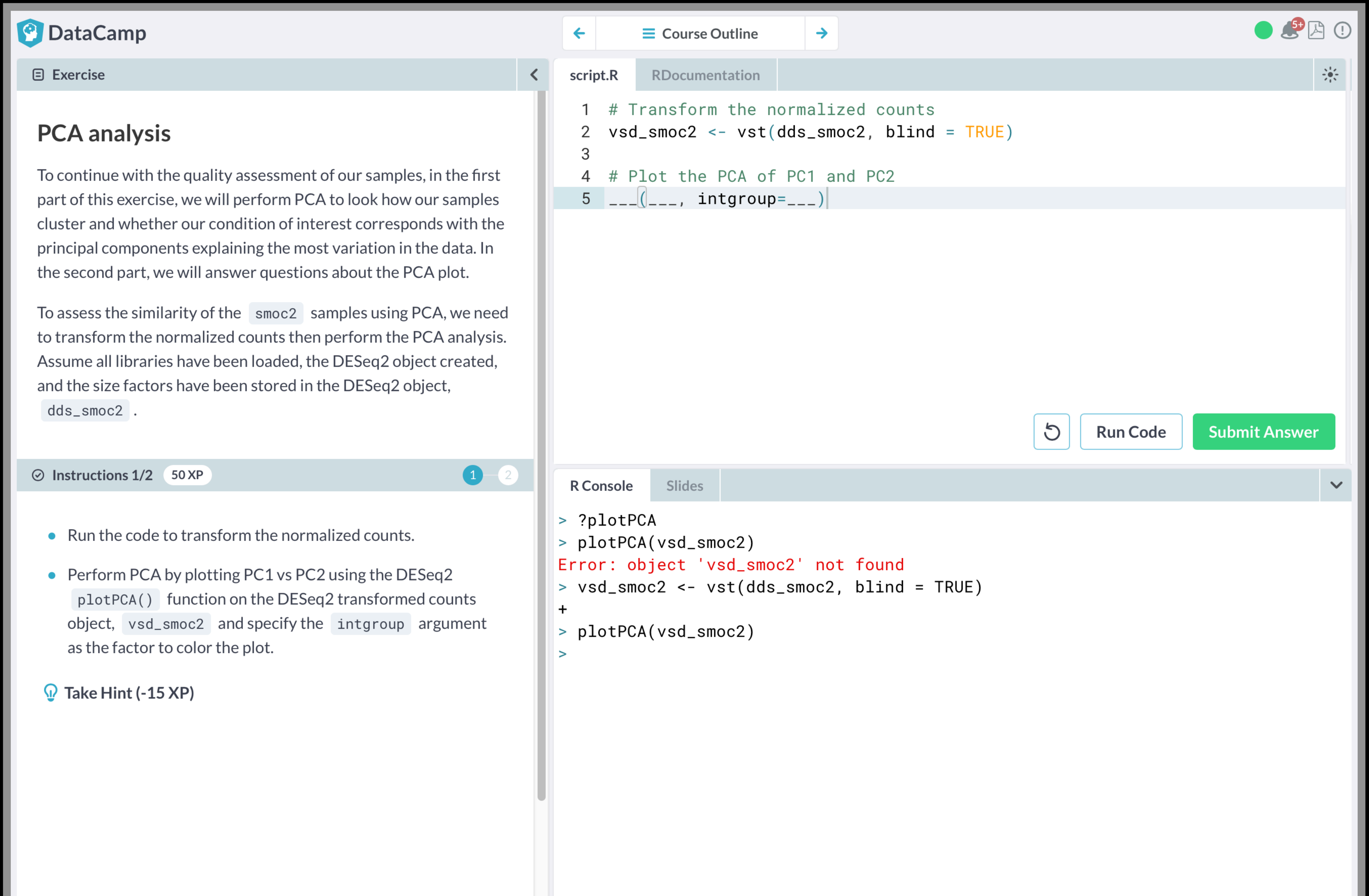
See all keyboard shortcuts

```
R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"
Copyright (C) 2016 The R Foundation for Statistical Comp
uting
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANT
Y.
You are welcome to redistribute it under certain conditi
ons.
Type 'license()' or 'licence()' for distribution detail
s.

  Natural language support but running in an English loc
ale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publicati
ons.

Type 'demo()' for some demos, 'help()' for on-line help,
or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Environment is empty

< https://www.datacamp.com/ >

Homework assignments will be via DataCamp



PCA analysis

To continue with the quality assessment of our samples, in the first part of this exercise, we will perform PCA to look how our samples cluster and whether our condition of interest corresponds with the principal components explaining the most variation in the data. In the second part, we will answer questions about the PCA plot.

To assess the similarity of the `smoc2` samples using PCA, we need to transform the normalized counts then perform the PCA analysis. Assume all libraries have been loaded, the DESeq2 object created, and the size factors have been stored in the DESeq2 object, `dds_smoc2` .

Instructions 1/2 50 XP

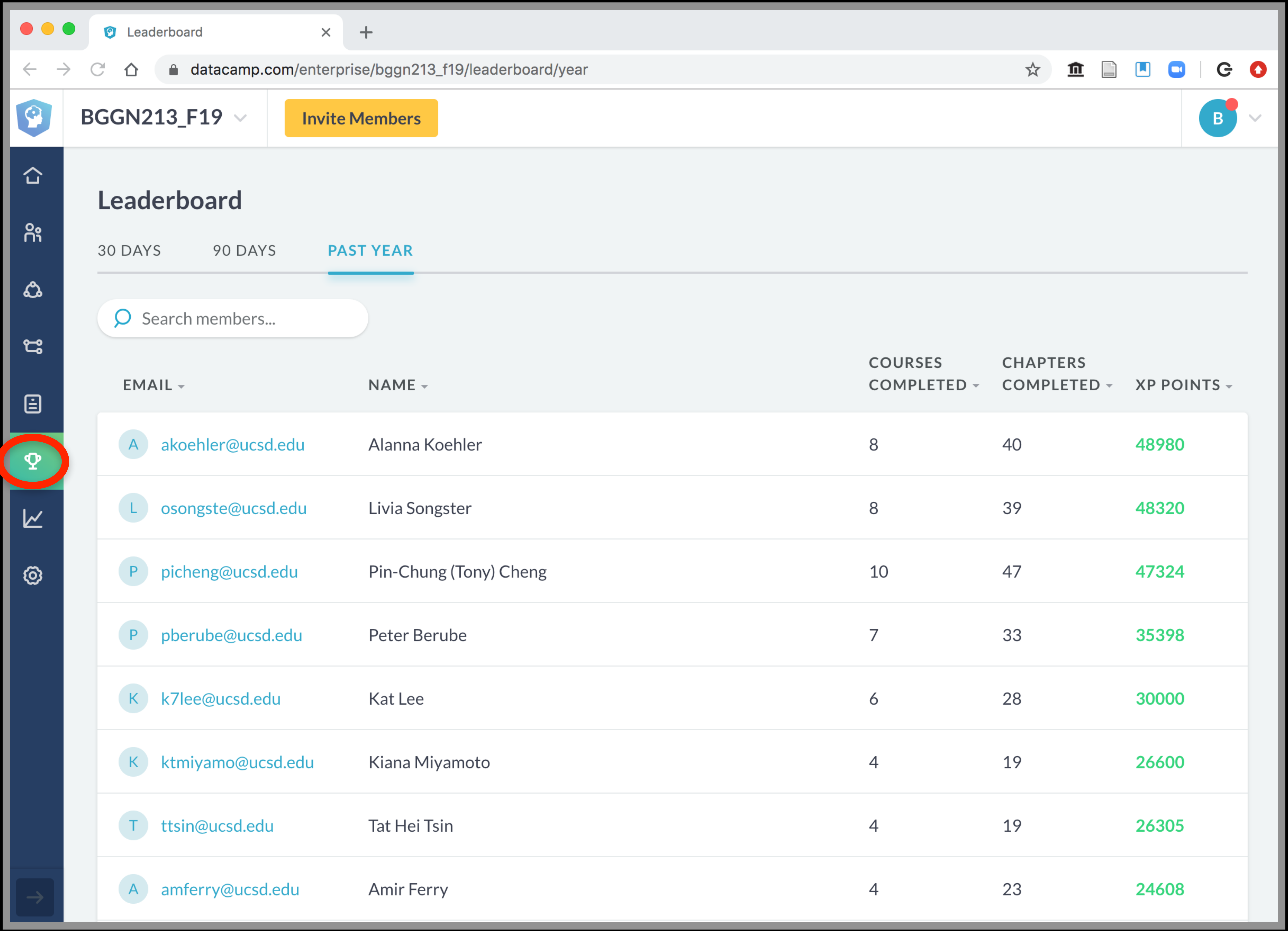
- Run the code to transform the normalized counts.
- Perform PCA by plotting PC1 vs PC2 using the DESeq2 `plotPCA()` function on the DESeq2 transformed counts object, `vsd_smoc2` and specify the `intgroup` argument as the factor to color the plot.

Take Hint (-15 XP)

```
# Transform the normalized counts
vsd_smoc2 <- vst(dds_smoc2, blind = TRUE)

# Plot the PCA of PC1 and PC2
___(___, intgroup=___)
```

```
> ?plotPCA
> plotPCA(vsd_smoc2)
Error: object 'vsd_smoc2' not found
> vsd_smoc2 <- vst(dds_smoc2, blind = TRUE)
+
> plotPCA(vsd_smoc2)
>
```

< https://www.datacamp.com/ >



Leaderboard

30 DAYS | 90 DAYS | PAST YEAR

Search members...

EMAIL	NAME	COURSES COMPLETED	CHAPTERS COMPLETED	XP POINTS
A akoehler@ucsd.edu	Alanna Koehler	8	40	48980
L osongste@ucsd.edu	Livia Songster	8	39	48320
P picheng@ucsd.edu	Pin-Chung (Tony) Cheng	10	47	47324
P pberube@ucsd.edu	Peter Berube	7	33	35398
K k7lee@ucsd.edu	Kat Lee	6	28	30000
K ktmiyamo@ucsd.edu	Kiana Miyamoto	4	19	26600
T ttsin@ucsd.edu	Tat Hei Tsin	4	19	26305
A amferry@ucsd.edu	Amir Ferry	4	23	24608

Today's Menu

Course Logistics	Website, screencasts, survey, ethics, assessment and grading.
Learning Objectives	What you need to learn to succeed in this course.
Course Structure	Major lecture topics and specific leaning goals.
Introduction to Bioinformatis	Introducing the *what*, *why* and *how* of bioinformatics?
Computer Setup	Ensuring your laptop is all set for future sections of this course.

“What is Bioinformatics?”

“*Bioinformatics is the application of computers to the collection, archiving, organization, and analysis of biological data.*”

… A hybrid of biology and computer science

“*Bioinformatics is the application of computers to the collection, archiving, organization, and analysis of biological data.*”

Bioinformatics is computer aided biology!

“*Bioinformatics is the application of computers to the collection, archiving, organization, and analysis of biological data*.”

Bioinformatics is computer aided biology!

Goal: Data to Knowledge

Side-Note:

There are many useful definitions...

- "Computer based management and analysis of biological and biomedical data with useful applications in many disciplines, particularly genomics, proteomics, metabolomics, and related fields."
  (BGGN-213)

- "Bioinformatics is conceptualizing biology in terms of macromolecules and then applying "informatics" techniques (derived from disciplines such as applied maths, computer science, and statistics) to understand and organize the information associated with these molecules, on a large-scale."
  (Luscombe *et al.* 2001)

- "Bioinformatics is research, development, or application of computational approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize and analyze such data ...<cut>..."
  (National Institutes of Health: http://tinyurl.com/l3gxr6b)

Side-Note:

There are many useful definitions...

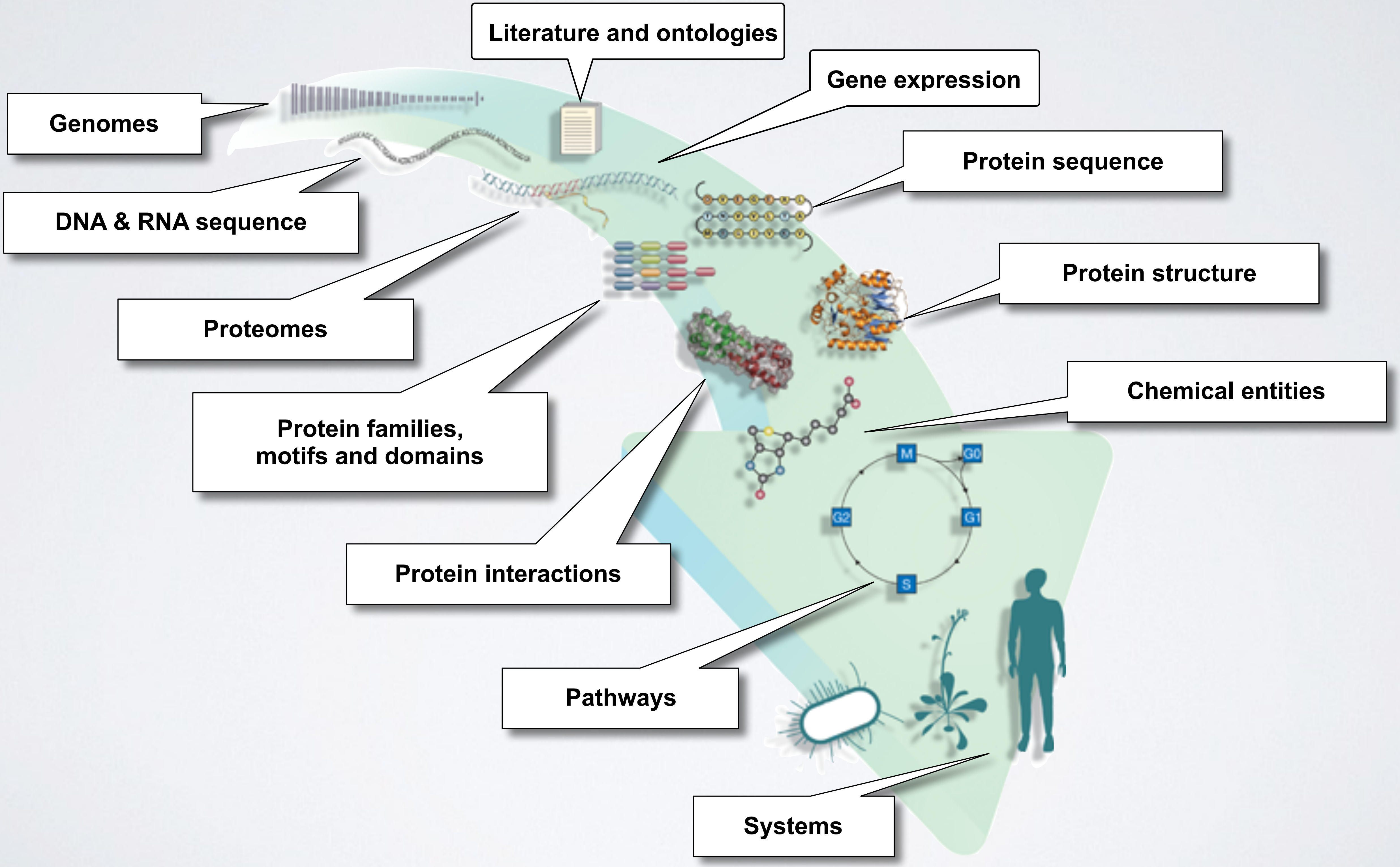
- "Computer based management and analysis of biological and biomedical data with useful applications in many disciplines, particularly genomics, proteomics, metabolomics, and rel[illegible]d fields."
