Big Data Applications
e534

Geoffrey C. Fox
Gregor von Laszewski
Editor

laszewski@gmail.com

https://cloudmesh-community.github.io/book/vonlaszewski-big-data-applications.epub

September 06, 2019 - 01:36 PM

Created by Cloudmesh & Cyberaide Bookmanager, https://github.com/cyberaide/bookmanager

Big Data Applications

Geoffrey C. Fox Gregor von Laszewski

Big Data Applications

1 PREFACE

Fri Sep 6 13:36:29 EDT 2019

1.1 CONTRIBUTORS

Contributors are sorted by the first letter of their combined Firstname and Lastname and if not available by their github ID. Please, note that the authors are identified through git logs in addition to some contributors added by hand. The git repository from which this document is derived contains more than the documents included in this document. Thus not everyone in this list may have directly contributed to this document. However if you find someone missing that has contributed (they may not have used this particular git) please let us know. We will add you. The contributors that we are aware of include:

> *Anand Sriramulu, Ankita Rajendra Alshi, Anthony Duer, Arnav, Averill Cate, Jr, Bertolt Sobolik, Bo Feng, Brad Pope, Dave DeMeulenaere, De'Angelo Rutledge, Eliyah Ben Zayin, Eric Bower, Fugang Wang, Geoffrey C. Fox, Gerald Manipon, Gregor von Laszewski, Hyungro Lee, Ian Sims, IzoldaIU, Javier Diaz, Jeevan Reddy Rachepalli, Jonathan Branam, Juliette Zerick, Keith Hickman, Keli Fine, Kenneth Jones, Mallik Challa, Mani Kagita, Miao Jiang, Mihir Shanishchara, Min Chen, Murali Cheruvu, Orly Esteban, Pulasthi Supun, Pulasthi Supun Wickramasinghe, Pulkit Maloo, Qianqian Tang, Ravinder Lambadi, Richa Rastogi, Ritesh Tandon, Saber Sheybani, Sachith Withana, Sandeep Kumar Khandelwal, Sheri Sanders, Silvia Karim, Swarnima H. Sowani, Tharak Vangalapat, Tim Whitson, Tyler Balson, Vafa Andalibi, Vibhatha Abeykoon, Vineet Barshikar, Yu Luo, ahilgenkamp, aralshi, bfeng, brandonfischer99, btpope, garbeandy, harshadpitkar, himanshu3jul, hrbahramian, isims1, janumudvari, joshish-iu, juaco77, karankotz, keithhickman08, kkp, mallik3006, manjunathsivan, niranda perera, qianqian tang, rajni-cs, rirasto, shilpasingh21, swsachith, toshreyanjain, trawat87, tvangalapat, varunjoshi01, vineetb-gh, xianghang mi, zhengyili4321*

1.2 ePub Readers

This document is distributed in ePub format. Every OS has a suitable ePub reader to view the document. Such readers can also be integrated into a Web browser so that when you click on an ePub it is automatically openedin your browser. As we use ePubs the document can be scaled based on the users preference If you ever see a content that does not fit on a page we recommend you zoom out to make sure you can see the entire content.

We have made good experiences with the following readers:

- **macOSX**: Books, which is a build in ebook reader
- **Windows 10**: Microsoft edge, but it must be the newest version, as older versions have bugs. Alternatively use calibre
- **Linux**: calibre

If your have an iPad or Tablet with enough memory, you may also be able to use them.

Sometimes you may want to adjust the zoom of you reader to increase or decrease it. Please adjust your zoom to a level that is comfortable for you. On macOS with a larger monitor we found that zooming out multiple times results in very good rendering allowing you to see the source code without horizontal scrolling.

1.3 Notation

The material here uses the following notation. This is especially helpful, if you contribute content, so we keep the content consistent.

if you like to see the details on how to create them in the markdown documents, you will have to look at the file source while clicking on the cloud in the heading of the Notation section (Section 1.3). This will bring you to the markdown tex, but you will still have to look at the raw content to see the details.

or 

```
![Github](images/github.png)
```

If you click on the  *or*  *in a heading, you can go directly to the > document in github that contains the next content. This is > convenient to fix errors or make additions to the content. The cloud will be automatically added upon inclusion of a new markdown file that includes in its first line a section header.*

$

Content in bash is marked with verbatim text and a dollar sign

```
$ This is a bash text
```

[1]

References are indicated with a number and are included in the > reference chapter [1]. Use it in markdown with `[@las14cloudmeshmultiple]`*. References must be added to the* `refernces.bib` *file in BibTex format.*

O or 

Chapters marked with this emoji are not yet complete or have some issue that we know about. These chapters need to be fixed. If you like to help us fixing this section, please let us know. Use it in markdown with `:o2:` *or if you like to use the image with* `![No](images/no.png)`*.*

 REST 36:02

Example for a video with the `![Video](images/video.png)` *emoji. Use it in markdown with* `[![Video](images/video.png) REST 36:02](https://youtu.be/xjFuA6q5N_U)`

 Slides 10

Example for slides with the `![Presentation](images/presentation.png)` *emoji. These slides may or may not include audio.*

 Slides 10

Slides without any audio. They may be faster to download. Use it in markdown with `[![Presentation](images/presentation.png) Slides 10](TBD)`.

A set of learning objectives with the `![Learning](images/learning.png)` *emoji.*

A section is release when it is marked with this emoji in the syllabus. Usie it in markdown with `![Ok](images/ok.png)`.

Indicates opportunities for contributions. Usie it in markdown with `![Question](images/question.png)`.

Indicates sections that are worked on by contributors. Use it in markdown with `![Construction](images/construction.png)`.

Sections marked by the contributor with this emoji `![Smiley](images/smile.png)` *when they are ready to be reviewed.*

Sections that need modifications are indicated with this emoji `![Comment](images/comment.png)`.

A warning that we need to look at in more detail `![Warning](images/warning.png)`

> *Notes are indicated with a bulb* `![Idea](images/idea.png)`

Other emojis

> *Other emojis can be found at https://gist.github.com/rxaviers/7360908. However, note that emojis may not be viewable in other formats or on all platforms. We know that some emojis do not show in calibre, but they do show in macOS iBooks and MS Edge*

This is the list of emojis that canbe converted to PDF. So if you like a PDF, please limit your emojis to

`:cloud:` ☁ `:o2:` 🅾 `:relaxed:` ☺ `:sunny:` ☀ `:baseball:` ⚾ `:spades:` ♠ `:hearts:` ♥ `:clubs:` ♣ `:diamonds:` ♦
`:hotsprings:` ♨ `:warning:` ⚠ `:parking:` 🅿 `:a:` 🅰 `:b:` 🅱 `:recycle:` ♻ `:copyright:` © `:registered:` ® `:tm:` ™
`:bangbang:` ‼ `:interrobang:` ⁉ `:scissors:` ✂ `:phone:` ☎

1.3.1 Figures

Figures have a caption and can be refereed to in the ePub simple with a number. We show such a reference pointer while referring to Figure 1.

Figure 1: Figure example

Figures must be written in the md as

```
![Figure example](images/code.png){#fig:code-example width=1in}
```

Note that the text must be in one line and must not be broken up even if it is longer than 80 characters. You can refer to them with `@fig:code-example`. Please note in order for numbering to work figure references must include the `#fig:` followed by a unique identifier. Please note that identifiers must be really unique and that identifies such as `#fig:cloud` or similar simple identifiers are a poor choice and will likely not work. To check, please list all lines with an identifier such as.

```
$ grep -R "#fig:" chapters
```

and see if your identifier is truly unique.

1.3.2 Hyperlinks in the document

To create hyperlinks in the document other than images, we need to use proper markdown syntax in the source. This is achieved with a refernce for example in sections headers. Let us discuss the refernce header for this section, e.g. Notation. We have augmented the section header as follows:

```
# Notation {#sec:notation}
```

Now we can use the refernce in the text as follows:

```
In @sec:notation we explain ...
```

It will be rendered as: In Section 1.3 we explain …

1.3.3 Equations

Equations can be written as

```
$$a^2+b^2=c^2$${#eq:pythagoras}
```

and used in text:

$$a^2 + b^2 = c^2 \qquad (1)$$

It will render as: As we see in Equation 1.

The equation number is optional. Inline equations just use one dollar sign and do not need an equation number:

```
This is the Pythagoras theorem: $a^2+b^2=c^2$
```

Which renders as:

This is the Pythagoras theorem: $a^2 + b^2 = c^2$.

1.3.4 Tables

Tables can be placed in text as follows:

```
: Sample Data Table {#tbl:sample-table}

x   y   z
--- --- ---
1   2   3
4   5   42
```

As usual make sure the label is unique. When compiling it will result in an error if labels are not unique. Additionally there are several md table generators available on the internet and make creating table more efficient.

2 ORGANIZATION

2.1 ORGANIZATION

This class is an online class. Online classes require you to be very disciplined in order to execute the tasks necessary for the class in time. It is your responsibility to organize the lessons so that you can complete them not only by the end of the semester, but also in time for conducting your assignments. This is a great opportunity for you to structure the class based on your availability. The classes are attended by two different set of students. One set are remote online students, while to other are residential students. For the residential students we have a mandatory in person meeting that takes place at the posted location and hours once a week. For pure online students we have weekly online hours that we will identify based on our availability and a doodle poll.

Figure *Components of the Class i523, i423, e534* showcases the different parts of the class. If you have taken a previous class with us you are able to continue your previous project upon approval. It must however be a significant improvement. Please note that the in i523 and i524 the project and it's report can be substituted by a longer term paper that does not require programming. As this is a significant reduction in work and goals, for that class, the maximum grade in this case for the entire class can only be an A-.

There will not be any bonus projects or tasks to improve grades. Instead make sure your deliverables of the few assignments are truly outstanding.

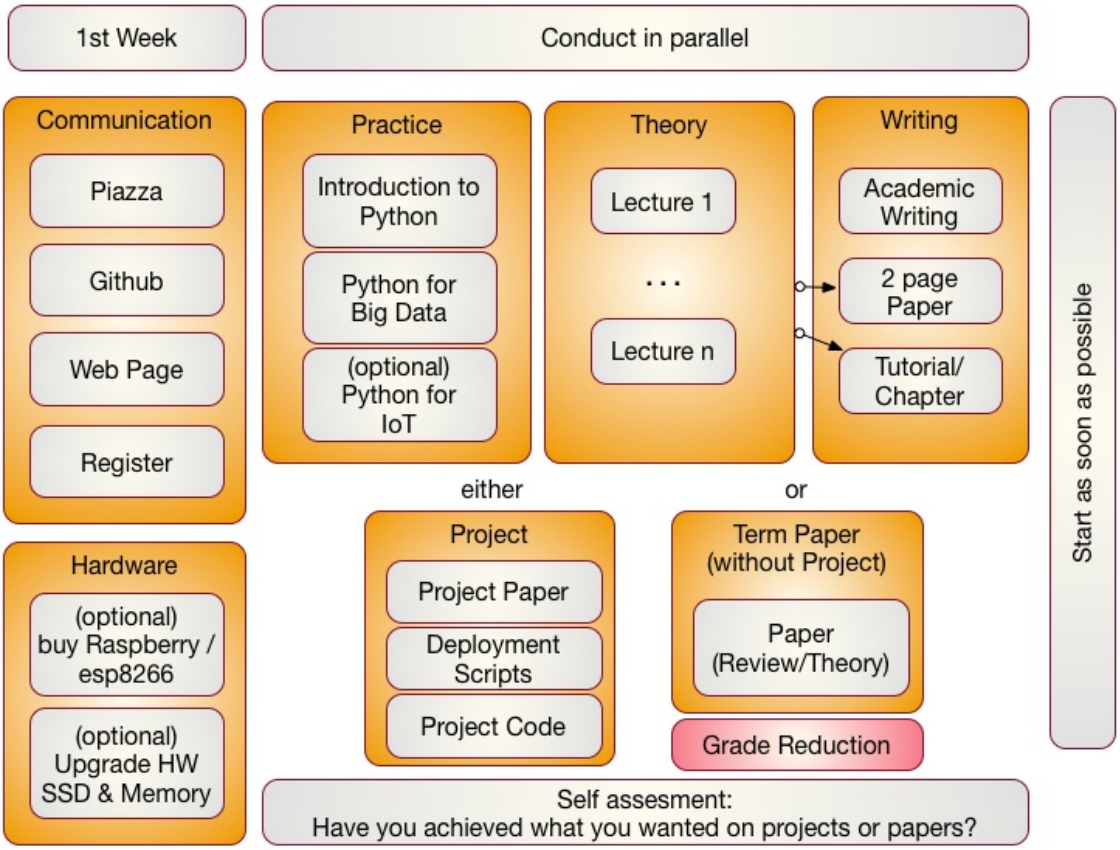

Figure: Components of the Class i523, i423, e534

The content for this class will be available through a series of documents that will be regularly updated and are linked from this document. All communication is done with Piazza. issues.

2.1.1 First Week

In the first week we will be introducing you how we communicate to you. Naturally you need to register for the class. Once you register you need to set up a number of services.

2.1.2 Access to Clouds

As part of the course you will also need access to a cloud. We will try our best to provide you with access to suitable computers for the class, but do be reminded that the amount of time and access to supercomputers and clouds we offer is limited. Our class policy is to use the compute resources only when you really need them. Thus you **must** shut down your VMs when they are not in use. It would be a violation of class policy if we would find out through an analysis of the cloud logs that you unnecessarily keep your VMs running. Thus we will implement a **strict policy** that you must record yourself how many hours you run VM's and provide this information to us. We will than compare that time with the time recorded by the computer system as well as with your target application and will deduct points form your project if you can not justify why you have not shut down your VMs. A resource section needs to be added to your report justifying the used resources.

Why is this such a big deal you may ask? For example we estimate if every student in class violates this policy it would cost about $200000 to rent the time for this on a public cloud. Due to this high cost, we no longer tolerate deliberate violations of the policy and will terminate your account. Furthermore, violators will have to find alternative resources to conduct their projects while not using our resources. In our case the problem is even beyond the issue of cost as our allocation on the clouds would be terminated due to abuse and **no student**, including those that follow policies, could use the cloud. It may take weeks to reestablish cloud access and would effect every student in class.

We will provide clarification for accessing cloud resources and teach you how to avoid getting in such a situation. I am sure that a future employer of yours will be real happy if you have a deep understanding of resource vs. cost estimate.

Listing the used computer time for your project is part of your report.

2.1.3 Using Your Own Computer

In many cases however you could and are recommended to use your own personal computer, but make sure the computer is up-to-date. We also like to make sure that you do not use a work computer as you need to avoid that when you develop a cloud program you do not by accident introduce a security risk on your machine. This does not mean that you need to buy a new computer, or need to upgrade it. However, if you consider an upgrade of an older machine please

consider the following.

These days we recommend that your computer has a solid-state drive and fast memory (put as much memory in your machine as is supported). We recommend 16 GB off main memory which gives you enough space to run containers, virtual machines and naturally the main operating system. We fond that students with only 8GB could do the work but it was slow. In some cases the memory to conduct their projects was not sufficient. Make sure you follow your upgrade guide to your computer and by suitable memory chips. In most cases you have to buy them in **pairs** and make sure all chips in your computer are the same. When it comes to buying a solid-state drive, make sure that you buy one that is compatible with motherboards bus speed. As you may want to reuse your solid-state drive at a later time I suggest to get a 6GB/s SSD and not a 3GB/s.

In case of Windows, your could also get yourself a UBS stick or external SSD drive and place ubuntu on it. You could than use your bios to boot of from that drive. This way you do not have to modify anything on your computer. This method works very well for most computers and allows you to use the maximum memory while for example using ubuntu.

Students that only had a chromebook and took this class gave us the feedback that they are too inconvenient as they do not allow you to program directly in python on them and the ssh terminals to login to other computer although working are not supporting the GUI tools.

Another option is (if money is an issue) you can buy a Raspberry Pi and edit your programs there and when satisfied run them on a cloud. However a PI is small and has only very limited memory and processing power.

We also like to remind you that this course does not require you to purchase expensive text books, thus the money you safe on this could be used in upgrading your hardware or renting yourself from your own money time on AWS. However, be careful with the cloud its easy to spend lots of money there if you are not careful.

2.1.3.1 Self Discipline

As this class has no graded tests and only few graded homework, we like that

you deliver an **exceptional** project report or paper. Instead of focusing on preparing for tests we provide you with the opportunity to **explore** without the pressure of grades. However you should not give up or take the easy way out or it will effect you in your project execution. Also, to achieve your best do not just say: *We do not have a test, so let me not do this weeks assignment, let me do it next week.* After a couple of times with this attitude you will be in big trouble. All this requires discipline. For example, if you believe you are so good that you can do a project within one week before deadline, you will **certainly fail**. To avoid this and to introduce discipline, you will also be monitored on progress and we check your github for activities which will be part of the participation grade.

It will be up to you to assess what you want to deliver before handing it in to us. Self assessment or a check with other students is a real good way to do that. You should not expect to get an A if you yourself are not convinced about your project or are unsure about it. Common sense prevails.

2.1.3.2 Fun

I hope you have fun and are able to integrate in the projects your own thoughts and interests.

We have quotes form students such as

> *"This is the best class I have taken …"*

or

> *"I really enjoyed taking this class and having maximum flexibility to schedule the lectures."*

or

> *"The lessons learned from this class were adopted within my company."*

Furthermore you should know that the way we teach the class has also been adopted in STEM classes. As a result a team coached by Gregor von Laszewski

won an award at the FLL Robotics World Championship. The certainly had lots of fun and integrated their own ideas into the project that won the award.

2.1.3.3 Uniqueness

We will try to have every project or paper to be non overlapping with another topic, If there are overlaps we may ask you to modify your focus.

2.1.3.4 Continuation

If you like to put additional effort in the project, the report could be made to a conference or workshop paper. Dr. von Laszewski is happy to help as co-author.

2.1.4 Parallel Tracks

In this class we have three parallel tracks.

2.1.4.1 Track 1: Practice

Track 1 introduces you to using python for Big Data. We recommend that you do know a programming language for any of our courses. Learning a programming language is not part of the hours you spend for this class. It is an additional time requirement that you must plan for. Maybe you want to take for example a python programming language class at the same time. This can also be done in self study. Although you do not need to know any programming language, it is certainly useful as it will make this course much easier for you. We had students that had no prior programming knowledge and successfully completed the course. So we know it can be done. The course is designed in such a fashion, that there is enough time to learn programming and do a project.

We provide you with a general introduction to Python. This includes enough knowledge so you can conduct a project with it. We will build on these technologies to introduce you to python libraries that can be used for big data. Not every section in the Python chapter will need to be used in this class. At minimum you must understand python classes, and the map reduce function.

2.1.4.2 Track 2: Theory

The theory track includes a number of online lectures that introduces you to a variety of topics related to Big Data. You have especially the opportunity to become part of a project that would contribute to the understanding and the development of a Big Data Architecture developed in collaboration with NIST. Other topics that are covered include IoT, Health Care, Physics, Science, Biology, Genomics, and so forth. We will update the Theory track and will release lectures in the specified areas. Some lectures may be used in multiple classes.

2.1.4.3 Track 3: Writing

You have a choice in this class between writing a two page review paper about a big data technology or application (area), or contribute a chapter to this document. We explain next the difference:

Review Paper: A review paper will introduce your into how to write an academic paper and conduct proper bibliography management. Knowing how to write is a preparation for your term project.

You will be writing a paper that is 2 pages long (in a particular format, typically ACM) possibly within a team. In case you work in a team you have to produce as many papers as you have team members. We like to avoid that all students take the same topic. We will use github to avoid that everyone chooses same topic. Knowing how to write is a preparation fo your project or term paper.

