Big Data Application and Analytics

Syllabus

By Dr. Geoffrey Fox

Section 1: Introduction

This section Contains Unit 1,2

Overview:

This section has a technical overview of course followed by a broad motivation for course.
The course overview covers it's content and structure. It presents the X-Informatics fields (defined values of X) and the Rallying cry of course: Use Clouds running Data Analytics Collaboratively processing Big Data to solve problems in X-Informatics (or e-X). The courses is set up as a MOOC divided into units that vary in length but are typically around an hour and those are further subdivided into 5-15 minute lessons. The course covers a mix of applications (the X in X-Informatics) and technologies needed to support the field electronically i.e. to process the application data. The overview ends with a discussion of course content at highest level. The course starts with a longish Motivation unit summarizing clouds and data science, then units describing applications (X = Physics, e-Commerce, Web Search and Text mining, Health, Sensors and Remote Sensing). These are interspersed with discussions of infrastructure (clouds) and data analytics (algorithms like clustering and collaborative filtering used in applications). The course uses either Python or Java and there are Side MOOCs discussing Python and Java tracks.
The course motivation starts with striking examples of the data deluge with examples from research, business and the consumer. The growing number of jobs in data science is highlighted. He describes industry trend in both clouds and big data. Then the cloud computing model developed at amazing speed by industry is introduced. The 4 paradigms of scientific research are described with growing importance of data oriented version.He covers 3 major X-informatics areas: Physics, e-Commerce and Web Search followed by a broad discussion of cloud applications. Parallel computing in general and particular features of MapReduce are described. He comments on a data science education and the benefits of using MOOC's.

Unit 1: Course Introduction

Go To Unit 1

Overview:

Geoffrey gives a short introduction to the course covering it's content and structure. He presents the X-Informatics fields (defined values of X) and the Rallying cry of course: Use Clouds running Data Analytics Collaboratively processing Big Data to solve problems in X-Informatics (or e-X). The courses is set up as a MOOC divided into units that vary in length but are typically around an hour and those are further subdivided into 5-15 minute lessons.
The course covers a mix of applications (the X in X-Informatics) and technologies needed to support the field electronically i.e. to process the application data. The introduction ends with a discussion of course content at highest level.
The course starts with a longish Motivation unit summarizing clouds and data science, then units describing applications (X = Physics, e-Commerce, Web Search and Text mining, Health, Sensors and Remote Sensing). These are interspersed with discussions of infrastructure (clouds) and data analytics (algorithms like clustering and collaborative filtering used in applications)
The course uses either Python or Java and there are Side MOOCs discussing Python and Java tracks.

Unit 2: Course Motivation

Go To Unit 2

Overview:

Geoffrey motivates the study of X-informatics by describing data science and clouds. He starts with striking examples of the data deluge with examples from research, business and the consumer. The growing number of jobs in data science is highlighted. He describes industry trend in both clouds and big data.
He introduces the cloud computing model developed at amazing speed by industry. The 4 paradigms of scientific research are described with growing importance of data oriented version. He covers 3 major X-informatics areas: Physics, e-Commerce and Web Search followed by a broad discussion of cloud applications. Parallel computing in general and particular features of MapReduce are described. He comments on a data science education and the benefits of using MOOC's.

Section 2: Overview of Data Science: What is Big Data, Data Analytics and X-Informatics?

This section Contains Unit 3,4,5

Overview:

This section has a technical overview of course followed by a broad motivation for course.
The course overview covers it's content and structure. It presents the X-Informatics fields (defined values of X) and the Rallying cry of course: Use Clouds running Data Analytics Collaboratively processing Big Data to solve problems in X-Informatics (or e-X). The courses is set up as a MOOC divided into units that vary in length but are typically around an hour and those are further subdivided into 5-15 minute lessons. The course covers a mix of applications (the X in X-Informatics) and technologies needed to support the field electronically i.e. to process the application data. The overview ends with a discussion of course content at highest level. The course starts with a longish Motivation unit summarizing clouds and data science, then units describing applications (X = Physics, e-Commerce, Web Search and Text mining, Health, Sensors and Remote Sensing). These are interspersed with discussions of infrastructure (clouds) and data analytics (algorithms like clustering and collaborative filtering used in applications). The course uses either Python or Java and there are Side MOOCs discussing Python and Java tracks.
The course motivation starts with striking examples of the data deluge with examples from research, business and the consumer. The growing number of jobs in data science is highlighted. He describes industry trend in both clouds and big data. Then the cloud computing model developed at amazing speed by industry is introduced. The 4 paradigms of scientific research are described with growing importance of data oriented version.He covers 3 major X-informatics areas: Physics, e-Commerce and Web Search followed by a broad discussion of cloud applications. Parallel computing in general and particular features of MapReduce are described. He comments on a data science education and the benefits of using MOOC's.