  (BGGN-213)

- "Bioinformatics is conceptualizing biolog[illegible] of macromolecules and then applyin[illegible]tics" techniques (derived from disciplines such as appl[illegible], computer science, and statistics) to understand [illegible]anize the information associated with these molecules, [illegible]e-scale."
  (Luscomb[illegible]001)

- "B[illegible]atics is research, development, or application of co[illegible]putational approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize and analyze such data ...<cut>..."
  (National Institutes of Health: http://tinyurl.com/l3gxr6b)

Key Point: Bioinformatics is Computer Aided Biology

Major types of Bioinformatics Data



Literature and ontologies

Genomes

Gene expression

Protein sequence

DNA & RNA sequence

Protein structure

Proteomes

Chemical entities

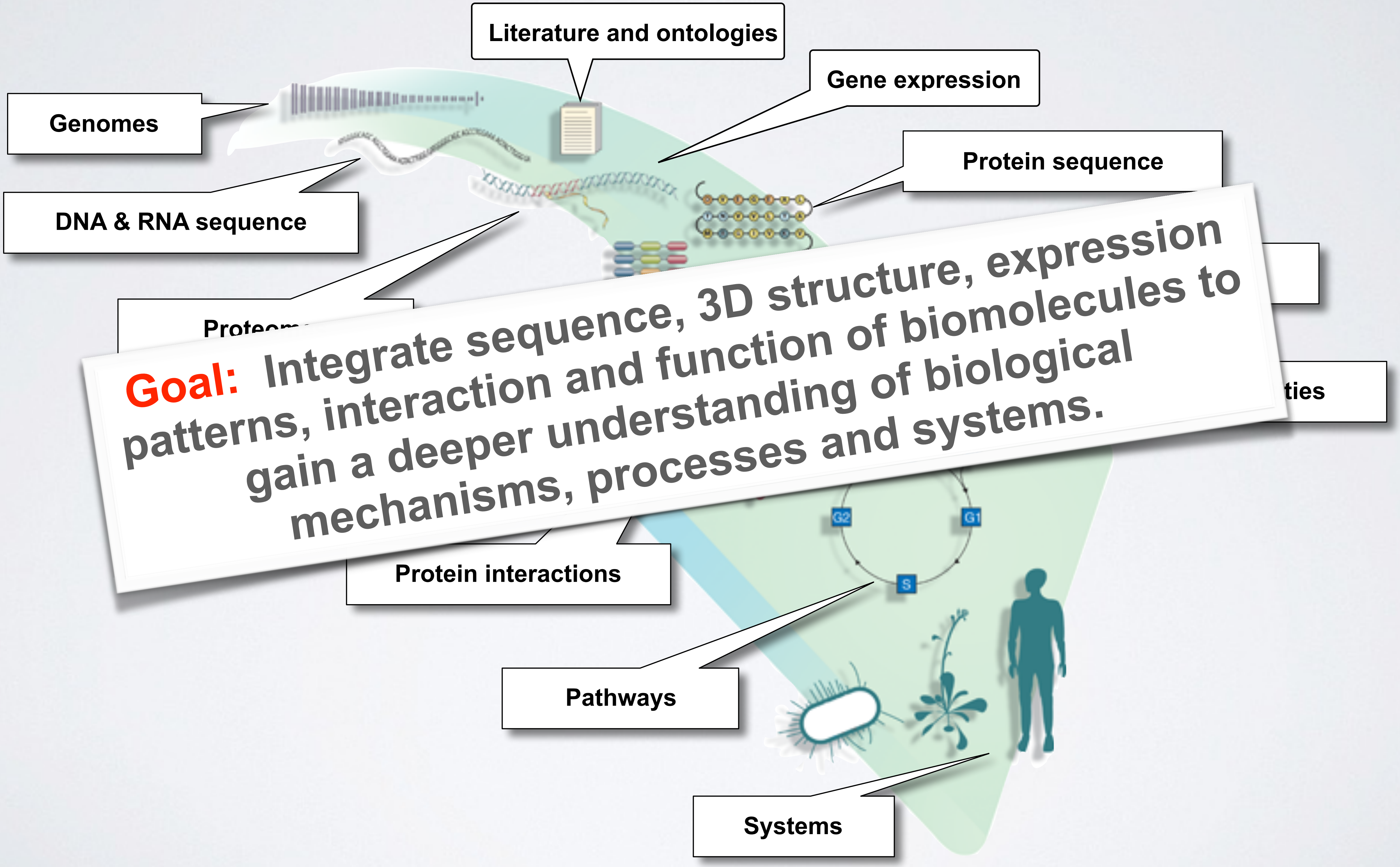
Protein families, motifs and domains

Protein interactions

Pathways

Systems

Major types of Bioinformatics Data



Literature and ontologies

Genomes

Gene expression

Protein sequence

DNA & RNA sequence

Proteom

ties

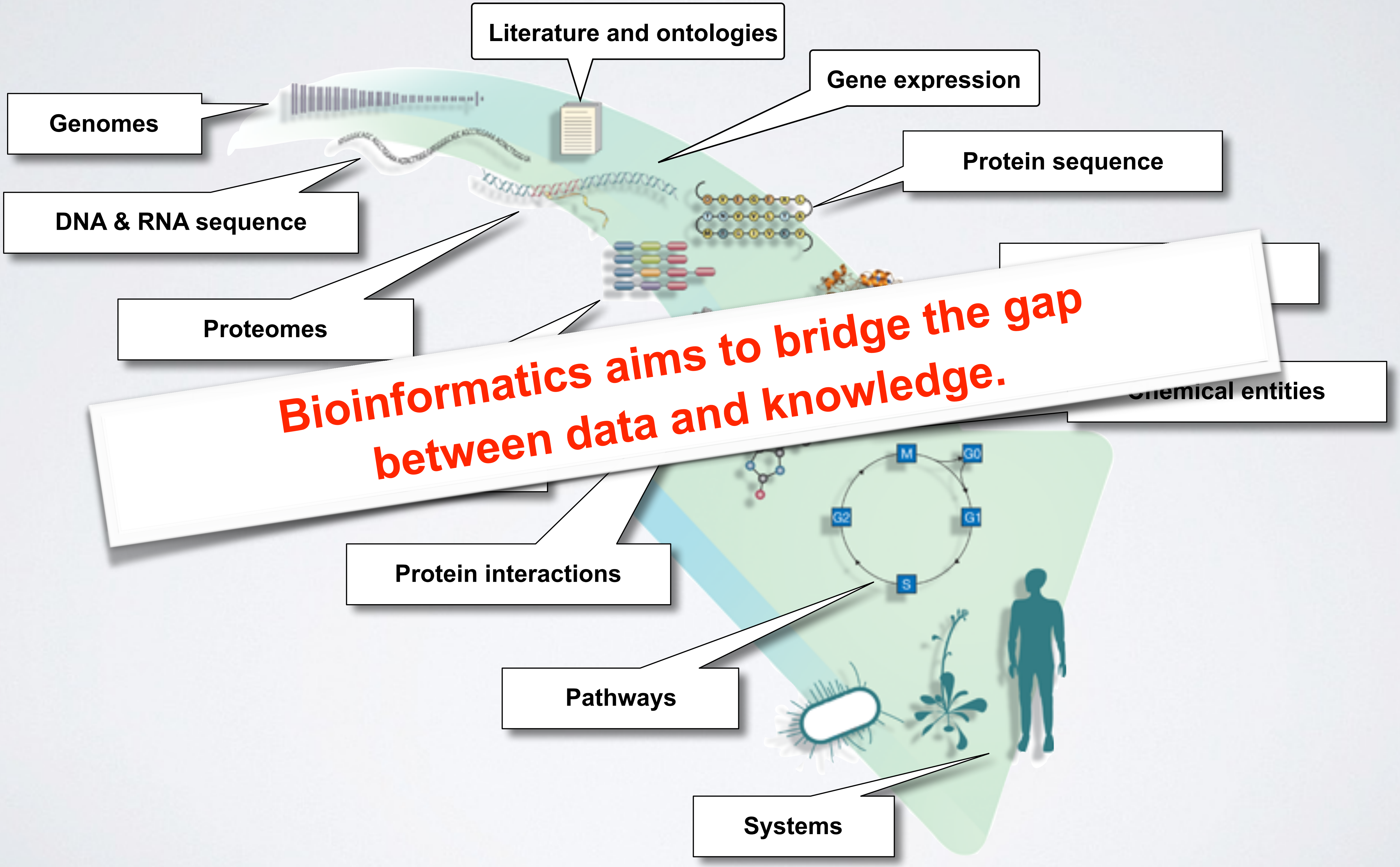
Protein interactions

Pathways

Systems

Goal: Integrate sequence, 3D structure, expression patterns, interaction and function of biomolecules to gain a deeper understanding of biological mechanisms, processes and systems.

Major types of Bioinformatics Data



Literature and ontologies

Genomes

Gene expression

Protein sequence

DNA & RNA sequence

Proteomes

Chemical entities

Protein interactions

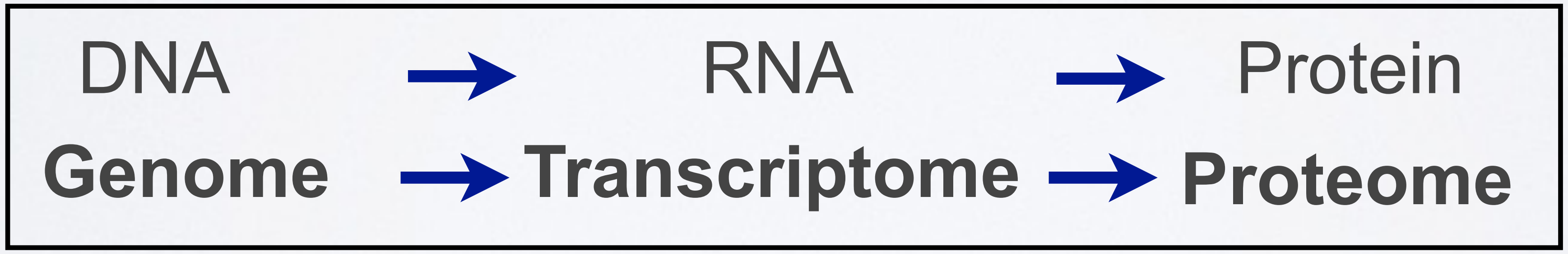
Pathways

Systems

Bioinformatics aims to bridge the gap between data and knowledge.

How do we do Bioinformatics?

- A “*bioinformatics approach*” involves the application of **computer algorithms**, **computer models** and **computer databases** with the broad goal of understanding the action of both individual genes, transcripts, proteins and large collections of these entities.



DNA → RNA → Protein

Genome → Transcriptome → Proteome

x 1,000

x 100,000

How do we *actually* do Bioinformatics?