We noticed a curious observation in previous classes. Any paper written in MSWord was inferior. Thus we no longer provide the choice to write papers in MSWord in order for you to achieve your best. Papers and Project reports must be written in LaTeX or markdown. For the classes starting in 2018 we do prefer markdown and may restrict all document to this format.

Chapter: A is chapter to a review paper, but is written in markdown and can be added to the lecture notes. A chapter should be formulated in a consistent form and is equivalent in length (number of words) to those of the 2 page paper. Bibliography management is conducted in bibtex and can be used in the

markdown document.

Important in both cases is that you stay focused. You can assume that if you write a document about “Big Data in Baseball”, you do not spend 1.5 pages to describe what big data is and only half a page where baseball fits in. What you should do is focus on the topic. A chapter could also include some practical lessons with real programming lessons.

2.1.4.4 Track 4: Term Paper/Project

The major deliverable of the course is a term project or paper. The exact details will be posted on the Web page in this document. The important part is that you start on this project once you are sufficiently familiar with Track 1-3. However you can also use the project to for example learn python and engage in a goal oriented learning activity while working towards implementing your project and integrating the python lessons that you encounter. The same is valid for the theory.

It is **expected** that you identify a suitable analysis and data set for the project and that you learn how to apply this analysis as well as justify it. It is part of the learning outcome that you determine this instead of us giving you a topic.

Furthermore, it is also important to note that if you do not do a project (this is your option) the maximum grade for the entire class is limited to an A-. This is achieved simply by reducing the grade of your term report by a full grade due to the distribution of the grade this will result in a fractional grade reduction and limits the maximum grade to an A-.

Starting in 2018 the paper format will be Markdown.

2.1.5 Plagiarism

In the first week(s) of class you will need to read the information about plagiarism. If there are any questions about plagiarism we require you to take a course offered from the IU educational department.

Warning:

> *If we find cheating or plagiarism, your assignment will be receiving an F. This especially includes copying text without proper attribution. We are required to follow IU policy and report your case to the dean of students who may elect to expel you form the university. Please understand that it is your doing and the instructors have no choice as to follow university policies. Thus, please do not blame the instructors for your actions. Excuses such as "I missed the lecture on plagiarism", "I forgot to include the original reference as I ran out of time", "I did not understand what plagiarism is" do not apply as we explicitly make the policies clear. This applies to all material prepared for class including assignments, exercises, code, sections, tutorials, papers, and projects. If there is no time, do not submit and instead of an F ask for an incomplete. In fact if you know you have plagiarized, do not even have us review your paper.*

For more information on this topic please see:

- https://studentaffairs.indiana.edu/student-conduct/misconduct-charges/academic-misconduct.shtml

Furthermore you are supposed to review our lecture material on plagiarism and take the plagiarism test. The information is located at:

- Scientific Writing with Markdown

In Piazza a form will be posted that will ask you for your passing ID. If the form is not yet posted, please be patient till it is.

2.2 Course Policies

We describe briefly some class policies.

2.2.1 Discussion via Piazza

1. All communication is done in Piazza. It is an IU approved communication tool and superior to CANVAS discussion list
2. CANVAS is only used with students that are not in Piazza. The only

messages they will get is to activate Piazza and use the class Piazza
3. You are allowed to use whatever calendar system you like.
4. We will not use CANVAS calendar, however you can manage that yourself. CANVASS allows you to add events such as assignment deadlines.
5. Piazza is FERPA compliant https://piazza.com/legal/ferpa

2.2.2 Managing Your Own Calendar

From time to time we get the question from a very small number of students why we are not using or uploading the assignment deadlines and the assignment descriptions to CANVAS. The reason for this is manifold. First, our class has different deadlines for different students within the same class. This is not supported by CANVAS and if we would use CANVAS leads to confusion and clearly shows the limitation of CANVAS. Second, we are teaching cloud computing. CANVAS is not a tool that you likely will use after graduating. Thus we are providing you the ability to explore industry standard tools such as github to maintaining your own tasks and deadlines, while for example using github issues (see the section about github). We highly recommend that you explore this as part of this class and you will see that managing the assignments in github is **superior** to CANVAS. Naturally you can not make that assessment if you are not trying it. Thus we like you to do so and it is part of any assignment in your class to use github issues to manage your assignments for this class.

However, if you still want to manage your tasks in CANVAS, you can do so. CANVAS allows you to create custom events, so if you see an assignment in piazza or the handbook, you are more than welcome to add that task yourself to your own CANVAS tasks. As we have only a very small number of assignments this will not pose a problem either for graduate or undergraduate. Being able to organize your deadlines and assignment with industry accepted tools is part of your general learning experience at IU.

Obviously, this makes it also possible to use any other task or calendar system that you may use such as google calendar, jira, microsoft project, and others.

As you can see through this strategy we provide the most flexible system for any student of the class, while giving each student the ability to chose the system they prefer for managing their assignment deadlines. It is obvious that this

strategy is superior to CANVAS as it is much more general.

2.2.3 Online and Office Hours

To support you we have established an open policy of sharing information not only as part of the class material, but also as part of how we conduct support. We establish the following principals:

- in case of doubt how to communicate address this early in class and attend online hours;
- all office hours if not of personal nature are open office hours meaning that any student in class can be joined by other students of the class and all meeting times are posted publicly. This includes in person office hours with TAs. Other students are allowed to listen in and participate.
- it is in your responsibility to attend in person classes and online hours as we found that those that do get better grades. For residential students participation in the residential classes may be mandatory. International students may need to check university policies.
- instructors of this class will attempt within reason to find suitable times for you to attend an online hour in case you are an online student.

2.2.3.1 Office Hour Calendar

Online Students:

- Online hours are prioritized for online students, residential students should attend the residential meetings.

Residential Students:

- Residential students participate in the official meeting times. If additional times are required, they have to be done by appointment. Office hours will be announced publicly. All technical office hours are public and can be attended by any student. Online hours are not an excuse not to come to the residential class.

However Residential students can in addition to the residential class use the online student meeting times. However, in that case online students will be served first. It is probably good to check into the zoom meeting and identify if the TA has time. They will be in zoom.

Meeting times and phone numbers are posted in your piazza in the `Resources` section

2.2.4 Class Material

As the class material will evolve during the semester it is obvious that some content will be improved and material will be added. This benefits everyone. To stay up to date, please, revisit this document on weekly basis. This is practice in any class.

2.2.5 HID

You will be assigned an hid (Homework IDentifier) which allows us to easily communicate with you and does allow us to not use your university ID to communicate with you.

You will receive the HID within the first couple of weeks of the semester by the TA's.

2.2.6 Class Directory

You will get a class directory on github.com and not the iu github. For that reason you will be asked to give us a github id so we can create a openly accessible directory for you in which you can collaborate with the students of this class. The directories are only used to store the artifacts of the class. As all artifacts are supposed to be open source github.com provides us with the service that millions of professionals and researchers use for their work.

2.2.7 Notebook

All students are required to maintain a *class notebook* in github in which they summarize their weekly activities for this course in bullet form. This includes a self maintained list of which lecture material they viewed and what they worked

on in each week of the class.

The notebook is maintained in the class github.com in your hid project folder. It is a file called `notebook.md` that uses markdown as format. Notebooks are expected to be set up as soon as the git repository was created.

You will be responsible to set up and maintain the `notebook.md` and update it accordingly. We suggest that you prepare sections such as: Logistic, Theory, Practice, Writing and put in bullet form what you have done into these sections during the week. We can see from the github logs when you changed the notebook.md file to monitor progress. The management of the notebook will be part of your discussion grade.

The format of the notebook is very simple markdown format and must follow these rules:

- use headings with the # character and have a space after the # Use `# Week X: mm/dd/yyyy - mm/dd/yyyy` as the subject line for each week
- use bullets in each topic.
- Do not refer to section numbers form the ePub in your notebook as they can change. Instead use the section name or headline and possibly a URL. When using URLS in md format they must be enclosed in `<>` or `[text](URL)`

Please examine carefully the sample note book is available at:

- https://github.com/cloudmesh-community/hid-sample/blob/master/notebook.md

The notebook.md is not a blog and should only contain a summary of what you have done.

2.2.8 Blog

You can maintain your own optional blog. However, the blog will not be used for grading. Do not include sensitive information in either the blog or the notebook. A blog is not a replacement for the notebook. If something does not go so well, do not focus on the negative things, but focus on how that experience

can be overcome and how you turn it to a positive experience. Be positive in general.

2.2.9 Waitlist

The waitlist contains students that are unable to enroll in a section of a course. Students choose to add themselves to the waitlist. They are not automatically added, but choose to do so intentionally based on the status of the course. There are two reasons for students to be on the waitlist. The first, and primary, reason is that the class is already at the scheduled, maximum capacity. Since there are no seats available, the student can elect to add themselves to the waitlist. The second reason is that the students' own schedule has a time conflict. This occurs when they are trying to enroll in a class that overlaps with the time of a class they are already enrolled in.

Students are moved from the waitlist to the regular section during a daily batch process, and not in real time. The process is not in realtime because the registrar receives many requests to increase capacity, decrease capacity, and change rooms. If the process were real time there would be a catastrophe of conflicts.

Students are moved from the waitlist in chronological order that they added themselves to the waitlist. If you are still on the waitlist there are no spaces free, the batch process has not run for the day, or the student in question has a schedule conflict.

Faculty are not able to selectively choose students from the waitlist.

How long does the waitlist process stay active?: The automated processing of the waitlist ends on Thursday of the first week of class At this time the waitlist will no longer be processed. As the residential class starts on Friday, this may cause issues. Either talk to the department on Thursday or show up on Friday. Most likely there will be spaces left. Students on the waitlist at that time will remain on the waitlist, but remain there until the student decides to change their registration. Students may not do that, because they get assessed a change schedule fee.

Students tell me they still want to enroll after the first week of classes. How do they do this?

Beginning Monday, after the first week of class students begin to use the eAdd process to do a late addition of the course. The request is routed to the professor of record on an eDoc and the faculty will be notified via email. Faculty can deny or approve based on whatever criteria they wish to apply. If the faculty member approves, the eDoc is electronically forwarded to the Academic Operations office and we will approve the late add **if the room capacity** allows the addition, otherwise we must deny the addition because of fire marshal regulations. Many times, there are seats in a classroom/discussion/lab, but because other students have not *officially* dropped, enrollment is still at capacity.

After everything, a student that was unable to enroll in the class attended all year and completed all course work as if they had enrolled. Can the student get credit and can I give the student a grade?

Yes. There is a provision for a late registration - contact our office if this occurs. Students will be assessed a tuition fee at the time of late or retroactive registration.

2.2.10 Registration

The Executive Associate Dean for Academic Affairs requires starting Spring 2018 that students that are not officially enrolled, can not register at the end of the class if they in-officially took the class. Please make sure that within the first month you have enrolled. If we do not see in CANVAS, you are not in the class. In case you are on a waitlist it is your responsibility to work with the administration after the waitlist is over to be added to the class by getting permission from the School.

2.2.11 Auditing the class

We no longer allow students to audit the class because:

- Seating in the lecture room is limited and we want foster students that enroll full time first.

- The best way to take the class is to conduct a project. As this can not be achieved without taking the class full time and as auditing the class does not provide the full value of the class, e.g. not more than 10% of the class.

Hence, we do not think it is useful to audit the class.

- Accounts and services have to be set up and require considerable resources that are not accessible to students that audit the class.

2.2.12 Resource restrictions

- It is not allowed to use our services we offer as part of the class for profit (e.g. just enrolling in the class to use our clouds).

- In case of abuse of available compute time on our clouds the student is aware that we will terminate the computer account on our clouds and the student may have to conduct the project on a public cloud or his own computer under own cost. There will be no guarantee that cloud services we offer will be available after the semester is over. Projects can be conducted as part of the class that do not require access to the cloud.

2.2.13 Incomplete

Incompletes have to be explicitly requested in piazza through a private mail to *instructors*. All incompletes have to be filed by DATE TO BE ANNOUNCED.

Incomplete's will receive a fractional Grade reduction: A will become A-, A- will become B+, and so forth. There is enough time in the course to complete all assignments without getting an incomplete.

Why do we have such a policy? As we teach state-of-the-art software this software is subject to change, not only within the course, but also after the course. As we may offer some services and only have access to the TA's during the semester it is obvious that we like all class projects and homework assignments to be completed within a semester. Services that were offered during the semester may no longer be available after the semester is over and could adversely effect your planing. It will be in the students responsibility to identify such services and provide alternatives if they become unavailable. We try hard to avoid this but we can not guarantee it.

Furthermore, once an incomplete is requested, you will have 10 month to complete it. We will need 2 month to grade. No grading will be conducted over

breaks including the summer. This may effect those that require student loans. Please plan ahead and avoid an incomplete.

The incomplete request needs to be off the following format in piazza:

```
Subject:
    INCOMPLETE REQUEST: HID000: Lastname, Firstname

Body:
    Firstname: TBD
    Lastname: TBD
    HID: TBD
    Semester: TBD
    Course: TBD
    Online: yes/no

    URL notebook: TBD
    URL assignment1:
    URL assignment2: TBD
    ....
    URL paper1: TBD
    URL project: TBD

    URL other1: TBD
```

Please make sure that the links are clickable in piazza. Also as classes have different assignments, make sure to include whatever is relevant for that class and add the appropriate artifacts.

In case of an incomplete you may be asked to do additional assignments and assignments that have been adapted based on experience from the class. Please note also that we could reject an assignment if it is identified to no longer reflect the state-of-the-art. All previously submitted assignments such as papers, sections, and so on will be reviewed on this criterion. For example, let us assume you developed a tutorial on technology visit version x. Let us assume that since you completed this task a version x+1 comes out. It will be your obligation to update the deliverable. This is also true if the tutorial has been graded previously. The incomplete and the change of the software have at this time negated the originally assigned grade. In most cases the changes may be small. In other cases the changes could be substantial. Hence avoid an incomplete.

Here is the process for how to deal with incompletes at IU are documented:

- http://registrar.indiana.edu/grades/grade-values/grade-of-incomplete.shtml

2.2.13.1 Exercises

E.Policy.1

> *Take the Plagiarism test, See the Scientific Writing I ePub for more details.*

2.3 Course Description

2.3.1 Big Data Applications and Big Data Applications Analytics

- Indiana University
- Fall 2019
- Course Numbers: E534, I523, I423
- Faculty: Dr. Geoffrey C. Fox
- Credits: 3
- Prerequisite(s): Knowledge of a programming language, the ability to pick up other programming languages as needed, willingness to enhance your knowledge from online resources and additional literature. You will need access to a "modern" computer that allows using virtual machines and/or containers. Knowledge of material taught by e516 is desirable and will make project execution easier. e516 and this class can be taken in parallel.
- This page is maintained and updated at e534-i523: Big Data Applications and Big Data Applications Analytics
- Course Description URL: https://github.com/cloudmesh-community/book/blob/master/chapters/class/e534-i523.md
- Registrar information and Other related classes

2.3.2 Course Objectives

This class investigates the use of clouds running data analytics collaboratively for processing Big Data to solve problems in Big Data Applications and Analytics. Case studies such as Netflix recommender systems, Genomic data, Sports, Health, and more will be discussed.

The course has the following objectives:

- Provide an introduction to Big Data
- Provide an introduction to Big Data Analytics
- Provide overviews of different Big Data Application areas
- Explore state-of-the-art big data and cloud technologies and services while

providing a write up about it and exploring it practically with a section that you develop
- Enforce the theoretical knowledge with a project that you conduct in one of the application areas.

2.3.3 Learning Outcomes

- Be able to explain the concepts of the big data paradigm including its paradigm shift, its characteristics, and the advantages. Contrast them with the challenges and disadvantages.
- Be able to identify bigdata applications and analytical methods needed to support real world applications.
- Be able to implement a real world application in big data.
- Be able to conduct sophisticated analysis of big data.
- Be able to communicate the results through sections, chapters, manuals, and reports.
- Be able to work in a team to develop collaboratively software or contribute collaboratively to develop sections using clouds.

2.3.4 Course Syllabus

.	Date[a]	Unit	Title	Description
.	TBD	1	Fundamentals	Introduction to Big data
.	TBD	1	Introduction	Intoduction to Clouds
.	TBD	1	Introduction	Overview of Big Data
.	TBD			Big Data Use Cases
.	TBD	2	Basic Math	Minimal Statistics
.	TBD		Applications	Physics and Astronomy
.	TBD	3		Lifestyle and e-Commerce
.	TBD			kNN and Clustering
.	TBD			k-means
.	TBD			Web Search
.	TBD			Sports

.	TBD			Health
.	TBD			Sensor
.	TBD			Radar
.	TBD	4	Basic Cloud	Cloud Computing for Big Data I
.	TBD			Cloud Computing for Big Data II
.	TBD	5	Applications	Edge Computing and the Cloud
.	TBD*	5	Technology Summaries	Contribute your assigned technology summaries while not plagiarizing.
.	TBD*	5	Example	Contribute a significant technical example related to Big Data technologies. Do not develop redundant or duplicated content.
.	TBD*	5	Chapter	Contribute a significant chapter while not plagiarizsing that may use your example to the class documentation. Do not develop redundant or duplicated content.
.	TBD*	6	Project Type A	Build a Big Data Application Project (maximum grade for class A or A+)
.	TBD*		Project Type B	Write a comprehensive Report without programming (maximum grade for entire class A-)

Students need only to do one project. The project is conducted thought the entire semester.

a. Dates may change as the semester evolves

(*) The project is a long term assignment (and are ideally worked on weekly by residential students). It is the major part of the course grade.

(*) Sections prepare you for documenting a technical aspect related to cloud computing. It is a preparation for a document that explains how to execute your project in a reproducible manner to others.

- all times are in EST

Additional Lectures will be added that allow easy management of the project. These lectures can be taken any time when needed

2.3.5 Assessment

This course is focusing on the principal *Learning by Doing* which is assessed by simple graded and non-graded activities. The assessment may include comprehension of the material taught, programming assignments, participation in online discussion forums, or the contribution of additional material to the class showcasing your comprehension.

The comprehension is also measured by the development of one or more sections in markdown that can be distributed and replicated to other students. This is done in preparation for the project that must include a simple deployment and runtime instruction set.

The main deliverable of the class is a project. The project is assessed through the following artifacts:

- a. Deployment and install instructions,
- b. Project report (typically 2-3 pages per month, sections and chapters can be reused if possible),
- c. Working project code that can be installed and executed in reproducible manner by a third party
- d. Code developed by the project team distributed in github.com
- e. Project progress notes checked into github throughout the semester. Each week the project progress is reported and will be integrated into the final grade.
- f. three discussions or progress reports with the instructors about your project

The grade distribution is as follows

- 10% Comprehension Activities
- 10% Section
- 10% Chapter
- 70% Project

As the project is the main deliverable of the course it is obvious that those starting a week before the deadline will not succeed in this class. The project will take a significant amount of time and fosters the principal of “Learning by Doing” at all stages throughout the semester.

The class will not have a regular midterm, but it is expected that you have worked on your project and can provide a snapshot of the progress outlining the goals of the project and how you will achieve these goals till the end of the semester.

The final Project is due Dec. 1st. Issues with your project ought to have been discussed before this deadline with the TA’s The TAs will in the next 14 days go over the projects and evaluate major and minor issues that you may be able to fix without penalty. Larger changes will receive a grade penalty. The last fix (upon approval) possible will be Dec 7th.

2.3.5.1 Incomplete

Please see the university regulations for getting an incomplete. However, as this class uses state-of-the-art technology that changes frequently, you must expect that an incomplete may result in significant additional work on your behalf as your project may need significant updates on infrastructure, technology, or even programming models used. It is best to complete the course within one semester.

2.3.5.2 Calendar

🅾 Note that the calendar will be updated in the second week of the semester.

