Handbook of Big Data Applications and Analytics

Theory and Practice

Gregor von Laszewski

laszewski@gmail.com

laszewski@gmail.com

HTTPS://GITHUB.COM/CLOUDMESH/CLASSES

First printing, October 2017

Contents

Preface

1. Introduction

F chapter/preface/about.tex

1.1 About

The document is based on selected material published at the following Web page

- https://cloudmesh.github.io/classes/

It is part of a class taught at Indiana University. The class communication takes place at:

- https://piazza.com/class/ix39m27czn5uw

The PDF version will be made in future available at

- https://github.com/laszewski/laszewski.github.io/raw/master/papers/vonLaszewski-bigdat
  pdf

This PDF document will be updated based on feedback from the students and once we have now material available. For a more complete set of information we recommend the students to visit the Web page.

1.2 Citation

The bibtex entry for this document is

```
@TechReport{las17handbook,
  author =       {Gregor von Laszewski},
  title =        {Handbook of Big Data Applications and Analytics},
  institution =  {Indiana University},
  year =         {2017},
```

```
  OPTtype =       {Draft},
  address =       {Smith Research Center, Bloomington, IN 47408},
  month =         dec,
  url={https://github.com/laszewski/laszewski.github.io/raw/master/papers/vonLaszewski-bigd
}
```

1.3 Contributors

We like to acknowledge the following contributors that helped on this document. Please notify us with your name and a brief commend on what you contributed:

John Doe He contributed to none of teh sections as this is just an example.

1.4 Videos

Videos to the class are refered to with embeded links into the PDF document as follows:

Test Video (25:36) 

An index will also be available in the index page that lists on which page the video has been added.

1.5 Images

The video icon was copied from `http://www.freeiconspng.com/img/8039`.

II Documenting Scientific Research

2. Documenting Scientific Research

F chapter/doc/report-book.tex

2.1 Writing a Scientific Article or Conference Paper

An important part of any scientific research is to document it. This is often done through scientific conferences or journal articles. Hence it is important to learn how to prepare and submit such papers. Most conferences accept typically the papers in PDF format but require the papers to be prepared on MSWord or in LaTeX. While working with many students in the past we noticed however that those students using Word often spend unnecessarily countless hours on trying to make there papers beautiful while actually violating the template provided by the conference. Furthermore, we noticed that the same students had issues with bibliography management. Instead of Word helping the student it provided the illusion to be easier than LaTeX but when adding up the time spend on the paper we found that LaTeX actually saved time. This has been especially true with the advent of collaborative editing services such as sharelatex [**??**] and overleaf [**??**].

In this section we provide you with a professional template that is used for either system based on the ACM standard that you can use to write papers. Naturally this will be extremely useful if the quality of your research is strong enough to be submitted to a conference. We structure this section as follows. Although we do not recommend that you use MSWord for your editing of a scientific paper, we have included a short section about it and outline some of its pitfalls that initially you may not think is problematic, but has proven to be an issue with students. Next we will focus on introducing you to LaTeX and showcasing you the advantages and disadvantages. We will dedicate an entire section on bibliography management and teach you jhow to use jabref which clearly has advantages for us.

Having a uniform report format not only helps the students but allow allows the comparison of paper length and effort as part of teaching a course. We have added an entire section to this chapter that discusses how we can manage a *Class Proceedings* form papers that are contributed by teams

in the class.

2.1.1 Professional Paper Format

The report format we suggest here is based on the standard ACM proccedings format. It is of very high quality and can be adapted for your own activities. Moreover, it is possible to use most of teh text to adapt to other formats in case the conference you intend to submit your paper to has a different format. The ACM format is always a good start.

Important is that you do not need to change the template but you can change some parameters in case you are not submitting the paper to a conference but use it for class papers. Certainly you should not change the spacing or the layout and instead focus on writing content. As for bibliography management we recommend you use jabref which we will introduce in Section **??**.

We recommend that you carefully study the requirements for the report format. We would nat want that your paper gets rejected by a journal, conference or the class just because you try to modify the format or do not follow the established publication guidlines.

The template we are providing is available from:

- `https://github.com/cloudmesh/classes/tree/master/docs/source/format/report`

Convenient compressed files are available at

- `https://github.com/cloudmesh/classes/tree/master/docs/source/format/report.tar.gz`
- `https://github.com/cloudmesh/classes/tree/master/docs/source/format/report.zip`

You will find in it a modified ACM proceedings templates for Word and for LaTeX that has an identification box removed on the lower left hand side of the first page. This is done for classes so that you have more space to write. In case you must submit to a conference you can use the original ACM template. This template can be found at

2.1.2 Submission Requirements

Although the initial requirement for some conferences or journals is the document PDF, in many cases you must be prepared to provide the source when submitting to the conference. This includes the submission of the original images in an images foder. You may ba asked to package the document into a folder with all of its sources and submit to the conference for professional publication.

2.1.3 Microsoft Word vs. LaTeX

Microsoft word will provide you with the initial impression that you will safe lots of time writing in it while you see the layout of the document. This will be initially true, but once you progress to the more challenging parts and later pages such as image menagement and bibliography management you will see some issues. Thes include that figure placement in Word need sto be done just right in order for images to be where they need. We have seen students spending hours with the placement of figures in a paper but when they did additional changes the images jumped around and were not at the place where teh students expected them to be. So if you work with images, make sure you understand how to place them. Also always use relative caption counters so that if an image gets placed elsewhere the counter stays consistent. So nefer use just the number, but a reference

to the figure when referring to it. Recently a new bibliography management system was added to Word. However, however it is not well documented and the references are placed in the system bibliography rather than a local managed bibliography. This mah have severe consequences when working with many authors on a paper. The same is true when using Endnote. We have heard in many occasions that the combination of endnote and Word destroyed documents. You certainly do not want that to happen the day before your deadline. Also in classes we observed that those using LaTeX deliver better structured and written papers as the focus is on text and not beautiful layout.

For all these reasons we do not recommend that you use Word.

In LaTeX where we have an easier time with this as we can just ignore all of these issues due to relative good image placement and excellent support for academic reference management. Hence, it is in your best interest to use LaTeX. The information we provide here will make it easy for you to get started and write a paper in no time as it is just like filling out a form.

2.1.4 Working in a Team

Today research is done in potentially large research teams. This also include the writing of a document. There are multiple ways this is done these days and depends on the system you chose.

In MSWord you can use skydrive, while for LaTeX you can use sharelatex and overleaf. However, in many cases the use of github is possible as the same groups that develop the code are also familiar with github. Thus we provide you here also with the introduction on how to write a document in github while group members can contribute.

Here are the options:

LaTeX and git: This option will likely safe you time as you can use jabref also for managing collaborative bibliographies and

sharelatex: an online tool to write latex documents

overleaf: an online tool to write latex documents

MS onedrive: It allows you to edit a word document in collaboration. We recommend that you use a local installed version of Word and do the editing with that, rather than using the online version. The online editor has some bugs. See also (untested): `http://www.paulkiddie.com/2009/07/jabref-exports-to-word-2007-xml/`, `http://usefulcodes.blogspot.com/2015/01/using-jabref-to-import-bib-to-microsoft.html`

Google Drive: google drive could be used to collaborate on text that is than pasted into document. However it is just a starting point as it does not support typically the format required by the publisher. Hence at one point you need to swithc to one of the other systems.

2.1.5 Timemanagement

Obviously writing a paper takes time and you need to car-fully make sure you devote enough time to it. The important part is that the paper should not be an after thought but should be the initial activity to conduct and execute your research. Remember that

1. It takes time to read the information
2. It takes time understand the information
3. It takes time to do the research

For deadlines the following will get you in trouble:

1. *There are still 10 weeks left till the deadline, so let me start in 4 week* . . . . Procrastination is your worst enemy.

2. If you work in a team that has time management issues address them immediately
3. Do not underestimate the time it takes to prepare the final submission into the submission system. Prepare automated scripts that can deliver the package for submission in minutes rather than hours by hand.

2.1.6 Paper Checklist

In this section we summarize a number of checks that you may perform to make sure your paper is properly formatted and in excellent shape. Naturally this list is just a partial list and if you find things we should add here, let us know.

- ☐ Have you written the report in the specified format?
- ☐ Have you included an acknowledgement section?
- ☐ Have you included the paper in the submission system (In our class it is git)?
- ☐ Have you specified proper identification in the submission system. THis is typically a form or ASCII text that needs to be filled out (In our case it is a README.md file that includes a homework ID, names of the authors, and e-mails)?
- ☐ Have you included all images in native and PDF format in the submission system?
- ☐ Have you added the bibliography file that you managed (In our case jabref to make it simple for you)?
- ☐ In case you used word have you also provided the jabref?
- ☐ In case of a class and if you do a multi-author paper, have you added an appendix describing who did what in the paper?
- ☐ Have you spellchecked the paper?
- ☐ Are you using **a** and **the** properly?
- ☐ Have you made sure you do not plagiarize?
- ☐ Is the title properly capitalized?
- ☐ Have you not used phrases such as shown in the Figure below, but instead used as shown in Figure 3 when referring to the 3rd figure?
- ☐ Have you capitalized “Figure 3”, “Table 1”, . . . ?
- ☐ Have you removed any figure that is not referred explicitly in the text (As shown in Figure ..)
- ☐ Are the figure captions bellow the figures and not on top. (Do not include the titles of the figures in the figure itself but instead use the caption or that information?
- ☐ When using tables have you put the table caption on top?
- ☐ Make the figures large enough so we can read the details. If needed make the figure over two columns?
- ☐ Do not worry about the figure placement if they are at a different location than you think. Figures are allowed to float. If you want you can place all figures at the end of the report?
- ☐ Are all figures and tables at the end?
- ☐ In case you copied a figure from another paper you need to ask for copyright permission. IN case of a class paper you **must** include a reference to the original in the caption.
- ☐ Do not use the word “I” instead use we even if you are the sole author?
- ☐ Do not use the phrase “In this paper/report we show” instead use “We show”. It is not important if this is a paper or a report and does not need to be mentioned.
- ☐ Do not artificially inflate your paper if you are bellow the page limit and have nothing to say anymore.
- ☐ If your paper limit is 12 pages but you want to hand in 120 pages, please check first ;-)
- ☐ Do not use the characters & # % in the paper if you use LaTeX. If you use them you probably need a in front of them.
- ☐ If you want to say and do not use & but use the word and.

- ☐ Latex uses double single open quotes and double single closed quotes for quotes. Have you made sure you replaced them?
- ☐ Pasting and copying from the Web often results in non ascii characters to be used in your text, please remove them and replace accordingly.

In case of a class

- ☐ Check in your current work of the paper on a weekly basis to show consistent progress.
- ☐ Please use the dedicated report format for class. It may not be the ACM or IEEE format, but may have some additions that make management of bibliographies easier. Do follow our instructions for bibliographies.

In case you are allowed to use word in class, such as the one we teach at IU, the following applies in addition:

- ☐ Are you managing your references in jabref and endnote (we need both)
- ☐ Are you using the right template we have a special 2 column template for the class that is a modified version from the 2 column ACM template
- ☐ Are you using build in numbered section management? MSWord has Sections that must be used
- ☐ Are you using real bulleted lists in Word and not just a "*" or a "-"?
- ☐ Have you carelessly pasted and copied into the document without using proper formats. E.g. in MSWord this is a problem. You need to fix the format and use the build in format. Not that if you paste wrong you effect the format styles.
- ☐ Have you created not only a docx document but also the PDF.
- ☐ Make sure you use .docx and not .doc

If you observe something missing let us know.

2.1.7 Example Paper

An example report in PDF format is available:

- report.pdf

2.1.8 Creating the PDF from LaTeX on your Computer

Latex can be easily installed on any computer as long as you have enough space. Furthermore if your machine can execute the make command we have provided in the standard report format a simple Makefile that allows you to do editing with immediate preview as documented in the LaTeX lesson.

2.1.9 Class Specific README.md

For the class we will manage all papers via github.com. You will be added to our github at

- `https://github.com/bigdata-i523`

and assigned an hid (homework index directory) directory with a unique hid number for you. In addition, once you decide for a project, you will aslso get a project id (pid) and a directory in which you place the projects. Projects must not be placed in hid directories as they are treated differently and a class proceedings is automatically created based on your submission.

As part of the hid directory, you will need to create a README.md file in it, that **must** follow a

specific format. The good news is that we have developed an easy template that with common sense you can modify easily. The template is located at

- `https://raw.githubusercontent.com/bigdata-i523/sample-hid000/master/README.md`

As the format may have been updated over time it does not hurt to revisit it and compare with your README.md and make corrections. It is important that you follow the format and not eliminate the lines with the three quotes. The text in the quotes is actually yaml. yaml is a data format the any data scientist must know. If you do not, you can look it up. However, if you follow our rules you should be good. If you find a rule missing for our purpose, let us know. We like to keep it simple and want you to fill out the *template* with your information.

Simple rules:

- replace the hid nimber with your hid number.
- naturally if you see sample- in the directory name you need to delete that as your directory name does not have sample- in it.
- do not ignore where the author is to be placed, it is in a list starting with a -
- there is always a space after a -
- do not introduce empty lines
- do not use TAB and make sure your editor does not bay accident automatically creates tabs. This is probably the most frequent error we see.
- do not use any : & _ in the attribute text including titles
- an object defined in the README.md must have on a single type field. for example in the project section. Make sure you select only one type and delete the other
- in case you have long paragraphs you can use the > after the abstract
- Once you understood how the README.md works, please delete the comment section.
- Add a chapter topic that your paper belongs to

2.1.10 Exercise

Report.1: Install latex and jabref on your system

Report.2: Check out the report example directory. Create a PDF and view it. Modify and recompile.

Report.4: Learn about the different bibliographic entry formats in bibtex

Report.5: What is an article in a magazine? Is it really an Article or a Misc?