Pre-packaged tools and databases

- Many online
- Most are free to use
- Time consuming methods require downloading…

Advanced tool application & development

- Mostly on a **UNIX** environment
- Knowledge of programing languages frequently required (*e.g.* **R**, Python, Perl, C, Java, Fortran)
- May require specialized high performance computing…

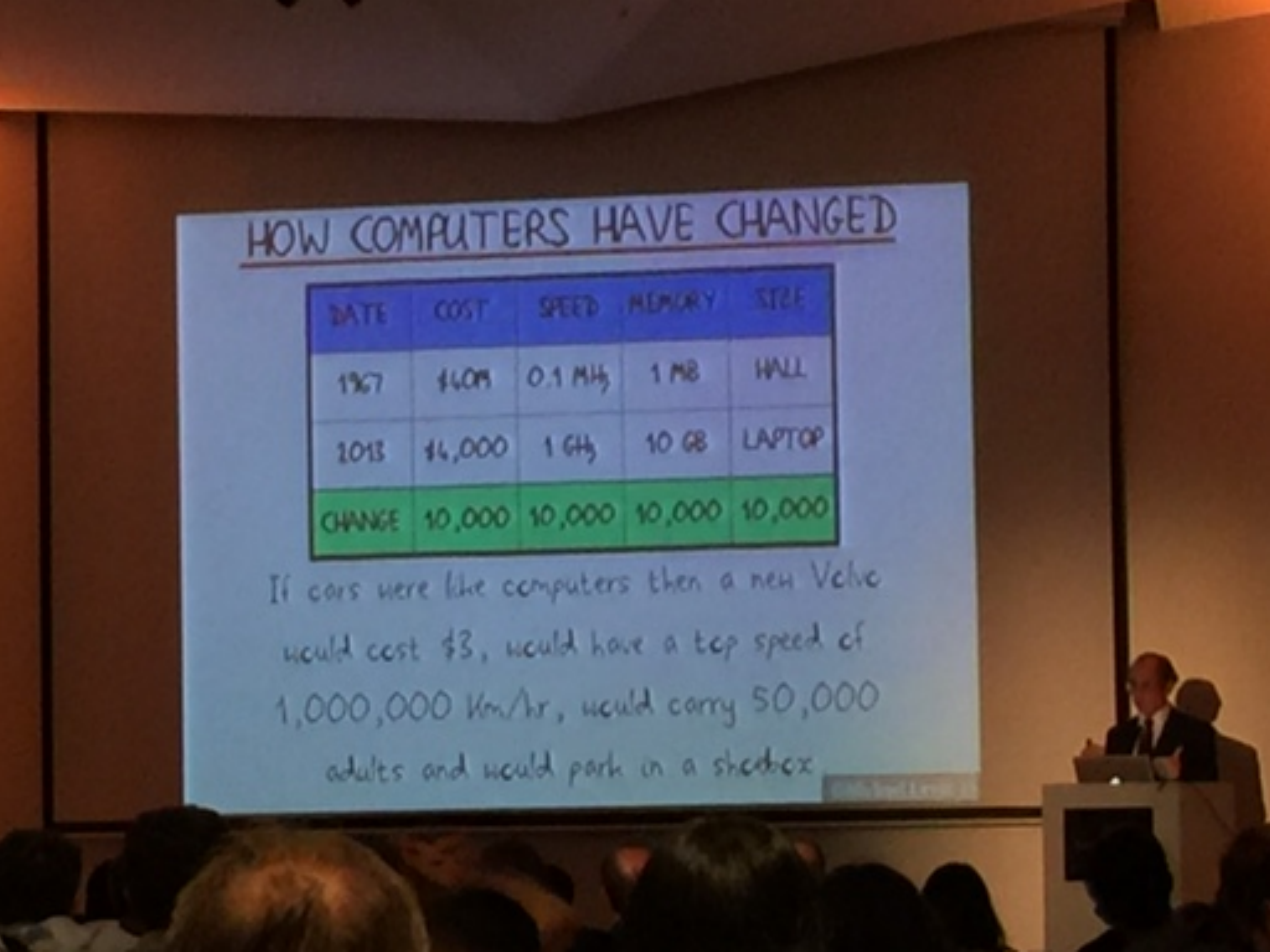
How do we *actually* do Bioinformatics?

Pre-packaged tools and databases

- Many online
- Most are free to use
- Time consuming methods require downloading…

Advanced tool application & development

- Mostly on a UNIX environment
- Knowledge of programing languages frequently required (*e.g.* **R**, Python, Perl, C, Java, Fortran)
- May require specialized high performance computing…

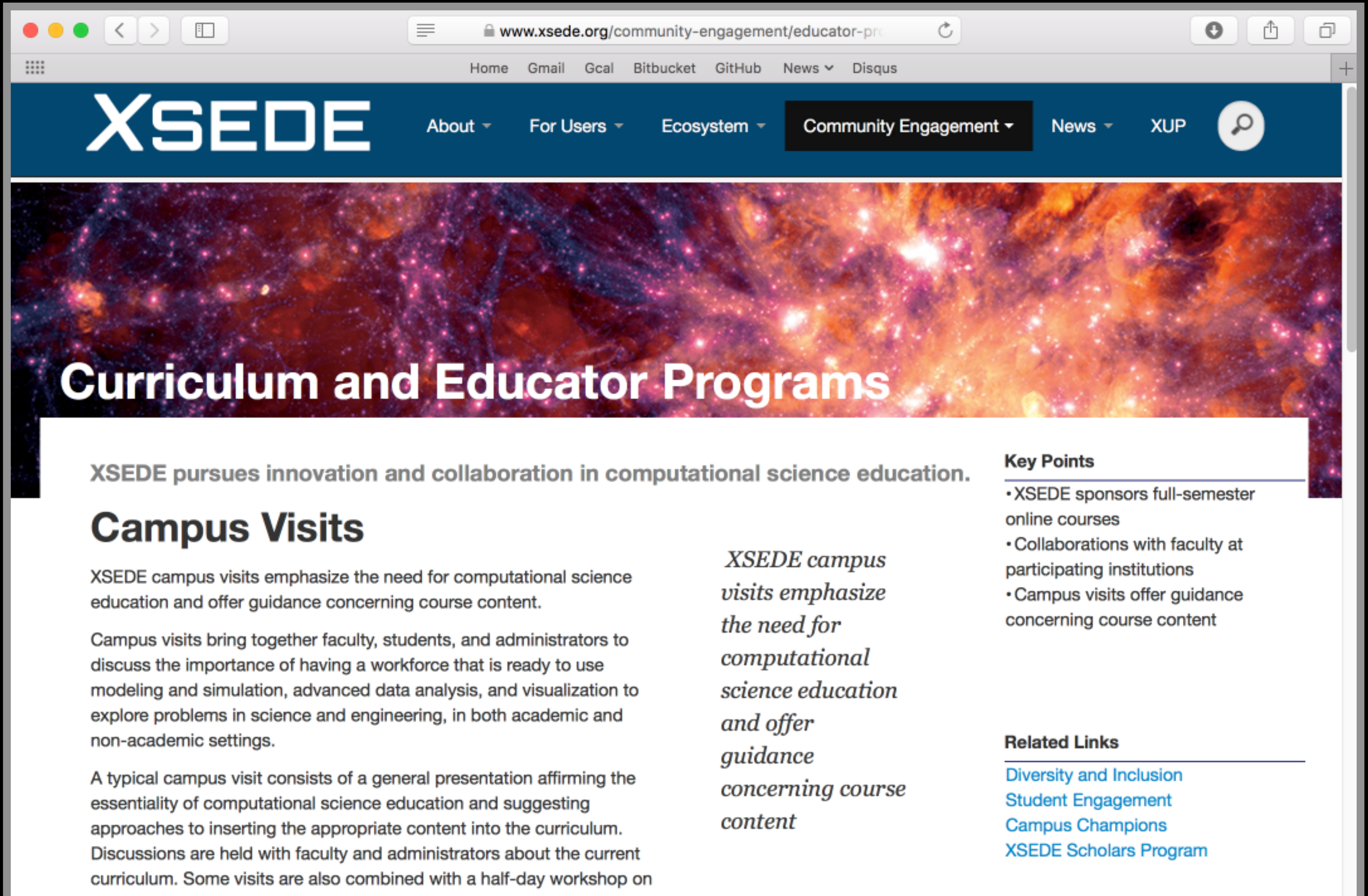


HOW COMPUTERS HAVE CHANGED

DATE	COST	SPEED	MEMORY	SIZE
1967	$60M	0.1 MHz	1 MB	HALL
2013	$6,000	1 GHz	10 GB	LAPTOP
CHANGE	10,000	10,000	10,000	10,000

If cars were like computers then a new Volvo would cost $3, would have a top speed of 1,000,000 Km/hr, would carry 50,000 adults and would park in a shoebox

NSF Extreme Science and Engineering Discovery Environment (XSEDE)



Curriculum and Educator Programs

XSEDE pursues innovation and collaboration in computational science education.

Campus Visits

XSEDE campus visits emphasize the need for computational science education and offer guidance concerning course content.

Campus visits bring together faculty, students, and administrators to discuss the importance of having a workforce that is ready to use modeling and simulation, advanced data analysis, and visualization to explore problems in science and engineering, in both academic and non-academic settings.

A typical campus visit consists of a general presentation affirming the essentiality of computational science education and suggesting approaches to inserting the appropriate content into the curriculum. Discussions are held with faculty and administrators about the current curriculum. Some visits are also combined with a half-day workshop on

XSEDE campus visits emphasize the need for computational science education and offer guidance concerning course content

Key Points

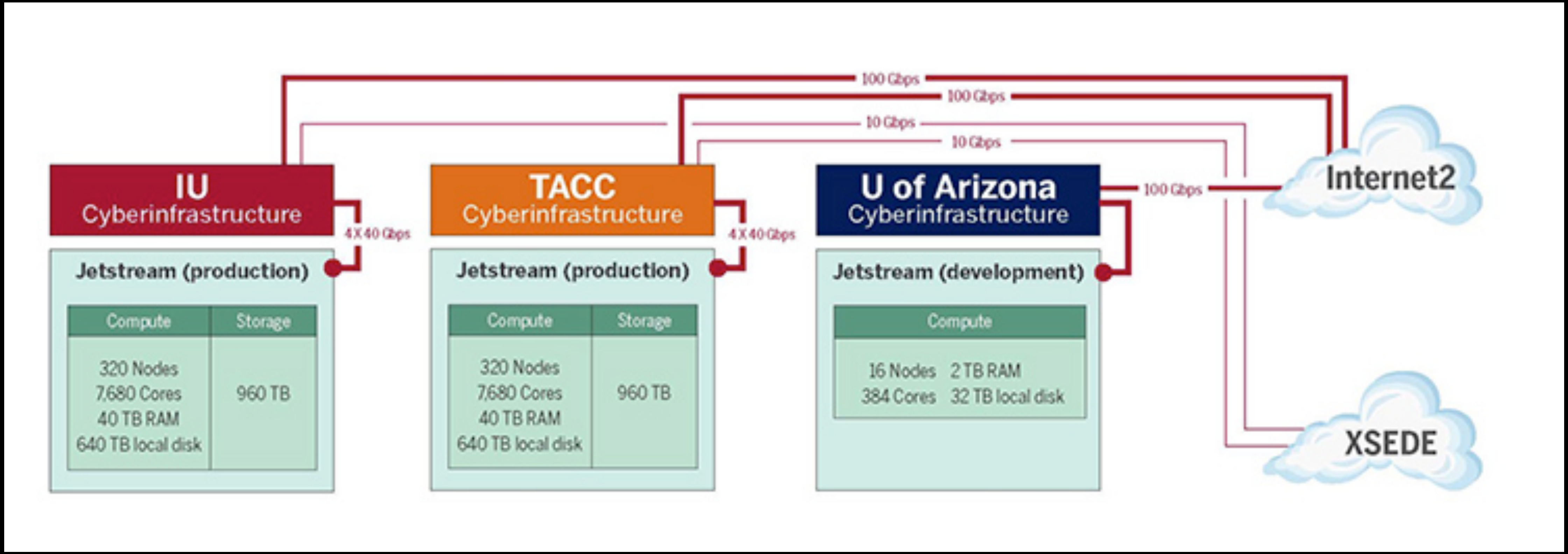
- XSEDE sponsors full-semester online courses
- Collaborations with faculty at participating institutions
- Campus visits offer guidance concerning course content

Related Links

Diversity and Inclusion
Student Engagement
Campus Champions
XSEDE Scholars Program

What is *Jetstream*?

- A new cloud computing environment based at Indiana University and the Texas Advanced Computing Center (TACC) providing on-demand access to interactive computing and data analysis resources.



