Web Scraping with Python

Garrett Dancik, PhD
Admitted Student Decision Day
Eastern Connecticut State University
Department of Computer Science
March 23, 2019

dancikg@easternct.edu
http://gdancik.github.io/bioinformatics/ASDD/

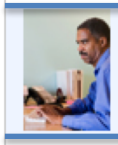
EASTERN
CONNECTICUT STATE UNIVERSITY

A Liberal Education. Practically Applied.

Visitors | Future Students | Current Students | Alumni and Friends | Faculty and Staff

Occupational Outlook – Data from the Burea of Labor Statistics

(http://www.bls.gov/ooh/computer-and-information-technology/home.htm)

	OCCUPATION	JOB SUMMARY	ENTRY-LEVEL EDUCATION	2012 MEDIAN PAY	Projected Growth (2022)
	Computer and Information Research Scientists	Computer and information research scientists invent and design new approaches to computing technology and find innovative uses for existing technology. They study and solve complex problems in computing for business, medicine, science, and other fields.	Doctoral or professional degree	$102,190	Faster than average
	Computer Network Architects	Computer network architects design and build data communication networks, including local area networks (LANs), wide area networks (WANs), and intranets. These networks range from a small connection between two offices to a multinational series of globally distributed communications systems.	Bachelor's degree	$91,000	Faster than average
	Computer Programmers	Computer programmers write code to create software programs. They turn the program designs created by software developers and engineers into instructions that a computer can follow.	Bachelor's degree	$74,280	Average
	Computer Support Specialists	Computer support specialists provide help and advice to people and organizations using computer software or equipment. Some, called computer network support specialists, support information technology (IT) employees within their organization. Others, called computer user support specialists, assist non-IT users who are having computer problems.	See How to Become One	$48,900	Faster than average
	Computer Systems Analysts	Computer systems analysts study an organization's current computer systems and procedures and design information systems solutions to help the organization operate more efficiently and effectively. They bring business and information technology (IT) together by understanding the needs and limitations of both.	Bachelor's degree	$79,680	Much faster than average

- The Computer Science degree program at Eastern Connecticut State University provides students the **foundations and skills for future work and careers in computing**.
- Upon graduation, students will:
  - Possess **practical and theoretical knowledge of computer science** sufficient to work professionally and contribute to the regional and global economic development.
  - Be able to **apply computational techniques to design and implement solutions to real-world problems**.
  - Be prepared **for advanced education in computer science and continued professional development**.
  - Possess the skills and the intellectual abilities that will enable them to **adapt in the ever-changing field of computer science**.

Occupational Outlook – Data from the Burea of Labor Statistics

(http://www.bls.gov/ooh/computer-and-information-technology/home.htm)

	OCCUPATION	JOB SUMMARY	ENTRY-LEVEL EDUCATION	2012 MEDIAN PAY	Projected Growth (2022)
	Database Administrators	Database administrators (DBAs) use specialized software to store and organize data, such as financial information and customer shipping records. They make sure that data are available to users and are secure from unauthorized access.	Bachelor's degree	$77,080	Faster than average
	Information Security Analysts	Information security analysts plan and carry out security measures to protect an organization's computer networks and systems. Their responsibilities are continually expanding as the number of cyberattacks increase.	Bachelor's degree	$86,170	Much faster than average
	Network and Computer Systems Administrators	Computer networks are critical parts of almost every organization. Network and computer systems administrators are responsible for the day-to-day operation of these networks.	Bachelor's degree	$72,560	Average
	Software Developers	Software developers are the creative minds behind computer programs. Some develop the applications that allow people to do specific tasks on a computer or other device. Others develop the underlying systems that run the devices or control networks.	Bachelor's degree	$93,350	Much faster than average
	Web Developers	Web developers design and create websites. They are responsible for the look of the site. They are also responsible for the site's technical aspects, such as performance and capacity, which are measures of a website's speed and how much traffic the site can handle. They also may create content for the site.	Associate's degree	$62,500	Faster than average