Unit 3: Part I: Data Science generics and Commercial Data Deluge

Go To Unit 3

Overview:

Geoffrey starts with X-Informatics and its rallying cry. The growing number of jobs in data science is highlighted. This unit offers a look at the phenomenon described as the Data Deluge starting with its broad features. Then he discusses data science and the famous DIKW (Data to Information to Knowledge to Wisdom) pipeline. Then more detail is given on the flood of data from Internet and Industry applications with eBay and General Electric discussed in most detail.

Unit 4 - Part II: Data Deluge and Scientific Applications and Methodology

Go To Unit 4

4**Overview:**

Geoffrey continues the discussion of the data deluge with a focus on scientific research. He takes a first peek at data from the Large Hadron Collider considered later as physics Informatics and gives some biology examples. He discusses the implication of data for the scientific method which is changing with the data-intensive methodology joining observation, theory and simulation as basic methods.
We discuss the long tail of sciences; many users with individually modest data adding up to a lot. The last lesson emphasizes how everyday devices -- the Internet of Things -- are

Section 2: Overview of Data Science: What is Big Data, Data Analytics and X-Informatics?

This section Contains Unit 3,4,5

Unit 5 - Part III: Clouds and Big Data Processing; Data Science Process and Analytics

Go To Unit 5

Overview:

Geoffrey starts with X-Informatics and its rallying cry. The growing number of jobs in data Geoffrey gives an initial technical overview of cloud computing as pioneered by companies like Amazon, Google and Microsoft with new centers holding up to a million servers. The benefits of Clouds in terms of power consumption and the environment are also touched upon, followed by a list of the most critical features of Cloud computing with a comparison to supercomputing.
He discusses features of the data deluge with a salutary example where more data did better than more thought. Examples are given of end to end systems to process big data. He introduces data science and one part of it -- data analytics -- the large algorithms that crunch the big data to give big wisdom. There are many ways to describe data science and several are discussed to give a good composite picture of this emerging field.

Section 3: Technology Training

This section Contains Unit 3,4,5

Overview:

This section is meant to give an overview of the python tools needed for doing for this course. These are really powerful tools which every data scientist who wishes to use python must know. This section covers. Canopy - Its is an IDE for python developed by EnThoughts. The aim of this IDE is to bring the various python libraries under one single framework or "Canopy" - that is why the name. NumPy - It is popular library on top of which many other libraries (like pandas, scipy) are built. It provides a way a vectorizing data. This helps to organize in a more intuitive fashion and also helps us use the various matrix operations which are popularly used by the machine learning community. Matplotlib: This a data visualization package. It allows you to create graphs charts and other such diagrams. It supports Images in JPEG, GIF, TIFF format. SciPy: SciPy is a library built above numpy and has a number of off the shelf algorithms / operations implemented. These include algorithms from calculus(like integration), statistics, linear algebra, image-processing, signal processing, machine learning, etc.

Unit 6 - Python for Big Data and X-Informatics: NumPy, SciPy, MatPlotlib

Go To Unit 6

Overview:

This section is meant to give an overview of the python tools needed for doing for this course. These are really powerful tools which every data scientist who wishes to use python must know.

Section 4 - Physics Case Study

This section Contains Unit 7,8,9,10

Overview:

This section starts by describing the LHC accelerator at CERN and evidence found by the experiments suggesting existence of a Higgs Boson. The huge number of authors on a paper, remarks on histograms and Feynman diagrams is followed by an accelerator picture gallery. The next unit is devoted to Python experiments looking at histograms of Higgs Boson production with various forms of shape of signal and various background and with various event totals. Then random variables and some simple principles of statistics are introduced with explanation as to why they are relevant to Physics counting experiments. The unit introduces Gaussian (normal) distributions and explains why they seen so often in natural phenomena. Several Python illustrations are given. Random Numbers with their Generators and Seeds lead to a discussion of Binomial and Poisson Distribution. Monte-Carlo and accept-reject methods. The Central Limit Theorem concludes discussion.

Unit 7 - Part I: Bumps in Histograms, Experiments and Accelerators

Go To Unit 7

Overview:

n this short unit Geoffrey describes the LHC accelerator at CERN and evidence found by the experiments ATLAS suggesting existence of a Higgs Boson. The huge number of authors on a paper, remarks on histograms and Feynman diagrams is followed by an accelerator picture gallery.

Unit 8 - Part II: Python Event Counting for Signal and Background (Python Track)

Go To Unit 8

Overview:

This unit is devoted to Python experiments with Geoffrey looking at histograms of Higgs Boson production with various forms of shape of signal and various background and with various event totals

Section 4 - Physics Case Study

This section Contains Unit 7,8,9,10

Unit 9 - Part III: Random Variables, Physics and Normal Distributions(Python Track)

Go To Unit 9

Overview:

Geoffrey introduces random variables and some simple principles of statistics and explains why they are relevant to Physics counting experiments. The unit introduces Gaussian (normal) distributions and explains why they seen so often in natural phenomena. Several Python illustrations are given.

Unit 10 - Part IV: Random Numbers, Distributions and Central Limit Theorem(Python Track)

Go To Unit 10

Overview:

Geoffrey discusses Random Numbers with their Generators and Seeds. It introduces Binomial and Poisson Distribution. Monte-Carlo and accept-reject methods are discussed. The Central Limit Theorem concludes discussion. Python examples and Physics applications are given.

Section 5: Technology Training

This section Contains Unit 11

Unit 11: Using Plotviz Software for Displaying Point Distributions in 3D

Go To Unit 11

Overview:

Geoffrey introduces Plotviz, a data visualization tool developed at Indiana University to display 2 and 3 dimensional data. The motivation is that the human eye is very good at pattern recognition and can "see" structure in data. Although most Big data is higher dimensional than 3, all can be transformed by dimension reduction techniques to 3D. He gives several examples to show how the software can be used and what kind of data can be visualized. This includes individual plots and the manipulation of multiple synchronized plots. Finally, he describes the download and software dependency of Plotviz.

Section 6 - e-Commerce and LifeStyle Case Study

This section Contains Unit 12,13,14,15,16

Overview:

Recommender systems operate under the hood of such widely recognized sites as Amazon, eBay, Monster and Netflix where everything is a recommendation. This involves a symbiotic relationship between vendor and buyer whereby the buyer provides the vendor with information about their preferences, while the vendor then offers recommendations tailored to match their needs. Kaggle competitions h improve the success of the Netflix and other recommender systems. Attention is paid to models that are used to compare how changes to the systems affect their overall performance. Geoffrey muses how the humble ranking has become such a dominant driver of the world's economy. More examples of recommender systems are given from Google News, Retail stores and in depth Yahoo! covering the multi-faceted criteria used in deciding recommendations on web sites. The formulation of recommendations in terms of points in a space or bag is given where bags of item properties, user properties, rankings and users are useful. Detail is given on basic principles behind recommender systems: user-based collaborative filtering, which uses similarities in user rankings to predict their interests, and the Pearson correlation, used to statistically quantify correlations between users viewed as points in a space of items. Items are viewed as points in a space of users in item-based collaborative filtering. The Cosine Similarity is introduced, the difference between implicit and explicit ratings and the k Nearest Neighbors algorithm. General features like the curse of dimensionality in high dimensions are discussed. A simple Python k Nearest Neighbor code and its application to an artificial data set in 3 dimensions is given. Results are visualized in Matplotlib in 2D and with Plotviz in 3D. The concept of a training and a testing set are introduced with training set pre labeled. Recommender system are used to discuss clustering with k-means based clustering methods used and their results examined in Plotviz. The original labelling is compared to clustering results and extension to 28 clusters given. General issues in clustering are discussed including local optima, the use of annealing to avoid this and value of heuristic algorithms.

Unit 12 - Part I: Recommender Systems: Introduction

Go To Unit 12

Overview:

Geoffrey introduces Recommender systems as an optimization technology used in a variety of applications and contexts online. They operate in the background of such widely recognized sites as Amazon, eBay, Monster and Netflix where everything is a recommendation. This involves a symbiotic relationship between vendor and buyer whereby the buyer provides the vendor with information about their preferences, while the vendor then offers recommendations tailored to match their needs, to the benefit of both.
There follows an exploration of the Kaggle competition site, other recommender systems and Netflix, as well as competitions held to improve the success of the Netflix recommender system. Finally attention is paid to models that are used to compare how changes to the systems affect their overall performance. Geoffrey muses how the humble ranking has become such a dominant driver of the world's economy.

Unit 13- Part II: Recommender Systems: Examples and Algorithms

Go To Unit 13

Overview:

Geoffrey continues the discussion of recommender systems and their use in e-commerce. More examples are given from Google News, Retail stores and in depth Yahoo! covering the multi-faceted criteria used in deciding recommendations on web sites. Then the formulation of recommendations in terms of points in a space or bag is given.
Here bags of item properties, user properties, rankings and users are useful. Then we go into detail on basic principles behind recommender systems: user-based collaborative filtering, which uses similarities in user rankings to predict their interests, and the Pearson correlation, used to statistically quantify correlations between users viewed as points in a space of items.

Section 6 - e-Commerce and LifeStyle Case Study

This section Contains Unit 12,13,14,15,16

Unit 14 - Part III: Item-based Collaborative Filtering and its Technologies

Go To Unit 14

Overview:

Geoffrey introduces Recommender systems as an optimization technology used in a Geoffrey moves on to item-based collaborative filtering where items are viewed as points in a space of users. The Cosine Similarity is introduced, the difference between implicit and explicit ratings and the k Nearest Neighbors algorithm. General features like the curse of dimensionality in high dimensions are discussed

Unit 15 - Part IV: k Nearest Neighbor Algorithm(Python Track)

Go To Unit 15

Overview:

Geoffrey discusses a simple Python k Nearest Neighbor code and its application to an artificial data set in 3 dimensions. Results are visualized in Matplotlib in 2D and with Plotviz in 3D. The concept of training and testing sets are introduced with training set pre-labelled.

Unit 16 - Part V: Clustering

Go To Unit 16

Overview:

Geoffrey uses example of recommender system to discuss clustering. The details of methods are not discussed but k-means based clustering methods are used and their results examined in Plotviz. The original labelling is compared to clustering results and extension to 28 clusters given. General issues in clustering are discussed including local optima, the use of annealing to avoid this and value of heuristic algorithms.

Section 7 - Infrastructure and Technologies for Big Data X-Informatics

This section Contains Unit 17,18,19,20

Overview:

Clouds and Big Data which is decomposed into lots of “Little data” running in individual cores. Many examples are given and it is stressed that issues in parallel computing are seen in day to day life for communication, synchronization, load balancing and decomposition. Cyberinfrastructure for e-moreorlessanything or moreorlessanything-Informatics and the basics of cloud computing are introduced. This includes virtualization and the important “as a Service” components and we go through several different definitions of cloud computing. Gartner’s Technology Landscape includes hype cycle and priority matrix and covers clouds and Big Data. Two simple examples of the value of clouds for enterprise applications are given with a review of different views as to nature of Cloud Computing. This IaaS (Infrastructure as a Service) discussion is followed by PaaS and SaaS (Platform and Software as a Service). Features in Grid and cloud computing and data are treated. Cloud (Data Center) Architectures with physical setup, Green Computing issues and software models are discussed followed by the Cloud Industry stakeholders and applications on the cloud including data intensive problems and comparison with high performance computing. Remarks on Security, Fault Tolerance and Synchronicity issues in cloud follow. The Big Data Processing from an application perspective with commercial examples including eBay concludes section.

Unit 17 - Parallel Computing: Overview of Basic Principles with familiar Examples

Go To Unit 17

Overview:

Geoffrey describes the central role of Parallel computing in Clouds and Big Data which is decomposed into lots of “Little data” running in individual cores. Many examples are given and it is stressed that issues in parallel computing are seen in day to day life for communication, synchronization, load balancing and decomposition.

Unit 18 - X-Informatics Cloud Technology Part I: Introduction

Go To Unit 18

Overview:

Geoffrey discusses Cyberinfrastructure for e-moreorlessanything or moreorlessanything-Informatics and the basics of cloud computing. This includes virtualization and the important “as a Service” components and we go through several different definitions of cloud computing.
Gartner’s Technology Landscape includes hype cycle and priority matrix and covers clouds and Big Data. Geoffrey reviews the nature of 48 technologies in 2012 emerging technology hype cycle. Gartner has specific predictions for cloud computing growth areas. The unit concludes with two simple examples of the value of clouds for enterprise applications.

Section 7 - Infrastructure and Technologies for Big Data X-Informatics

This section Contains Unit 17,18,19,20

Unit 19 - X-Informatics Cloud Technology Part II: Introduction

Go To Unit 19

Overview:

Geoffrey covers different views as to nature of Cloud Computing. This IaaS (Infrastructure as a Service) discussion is followed by PaaS and SaaS (Platform and Software as a Service). The unit discusses features introduced in Grid computing and features introduced by clouds. The unit concludes with the treatment of data in the cloud from an architecture perspective.

Unit 20 - X-Informatics Cloud Technology Part III: Introduction

Go To Unit 20

Overview:

Geoffrey opens up with a discussion of Cloud (Data Center) Architectures with physical setup, Green Computing issues and software models. A discussion of Cloud Industry stakeholders is followed by applications on the cloud including data intensive problems and comparison with high performance computing. Remarks on Security, Fault Tolerance and Synchronicity issues in cloud follow. The Big Data Processing from an application perspective with commercial examples including eBay concludes unit.

Section 8 - Web Search Informatics

This section Contains Unit 21,22

Overview:

This section starts with an overview of data mining and puts our study of classification, clustering and exploration methods in context. We examine the problem to be solved in web and text search and note the relevance of history with libraries, catalogs and concordances. An overview of web search is given describing the continued evolution of search engines and the relation to the field of Information Retrieval. The importance of recall, precision and diversity is discussed. The important Bag of Words model is introduced and both Boolean queries and the more general fuzzy indices. The important vector space model and revisiting the Cosine Similarity as a distance in this bag follows. The basic TF-IDF approach is dis cussed. Relevance is discussed with a probabilistic model while the distinction between Bayesian and frequency views of probability distribution completes this unit. Geoffrey starts with an overview of the different steps (data analytics) in web search and then goes key steps in detail starting with document preparation. An inverted index is described and then how it is prepared for web search. The Boolean and Vector Space approach to query processing follow. This is followed by Link Structure Analysis including Hubs, Authorities and PageRank. The application of PageRank ideas as reputation outside web search is covered. The web graph structure, crawling it and issues in web advertising and search follow. The use of clustering and topic models completes section.

Unit 21- X-Informatics Web Search and Text Mining I

Go To Unit 21

Overview:

Geoffrey starts this unit with an overview of data mining and noting our study of classification, clustering and exploration methods. We examine the problem to be solved in web and text search and note the relevance of history with libraries, catalogs and concordances. An overview of web search is given describing the continued evolution of search engines and the relation to the field of Information Retrieval.
The importance of recall, precision and diversity is discussed. The important Bag of Words model is introduced and both Boolean queries and the more general fuzzy indices. The important vector space model and revisiting the Cosinne Similarity as a distance in this bag follows. The basic TF-IDF approach is discussed. Relevance is discussed with a probabilistic model while the distinction between Bayesian and frequency views of probability distribution completes this unit.

Unit 22 - X-Informatics Web Search and Text Mining II

Go To Unit 22

Overview:

Geoffrey starts with an overview of the different steps (data analytics) in web search and then goes key steps in detail starting with document preparation. An inverted index is described and then how it is prepared for web search. The Boolean and Vector Space approach to query processing follow.
This is followed by Link Structure Analysis including Hubs, Authorities and PageRank. The application of PageRank ideas as reputation outside web search is covered. The web graph structure, crawling it and issues in web advertising and search follow. The use of clustering and topic models completes unit.

Section 9 - Technology for X-Informatics

This section Contains Unit 23,24,25,26

Overview:

Geoffrey uses the K-means Python code in SciPy package to show real code for clustering. After a simple example we generate 4 clusters of distinct centers and various choice for sizes using Matplotlib tor visualization. We show results can sometimes be incorrect and sometimes make different choices among comparable solutions. We discuss the “hill” between different solutions and rationale for running K-means many times and choosing best answer. Then we introduce MapReduce with the basic architecture and a homely example. The discussion of advanced topics includes an extension to Iterative MapReduce from Indiana University called Twister and a generalized Map Collective model. Some measurements of parallel performance are given. The SciPy K-means code is modified to support a MapReduce execution style. This illustrates the key ideas of mappers and reducers. With appropriate runtime this code would run in parallel but here the “parallel” maps run sequentially. This simple 2 map version can be generalized to scalable parallelism. Python is used to Calculate PageRank from Web Linkage Matrix showing several different formulations of the basic matrix equations to finding leading eigenvector. The unit is concluded by a calculation of PageRank for general web pages by extracting the secret from Google.

Unit 23 - PageRank (Python Track)

Go To Unit 23

Overview:

Geoffrey uses Python to Calculate PageRank from Web Linkage Matrix showing several different formulations of the basic matrix equations to finding leading eigenvector. The unit is concluded by a calculation of PageRank for general web pages by extracting the secret from Google.

Unit 24 - K-means (Python Track)

Go To Unit 24

Overview:

Geoffrey uses the K-means Python code in SciPy package to show real code for clustering. After a simple example we generate 4 clusters of distinct centers and various choice for sizes using Matplotlib tor visualization. We show results can sometimes be incorrect and sometimes make different choices among comparable solutions. We discuss the “hill” between different solutions and rationale for running K-means many times and choosing best answer.

Unit 25 - MapReduce

Go To Unit 25

Overview:

Geoffrey’s introduction to MapReduce describes the basic architecture and a homely example. The discussion of advanced topics includes extension to Iterative MapReduce from Indiana University called Twister and a generalized Map Collective model. Some measurements of parallel performance are given.

Section 9 - Technology for X-Informatics

This section Contains Unit 23,24,25,26

Unit 26 - Kmeans and MapReduce Parallelism (Python Track)

Go To Unit 26

Overview:

Geoffrey modifies the SciPy K-means code to support a MapReduce execution style and runs it in this short unit. This illustrates the key ideas of mappers and reducers. With appropriate runtime this code would run in parallel but here the “parallel” maps run sequentially. Geoffrey stresses that this simple 2 map version can be generalized to scalable parallelism.

Section 10 - Health Informatics

This section Contains Unit 27

Unit 27 - Health Informatics

Go To Unit 27

Overview:

Geoffrey starts by discussing general aspects of Big Data and Health including data sizes, different areas including genomics, EBI, radiology and the Quantified Self movement. We survey an April 2013 McKinsey report on the Big Data revolution in US health care; a Microsoft report in this area and a European Union report on how Big Data will allow patient centered care in the future. Some remarks on Cloud computing and Health focus on security and privacy issues. The final topic is Genomics, Proteomics and Information Visualization.

Section 11 - Sensor Informatics

This section Contains Unit 28

Unit 28 - Sensors Informatics

Go To Unit 28

Overview:

Geoffrey starts with the Internet of Things giving examples like monitors of machine operation, QR codes, surveillance cameras, scientific sensors, drones and self driving cars and more generally transportation systems. Sensor clouds control these many small distributed devices. More detail is given for radar data gathered by sensors; ubiquitous or smart cities and homes including U-Korea; and finally the smart electric grid.

Section 12 - Radar Informatics

This section Contains Unit 29

Unit 29 - Radar Informatics

Go To Unit 29

Overview:

The changing global climate is suspected to have long-term effects on much of the world's inhabitants. Among the various effects, the rising sea level will directly affect many people living in low-lying coastal regions. While the ocean-s thermal expansion has been the dominant contributor to rises in sea level, the potential contribution of discharges from the polar ice sheets in Greenland and Antarctica may provide a more significant threat due to the unpredictable response to the changing climate. The Radar-Informatics unit provides a glimpse in the processes fueling global climate change and explains what methods are used for ice data acquisitions and analysis.