IU Cyberinfrastructure

Jetstream (production)

Compute	Storage
320 Nodes 7,680 Cores 40 TB RAM 640 TB local disk	960 TB

TACC Cyberinfrastructure

Jetstream (production)

Compute	Storage
320 Nodes 7,680 Cores 40 TB RAM 640 TB local disk	960 TB

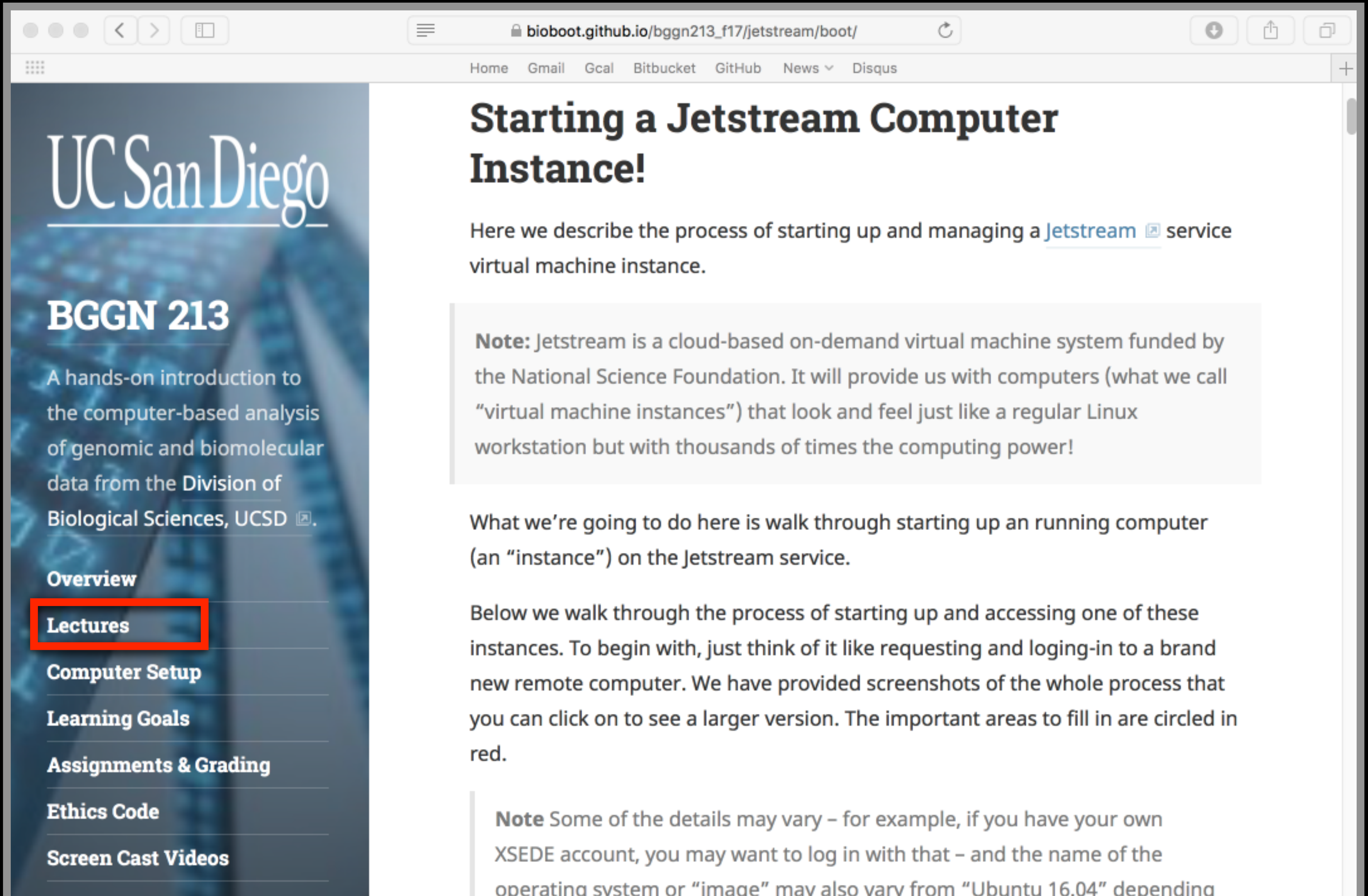
U of Arizona Cyberinfrastructure

Jetstream (development)

Compute	
16 Nodes 384 Cores	2 TB RAM 32 TB local disk

Jetstream tutorials

Developed *user friendly* labs for Jetstream basics



bioboot.github.io/bggn213_f17/jetstream/boot/
Home Gmail Gcal Bitbucket GitHub News Disqus
UC San Diego

BGGN 213

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

- Overview
- Lectures
- Computer Setup
- Learning Goals
- Assignments & Grading
- Ethics Code
- Screen Cast Videos

Starting a Jetstream Computer Instance!

Here we describe the process of starting up and managing a Jetstream service virtual machine instance.

> **Note:** Jetstream is a cloud-based on-demand virtual machine system funded by the National Science Foundation. It will provide us with computers (what we call "virtual machine instances") that look and feel just like a regular Linux workstation but with thousands of times the computing power!

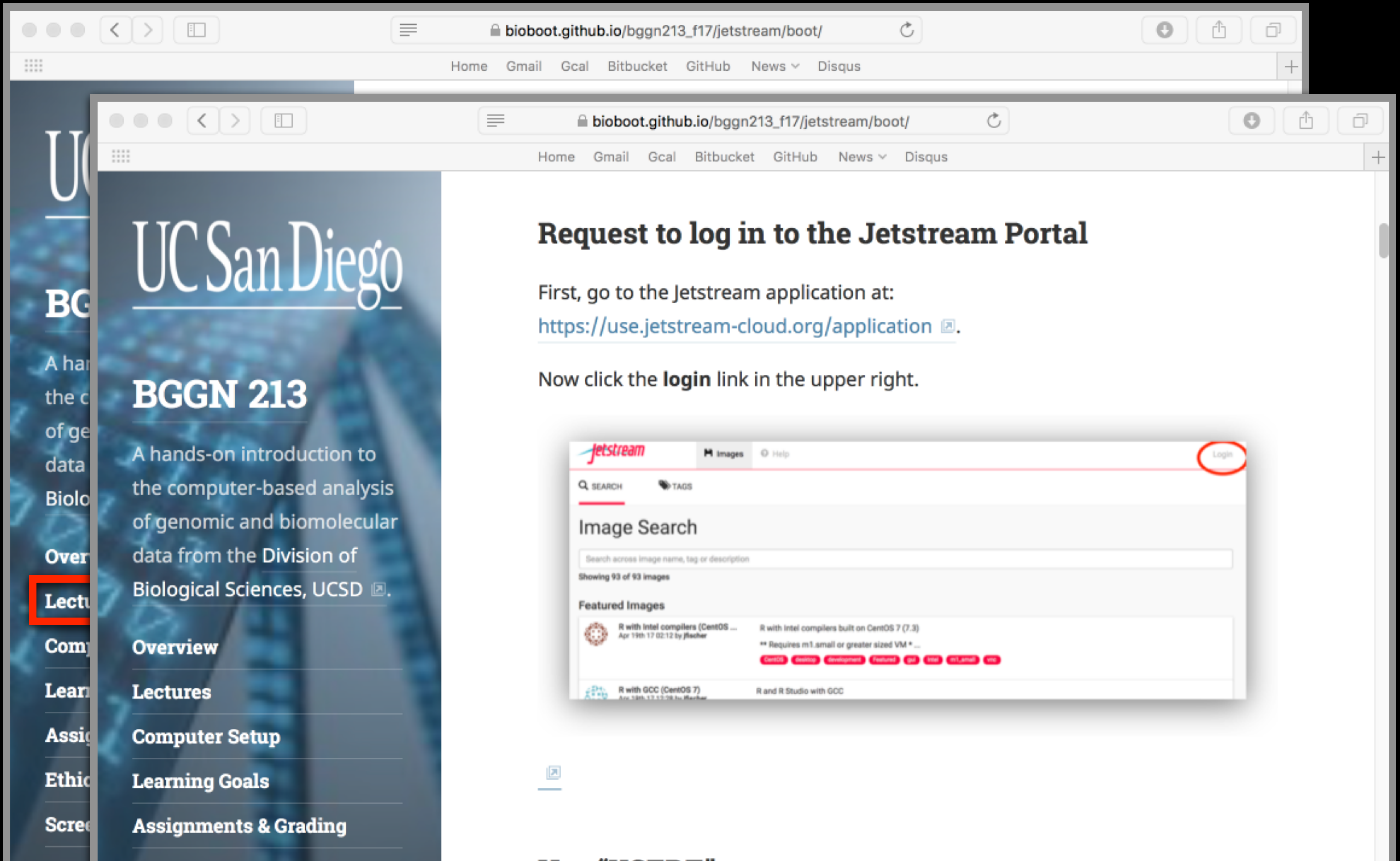
What we're going to do here is walk through starting up an running computer (an "instance") on the Jetstream service.

Below we walk through the process of starting up and accessing one of these instances. To begin with, just think of it like requesting and loging-in to a brand new remote computer. We have provided screenshots of the whole process that you can click on to see a larger version. The important areas to fill in are circled in red.

> **Note** Some of the details may vary – for example, if you have your own XSEDE account, you may want to log in with that – and the name of the operating system or "image" may also vary from "Ubuntu 16.04" depending

Jetstream tutorials

Developed *user friendly* labs for Jetstream basics



BGGN 213

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

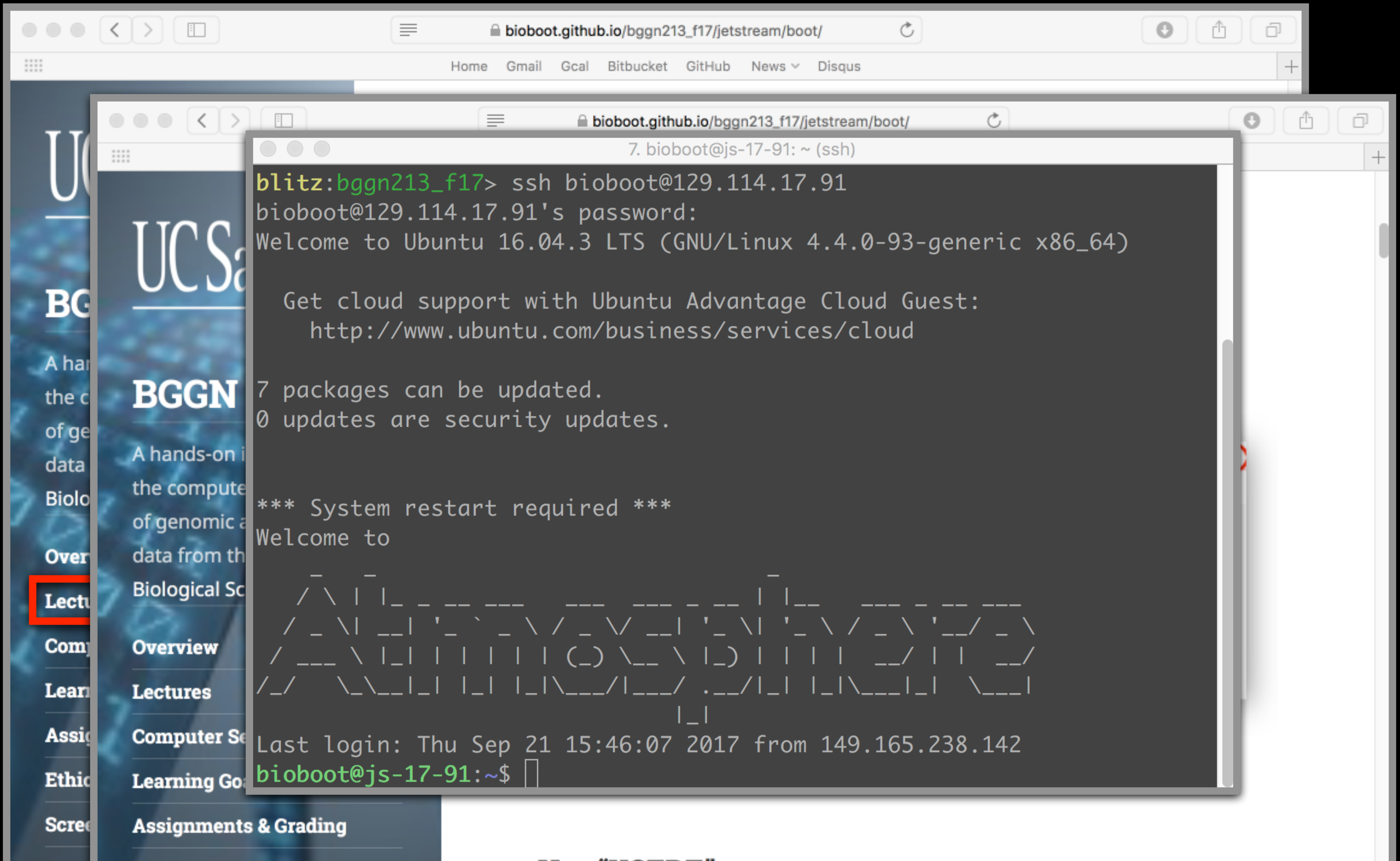
Request to log in to the Jetstream Portal

First, go to the Jetstream application at: https://use.jetstream-cloud.org/application.

Now click the **login** link in the upper right.

Jetstream tutorials

Developed *user friendly* labs for Jetstream basics



```
blitz:bggn213_f17> ssh bioboot@129.114.17.91
bioboot@129.114.17.91's password:
Welcome to Ubuntu 16.04.3 LTS (GNU/Linux 4.4.0-93-generic x86_64)

  Get cloud support with Ubuntu Advantage Cloud Guest:
    http://www.ubuntu.com/business/services/cloud

7 packages can be updated.
0 updates are security updates.


*** System restart required ***
Welcome to
    _   _                             _
   / \ | |_ _ __ ___   ___  ___ _ __ | |__   ___ _ __ ___
  / _ \| __| '_ ` _ \ / _ \/ __| '_ \| '_ \ / _ \ '__/ _ \
 / ___ \ |_| | | | | | (_) \__ \ |_) | | | |  __/ | |  __/
/_/   \_\__|_| |_| |_|\___/|___/ .__/|_| |_|\___|_|  \___|
                               |_|
Last login: Thu Sep 21 15:46:07 2017 from 149.165.238.142
bioboot@js-17-91:~$
```

Skepticism & Bioinformatics

We have to approach computational results the same way we do wet-lab results:

- Do they make sense?
- Is it what we expected?
- Do we have adequate controls, and how did they come out?
- Modeling is modeling, but biology is different...

  What does this model actually contribute?

- Avoid the miss-use of ‘black boxes’

Skepticism & Bioinformatics

Gunnar von Heijne in *"Sequence Analysis in Molecular Biology; Treasure Trove or Trivial Pursuit"* states:

- ➡"Think about what you're doing; use your knowledge of the molecular system involved to guide both your interpretation of results and your direction of inquiry; use as much information as possible; and do not blindly accept everything the computer offers you".