- Computer Science major Requirements
  - http://www.easternct.edu/computerscience/catalog/

- **Minors:**
  - Computer Science
  - Computer Engineering
  - Game Design
  - Management Information Systems
  - Bioinformatics

Web development overview

- **Hypertext Markup Language** (HTML) describes the structure of a web page
  - HTML pages are composed of elements that are specified using tags
    - This is a paragraph
    - <h1> This is a header </h1>

- **Cascading style sheets (CSS)** describe how HTML elements are displayed

   This is a paragraph with yellow background

Examples

- See example web page at:

  https://gdancik.github.io/CSC-360/data/notes/schedule.html

- Let's use the Web Inspector to look at the structure of this page

Web Scraping Overview

- Web scraping is the process of retrieving web pages and extracting relevant data from them
- Why?
  - Search engines collect data to index web pages
  - Collecting weather and climate data for research
    - https://www.sciencedirect.com/science/article/pii/S0168169909002348
  - For businesses and consumers to keep track of products
  - For research in economics
    - https://www.aeaweb.org/articles?id=10.1257/jep.30.2.151

Steps for web scraping

- Identify a page you want to scrape
- Understand the structure of the page (e.g., by using the Web Inspector)
- Write a script that
  - Retrieves the web page
    - From a URL using the Python requests library
    - From a file using *open*
  - Extracts information from the web page
    - Using Python's BeautifulSoup library

What else can we do with data?

OPEN ACCESS Freely available online

PLOS ONE

Influenza Forecasting with Google Flu Trends

Andrea Freyer Dugas[1]*, Mehdi Jalalpour[2], Yulia Gel[2,3], Scott Levin[1,2], Fred Torcaso[2], Takeru Igusa[2], Richard E. Rothman[1]

1 Department of Emergency Medicine, Johns Hopkins University, Baltimore, Maryland, United States of America, **2** Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland, United States of America, **3** University of Waterloo, Waterloo, Ontario, Canada

What else can we do with data?

APRIL 24, 2014

Update: Single Fire Hydrant Nets NYC $33,000 a Year, not just $25,000

Since there was some interest on my last post on the most lucrative hydrant in NYC, I wanted to download a bit more data and update the figures. My new download encompasses August 1, 2013 - March 26, 2014.

So with this fresh download, thanks to NYC Open Data, we see that the second confusing hydrant on Forsyth place has moved into second place! I also made a Top 10 list of hydrants to avoid at all costs:

Top Ticketed NYC Hydrants, August 1 2013 - March 26 2014				
Address	Borough	# Tickets	Total Fines ($)	Annualized Fines ($)
Opposite 152 Forsyth St	Manhattan	187	21,505	33,118
Opposite 104 Forsyth St	Manhattan	139	15,985	24,617
Front of 44 Court St	Brooklyn	101	11,615	17,887
Opposite 122 Montague St	Brooklyn	95	10,925	16,825
Front of 21 W 58th St	Manhattan	91	10,465	16,116
Opposite 100 Overlook Ter	Manhattan	85	9,775	15,054
Front of 720 Lenox Ave	Manhattan	81	9,315	14,345
Front of 2960 Fredrick Douglas Blv	Manhattan	80	9,200	14,168
Front of 2711 Valentine Ave	Bronx	77	8,855	13,637
Front of 1450 3rd Ave	Manhattan	76	8,740	13,460
Top 10 Total		**1,012**	**116,380**	**179,225**
All Total		**314,637**	**36,183,255**	**55,722,213**

LETTER

doi:10.1038/nature21056

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva[1]*, Brett Kuprel[1]*, Roberto A. Novoa[2,3], Justin Ko[2], Susan M. Swetter[2,4], Helen M. Blau[5] & Sebastian Thrun[6]