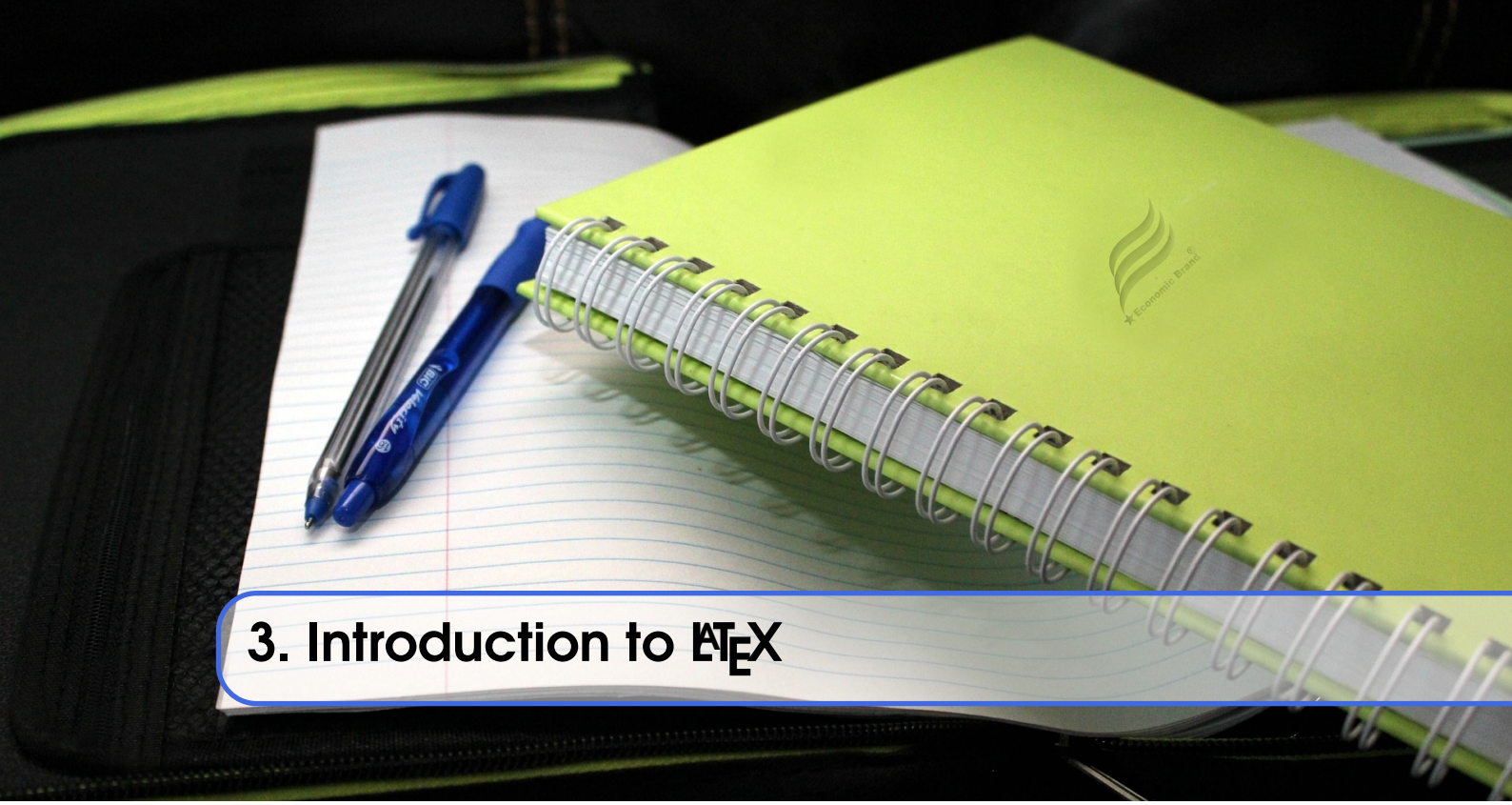
Report.6: What is an InProceedings and how does it differ from Conference?

Report.7: What is a Misc?

Report.8: Why are spaces, underscores in directory names problematic and why should you avoid using them for your projects

Report.9: Write an objective report about the advantages and disadvantages of programs to write reports.

Report.10: Why is it advantageous that directories are lowercase have no underscore or space in the name?



3. Introduction to LaTeX

F chapter/doc/latex.tex

Mastering a text processing system is an essential part of a researcher's life. Not knowing how to use a text processing system can slow down the productivity of research drastically.

3.1 Installation

LaTeX is available on all modern computer systems. A very good installation for OSX is available at:

- https://tug.org/mactex/

However, if you have older versions on your systems you may have to first completely uninstall them.

3.1.1 Local Install

Installing LaTeX is trivial, and is documented on the internet very well. However, it requires sufficient space and time as it is a large environment. A system such as TeX Live takes in full install about 5.5 GB. In addition to LaTeX we recommend that you install jabref and use it for bibliography management.

Thus you will have the most of them on your system.

- pdflatex: the latex program producing pdf
- bibtex: to create bibliographies
- jabref: GUI application to bibtex files (http://www.jabref.org/)

Make sure you check that these programs are there, for example with the Linux commands:

```
which pdflatex
which bibtex
which jabref (on OSX you may have an icon for it)
```

If these commands are missing, please install them. For the newest documentation on installation of LaTeX we recommend you look up the installation for your specific OS.

Install on Ubuntu 16.04

The easiest way to install it on ubuntu is to use the terminal and type in (make sure you have enough space):

```
sudo apt-get install texlive-full
```

One of the best editors for LaTeX is emacs as you can also do bibliography management with it and not just LaTeX. However, other editors are available including:

- Kile, TeXworks, JLatexEditor, Gedit LaTeX Plugin, TeXMaker

Please look up how to install them if you like to use them. TeXMaker is popular, However I find the combination of emacs and latexmk superior. TeXmaker is installed with:

```
sudo apt-get install texmaker
```

Other installations:

- kile is installed by default
- `https://www.tug.org/texworks/` (Works on ubuntu, Windows, OSX)

LaTeX for OSX

- `https://www.latex-project.org/get/`

LaTeX for Windows

- `https://www.latex-project.org/get/`

3.1.2 Online Services

Sharelatex

ShareLaTeX is an online, collaborative LaTeX editor that makes the creation, preview, and sharing of LaTeX documents easy through a web-based interface. Those that like to use latex, but do not have it installed on their computers may want to look at the following video:

Video: `https://youtu.be/PfhSOjuQk8Y`

Video with cc: `https://www.youtube.com/watch?v=8IDCGTFXoBs`

ShareLaTeX not only allows you to edit online, but allows you to share your documents in a group of up to three. Licenses are available if you need more than three people in a team.

IU Licensed ShareLaTeX

At IU we has a license for the ShareLaTeX service available to School of Informatics and Computing and Engeneering students, faculty, and staff only on the Bloomington campus.

You can create a free ShareLaTeX account but the free accounts have limitations. Adding your account to the IU license will give you access to advanced features, including unlimited sharing. It

will also allow GitHub integration. THis however only works with the commercial github.com and not the IU Enterprise GitHub at github.iu.edu. As we require in our courses github.com you will be able to use it.

Please note that this license is only available to School of Informatics and Computing students, faculty, and staff on the Bloomington campus. Students must be enrolled in one of the SoIC degree programs on the Bloomington campus to be eligible. Students in other degree programs (even those taking SoIC classes) are not eligible.

If you want to use this service, please do and be aware of the following:

1. Go to the ShareLaTeX site and register. Please note that you **must** use either an `@indiana.edu` or `@iu.edu` email address when you register. If you use any other email address, we will not be able to add you to our site license. You are also required to use your IU passphrase as your ShareLaTeX password. Once you have registered, send an email to `soichelp@indiana.edu` asking to have your sharelatex account added to the IU license.
2. In your request, you must include the following: The IU email address you used when you registered (which must be in either the `@indiana.edu` or `@iu.edu` domain) A statement indicating that you understand that the ShareLaTeX service cannot be used for any sensitive data
3. Note that the ShareLaTeX service is **not** qualified for any sensitive data. This includes all data in the Critical, Restricted, and University-Internal categories as defined in the Data Classifications Page.

Overleaf

Overleaf.com is a collaborative latex editor. In its free version it has a very limited disk space. However it comes with a Rich text mode that allows you to edit the document in a preview mode. The free templates provided do not include ACM template, put you are allowed to use the OSA template.

Features of overleaf are documented at: `https://www.overleaf.com/benefits`

Paperia

We do not know where this service is located. However it offers similar services as ShareLatex and Overleaf.

- `https://papeeria.com/`

3.2 Basic LaTeX Elements

Often researchers may be initially overwhelmed with all the features that LaTeXprovides. However, it is much simpler than you initially believe. In Chapter ?? we introduced you towards using an article template. As a template is provided you can just look at the elements in that article and modify or copy them while adapting the content. Thus, it is more like filling out a form. You do not have to learn much and you can learn as you go. We are providing in this chapter some basic LaTeXelements that will help you getting started quickly while serving you as a reminder what how to do certain things in LaTeX.

3.2.1 Characters

LaTeX is a command language and as such uses some special characters as part of the language. Thus if you want to use these characters either in your text or bibliography you need to be especially carful about. These characters include % $ # _

Other than in hypreref links and urls you need to put a backslash in front of them. For example to print a % in the text you need to use:

```
\%
```

Furthermore tthe character " is not at all used as discussed in the next section.

3.2.2 Highlighting Text

Quotes are not written with the " character, but are embedded in two left single quotes and two right single quotes:

```
‘‘This is a quote’’
```

which will result in:

“This is a quote”

In many papers we see that the quote is misused while putting quotes around a word. However quotes are often just used to quote a text from another paper. Instead of using quotes authors may actually emphasize a word. LaTeX has a special command for that using:

```
{\em this is emphasized}
```

resulting in

this is emphasized

To write a text as bold (which should also be avoided as bold is typically used in section headers), you can use:

```
{\em this is bold fett}
```

resulting in

this is bold fett

3.2.3 Sections

LaTeX provides a convenient mechanism to structure a paper with sections and subsections THis is achieved with the following commands:

```
\section{This is a Section}
\subsection{This is a Subsection}
\subsubsection{This is a Subsubsection}
```

Once you use one of these commands the next paragraph will start bellow the section command.

In addition you have the command:

```
\paragraph{This is a paragraph.}

The line is behind the paragraph heading
```

The command is special as it does not introduce a new line between the Heading and the next line even if you include empty lines

3.2.4 Empty Lines

Multiple empty lines will be reduced to a single empty line.

3.2.5 Itemize

Itemized lists can be written as:

```
\begin{itemize}
   \item First item
   \item Second item
\begin{itemize}
```

resulting in

- First item
- Second item

3.2.6 Enumerate

Enumerations can be written as:

```
\begin{enumerate}
   \item First item
   \item Second item
\begin{enumerate}
```

resulting in

1. First item
2. Second item

3.2.7 Descriptions

Description lists can be written as:

```
\begin{itemize}
   \item[Cloud] My definition of a Cloud.
   \item[Big Data] My definition of Big Data
\begin{itemize}
```

Cloud: My definition of a Cloud
Big Data: My definition of Big Data

3.2.8 Images

Figures are extremely easy to handle by including them from source. We never worry about the placement as LaTeX does typically a very good job of doing this.:

```
In Figure \ref{F:graph} we show a black and white graph about ... .
```

```
\begin{figure}
  \includegraphics[width=\columnwidth]{images/graph.pdf}
  \caption{A sample black and white graphic. \cite{las17graph}}
  \label{F:graph}
\end{figure}
```

Note that las17graph must be a label of a valid bibtex entry. This is needed if you have copied the image from elsewhere to avoid plagiarism. However, if you came up with the graph yourself than you do not need a citation.

We recommend that you place in your paper drafts all images at the which can be done with the endfloat package

This can be enabled if you include the following lines before begin document command:

```
\usepackage{endfloat}
\renewcommand{\efloatseparator}{\mbox{}}

\begin{document}
```

3.2.9 Tables

tables from csv tables by hand

3.2.10 Labels

As we saw already for figures and tables it is recommended to use the label and ref commands to refer to figure or table numbers. This applies also to sections. Thus I can place a label after a section:

```
\section{Introduction}\label{S:introduction}
```

and write elsewhere in the paper:

```
As we showcased in Section \ref{S:introduction}
```

Furthermore to conveniently distinguish sections tables and figures, we use the prefix S T F followed by a colon for the label. This helps organizing your paper in case you have many labels.

3.2.11 Mathematics

One of the strength of LaTeX thi the ability to write easily sophisticated mathematical expressions on paper with high quality. A good online resouce is provided by the following online resource from which we have copied some examples:

- `https://en.wikibooks.org/wiki/LaTeX/Mathematics`

To activate them use

```
\usepackage{amsmath}
```

at the beginnning of the document after the document class

Exponents are using the ^ character:

```
$ (a + b)^2 = a^2 + 2ab + b^{c+2} $
```

$$(a-b)^2 = a^2 - 2ab + b^2$$

Greek letters are referred to by their name proceeded by the slash:

```
$$ \alpha \beta \gamma \Gamma \pi \Pi \phi $$
```

$$\alpha\beta\gamma\Gamma\pi\Pi\phi$$

Limits can be written as follows:

```
$$ \lim_{x \to \infty} \exp(-x) = 0 $$
```

$$\lim_{x\to\infty} \exp(-x) = 0$$

Fractions are indicated by the frac command, and binomials by binom:

```
$ \frac{n!}{k!(n-k)!} = \binom{n}{k} $
```

$$\frac{n!}{k!(n-k)!} = \binom{n}{k}$$

Matricies can be created as follows:

```
A_{m,n} =
\begin{pmatrix}
  a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\
  a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\
  \vdots  & \vdots  & \ddots & \vdots  \\
  a_{m,1} & a_{m,2} & \cdots & a_{m,n}
\end{pmatrix}
```

$$A_{m,n} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix}$$

3.3 Advanced topics

3.3.1 ACM and IEEE Proceedings Format

- `http://www.acm.org/publications/proceedings-template`
- `https://www.ieee.org/conferences_events/conferences/publishing/templates.html`

3.3.2 Generating and Managing Images

To produce high quality images the programs PowerPoint and omnigraffle on OSX are recommended. When using powerpoint please keep the image ratio to 4x3 as they produce nice size

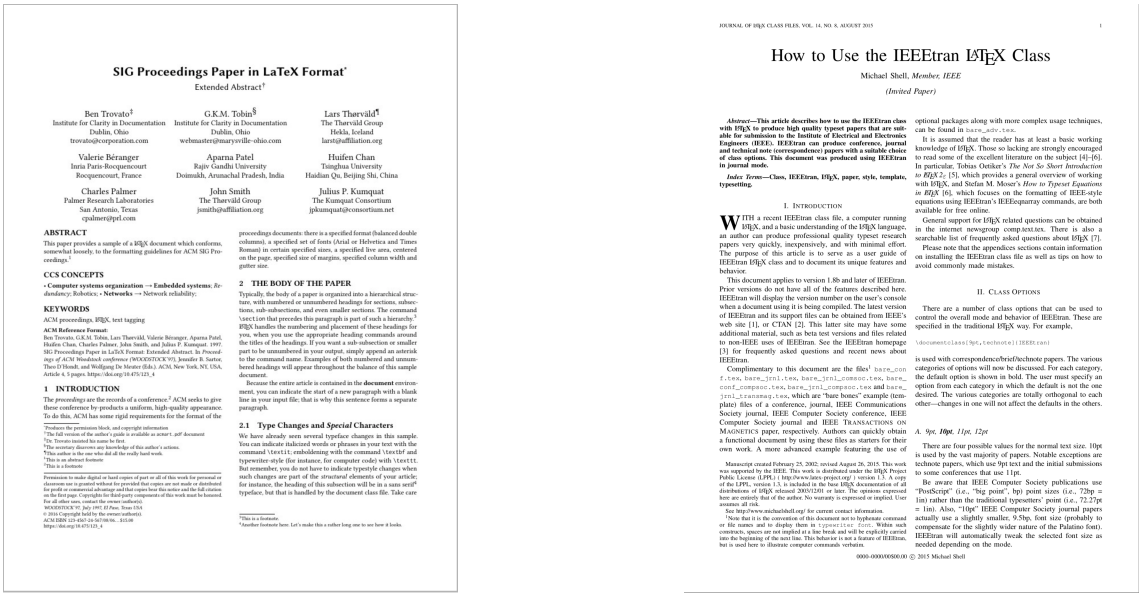


Figure 3.1: The look of the ACM and IEEE format templates

graphics which you also can use in your presentations. When using other rations they may not fit in presentations and thus you may increase unnecessarily your work. We do not recommend vizio as it is not universally available and produces images that in case you have to present them in a slide presentation does not easily reformat if you do not use 4x3 aspect ratio.

Naturally, graphics should be provided in SVG or PDF format so they can scale well when we look at the final PDF. Including PNG, gif, or jpeg files often do not result in the necessary resolution or the files become real big. For this reason we for example can also not recommend tools such as tablaeu as they do not provide proper exports to high quality publication formats. For interactive display such tool may be good, but for publications it produces inferior formatted images.

We recommend that all images be stored into a folder called images in the same directory where your LaTeXmain document resides.

3.3.3 Slides

Slides are best produced with the seminar package:

```
\documentclass{seminar}

\begin{slide}

    Hello World on slide 1

\end{slide}

The text between slides is ignored

\begin{slide}

    Hello World on slide 2
```

```
\end{slide}
```

However, in case you need to have a slide presentation we recommend you use ppt. Just paste and copy content from your PDF or your LaTeX source file into the ppt.

3.3.4 Useful Online Information about LaTeX

Latex Sheet: `https://wch.github.io/latexsheet/latexsheet.pdf`
Latex Short: `http://tug.ctan.org/info/lshort/english/lshort.pdf`
Wikibook: `https://en.wikibooks.org/wiki/LaTeX`
Wikibook (PDF) : `https://upload.wikimedia.org/wikipedia/commons/2/2d/LaTeX.pdf`
Links to books: `https://latexforhumans.wordpress.com/2008/10/11/the-best-guides-to-latex/`
Links to books: `https://www.latex-project.org/help/books/`
LaTeX2e: The LaTeX Reference Manual provides a good introduction to Latex.

- LaTeX Users and Reference Guide, by Leslie Lamport `https://www.amazon.com/LaTeX-Document-Prepara dp/0201529831/ref=sr_1_2?s=books&ie=UTF8&qid=1507114870&sr=1-2&keywords= lamport`
- LaTeX an Introduction, by Helmut Kopka `https://www.amazon.com/Guide-LaTeX-4th-Helmut-Kopka/ dp/0321173856/ref=pd_lpo_sbs_14_t_0?_encoding=UTF8&psc=1&refRID=2BB4APDFEX34A4JM65ZB`
- The LaTeX Companion, by Frank Mittelbach `https://www.amazon.com/LaTeX-Companion-Techniques-C dp/0201362996`

3.3.5 LaTeX vs. X

We will refrain from providing a detailed analysis on why we use LaTeX in many cases versus other technologies. In general, we find that LaTeX:

- is incredibly stable
- produces high-quality output
- is platform independent
- has lots of templates
- has been around for many years so it works well
- removes you from the pain of figure placements
- focusses you on content rather tan the appearance of the paper
- integrates well with code repositories such as git to write collaborative papers.
- has superior bibliography integration
- has a rich set of tools that make using LaTeX easier
- authors do not play with layouts much so papers in a format are uniform

In case you need a graphical view to edit LaTeX or LateX exportable files you also find AucTeX and Lyx.

Word

Word is arguably available to many, but if you work on Linux you may be out of luck. Also Word often focusses not on structure of the text but on its appearance. Many students abuse Word and the documents in Word become a pain to edit with multiple users. Recently Microsoft has offered online services to collaborate on writing documents in groups which work well. Integration with bibliography managers such as endnote or Mendeley is possible.

However, we ran into issues whenever we use word:

- Word tends sometimes to crash for unknown reasons and we lost a lot of work
- Word has some issues with the bibliography managers and tends to crash sometimes for unknown reasons.
- Word is slow with integration to large bibliographies.
- Figure placement in Word in some formats is a disaster and you will spend many hours to correct things just to find out that if you make small changes you have to spend additional many hours to get used to the new placement. We have not yet experienced a word version where we have not lost images. Maybe that has changed, so let us know

However, we highly recommend the collaborative editing features of Word that work on a paragraph and not letter level. Thus saving is essential so you do not block other people from editing the paragraph.

Google Docs

Unfortunately, many useful features got lost in the new google docs. However, it is great to collaborate quickly online, share thoughts and even write your latex documents together if you like (just copy your work in a file offline and use latex to compile it ;-))

The biggest issue we have with Google Docs is that it does not allow the support of 2 column formats, that the bibliography integration is non-existent and that paste and copy from web pages and images encourages unintended plagiarism when collecting information without annotations (LaTeX and Word are prone to this too, but we found from experience that it tends to happen more with Google docs users.

A Place for Each

When looking at the tools we find a place for each:

Google docs: Short meeting notes, small documents, quick online collaborations to develop documents collaboratively at the same time.

Word: Available to many, supports 2 column format, supports paragraph based collaborative editing, Integrates with bibliography managers.

LaTeX: Reduces failures, great offline editing, superior bibliography management, superior image placement, runs everywhere. Great collaborative editing with sharelatex, allows easy generation of proceedings written by hundreds of people with shared index.

The best choice for your class: LaTeX

3.4 Editing

3.4.1 Emacs

The text editor emacs provides a great basis for editing TeX and LaTeX documents. Both modes are supported. In addition there exists a color highlight module enabling the color display of LaTeX and TeX commands. On OSX aquaemacs and carbon emacs have build in support for LaTeX. Spell checking is done with flyspell in emacs.

Aquamacs

Aquamacs is an editor based on GNU Emacs that runs on OSX and integrates with the OSX desktop. This is for many the preferred editor on OSX for LaTeX.

`http://aquamacs.org`

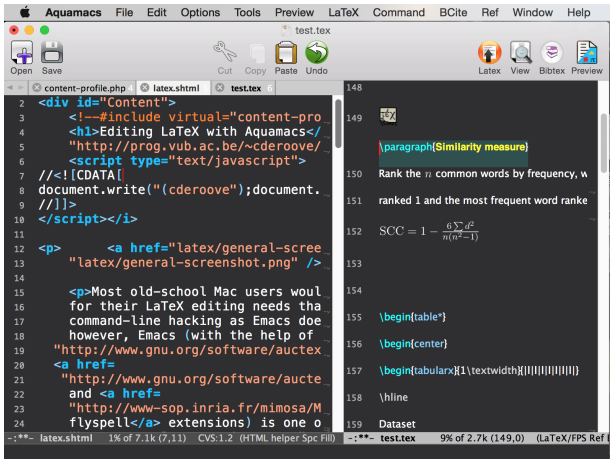


Figure 3.2: Aquamacs

3.4.2 Vi/Vim

Another popular editor is vi or vim. It is less feature rich but many programmers ar using it. As it can edit ASCII text you can edit LaTeX. With the LaTeX add-ons to vim, vim becomes similar powerful while offering help and syntax highlighting for LaTeX as emacs does. (The authors still prefer emacs)

3.4.3 TeXshop

Other editors such as TeXshop are available which provide a more integrated experience. However, we find them at times to stringent and prefer editors such as emacs.

3.4.4 LyX

We have made very good experiences with Lyx. You must assure that the team you work with uses it consistently and that you all use the same version.

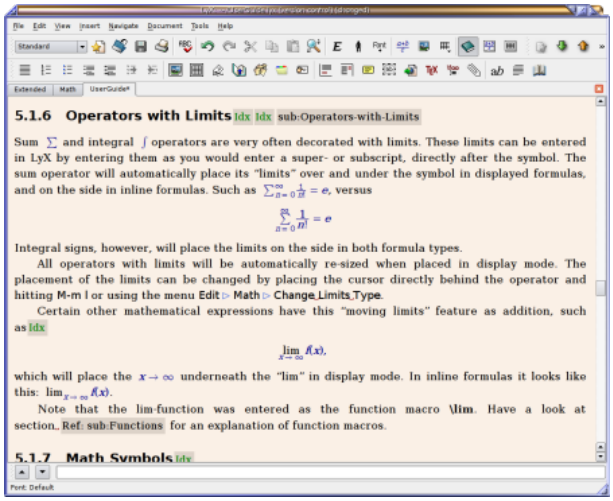


Figure 3.3: Lyx

Using the ACM templates is documented here:

- `https://wiki.lyx.org/Examples/AcmSiggraph`

On OSX it is important that you have a new version of LaTeX and Lyx installed. As it takes up quite some space, you ma want to delete older versions. The new version of LyX comes with the acmsigplan template included. However on OSX and other platforms the .cls file is not included by default. However the above link clearly documents how to fix this.

3.4.5 WYSIWYG locally

We have found that editors such as Lyx and Auctex provide very good WYSIWYG alike features. However, we found an even easier way while using skim, a pdf previewer, in conjunction with emacs and latexmk. This can be achieved while using the following command assuming your latex file is called 'report.tex':

```
latexmk -pvc -view=pdf report
```

This command will update your pdf previewer (make sure to use skim) whenever you edit the file report.tex and save it. It will maintain via skim the current position, thus you have a real great way of editing in one window, while seeing the results in the other.

Skim can be found at: `http://skim-app.sourceforge.net/`

3.4.6 Markdown and LaTeX

It may come as a surprise to many that one can actually write simple LaTeX documents also in markdown Syntax or mix section written in markdown while others are written in LaTeX. To do so all you ahve to do is place the markdown text in a separate file. Let us call the file `content.md` which has the following lines included in it:

```
# Section

* item a
* item b
```

Obviously, we would have to convert this to LaTeX. Luckily there is a very useful program called *pandoc* that does this for you. YOu could make the translation in the shell, but you could also make the translation locally on your computer while allowing LaTeXto start up external programs. This is achieved with the *write18* command and allowing LaTeX explicitly to call external programs. Please inspect the following latex file that includes a template on how to do this. We assume the file is called markdown.tex for our example.

```
\documentclass{article}

\include{graphicx}
\newcommand{\tightlist}{}

\begin{document}
\immediate\write18{pandoc content.md -o content.tex}

\input{content}

\end{document}
```

Now to generate the PDF we simply have to call the following command that include the *-shell-*

escape flag to allow the execution of write18 embedded commands:

```
pdflatex -shell-escape markdown-test
```

The output will be *markdown.pdf* with the content from the markdown file translated. Doing this naturally allows you to write large portions in markdown and automatically include them in your LaTeX document. Hence, you can use editors such as Macdown to initially work in semi WYSIWYG mode and do fairly straight forward edition. Naturally the same can be done in RST. Naturally the most elementary features are supported. For more sophisticated features, please use LaTeX directly.

3.4.7 Including RST into LaTeX

content.rst:

```
Section
-------

* item a
* item b
```

sample.tex:

```
\documentclass{article}

\include{graphicx}
\newcommand{\tightlist}{}

\begin{document}
\immediate\write18{pandoc content.rst -o content.tex}

\input{content}

\end{document}
```

3.4.8 pyCharm

TODO: comment on how we can use pycharm for editing and what the limitations are.

3.4.9 MSWord

it is possible to use Word.

be careful with

3.5 The LaTeX Cycle

To create a PDF file from latex yo need to generate it following a simple development and improvement cycle.

First, Create/edit ASCII source file with `file.tex` file:

```
emacs file.tex
```

Create/edit bibliography file:

```
jabref refs.bib
```

Create the PDF:

```
pdflatex file
bibtex file
pdflatex file
pdflatex file
```

View the PDF:

```
open file
```

It not only showcases you an example file in ACM 2 column format, but also integrates with a bibliography. Furthermore, it provides a sample Makefile that you can use to generate view and recompile, or even autogenerate. A compilation would look like:

```
make
make view
```

If however you want to do things on change in the tex file you can do this automatically simply with:

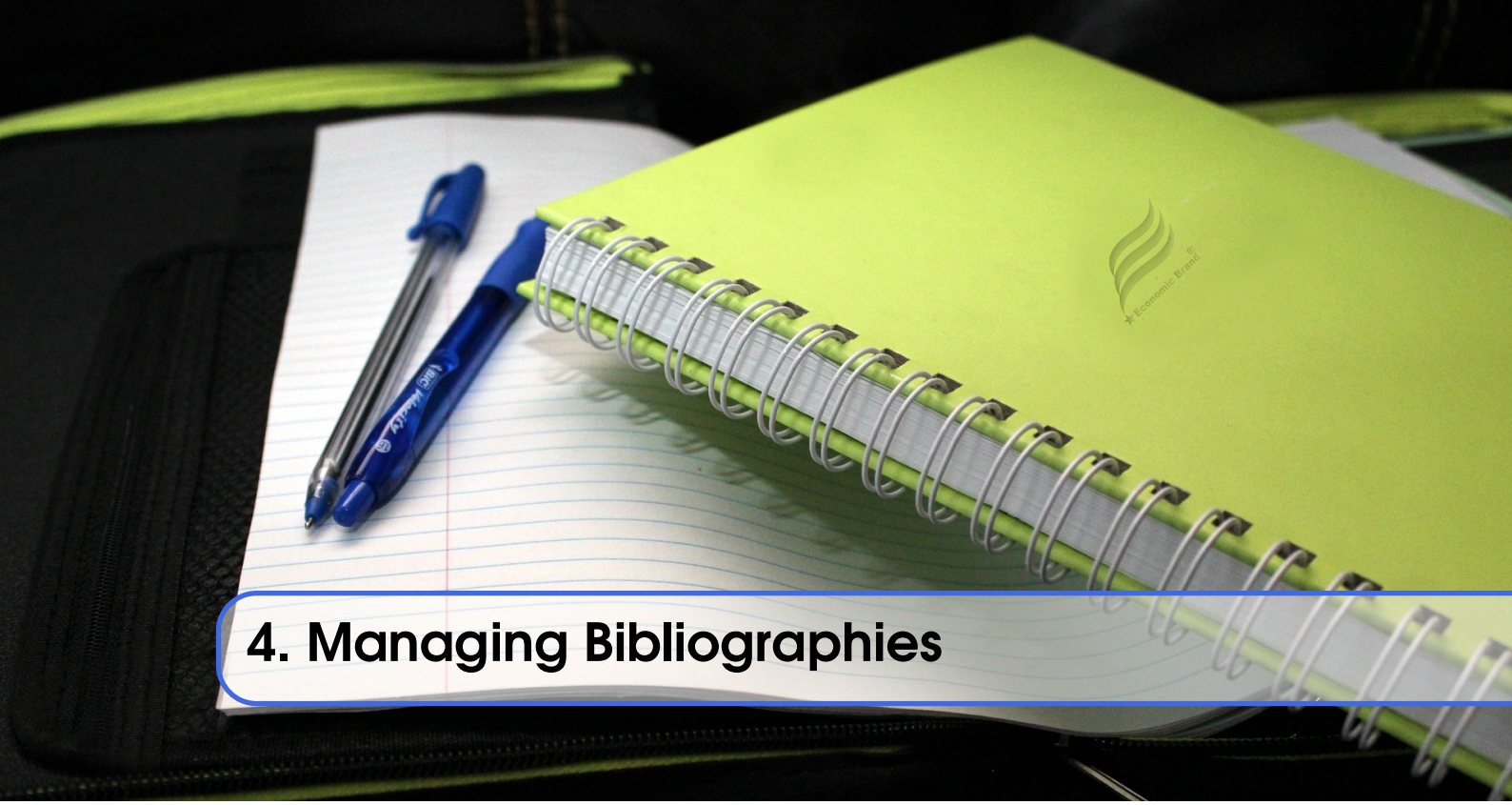
```
make watch
```

for make watch its best to use skim as pdf previewer

3.6 Tips

Including figures over two columns:

- `http://tex.stackexchange.com/questions/30985/displaying-a-wide-figure-in-a-two-column`
- positioning figures with textwidth and columnwidth `https://www.sharelatex.com/learn/Positioning_images_and_tables`
- An organization as the author. Assume the author is National Institute of Health and want to have the author show up, please do:
  `key= {National Institute of Health},`
  `author= {{National Institute of Health}},`
  Please note the {{ }}
- words containing 'fi' or 'ffi' showing blank places like below after recompiling it: find as nd efficiency as e ciency
  You copied from word or PDF ff which is actually not an ff, but a condensed character, change it to ff and ffi, you may find other such examples such as any non ASCII character. A degree is for example another common issue in data science.
- do not use | & and other latex characters in bibtex references, instead use , and the word and
- If you need to use _ it is _ but if you use urls leave them as is
- We do recommend that you use sharelatex and jabref for writing papers. This is the easiest solution and beats in many cases MSWord as you can focus on writing and not on formatting.



4. Managing Bibliographies

F https://github.com/cloudmesh/classes/blob/master/docs/source/lesson/doc/bibtex.rst

4.0.1 Integrating Bibliographies

LaTeX integrates very well with bibtex. There are several pre-formated styles available. It includes also styles for ACM and IEEE bibliographies. For the ACM style we recommend that you replace abbrv.bst with abbrvurl.bst, add hyperref to your usepackages so you can also display URLs in your citations:

```
\bibliographystyle{IEEEtran}
\bibliography{references.bib}
```

Then you have to run latex and bibtex in the following order:

```
latex  file
bibtex file
latex  file
latex  file
```

or simply call make from our makefile.

The reason for the multiple execution of the latex program is to update all cross-references correctly. In case you are not interested in updating the library every time in the writing progress just postpone it till the end. Missing citations are viewed as [?].

Two programs stand out when managing bibliographies: emacs and jabref:

- `http://www.jabref.org/`

Other programs such as Mendeley, Zotero, and even endnote integrate with bibtex. However their support is limited, so we recommend that you just use jabref. Furthermore its free and runs on all

platforms.

jabref

Jabref is a very simple to use bibliography manager for LaTeX and other systems. It can create a multitude of bibliography file formats and allows upload in other online bibliography managers.

- Installation: Go to `http://www.jabref.org/` and click download
- Video: `https://youtu.be/cMtYOHCHZ3k`
- Video with cc: `https://www.youtube.com/watch?v=QVbifcLgMic`

jabref and MSWord

According to others it is possible to integrate jabref references directly into MSWord. This has been conducted so far however only on a Windows computer.

We have not tried this ourselves, but give it as a potential option.

Here are the steps the need to be done:

1. Create the Jabref bibliography just like in presented in the Jabref video
2. After finishing adding your sources in Jabref, click File -> export
3. Name your bibliography and choose MS Office 2007(*.xml) as the file format. Remember the location of where you saved your file.
4. Open up your word document. If you are using the ACM template, go ahead and remove the template references listed under Section 7. References
5. In the MS Word ribbon choose ‘References’
6. Choose ‘Manage Sources’
7. Click ‘Browse’ and locate/select your Jabref xml file
8. You should now see your references appear in the left side window. Select the references you want to add to your document and click the ‘copy’ button to move them from the left side window to the right window.
9. Click the ‘Close’ button
10. In the MS Word Ribbon, select ‘Bibliography’ under the References tab
11. Click ‘Insert Bibliography’ and your references should appear in the document
12. Ensure references are of Style: IEEE. Styles are located in the References tab under ‘Manage Sources’

As you can see there is significant effort involve, so we do recommend you use LaTeX as you can focus there on content rather than dealing with complex layout decisions. This is especially true, if your papers have figures or tables, or you need to add references.

4.1 Entry types

In this section we will explain how to find and properly generate bibliographic entries. We are using bibtex for this as it is easy to use and generates reasonable entries that can be included in papers. What we like to achieve in this section is not to just show you a final entry, but to document the process on how that entry was derived. This will allow you to replicate or learn from the process to apply to your own entries.

We will address a number of important entry types which includes:

- wikipedia entries
- github entries

- books
- articles in a scientific journal
- articles in a conference
- articles in magazines (non scientific)
- blogs

4.1.1 Source code References

We will learn how to cite a source code from a publicly hosted repository. Such repositories are frequently used and include, for example github, bitbucket, sourcefore, or your Universities code repository as long as it is publicly reachable. As changes can occur on these repositories, it is important that the date ov access is listed in the entry or even the release version of the source code.

Let us without bias chose a random source dode entry that has been contributed by a student as follows:

```
@Misc{gonzalez_2015,
  Title =  {Buildstep},
  Author =     {Gonzalez, Jose and Lindsay,Jeff},
  HowPublished = {Web Page},
  Month =  {Jul},
  Note =   {Accessed: 2017-1-24},
  Year =   2015,
  Key =        {www-buildstep},
  Url =        {https://github.com/progrium/buildstep}
}
```

Is this entry correct? Let us analyse.

Entry type Misc

First, it seems appropriate to use a *@misc* entry. We correctly identify this is a misc entry as it is online available. More recent version of bibtex include also the type *@online* for it. However, in order to maintain compatibility to older formats we chose simply Misc here and if we really would need to we could replace it easily

Label

Typically the Label should contain 3 letters from an author name, short year and the short name of the publication to provide maximum information regarding the publication. Underscores need to be replaced by dashes or removed. However as this is a github repository it is better to integrate this into the label. Hence, we simply use the github-projectname (in our case github-buildstep, out of convention we only use lower case letters.

Author

Unless the last name contains spaces, it should be first name followed by the last name with multiple authors separated with “and”.

Key

In this case the key field can be removed as the entry has an author field entry. If there was no author field, we could use key to specify the alphabetical ordering based on the specified key. Note

that a key is not the label. In fact in our original entry the key field was wrongly used and the student did not understand that the key is used for sorting.

Howpublished

Since the source is a github project repository, the howpublished field shall hold the value {Code Repository} rather than a web page. If the url specified was a normal webpage, the {Web Page} entry would be valid.

Month

The lowercase month is, used for international notation since months are not capitalized in some other languages.

Owner

In class we introduced the convention to put the student HID in it. If multiple students contributed, add them with space separation.

Accessed

As we do not yet typically an accessed field, we simply include it in the note field. This is absolutely essential as code can change and when we read the code we looked at a particular snapshot in time. In addition it is often necessary to record the actual version of the code. Typically this can also be done with the month and year field while relying on a release date

Final Entry

Filling out as many fields as possible with information for this entry we get:

```
@Misc{github-buildstep,
  Title =  {Buildstep},
  Author =     {Jose Gonzalez and Jeff Lindsay},
  HowPublished = {Code Repository},
  Year =   {2015},
  Month =  jul,
  Note =   {Accessed: 2017-1-24},
  Url =    {https://github.com/progrium/buildstep},
  Owner =  {S17-IO-3025},
}
```

We are using the release date in the year and month field as this project uses this for organizing releases. However, other project may have release versions so you would have in addition to using the data also to include the version in the note field such as:

```
Note =     {Version: 1.2.3, Accessed: 2017-1-24},
```

All those that helped should add your HID to this entry with a space separated from each other

4.1.2 Researching proper bibtex entries

Article in a journal

Many online bibtex entries are wrong or incomplete. Often you may find via google a bibtex entry that may need some more research. Lets assume your first google query returns a publication and

you cite it such as this:

```
@Unpublished{unpublished-google-sawzall,
    Title = {{Interpreting the Data: Parallel Analysis with Sawzall}},
    Author = {{Rob Pike, Sean Dorward, Robert Griesemer, Sean Quinlan}},
    Note = {accessed 2017-01-28},
    Month = {October},
    Year = {2005},
    Owner = {for the purpose of this discussion removed},
    Timestamp = {2017.01.31}
}
```

Could we improve this entry to achieve your best? We observe:

1. The author field has a wrong entry as the , is to be replaced by an and.
2. The author feild has authors and thus must not have a {{ }}
3. The url is missing, as the simple google search actually finds a PDF document.

Let us investigate a bit more while searching for the title. We find

A) `https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwj_ytSA-PDRAhUH8IMKHaomC-oQFggaMAA&url=https%3A%2F%2Fresearch.google.com%2Farchive%2Fsawzall-sciprog.pdf&usg=AFQjCNHSSfKBwbxVAVPQ0td4rTjitKucpA&sig2=vbiVzi36B3gGFjIzlUKBDA&bvm=bv.14`

B) `https://research.google.com/pubs/pub61.html`

C) `http://dl.acm.org/citation.cfm?id=1239658`

Let us look at A)

As you can see from the url this is actually some redirection to a google web page which probably is replaced by B as its from google research. So let us look at B)

Now when you look at the link we find the url `https://research.google.com/archive/sawzall-sciprog.pdf` which redirects you to the PDF paper.

When we go to B) we find surprisingly a bibtex entry as follows:

```
@article{61,
  title = {Interpreting the Data: Parallel Analysis with Sawzall},
  author = {Rob Pike and Sean Dorward and Robert Griesemer and Sean Quinlan},
  year = 2005,
  URL = {https://research.google.com/archive/sawzall.html},
  journal = {Scientific Programming Journal},
  pages = {277--298},
  volume = {13}
}
```

Now we could say lets be satisfied, but C) seems to be even more interesting as its from a major publisher. So lats just make sure we look at C)

If you go to C, you find under the colored box entitled Tools and Resources a link called **bibtex**. Thus it seems a good idea to click on it. This will give you:

```
@article{Pike:2005:IDP:1239655.1239658,
    author = {Pike, Rob and Dorward, Sean and Griesemer, Robert and Quinlan, Sean},
    title = {Interpreting the Data: Parallel Analysis with Sawzall},
    journal = {Sci. Program.},
    issue_date = {October 2005},
```

```
    volume = {13},
    number = {4},
    month = oct,
    year = {2005},
    issn = {1058-9244},
    pages = {277--298},
    numpages = {22},
    url = {http://dx.doi.org/10.1155/2005/962135},
    doi = {10.1155/2005/962135},
    acmid = {1239658},
    publisher = {IOS Press},
    address = {Amsterdam, The Netherlands, The Netherlands},
}
```

Now we seem to be at a position to combine our search result as neither entry is sufficient. As the doi number properly specifies a paper (look up what a doi is) we can replace the url with one that we find online, such as the one we found in A) Next we see that all field sin B are already covered in C, so we take C) and add the url. Now as the label is great and uniform for ACM, but for us a bit less convenient as its difficult to remember, we just change it while for example using authors, title, and year information. lets also make sure to do mostly lowercase in the label just as a convention. Thus our entry looks like:

```
@article{pike05swazall,
    author = {Pike, Rob and Dorward, Sean and Griesemer, Robert and Quinlan, Sean},
    title = {Interpreting the Data: Parallel Analysis with Sawzall},
    journal = {Sci. Program.},
    issue_date = {October 2005},
    volume = {13},
    number = {4},
    month = oct,
    year = {2005},
    issn = {1058-9244},
    pages = {277--298},
    numpages = {22},
    url = {https://research.google.com/archive/sawzall-sciprog.pdf},
    doi = {10.1155/2005/962135},
    acmid = {1239658},
    publisher = {IOS Press},
    address = {Amsterdam, The Netherlands, The Netherlands},
}
```

As you can see properly specifying a reference takes multiple google queries and merging of the results you find from various returns. As you still have time to correct things I advise that you check your references and correct them. If the original reference would have been graded it would have been graded with a “fail” instead of a “pass”.

4.1.3 Article in a conference proceedings

Lets look at a second obvious example that needs improvement:

```
@InProceedings{wettinger-any2api,
  Title                    = {Any2API - Automated APIfication},
```

```
  Author                    = {Wettinger, Johannes and
                               Uwe Breitenb{\"u}cher
                               and Frank Leymann},
  Booktitle                 = {Proceedings of the 5th International
                               Conference on Cloud Computing and
            Services Science},
  Year                      = {2015},
  Pages                     = {475486},
  Publisher                 = {SciTePress},

  ISSN                      = {2326-7550},
  Owner                     = {S17-IO-3005},
  Url                       = {https://pdfs.semanticscholar.org/1cd4/4b87be8cf68ea5c4c642d38
}
```

As you can see this entry seems to define all required fields, so we could be tempted to stop here. But its good to double check. Lets do some queries against ACM, . and google scholar, so we just type in the title, and if this is in a proceedings they should return hopefully a predefined bibtex record for us.

Lets query:

`google: googlescholar Any2API Automated APIfication`

We get:

- `https://scholar.google.de/citations?view_op=view_citation&hl=en&user=j6lIXt0AAAAJ&citation_for_view=j6lIXt0AAAAJ:8k81kl-MbHgC`

On that page we see Cite

So we find a PDF at `https://pdfs.semanticscholar.org/1cd4/4b87be8cf68ea5c4c642d38678a7b40a86de.pdf`

Lets click on this and the document includes a bibtex entry such as:

```
@inproceedings{Wettinger2015,
  author= {Johannes Wettinger and Uwe Breitenb{\"u}cher and Frank
       Leymann},
  title = {Any2API - Automated APIfication},
  booktitle = {Proceedings of the 5th International Conference on Cloud
       Computing and Service Science (CLOSER)},
  year = {2015},
  pages = {475--486},
  publisher = {SciTePress}
}
```

Now lets add the URL and owner:

```
@inproceedings{Wettinger2015,
  author= {Johannes Wettinger and Uwe Breitenb{\"u}cher and Frank
       Leymann},
  title = {Any2API - Automated APIfication},
  booktitle = {Proceedings of the 5th International Conference on Cloud
       Computing and Service Science (CLOSER)},
```

```
  year = {2015},
  pages = {475--486},
  publisher = {SciTePress},
  url ={https://pdfs.semanticscholar.org/1cd4/4b87be8cf68ea5c4c642d38678a7b40a86de.pdf},
  owner = {S17-IO-3005},
}
```

Should we be satisfied? No, even our original information we gather provided more information. So lets continue. Lets googlesearch different queries with ACM or IEEE and the title. When doing the IEEE in the example we find an entry called

dlp: Frank Leyman

Lets look at it and we find two entries:

```
@inproceedings{DBLP:conf/closer/WettingerBL15,
  author    = {Johannes Wettinger and
       Uwe Breitenb{\"{u}}cher and
       Frank Leymann},
  title     = {{ANY2API} - Automated APIfication - Generating APIs for Executables
       to Ease their Integration and Orchestration for Cloud Application
       Deployment Automation},
  booktitle = {{CLOSER} 2015 - Proceedings of the 5th International Conference on
       Cloud Computing and Services Science, Lisbon, Portugal, 20-22 May,
       2015.},
  pages     = {475--486},
  year      = {2015},
  crossref  = {DBLP:conf/closer/2015},
  url       = {http://dx.doi.org/10.5220/0005472704750486},
  doi       = {10.5220/0005472704750486},
  timestamp = {Tue, 04 Aug 2015 09:28:21 +0200},
  biburl    = {http://dblp.uni-trier.de/rec/bib/conf/closer/WettingerBL15},
  bibsource = {dblp computer science bibliography, http://dblp.org}
}

@proceedings{DBLP:conf/closer/2015,
  editor    = {Markus Helfert and
       Donald Ferguson and
       V{\'{\i}}ctor M{\'{e}}ndez Mu{\-{n}}oz},
  title     = {{CLOSER 2015 - Proceedings of the 5th International Conference on
       Cloud Computing and Services Science, Lisbon, Portugal, 20-22 May,
       2015}},
  publisher = {SciTePress},
  year      = {2015},
  isbn      = {978-989-758-104-5},
  timestamp = {Tue, 04 Aug 2015 09:17:34 +0200},
  biburl    = {http://dblp.uni-trier.de/rec/bib/conf/closer/2015},
  bibsource = {dblp computer science bibliography, http://dblp.org}
}
```

So lets look at the entry and see how to get a better one for our purpose to combine them. When using jabref, you see optional and required fields, we want to add as many as possible, regardless if

optional or required, so Lets do that (I I write here in ASCII as easier to document:

```
@InProceedings{,
  author =   {},
  title =    {},
  OPTcrossref =  {},
  OPTkey =   {},
  OPTbooktitle = {},
  OPTyear =      {},
  OPTeditor =    {},
  OPTvolume =    {},
  OPTnumber =    {},
  OPTseries =    {},
  OPTpages =     {},
  OPTmonth =     {},
  OPTaddress =   {},
  OPTorganization = {},
  OPTpublisher = {},
  OPTnote =      {},
  OPTannote =    {},
  url = {}
}
```

So lets copy and fill out the **form** from our various searches:

```
@InProceedings{Wettinger2015any2api,
  author    = {Johannes Wettinger and
     Uwe Breitenb{\"{u}}cher and
     Frank Leymann},
  title     = {{ANY2API - Automated APIfication - Generating APIs for Executables
     to Ease their Integration and Orchestration for Cloud Application
     Deployment Automation}},
  booktitle = {{CLOSER 2015 - Proceedings of the 5th International Conference on
       Cloud Computing and Services Science}},
  year =     {2015},
  editor    = {Markus Helfert and
       Donald Ferguson and
       V{\'{\i}}ctor M{\'{e}}ndez Mu{\-{n}}oz},
  publisher = {SciTePress},
  isbn      = {978-989-758-104-5},
  pages = {475--486},
  month = {20-22 May},
  address =      {Lisbon, Portugal},
  doi       = {10.5220/0005472704750486},
  url ={https://pdfs.semanticscholar.org/1cd4/4b87be8cf68ea5c4c642d38678a7b40a86de.pdf},
  owner = {S17-IO-3005},
}
```

4.1.4 What are the differnt entry types and fields

We were asked what are the different entry types and fields, so we did a google query and found the following useful information. please remember that we also have fields such as doi, owner, we will add status ={pass/fail} at time of grading to indicate if the reference passes or fails. We may assign this to you so you get familiar with the identification if a refernce is ok or not.

Please see `https://en.wikipedia.org/wiki/BibTeX`

4.1.5 InProceedings

Please fill out

```
@InProceedings{,
  author =       {},
  title =        {},
  OPTcrossref =  {},
  OPTkey =       {},
  OPTbooktitle = {},
  OPTyear =      {},
  OPTeditor =    {},
  OPTvolume =    {},
  OPTnumber =    {},
  OPTseries =    {},
  OPTpages =     {},
  OPTmonth =     {},
  OPTaddress =   {},
  OPTorganization = {},
  OPTpublisher = {},
  OPTnote =      {},
  OPTannote =    {},
  url = {}
}
```

```
@inproceedings{vonLaszewski15tas,
  author =     {DeLeon, Robert L. and Furlani, Thomas R. and Gallo,
                Steven M. and White, Joseph P. and Jones, Matthew
                D. and Patra, Abani and Innus, Martins and Yearke,
                Thomas and Palmer, Jeffrey T. and Sperhac, Jeanette
                M. and Rathsam, Ryan and Simakov, Nikolay and von
                Laszewski, Gregor and Wang, Fugang},
  title =  {{TAS View of XSEDE Users and Usage}},
  booktitle =  {Proceedings of the 2015 XSEDE Conference: Scientific
                Advancements Enabled by Enhanced
                Cyberinfrastructure},
  series =     {XSEDE '15},
  year =   2015,
  isbn =   {978-1-4503-3720-5},
  location =   {St. Louis, Missouri},
  pages =  {21:1--21:8},
  articleno =  21,
```

```
  numpages =   8,
  url =        {http://doi.acm.org/10.1145/2792745.2792766},
  doi =        {10.1145/2792745.2792766},
  acmid =  2792766,
  publisher =  {ACM},
  address =    {New York, NY, USA},
  keywords =   {HPC, SUPReMM, TAS, XDMoD, XSEDE usage, XSEDE users},
}
```

4.1.6 TechReport

Please fill out

```
@TechReport{,
  author =       {},
  title =        {},
  institution =  {},
  year =         {},
  OPTkey =       {},
  OPTtype =      {},
  OPTnumber =    {},
  OPTaddress =   {},
  OPTmonth =     {},
  OPTnote =      {},
  OPTannote =    {},
  url = {}
}
```

```
@TechReport{las05exp,
  title =  {{The Java CoG Kit Experiment Manager}},
  Author =     {von Laszewski, Gregor},
  Institution =    {Argonne National Laboratory},
  Year =   2005,
  Month =  jun,
  Number =     {P1259},
  url = {https://laszewski.github.io/papers/vonLaszewski-exp.pdf}
}
```

4.1.7 Article

Please fill out

```
@Article{,
  author =       {},
  title =        {},
  journal =      {},
  year =         {},
  OPTkey =       {},
  OPTvolume =    {},
  OPTnumber =    {},
  OPTpages =     {},
```

```
  OPTmonth =      {},
  OPTnote =       {},
  OPTannote =     {},,
  url = {}
}
```

```
@Article{las05gridhistory,
  title =  {{The Grid-Idea and Its Evolution}},
  author =      {von Laszewski, Gregor},
  journal =     {Journal of Information Technology},
  year =   2005,
  month =  jun,
  number =      6,
  pages =  {319-329},
  volume =      47,
  doi =         {10.1524/itit.2005.47.6.319},
  url = {https://laszewski.github.io/papers/vonLaszewski-grid-idea.pdf}
}
```

4.1.8 Proceedings

Please fill out

```
@Proceedings{,
  title =         {},
  year =          {},
  OPTkey =        {},
  OPTbooktitle = {},
  OPTeditor =     {},
  OPTvolume =     {},
  OPTnumber =     {},
  OPTseries =     {},
  OPTaddress =    {},
  OPTmonth =      {},
  OPTorganization = {},
  OPTpublisher = {},
  OPTnote =       {},
  OPTannote =     {},
  url = {}
}
```

```
@Proceedings{las12fedcloud-proc,
  title =  {{FederatedClouds '12: Proceedings of the 2012
                 Workshop on Cloud Services, Federation, and the 8th
                 Open Cirrus Summit}},
  year =   2012,
  address =     {New York, NY, USA},
  editor =      {vonLaszewski, Gregor and Robert Grossman and Michael
                 Kozuchand Rick McGeerand Dejan Milojicic},
  publisher =  {ACM},
  iSBN =   {978-1-4503-1754-2},
```

```
  location =   {San Jose, California, USA},
  url =
                  {http://dl.acm.org/citation.cfm?id=2378975&picked=prox&cfid=389635474&cft
}
```

4.1.9 Wikipedia Entry

Please fill out

```
@Misc{,
  OPTkey =       {},
  OPTauthor =    {},
  OPTtitle =     {},
  OPThowpublished = {},
  OPTmonth =     {},
  OPTyear =      {},
  OPTnote =      {},
  OPTannote =    {},
  url = {}
}
```

```
@Misc{www-ode-wikipedia,
  Title =  {Apache ODE},
  HowPublished = {Web Page},
  Note =   {Accessed: 2017-2-11},
  Key =        {Apache ODE},
  Url =        {https://en.wikipedia.org/wiki/Apache_ODE}
}
```

4.1.10 Blogs

Please fill out

```
@Misc{,
  OPTkey =       {},
  OPTauthor =    {},
  OPTtitle =     {},
  OPThowpublished = {},
  OPTmonth =     {},
  OPTyear =      {},
  OPTnote =      {},
  OPTannote =    {},
  OPTurl = {}
}
```

```
@Misc{www-clarridge-discoproject-blog,
  title =  {Disco - A Powerful Erlang and Python Map/Reduce
                  Framework},
  uthor =  {Clarridge, Tait},
  howpublished = {Blog},
  month =  may,
```

```
  note =   {Accessed: 25-feb-2017},
  year =   2014,
  url =  {http://www.taitclarridge.com/techlog/2014/05/disco-a-powerful-erlang-and-python-m
}
```

4.1.11 Web Page

Please fill out

```
@Misc{,
  OPTkey =       {},
  OPTauthor =    {},
  OPTtitle =     {},
  OPThowpublished = {},
  OPTmonth =     {},
  OPTyear =      {},
  OPTnote =      {},
  OPTannote =    {},
  url = {}
}
```

```
@Misc{www-cloudmesh-classes,
  OPTkey =       {},
  author =    {von Laszewski, Gregor},
  title =     {Cloudmesh Classes},
  howpublished = {Web Page},
  OPTmonth =     {},
  OPTyear =      {},
  OPTnote =      {},
  OPTannote =    {},
  url = {https://cloudmesh.github.io/classes/}
}
```

```
@Misc{www-awslambda,
  title =  {AWS Lambda},
  author =     {{Amazon}},
  key =        {AWS Lambda},
  howpublished = {Web Page},
  url =        {https://aws.amazon.com/lambda/faqs/}
}
```

4.1.12 Book

Given the following entry. What is the proper entry for this book. Provide rationale:

```
@Book{netty-book,
    Title = {Netty in Action},
    Author = {Maurer, Norman and Wolfthal, Marvin},
    Publisher = {Manning Publications},
    Year = {2016},
}
```

To obtain the record of a book you can look at many information sources. The can include:

- https://www.manning.com/books/netty-in-action
- https://www.amazon.com/Netty-Action-Norman-Maurer/dp/1617291471
- http://www.barnesandnoble.com/w/netty-in-action-norman-maurer/1117342155?ean=9781617291470#productInfoTabs
- http://www.powells.com/book/netty-in-action-9781617291470/1-0

Furthermore, we need to consider the entry of a book, we simply look it up in emacs where we find the following but add the owner and the url field:

```
@Book{,
  ALTauthor =      {},
  ALTeditor =      {},
  title =      {},
  publisher =      {},
  year =   {},
  OPTkey =     {},
  OPTvolume =      {},
  OPTnumber =      {},
  OPTseries =      {},
  OPTaddress =     {},
  OPTedition =     {},
  OPTmonth =   {},
  OPTnote =    {},
  OPTannote =      {},
  ownwer =       {},
  url = {}
}
```

In summary we find the following fields:

Required fields: author/editor, title, publisher, year
Optional fields: volume/number, series, address, edition, month, note, key

We apply the following to fill out the fields.

address: The address is the Publisher's address. Usually just the city, but can be the full address for lesser-known publishers.

author: The name(s) of the author(s) (in the case of more than one author, separated by and) Names can be written in one of two forms: Donald E. Knuth or Knuth, Donald E. or van Halen, Eddie. Please note that Eddie van Halen would result in a wrong name. For our purpose we keep nobelity titles part of the last name.

edition: The edition of a book, long form (such as "First" or "Second")

editor: The name(s) of the editor(s)

key: A hidden field used for specifying or overriding the alphabetical order of entries (when the "author" and "editor" fields are missing). Note that this is very different from the key that is used to cite or cross-reference the entry.

label: The label field should contain three letters from the auth field, a short year reference and a short name of the publication to provide the maximum information regarding the publication. Underscores should be replaced with dashes or removed completely.

month: The month of publication or, if unpublished, the month of creation. Use three-letter abbreviations for this field in order to account for languages that do not capitalize month

names. Additional information for the day can be included as follows: aug #"~10,"

publisher: The publisher's name

series: The series of books the book was published in (e.g. "The Hardy Boys" or "Lecture Notes in Computer Science")

title: The title of the work. As the capitalization depends on the bibliography style and the language used we typically use camel case. To force capitalization of a word or its first letter you can use the curly braces, '{ }'. To keep the title in camel case simple use title = {{My Title}}

type: The field overriding the default type of publication (e.g. "Research Note" for techreport, "{PhD} dissertation" for phdthesis, "Section" for inbook/incollection) volume The volume of a journal or multi-volume book year The year of publication (or, if unpublished, the year of creation)

While applying the above rules and tips we summarize what we have done for this entry:

1. Search for the book by title/Author on ACM (`http://dl.acm.org/`) or Amazon or barnesandnoble or upcitemdb (`http://upcitemdb.com`). These services return bibtex entrie that you can improve.
2. Hence one option is t get the ISBN of the book. For "Mesos in action" from upcitemdb we got the ISBN as "9781617 292927". This is the 13 digit ISBN. The first 3 digits (GS1 code) can be skipped. Using the rest of 10 digits "1617 292927", Add in JabRef in Optional Fields->ISBN.
   However it is fine to youst specify the full number.
   We can also return a bibtex entry generated while using Click on the "Get BibTex from ISBN".
   Now we get more information on this book entry from ISBN. We can opt either the original or newly searched entry for the below bibtex fields or merge as appropriate. URL may not match from where we initially read the book, however there is option to put your original url or newly searched url. EAN, Edition, Pages,url,published date etc. Do a search on amazon for "ASIN". Can skip if not available. Sometime we get ASIN for a different publication, maybe a paperback ASIN={B01MT311CU} We can add it as it becomes easier to search

doi: If you can find a doi numer you should also add it. IN this case we could not locate one.

As a result we obtain the entry:

```
@Book{netty-book,
  title = {Netty in Action},
  publisher = {Manning Publications Co.},
  year = {2015},
  author = {Maurer, Norman and Wolfthal, Marvin Allen},
  address = {Greenwich, CT, USA},
  edition = {1st},
  isbn = {1617291471},
  asin = {1617291471},
  date = {2015-12-23},
  ean = {9781617291470},
  owner = {S17-IO-3022 S17-IO-3010 S17-IO-3012},
  pages = {296},
  url = {http://www.ebook.de/de/product/21687528/norman_maurer_netty_in_action.html},
}
```

4.2 Integrating Bibtex entries into Other Systems

We have not tested any of this

4.2.1 Bibtex import to MSWord

XML import

Plaease respond back to us if you have used this and give feedback.

1. In JabRef, export the bibliography in MS Word 2008 xml format
2. Name the file Sources.xml (case sensitive)
3. In OSX with MS Word 2015: Go to `/Library/Containers/com.microsoft.word/Data/Library/Applica`
4. Rename the original Sources.xml file to Sources.xml.bak
5. Copy the generated Sources.xml in this folder
6. Restart MS Word.

We do not know what needs to be done in case you need to make changes to the refernces. Please report back your experiences. To avoid issues we recommend that you use LaTeX. and not MSWord.

BibTex4Word

We have not tried this:

- `http://www.ee.ic.ac.uk/hp/staff/dmb/perl/index.html`

You are highly recommended to use Jabref for bibliography management in this class. Here is an introductory video on Jabref: `https://youtu.be/roi7vezNmfo?t=8m6s`

4.3 Other Reference Managers

Please note that you should first decide which reference manager you like to use. In case you for example install zotero and mendeley, that may not work with word or other programs.

4.3.1 Endnote

Endnote os a reference manager that works with Windows. Many people use Endnote. However, in the past, Endnote has caused complications when dealing with collaborative management of references. Its price is considerable. We have lost many hours of work because of instability of Endnote in some cases. As a student, you may be able to use Endnote for free at Indiana University.

- `http://endnote.com/`

4.3.2 Mendeley

Mendeley is a free reference manager compatible with Windows Word 2013, Mac Word 2011, LibreOffice, BibTeX. Videos on how to use it are available at:

- `https://community.mendeley.com/guides/videos`

Installation instructions are available at

- `https://www.mendeley.com/features/reference-manager/`

When dealing with large databases, we found the integration of Mendeley into word slow.

4.3.3 Zotero

Zotero is a free tool to help you collect, organize, cite, and share your research sources. Documentation is available at

- `https://www.zotero.org/support/`

The download link is available from

- `https://www.zotero.org/`

We have limited experience with Zotero

5. Editors

F https://github.com/cloudmesh/classes/blob/master/docs/source/lesson/doc/emacs.rst

5.1 Basic Emacs

One of the most useful short manuals for emacs is the following refrence card. It takes some time to use this card efficiently, but the most important commands are written on it. Generations of students have litterally been just presented with this card and they learned emacs from it.

- https://www.gnu.org/software/emacs/refcards/pdf/refcard.pdf

There is naturally also additional material available and a great manual. You could also look at

- https://www.gnu.org/software/emacs/tour/

From the last page we have summarized the most useful and **simple** features. And present them here. One of the hidden gems of emacs is the ability to recreate replay able macros which we include here also. You ought to try it and you will find that for data science and the cleanup of data emacs (applied to smaller datasets) is a gem.

Notation

Key	Description
C	Control
M	Esc (meta character)

In the event of an emergency...

Here's what to do if you've accidentally pressed a wrong key:

If you executed a command and Emacs has modified your buffer, use C-/ to undo that change. If you pressed a prefix key (e.g. C-x) or you invoked a command which is now prompting you for input (e.g. Find file: ...), type C-g, repeatedly if necessary, to cancel. C-g also cancels a long-running operation if it appears that Emacs has frozen.

Moving around in buffers can be done with cursor keys, or with the following key combinations:

Key	Description
C-f Forw	ard one character
C-n Next	line
C-b Back	one character
C-p Prev	ious line

Here are some ways to move around in larger increments:

Key	Description
C-a Begi	nning of line
M-f Forw	ard one word
M-a Prev	ious sentence
M-v Prev	ious screen
M-< Begi	nning of buffer
C-e End	of line
M-b Back	one word
M-e Next	sentence
C-v Next	screen
M-> End	of buffer

You can jump directly to a particular line number in a buffer:

Key	Description
M-g g	Jump to specified line

Searching is easy with the following commands

Key	Description
C-s Incr	emental search forward
C-r Incr	emental search backward

Replace

Key	Description
M-% Quer	y replace

Killing (“cutting”) text

Key	Description
C-k Kill	line

Yanking

Key	Description
C-y Yank	s last killed text

Macros

Keyboard Macros

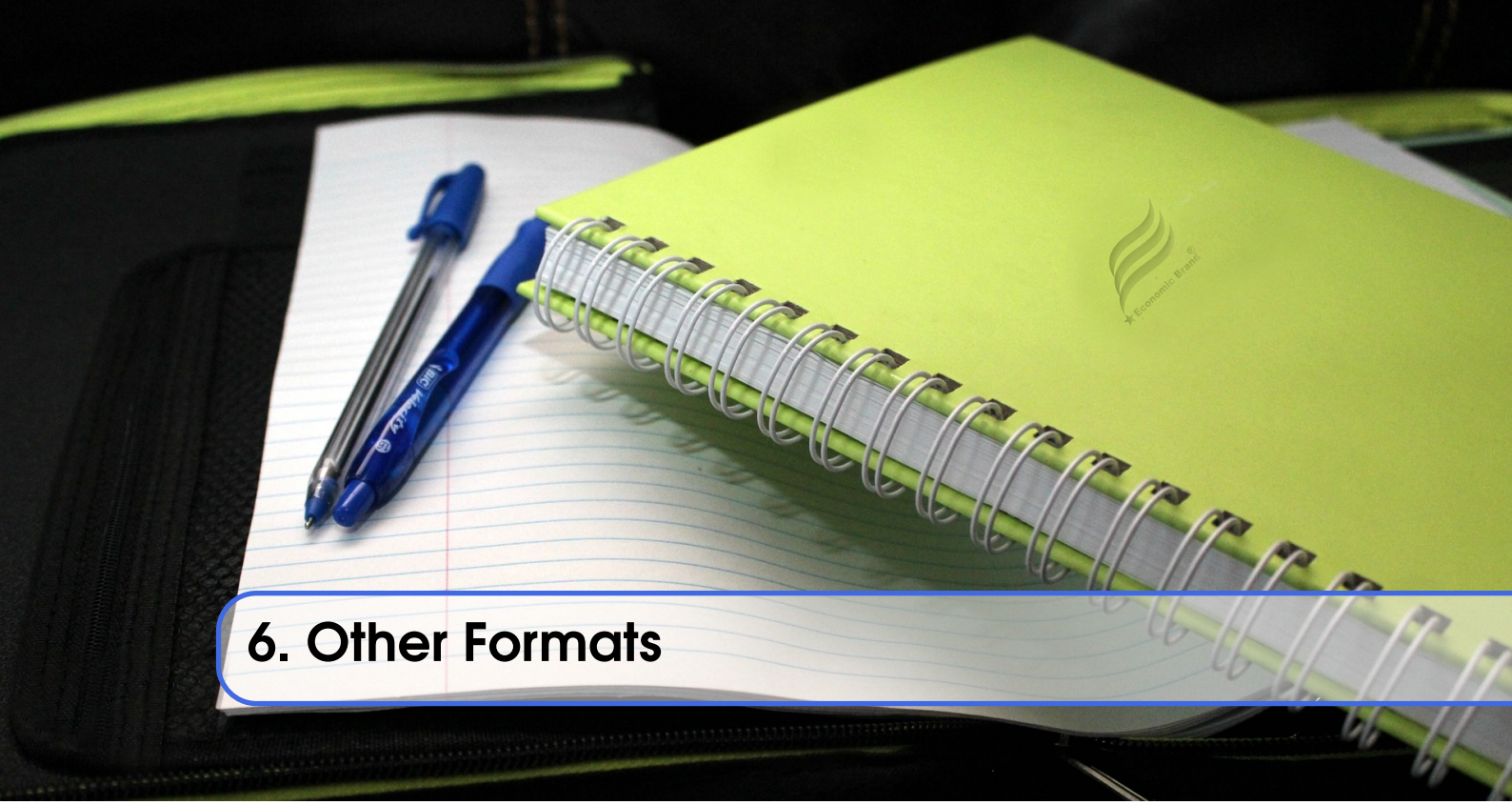
Keyboard macros are a way to remember a fixed sequence of keys for later repetition. They're handy for automating some boring editing tasks.

Key	Description
M-x (	Start recording macro
M-x)	Stop recording macro
M-x e	Play back macro once
M-5 M-x-e	Play back macro 5 times

Modes

"Every buffer has an associated major mode, which alters certain behaviors, key bindings, and text display in that buffer. The idea is to customize the appearance and features available based on the contents of the buffer." modes are typically activated by ending such as .py, .java, .rst, ...

Key	Description
M-x python-mode	Mode for editing Python files
M-x auto-fill-mode	Wraps your lines automatically when they get longer than 70 characters.
M-x flyspell-mode	Highlights misspelled words as you type.



6. Other Formats

F https://github.com/cloudmesh/classes/blob/master/docs/source/lesson/doc/rst.rst

6.1 reStructuredText

reStructuredText (RST) pur[pose is to provide an easy-to-read, what-you-see-is-what-you-get plaintext markup syntax and parser system. With its help you can develop documentation not only for stand aone documentation, simple web pages, an in-line program documentation (such as Python). RST is extensible and new features can be added. It is used in sphinx as one of its supported formats.

6.1.1 Links

- RST Sphinx documentation: http://www.sphinx-doc.org/en/stable/rest.html
- RST Syntax: http://docutils.sourceforge.net/rst.html
- Important extensions: http://sphinx-doc.org/ext/todo.html

Cheatcheat:
- http://github.com/ralsina/rst-cheatsheet/raw/master/rst-cheatsheet.pdf
- http://docutils.sourceforge.net/docs/ref/rst/directives.html

6.1.2 Source

The source for this page is located at

- https://raw.githubusercontent.com/cloudmesh/classes/master/docs/source/lesson/doc/rst.rst

This way you can look at the source on how we create this page.

6.1.3 Sections

with overline, for parts * with overline, for chapters =, for sections -, for subsections ^, for subsubsections ", for paragraphs

RST allows to specify a number of sections. You can do this with the various underlines:

```
**********
Chapter
**********
Section
==========
Subsection
----------
Subsubsection
^^^^^^^^^^
Paragraph
~~~~~~~~~~
```

6.1.4 Listtable

```
.. csv-table:: Eye colors
   :header: "Name", "Firstname", "eyes"
   :widths: 20, 20, 10

   "von Laszewski", "Gregor", "gray"
```

6.1.5 Exceltable

we have integrated Excel table from `http://pythonhosted.org//sphinxcontrib-exceltable/` intou our sphinx allowing the definition of more elaborate tables specified in excel. Howere the most convenient way may be to use list-tables. The documentation to list tables can be found at `http://docutils.sourceforge.net/docs/ref/rst/directives.html#list-table`

6.1.6 Boxes

Seealso

```
.. seealso:: This is a simple **seealso** note.
```

Note

This is a **note** box.

```
.. note::  This is a **note** box.
```

Warning

note the space between the directive and the text

```
.. warning:: note the space between the directive and the text
```

Others

This is an **attention** box.

```
.. attention:: This is an **attention** box.
```

This is a **caution** box.

```
.. caution:: This is a **caution** box.
```

This is a **danger** box.

```
.. danger:: This is a **danger** box.
```

This is a **error** box.

```
.. error:: This is a **error** box.
```

This is a **hint** box.

```
.. hint:: This is a **hint** box.
```

This is an **important** box.

```
.. important:: This is an **important** box.
```

This is a **tip** box.

```
.. tip:: This is a **tip** box.
```

6.1.7 Sidebar directive

It is possible to create sidebar using the following code:

```
.. sidebar:: Sidebar Title
    :subtitle: Optional Sidebar Subtitle

    Subsequent indented lines comprise
    the body of the sidebar, and are
    interpreted as body elements.
```

Sidebar Title: Optional Sidebar Subtitle

Subsequent indented lines comprise the body of the sidebar, and are interpreted as body elements.

6.1.8 Sphinx Prompt

```
.. prompt:: bash, cloudmesh$

   wget -O cm-setup.sh http://bit.ly/cloudmesh-client-xenial
   sh cm-setup.sh
```

6.1.9 Programm examples

You can include code examples and bash commands with two colons.

This is an example for python:

```
print ("Hallo World")
```

This is an example for a shell command:

```
$ ls -lisa
```

6.1.10 Hyperlinks

Direct links to html pages can ve done with:

```
‘This is a link to an html page <hadoop.html>‘_
```

Note that this page could be generated from an rst page

Links to the FG portal need to be formulated with the portal tag:

```
:portal:‘List to FG projects </projects/all>‘
```

In case a subsection has a link declared you can use :ref: (this is the prefered way as it can be used to point even to subsections:

```
:ref:‘Connecting private network VMs  clusters <_s_vpn>‘
```

A html link can be created anywhere in the document but must be unique. for example if you place:

```
.. _s_vpn:
```

in the text it will create a target to which the above link points when you click on it

6.1.11 Todo

```
.. todo:: an example
```

Todo
an example

F `https://github.com/cloudmesh/classes/blob/master/docs/source/lesson/doc/markdown.rst`

6.2 Markdown

TBD. Section about Markdown

see: `https://en.wikipedia.org/wiki/Markdown`

F `https://github.com/cloudmesh/classes/blob/master/docs/source/lesson/doc/type.rst`

6.3 Communicating Research in Other Ways

Naturally, writing papers is not the only way to communicate your research with others. We find that today we see additional pathways for cumminication includine blogs, twitter, facebook, e-mail, Web pages, and electronic notebooks. Let us refisit some of them and identify when they are helpful.

6.3.1 Blogs

blog: noun, a regularly updated website or web page, typically one run by an individual or small group, that is written in an informal or conversational style.

Advantages:

- encourages spontaneous posts
- encourages small short contributions
- chronologically ordered
- standard software exists to set up blogs
- online services exists to set up blogs

Disadvantages:

- structuring data is difficult (some blog software support it)
- not suitable for formal development of a paper
- often lack of sophisticated track change features
- no collaborative editing features

6.3.2 Sphinx

Sphinx (`http://www.sphinx-doc.org/`) is a tool that to create integrated documentation from a markup language whlie.

Advantages:

- output formats: html, LaTeX, PDF, ePub
- integrates well with directory structure
- powerful markup language (reStructuredText)
- can be hosted on github via github pages
- can integare other renderers such as Markdown
- automatic table of content, tebale of index
- code documentation integration
- search
- written in python and using bash, so extensions and custom automation are possible

Disadvantage:

- requires compile step
- When using markdown github can render individual page

Others:

- Read the Docs (`https://readthedocs.org/`)
- Doxygen (`http://www.stack.nl/~dimitri/doxygen/`)
- MkDocs (`http://www.mkdocs.org/`)

6.3.3 Notebooks

Jupyter

The Jupyter Notebook (`http://jupyter.org/`) is an open-source web application allowing users to create and share documents that contain live code, equations, visualizations and explanatory text. Use cases include data cleaning and transformation, numerical simulation, statistical modeling, machine learning.

Advantages:

- Integrates with python
- Recently other programming languages have been integrated
- Allows experimenting with settings
- Allows a form of literate programming while mixing documentation with code
- automatically renders on github
- comes with web service that allows hosting

Disadvantage:

- mostly encourages short documents
- mark up language is limited
- editing in ASCII is complex and Web editing is prefered

Apache Zeppilin

A Web-based notebook that enables data-driven, interactive data analytics and collaborative documents with SQL, Scala and hadoop. It integrates a web-based notebook with data ingestion, data exploration, visualization, sharing and collaboration features to Hadoop and Spark.

Advantages:

- integration to various framework
- Web framework
- integration with spark, hadoop

Disadvantages:

- larger framework
- must leverages existing deployments of spak, hadoop

III Big Data Applications

F chapter/theory/introduction.tex

6.4 Introduction

You may find that some videos may have a different lesson, section or unit number. Please ignore this. In case the content does not correspond to the title, please let us know.

This section has a technical overview of course followed by a broad motivation for course hosted at www-cloudmesh-classes.

The course overview covers it's content and structure. It presents an introduction to general field of Big Data and Analytics. We are especially analysing the many different application areas in which Big Data can be applied. As Big Datais typically not just used in isolation but is part of a larger Informatics issue for a particular field we also use the term X-Informatics, where X defines a usecase or area of specialization in which Big Data is applied to. As such we organize the class around the the *Rallying Cry* of course: Use Clouds running Data Analytics Collaboratively processing Big Data to solve problems in X-Informatics.

The courses is set up as a number of lessons that are typically between 20 minutes to an hour. The lessons are either provided as written documents or as video lectures. They are enhanced by an in person meeting that takes place either in a lecture room for residential students or as online meeting for online students.

The course covers a mix of applications (the X in X-Informatics) and technologies needed to support the field electronically i.e. to process the application data. The overview ends with a discussion of course content at highest level. The course starts with a motivation summarizing clouds and data science, then units describing applications in areas such as Physics, e-Commerce, Web Search and Text mining, Health, Sensors and Remote Sensing). These are interspersed with discussions of infrastructure (clouds) and data analytics (algorithms like clustering and collaborative filtering used in applications). The course uses Python as primary programming language. We will be introducing practical use of cloud resources so that you have the oportunity to explore example analytics applications on smaller data sets that you define.

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

We discuss in this course include the following topics. We may change the order of the topics to allow for maximal flexibility and parallel learning experiences.

Writing Track:

- Writing a short review article
- Writing a porject or term report

Theory Track:

- Motivation: Big Data and the Cloud; Centerpieces of the Future Economy
- Introduction: What is Big Data, Data Analytics
- Use Cases: Big Data Use Cases Survey

  - Use Case, Physics Discovery of Higgs Particle
  - Use Case: e-Commerce and Lifestyle with recommender systems
  - Use Case: Web Search and Text Mining and their technologies
  - Use Case: Sports
  - Use Case: Health
  - Use Case: Sensors
  - Use Case: Radar for Remote Sensing.
- Parallel Computing Overview and familiar examples
- Cloud Technology for Big Data Applications & Analytics

Practice Track:

- Python for Big Data Applications and Analytics: NumPy, SciPy, MatPlotlib
- Using FutureGrid for Big Data Applications and Analytics Course
- Using Chameleon Cloud for Big Data Applications and Analytics Course
- [optional] Using Plotviz Software for Displaying Point Distributions in 3D
- Recommender Systems - K-Nearest Neighbors, Clustering and heuristic methods
- PageRank
- Kmeans
- MapReduce
- Kmeans and MapReduce Parallelism

6.4.1 Course Motivation

We motivate the study of X-informatics by describing data science and clouds. He starts with striking examples of the data deluge with examples from research, business and the consumer. The growing number of jobs in data science is highlighted. He describes industry trend in both clouds and big data.

He introduces the cloud computing model developed at amazing speed by industry. The 4 paradigms of scientific research are described with growing importance of data oriented version. He covers 3 major X-informatics areas: Physics, e-Commerce and Web Search followed by a broad discussion of cloud applications. Parallel computing in general and particular features of MapReduce are described. He comments on a data science education and the benefits of using MOOC's.

Emerging Technologies

This presents the overview of talk, some trends in computing and data and jobs. Gartner's emerging technology hype cycle shows many areas of Clouds and Big Data. We highlight 6 issues of importance: economic imperative, computing model, research model, Opportunities in advancing computing, Opportunities in X-Informatics, Data Science Education

Motivation (40:14) 

Motivation (30) 

Data Deluge

We give some amazing statistics for total storage; uploaded video and uploaded photos; the social media interactions every minute; aspects of the business big data tidal wave; monitors of aircraft

engines; the science research data sizes from particle physics to astronomy and earth science; genes sequenced; and finally the long tail of science. The next slide emphasizes applications using algorithms on clouds. This leads to the rallying cry "Use Clouds running Data Analytics Collaboratively processing Big Data to solve problems in X-Informatics educated in data science' 'with a catalog of the many values of X"Astronomy, Biology, Biomedicine, Business, Chemistry, Climate, Crisis, Earth Science, Energy, Environment, Finance, Health, Intelligence, Lifestyle, Marketing, Medicine, Pathology, Policy, Radar, Security, Sensor, Social, Sustainability, Wealth and Wellness"

Data Deluge (30:38) 

Data Deluge (20) 

Jobs

Jobs abound in clouds and data science. There are documented shortages in data science, computer science and the major tech companies advertise for new talent.

Jobs (9:39) 

Jobs (8) 

Industrial Trends

Trends include the growing importance of mobile devices and comparative decrease in desktop access, the export of internet content, the change in dominant client operating systems, use of social media, thriving Chinese internet companies.

Industrial Trends (19:25) 

Industrial Trends (16) 

Industrial Trends II (16:54) 

Indusrial Trends II (16) 

Indusrial Trends III (30:13) 

Industrial Trends III (21) 

Digital Disruption of Old Favorites

Not everything goes up. The rise of the Internet has led to declines in some traditional areas including Shopping malls and Postal Services.

Digital Distruption and transformation (32:54) 

Digital Distruption and transformation (28) 

Computing Model

Industry adopted clouds which are attractive for data analytics

Clouds and Big Data are transformational on a 2-5 year time scale. Already Amazon AWS is a lucrative business with almost a $4B revenue. We describe the nature of cloud centers with economies of scale and gives examples of importance of virtualization in server consolidation. Then key characteristics of clouds are reviewed with expected high growth in Infrastructure, Platform and Software as a Service.

Computing Model I (24:03) 

Computing Model I (14) 

Computing Model II (28:18) 

Computing Model II (27) 

Research Model

4th Paradigm; From Theory to Data driven science?

We introduce the 4 paradigms of scientific research with the focus on the new fourth data driven methodology.

Research Model (7:33) 

Research Model (4) 

Data Science Process

We introduce the DIKW data to information to knowledge to wisdom paradigm. Data flows through cloud services transforming itself and emerging as new information to input into other transformations.

Data Science Process (15:42) 

Data Science Process (10) 

Physics-Informatics

Looking for Higgs Particle with Large Hadron Collider LHC

We look at important particle physics example where the Large hadron Collider has observed the Higgs Boson. He shows this discovery as a bump in a histogram; something that so amazed him 50 years ago that he got a PhD in this field. He left field partly due to the incredible size of author lists on papers.

Physics-informatics (13:27) 

Physics-inforamtics (6) 

Recommender Systems

Many important applications involve matching users, web pages, jobs, movies, books, events etc. These are all optimization problems with recommender systems one important way of performing this optimization. We go through the example of Netflix ~~ everything is a recommendation and muses about the power of viewing all sorts of things as items in a bag or more abstractly some space with funny properties.

Recommender Systems I (12:21) 

Recommender Systems I (9) 

Recommender Systems II (9:44) 

Recommender Systems II (6) 

Web Search and Information Retrieval

This course also looks at Web Search and here we give an overview of the data analytics for web search, Pagerank as a method of ranking web pages returned and uses material from Yahoo on the subtle algorithms for dynamic personalized choice of material for web pages.

Web Search and Information Retrieval (12:05) 

Web Search and Information Retrieval (6) 

Cloud Application in Research

We describe scientific applications and how they map onto clouds, supercomputers, grids and high throughput systems. He likes the cloud use of the Internet of Things and gives examples.

Cloud Applications in Research (33:51) 

Cloud Applications in Research (20) 

Parallel Computing and MapReduce

We define MapReduce and gives a homely example from fruit blending.

Computing and MapReduce (14:02) 

Computing and MapReduce (9) 

Data Science Education

We discuss one reason you are taking this course ~~ Data Science as an educational initiative and aspects of its Indiana University implementation. Then general; features of online education are discussed with clear growth spearheaded by MOOC's where we use this course and others as an example. He stresses the choice between one class to 100,000 students or 2,000 classes to 50 students and an online library of MOOC lessons. In olden days he suggested ''hermit's cage virtual university" ~~ gurus in isolated caves putting together exciting curricula outside the traditional university model. Grading and mentoring models and important online tools are discussed. Clouds have MOOC's describing them and MOOC's are stored in clouds; a pleasing symmetry.

Data Science Education (28:08) 

Data Science Education (19) 

Conclusions

The conclusions highlight clouds, data-intensive methodology, employment, data science, MOOC's and never forget the Big Data ecosystem in one sentence "Use Clouds running Data Analytics Collaboratively processing Big Data to solve problems in X-Informatics educated in data science"

Conclusions (4:59) 

Conclusions (4) 

Resources

- `http://www.gartner.com/technology/home.jsp` and many web links
- Meeker/Wu May 29 2013 Internet Trends D11 Conference `http://www.slideshare.net/kleinerperkins/kpcb-internet-trends-2013`
- `http://cs.metrostate.edu/~sbd/slides/Sun.pdf`
- Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics, Bill Franks Wiley ISBN: 978-1-118-20878-6
- Bill Ruh `http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html`
- `http://www.genome.gov/sequencingcosts/`

- 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://www.microsoft.com/en-us/news/features/2012/mar12/03-05CloudComputingJobs.aspx
- http://www.mckinsey.com/mgi/publications/big_data/index.asp
- Tom Davenport http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html
- http://research.microsoft.com/en-us/people/barga/sc09_cloudcomp_tutorial.pdf
- http://research.microsoft.com/pubs/78813/AJ18_EN.pdf
- http://www.google.com/green/pdfs/google-green-computing.pdf
- http://www.wired.com/wired/issue/16-07
- http://research.microsoft.com/en-us/collaboration/fourthparadigm/
- Jeff Hammerbacher http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf
- http://grids.ucs.indiana.edu/ptliupages/publications/Where%20does%20all%20the%20data%20come%20from%20v7.pdf
- http://www.interactions.org/cms/?pid=1032811
- http://www.quantumdiaries.org/2012/09/07/why-particle-detectors-need-a-trigger/atlasmgg/
- http://www.sciencedirect.com/science/article/pii/S037026931200857X
- http://www.slideshare.net/xamat/building-largescale-realworld-recommender-systems-recs
- http://www.ifi.uzh.ch/ce/teaching/spring2012/16-Recommender-Systems_Slides.pdf
- http://en.wikipedia.org/wiki/PageRank
- http://pages.cs.wisc.edu/~beechung/icml11-tutorial/
- https://sites.google.com/site/opensourceiotcloud/
- http://datascience101.wordpress.com/2013/04/13/new-york-times-data-science-articles/
- http://blog.coursera.org/post/49750392396/on-the-topic-of-boredom
- http://x-informatics.appspot.com/course
- http://iucloudsummerschool.appspot.com/preview
- https://www.youtube.com/watch?v=M3jcSCA9_hM

F chapter/theory/overview.tex

6.5 Overview of Data Science

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

The course introduction starts 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.

6.5.1 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.

TBD (45) 

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 introduce the motto of the course: Use Clouds running Data Analytics Collaboratively processing Big Data to solve problems in X-Informatics. We list many values of X you can defined in various activities across the world.

TBD (9:49) 

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.

TBD (2:58) 

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.

TBD (13:04) 

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.

TBD (4:27) 

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.

TBD (3:42) 

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.

TBD (6:00) 

TBD (7:34) 

TBD (9:37) 

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 `http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html`
- Anjul Bhambhri `http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html`
- Jeff Hammerbacher `http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf`
- `http://www.economist.com/node/15579717`
- `http://cs.metrostate.edu/~sbd/slides/Sun.pdf`
- `http://jess3.com/geosocial-universe-2/`
- Bill Ruh `http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html`
- `http://www.hsph.harvard.edu/ncb2011/files/ncb2011-z03-rodriguez.pptx`
- Hugh Williams `http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html`

6.5.2 Data Deluge and Scientific Applications and Methodology

Overview

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.

TBD (22)  PDF

Science & 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.

TBD (11:27) 

TBD (11:49) 

Implications for Scientific Method

We discuss the emergences 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.

TBD (5:07) 

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 don't generate a lot of data each but together they add up to Big data.

TBD (2:10) 

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.

TBD (5:45) 

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_Computing_FoxGannonv2.pdf
- http://grids.ucs.indiana.edu/ptliupages/publications/Where%20does%20all%20the%20data%20come%20from%20v7.pdf
- 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 https://sites.google.com/site/opensourceiotcloud/

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

Overview

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.

TBD (35) PDF

6.5.4 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.

TBD (16:04) MP4

Features of Data Deluge I

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

TBD (8:02) 

TBD (6:24) 

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.

TBD (11:28) 

Data Analytics

TBD (30) 

We stress the importance of data analytics giving 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.

TBD (7:28) 

TBD (6:51) 

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 Wolskihttp://research.microsoft.com/en-us/people/barga/sc09_cloud
- http://www.datacenterknowledge.com/archives/2011/05/10/uptime-institute-the-average-pu
- http://loosebolts.wordpress.com/2008/12/02/our-vision-for-generation-4-modular-data-c
- http://www.mediafire.com/file/zzqna34282frr2f/koomeydatacenterelectuse2011finalversion pdf
- Bina Ramamurthy http://www.cse.buffalo.edu/~bina/cse487/fall2011/
- Jeff Hammerbacher http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf
- Jeff Hammerbacher http://berkeleydatascience.files.wordpress.com/2012/01/20120119berkeley.pdf
- Anjul Bhambhri http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html
- http://cs.metrostate.edu/~sbd/slides/Sun.pdf
- Hugh Williams http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html
- Tom Davenport http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html
- http://www.mckinsey.com/mgi/publications/big_data/index.asp

- http://cra.org/ccc/docs/nitrdsymposium/pdfs/keyes.pdf

F chapter/theory/health.tex

6.6 Health Informatics Case Study

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.

6.6.1 X-Informatics Case Study: Health Informatics

Overview

131 (Health) 

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.

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) 

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) 

Telemedicine (Virtual Health)

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

Telemedicine (8:21) 

Big Data and Healthcare Industry

Summary of an industry survey by GE and Accenture.

Big Data and Healthcare Indusry (10:02) 

Medical Big Data in the Clouds

An impressive exemplar Cloud-based medicine system from Potsdam.

Medical Big Data in the Clouds (15:02) 

Medical image Big Data

Midical Image Big Data (6:33) 

Clouds and Health

Clouds and Health (4:35) 

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) 

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) 

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) 

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) 

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) 

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) 

CC) 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) 

131 (Proteomics and Information Visualization) 

Resources

- https://wiki.nci.nih.gov/display/CIP/CIP+Survey+of+Biomedical+Imaging+Archives
- http://grids.ucs.indiana.edu/ptliupages/publications/Where%20does%20all%20the%20data%20come%20from%20v7.pdf
- http://www.ieee-icsc.org/ICSC2010/Tony%20Hey%20-%2020100923.pdf
- http://quantifiedself.com/larry-smarr/
- http://www.ebi.ac.uk/Information/Brochures/
- http://www.kpcb.com/internet-trends
- http://www.slideshare.net/drsteventucker/wearable-health-fitness-trackers-and-the-qua
- http://www.siam.org/meetings/sdm13/sun.pdf
- http://en.wikipedia.org/wiki/Calico_%28company%29
- http://www.slideshare.net/GSW_Worldwide/2015-health-trends
- http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture-Industrial-Internet-Cha pdf
- http://www.slideshare.net/schappy/how-realtime-analysis-turns-big-medical-data-into-p
- http://medcitynews.com/2013/03/the-body-in-bytes-medical-images-as-a-source-of-health
- http://healthinformatics.wikispaces.com/file/view/cloud_computing.ppt
- http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights/Health%20care/The%20big-data%20revolution%20in%20US%20health%20care/The%20big-data%20revolution%20in%20US%20health%20care%20Accelerating%20value%20and%20innovation.ashx
- https://partner.microsoft.com/download/global/40193764
- http://ec.europa.eu/information_society/activities/health/docs/policy/taskforce/redesigning_health-eu-for2020-ehtf-report2012.pdf
- http://www.kpcb.com/internet-trends
- http://www.liveathos.com/apparel/app
- http://debategraph.org/Poster.aspx?aID=77
- http://www.oerc.ox.ac.uk/downloads/presentations-from-events/microsoftworkshop/gannon
- http://www.delsall.org
- http://salsahpc.indiana.edu/millionseq/mina/16SrRNA_index.html

- http://www.geatbx.com/docu/fcnindex-01.html
- https://wiki.nci.nih.gov/display/CIP/CIP+Survey+of+Biomedical+Imaging+Archives
- http://grids.ucs.indiana.edu/ptliupages/publications/Where%20does%20all%20the%20data%20come%20from%20v7.pdf
- http://www.ieee-icsc.org/ICSC2010/Tony%20Hey%20-%2020100923.pdf
- http://quantifiedself.com/larry-smarr/
- http://www.ebi.ac.uk/Information/Brochures/
- http://www.kpcb.com/internet-trends
- http://www.slideshare.net/drsteventucker/wearable-health-fitness-trackers-and-the-qua
- http://www.siam.org/meetings/sdm13/sun.pdf
- http://en.wikipedia.org/wiki/Calico_%28company%29
- http://www.slideshare.net/GSW_Worldwide/2015-health-trends
- http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture-Industrial-Internet-Cha pdf
- http://www.slideshare.net/schappy/how-realtime-analysis-turns-big-medical-data-into-p
- http://medcitynews.com/2013/03/the-body-in-bytes-medical-images-as-a-source-of-health
- http://healthinformatics.wikispaces.com/file/view/cloud_computing.ppt
- http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights/Health%20care/The%20big-data%20revolution%20in%20US%20health%20care/The%20big-data%20revolution%20in%20US%20health%20care%20Accelerating%20value%20and%20innovation.ashx
- https://partner.microsoft.com/download/global/40193764
- http://ec.europa.eu/information_society/activities/health/docs/policy/taskforce/redesigning_health-eu-for2020-ehtf-report2012.pdf
- http://www.kpcb.com/internet-trends
- http://www.liveathos.com/apparel/app
- http://debategraph.org/Poster.aspx?aID=77
- http://www.oerc.ox.ac.uk/downloads/presentations-from-events/microsoftworkshop/gannon
- http://www.delsall.org
- http://salsahpc.indiana.edu/millionseq/mina/16SrRNA_index.html
- http://www.geatbx.com/docu/fcnindex-01.html

F chapter/theory/lifestyle.tex

6.7 e-Commerce and LifeStyle Case Study

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.

6.7.1 Recommender Systems: Introduction

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.

45 (Recommender)  PDF

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) 

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 aren't 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) 

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) 

Examples of Recommender Systems

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

Examples of Recommender Systems (1:00) 

Netflix on Recommender Systems

This is Part 1.

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) 

Consumer Data Science

Here 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) 

Resources

- http://www.slideshare.net/xamat/building-largescale-realworld-recommender-systems-recs
- http://www.ifi.uzh.ch/ce/teaching/spring2012/16-Recommender-Systems_Slides.pdf
- https://www.kaggle.com/
- http://www.ics.uci.edu/~welling/teaching/CS77Bwinter12/CS77B_w12.html
- Jeff Hammerbacher https://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf
- http://www.techworld.com/news/apps/netflix-foretells-house-of-cards-success-with-cass

- https://en.wikipedia.org/wiki/A/B_testing
- http://www.infoq.com/presentations/Netflix-Architecture

6.7.2 Recommender Systems: Examples and Algorithms

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.

49 (Recommender) PDF

Recap and Examples of Recommender Systems

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)

Examples of Recommender Systems

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

Examples of Recommender Systems (8:34)

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)

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) 

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) 

Resources

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

6.7.3 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.

18 (Filtering)  PDF

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) 

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) 

F chapter/theory/physics.tex

6.8 Physics Case Study

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.

6.8.1 Looking for Higgs Particles, Bumps in Histograms, Experiments and Accelerators (Part 1)

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.

20 (Higgs) 

Files: HiggsClassI-Sloping.py </files/python/physics/mr_higgs/higgs_classI_sloping.py>_

Looking for Higgs Particle and Counting Introduction

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) 

Physics-Informatics Looking for Higgs Particle Experiments

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) 

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) 

Resources

- `http://grids.ucs.indiana.edu/ptliupages/publications/Where%20does%20all%20the%20data%20come%20from%20v7.pdf`
- `http://www.sciencedirect.com/science/article/pii/S037026931200857X`
- `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.

29 (Higgs II)  PDF

Files:

- HiggsClassI-Sloping.py </files/python/physics/mr_higgs/higgs_classI_sloping.py>
- HiggsClassIII.py </files/python/physics/number_theory/higgs_classIII.py>
- HiggsClassIIUniform.py </files/python/physics/mr_higgs/higgs_classII_uniform.py>

Physics Use Case II 1: Class Software

We discuss how this unit uses Java and Python on both a backend server (FutureGrid) or a local client. WE point out useful book on Python for data analysis. This builds on technology training in Section 3.

Higgs Particle Events and Counting (9:30) 

This video contains Java information, but we are no longer using Java in this class.

Physics Use Case II 2: 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) 

Physics Use Case II 3: With Python examples of Signal plus Background

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) 

Physics Use Case II 4: Change shape of background & num of Higgs Particles

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) 

Resources

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

6.8.2 Looking for Higgs Particles: 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.

39 (Higgs) 

HiggsClassIII.py </files/python/physics/number_theory/higgs_classIII.py>

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) 

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) 

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) 

Gaussian Distributions

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

Gaussian Distributions (9:08) 

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) 

Resources

- `http://indico.cern.ch/event/20453/session/6/contribution/15?materialId=slides`
- `http://www.atlas.ch/photos/events.html`
- `https://cms.cern/`

6.8.3 Looking for Higgs Particles: Random Numbers, Distributions and Central Limit Theorem (Part 3)

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.

44 (Higgs III) PDF

Files:

- HiggsClassIII.py </files/python/physics/calculated_dice_roll/higgs_classIV_seeds.py>

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)

Binomial Distribution

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

Binomial Distribution: (12:38)

Accept-Reject

We introduce an advanced method **accept/reject** for generating random variables with arbitrary distributions.

Accept-Reject (5:54)

Monte Carlo Method

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

Monte Carlo Method (2:23)

Poisson Distribution

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

Poisson Distribution (4:37)

Central Limit Theorem

We introduce Central Limit Theorem and give examples from Wikipedia.

Central Limit Theorem (4:47)

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) 

Resources

..bibliography:: physics-references.bib

F chapter/theory/radar.tex

6.9 Radar Case Study

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.

58 (Radar) 

6.9.1 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) 

6.9.2 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) 

6.9.3 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) 

6.9.4 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) 

6.9.5 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) 

6.9.6 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) 

F chapter/theory/sensor.tex

6.10 Sensors Case Study

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.

31 (Sensor I) 

44 (Sensor II) 

6.10.1 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) 

6.10.2 Robotics and IOT Expectations

Examples of Robots and Drones.

Robotics and IoT Expectations (8:05) 

6.10.3 Industrial Internet of Things

We summarize surveys and expectations Industry wide.

Industrial Internet of Things (1:24:02) 

6.10.4 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) 

6.10.5 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) 

6.10.6 Ubiquitous/Smart Cities

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

Ubiquitous/Smart Cities (1:44) 

6.10.7 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) 

6.10.8 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) 

6.10.9 Resources

- 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/
- http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture-Industrial-Internet-Cha pdf
- http://www.gesoftware.com/ge-predictivity-infographic
- http://www.getransportation.com/railconnect360/rail-landscape
- http://www.gesoftware.com/sites/default/files/GE-Software-Modernizing-Machine-to-Machi pdf

F chapter/theory/sport.tex

6.11 Sports Case Study

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.

6.11.1 Sports Informatics I : Sabermetrics (Basic)

Unit Overview

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.

40 (Overview) 

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) 

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) 

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) 

Resources

- http://www.slideshare.net/BrandEmotivity/sports-analytics-innovation-summit-data-powe
- http://www.sloansportsconference.com/
- http://sabr.org/
- http://en.wikipedia.org/wiki/Sabermetrics
- http://en.wikipedia.org/wiki/Baseball_statistics
- http://www.sportvision.com/baseball
- http://m.mlb.com/news/article/68514514/mlbam-introduces-new-way-to-analyze-every-play
- http://www.fangraphs.com/library/offense/offensive-statistics-list/
- http://en.wikipedia.org/wiki/Component_ERA
- http://www.fangraphs.com/library/pitching/fip/
- http://nomaas.org/2012/05/a-look-at-the-defense-the-yankees-d-stinks-edition/
- http://en.wikipedia.org/wiki/Wins_Above_Replacement
- http://www.fangraphs.com/library/misc/war/
- http://www.baseball-reference.com/about/war_explained.shtml

- http://www.baseball-reference.com/about/war_explained_comparison.shtml
- http://www.baseball-reference.com/about/war_explained_position.shtml
- http://www.baseball-reference.com/about/war_explained_pitch.shtml
- http://www.fangraphs.com/leaders.aspx?pos=all&stats=bat&lg=all&qual=y&type=8&season=2014&month=0&season1=1871&ind=0
- http://battingleadoff.com/2014/01/08/comparing-the-three-war-measures-part-ii/
- http://battingleadoff.com/2014/01/08/comparing-the-three-war-measures-part-ii/
- http://en.wikipedia.org/wiki/Coefficient_of_determination
- http://www.sloansportsconference.com/wp-content/uploads/2014/02/2014_SSAC_Data-driven-Method-for-In-game-Decision-Making.pdf
- https://courses.edx.org/courses/BUx/SABR101x/2T2014/courseware/10e616fc7649469ab4457ae

6.11.2 Sports Informatics II : Sabermetrics (Advanced)

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

41 (Sporta II) 

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) 

Pitcher Quality

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

Pitcher Quality (10:02) 

6.11.3 PITCHf/X

Examples of use of PITCHf/X.

PITCHf/X (10:39) 

Other Video Data Gathering in Baseball

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

Other Video Data Gathering in Baseball (18:5) 

Resources

- http://vincegennaro.mlblogs.com/
- https://www.youtube.com/watch?v=H-kx-x_d0Mk

- http://www.sportvision.com/media/pitchfx-how-it-works
- http://www.baseballprospectus.com/article.php?articleid=13109
- http://baseball.physics.illinois.edu/FastPFXGuide.pdf
- http://baseball.physics.illinois.edu/FieldFX-TDR-GregR.pdf
- http://www.sportvision.com/baseball/fieldfx
- http://regressing.deadspin.com/mlb-announces-revolutionary-new-fielding-tracking-syst
- http://grantland.com/the-triangle/mlb-advanced-media-play-tracking-bob-bowman-intervi
- http://www.sportvision.com/baseball/hitfx
- https://www.youtube.com/watch?v=YkjtnuNmK74

6.11.4 Sports Informatics III : 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.

44 (Sports III) 

Wearables

Consumer Sports, Stake Holders, and Multiple Factors.

Wearables (22:2) 

Soccer and the Olympics

Soccer, Tracking Players and Balls, Olympics.

Soccer and the Olympics (8:28) 

Spatial Visualization in NFL and NBA

NFL, NBA, and Spatial Visualization.

Spatial Visualization in NFL and NBA (15:19) 

Tennis and Horse Racing

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

V̌ideo8:52Tennis and Horse Racinghttps://www.youtube.com/watch?v=2P-pismFSrI

Resources

- http://www.sloansportsconference.com/?page_id=481&sort_cate=Research%20Paper
- http://www.slideshare.net/Tricon_Infotech/big-data-for-big-sports
- http://www.slideshare.net/BrandEmotivity/sports-analytics-innovation-summit-data-powe
- http://www.liveathos.com/apparel/app
- http://www.slideshare.net/elew/sport-analytics-innovation

- http://www.wired.com/2013/02/catapault-smartball/
- http://www.sloansportsconference.com/wp-content/uploads/2014/06/Automated_Playbook_Generation.pdf
- http://autoscout.adsc.illinois.edu/publications/football-trajectory-dataset/
- http://www.sloansportsconference.com/wp-content/uploads/2012/02/Goldsberry_Sloan_Submission.pdf
- http://gamesetmap.com/
- http://www.trakus.com/technology.asp#tNetText

F chapter/theory/usecases.tex

6.12 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.

6.12.1 Overview of NIST Big Data Public Working Group (NBD-PWG) Process and1 Results

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.

45 (Overview) 

Introduction to NIST Big Data Public Working Group (NBD-PWG) Process

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) 

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) 

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) 

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) 

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) 

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 refernce architecture form the Refernce Architecture Interface specifications directly from the document.

The workshop Web page is located at

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

The agenda of teh workshop is as follows:

- `https://bigdatawg.nist.gov/2017_NIST_Big_Data_PWG_WorkshopAgenda_with_Speakers_Bio.pdf`

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.1500-8-draft.pdf`

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

Requirements and Use Case Subgroup Introduction

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) 

6.12.2 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.

100 (51) 

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) 

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) 

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) 

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) 

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) 

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) 

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) 

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) 

Energy Use Case

This covers Consumption forecasting in Smart Grids.

Energy Use Case (4:01) 

6.12.3 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.

43 (Features) 

Summary of Use Case Classification I

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) 

Database(SQL) Use Case Classification

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

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

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) 

Use Case Classifications I

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) 

Use Case Classifications II

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) 

Use Case Classifications III

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) 

Resources

- NIST Big Data Public Working Group (NBD-PWG) Process https://www.nist.gov/el/cyber-physical-systems/big-data-pwg
- Big Data Definitions: http://dx.doi.org/10.6028/NIST.SP.1500-1 (link is external)
- Big Data Taxonomies: http://dx.doi.org/10.6028/NIST.SP.1500-2 (link is external)
- Big Data Use Cases and Requirements: http://dx.doi.org/10.6028/NIST.SP.1500-3 (link is external)
- Big Data Security and Privacy: http://dx.doi.org/10.6028/NIST.SP.1500-4 (link is external)
- Big Data Architecture White Paper Survey: http://dx.doi.org/10.6028/NIST.SP.1500-5 (link is external)
- Big Data Reference Architecture: http://dx.doi.org/10.6028/NIST.SP.1500-6 (link is external)
- Big Data Standards Roadmap: http://dx.doi.org/10.6028/NIST.SP.1500-7 (link is external)

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: https://www.youtube.com/watch?v=l4Qii7T8zeg
- On line 51 Use Cases http://bigdatawg.nist.gov/usecases.php
- Summary of Requirements Subgroup http://bigdatawg.nist.gov/_uploadfiles/M0245_v5_6066621242.docx
- Use Case 6 Mendeley http://mendeley.com%20http//dev.mendeley.com
- Use Case 7 Netflix http://www.slideshare.net/xamat/building-largescale-realworld-recommend
- 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) http://www.disasterrecovery.org/
- Use Case 11 and Use Case 12 Simulation driven Materials Genomics https://www.materialsproject.

org/
- 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-extrac htm, http://www.defencetalk.com/wide-area-persistent-surveillance-revolutionizes-tactic
- Use Case 15 Intelligence Data Processing and Analysis http://www.afcea-aberdeen.org/files/presentations/AFCEAAberdeen_DCGSA_COLWells_PS.pdf, http://stids.c4i.gmu.edu/papers/STIDSPapers/STIDS2012/_T14/_SmithEtAl/_HorizontalIntegrationOfWarfi pdf, http://stids.c4i.gmu.edu/STIDS2011/papers/STIDS2011_CR_T1_SalmenEtAl.pdf, 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/PAIS , 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 http://img.jgi.doe.gov/
- Use Case 25 Biodiversity and 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-j html , http://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evide html , http://www.wired.com/2013/06/andrew_ng/,
  A recent research paper on HPC for Deep Learning: http://www.stanford.edu/~acoates/papers/CoatesHuvalWangWuNgCatanzaro_icml2013.pdf, 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 http://vision.soic.indiana.edu/projects/disco/
- Use Case 28 Truthy: Information diffusion research from Twitter Data http://truthy.indiana.edu/ , 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 http://cinet.vbi.vt.edu/cinet_new/
- Use Case 31 NIST Information Access Division analytic technology performance measurement, evaluations, and standards http://www.nist.gov/itl/iad/
- Use Case 32 DataNet Federation Consortium DFC: The DataNet Federation Consortium, iRODS
- Use Case 33 The 'Discinnet process', metadata < - > big data global experiment http://www.discinnet.org/
- 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, see, e.g., http://arxiv.org/abs/1209.1681
- 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%20all%20the%20data%20come%20from%20v7.pdf, 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 http://belle2.kek.jp/
- Use Case 41 EISCAT 3D incoherent scatter radar system https://www.eiscat3d.se/
- 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 https://www.cresis.ku.edu/
- 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-pr
- 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/, 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=null#%40%3F_afrWindowId%3Dnull%26_afrLoop%3D157401916661989%26_afrWindowMode%3D0%26_adf.ctrl-state%3Db7yulr4rl_17, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6475927

F chapter/theory/web.tex

6.13 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.

6.13.1 Web Search and Text Mining I

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.

56 (Web Search and Text Mining) 

6.13.2 Web and Document/Text Search: 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.

6.13.3 Information Retrieval leading to Web Search

Information Retrival (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.

6.13.4 History behind Web Search

Web Search History (5:48) 

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

6.13.5 Key Fundamental Principles behind Web Search

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.

6.13.6 Information Retrieval (Web Search) Components

Fundametal 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.

6.13.7 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.

6.13.8 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.

6.13.9 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.

6.13.10 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.

6.13.11 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.

6.13.12 Resources

- `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://www.youtube.com/watch?v=yPFi6xFnDHE&feature=youtu.be Jean-Loup Guillaume and Matthieu Latapy `https://hal.archives-ouvertes.fr/file/index/docid/54458/filename/webgraph.pdf`
- 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.sciencedirect.com/science/article/pii/S0167923612001844`
- `http://www.ifis.cs.tu-bs.de/teaching/ss-11/irws`

6.13.13 Web Search and Text Mining II

33 (Text Mining)  PDF

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.

6.13.14 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.

6.13.15 Link Structure Analysis including PageRank

Realated 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.

6.13.16 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.

6.13.17 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.

6.13.18 Resources

- `http://www.ifis.cs.tu-bs.de/teaching/ss-11/irws`
- `https://en.wikipedia.org/wiki/PageRank`
- `http://webcourse.cs.technion.ac.il/236621/Winter2011-2012/en/ho_Lectures.html`
- Meeker/Wu May 29 2013 Internet Trends D11 Conference `http://www.slideshare.net/kleinerperkins/kpcb-internet-trends-2013`

Index

V