Key-Point: **Avoid the miss-use of 'black boxes'!**

Common problems with Bioinformatics

Confusing multitude of tools available

- Each with many options and settable parameters

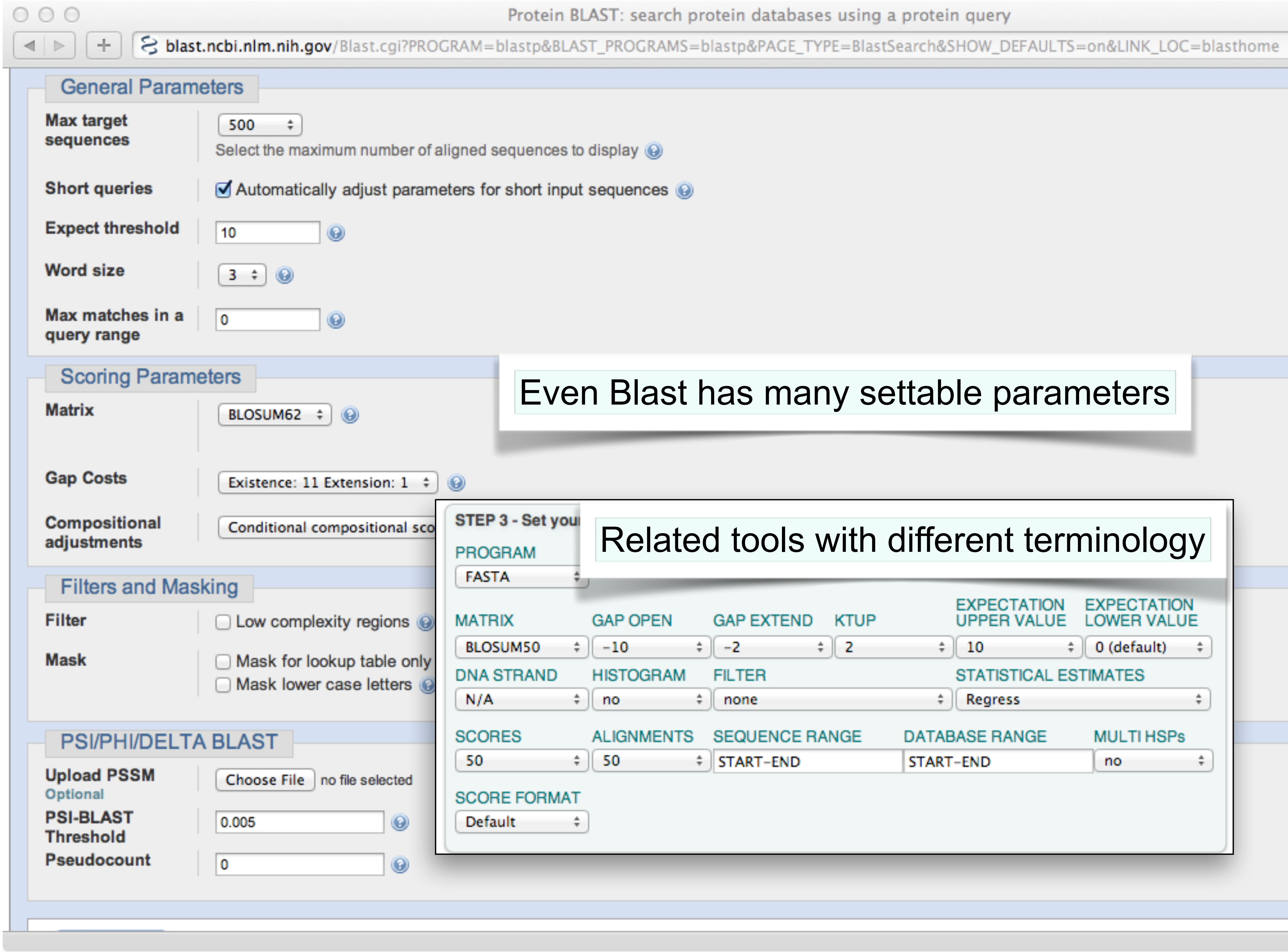
Most tools and databases are written by and for nerds

- Same is true of documentation - if any exists!

Most are developed independently

Notable exceptions are found at the:

- **EBI** (European Bioinformatics Institute) and
- **NCBI** (National Center for Biotechnology Information)



Protein BLAST: search protein databases using a protein query

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome

General Parameters

Max target sequences: 500
Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

Max matches in a query range: 0

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional sco

Filters and Masking

Filter: □ Low complexity regions

Mask: □ Mask for lookup table only
□ Mask lower case letters

PSI/PHI/DELTA BLAST

Upload PSSM Optional: Choose File no file selected

PSI-BLAST Threshold: 0.005

Pseudocount: 0

Even Blast has many settable parameters

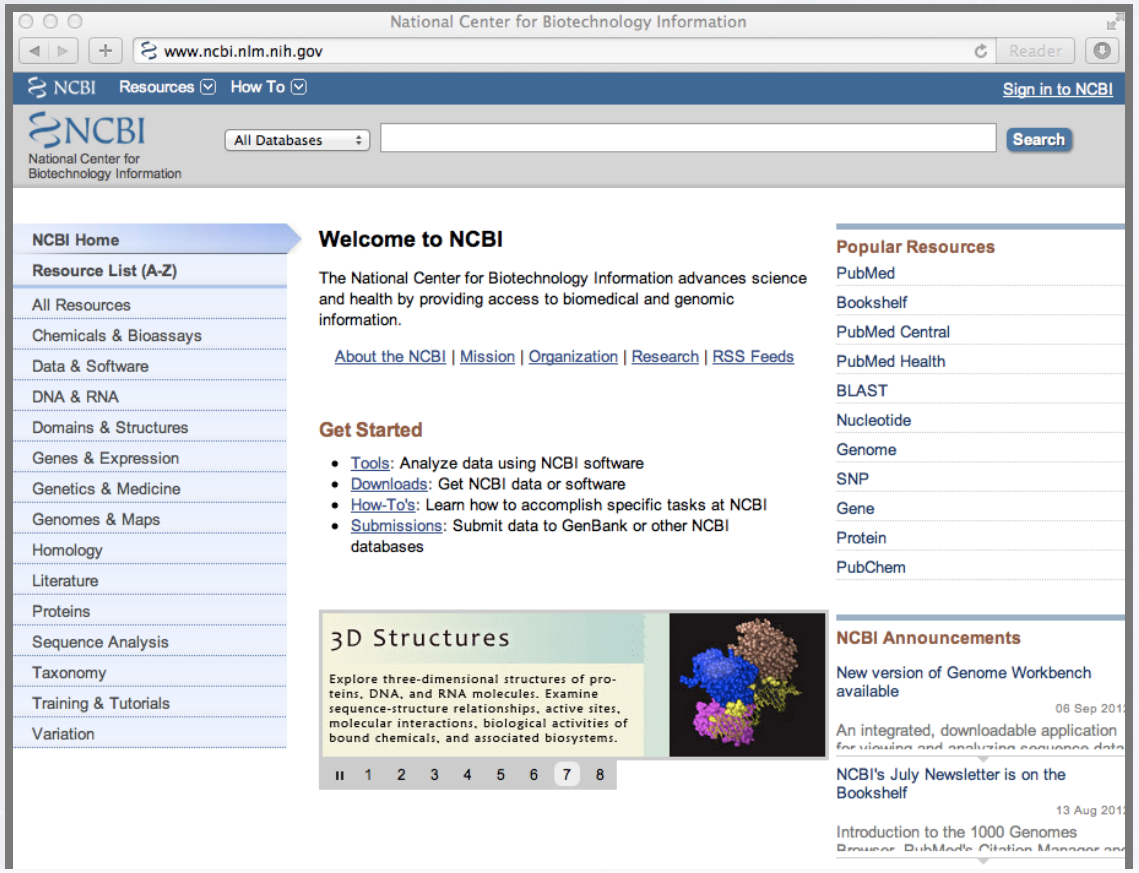
STEP 3 - Set you

PROGRAM					
FASTA					
MATRIX	GAP OPEN	GAP EXTEND	KTUP	EXPECTATION UPPER VALUE	EXPECTATION LOWER VALUE
BLOSUM50	-10	-2	2	10	0 (default)
DNA STRAND	HISTOGRAM	FILTER		STATISTICAL ESTIMATES	
N/A	no	none		Regress	
SCORES	ALIGNMENTS	SEQUENCE RANGE	DATABASE RANGE	MULTI HSPs	
50	50	START-END	START-END	no	
SCORE FORMAT					
Default					

Related tools with different terminology

Key Online Bioinformatics Resources: NCBI & EBI

The NCBI and EBI are invaluable, publicly available resources for biomedical research



Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | Research | RSS Feeds

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

3D Structures

Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of bound chemicals, and associated biosystems.

NCBI Announcements

New version of Genome Workbench available

06 Sep 201

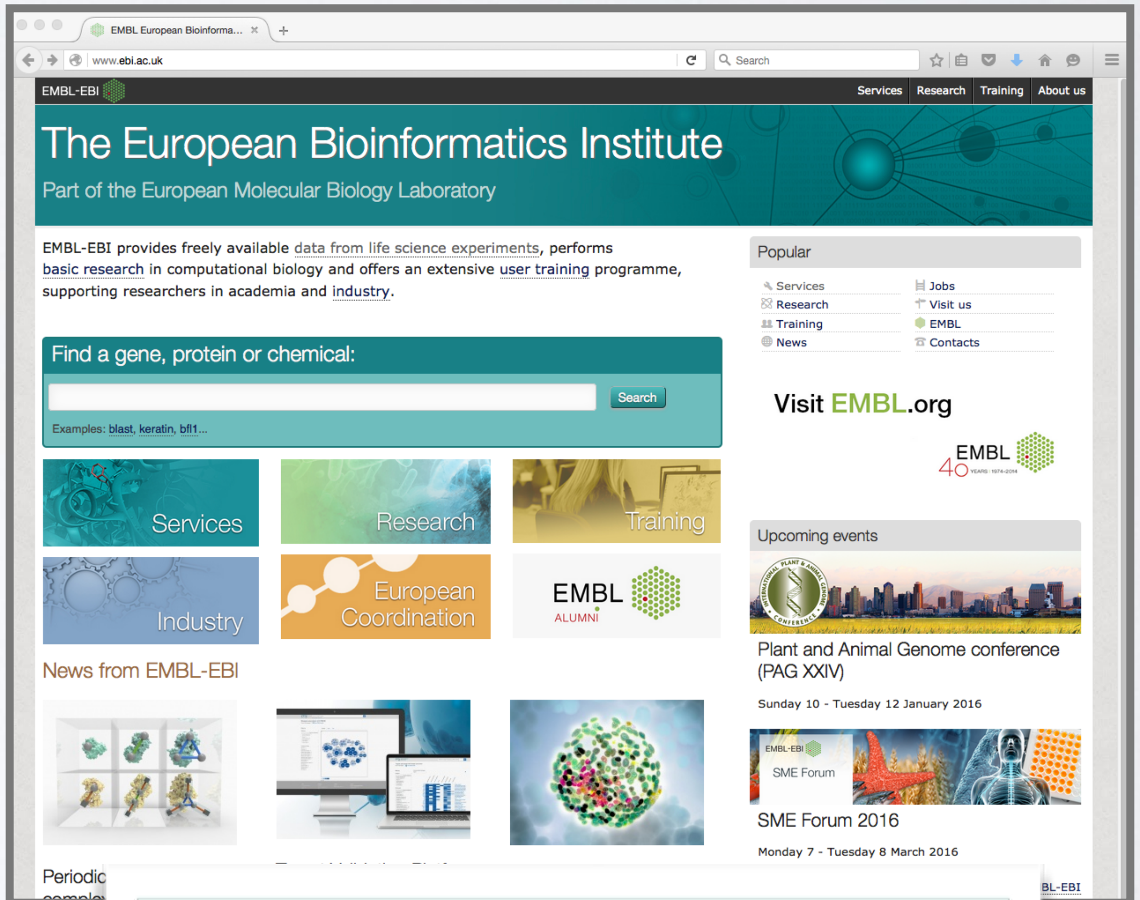
An integrated, downloadable application for viewing and analyzing sequence data

NCBI's July Newsletter is on the Bookshelf

13 Aug 201

Introduction to the 1000 Genomes Browser, PubMed's Citation Manager an

http://www.ncbi.nlm.nih.gov



The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry.

Find a gene, protein or chemical:

Examples: blast, keratin, bfl1...