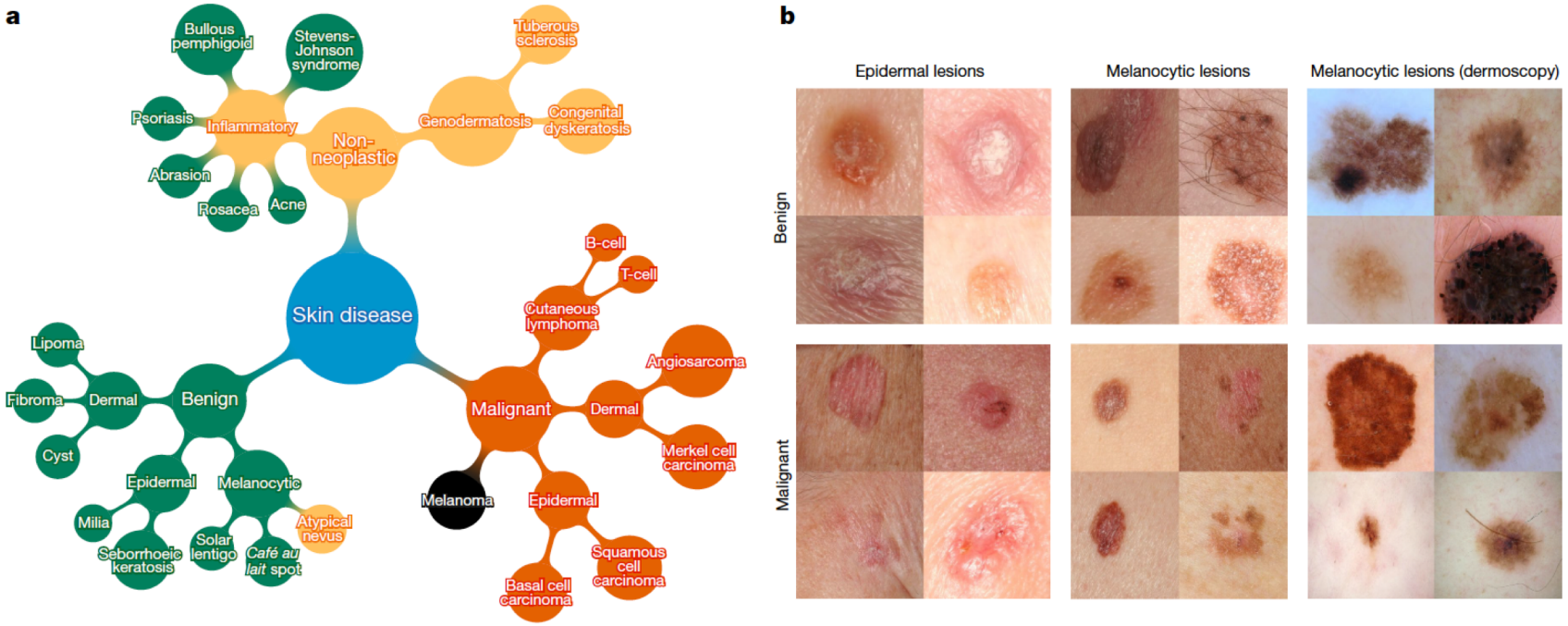