All dates assume submisison of the deliverable at 9am that day.

Assignment #	Event		Date
.	Full Term		16 Weeks
.	*Begins*		Mon 08/29
1, 2	**Bio, Notebook**	assigned	Mon 08/29
1, 2	**Bio, Notebook**	due	Mon 10/06 9am
3	**Section**	assigned	Mon 10/06
4	**Chapter**	assigned	Mon 10/06
5	**Project**	selection or proposal	Mon 09/18
.	*Labor Day*		Mon TBD
5	**Project**	update	TBD
.	*Fall Break*	TBD	
.	*Auto W*	TBD	
5	**Project**	update	TBD
3	**Section**	due	TBD
4	**Chapter**	due	TBD
.	*Thanksgiving*		TBD
5	**Project**	due	TBD
5	**Project** (no penalty)	improvements	TBD
5	**Project** (with penalty)	improvements	TBD
.	**Final Deliverables due**		TBD
.	*Grading*		TBD
.	*Ends*		TBD

- TA’s must be available till all grades have been submitted.
- Bio: a formal 3 paragraph Bio
- Notebook: a markdown in which you record your progress of this class in bullet form
- All times are in EST

- Dependent on class progress Comprehension Assignments may be added

2.4 Example Artifacts

Learning Objectives

- Identify what other students have done previously.
- Look at previous chapters, which are collected as technology reviews.
- Look at previous project reports.
- Looking at the documents provides you with an initial overview of the scope for the artifacts.

As part of this class you will be delivering some artifacts that are being graded. Some of them include writing a *chapter* that can be contributed to the class lecture and a project. To showcase you some example artifacts take a closer look at the documents listed in this section. Please also note that you can not duplicate or replicate a students previous work without significant improvements. All material listed here is available online, including all source code.

2.4.1 Technology Summaries

We are maintaining a large list of technologies related to clouds and Big Data at

- https://github.com/cloudmesh/technologies

This repository generates the following epub

- https://github.com/cloudmesh/technologies/blob/master/vonLaszewski-cloud-technologies.epub

Students that have to contribute as part of their class Technology summaries are asked to produce meaningful, advertisement free summaries of the technology and indicate in some cases if not obvious show they relate to cloud or big data. The length of such summaries is about 300 words. Students of E516 do not have to contribute to this and will instead focus on programming. Students of I423,

I523 and other sections must contribute to it and will get an assignment related to it. We post here the existence of this document also for 516 students. They can voluntarily improve or add sections if they like which will go into their discussion credit.

Please use the following indicators to mark the progress of summaries that you are working on.

 ready for review

 selected by student so others do not select it and we know what is worked on

 needs revision (only assigned by ta, after smiley)

The signs are put as follows. You can view an example at https://github.com/cloudmesh/technologies/blob/master/chapters/tech/bioconduct

Ex - Title Of Summary  fa18-xxx-xx

2.4.2 Chapters

Previously we asked students to write a 2 page paper on a topic related to bigdata analytics or cloud technologies (dependent on the course). Example papers are listed bellow

- Use Cases in Big Data Software and Analytics Vol. 1, Gregor von Laszewski, Fall 2017, http://cyberaide.org/papers/vonLaszewski-i523-v1.pdf
- Use Cases in Big Data Software and Analytics Vol. 2, Gregor von Laszewski, Fall 2017, http://cyberaide.org/papers/vonLaszewski-i523-v2.pdf
- Big Data Software Vol 1., Gregor von Laszewski, Spring 2017, https://github.com/cloudmesh/sp17-

i524/blob/master/paper1/proceedings.pdf

- Big Data Software Vol 2., Gregor von Laszewski, Spring 2017, https://github.com/cloudmesh/sp17-i524/blob/master/paper2/proceedings.pdf

- Vol 8, Gregor von Laszewski, Spring 2018, http://cyberaide.org/papers/vonLaszewski-cloud-vol-8.pdf

This has however resulted in a large number of duplicated material especially in the introductions and motivations. Thus we like this year to have you more focused on the topic and do not write a large introduction on what big data or cloud computing is. Therefore we renamed the 2 paper to a chapter, while you could assume certain things that have already been taught to you and you do not have to repeat it.

2.4.3 Project Reports

The goal of the class is to use open source technology to also write your technical reports. As a beneficial side product, we are able to distribute all previous reports from students to you. In your reports you will be doing a similar report, but will not use the same topic, without a significant improvement from a report already delivered in that area. For big data we have more than 1000 data sets we point to. I am sure you can do a unique project. For engineering cloud there are recently so many new technologies that there is not much chance of an overlap. TA's will review your project proposal, but it is your responsibility to make sure they are unique.

Please note that we do not make any quality assumptions to the published papers. It is up to you to identify outstanding papers.

- Use Cases in Big Data Software and Analytics Vol. 3, Gregor von Laszewski, Fall 2017, http://cyberaide.org/papers/vonLaszewski-i523-v3.pdf

- Big Data Projects, Gregor von Laszewski, Spring 2017, https://github.com/cloudmesh/sp17-i524/blob/master/project/projects.pdf

- Vol 9, Gregor von Laszewski, Spring 2018, http://cyberaide.org/papers/vonLaszewski-cloud-vol-9.pdf

2.5 Datasets

Given next are links to collections of datasets that may be of use for homework assignments or projects.

- https://www.data.gov/
- https://github.com/caesar0301/awesome-public-datasets
- https://aws.amazon.com/public-data-sets/
- https://www.kaggle.com/datasets
- https://cloud.google.com/bigquery/public-data/github
- https://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public

For NIST Projects:

- NIST Special Database 27A - 4GB
- INRIA Person Dataset
- Healthcare data from CMS
- Uber Ride Sharing GPS Data
- Census Data

2.6 Assignments

2.6.1 Due dates

For due dates see the calendar section.

2.6.2 Terminology

Dependent on the class you need to do different assignments. The assignments will be clearly posted. In case of questions, we will update this document to provide clarifications if needed. We use the following terminology:

License:

> All projects are developed under an open source license such as Apache 2.0 License. You will be required to add a LICENCE.txt file and if you use other software identify how it can be reused in your project. If your project uses different licenses, please add in a README.md file which packages are used and which license these packages have.

Sections:

> Sections are written in markdown and include information on a particular technical issue that is in general helpful for other students. Sections must be about a substantial topic and include an introduction a section that teaches a reader a significant issue, as well as practical code examples. Multiple small sections can lead to a substantial contribution. We expect that the sections are of high quality and can be included in our handbooks. Please be careful of plagiarism and do not just copy the sections from tutorials or code or from elsewhere.

Technology or Review Paper :

> A technology paper is a summary paper about a technology, application, or topic that is not yet covered in other technology papers delivered by previous students of this class. A review paper is a paper that reviews a specific topic related to this class.

In either case includes useful information that provides an overview of what you are trying to describe and analyses its relationship to the class topic. Be mindful about plagiarism. The paper is written in LaTeX or Markdown and uses bibtex for bibliography management. It uses the same format as your report paper. The format is discussed in the Section Report Format.

A technology paper is 2 pages long. This will make it between 2000-2400 words.

Note: that for the 2018 we decided to just us Markdown and not LaTeX. We will calculate the exact number of words needed.

Project:

> We refer with the term project to the major activity that you chose as part of your class. The default case is an implementation project that requires a *project report* and project code. In case you have issues with code development you can also chose a *term paper* as project.

Term Paper:

> A term paper is an enhanced topic paper (only available for I523). The difference is in length and depth of coverage. Comparative or review papers can also become term papers. In case you chose the term paper, you or your team will pick a topic relevant for the class. Term papers should have the quality to be publishable either in a workshop or as part of the handbook. Not all classes allow you to do a term paper, but require you to do a project. Please confirm with your class. For the classes listed here the term paper wil result in a quarter reduction in grade for the entire class not just the paper. Remember tables and figures do not count towards the paper length. A term paper has the following length.

- 8 pages, one student in the project
- 10 pages, two student in the project
- 12 pages, three student in the project

We estimate that a single page is between 1000-1200 words. Please note that for 2018 the format will be markdown, so the word count will be used instead.

Project Report:

> A project report is an enhanced topic paper that includes not just the analysis of a topic, but an actual code, with **benchmark** and demonstrated application use. Obviously it is longer than a term paper and includes descriptions about reproducibility of the application. A README is

provided that describes in a section how others can reproduce your project and run it. Term papers should have the quality to be publishable either in a workshop or as part of the handbook. The format is discussed in the Section Report Format. Remember tables and figures do not count towards the paper length. The following length is required:

- 6 pages, one student in the project
- 8 pages, two students in the project
- 10 pages, three students in the project

We estimate that a single page is between 1000-1200 words. Please note that for 2018 the format will be markdown, so the word count will be used instead.

Project Code:

This is the **documented** and **reproducible** code and scripts that allows a TA do replicate the project. In case you use images they must be created from scratch locally and may not be uploaded to services such as dockerhub. You can however reuse vendor uploaded images such as from ubuntu or centos. All code, scripts, and documentation must be uploaded to github.com under the class specific github directory.

Data:

Data is to be hosted on IUs google drive if needed. If you have larger data, it should be downloaded from the internet. It is in your responsibility to develop a download program. The data **must** not be stored in github. You will be expected to write a python program that downloads the data.

Work Breakdown:

This is an appendix to the document that describes in detail who did what in the project. This section comes in a new page after the references. It does not count towards the page length of the document. It also includes explicit URLs to the git history that documents the statistics to demonstrate not only one student has worked on the project. If you can not provide such a statistic or all check-ins have been made by a single student, the project has shown that they have not properly used git. Thus points will be deducted

from the project. Furthermore, if we detect that a student has not contributed to a project we may invite the student to give a detailed presentation of the project.

Bibliography:

All bibliography has to be provided in a jabref/bibtex file. This is regardless if you use LaTeX or Word. There is **NO EXCEPTION** to this rule. Please be advised doing references right takes some time so you want to do this early. Please note that exports of Endnote or other bibliography management tools do not lead to properly formatted bibtex files, despite they claiming to do so. You will have to clean them up and we recommend to do it the other way around. Manage your bibliography with jabref, and if you like to use it import them to endnote or other tools. Naturally you may have to do some cleanup to. If you use LaTeX and jabref, you have naturally much less work to do. What you chose is up to you.

2.6.2.1 Project Deliverables

The objective of the project is to define a clear problem statement and create a framework to address that problem as it relates to big data your project must address the reproducibility of the deployment and the application. A dataset must be chosen and you can analyze the data. YOu must make sure your project can be deployed on the TAs computer through scripts that make your project reproducible.

You have plenty of time to make this choice and if you find you struggle with programming you may want to consider a term paper instead of a project.

In case you chose a project your maximum grade for the entire class could be an A+. However, an A+ project must be truly outstanding and include an exceptional project report. Such a project and report will have the potential quality of being able to be published in a conference or workshop/

In case you chose a term paper your maximum grade for the *entire* class will be an A-.

2.6.2.1.0.1 Deliverables

- Find a data set with reasonable size (this may depend on your resources and needs to include a benchmark in your paper for justification).
- Clean up the data set or make it smaller or find a bigger data set
- Identify existing algorithms and tools and technologies that you can use to analyze your data
- Provide benchmarks.
- Take results in two different cloud services and your local PC (ex: Chameleon Cloud, echo kubernetes). Make sure your system can be created and deployed based on your documentation.
- Create a Makefile with the tags deploy, run, kill, view, clean that deploys your environment, runs application, kills it, views the result and cleans up after wards. You are allowed to have different makefiles for the different clouds and different directories. Keep the code and directory structure clean and document how to reproduce your results.
- For python use a requirements.txt file also
- For docker use a Dockerfile also
- Write a report that includes the following sections
    - Abstract
    - Introduction
    - Architecture
    - Implementation
        - Technologies Used
    - Design
    - Implementation
    - Results
        - Deployment Benchmarks
        - Application Benchmarks
    - (Limitations)
    - Conclusion
    - (Work Breakdown)

- Your paper will not have a future work section as this implies that you will do work in future, instead you can use an optional limitations section.

2.6.2.2 Group work

You are allowed to work on any assignment in class in groups up to 3 team members. We will not allow more team members as previous examples showed that more team members do not result in better projects than thise delivered with 3 team members.

The assignment is only to be added into github by one team member. Please make sure that you do pull requests to the repository of that team member. If your team likes direct access to the repo form the lead, please communicate this in a private post to piazza with the github user names and we will add the team members. The lea should be aware that in this case all team members have access to all files from the team leader, not only that assignment. If the team leader does not like this, the team should stick with pull requests that the team lead coordinates and integrates.

Groups are expected to have significantly better artifacts than a single student. It is not the goal of the group to deliver a project or paper that is done by n people but could have been done by a single person. Therefore the requirements of all group projects are increased accordingly. Typically this is not an issue. Make sure all team members contribute.

Group requirements Technology summaries:

> If you have n team members together you need to have at least 4 * n technologies. The technologies that have been assigned to each team member will have to be completed. YOu can work in collaboration on the technologies, but you need to place all hids that worked on the technology in the headline.

Group requirements 2-page Technology or Review paper:

> Option multiple papers. If you have n team members you need to write n different papers each of which has 2 pages. The n team members can be

authoring the n papers jointly. Put your hids and names in the paper

Option one large paper. If you have n team members you need to write one large paper with 2 * n pages. The paper must be well written and integrated and not just the concatenation of 2 pages from each author.

Group requirements project paper:

The requirements are clearly stated in another section of the ePub.

3 DETAILS

3.1 Introduction to Big Data Applications

This is an overview course of Big Data Applications covering a broad range of problems and solutions. It covers cloud computing technologies and includes a project. Also, algorithms are introduced and illustrated.

3.1.1 General Remarks Including Hype cycles

This is Part 1 of the introduction. We start with some general remarks and take a closer look at the emerging technology hype cycles.

1.a Gartner's Hypecycles and especially those for emerging technologies between 2016 and 2018

- .

1.b Gartner's Hypecycles with Emerging technologies hypecycles and the priority matrix at selected times 2008-2015

- .

1.a + 1.b:

-

- Technology trends
- Industry reports

3.1.2 Data Deluge

This is Part 2 of the introduction.

2.a Business usage patterns from NIST

- 

2.b Cyberinfrastructure and AI

- 

2.a + 2.b

- 

- Several examples of rapid data and information growth in different areas
- Value of data and analytics

3.1.3 Jobs

This is Part 3 of the introduction.

- 

- Jobs opportunities in the areas: data science, clouds and computer science and computer engineering
- Jobs demands in different countries and companies.
- Trends and forecast of jobs demands in the future.

3.1.4 Industry Trends

This is Part 4 of the introduction.

4a. Industry Trends: Technology Trends by 2014

- 

4b. Industry Trends: 2015 onwards

- 

An older set of trend slides is available from:

4a. Industry Trends: Technology Trends by 2014

- 

A current set is available at:

4b. Industry Trends: 2015 onwards

-  

4c. Industry Trends: Voice and HCI, cars,Deep learning

- 

- Many technology trends through end of 2014 and 2015 onwards, examples in different fields
- Voice and HCI, Cars Evolving and Deep learning

3.1.5 Digital Disruption and Transformation

This is Part 5 of the introduction.

5. Digital Disruption and Transformation

-   

- The past displaced by digital disruption

3.1.6 Computing Model

This is Part 6 of the introduction.

6a. Computing Model: earlier discussion by 2014:

- .

6b. Computing Model: developments after 2014 including Blockchain:

- .

- Industry adopted clouds which are attractive for data analytics, including big companies, examples are Google, Amazon, Microsoft and so on.
- Some examples of development: AWS quarterly revenue, critical capabilities public cloud infrastructure as a service.

- Blockchain: ledgers redone, blockchain consortia.

3.1.7 Research Model

This is Part 7 of the introduction.

Research Model: 4th Paradigm; From Theory to Data driven science?

- .

- The 4 paradigm of scientific research: Theory,Experiment and observation,Simulation of theory or model,Data-driven.

3.1.8 Data Science Pipeline

This is Part 8 of the introduction. 8. Data Science Pipeline

- .

- DIKW process:Data, Information, Knowledge, Wisdom and Decision.
- Example of Google Maps/navigation.

- Criteria for Data Science platform.

3.1.9 Physics as an Application Example

This is Part 9 of the introduction.

- 9. Physics as an Application Example
- Physics as an application example.

3.1.10 Technology Example

This is Part 10 of the introduction.

- 10. Technology Example: Recommender Systems I
- Overview of many informatics areas, recommender systems in detail.
- NETFLIX on personalization, recommendation, datascience.

3.1.11 Exploring Data Bags and Spaces

This is Part 11 of the introduction.

11. Exploring data bags and spaces: Recommender Systems II

- .
- Distances in funny spaces, about “real” spaces and how to use distances.

3.1.12 Another Example: Web Search Information Retrieval

This is Part 12 of the introduction. 12. Another Example: Web Search Information Retrieval

- .

3.1.13 Cloud Application in Research

This is Part 13 of the introduction discussing cloud applications in research.

13. Cloud Applications in Research: Science Clouds and Internet of Things

- 

3.1.14 Software Ecosystems: Parallel Computing and MapReduce

This is Part 14 of the introduction discussing the software ecosystem

14. Software Ecosystems: Parallel Computing and MapReduce

- 

3.1.15 Conclusions

This is Part 15 of the introduction with some concluding remarks. 15. Conclusions

- 

3.2 Overview of Data Science

What is Big Data, Data Analytics and X-Informatics?

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

In the next unit, we continue 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. Two broad classes of data are the long tail of sciences: many users with individually modest data adding up to a lot; and a myriad of Internet connected devices – the Internet of Things.

We give 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. Features of the data deluge are discussed with a salutary example where more data did better than more thought. Then comes 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.

3.2.1 Data Science generics and Commercial Data Deluge

We start 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.

 Commercial Data Deluge (45)

3.2.1.1 What is X-Informatics and its Motto

This discusses trends that are driven by and accompany Big data. We give some key terms including data, information, knowledge, wisdom, data analytics and data science. We discuss how clouds running Data Analytics Collaboratively processing Big Data can solve problems in X-Informatics. We list many values of X you can defined in various activities across the world.

-  X Informatics (9:49)

3.2.1.2 Jobs

Big data is especially important as there are some many related jobs. We illustrate this for both cloud computing and data science from reports by Microsoft and the McKinsey institute respectively. We show a plot from LinkedIn showing rapid increase in the number of data science and analytics jobs as a function of time.

-  Jobs (2:58)

3.2.1.3 Data Deluge: General Structure

We look at some broad features of the data deluge starting with the size of data in various areas especially in science research. We give examples from real world of the importance of big data and illustrate how it is integrated into an enterprise IT architecture. We give some views as to what characterizes Big data and why data science is a science that is needed to interpret all the data.

-  Data Deluge (13:04)

3.2.1.4 Data Science: Process

We stress the DIKW pipeline: Data becomes information that becomes knowledge and then wisdom, policy and decisions. This pipeline is illustrated with Google maps and we show how complex the ecosystem of data, transformations (filters) and its derived forms is.

-  Data Science Process (4:27)

3.2.1.5 Data Deluge: Internet

We give examples of Big data from the Internet with Tweets, uploaded photos

and an illustration of the vitality and size of many commodity applications.

- Internet (3:42)

3.2.1.6 Data Deluge: Business

We give examples including the Big data that enables wind farms, city transportation, telephone operations, machines with health monitors, the banking, manufacturing and retail industries both online and offline in shopping malls. We give examples from ebay showing how analytics allowing them to refine and improve the customer experiences.

- Business I (6:00)
- Business II (7:34)
- Business III (9:37)

3.2.1.7 Resources

- http://www.microsoft.com/en-us/news/features/2012/mar12/03-05CloudComputingJobs.aspx
- http://www.mckinsey.com/mgi/publications/big_data/index.asp
- Tom Davenport
- Anjul Bhambhri
- Jeff Hammerbacher
- http://www.economist.com/node/15579717
- http://cs.metrostate.edu/~sbd/slides/Sun.pdf
- http://jess3.com/geosocial-universe-2/
- Bill Ruh
- http://www.hsph.harvard.edu/ncb2011/files/ncb2011-z03-rodriguez.pptx
- Hugh Williams

3.2.2 Data Deluge and Scientific Applications and Methodology

3.2.2.1 Overview of Data Science

We continue 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 being used to create a wealth of data.

 Methodology (22)

3.2.2.2 Science and Research

We look into more big data examples with a focus on science and research. We give astronomy, genomics, radiology, particle physics and discovery of Higgs particle (Covered in more detail in later lessons), European Bioinformatics Institute and contrast to Facebook and Walmart.

-  Science and Research (11:27)
-  Science and Research (11:49)

3.2.2.3 Implications for Scientific Method

We discuss the emergencies of a new fourth methodology for scientific research based on data driven inquiry. We contrast this with third ~~ computation or simulation based discovery - methodology which emerged itself some 25 years ago.

-  Scientific Methods (5:07)

3.2.2.4 Long Tail of Science

There is big science such as particle physics where a single experiment has 3000 people collaborate!.Then there are individual investigators who do not generate a lot of data each but together they add up to Big data.

-  Long Tail of Science (2:10)

3.2.2.5 Internet of Things

A final category of Big data comes from the Internet of Things where lots of small devices ~~ smart phones, web cams, video games collect and disseminate data and are controlled and coordinated in the cloud.

-  Internet of Things (5:45)

3.2.2.6 Resources

- http://www.economist.com/node/15579717
- Geoffrey Fox and Dennis Gannon Using Clouds for Technical Computing To be published in Proceedings of HPC 2012 Conference at Cetraro, Italy June 28 2012
- http://grids.ucs.indiana.edu/ptliupages/publications/Clouds_Technical_Com
- http://grids.ucs.indiana.edu/ptliupages/publications/Where%20does%20all%
- http://www.genome.gov/sequencingcosts/
- http://www.quantumdiaries.org/2012/09/07/why-particle-detectors-need-a-trigger/atlasmgg
- http://salsahpc.indiana.edu/dlib/articles/00001935/
- http://en.wikipedia.org/wiki/Simple_linear_regression
- http://www.ebi.ac.uk/Information/Brochures/
- http://www.wired.com/wired/issue/16-07
- http://research.microsoft.com/en-us/collaboration/fourthparadigm/
- CSTI General Assembly 2012, Washington, D.C., USA Technical Activities Coordinating Committee (TACC) Meeting, Data Management, Cloud Computing and the Long Tail of Science October 2012 Dennis Gannon

3.2.3 Clouds and Big Data Processing; Data Science Process and Analytics

3.2.3.1 Overview of Data Science

We give 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. 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.

 Clouds (35)

3.2.3.2 Clouds

We describe cloud data centers with their staggering size with up to a million servers in a single data center and centers built modularly from shipping containers full of racks. 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 and a comparison to supercomputing.

-  Clouds (16:04){MP4}

3.2.3.3 Aspect of Data Deluge

Data, Information, intelligence algorithms, infrastructure, data structure, semantics and knowledge are related. The semantic web and Big data are compared. We give an example where “More data usually beats better algorithms”. We discuss examples of intelligent big data and list 8 different types of data deluge

-  Data Deluge (8:02)
-  Data Deluge (6:24)

3.2.3.4 Data Science Process

We describe and critique one view of the work of a data scientists. Then we discuss and contrast 7 views of the process needed to speed data through the DIKW pipeline.

- Scientific Process (11:28)

3.2.3.5 Data Analytics

Data Analytics (30) We stress the importance of data analytics givi ng examples from several fields. We note that better analytics is as important as better computing and storage capability. In the second video we look at High Performance Computing in Science and Engineering: the Tree and the Fruit.

- Data Analytics (7:28)
- Data Analytics (6:51)

3.2.3.6 Resources

- CSTI General Assembly 2012, Washington, D.C., USA Technical Activities Coordinating Committee (TACC) Meeting, Data Management, Cloud Computing and the Long Tail of Science October 2012 Dennis Gannon
- Dan Reed Roger Barga Dennis Gannon Rich Wolski http://research.microsoft.com/en-us/people/barga/sc09_cloudcomp_tutorial.pdf
- http://www.datacenterknowledge.com/archives/2011/05/10/uptime-institute-the-average-pue-is-1-8/
- http://loosebolts.wordpress.com/2008/12/02/our-vision-for-generation-4-modular-data-centers-one-way-of-getting-it-just-right/
- http://www.mediafire.com/file/zzqna34282frr2f/koomeydatacenterelectuse2
- Bina Ramamurthy
- Jeff Hammerbacher
- Jeff Hammerbacher

- Anjul Bhambhri
- http://cs.metrostate.edu/~sbd/slides/Sun.pdf
- Hugh Williams
- Tom Davenport
- http://www.mckinsey.com/mgi/publications/big_data/index.asp
- http://cra.org/ccc/docs/nitrdsymposium/pdfs/keyes.pdf

3.3 PHYSICS

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.

3.3.1 Looking for Higgs Particles

3.3.1.1 Bumps in Histograms, Experiments and Accelerators

This unit is devoted to Python and Java experiments looking at histograms of Higgs Boson production with various forms of shape of signal and various background and with various event totals. The lectures use Python but use of Java is described.

-  Higgs (20)
- <{gitcode}/physics/mr-higgs/higgs-classI-sloping.py>

3.3.1.2 Particle Counting

We return to particle case with slides used in introduction and stress that particles often manifested as bumps in histograms and those bumps need to be large enough to stand out from background in a statistically significant fashion.

-  Discovery of Higgs Particle (13:49)

We give a few details on one LHC experiment ATLAS. Experimental physics papers have a staggering number of authors and quite big budgets. Feynman diagrams describe processes in a fundamental fashion.

-  Looking for Higgs Particle and Counting Introduction II (7:38)

3.3.1.3 Experimental Facilities

We give a few details on one LHC experiment ATLAS. Experimental physics papers have a staggering number of authors and quite big budgets. Feynman diagrams describe processes in a fundamental fashion.

-  Looking for Higgs Particle Experiments (9:29)

3.3.1.4 Accelerator Picture Gallery of Big Science

This lesson gives a small picture gallery of accelerators. Accelerators, detection chambers and magnets in tunnels and a large underground laboratory used fpr experiments where you need to be shielded from background like cosmic rays.

-  Accelerator Picture Gallery of Big Science (11:21)

3.3.1.5 Resources

- http://grids.ucs.indiana.edu/ptliupages/publications/Where%20does%20all% [2]
- http://www.sciencedirect.com/science/article/pii/S037026931200857X [3]

- http://www.nature.com/news/specials/lhc/interactive.html

Looking for Higgs Particles: Python Event Counting for Signal and Background (Part 2)

This 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.

-  Higgs II (29)

Files:

- <{gitcode}/physics/mr-higgs/higgs-classI-sloping.py>
- <{gitcode}/physics/number-theory/higgs-classIII.py>
- <{gitcode}/physics/mr-higgs/higgs-classII-uniform.py>

3.3.1.6 Event Counting

We define *event counting* data collection environments. We discuss the python and Java code to generate events according to a particular scenario (the important idea of Monte Carlo data). Here a sloping background plus either a Higgs particle generated similarly to LHC observation or one observed with better resolution (smaller measurement error).

-  Event Counting (7:02)

3.3.1.7 Monte Carlo

This uses Monte Carlo data both to generate data like the experimental observations and explore effect of changing amount of data and changing measurement resolution for Higgs.

-  With Python examples of Signal plus Background (7:33) This lesson continues the examination of Monte Carlo data looking at effect of change in number of Higgs particles produced and in change in shape of

background.

-  Change shape of background & num of Higgs Particles (7:01)

3.3.1.8 Resources

- Python for Data Analysis: Agile Tools for Real World Data By Wes McKinney, Publisher: O'Reilly Media, Released: October 2012, Pages: 472. [4]
- http://jwork.org/scavis/api/ [5]
- https://en.wikipedia.org/wiki/DataMelt **???**

3.3.1.9 Random Variables, Physics and Normal Distributions

We introduce 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. Java is currently not available in this unit.

-  Higgs (39)
- <{gitcode}/physics/number-theory/higgs-classIII.py>

3.3.1.10 Statistics Overview and Fundamental Idea: Random Variables

We go through the many different areas of statistics covered in the Physics unit. We define the statistics concept of a random variable.

-  Random variables and normal distributions (8:19)

3.3.1.11 Physics and Random Variables

We describe the DIKW pipeline for the analysis of this type of physics experiment and go through details of analysis pipeline for the LHC ATLAS experiment. We give examples of event displays showing the final state particles

seen in a few events. We illustrate how physicists decide whats going on with a plot of expected Higgs production experimental cross sections (probabilities) for signal and background.

- Physics and Random Variables I (8:34)
- Physics and Random Variables II (5:50)

3.3.1.12 Statistics of Events with Normal Distributions

We introduce Poisson and Binomial distributions and define independent identically distributed (IID) random variables. We give the law of large numbers defining the errors in counting and leading to Gaussian distributions for many things. We demonstrate this in Python experiments.

- Statistics of Events with Normal Distributions (11:25)

3.3.1.13 Gaussian Distributions

We introduce the Gaussian distribution and give Python examples of the fluctuations in counting Gaussian distributions.

- Gaussian Distributions (9:08)

3.3.1.14 Using Statistics

We discuss the significance of a standard deviation and role of biases and insufficient statistics with a Python example in getting incorrect answers.

- Using Statistics (14:02)

3.3.1.15 Resources

- http://indico.cern.ch/event/20453/session/6/contribution/15?materialId=slides

- ~~http://www.atlas.ch/photos/events.html~~ (this link is outdated)
- https://cms.cern/ [6]

3.3.1.16 Random Numbers, Distributions and Central Limit Theorem

We discuss 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 and Bayes law concludes discussion. Python and Java (for student - not reviewed in class) examples and Physics applications are given.

- Higgs III (44)

Files:

- <{gitcode}/physics/calculated-dice-roll/higgs-classIV-seeds.py>

3.3.1.16.1 Generators and Seeds

We define random numbers and describe how to generate them on the computer giving Python examples. We define the seed used to define to specify how to start generation.

- Higgs Particle Counting Errors (6:28)
- Generators and Seeds II (7:10)

3.3.1.16.2 Binomial Distribution

We define binomial distribution and give LHC data as an example of where this distribution valid.

- Binomial Distribution: (12:38)

3.3.1.16.3 Accept-Reject

We introduce an advanced method **accept/reject** for generating random

variables with arbitrary distributions.

- Accept-Reject (5:54)

3.3.1.16.4 Monte Carlo Method

We define Monte Carlo method which usually uses accept/reject method in typical case for distribution.

- Monte Carlo Method (2:23)

3.3.1.16.5 Poisson Distribution

We extend the Binomial to the Poisson distribution and give a set of amusing examples from Wikipedia.

- Poisson Distribution (4:37)

3.3.1.16.6 Central Limit Theorem

We introduce Central Limit Theorem and give examples from Wikipedia.

- Central Limit Theorem (4:47)

3.3.1.16.7 Interpretation of Probability: Bayes v. Frequency

This lesson describes difference between Bayes and frequency views of probability. Bayes's law of conditional probability is derived and applied to Higgs example to enable information about Higgs from multiple channels and multiple experiments to be accumulated.

- Interpretation of Probability (12:39)

3.3.1.16.8 Resources

3.3.2 SKA – Square Kilometer Array

Professor Diamond, accompanied by Dr. Rosie Bolton from the SKA Regional Centre Project gave a presentation at SC17 “into the deepest reaches of the observable universe as they describe the SKA’s international partnership that will map and study the entire sky in greater detail than ever before.”

- http://sc17.supercomputing.org/presentation/?id=inspkr101&sess=sess263

A summary article about this effort is available at:

- https://www.hpcwire.com/2017/11/17/sc17-keynote-hpc-powers-ska-efforts-peer-deep-cosmos/ The video is hosted at
- http://sc17.supercomputing.org/presentation/?id=inspkr101&sess=sess263 Start at about 1:03:00 (e.g. the one hour mark)

3.4 e-Commerce and LifeStyle

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. It is interesting that 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.

3.4.1 Recommender Systems

We introduce 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. It is interesting how the humble ranking has become such a dominant driver of the world's economy.

 Lifestyle Recommender (45)

3.4.1.1 Recommender Systems as an Optimization Problem

We define a set of general recommender systems as matching of items to people or perhaps collections of items to collections of people where items can be other people, products in a store, movies, jobs, events, web pages etc. We present this as "yet another optimization problem".

 Recommender Systems I (8:06)

3.4.1.2 Recommender Systems Introduction

We give a general discussion of recommender systems and point out that they are particularly valuable in long tail of tems (to be recommended) that are not commonly known. We pose them as a rating system and relate them to information retrieval rating systems. We can contrast recommender systems based on user profile and context; the most familiar collaborative filtering of others ranking; item properties; knowledge and hybrid cases mixing some or all of these.

 Recommender Systems Introduction (12:56)

3.4.1.3 Kaggle Competitions

We look at Kaggle competitions with examples from web site. In particular we discuss an Irvine class project involving ranking jokes.

 Kaggle Competitions: (3:36)

> ⚠ *Please not that we typically do not accept any projects using kaggle data for this classes. This class is not about winning a kaggle competition and if done wrong it does not fullfill the minimum requiremnt for this class. Please consult with the instructor.*

3.4.1.4 Examples of Recommender Systems

We go through a list of 9 recommender systems from the same Irvine class.

 Examples of Recommender Systems (1:00)

3.4.1.5 Netflix on Recommender Systems

We summarize some interesting points from a tutorial from Netflix for whom *everything is a recommendation*. Rankings are given in multiple categories and categories that reflect user interests are especially important. Criteria used include explicit user preferences, implicit based on ratings and hybrid methods as well as freshness and diversity. Netflix tries to explain the rationale of its recommendations. We give some data on Netflix operations and some methods used in its recommender systems. We describe the famous Netflix Kaggle competition to improve its rating system. The analogy to maximizing click through rate is given and the objectives of optimization are given.

 Netflix on Recommender Systems (14:20)

Next we go through Netflix's methodology in letting data speak for itself in optimizing the recommender engine. An example iis given on choosing self produced movies. A/B testing is discussed with examples showing how testing does allow optimizing of sophisticated criteria. This lesson is concluded by comments on Netflix technology and the full spectrum of issues that are involved including user interface, data, AB testing, systems and architectures. We comment on optimizing for a household rather than optimizing for individuals in household.

 Consumer Data Science (13:04)

3.4.1.6 Other Examples of Recommender Systems

We continue 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.

 Lifestyle Recommender (49)

We start with a quick recap of recommender systems from previous unit; what they are with brief examples.

 Recap and Examples of Recommender Systems (5:48)

3.4.1.6.1 Examples of Recommender Systems

We give 2 examples in more detail: namely Google News and Markdown in Retail.

 Examples of Recommender Systems (8:34)

3.4.1.6.2 Recommender Systems in Yahoo Use Case Example

We describe in greatest detail the methods used to optimize Yahoo web sites. There are two lessons discussing general approach and a third lesson examines a particular personalized Yahoo page with its different components. We point out the different criteria that must be blended in making decisions; these criteria include analysis of what user does after a particular page is clicked; is the user satisfied and cannot that we quantified by purchase decisions etc. We need to choose Articles, ads, modules, movies, users, updates, etc to optimize metrics such as relevance score, CTR, revenue, engagement.These lesson stress that if though we have big data, the recommender data is sparse. We discuss the approach that involves both batch (offline) and on-line (real time) components.

 Recap of Recommender Systems II (8:46)

 Recap of Recommender Systems III (10:48)

 Case Study of Recommender systems (3:21)

3.4.1.6.3 User-based nearest-neighbor collaborative filtering

Collaborative filtering is a core approach to recommender systems. There is user-based and item-based collaborative filtering and here we discuss the user-based case. Here similarities in user rankings allow one to predict their interests, and typically this quantified by the Pearson correlation, used to statistically quantify correlations between users.

 User-based nearest-neighbor collaborative filtering I (7:20)

 User-based nearest-neighbor collaborative filtering II (7:29)

3.4.1.6.4 Vector Space Formulation of Recommender Systems

We go through recommender systems thinking of them as formulated in a funny vector space. This suggests using clustering to make recommendations.

 Vector Space Formulation of Recommender Systems new (9:06)

3.4.1.7 Resources

- http://pages.cs.wisc.edu/~beechung/icml11-tutorial/

3.4.2 Item-based Collaborative Filtering and its Technologies

We move 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.

 Lifestyle Filtering (18)

3.4.2.1 Item-based Collaborative Filtering

We covered user-based collaborative filtering in the previous unit. Here we start by discussing memory-based real time and model based offline (batch)

approaches. Now we look at item-based collaborative filtering where items are viewed in the space of users and the cosine measure is used to quantify distances. WE discuss optimizations and how batch processing can help. We discuss different Likert ranking scales and issues with new items that do not have a significant number of rankings.

 Item Based Filtering (11:18)

 k Nearest Neighbors and High Dimensional Spaces (7:16)

3.4.2.2 k-Nearest Neighbors and High Dimensional Spaces

We define the k Nearest Neighbor algorithms and present the Python software but do not use it. We give examples from Wikipedia and describe performance issues. This algorithm illustrates the curse of dimensionality. If items were a real vectors in a low dimension space, there would be faster solution methods.

 k Nearest Neighbors and High Dimensional Spaces (10:03)

3.4.2.2.1 Recommender Systems - K-Neighbors

Next we provide some sample Python code for the k Nearest Neighbor 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. This lesson is adapted from the Python k Nearest Neighbor code found on the web associated with a book by Harrington on Machine Learning [??]. There are two data sets. First we consider a set of 4 2D vectors divided into two categories (clusters) and use k=3 Nearest Neighbor algorithm to classify 3 test points. Second we consider a 3D dataset that has already been classified and show how to normalize. In this lesson we just use Matplotlib to give 2D plots.

The lesson goes through an example of using k NN classification algorithm by dividing dataset into 2 subsets. One is training set with initial classification; the other is test point to be classified by k=3 NN using training set. The code records fraction of points with a different classification from that input. One can

experiment with different sizes of the two subsets. The Python implementation of algorithm is analyzed in detail.

3.4.2.2.2 Plotviz

The 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.

3.4.2.2.3 Files

- https://github.com/cloudmesh-community/book/blob/master/examples/python/knn/kNN.py
- https://github.com/cloudmesh-community/book/blob/master/examples/python/knn/kNN_Driver.py
- https://github.com/cloudmesh-community/book/blob/master/examples/python/knn/dating_test_set2.txt
- https://github.com/cloudmesh-community/book/blob/master/examples/python/knn/clusterFinal-M3-C3Dating-ReClustered.pviz
- https://github.com/cloudmesh-community/book/blob/master/examples/python/knn/dating_rating_original_
- https://github.com/cloudmesh-community/book/blob/master/examples/python/knn/clusterFinal-M30-C28.pviz
- https://github.com/cloudmesh-community/book/blob/master/examples/python/plotviz/clusterfinal_m3_c3d
- https://github.com/cloudmesh-community/book/blob/master/examples/python/plotviz/fungi_lsu_3_15_to_3

3.4.2.3 Resources k-means

- http://www.slideshare.net/xamat/building-largescale-realworld-recommender-systems-recsys2012-tutorial [7]
- http://www.ifi.uzh.ch/ce/teaching/spring2012/16-Recommender-Systems_Slides.pdf [8]

- https://www.kaggle.com/ [9]
- http://www.ics.uci.edu/~welling/teaching/CS77Bwinter12/CS77B_w12.htm [10]
- Jeff Hammerbacher[11]
- http://www.techworld.com/news/apps/netflix-foretells-house-of-cards-success-with-cassandra-big-data-engine-3437514/ [12]
- https://en.wikipedia.org/wiki/A/B_testing [13]
- http://www.infoq.com/presentations/Netflix-Architecture [14]

3.5 Sports

Sports sees significant growth in analytics with pervasive statistics shifting to more sophisticated measures. We start with baseball as game is built around segments dominated by individuals where detailed (video/image) achievement measures including PITCHf/x and FIELDf/x are moving field into big data arena. There are interesting relationships between the economics of sports and big data analytics. We look at Wearables and consumer sports/recreation. The importance of spatial visualization is discussed. We look at other Sports: Soccer, Olympics, NFL Football, Basketball, Tennis and Horse Racing.

3.5.1 Basic Sabermetrics

This unit discusses baseball starting with the movie Moneyball and the 2002-2003 Oakland Athletics. Unlike sports like basketball and soccer, most baseball action is built around individuals often interacting in pairs. This is much easier to quantify than many player phenomena in other sports. We discuss Performance-Dollar relationship including new stadiums and media/advertising. We look at classic baseball averages and sophisticated measures like Wins Above Replacement.

Overview (40)

3.5.1.1 Introduction and Sabermetrics (Baseball Informatics) Lesson

Introduction to all Sports Informatics, Moneyball The 2002-2003 Oakland Athletics, Diamond Dollars economic model of baseball, Performance - Dollar

relationship, Value of a Win.

 Introduction and Sabermetrics (Baseball Informatics) Lesson (31:4)

3.5.1.2 Basic Sabermetrics

Different Types of Baseball Data, Sabermetrics, Overview of all data, Details of some statistics based on basic data, OPS, wOBA, ERA, ERC, FIP, UZR.

 Basic Sabermetrics (26:53)

3.5.1.3 Wins Above Replacement

Wins above Replacement WAR, Discussion of Calculation, Examples, Comparisons of different methods, Coefficient of Determination, Another, Sabermetrics Example, Summary of Sabermetrics.

 Wins Above Replacement (30:43)

3.5.2 Advanced Sabermetrics

This unit discusses ‘advanced sabermetrics’ covering advances possible from using video from PITCHf/X, FIELDf/X, HITf/X, COMMANDf/X and MLBAM.

 Sporta II (41)

3.5.2.1 Pitching Clustering

A Big Data Pitcher Clustering method introduced by Vince Gennaro, Data from Blog and video at 2013 SABR conference.

 Pitching Clustering (20:59)

3.5.2.2 Pitcher Quality

Results of optimizing match ups, Data from video at 2013 SABR conference.

Pitcher Quality (10:02)

3.5.3 PITCHf/X

Examples of use of PITCHf/X.

PITCHf/X (10:39)

3.5.3.1 Other Video Data Gathering in Baseball

FIELDf/X, MLBAM, HITf/X, COMMANDf/X.

Other Video Data Gathering in Baseball (18:5) Other Sports

We look at Wearables and consumer sports/recreation. The importance of spatial visualization is discussed. We look at other Sports: Soccer, Olympics, NFL Football, Basketball, Tennis and Horse Racing.

Sport Sports III (44)

3.5.3.2 Wearables

Consumer Sports, Stake Holders, and Multiple Factors.

Wearables (22:2)

3.5.3.3 Soccer and the Olympics

Soccer, Tracking Players and Balls, Olympics.

 Soccer and the Olympics (8:28)

3.5.3.4 Spatial Visualization in NFL and NBA

NFL, NBA, and Spatial Visualization.

 Spatial Visualization in NFL and NBA (15:19)

3.5.3.5 Tennis and Horse Racing

Tennis, Horse Racing, and Continued Emphasis on Spatial Visualization.

 Tennis and Horse Racing (8:52)

3.5.3.6 Resources

- http://www.slideshare.net/Tricon_Infotech/big-data-for-big-sports [15]
- http://www.slideshare.net/BrandEmotivity/sports-analytics-innovation-summit-data-powered-storytelling **???**
- http://www.slideshare.net/elew/sport-analytics-innovation [16]
- http://www.wired.com/2013/02/catapault-smartball/ [17]
- http://www.sloansportsconference.com/wp-content/uploads/2014/06/Automated_Playbook_Generation.pdf [18]
- http://autoscout.adsc.illinois.edu/publications/football-trajectory-dataset/ [19]
- http://www.sloansportsconference.com/wp-content/uploads/2012/02/Goldsberry_Sloan_Submission.pdf [20]
- http://gamesetmap.com/ [21]
- http://www.slideshare.net/BrandEmotivity/sports-analytics-innovation-summit-data-powered-storytelling [22]
- http://www.sloansportsconference.com/ [23]
- http://sabr.org/ [24]
- http://en.wikipedia.org/wiki/Sabermetrics [25]
- http://en.wikipedia.org/wiki/Baseball_statistics [26]
- http://m.mlb.com/news/article/68514514/mlbam-introduces-new-way-to-

analyze-every-play [27]
- http://www.fangraphs.com/library/offense/offensive-statistics-list/ [28]
- http://en.wikipedia.org/wiki/Component_ERA [29]
- http://www.fangraphs.com/library/pitching/fip/ **???**
- http://en.wikipedia.org/wiki/Wins_Above_Replacement [30]
- http://www.fangraphs.com/library/misc/war/ [31]
- http://www.baseball-reference.com/about/war_explained.shtml [32]
- http://www.baseball-reference.com/about/war_explained_comparison.shtml [33]
- http://www.baseball-reference.com/about/war_explained_position.shtml [34]
- http://www.baseball-reference.com/about/war_explained_pitch.shtml [35]
- http://www.fangraphs.com/leaders.aspx?pos=all&stats=bat&lg=all&qual=y&type=8&season=2014&month=0&seas [36]
- http://battingleadoff.com/2014/01/08/comparing-the-three-war-measures-part-ii/ [37]
- http://en.wikipedia.org/wiki/Coefficient_of_determination [38]
- http://www.sloansportsconference.com/wp-content/uploads/2014/02/2014_SSAC_Data-driven-Method-for-In-game-Decision-Making.pdf [39]
- https://courses.edx.org/courses/BUx/SABR101x/2T2014/courseware/10e61
- http://vincegennaro.mlblogs.com/ [40]
- https://www.youtube.com/watch?v=H-kx-x_d0Mk **???**
- http://www.baseballprospectus.com/article.php?articleid=13109 [41]
- http://baseball.physics.illinois.edu/FastPFXGuide.pdf [42]
- http://baseball.physics.illinois.edu/FieldFX-TDR-GregR.pdf [43]
- http://regressing.deadspin.com/mlb-announces-revolutionary-new-fielding-tracking-syste-1534200504 [44]
- http://grantland.com/the-triangle/mlb-advanced-media-play-tracking-bob-bowman-interview/ [45]

- https://www.youtube.com/watch?v=YkjtnuNmK74 [46]

These resources do not exsit:
- ~~http://www.sportvision.com/baseball~~
- ~~http://www.sportvision.com/media/pitchfx-how-it-works~~

- http://www.sportvision.com/baseball/fieldfx
- http://www.sportvision.com/baseball/hitfx
- http://www.trakus.com/technology.asp#tNetText
- http://www.sloansportsconference.com/?page_id=481&sort_cate=Research%20Paper

- http://www.liveathos.com/apparel/app

3.6 Cloud Computing

No

We describe 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. 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. We summarize the 21 layers and almost 300 software packages in the HPC-ABDS Software Stack explaining how they are used.

Cloud (Data Center) Architectures with physical setup, Green Computing issues and software models are discussed followed by the Cloud Industry stakeholders with a 2014 Gartner analysis of Cloud computing providers. This is followed by applications on the cloud including data intensive problems, comparison with high performance computing, science clouds and the Internet of Things. Remarks on Security, Fault Tolerance and Synchronicity issues in cloud follow.

We describe the way users and data interact with a cloud system. The Big Data Processing from an application perspective with commercial examples including eBay concludes section after a discussion of data system architectures.

3.6.1 Parallel Computing (Outdated)

We describe 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.

 Parallel Computing (33)

3.6.1.1 Decomposition

We describe why parallel computing is essential with Big Data and distinguishes parallelism over users to that over the data in problem. The general ideas behind data decomposition are given followed by a few often whimsical examples dreamed up 30 years ago in the early heady days of parallel computing. These include scientific simulations, defense outside missile attack and computer chess. The basic problem of parallel computing – efficient coordination of separate tasks processing different data parts – is described with MPI and MapReduce as two approaches. The challenges of data decomposition in irregular problems is noted.

-  Decomposition (8:51)
-  Examples of Application (13:22)
-  Decomposition Strategies (9:22)

3.6.1.2 Parallel Computing in Society

This lesson from the past notes that one can view society as an approach to parallel linkage of people. The largest example given is that of the construction of a long wall such as that (Hadrian's wall) between England and Scotland.

Different approaches to parallelism are given with formulae for the speed up and efficiency. The concepts of grain size (size of problem tackled by an individual processor) and coordination overhead are exemplified. This example also illustrates Amdahl's law and the relation between data and processor topology. The lesson concludes with other examples from nature including collections of neurons (the brain) and ants.

- Parallel Computing in Society I (8:24)
- Parallel Computing in Society II (8:01)

3.6.1.3 Parallel Processing for Hadrian's Wall

This lesson returns to Hadrian's wall and uses it to illustrate advanced issues in parallel computing. First We describe the basic SPMD – Single Program Multiple Data – model. Then irregular but homogeneous and heterogeneous problems are discussed. Static and dynamic load balancing is needed. Inner parallelism (as in vector instruction or the multiple fingers of masons) and outer parallelism (typical data parallelism) are demonstrated. Parallel I/O for Hadrian's wall is followed by a slide summarizing this quaint comparison between Big data parallelism and the construction of a large wall.

- Processing for Hadrian's Wall (9:24)

3.6.1.4 Resources

- Solving Problems in Concurrent Processors-Volume 1, with M. Johnson, G. Lyzenga, S. Otto, J. Salmon, D. Walker, Prentice Hall, March 1988.
- Parallel Computing Works!, with P. Messina, R. Williams, Morgan Kaufman (1994).
- The Sourcebook of Parallel Computing book edited by Jack Dongarra, Ian Foster, Geoffrey Fox, William Gropp, Ken Kennedy, Linda Torczon, and Andy White, Morgan Kaufmann, November 2002.
- Geoffrey Fox Computational Sciences and Parallelism to appear in

Encyclopedia on Parallel Computing edited by David Padua and published by Springer.

3.6.2 Introduction

We discuss 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. The unit concludes with two simple examples of the value of clouds for enterprise applications. Gartner also has specific predictions for cloud computing growth areas.

 Introduction (45)

3.6.2.1 Cyberinfrastructure for E-Applications

This introduction describes Cyberinfrastructure or e-infrastructure and its role in solving the electronic implementation of any problem where e-moreorlessanything is another term for moreorlessanything-Informatics and generalizes early discussion of e-Science and e-Business.

-  Cloud Computing Introduction Part1 (13:34)

3.6.2.2 What is Cloud Computing: Introduction

Cloud Computing is introduced with an operational definition involving virtualization and efficient large data centers that can rent computers in an elastic fashion. The role of services is essential – it underlies capabilities being offered in the cloud. The four basic aaS’s – Software (SaaS), Platform (Paas), Infrastructure (IaaS) and Network (NaaS) – are introduced with Research aaS and other capabilities (for example Sensors aaS are discussed later) being built on top of these.

-  What is Cloud Computing Intro (12:01)

3.6.2.3 What and Why is Cloud Computing: Other Views I

This lesson contains 5 slides with diverse comments on ‘’what is cloud computing’’ from the web.

- Other Views I (5:25)
- Other Views II (6:41)
- Other Views III (7:27)

3.6.2.4 Gartner’s Emerging Technology Landscape for Clouds and Big Data

This lesson gives Gartner’s projections around futures of cloud and Big data. We start with a review of hype charts and then go into detailed Gartner analyses of the Cloud and Big data areas. Big data itself is at the top of the hype and by definition predictions of doom are emerging. Before too much excitement sets in, note that spinach is above clouds and Big data in Google trends.

- Gartners Emerging Technology Landscape (11:26)

3.6.2.5 Simple Examples of use of Cloud Computing

This short lesson gives two examples of rather straightforward commercial applications of cloud computing. One is server consolidation for multiple Microsoft database applications and the second is the benefits of scale comparing gmail to multiple smaller installations. It ends with some fiscal comments.

- Examples (3:26)

3.6.2.6 Value of Cloud Computing

Some comments on fiscal value of cloud computing.

- Value of Cloud Computing (4:20)

3.6.2.7 Resources

- http://www.slideshare.net/woorung/trend-and-future-of-cloud-computing
- http://www.slideshare.net/JensNimis/cloud-computing-tutorial-jens-nimis
- https://setandbma.wordpress.com/2012/08/10/hype-cycle-2012-emerging-technologies/
- http://insights.dice.com/2013/01/23/big-data-hype-is-imploding-gartner-analyst-2/
- http://research.microsoft.com/pubs/78813/AJ18_EN.pdf
- http://static.googleusercontent.com/media/www.google.com/en//green/pdfs/green-computing.pdf

3.6.3 Software and Systems

We cover different views as to nature of architecture and application for Cloud Computing. Then we discuss cloud software for the cloud starting at virtual machine management (IaaS) and the broad Platform (middleware) capabilities with examples from Amazon and academic studies. We summarize the 21 layers and almost 300 software packages in the HPC-ABDS Software Stack explaining how they are used.

 Software and Systems (32)

3.6.3.1 What is Cloud Computing

This lesson gives some general remark of cloud systems from an architecture and application perspective.

-  What is Cloud Computing (6:20)

3.6.3.2 Introduction to Cloud Software Architecture: IaaS and PaaS I

We discuss cloud software for the cloud starting at virtual machine management (IaaS) and the broad Platform (middleware) capabilities with examples from Amazon and academic studies. We cover different views as to nature of architecture and application for Cloud Computing. Then we discuss cloud

software for the cloud starting at virtual machine management (IaaS) and the broad Platform (middleware) capabilities with examples from Amazon and academic studies. We summarize the 21 layers and almost 300 software packages in the HPC-ABDS Software Stack explaining how they are used.

-  Intro to IaaS and PaaS I (7:42)
-  Intro to IaaS and PaaS II (6:42)

We discuss cloud software for the cloud starting at virtual machine management (IaaS) and the broad Platform (middleware) capabilities with examples from Amazon and academic studies. We cover different views as to nature of architecture and application for Cloud Computing. Then we discuss cloud software for the cloud starting at virtual machine management (IaaS) and the broad Platform (middleware) capabilities with examples from Amazon and academic studies. We summarize the 21 layers and almost 300 software packages in the HPC-ABDS Software Stack explaining how they are used.

-  Software Architecture: (7:42)
-  IaaS and PaaS II: (6:43)

3.6.3.3 Using the HPC-ABDS Software Stack

Using the HPC-ABDS Software Stack.

-  ABDS (27:50)

3.6.3.4 Resources

- http://www.slideshare.net/JensNimis/cloud-computing-tutorial-jens-nimis
- http://research.microsoft.com/en-us/people/barga/sc09_cloudcomp_tutorial.pdf
- http://research.microsoft.com/en-us/um/redmond/events/cloudfutures2012/tuesday/Keynote_OpportunitiesAn
- http://cloudonomic.blogspot.com/2009/02/cloud-taxonomy-and-ontology.html

3.6.4 Architectures, Applications and Systems

We start with a discussion of Cloud (Data Center) Architectures with physical setup, Green Computing issues and software models. We summarize a 2014 Gartner analysis of Cloud computing providers. This is followed by applications on the cloud including data intensive problems, comparison with high performance computing, science clouds and the Internet of Things. Remarks on Security, Fault Tolerance and Synchronicity issues in cloud follow.

scroll: Architectures (64)

3.6.4.1 Cloud (Data Center) Architectures

Some remarks on what it takes to build (in software) a cloud ecosystem, and why clouds are the data center of the future are followed by pictures and discussions of several data centers from Microsoft (mainly) and Google. The role of containers is stressed as part of modular data centers that trade scalability for fault tolerance. Sizes of cloud centers and supercomputers are discussed as is "green" computing.

-  Coud Architecture (8:38)
-  Cloud Data Center Architecture (9:59)

3.6.4.2 Analysis of Major Cloud Providers

Gartner 2014 Analysis of leading cloud providers.

-  Analysis of Major Cloud Providers (21:40)

3.6.4.3 Commercial Cloud Storage Trends

Use of Dropbox, iCloud, Box etc.

-  Commercial Storage Trends (3:07)

3.6.4.4 Cloud Applications I

This short lesson discusses the need for security and issues in its implementation. Clouds trade scalability for greater possibility of faults but here clouds offer good support for recovery from faults. We discuss both storage and program fault tolerance noting that parallel computing is especially sensitive to faults as a fault in one task will impact all other tasks in the parallel job.

- Cloud Applications I (7:57)
- Cloud Applications II (7:44)

3.6.4.5 Science Clouds

Science Applications and Internet of Things.

- Science Clouds (19:26)

3.6.4.6 Security

This short lesson discusses the need for security and issues in its implementation.

- Security (2:34)

3.6.4.7 Comments on Fault Tolerance and Synchronicity Constraints

Clouds trade scalability for greater possibility of faults but here clouds offer good support for recovery from faults. We discuss both storage and program fault tolerance noting that parallel computing is especially sensitive to faults as a fault in one task will impact all other tasks in the parallel job.

- Comments on Fault Tolerance and Synchronicity Constraints (8:55)

3.6.4.8 Resources

- http://www.slideshare.net/woorung/trend-and-future-of-cloud-computing
- http://www.eweek.com/c/a/Cloud-Computing/AWS-Innovation-Means-Cloud-Domination-307831
- CSTI General Assembly 2012, Washington, D.C., USA Technical Activities Coordinating Committee (TACC) Meeting, Data Management, Cloud Computing and the Long Tail of Science October 2012 Dennis Gannon.
- http://research.microsoft.com/en-us/um/redmond/events/cloudfutures2012/tuesday/Keynote_OpportunitiesAn
- http://www.datacenterknowledge.com/archives/2011/05/10/uptime-institute-the-average-pue-is-1-8/
- https://loosebolts.wordpress.com/2008/12/02/our-vision-for-generation-4-modular-data-centers-one-way-of-getting-it-just-right/
- http://www.mediafire.com/file/zzqna34282frr2f/koomeydatacenterelectuse2
- http://www.slideshare.net/JensNimis/cloud-computing-tutorial-jens-nimis
- http://www.slideshare.net/botchagalupe/introduction-to-clouds-cloud-camp-columbus
- http://www.venus-c.eu/Pages/Home.aspx
- Geoffrey Fox and Dennis Gannon Using Clouds for Technical Computing To be published in Proceedings of HPC 2012 Conference at Cetraro, Italy June 28 2012
- https://berkeleydatascience.files.wordpress.com/2012/01/20120119berkeley.
- Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics, Bill Franks Wiley ISBN: 978-1-118-20878-6
- Anjul Bhambhri, VP of Big Data, IBM
- Conquering Big Data with the Oracle Information Model, Helen Sun, Oracle
- Hugh Williams VP Experience, Search & Platforms, eBay
- Dennis Gannon, Scientific Computing Environments
- http://research.microsoft.com/en-us/um/redmond/events/cloudfutures2012/tuesday/Keynote_OpportunitiesAn
- http://www.datacenterknowledge.com/archives/2011/05/10/uptime-institute-the-average-pue-is-1-8/
- https://loosebolts.wordpress.com/2008/12/02/our-vision-for-generation-4-modular-data-centers-one-way-of-getting-it-just-right/
- http://www.mediafire.com/file/zzqna34282frr2f/koomeydatacenterelectuse2

- http://searchcloudcomputing.techtarget.com/feature/Cloud-computing-experts-forecast-the-market-climate-in-2014
- http://www.slideshare.net/botchagalupe/introduction-to-clouds-cloud-camp-columbus
- http://www.slideshare.net/woorung/trend-and-future-of-cloud-computing
- http://www.venus-c.eu/Pages/Home.aspx
- http://www.kpcb.com/internet-trends

3.6.5 Data Systems

We describe the way users and data interact with a cloud system. The unit concludes with the treatment of data in the cloud from an architecture perspective and Big Data Processing from an application perspective with commercial examples including eBay.

 Data Systems (49)

3.6.5.1 The 10 Interaction scenarios (access patterns) I

The next 3 lessons describe the way users and data interact with the system.

-  The 10 Interaction scenarios I (10:26)

3.6.5.2 The 10 Interaction scenarios. Science Examples

This lesson describes the way users and data interact with the system for some science examples.

-  The 10 Interaction scenarios. Science Examples (16:34)

3.6.5.3 Remaining general access patterns

This lesson describe the way users and data interact with the system for the final set of examples.

 Access Patterns (11:36)

3.6.5.4 Data in the Cloud

Databases, File systems, Object Stores and NOSQL are discussed and compared. The way to build a modern data repository in the cloud is introduced.

 Data in the Cloud (10:24)

3.6.5.5 Applications Processing Big Data

This lesson collects remarks on Big data processing from several sources: Berkeley, Teradata, IBM, Oracle and eBay with architectures and application opportunities.

 Processing Big Data (8:45)

3.6.6 Resources

- http://bigdatawg.nist.gov/_uploadfiles/M0311_v2_2965963213.pdf
- https://dzone.com/articles/hadoop-t-etl
- http://venublog.com/2013/07/16/hadoop-summit-2013-hive-authorization/
- https://indico.cern.ch/event/214784/session/5/contribution/410
- http://asd.gsfc.nasa.gov/archive/hubble/a_pdf/news/facts/FS14.pdf
- http://blogs.teradata.com/data-points/announcing-teradata-aster-big-analytics-appliance/
- http://wikibon.org/w/images/2/20/Cloud-BigData.png
- http://hortonworks.com/hadoop/yarn/
- https://berkeleydatascience.files.wordpress.com/2012/01/20120119berkeley
- http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html

3.7 Big Data Use Cases Survey

This section covers 51 values of X and an overall study of Big data that emerged from a NIST (National Institute for Standards and Technology) study of Big

data. The section covers the NIST Big Data Public Working Group (NBD-PWG) Process and summarizes the work of five subgroups: Definitions and Taxonomies Subgroup, Reference Architecture Subgroup, Security and Privacy Subgroup, Technology Roadmap Subgroup and the Requirements andUse Case Subgroup. 51 use cases collected in this process are briefly discussed with a classification of the source of parallelism and the high and low level computational structure. We describe the key features of this classification.

3.7.1 NIST Big Data Public Working Group

This unit covers the NIST Big Data Public Working Group (NBD-PWG) Process and summarizes the work of five subgroups: Definitions and Taxonomies Subgroup, Reference Architecture Subgroup, Security and Privacy Subgroup, Technology Roadmap Subgroup and the Requirements and Use Case Subgroup. The work of latter is continued in next two units.

 Overview (45)

3.7.1.1 Introduction to NIST Big Data Public Working

The focus of the (NBD-PWG) is to form a community of interest from industry, academia, and government, with the goal of developing a consensus definitions, taxonomies, secure reference architectures, and technology roadmap. The aim is to create vendor-neutral, technology and infrastructure agnostic deliverables to enable big data stakeholders to pick-and-choose best analytics tools for their processing and visualization requirements on the most suitable computing platforms and clusters while allowing value-added from big data service providers and flow of data between the stakeholders in a cohesive and secure manner.

 Introduction (13:02)

3.7.1.2 Definitions and Taxonomies Subgroup

The focus is to gain a better understanding of the principles of Big Data. It is important to develop a consensus-based common language and vocabulary terms

used in Big Data across stakeholders from industry, academia, and government. In addition, it is also critical to identify essential actors with roles and responsibility, and subdivide them into components and sub-components on how they interact/ relate with each other according to their similarities and differences.

For Definitions: Compile terms used from all stakeholders regarding the meaning of Big Data from various standard bodies, domain applications, and diversified operational environments. For Taxonomies: Identify key actors with their roles and responsibilities from all stakeholders, categorize them into components and subcomponents based on their similarities and differences. In particular data Science and Big Data terms are discussed.

 Taxonomies (7:42)

3.7.1.3 Reference Architecture Subgroup

The focus is to form a community of interest from industry, academia, and government, with the goal of developing a consensus-based approach to orchestrate vendor-neutral, technology and infrastructure agnostic for analytics tools and computing environments. The goal is to enable Big Data stakeholders to pick-and-choose technology-agnostic analytics tools for processing and visualization in any computing platform and cluster while allowing value-added from Big Data service providers and the flow of the data between the stakeholders in a cohesive and secure manner. Results include a reference architecture with well defined components and linkage as well as several exemplars.

 Architecture (10:05)

3.7.1.4 Security and Privacy Subgroup

The focus is to form a community of interest from industry, academia, and government, with the goal of developing a consensus secure reference architecture to handle security and privacy issues across all stakeholders. This includes gaining an understanding of what standards are available or under

development, as well as identifies which key organizations are working on these standards. The Top Ten Big Data Security and Privacy Challenges from the CSA (Cloud Security Alliance) BDWG are studied. Specialized use cases include Retail/Marketing, Modern Day Consumerism, Nielsen Homescan, Web Traffic Analysis, Healthcare, Health Information Exchange, Genetic Privacy, Pharma Clinical Trial Data Sharing, Cyber-security, Government, Military and Education.

 Security (9:51)

3.7.1.5 Technology Roadmap Subgroup

The focus is to form a community of interest from industry, academia, and government, with the goal of developing a consensus vision with recommendations on how Big Data should move forward by performing a good gap analysis through the materials gathered from all other NBD subgroups. This includes setting standardization and adoption priorities through an understanding of what standards are available or under development as part of the recommendations. Tasks are gather input from NBD subgroups and study the taxonomies for the actors' roles and responsibility, use cases and requirements, and secure reference architecture; gain understanding of what standards are available or under development for Big Data; perform a thorough gap analysis and document the findings; identify what possible barriers may delay or prevent adoption of Big Data; and document vision and recommendations.

 Technology (4:14)

3.7.1.6 Interfaces Subgroup

This subgroup is working on the following document: *NIST Big Data Interoperability Framework: Volume 8, Reference Architecture Interface*.

This document summarizes interfaces that are instrumental for the interaction with Clouds, Containers, and HPC systems to manage virtual clusters to support the NIST Big Data Reference Architecture (NBDRA). The Representational State Transfer (REST) paradigm is used to define these interfaces allowing easy

integration and adoption by a wide variety of frameworks. . This volume, Volume 8, uses the work performed by the NBD-PWG to identify objects instrumental for the NIST Big Data Reference Architecture (NBDRA) which is introduced in the NBDIF: Volume 6, Reference Architecture.

This presentation was given at the *2nd NIST Big Data Public Working Group (NBD-PWG) Workshop* in Washington DC in June 2017. It explains our thoughts on deriving automatically a reference architecture form the Reference Architecture Interface specifications directly from the document.

The workshop Web page is located at

- https://bigdatawg.nist.gov/workshop2.php

The agenda of the workshop is as follows:

- https://bigdatawg.nist.gov/2017_NIST_Big_Data_PWG_WorkshopAgenda_

The Web cas of the presentation is given bellow, while you need to fast forward to a particular time

- Webcast: Interface subgroup: https://www.nist.gov/news-events/events/2017/06/2nd-nist-big-data-public-working-group-nbd-pwg-workshop
    - see: Big Data Working Group Day 1, part 2 Time start: 21:00 min, Time end: 44:00
- Slides: https://github.com/cloudmesh/cloudmesh.rest/blob/master/docs/NBDPWG-vol8.pptx?raw=true
- Document: https://github.com/cloudmesh/cloudmesh.rest/raw/master/docs/NIST.SP.150 8-draft.pdf

You are welcome to view other presentations if you are interested.

3.7.1.7 Requirements and Use Case Subgroup

The focus is to form a community of interest from industry, academia, and government, with the goal of developing a consensus list of Big Data requirements across all stakeholders. This includes gathering and understanding various use cases from diversified application domains.Tasks are gather use case input from all stakeholders; derive Big Data requirements from each use case; analyze/prioritize a list of challenging general requirements that may delay or prevent adoption of Big Data deployment; develop a set of general patterns capturing the *essence* of use cases (not done yet) and work with Reference Architecture to validate requirements and reference architecture by explicitly implementing some patterns based on use cases. The progress of gathering use cases (discussed in next two units) and requirements systemization are discussed.

 Requirements (27:28)

3.7.2 51 Big Data Use Cases

This units consists of one or more slides for each of the 51 use cases - typically additional (more than one) slides are associated with pictures. Each of the use cases is identified with source of parallelism and the high and low level computational structure. As each new classification topic is introduced we briefly discuss it but full discussion of topics is given in following unit.

 51 Use Cases (100)

3.7.2.1 Government Use Cases

This covers Census 2010 and 2000 - Title 13 Big Data; National Archives and Records Administration Accession NARA, Search, Retrieve, Preservation; Statistical Survey Response Improvement (Adaptive Design) and Non-Traditional Data in Statistical Survey Response Improvement (Adaptive Design).

 Government Use Cases (17:43)

3.7.2.2 Commercial Use Cases

This covers Cloud Eco-System, for Financial Industries (Banking, Securities & Investments, Insurance) transacting business within the United States; Mendeley - An International Network of Research; Netflix Movie Service; Web Search; IaaS (Infrastructure as a Service) Big Data Business Continuity & Disaster Recovery (BC/DR) Within A Cloud Eco-System; Cargo Shipping; Materials Data for Manufacturing and Simulation driven Materials Genomics.

 Commercial Use Cases (17:43)

3.7.2.3 Defense Use Cases

This covers Large Scale Geospatial Analysis and Visualization; Object identification and tracking from Wide Area Large Format Imagery (WALF) Imagery or Full Motion Video (FMV) - Persistent Surveillance and Intelligence Data Processing and Analysis.

 Defense Use Cases (15:43)

3.7.2.4 Healthcare and Life Science Use Cases

This covers Electronic Medical Record (EMR) Data; Pathology Imaging/digital pathology; Computational Bioimaging; Genomic Measurements; Comparative analysis for metagenomes and genomes; Individualized Diabetes Management; Statistical Relational Artificial Intelligence for Health Care; World Population Scale Epidemiological Study; Social Contagion Modeling for Planning, Public Health and Disaster Management and Biodiversity and LifeWatch.

 Healthcare and Life Science Use Cases (30:11)

3.7.2.5 Deep Learning and Social Networks Use Cases

This covers Large-scale Deep Learning; Organizing large-scale, unstructured collections of consumer photos; Truthy: Information diffusion research from Twitter Data; Crowd Sourcing in the Humanities as Source for Bigand Dynamic Data; CINET: Cyberinfrastructure for Network (Graph) Science and Analytics

and NIST Information Access Division analytic technology performance measurement, evaluations, and standards.

 Deep Learning and Social Networks Use Cases (14:19)

3.7.2.6 Research Ecosystem Use Cases

DataNet Federation Consortium DFC; The 'Discinnet process', metadata -big data global experiment; Semantic Graph-search on Scientific Chemical and Text-based Data and Light source beamlines.

 Research Ecosystem Use Cases (9:09)

3.7.2.7 Astronomy and Physics Use Cases

This covers Catalina Real-Time Transient Survey (CRTS): a digital, panoramic, synoptic sky survey; DOE Extreme Data from Cosmological Sky Survey and Simulations; Large Survey Data for Cosmology; Particle Physics: Analysis of LHC Large Hadron Collider Data: Discovery of Higgs particle and Belle II High Energy Physics Experiment.

 Astronomy and Physics Use Cases (17:33)

3.7.2.8 Environment, Earth and Polar Science Use Cases

EISCAT 3D incoherent scatter radar system; ENVRI, Common Operations of Environmental Research Infrastructure; Radar Data Analysis for CReSIS Remote Sensing of Ice Sheets; UAVSAR Data Processing, DataProduct Delivery, and Data Services; NASA LARC/GSFC iRODS Federation Testbed; MERRA Analytic Services MERRA/AS; Atmospheric Turbulence - Event Discovery and Predictive Analytics; Climate Studies using the Community Earth System Model at DOE's NERSC center; DOE-BER Subsurface Biogeochemistry Scientific Focus Area and DOE-BER AmeriFlux and FLUXNET Networks.

Environment, Earth and Polar Science Use Cases (25:29)

3.7.2.9 Energy Use Case

This covers Consumption forecasting in Smart Grids.

Energy Use Case (4:01)

3.7.3 Features of 51 Big Data Use Cases

This unit discusses the categories used to classify the 51 use-cases. These categories include concepts used for parallelism and low and high level computational structure. The first lesson is an introduction to all categories and the further lessons give details of particular categories.

Features (43)

3.7.3.1 Summary of Use Case Classification

This discusses concepts used for parallelism and low and high level computational structure. Parallelism can be over People (users or subjects), Decision makers; Items such as Images, EMR, Sequences; observations, contents of online store; Sensors – Internet of Things; Events; (Complex) Nodes in a Graph; Simple nodes as in a learning network; Tweets, Blogs, Documents, Web Pages etc.; Files or data to be backed up, moved or assigned metadata; Particles/cells/mesh points. Low level computational types include PP (Pleasingly Parallel); MR (MapReduce); MRStat; MRIter (Iterative MapReduce); Graph; Fusion; MC (Monte Carlo) and Streaming. High level computational types include Classification; S/Q (Search and Query); Index; CF (Collaborative Filtering); ML (Machine Learning); EGO (Large Scale Optimizations); EM (Expectation maximization); GIS; HPC; Agents. Patterns include Classic Database; NoSQL; Basic processing of data as in backup or metadata; GIS; Host of Sensors processed on demand; Pleasingly parallel processing; HPC assimilated with observational data; Agent-based models; Multi-modal data fusion or Knowledge Management; Crowd Sourcing.

 Summary of Use Case Classification (23:39)

3.7.3.2 Database(SQL) Use Case Classification

This discusses classic (SQL) database approach to data handling with Search&Query and Index features. Comparisons are made to NoSQL approaches.

 Database (SQL) Use Case Classification (11:13)

3.7.3.3 NoSQL Use Case Classification

This discusses NoSQL (compared in previous lesson) with HDFS, Hadoop and Hbase. The Apache Big data stack is introduced and further details of comparison with SQL.

 NoSQL Use Case Classification (11:20)

3.7.3.4 Other Use Case Classifications

This discusses a subset of use case features: GIS, Sensors. the support of data analysis and fusion by streaming data between filters.

 Use Case Classifications I (12:42) This discusses a subset of use case features: Pleasingly parallel, MRStat, Data Assimilation, Crowd sourcing, Agents, data fusion and agents, EGO and security.

 Use Case Classifications II (20:18)

This discusses a subset of use case features: Classification, Monte Carlo, Streaming, PP, MR, MRStat, MRIter and HPC(MPI), global and local analytics (machine learning), parallel computing, Expectation Maximization, graphs and Collaborative Filtering.

 Use Case Classifications III (17:25)

3.7.3.5 Resources

- NIST Big Data Public Working Group (NBD-PWG) Process
- Big Data Definitions
- Big Data Taxonomies
- Big Data Use Cases and Requirements
- Big Data Security and Privacy
- Big Data Architecture White Paper Survey
- Big Data Reference Architecture
- Big Data Standards Roadmap

Some of the links bellow may be outdated. Please let us know the new links and notify us of the outdated links.

- DCGSA Standard Cloud
- On line 51 Use Cases

- Summary of Requirements Subgroup

- ~~Use Case 6 Mendeley~~ **(this link does not exist any longer)**
- ~~Use Case 7 Netflix~~
- Use Case 8 Search
  - http://www.slideshare.net/kleinerperkins/kpcb-internet-trends-2013,
  - http://webcourse.cs.technion.ac.il/236621/Winter2011-2012/en/ho_Lectures.html,
  - http://www.ifis.cs.tu-bs.de/teaching/ss-11/irws,
  - http://www.slideshare.net/beechung/recommender-systems-tutorialpart1intro,
  - http://www.worldwidewebsize.com/
- Use Case 9 IaaS (Infrastructure as a Service) Big Data Business Continuity & Disaster Recovery (BC/DR) Within A Cloud Eco-System provided by Cloud Service Providers (CSPs) and Cloud Brokerage Service Providers (CBSPs)
- Use Case 11 and Use Case 12 Simulation driven Materials Genomics
- Use Case 13 Large Scale Geospatial Analysis and Visualization

    - http://www.opengeospatial.org/standards
    - http://geojson.org/
    - http://earth-info.nga.mil/publications/specs/printed/CADRG/cadrg.html
- Use Case 14 Object identification and tracking from Wide Area Large Format Imagery (WALF) Imagery or Full Motion Video (FMV) - Persistent Surveillance
    - http://www.militaryaerospace.com/topics/m/video/79088650/persistent surveillance-relies-on-extracting-relevant-data-points-and-connecting-the-dots.htm,
    - http://www.defencetalk.com/wide-area-persistent-surveillance-revolutionizes-tactical-isr-45745/
- Use Case 15 Intelligence Data Processing and Analysis
    - http://www.afcea-aberdeen.org/files/presentations/AFCEAAberdeen_DCGSA_COLWell
    - http://stids.c4i.gmu.edu/STIDS2011/papers/STIDS2011_CR_T1_Salm
    - ~~http://stids.c4i.gmu.edu/papers/STIDSPapers/STIDS2012/_T14/_Smith~~
    - ~~https://www.youtube.com/watch?v=l4Qii7T8zeg~~
    - ~~http://dcgsa.apg.army.mil/~~
- Use Case 16 Electronic Medical Record (EMR) Data:
    - Regenstrief Institute
    - Logical observation identifiers names and codes
    - Indiana Health Information Exchange
    - ~~Institute of Medicine Learning Healthcare System~~
- Use Case 17
    - Pathology Imaging/digital pathology
    - https://web.cci.emory.edu/confluence/display/HadoopGIS
- Use Case 19 Genome in a Bottle Consortium:
    - ~~www.genomeinabottle.org~~
- Use Case 20 Comparative analysis for metagenomes and genomes

- Use Case 25
  - Biodiversity
  - LifeWatch
- Use Case 26 Deep Learning: Recent popular press coverage of deep learning technology:
  - http://www.nytimes.com/2012/11/24/science/scientists-see-advances-in-deep-learning-a-part-of-artificial-intelligence.html
  - http://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-machine-learning.html
  - http://www.wired.com/2013/06/andrew_ng/,
  - ~~A recent research paper on HPC for Deep Learning~~
  - Widely-used tutorials and references for Deep Learning:
    - http://ufldl.stanford.edu/wiki/index.php/Main_Page
    - http://deeplearning.net/
- Use Case 27 Organizing large-scale, unstructured collections of consumer photos
- Use Case 28
  - Truthy: Information diffusion research from Twitter Data
  - http://cnets.indiana.edu/groups/nan/truthy/
  - http://cnets.indiana.edu/groups/nan/despic/
- ~~Use Case 30 CINET: Cyberinfrastructure for Network (Graph) Science and Analytics~~
- Use Case 31 NIST Information Access Division analytic technology performance measurement, evaluations, and standards
- Use Case 32
  - DataNet Federation Consortium DFC: The DataNet Federation Consortium,
  - iRODS
- Use Case 33 The ‘Discinnet process’, big data global experiment
- Use Case 34 Semantic Graph-search on Scientific Chemical and Text-based Data
  - http://www.eurekalert.org/pub_releases/2013-07/aiop-ffm071813.php
  - http://xpdb.nist.gov/chemblast/pdb.pl
- Use Case 35 Light source beamlines
  - http://www-als.lbl.gov/
  - https://www1.aps.anl.gov/
- Use Case 36

    - CRTS survey
    - CSS survey
    - For an overview of the classification challenges
- Use Case 37 DOE Extreme Data from Cosmological Sky Survey and Simulations
    - http://www.lsst.org/lsst/
    - http://www.nersc.gov/
    - http://www.nersc.gov/assets/Uploads/HabibcosmosimV2.pdf
- Use Case 38 Large Survey Data for Cosmology
    - http://desi.lbl.gov/
    - http://www.darkenergysurvey.org/
- Use Case 39 Particle Physics: Analysis of LHC Large Hadron Collider Data: Discovery of Higgs particle
    - http://grids.ucs.indiana.edu/ptliupages/publications/Where%20does%2(
    - http://www.es.net/assets/pubs_presos/High-throughput-lessons-from-the-LHC-experience.Johnston.TNC2013.pdf
- ~~Use Case 40 Belle II High Energy Physics Experiment~~ **(old link does not exist, new link: https://www.belle2.org)**
- Use Case 41 EISCAT 3D incoherent scatter radar system
- Use Case 42 ENVRI, Common Operations of Environmental Research Infrastructure
    - ENVRI Project website
    - ~~ENVRI Reference Model~~
    - ~~ENVRI deliverable D3.2 : Analysis of common requirements of Environmental Research Infrastructures~~
    - ICOS,
    - Euro-Argo
    - EISCAT 3D
    - LifeWatch
    - EPOS
    - EMSO
- Use Case 43 Radar Data Analysis for CReSIS Remote Sensing of Ice Sheets
- Use Case 44 UAVSAR Data Processing, Data Product Delivery, and Data Services
    - http://uavsar.jpl.nasa.gov/,
    - http://www.asf.alaska.edu/program/sdc,

    - http://geo-gateway.org/main.html
- Use Case 47 Atmospheric Turbulence - Event Discovery and Predictive Analytics
    - http://oceanworld.tamu.edu/resources/oceanography-book/teleconnections.htm
    - http://www.forbes.com/sites/toddwoody/2012/03/21/meet-the-scientists-mining-big-data-to-predict-the-weather/
- Use Case 48 Climate Studies using the Community Earth System Model at DOE's NERSC center
    - http://www-pcmdi.llnl.gov/
    - http://www.nersc.gov/
    - http://science.energy.gov/ber/research/cesd/
    - http://www2.cisl.ucar.edu/
- Use Case 50 DOE-BER AmeriFlux and FLUXNET Networks
    - http://ameriflux.lbl.gov/
    - http://www.fluxdata.org/default.aspx
- Use Case 51 Consumption forecasting in Smart Grids
    - ~~http://smartgrid.usc.edu/~~ **(old link does not exsit, new link: http://dslab.usc.edu/smartgrid.php)**
    - http://ganges.usc.edu/wiki/Smart_Grid
    - https://www.ladwp.com/ladwp/faces/ladwp/aboutus/a-power/a-p-smartgridla?_afrLoop=157401916661989&_afrWindowMode=0&_afrWindowId=r&_adf.ctrl-state%3Db7yulr4rl_17
    - http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6475927

3.8 Sensors

We start with the Internet of Things IoT giving examples like monitors of machine operation, QR codes, surveillance cameras, scientific sensors, drones and self driving cars and more generally transportation systems. We give examples of robots and drones. We introduce the Industrial Internet of Things IIoT and summarize surveys and expectations Industry wide. We give examples from General Electric. Sensor clouds control the many small distributed devices of IoT and IIoT. 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.

 Sensor I (31)

 Sensor II (44)

3.8.1 Internet of Things

There are predicted to be 24-50 Billion devices on the Internet by 2020; these are typically some sort of sensor defined as any source or sink of time series data. Sensors include smartphones, webcams, monitors of machine operation, barcodes, surveillance cameras, scientific sensors (especially in earth and environmental science), drones and self driving cars and more generally transportation systems. The lesson gives many examples of distributed sensors, which form a Grid that is controlled by a cloud.

 Internet of Things (12:36)

3.8.2 Robotics and IoT

Examples of Robots and Drones.

 Robotics and IoT Expectations (8:05)

3.8.3 Industrial Internet of Things

We summarize surveys and expectations Industry wide.

 Industrial Internet of Things (24:02)

3.8.4 Sensor Clouds

We describe the architecture of a Sensor Cloud control environment and gives example of interface to an older version of it. The performance of system is measured in terms of processing latency as a function of number of involved sensors with each delivering data at 1.8 Mbps rate.

 Sensor Clouds (4:40)

3.8.5 Earth/Environment/Polar Science data gathered by Sensors

This lesson gives examples of some sensors in the Earth/Environment/Polar Science field. It starts with material from the CReSIS polar remote sensing project and then looks at the NSF Ocean Observing Initiative and NASA's MODIS or Moderate Resolution Imaging Spectroradiometer instrument on a satellite.

 Earth/Environment/Polar Science data gathered by Sensors (4:58)

3.8.6 Ubiquitous/Smart Cities

For Ubiquitous/Smart cities we give two examples: Iniquitous Korea and smart electrical grids.

 Ubiquitous/Smart Cities (1:44)

3.8.7 U-Korea (U=Ubiquitous)

Korea has an interesting positioning where it is first worldwide in broadband access per capita, e-government, scientific literacy and total working hours. However it is far down in measures like quality of life and GDP. U-Korea aims to improve the latter by Pervasive computing, everywhere, anytime i.e. by spreading sensors everywhere. The example of a 'High-Tech Utopia' New Songdo is given.

 U-Korea (U=Ubiquitous) (2:49)

3.8.8 Smart Grid

The electrical Smart Grid aims to enhance USA's aging electrical infrastructure by pervasive deployment of sensors and the integration of their measurement in a cloud or equivalent server infrastructure. A variety of new instruments include

smart meters, power monitors, and measures of solar irradiance, wind speed, and temperature. One goal is autonomous local power units where good use is made of waste heat.

 Smart Grid (6:04)

3.8.9 Resources

- http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture-Industrial-Internet-Changing-Competitive-Landscape-Industries.pdf [47]
- http://www.gesoftware.com/ge-predictivity-infographic [48]
- http://www.getransportation.com/railconnect360/rail-landscape [49]

- http://www.gesoftware.com/sites/default/files/GE-Software-Modernizing-Machine-to-Machine-Interactions.pdf **???**

  These resources do not exsit:
- ~~https://www.gesoftware.com/minds-and-machines~~
- ~~https://www.gesoftware.com/predix~~
- ~~https://www.gesoftware.com/sites/default/files/the-industrial-internet/index.html~~

- ~~https://developer.cisco.com/site/eiot/discover/overview/~~

3.9 RADAR

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.

 Radar (58)

3.9.1 Introduction

This lesson motivates radar-informatics by building on previous discussions on why X-applications are growing in data size and why analytics are necessary for acquiring knowledge from large data. The lesson details three mosaics of a changing Greenland ice sheet and provides a concise overview to subsequent lessons by detailing explaining how other remote sensing technologies, such as the radar, can be used to sound the polar ice sheets and what we are doing with radar images to extract knowledge to be incorporated into numerical models.

-  Radar Informatics (3:31)

3.9.2 Remote Sensing

This lesson explains the basics of remote sensing, the characteristics of remote sensors and remote sensing applications. Emphasis is on image acquisition and data collection in the electromagnetic spectrum.

-  Remote Sensing (6:43)

3.9.3 Ice Sheet Science

This lesson provides a brief understanding on why melt water at the base of the ice sheet can be detrimental and why it’s important for sensors to sound the bedrock.

-  Ice Sheet Science (1:00)

3.9.4 Global Climate Change

This lesson provides an understanding and the processes for the greenhouse effect, how warming effects the Polar Regions, and the implications of a rise in sea level.

-  Global Climate Change (2:51)

3.9.5 Radio Overview

This lesson provides an elementary introduction to radar and its importance to remote sensing, especially to acquiring information about Greenland and Antarctica.

-  Radio Overview (4:16)

3.9.6 Radio Informatics

This lesson focuses on the use of sophisticated computer vision algorithms, such as active contours and a hidden markov model to support data analysis for extracting layers, so ice sheet models can accurately forecast future changes in climate.

-  Radio Informatics (3:35)

3.10 Web Search and Text Mining

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.

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.

We start 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 the section.

3.10.1 Web Search and Text Mining

The unit starts with the web with its size, shape (coming from the mutual linkage of pages by URL's) and universal power laws for number of pages with particular number of URL's linking out or in to page. Information retrieval is introduced and compared to web search. A comparison is given between semantic searches as in databases and the full text search that is base of Web search. The origin of web search in libraries, catalogs and concordances is summarized. DIKW – Data Information Knowledge Wisdom – model for web search is discussed. Then features of documents, collections and the important Bag of Words representation. Queries are presented in context of an Information Retrieval architecture. The method of judging quality of results including recall, precision and diversity is described. A time line for evolution of search engines is given.

Boolean and Vector Space models for query including the cosine similarity are introduced. Web Crawlers are discussed and then the steps needed to analyze data from Web and produce a set of terms. Building and accessing an inverted index is followed by the importance of term specificity and how it is captured in TF-IDF. We note how frequencies are converted into belief and relevance.

 Web Search and Text Mining (56)

3.10.1.1 The Problem

 Text Mining (9:56)

This lesson starts with the web with its size, shape (coming from the mutual linkage of pages by URL's) and universal power laws for number of pages with particular number of URL's linking out or in to page.

3.10.1.2 Information Retrieval

 Information Retrieval (6:06)

Information retrieval is introduced A comparison is given between semantic searches as in databases and the full text search that is base of Web search. The ACM classification illustrates potential complexity of ontologies. Some differences between web search and information retrieval are given.

3.10.1.3 History

 Web Search History (5:48)

The origin of web search in libraries, catalogs and concordances is summarized.

3.10.1.4 Key Fundamental Principles

 Principles (9:30)

This lesson describes the DIKW – Data Information Knowledge Wisdom – model for web search. Then it discusses documents, collections and the important Bag of Words representation.

3.10.1.5 Information Retrieval (Web Search) Components

 Fundamental Principals of Web Search (5:06)

This describes queries in context of an Information Retrieval architecture. The method of judging quality of results including recall, precision and diversity is described.

3.10.2 Search Engines

 Search Engines (3:08)

This short lesson describes a time line for evolution of search engines. The first web search approaches were directly built on Information retrieval but in 1998 the field was changed when Google was founded and showed the importance of URL structure as exemplified by PageRank.

3.10.2.1 Boolean and Vector Space Models

 Boolean and Vector Space Model (6:17)

This lesson describes the Boolean and Vector Space models for query including the cosine similarity.

3.10.2.2 Web crawling and Document Preparation

 Web crawling and Document Preparation (4:55)

This describes a Web Crawler and then the steps needed to analyze data from Web and produce a set of terms.

3.10.2.3 Indices

 Indices (5:44)

This lesson describes both building and accessing an inverted index. It describes how phrases are treated and gives details of query structure from some early logs.

3.10.2.4 TF-IDF and Probabilistic Models

 TF-IDF and Probabilistic Models (3:57)

It describes the importance of term specificity and how it is captured in TF-IDF. It notes how frequencies are converted into belief and relevance.

3.10.3 Topics in Web Search and Text Mining

 Text Mining (33)

We start with an overview of the different steps (data analytics) in web search. This is followed by Link Structure Analysis including Hubs, Authorities and PageRank. The application of PageRank ideas as reputation outside web search is covered. Issues in web advertising and search follow. his leads to emerging field of computational advertising. The use of clustering and topic models completes unit with Google News as an example.

3.10.3.1 Data Analytics for Web Search

 Web Search and Text Mining II (6:11)

This short lesson describes the different steps needed in web search including: Get the digital data (from web or from scanning); Crawl web; Preprocess data to get searchable things (words, positions); Form Inverted Index mapping words to documents; Rank relevance of documents with potentially sophisticated techniques; and integrate technology to support advertising and ways to allow or stop pages artificially enhancing relevance.

3.10.3.2 Link Structure Analysis including PageRank

 Related Applications (17:24)

The value of links and the concepts of Hubs and Authorities are discussed. This leads to definition of PageRank with examples. Extensions of PageRank viewed as a reputation are discussed with journal rankings and university department rankings as examples. There are many extension of these ideas which are not discussed here although topic models are covered briefly in a later lesson.

3.10.3.3 Web Advertising and Search

 Web Advertising and Search (9:02)

Internet and mobile advertising is growing fast and can be personalized more than for traditional media. There are several advertising types Sponsored search, Contextual ads, Display ads and different models: Cost per viewing, cost per clicking and cost per action. This leads to emerging field of computational advertising.

3.10.3.4 Clustering and Topic Models

 Clustering and Topic Models (6:21)

We discuss briefly approaches to defining groups of documents. We illustrate this for Google News and give an example that this can give different answers from word-based analyses. We mention some work at Indiana University on a Latent Semantic Indexing model.

3.10.3.5 Resources

All resources accessed March 2018.

- http://saedsayad.com/data_mining_map.htm
- http://webcourse.cs.technion.ac.il/236621/Winter2011-2012/en/ho_Lectures.html
- The Web Graph: an Overviews
- Jean-Loup Guillaume and Matthieu Latapy
- Constructing a reliable Web graph with information on browsing behavior, Yiqun Liu, Yufei Xue, Danqing Xu, Rongwei Cen, Min Zhang, Shaoping Ma, Liyun Ru
- http://www.ifis.cs.tu-bs.de/teaching/ss-11/irws
- https://en.wikipedia.org/wiki/PageRank
- Meeker/Wu May 29 2013 Internet Trends D11 Conference

3.11 Health Informatics

 Health Informatics (131)

This section 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 review current state of health care and trends associated with it including increased use of Telemedicine. We summarize an industry survey by GE and Accenture and an impressive exemplar Cloud-based medicine system from Potsdam. We give some details of big data in medicine. Some remarks on Cloud computing and Health focus on security and privacy issues.

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. Examples are given of the Internet of Things, which will have great impact on health including wearables. A study looks at 4 scenarios for healthcare in 2032. Two are positive, one middle of the road and one negative. The final topic is Genomics, Proteomics and Information Visualization.

3.11.1 Big Data and Health

This lesson starts with general aspects of Big Data and Health including listing subareas where Big data important. Data sizes are given in radiology, genomics, personalized medicine, and the Quantified Self movement, with sizes and access to European Bioinformatics Institute.

 Big Data and Health (10:02)

3.11.2 Status of Healthcare Today

This covers trends of costs and type of healthcare with low cost genomes and an aging population. Social media and government Brain initiative.

 Status of Healthcare Today (16:09)

3.11.3 Telemedicine (Virtual Health)

This describes increasing use of telemedicine and how we tried and failed to do this in 1994.

Telemedicine (8:21)

3.11.4 Medical Big Data in the Clouds

An impressive exemplar Cloud-based medicine system from Potsdam.

Medical Big Data in the Clouds (15:02)

3.11.4.1 Medical image Big Data

Medical Image Big Data (6:33)

3.11.4.2 Clouds and Health

Clouds and Health (4:35)

3.11.4.3 McKinsey Report on the big-data revolution in US health care

This lesson covers 9 aspects of the McKinsey report. These are the convergence of multiple positive changes has created a tipping point for

innovation; Primary data pools are at the heart of the big data revolution in healthcare; Big data is changing the paradigm: these are the value pathways; Applying early successes at scale could reduce US healthcare costs by $300 billion to $450 billion; Most new big-data applications target consumers and providers across pathways; Innovations are weighted towards influencing individual decision-making levers; Big data innovations use a range of public, acquired, and proprietary data

types; Organizations implementing a big data transformation should provide the

leadership required for the associated cultural transformation; Companies must develop a range of big data capabilities.

 McKinsey Report (14:53)

3.11.4.4 Microsoft Report on Big Data in Health

This lesson identifies data sources as Clinical Data, Pharma & Life Science Data, Patient & Consumer Data, Claims & Cost Data and Correlational Data. Three approaches are Live data feed, Advanced analytics and Social analytics.

 Microsoft Report on Big Data in Health (2:26)

3.11.4.5 EU Report on Redesigning health in Europe for 2020

This lesson summarizes an EU Report on Redesigning health in Europe for 2020. The power of data is seen as a lever for change in My Data, My decisions; Liberate the data; Connect up everything; Revolutionize health; and Include Everyone removing the current correlation between health and wealth.

 EU Report on Redesigning health in Europe for 2020 (5:00)

3.11.4.6 Medicine and the Internet of Things

The Internet of Things will have great impact on health including telemedicine and wearables. Examples are given.

 Medicine and the Internet of Things (8:17)

3.11.4.7 Extrapolating to 2032

A study looks at 4 scenarios for healthcare in 2032. Two are positive, one middle of the road and one negative.

 Extrapolating to 2032 (15:13)

3.11.4.8 Genomics, Proteomics and Information Visualization

A study of an Azure application with an Excel frontend and a cloud BLAST backend starts this lesson. This is followed by a big data analysis of personal genomics and an analysis of a typical DNA sequencing analytics pipeline. The Protein Sequence Universe is defined and used to motivate Multi dimensional Scaling MDS. Sammon's method is defined and its use illustrated by a metagenomics example. Subtleties in use of MDS include a monotonic mapping of the dissimilarity function. The application to the COG Proteomics dataset is discussed. We note that the MDS approach is related to the well known chisq method and some aspects of nonlinear minimization of chisq (Least Squares) are discussed.

 Genomics, Proteomics and Information Visualization (6:56)

Next we continue the discussion of the COG Protein Universe introduced in the last lesson. It is shown how Proteomics clusters are clearly seen in the Universe browser. This motivates a side remark on different clustering methods applied to metagenomics. Then we discuss the Generative Topographic Map GTM method that can be used in dimension reduction when original data is in a metric space and is in this case faster than MDS as GTM computational complexity scales like N not N squared as seen in MDS.

Examples are given of GTM including an application to topic models in Information Retrieval. Indiana University has developed a deterministic annealing improvement of GTM. 3 separate clusterings are projected for visualization and show very different structure emphasizing the importance of visualizing results of data analytics. The final slide shows an application of MDS to generate and visualize phylogenetic trees.

 Genomics, Proteomics and Information Visualization I (10:33)

 Genomics, Proteomics and Information Visualization: II (7:41)

 Proteomics and Information Visualization (131)

3.11.4.9 Resources

- https://wiki.nci.nih.gov/display/CIP/CIP+Survey+of+Biomedical+Imaging+ [50]
- http://grids.ucs.indiana.edu/ptliupages/publications/Where\%20does\%20all\ [2]
- ~~http://www.ieee-icsc.org/ICSC2010/Tony\%20Hey\%20-\%2020100923.pdf~~(this link does not exist any longer)
- http://quantifiedself.com/larry-smarr/ [51]
- http://www.ebi.ac.uk/Information/Brochures/ [52]
- http://www.kpcb.com/internet-trends **???**
- http://www.slideshare.net/drsteventucker/wearable-health-fitness-trackers-and-the-quantified-self [53]
- http://www.siam.org/meetings/sdm13/sun.pdf **???** –big-data-analytics-healthcare
- http://en.wikipedia.org/wiki/Calico_\%28company\%29 [54]
- http://www.slideshare.net/GSW_Worldwide/2015-health-trends **???** trends
- http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture-Industrial-Internet-Changing-Competitive-Landscape-Industries.pdf [55]
- http://www.slideshare.net/schappy/how-realtime-analysis-turns-big-medical-data-into-precision-medicine [56]
- http://medcitynews.com/2013/03/the-body-in-bytes-medical-images-as-a-source-of-healthcare-big-data-infographic/ [57]
- ~~http://healthinformatics.wikispaces.com/file/view/cloud_computing.ppt~~ (this link does not exist any longer)
- https://www.mckinsey.com/~/media/mckinsey/industries/healthcare%20syst **???**
- ~~https://partner.microsoft.com/download/global/40193764~~ (this link does not exist any longer)
- https://ec.europa.eu/eip/ageing/file/353/download_en?token=8gECi1RO
- ~~http://www.liveathos.com/apparel/app~~
- http://debategraph.org/Poster.aspx?aID=77 [58]
- ~~http://www.oerc.ox.ac.uk/downloads/presentations-from-events/microsoftworkshop/gannon~~(this link does not exist any longer)

- ~~http://www.delsall.org~~ (this link does not exist any longer)
- http://salsahpc.indiana.edu/millionseq/mina/16SrRNA_index.html [59]
- http://www.geatbx.com/docu/fcnindex-01.html **???**

4 TECHNOLOGIES

4.1 STATISTICS

We assume that you are familiar with elementary statistics including

- mean, minimum, maximum
- standard deviation
- probability
- distribution
- frequency distribution
- Gaussian distribution
- bell curve
- standard normal probabilities
- tables (z table)
- Regression
- Correlation

Some of these terms are explained in various sections throughout our application discussion. This includes especially the Physics section. However these terms are so elementary that any undergraduate or highschool book will provide you with a good introduction.

It is expected from you to identify these terms and you can contribute to this section with non plagiarized subsections explaining these topics for credit.

 Topics identified by a :?: can be contributed by students. If you are interested, use piazza for announcing your willingness to do so.

Mean, minimum, maximum:

Standard deviation:

Probability:

Distribution:

Frequency distribution:

Gaussian distribution:

Bell curve:

Standard normal probabilities:

Tables (z-table):

Regression:

Correlation:

4.1.1 Exercise

E.Statistics.1:

> *Pick a term from the previous list and define it while not plagiarizing. Create a pull request. Coordinate on piazza as to not duplicate someone else's contribution. Also look into outstanding pull requests.*

E.Statistics.2:

> *Pick a term from the previous list and develop a python program demonstrating it and create a pull request for a contribution into the examples directory. Make links to the github location. Coordinate on piazza as to not duplicate someone else's contribution. Also look into outstanding pull requests.*

4.2 Practical K-Means, Map Reduce, and Page Rank for Big Data Applications and Analytics

We use 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.

 K-Means I (11:42)

 K-Means II (11:54)

4.2.1 K-means in Practice

We introduce the k means algorithm in a gentle fashion and describes its key features including dangers of local minima. A simple example from Wikipedia is examined.

We use 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.

Files:

- https://github.com/cloudmesh-community/book/blob/master/examples/python/kmeans/xmean.py
- https://github.com/cloudmesh-community/book/blob/master/examples/python/kmeans/sample.csv
- https://github.com/cloudmesh-community/book/blob/master/examples/python/kmeans/parallel-kmeans.py
- https://github.com/cloudmesh-community/book/blob/master/examples/python/kmeans/kmeans-extra.py

4.2.1.1 K-means in Python

We use the K-means Python code in SciPy package to show real code for clustering and applies it a set of 85 two dimensional vectors – officially sets of weights and heights to be clustered to find T-shirt sizes. We run through Python

code with Matplotlib displays to divide into 2-5 clusters. Then we discuss Python to generate 4 clusters of varying sizes and centered at corners of a square in two dimensions. We formally give the K means algorithm better than before and make definition consistent with code in SciPy.

4.2.1.2 Analysis of 4 Artificial Clusters

We present clustering results on the artificial set of 1000 2D points described in previous lesson for 3 choices of cluster sizes *small large* and *very large*. We emphasize the SciPy always does 20 independent K means and takes the best result – an approach to avoiding local minima. We allow this number of independent runs to be changed and in particular set to 1 to generate more interesting erratic results. We define changes in our new K means code that also has two measures of quality allowed. The slides give many results of clustering into 2 4 6 and 8 clusters (there were only 4 real clusters). We show that the *very small* case has two very different solutions when clustered into two clusters and use this to discuss functions with multiple minima and a hill between them. The lesson has both discussion of already produced results in slides and interactive use of Python for new runs.

4.2.2 Parallel K-means

We modify 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. We stress that this simple 2 map version can be generalized to scalable parallelism.

Files:

- https://github.com/cloudmesh-community/book/blob/master/examples/python/kmeans/parallel-kmeans.py

4.2.3 PageRank in Practice

We use 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.

Files:

- https://github.com/cloudmesh-community/book/blob/master/examples/python/page-rank/pagerank1.py
- https://github.com/cloudmesh-community/book/blob/master/examples/python/page-rank/pagerank2.py

4.2.4 Resources

- https://en.wikipedia.org/wiki/Kmeans
- http://grids.ucs.indiana.edu/ptliupages/publications/DACIDR_camera_ready
- http://salsahpc.indiana.edu/millionseq/
- http://salsafungiphy.blogspot.com/
- https://en.wikipedia.org/wiki/Heuristic

4.3 PLOTVIZ

NOTE: This an legacy application this has now been replaced by WebPlotViz which is a web browser based visualization tool which provides added functionality's.

We introduce 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.

4.3.1 Using Plotviz Software for Displaying Point Distributions in 3D

We introduce 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.

 Plotviz (34)

Files:

- https://github.com/cloudmesh-community/book/blob/master/examples/python/plotviz/fungi-lsu-3-15-to-3-26-zeroidx.pviz
- https://github.com/cloudmesh-community/book/blob/master/examples/python/plotviz/datingrating-originallabels.pviz
- https://github.com/cloudmesh-community/book/blob/master/examples/python/plotviz/clusterFinal-M30-C28.pviz
- https://github.com/cloudmesh-community/book/blob/master/examples/python/plotviz/clusterfinal-m3-c3dating-reclustered.pviz

4.3.1.1 Motivation and Introduction to use

The motivation of Plotviz 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 data can be transformed by dimension reduction techniques to 3D and one can check analysis like clustering and/or see structure missed in a computer analysis. The motivations shows some Cheminformatics examples. The use of Plotviz is started in slide 4 with a discussion of input file which is either a simple text or more features (like colors) can be specified in a rich XML syntax. Plotviz deals with points and their classification (clustering). Next the protein sequence browser in 3D shows the basic structure of Plotviz interface. The next two slides explain the core 3D and 2D manipulations

respectively. Note all files used in examples are available to students.

 Motivation (7:58)

4.3.1.2 Example of Use I: Cube and Structured Dataset

Initially we start with a simple plot of 8 points – the corners of a cube in 3 dimensions – showing basic operations such as size/color/labels and Legend of points. The second example shows a dataset (coming from GTM dimension reduction) with significant structure. This has .pviz and a .txt versions that are compared.

 Example I (9:45)

4.3.1.3 Example of Use II: Proteomics and Synchronized Rotation

This starts with an examination of a sample of Protein Universe Browser showing how one uses Plotviz to look at different features of this set of Protein sequences projected to 3D. Then we show how to compare two datasets with synchronized rotation of a dataset clustered in 2 different ways; this dataset comes from k Nearest Neighbor discussion.

 Proteomics and Synchronized Rotation (9:14)

4.3.1.4 Example of Use III: More Features and larger Proteomics Sample

This starts by describing use of Labels and Glyphs and the Default mode in Plotviz. Then we illustrate sophisticated use of these ideas to view a large Proteomics dataset.

 Larger Proteomics Sample (8:37)

4.3.1.5 Example of Use IV: Tools and Examples

This lesson starts by describing the Plotviz tools and then sets up two examples –

Oil Flow and Trading – described in PowerPoint. It finishes with the Plotviz viewing of Oil Flow data.

 Plotviz I (10:17)

4.3.1.6 Example of Use V: Final Examples

This starts with Plotviz looking at Trading example introduced in previous lesson and then examines solvent data. It finishes with two large biology examples with 446K and 100K points and each with over 100 clusters. We finish remarks on Plotviz software structure and how to download. We also remind you that a picture is worth a 1000 words.

 Plotviz II (14:58)

4.3.2 Resources

Download

5 REFERENCES

[1] G. von Laszewski, F. Wang, H. Lee, H. Chen, and G. C. Fox, “Accessing Multiple Clouds with Cloudmesh,” in *Proceedings of the 2014 acm international workshop on software-defined ecosystems*, 2014, p. 8 [Online]. Available: https://github.com/cyberaide/paper-cloudmesh/raw/master/vonLaszewski-cloudmesh.pdf

[2] G. Fox, T. Hey, and A. Trefethen, “Where does all the data come from,” *Data-Intensive Science. Chapman and Hall/CRC*, pp. 15–51, 2011 [Online]. Available: http://grids.ucs.indiana.edu/ptliupages/publications/Where%20does%20all%20th

[3] G. Aad *et al.*, “Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc,” *Physics Letters B*, vol. 716, no. 1, pp. 1–29, 2012 [Online]. Available: http://www.sciencedirect.com/science/article/pii/S037026931200857X

[4] W. McKinney, *Python for data analysis: Data wrangling with pandas, numpy, and ipython*. O’Reilly Media, Inc., 2012.

[5] jwork, “Welcome to datamelt.” 2018 [Online]. Available: http://jwork.org/scavis/api/

[6] cms.cern, “Observation of higgs boson decay to bottom quakers.” Web Page, Aug-2018 [Online]. Available: https://cms.cern/

[7] X. Amatriain, “Building large-scale real-world recommender systems - recsys2012 tutorial.” Web Page, Sep-2012 [Online]. Available: https://www.slideshare.net/xamat/building-largescale-realworld-recommender-systems-recsys2012-tutorial

[8] S. Seuken, “PowerPoint presentation.” Web Page, Oct-2012 [Online]. Available: https://www.ifi.uzh.ch/ce/teaching/spring2012/16-Recommender-Systems_Slides.pdf

[9] Kaggle Inc, “Kaggle: Your home for data science.” Web Page, 2018 [Online]. Available: https://www.kaggle.com/

[10] M. Welling, “ICS175winter11.” Web Page, 2012 [Online]. Available: https://www.ics.uci.edu/~welling/teaching/CS77Bwinter12/CS77B_w12.html

[11] J. Hammerbacher, “Introduction to data science.” Web Page, Jan-2012 [Online]. Available: https://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf

[12] S. Curtis, “Netflix foretells ’house of cards’ success with cassandra big data engine | apps & wearables | techworld.” Web Page, Mar-2013 [Online]. Available: https://www.techworld.com/news/apps-wearables/netflix-foretells-house-of-cards-success-with-cassandra-big-data-engine-3437514/

[13] Wikipedia, “A/b testing.” Web Page, 2018 [Online]. Available: https://en.wikipedia.org/wiki/A/B_testing

[14] A. Cockcroft, “Architectural patterns for high availability.” Web Page, Apr-2013 [Online]. Available: https://www.infoq.com/presentations/Netflix-Architecture

[15] T. Infotech, “Big data for big sports.” Web Page, Aug-2014 [Online]. Available: https://www.slideshare.net/Tricon_Infotech/big-data-for-big-sports

[16] E. Lewallen, “Sport analytics innovation summit.” Web Page, Dec-2013 [Online]. Available: https://www.slideshare.net/elew/sport-analytics-innovation

[17] J. Beckham, “SmartBall keeps an eye inside the ball.” Web Page, Feb-2013 [Online]. Available: https://www.wired.com/2013/02/catapault-smartball/

[18] J. Varadarajan, “Automated playbook generation in football through videos.” Presentation, Jun-2014 [Online]. Available: http://www.sloansportsconference.com/wp-content/uploads/2014/06/Automated_Playbook_Generation.pdf

[19] A. D. S. Center, “Football trajectory dataset - interactive digital media: Semantic analysis of video.” Web Page, Oct-2018 [Online]. Available: http://autoscout.adsc.illinois.edu/publications/football-trajectory-dataset/

[20] K. Goldsberry, “CourtVision: New visual and spatial analytics for the nba.” Presentation, Feb-2012 [Online]. Available: http://www.sloansportsconference.com/wp-content/uploads/2012/02/Goldsberry_Sloan_Submission.pdf

[21] GameSetMap, “GameSetMap acquired by golden set analytics.” Web Page, Nov-2017 [Online]. Available: http://gamesetmap.com/

[22] N. Santos, “Sports analytics innovation summit - data powered storytelling.” Web Page, Sep-2013 [Online]. Available: https://www.slideshare.net/BrandEmotivity/sports-analytics-innovation-summit-data-powered-storytelling

[23] ESPN, “MIT sloan sports analytics conference.” Web Page, Mar-2019 [Online]. Available: http://www.sloansportsconference.com/

[24] SABR, “Society for american baseball research.” Web Page, Oct-2018 [Online]. Available: https://sabr.org/

[25] Wikipedia, “Sabermetrics.” Web Page, 2018 [Online]. Available: https://en.wikipedia.org/wiki/Sabermetrics

[26] Wikipedia, “Baseball statistics.” Web Page, Oct-2018 [Online]. Available: https://en.wikipedia.org/wiki/Baseball_statistics

[27] M. Newman, “MLBAM introduces new way to analyze every play.” Web Page, Mar-2014 [Online]. Available: https://www.mlb.com/news/mlbam-introduces-new-way-to-analyze-every-play/c-68514514

[28] FANGRAPHS, “Complete list (offense).” Web Page, Oct-2018 [Online]. Available: https://www.fangraphs.com/library/offense/offensive-statistics-list/

[29] Wikipedia, “Component era.” Web Page, Dec-2016 [Online]. Available: https://en.wikipedia.org/wiki/Component_ERA

[30] Wikipedia, “Wins above replacement.” Web Page, Sep-2016 [Online]. Available: https://en.wikipedia.org/wiki/Wins_Above_Replacement

[31] FANGRAPHS, “What is war.” Web Page, Oct-2018 [Online]. Available:

https://www.fangraphs.com/library/misc/war/

[32] Sports Reference LLC, “Baseball-reference.com war explained.” Web Page, Oct-2018 [Online]. Available: https://www.baseball-reference.com/about/war_explained.shtml

[33] Sports Reference LLC, “War cpmparison chart.” Web Page, Oct-2018 [Online]. Available: https://www.baseball-reference.com/about/war_explained_comparison.shtml

[34] Sports Reference LLC, “Position player war calculations and details.” Web Page, Oct-2018 [Online]. Available: https://www.baseball-reference.com/about/war_explained_position.shtml

[35] Sports Reference LLC, “Pitcher war calculations and details.” Web Page, Oct-2018 [Online]. Available: https://www.baseball-reference.com/about/war_explained_pitch.shtml

[36] FANGRAPHS, “2018 fans’ scouting report.” Web Page, Oct-2018 [Online]. Available: https://www.fangraphs.com/leaders.aspxpos=all&stats=bat&lg=all&qual=y&type

[37] Batting Leadoff, “Coffee as energizer of baseball players?” Web Page, Aug-2018 [Online]. Available: http://battingleadoff.com/

[38] Wikipedia, “Coefficient of determination.” Web Page, Sep-2018 [Online]. Available: https://en.wikipedia.org/wiki/coefficient_of_determination

[39] G. Ganeshapillai and J. Guttag, “A data-driven method for in-game decision making in mlb,” in *Sport analytics conf*, 2014.

[40] MLB Advanced Media, “Opening day, mike trout and 2014 season expectations.” Web Page, Mar-2014 [Online]. Available: http://vincegennaro.mlblogs.com/

[41] M. Fast, “Spinning yarn: How accurate is pitchtrax.” Web Page, Mar-2011 [Online]. Available: https://www.baseballprospectus.com/news/article/13109/spinning-yarn-how-accurate-is-pitchtrax/

[42] M. Fast, “What the heck is pitchf/x,” illinois university of urbana-champaign, 2010 [Online]. Available: http://baseball.physics.illinois.edu/FastPFXGuide.pdf

[43] K. McSurley and G. Rybarczyk, “An introduction to fieldf/x,” illinois university of urbana-champaign, 2011 [Online]. Available: http://baseball.physics.illinois.edu/FieldFX-TDR-GregR.pdf

[44] K. Wagner, “MLB announces revolutionary new fielding-tracking system.” Web Page, Mar-2014 [Online]. Available: https://deadspin.com/mlb-announces-revolutionary-new-fielding-tracking-syste-1534200504

[45] J. Keri, “Q&A: MLB advanced media’s bob bowman discusses revolutionary new play-tracking system.” Web Page, Mar-2014 [Online]. Available: http://grantland.com/the-triangle/mlb-advanced-media-play-tracking-bob-bowman-interview/

[46] A. Andres, “The science of a home run: Andy andres at tedxyouth@BeaconStreet.” Apr-2013 [Online]. Available: https://www.youtube.com/watch?v=YkjtnuNmK74

[47] Accenture, “Winning with the industrial internet of things.” Web Page, Oct-2018 [Online]. Available: https://www.accenture.com/us-en/insight-industrial-internet-of-things

[48] GE digital, “No unplanned downtime.” Web Page, Oct-2015 [Online]. Available: https://www.predix.com/ge-industrial-internet-infographic

[49] GE transportation, “Driving the digital transformation of transportation.” Web Page, 2016 [Online]. Available: http://www.getransportation.com/digital-solutions

[50] J. (. [. Freymann, “CIP survey of biomedical imaging archives.” Web Page, Oct-2016 [Online]. Available: https://wiki.nci.nih.gov/display/CIP/CIP+Survey+of+Biomedical+Imaging+Arch

[51] E. Ramirez, “Larry smarr archives - quantified self.” Web Page, Feb-2013 [Online]. Available: http://quantifiedself.com/larry-smarr/

[52] E. B. Institute, “About us.” Web Page, 2018 [Online]. Available: https://www.ebi.ac.uk/about

[53] S. Tucker, “Wearable health, fitness trackers, and the quantified self.” Web Page, Feb-2014 [Online]. Available: https://www.slideshare.net/drsteventucker/wearable-health-fitness-trackers-and-the-quantified-self

[54] Wikipedia, “Calico(company).” Web Page, Sep-2018 [Online]. Available: https://en.wikipedia.org/wiki/Calico_%28company%29

[55] accenture, “Winning with the indusrial internet of things.” Web Page, 2018 [Online]. Available: https://www.accenture.com/us-en/insight-industrial-internet-of-things

[56] M. Schapranow, “How real-time analysis turns big medical data into precision medicine.” Web Page, Sep-2014 [Online]. Available: https://www.slideshare.net/schappy/how-realtime-analysis-turns-big-medical-data-into-precision-medicine

[57] D. Pogorelc, “The body in bytes: Medical images as a source of healthcare big data (infographic).” Web Page, Mar-2013 [Online]. Available: https://medcitynews.com/2013/03/the-body-in-bytes-medical-images-as-a-source-of-healthcare-big-data-infographic/

[58] debategraph, “RWJF symposium – june 2012.” Web Page, Jun-2012 [Online]. Available: https://debategraph.org/Poster.aspx?aID=77

[59] Indiana university Bloomington, “Million sequence clustering.” Web Page, 2008 [Online]. Available: http://salsahpc.indiana.edu/millionseq/mina/16SrRNA_index.html