News from EMBL-EBI

Periodic complex

Plant and Animal Genome conference (PAG XXIV)

Sunday 10 - Tuesday 12 January 2016

SME Forum 2016

Monday 7 - Tuesday 8 March 2016

https://www.ebi.ac.uk

National Center for Biotechnology Information (NCBI)

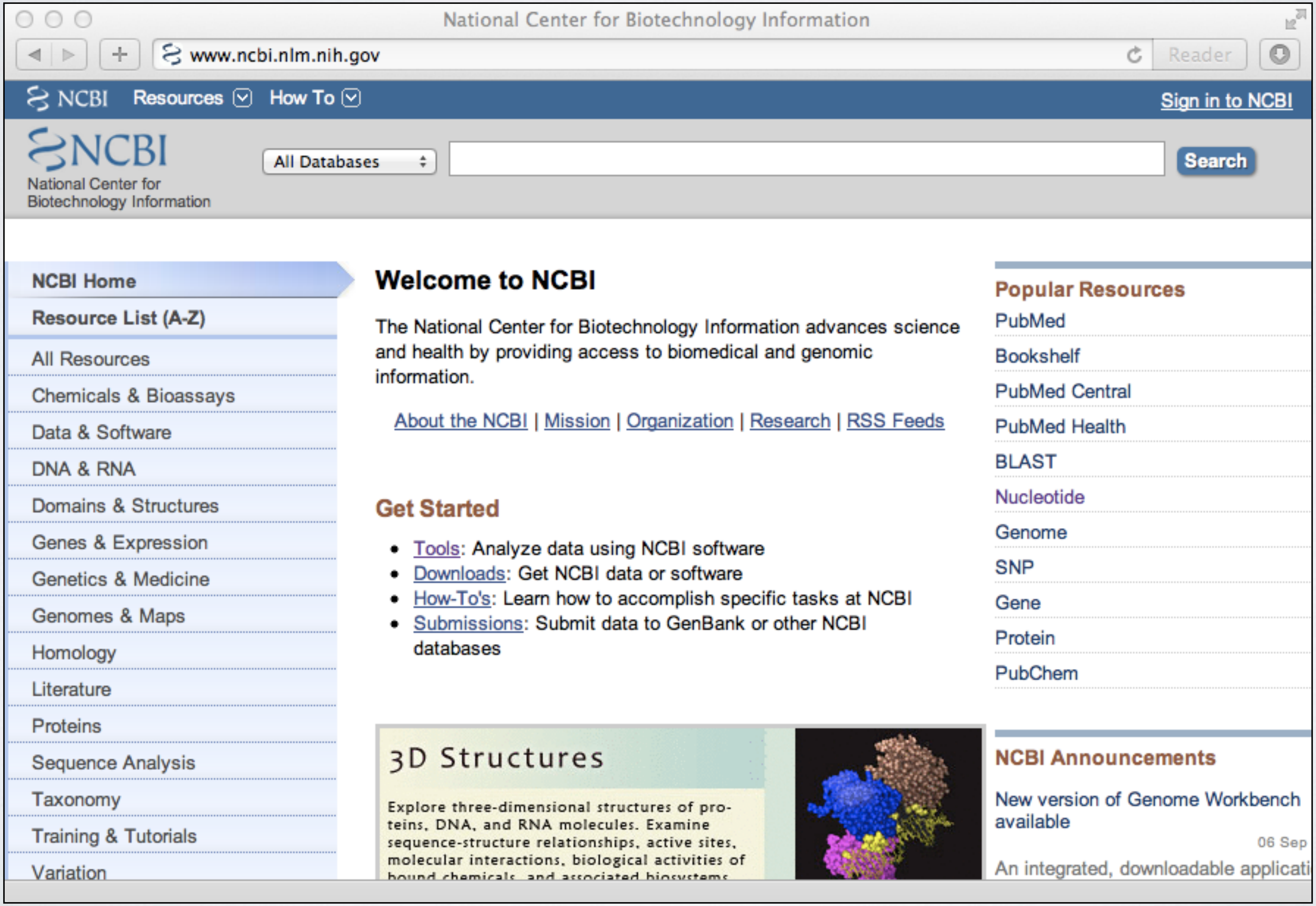
- Created in 1988 as a part of the National Library of Medicine (NLM) at the National Institutes of Health

- NCBI's mission includes:
  - Establish **public databases**
  - Develop **software tools**
  - **Education** on and dissemination of biomedical information

Bethesda,MD

- We will cover a number of core NCBI databases and software tools in the lecture

http://www.ncbi.nlm.nih.gov



Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | Research | RSS Feeds

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

3D Structures

Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of bound chemicals, and associated biosystems

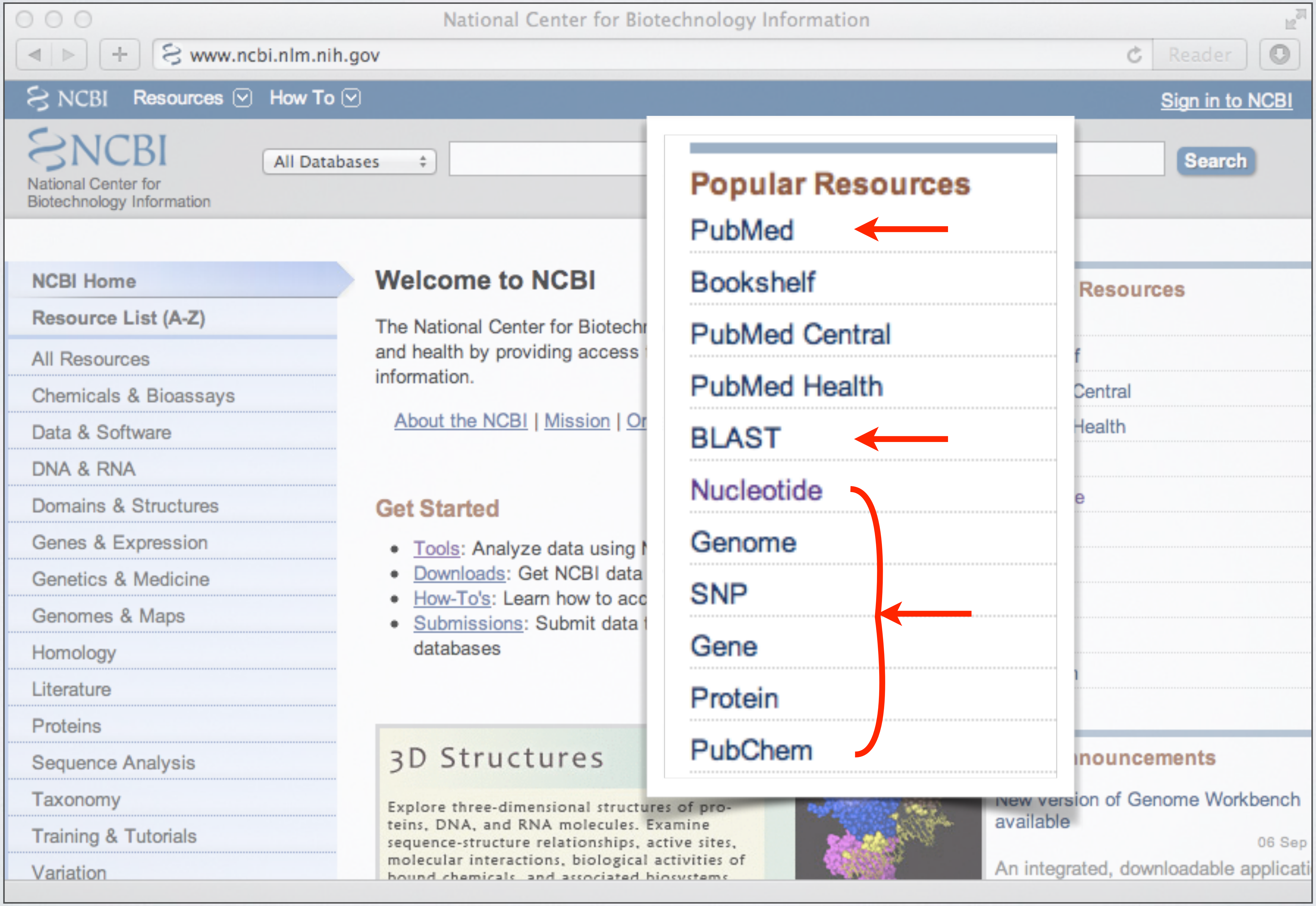
NCBI Announcements

New version of Genome Workbench available

06 Sep

An integrated, downloadable applicati

http://www.ncbi.nlm.nih.gov



National Center for Biotechnology Information

www.ncbi.nlm.nih.gov

Reader

NCBI Resources How To

Sign in to NCBI

NCBI
National Center for
Biotechnology Information

All Databases

Search

NCBI Home

Resource List (A-Z)

- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Welcome to NCBI

The National Center for Biotechn
and health by providing access
information.

About the NCBI | Mission | Or

Get Started

- Tools: Analyze data using N
- Downloads: Get NCBI data
- How-To's: Learn how to acc
- Submissions: Submit data t
databases

3D Structures

Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of bound chemicals, and associated biosystems

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

Resources

f

Central

Health

e

n

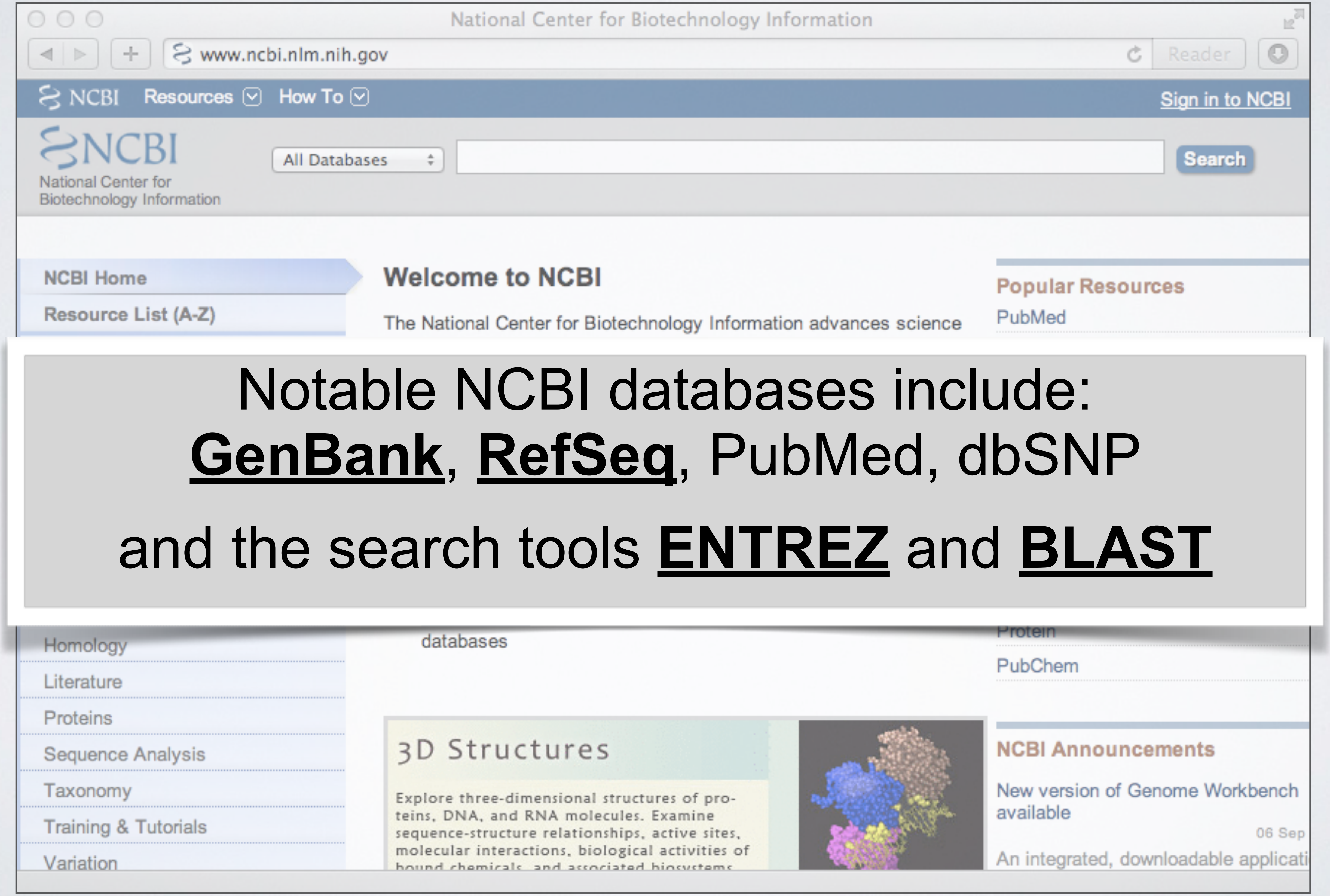
nouncements

New version of Genome Workbench available

06 Sep

An integrated, downloadable applicati

http://www.ncbi.nlm.nih.gov



National Center for Biotechnology Information

www.ncbi.nlm.nih.gov

NCBI Resources How To Sign in to NCBI

NCBI National Center for Biotechnology Information

All Databases Search

NCBI Home

Resource List (A-Z)

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science

databases

3D Structures

Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of bound chemicals, and associated biosystems

Popular Resources

PubMed

Protein

PubChem

NCBI Announcements

New version of Genome Workbench available

06 Sep

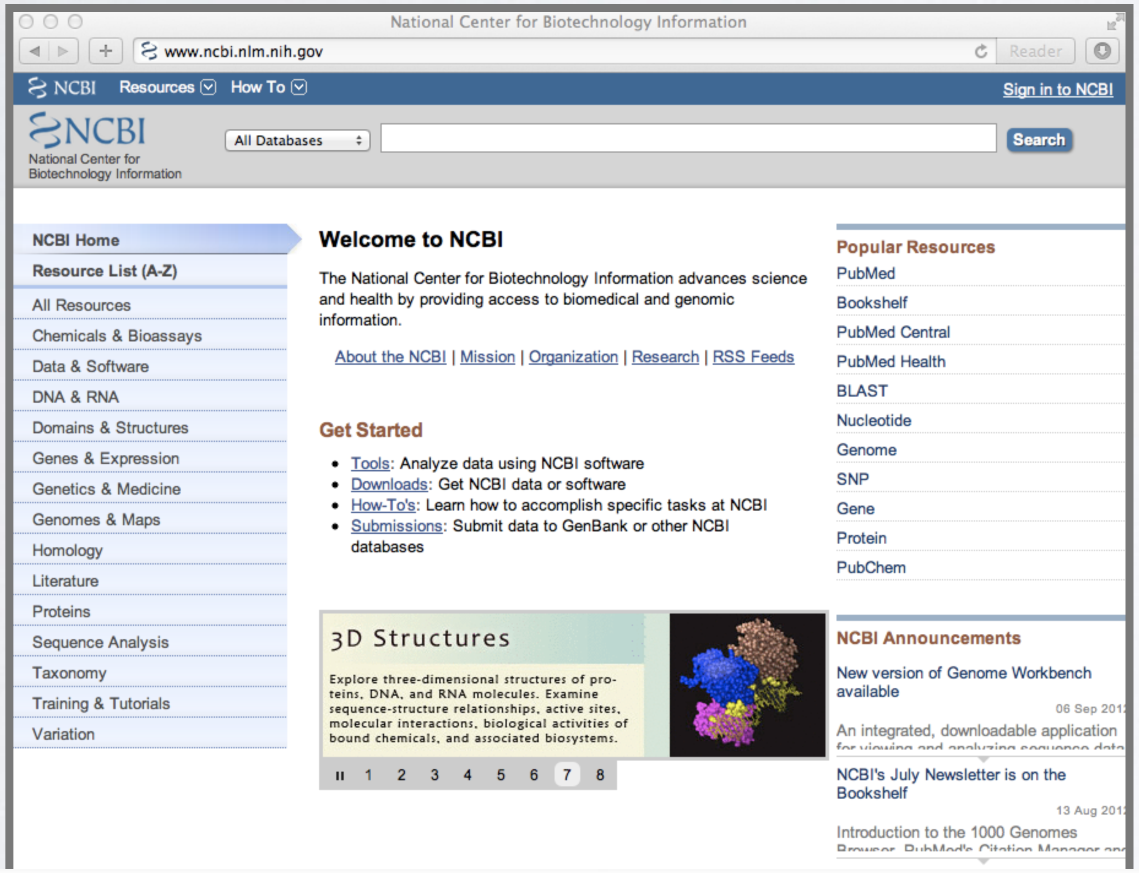
An integrated, downloadable applicati

Notable NCBI databases include:
<u>GenBank</u>, **<u>RefSeq</u>**, PubMed, dbSNP

and the search tools **<u>ENTREZ</u>** and **<u>BLAST</u>**

Key Online Bioinformatics Resources: NCBI & EBI

The NCBI and EBI are invaluable, publicly available resources for biomedical research


National Center for Biotechnology Information
www.ncbi.nlm.nih.gov
Reader
NCBI
Resources
How To
Sign in to NCBI
NCBI
National Center for Biotechnology Information
All Databases
Search

NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | Research | RSS Feeds

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

3D Structures

Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of bound chemicals, and associated biosystems.