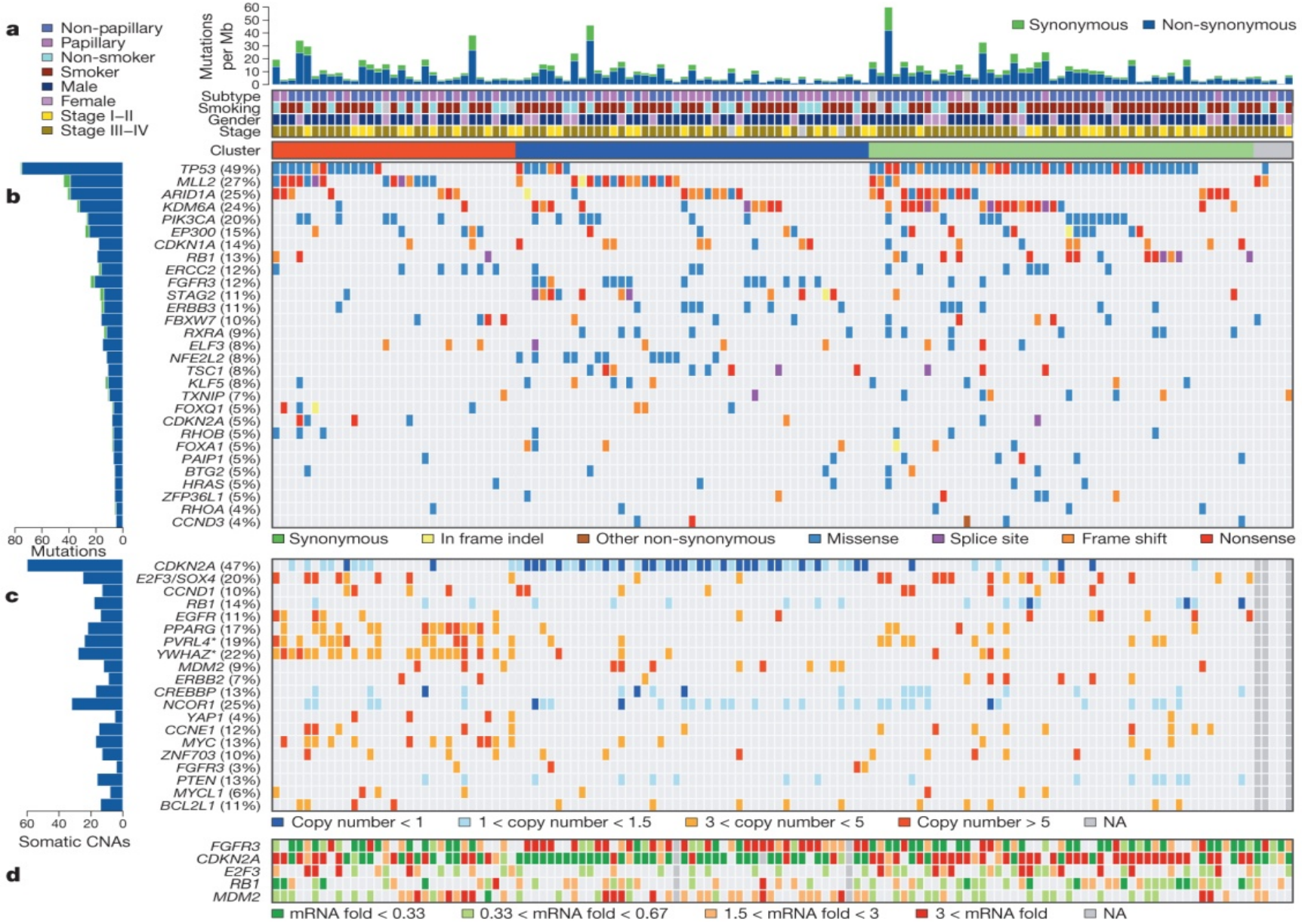


Figure 2 | A schematic illustration of the taxonomy and example test set images. a, A subset of the top of the tree-structured taxonomy of skin disease. The full taxonomy contains 2,032 diseases and is organized based on visual and clinical similarity of diseases. Red indicates malignant, green indicates benign, and orange indicates conditions that can be either

example images from two disease classes. These test images highlight the difficulty of malignant versus benign discernment for the three medically critical classification tasks we consider: epidermal lesions, melanocytic lesions and melanocytic lesions visualized with a dermoscope. Example images reprinted with permission from the Edinburgh Dermofit Library

The genomic landscape of bladder cancer



The Cancer Genome Atlas Research Network *Nature* **507**, 315-322 (2014) doi:10.1038/nature12965

Thank You!

- Any questions?
- Contact:
  - dancikg@easternct.edu
  - http://gdancik.github.io/

Computer Science Minor: Degree Requirements

This minor is designed for students who anticipate that computer science will have a prominent role to play in their academic and professional career. The minor emphasizes fundamental programming skills and hands-on experience applying those skills to computer-related projects.

I. Required Courses:			
CSC	210	Computer Science and Programming I	3
CSC	231	Computer Science and Programming II	3
CSC	330	Data Structures	3
II. Select two additional CSC courses number 2XX or above (except CSC 200) or two additional courses in the discipline of computing that are approved by the Computer Science Program Chair.			
Total Credits			

Computer Engineering Sciences Minor: Degree Requirements

Objectives

This minor is designed to provide students with the mathematical background and practical experience expected ofcomputer engineering majors. The objectives of the computer engineering sciences minor are the following:

1. To give students a background in engineering to assist them in graduate engineering schools.
2. To assist students in pursuing careers in engineering.
3. To afford science and mathematics students an experience in engineering.

I. Required Courses:			
CSC	350	Numerical Analysis	3
CSC	351	Signals and Systems	4
CSC	355	Digital Logic Lecture	3
II. Select one Computer Science elective numbered 300 or above.			
III. Complete the following Mathematics courses:			
MAT	310	Applied Linear Algebra	3
MAT	340	Calculus III	4
Total Credits			

Game Design Minor: Degree Requirements

Objectives

The computer gaming minor addresses the needs of students interested in simulation, human machine interaction and gaming. This is an interdisciplinary minor covering both the artistic and computational needs of the field. Students with such a minor may work with animation, game engines, mathematics, modeling, network design, and state-of-the-art hardware and software.

I. Required courses:			
ART	343	Introduction to 3D Animation	3
CSC	311	Computer & Video Games Development	3
II. Select additional nine credits from the following list:			
ART	403	3D Imaging/Animation I	3
CSC	312	Computer Graphics	3
CSC	337	Computer Networks and Distributed Processing	4
CSC/MAT	350	Numerical Methods	3
MUS	372	Multimedia Composition	3
SOC	320	Video Games and Society	3
		Any 300/400 level course by arrangement with coordinator	
Total Credits			**15**

Management Information Systems Minor: Degree Requirements

The goal of the Management Information Systems minor is to prepare students to contribute to an increase in productivity and the generation of new products, services and ventures, using state-of-the-art computer applications for better communication, problem diagnosis and decision making.

I. The MIS minor requires a total of 15*/18 credits as follows:			
CSC	110	Introduction to Computing and Problem Solving	3
		Or Equivalent Computing Experience	
II. Two Business Courses:			
BUS/BIS	442	Information Technology Project Management	3
BIS	361	Business Information Systems and Web Technologies	3
III. Any three of the following seven options:			
BIS	370	Systems Analysis and Design	3
CSC	200	Management Systems	3
CSC	210	Computer Science and Programming I	3
CSC	231	Computer Science and Programming II	3
CSC	249	Visual Basic or CSC 259 Advanced Visual Basic	3
CSC	251	Net-centric Computing	3
		Any 300/400-level computer science course except internships	
Total Credits			**15*/18**

Bioinformatics Minor (pending): Degree Requirements

Bioinformatics is an interdisciplinary science that involves the development and use of computational, statistical, and mathematical tools to store and analyze large biological datasets such as genomic sequences. Bioinformatics is routinely used in genomics research and in personalized medicine to help decide on an appropriate treatment for a cancer patient. The minor will prepare students who want to pursue graduate studies in Bioinformatics or Computational Biology and will assist students in pursuing related careers.

I.	**Required Courses**			
	CSC	210	Computer Science and Programming I	3
	MAT	216	Statistical Data Analysis OR	3-4
		315	Applied Probability and Statistics	
	BIO	230	Genetics w/ Laboratory OR	4
		304/314	Genetics and Society Lecture/ Lab	
	CSC	314	Introduction to Bioinformatics	3
	CSC	315	Bioinformatics Programming and Analysis	3
II.	**Select one of the following courses:**			
	CSC	342	Advanced Database Systems	3
	CSC	305	Data Mining and Applications	3
	CSC/MAT	350	Numerical Analysis	3
	MAT	373	Explorations - Mathematical Biology	3
	BIO	450	Biotechnology with Lab	4
	BIO	436	Molecular Genetics with Lab	4
Total Credits				**19-21**