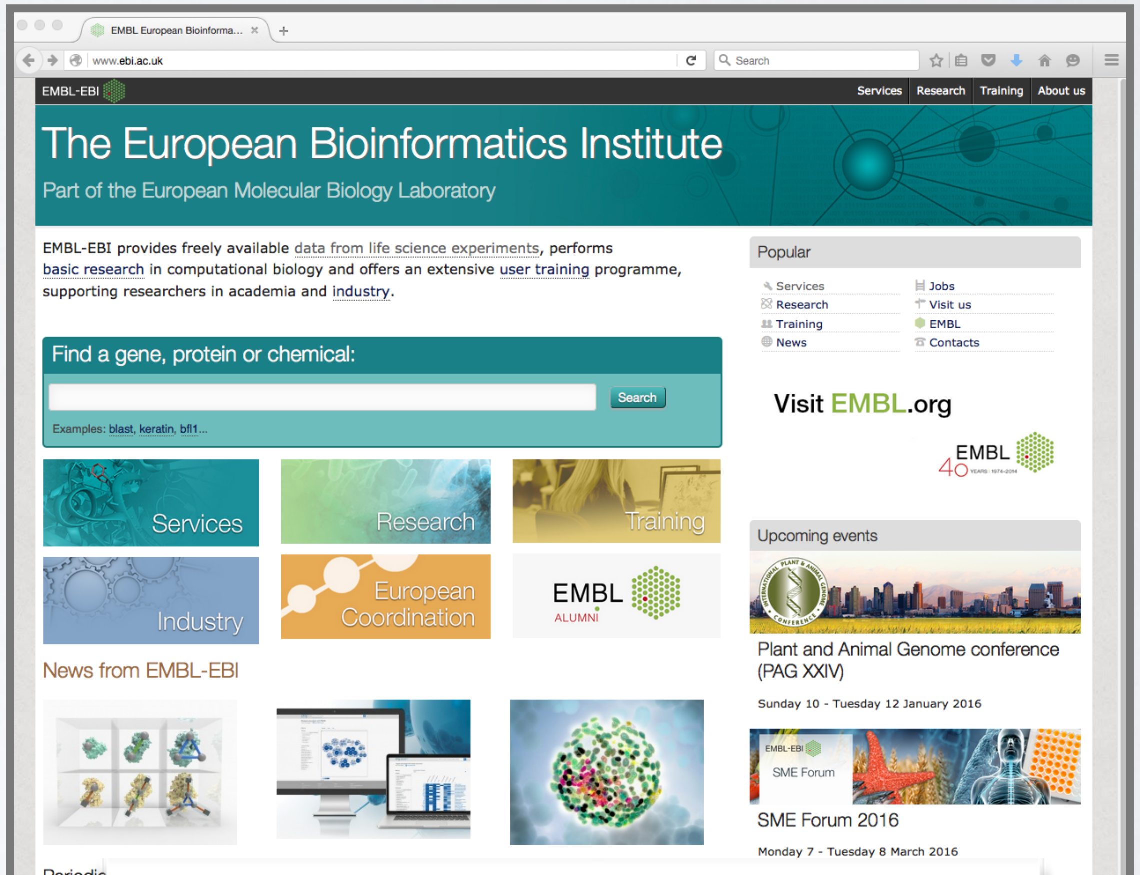
Popular Resources
PubMed
Bookshelf
PubMed Central
PubMed Health
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

NCBI Announcements

New version of Genome Workbench available
06 Sep 201
An integrated, downloadable application for viewing and analyzing sequence data

NCBI's July Newsletter is on the Bookshelf
13 Aug 201
Introduction to the 1000 Genomes Browser, PubMed's Citation Manager and

http://www.ncbi.nlm.nih.gov



The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry.

Find a gene, protein or chemical:

Examples: blast, keratin, bfl1...

News from EMBL-EBI

Periodic complex

Popular
Services
Research
Training
News
Jobs
Visit us
EMBL
Contacts

Visit EMBL.org

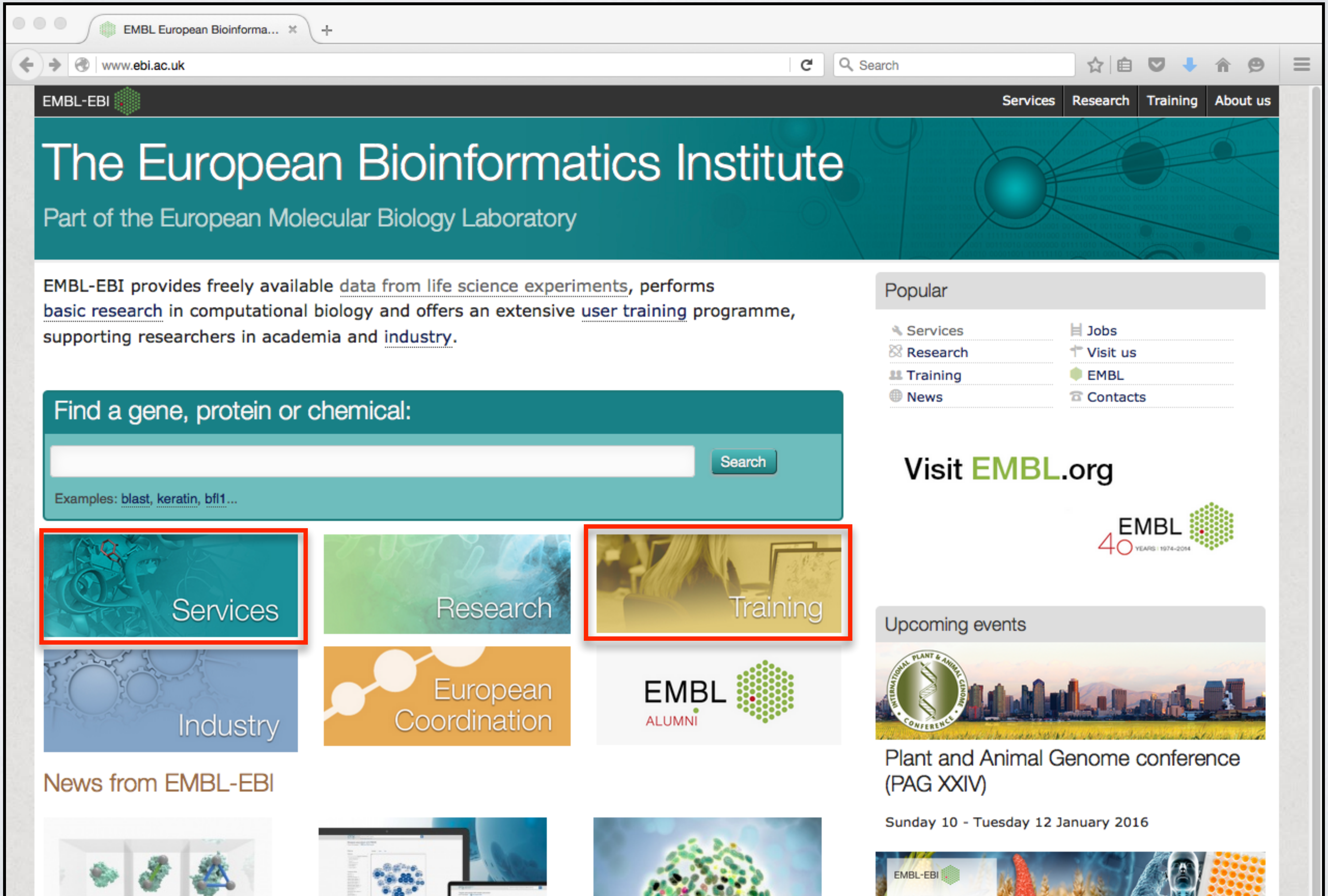
Upcoming events

Plant and Animal Genome conference (PAG XXIV)
Sunday 10 - Tuesday 12 January 2016

SME Forum 2016
Monday 7 - Tuesday 8 March 2016

https://www.ebi.ac.uk

The EBI maintains a number of high quality curated **secondary databases** and associated tools



The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry.

Find a gene, protein or chemical:

Search

Examples: blast, keratin, bfl1...

News from EMBL-EBI

Popular

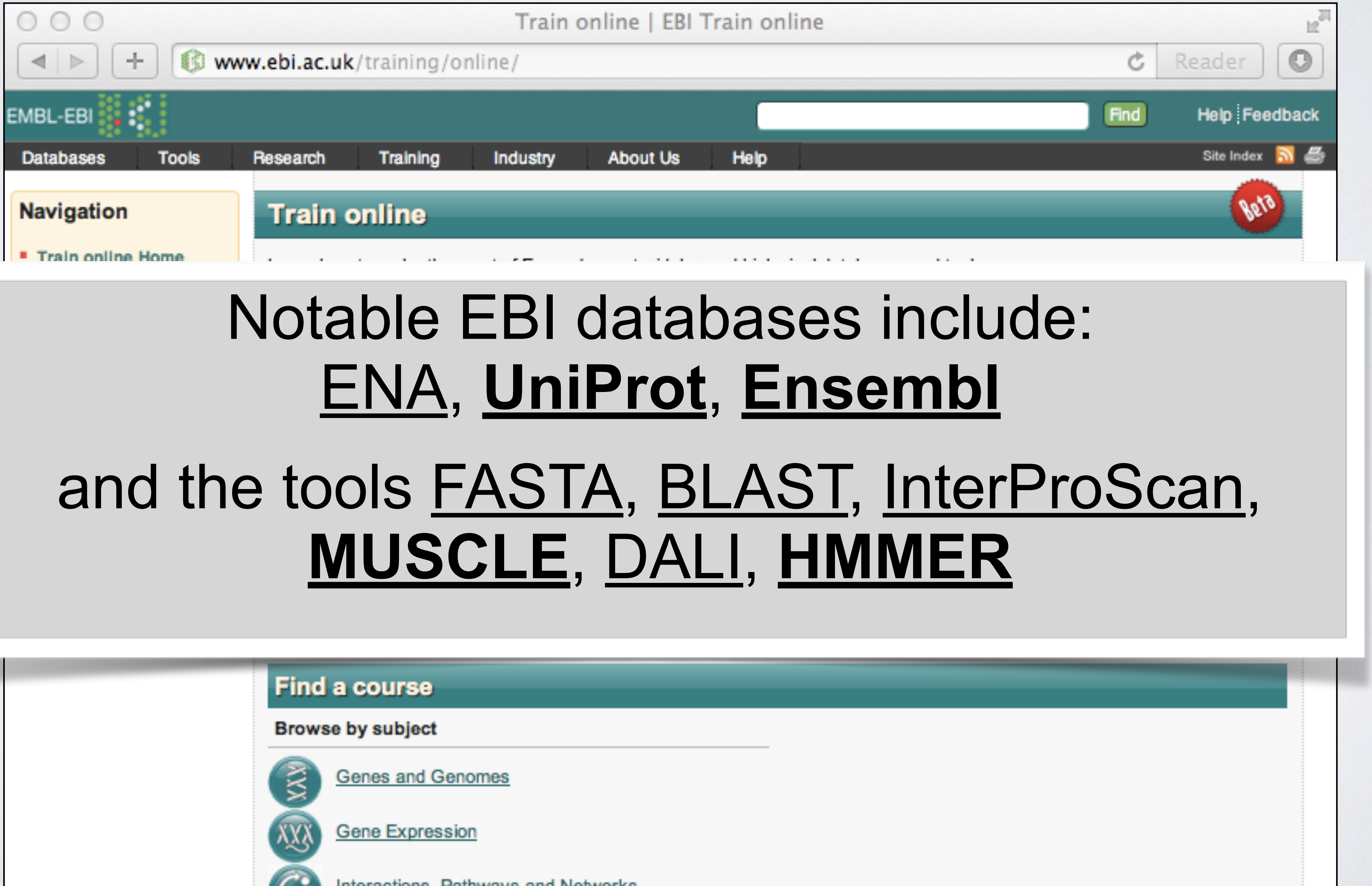
Visit EMBL.org

Upcoming events

Plant and Animal Genome conference (PAG XXIV)

Sunday 10 - Tuesday 12 January 2016

The EBI also provides a growing selection of **online tutorials** on EBI databases and tools



Notable EBI databases include:
ENA, **UniProt**, **Ensembl**

and the tools FASTA, BLAST, InterProScan, **MUSCLE**, DALI, **HMMER**

Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb,BBDB, BCGD, Beanref, BioImage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAP, BSORF, BTKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, CCDC, CD4OLbase, CGAP, ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG, CyanoBase, dbCFC, dbEST, dbSTS, DDBJ, DGP, DictyDb, Picty_cDB, DIP, DOGS, DOMO, DPD, DPInteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genlilesne, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB, HAEMB, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD, HIDB, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5 Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-Us, MPDB, MRR, MutBase, MycDB, NDB, NRSub, 0-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene,Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS- MODEL Repository, SWISS-PROT,TelDB,TGN, tmRDB,TOPS,TRANSFAC,TRR, UniGene, URNADB,V BASE, VDRR,VectorDB,WDCM,WIT,WormPep, etc .... !!!!

Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb,BBDB, BC Beanref,
BioImage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BO TKbase,
CANSITE, CarbBank, CARBHYD, CATH, CAZY, C AP,
ChickGBASE, Colibri, COPE, CottonDB, C bEST,
dbSTS, DDBJ, DGP, DictyDb, Pi CDC,
ECGC, EC02DBASE, E THER,
FlyBase, Fl Link,
G DB, HAEMB,
H HIVdb, HotMolecBase,
H IL2RGbase, IMGT, Kabat, KDNA,
K DB, Medline, Mendel, MEROPS, MGDB, MGI,
MHC MOMAP, MJDB, MmtDB, Mol-R-Us, MPDB, MRR, MutBase,
My Sub, 0-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase,
PD PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS,
ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP,
SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK,
StyGene,Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS- MODEL
Repository, SWISS-PROT, TelDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene,
URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc .... !!!!

There are lots of Bioinformatics Databases

For a annotated listing of major bioinformatics databases please see the online handout

< Major_Databases.pdf >

Side-note: Databases come in all shapes and sizes

Databases can be of variable quality and often there are multiple databases with overlapping content.

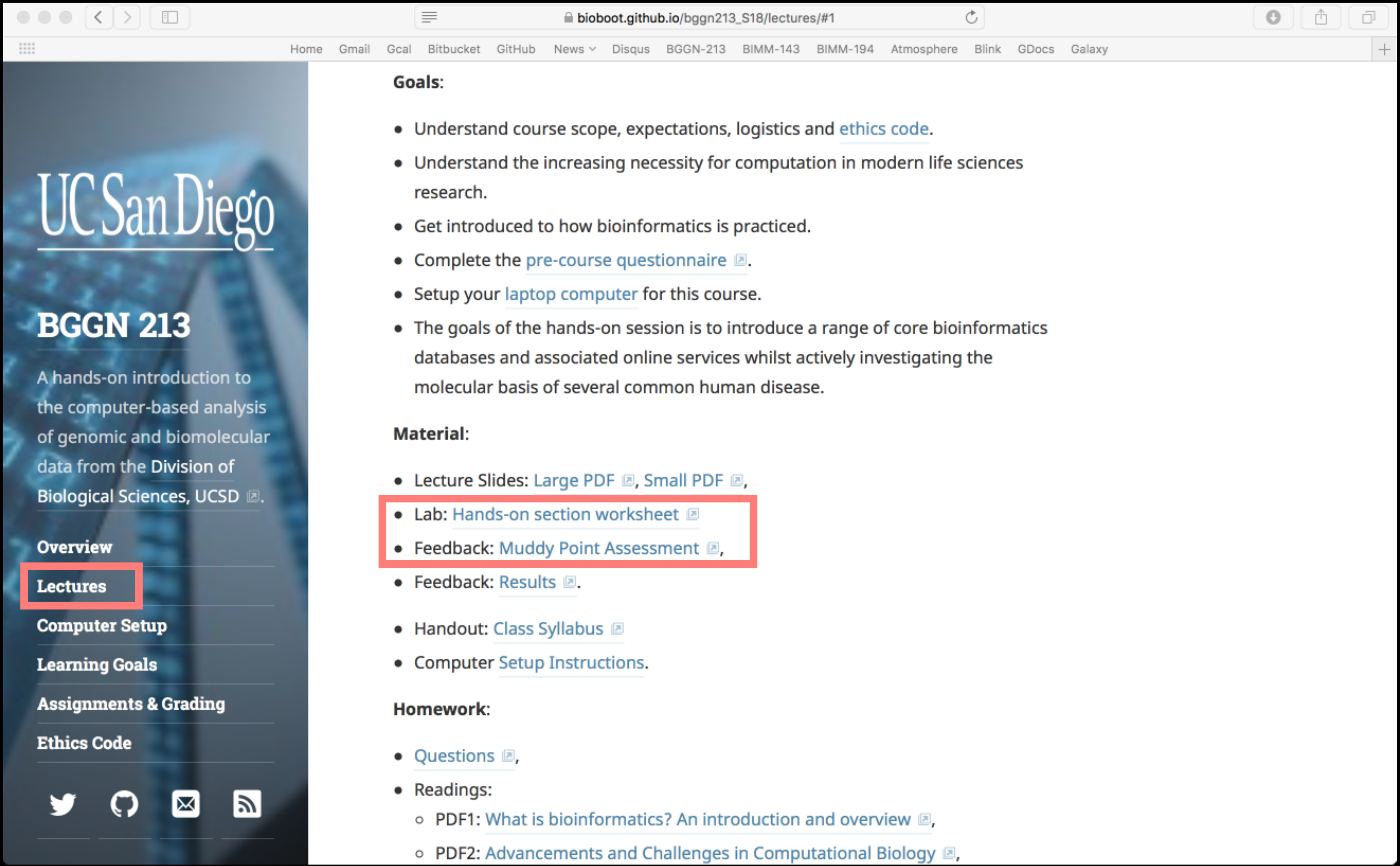
Today’s Menu

Course Logistics	Website, screencasts, survey, ethics, assessment and grading.
Learning Objectives	What you need to learn to succeed in this course.
Course Structure	Major lecture topics and specific leaning goals.
Introduction to Bioinformatis	Introducing the *what*, *why* and *how* of bioinformatics?
Bioinformatics Database	**Hands-on** exploration of several major databases and their associated tools.

Hands-on section

Your Turn!

http://thegrantlab.org/bggn213/



UC San Diego

BGGN 213

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

- Overview
- Lectures
- Computer Setup
- Learning Goals
- Assignments & Grading
- Ethics Code

Goals:

- Understand course scope, expectations, logistics and ethics code.
- Understand the increasing necessity for computation in modern life sciences research.
- Get introduced to how bioinformatics is practiced.
- Complete the pre-course questionnaire.
- Setup your laptop computer for this course.
- The goals of the hands-on session is to introduce a range of core bioinformatics databases and associated online services whilst actively investigating the molecular basis of several common human disease.

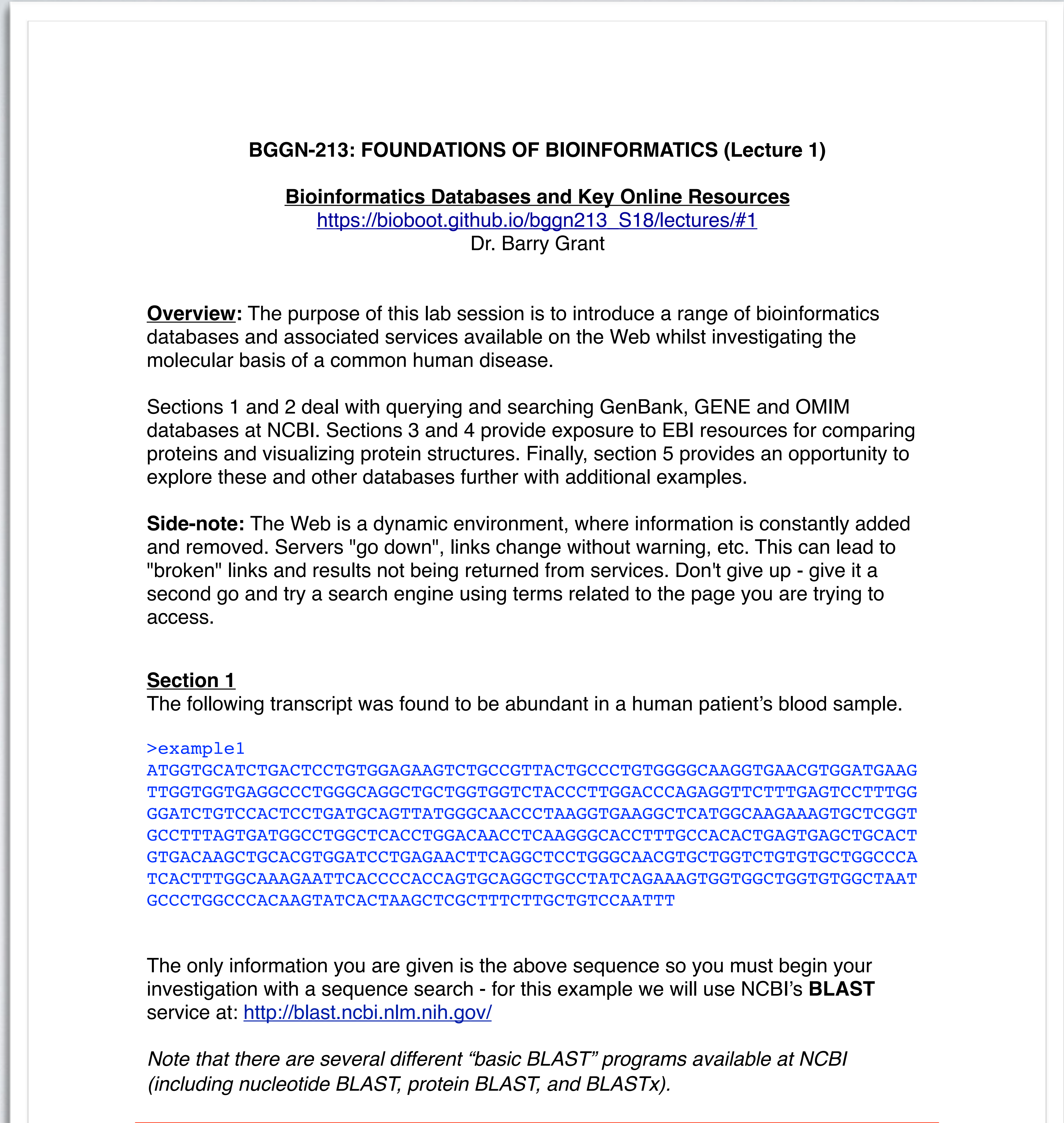
Material:

- Lecture Slides: Large PDF, Small PDF,
- Lab: Hands-on section worksheet
- Feedback: Muddy Point Assessment,
- Feedback: Results.

- Handout: Class Syllabus
- Computer Setup Instructions.

Homework:

- Questions,
- Readings:
  - PDF1: What is bioinformatics? An introduction and overview,
  - PDF2: Advancements and Challenges in Computational Biology,



BGGN-213: FOUNDATIONS OF BIOINFORMATICS (Lecture 1)

Bioinformatics Databases and Key Online Resources
https://bioboot.github.io/bggn213_S18/lectures/#1
Dr. Barry Grant

Overview: The purpose of this lab session is to introduce a range of bioinformatics databases and associated services available on the Web whilst investigating the molecular basis of a common human disease.

Sections 1 and 2 deal with querying and searching GenBank, GENE and OMIM databases at NCBI. Sections 3 and 4 provide exposure to EBI resources for comparing proteins and visualizing protein structures. Finally, section 5 provides an opportunity to explore these and other databases further with additional examples.

Side-note: The Web is a dynamic environment, where information is constantly added and removed. Servers "go down", links change without warning, etc. This can lead to "broken" links and results not being returned from services. Don't give up - give it a second go and try a search engine using terms related to the page you are trying to access.

Section 1

The following transcript was found to be abundant in a human patient's blood sample.

```
>example1
ATGGTGCATCTGACTCCTGTGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG
TTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGG
GGATCTGTCCACTCCTGATGCAGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGT
GCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACT
GTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCA
TCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAAT
GCCCTGGCCCACAAGTATCACTAAGCTCGCTTTCTTGCTGTCCAATTT
```

The only information you are given is the above sequence so you must begin your investigation with a sequence search - for this example we will use NCBI's **BLAST** service at: http://blast.ncbi.nlm.nih.gov/

Note that there are several different "basic BLAST" programs available at NCBI (including nucleotide BLAST, protein BLAST, and BLASTx).

YOUR TURN!

- There are five major hands-on sections including:

1. BLAST, GenBank and OMIM @ **NCBI**	[~35 mins]
2. GENE database @ **NCBI**	[~15 mins]
— BREAK —	
3. UniProt & Muscle @ **EBI**	[~25 mins]
4. PFAM, PDB & NGL	[~30 mins]
— BREAK —	
5. Extension exercises	[~30 mins]

- Please do answer the last review question (**Q19**).
- We encourage discussion and exploration!

YOUR TURN!

- There are five major hands-on sections including:

	End times:
1. BLAST, GenBank and OMIM @ **NCBI**	[2:35 pm]
2. GENE database @ **NCBI**	[2:55 pm]
— BREAK —	— 3:10 pm —
3. UniProt & Muscle @ **EBI**	[3:30 pm]
4. PFAM, PDB & NGL	[4:00 pm]
— BREAK —	— 4:10 pm —
5. Extension exercises	[4:40 pm]

- Please do answer the last review question (**Q19**).
- We encourage discussion and exploration!

SUMMARY

- Bioinformatics is computer aided biology.
- Bioinformatics deals with the collection, archiving, organization, and interpretation of a wide range of biological data.
- The NCBI and EBI are major online bioinformatics service providers.
- Introduced Gene, UniProt, PDB databases as well as a number of ‘boutique’ databases including PFAM and OMIM.

HOMEWORK

http://thegrantlab.org/bggn213/

- ☑ Complete the initial **course questionnaire**:
- ☑ Check out the "**background reading**" material online:
- ☑ Complete the lecture 1 **homework questions**: