



FIT1043 Lecture 9

Introduction to Data Science

Mahsa Salehi*

Faculty of Information Technology, Monash University

Semester 2, 2022

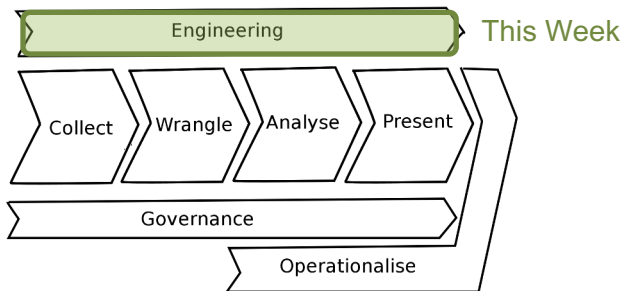
Assignment 2

- Due Monday 3rd October 11:55pm
- Using libraries/packages in Python
- Any questions:
 - Post to Ed discussion
 - Email: : fit1043.clayton-x@monash.edu
 - Email your tutors

Unit Schedule

Week	Activities	Assignments
1	Overview of data science	Weekly Lecture/tutorial active participation assessment
2	Introduction to Python for data science	
3	Data visualisation and descriptive statistics	
4	Data sources and data wrangling	
5	Data analysis theory	Assignment 1
6	Regression analysis	
7	Classification and clustering	
8	Introduction to R for data science	
9	Characterising data and "big" data	
10	Big data processing	Assignment 2 (Monday)
11	Issues in data management	
12	Industry guest lecture	Assignment 3

Our Standard Value Chain



Last Week

Tools for
data science

Outline

- Characterising data and “big data”
 - the V's
 - Metadata
 - Dimensions of data
 - Growth laws
- Introduction to Unix Shell for data science
 - Why Unix shell
 - Useful commands to read/manipulate large data files

Learning Outcomes (Week 9)

By the end of this week you should be able to:

- ▶ Characterize data sets used to assess a data science project
- ▶ Explain what Big data is
- ▶ Understand the V's in Big data
- ▶ Understand and analyse the growth laws: Moore's Law, Koomey's Law, Bell's Law and Zimmerman's Law
- ▶ Analyze and use shell commands to read and manipulate big data



Characterising Data

Characterising Data

Some general characterisations of data sets used to assess a project:

- **The V's**
 - The first characterisations by someone with a penchant for alliteration
- **Metadata**
 - Data about data is critical to understanding
- **Dimensions of data**
 - Infographics on data dimensions (how big is “big”)
- **Growth laws**
 - Understanding the exponential growth

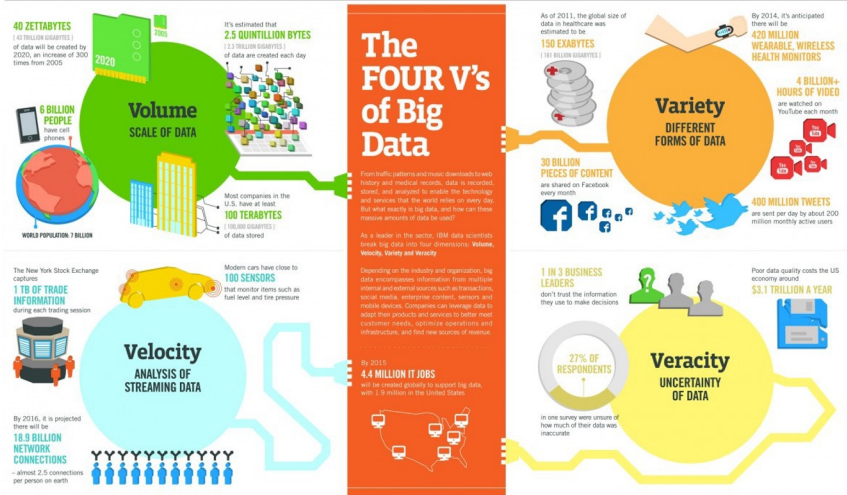
Characterising Data

The V's

The first characterisations of big data were by someone with a penchant for alliteration ... others followed

The Four V's of Big Data

"The Four V's of Big Data," by IBM (infographic)



Big Data

From [Big data](#) on Wikipedia:

*Big data usually includes data sets with **sizes beyond the ability of commonly used software tools** to capture, curate, manage, and process data within a tolerable elapsed time. Big data "size" is a constantly moving target, ...*

- Don't always ask why, can simply detect patterns
- A cost-free byproduct of digital interaction
- Enabled by the cloud: affordability, extensibility, agility

Big Data and “V”s

- 2001 Doug Laney produced report describing 3 V's:
“3-D Data Management: Controlling Data Volume, Velocity and Variety”
- These characterise bigness, adequately
- Other V's characterise problems with analysis and understanding
 - Veracity: correctness, truth, *i.e.* lack of ...
 - Variability: change in meaning over time, *e.g.*, natural language
- Other V's characterise aspirations
 - Visualisation: one method for analysis
 - Value: what we want to get out of the data
- Think of any more? write a blog!

FLUX Question

The 3Vs of big data are important because:

- A. They are an industry standard
- B. They are the basis for the development of more Vs (e.g. Value)
- C. They are used to describe in what way a dataset may be too big to handle
- D. They are from the influential Gartner Inc

FLUX Question

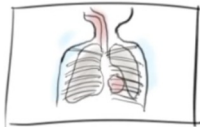
Which of the following is considered as big data?

40MB



A. power point presentation

1TB



B. Healthcare image

1PB



C. Movie

Summary

BIG DATA is ANY attribute that challenges
CONSTRAINTS of a system CAPABILITY or BUSINESS
NEED

Characterising Data Metadata

Data about data is critical to understanding

Metadata

MetaData ::= structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource.

MetaData is:

- **Data about data**
- **Structured** so that a computer can process & interpret it

Metadata (cont.)

Metadata can be:

Descriptive: Describes content for identification and retrieval
e.g. title, author of a book

Structural: Documents relationships and links
e.g. chapters in a book, elements in XML, containers in MPEG

Administrative: Helps to manage information
e.g. version number, archiving date, Digital Rights Management (DRM)

Why Use Metadata?

- Facilitate data discovery
- Help users determine the applicability of the data
- Enable interpretation and reuse
- Clarify ownership and restrictions on reuse

FLUX Question

Name a type of metadata might be associated with an image.

EXIF Metadata


Photo Data Explorer

File Edit View Photo Help

Open Folder Save as Recycle First Prior Next Last Rotate L Rotate R Flip Fit Actual Zoom In Zoom Out Comment

Photo Data Explorer

Tabby.jpg tilt.jpg
Taj Mahal.jpg Toco Toucan.jpg
Target.jpg tower.jpg
The Kiss.jpg Tree.jpg
The Way.jpg Trees.jpg
Thoughtful.jpg Tropical.jpg
Thunder.jpg Tube.jpg



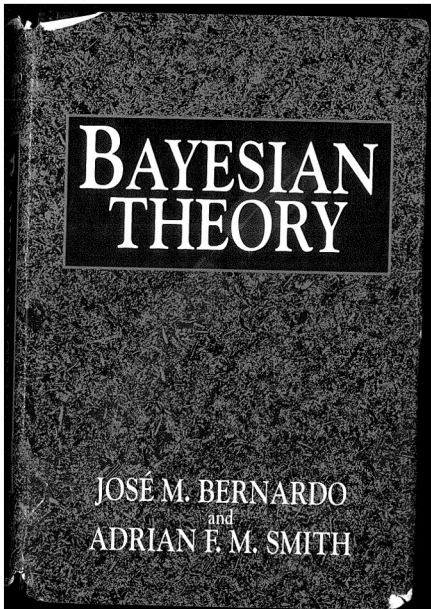
Exif Properties | Maker Data | Summary

Item	Details
Image Description	DCF 1.0
Make	Minolta Co., Ltd.
Model	DiMAGE S304
Orientation	Normal
XResolution	72.00
YResolution	72.00
Resolution Unit	Inch
Software	Adobe Photoshop CS Win
Date Time	2005:02:28 10:08:32
YCb Cr Positioning	Centered
Exposure Program	Normal
ISO Speed Ratings	100
Exif Version	"0210"
Date Time Original	2001:01:02 15:43:30
Date Time Digitized	2001:01:02 15:43:30
Components Configuration	YCbCr
Shutter Speed Value	0.0039 sec (1/256)
Aperture Value	F6.0
Exposure Bias Value	0/10
Max Aperture Value	F3.7
Metering Mode	MultiSegment
Light Source	Unidentified
Flash	Off
Focal Length	11.81 mm
Flash Pix Version	"0100"

Filename: Taj Mahal.jpg
Folder: C:\Users\Mike\Pictures\Slide Shows\

<http://www.alexnolan.net/photodata>

Book Metadata



Copyright © 1994 by John Wiley & Sons Ltd.
Baffins Lane, Chichester
West Sussex PO19 1UD, England
National Chichester (0243) 779777
International (+44) 243 779777

All rights reserved.

No part of this book may be reproduced by any means,
or transmitted, or translated into a machine language
without the written permission of the publisher.

Other Wiley Editorial Offices

John Wiley & Sons, Inc., 605 Third Avenue,
New York, NY 10158-0012, USA

Jacarana Wiley Ltd, 33 Park Road, Milton,
Queensland 4064, Australia

John Wiley & Sons (Canada) Ltd, 22 Worcester Road,
Rexdale, Ontario M9W 1L1, Canada

John Wiley & Sons (SEA) Pte Ltd, 37 Jalan Pemimpin #05-04,
Block B, Union Industrial Building, Singapore 2057

book metadata listed
on about third page

Library of Congress Cataloging-in-Publication Data

Bernardo, José M.

Bayesian theory / José M. Bernardo, Adrian F.M. Smith.
p. cm. — (Wiley series in probability and mathematical
statistics)

Includes bibliographical references and indexes.

ISBN 0 471 92416 4

I. Bayesian statistical decision theory. I. Smith, Adrian F.M.

II. Title. III. Series.

QA279.5.B47 1993

519.5'42—dc20

93-37554
CIP

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0 471 92416 4

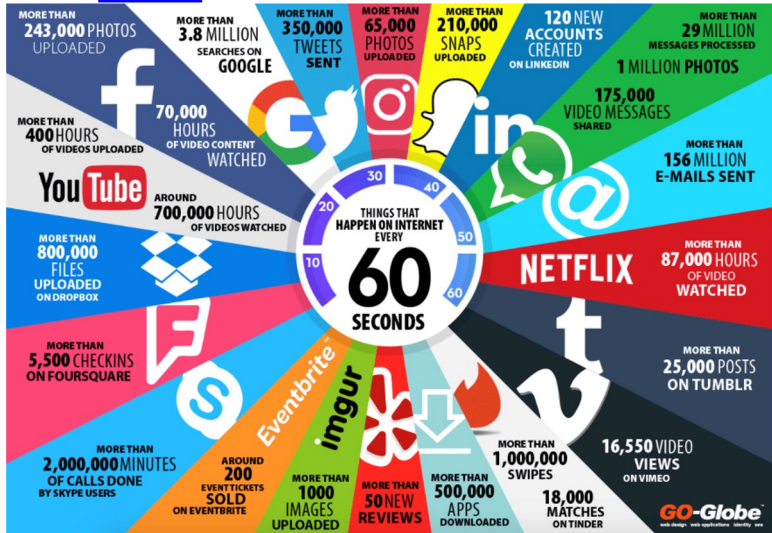
Characterising Data

Dimensions of data

Infographics on data dimensions (how big is “big”)

Things that happen in 60secs

from [GO-globe](#)



Infographics on Data

- [*"Data Science Matters"*](#) from the datascience@berkeley Blog
- Social Media Prisma from the [*Ethority.de site*](#)

Characterising Data

Growth laws

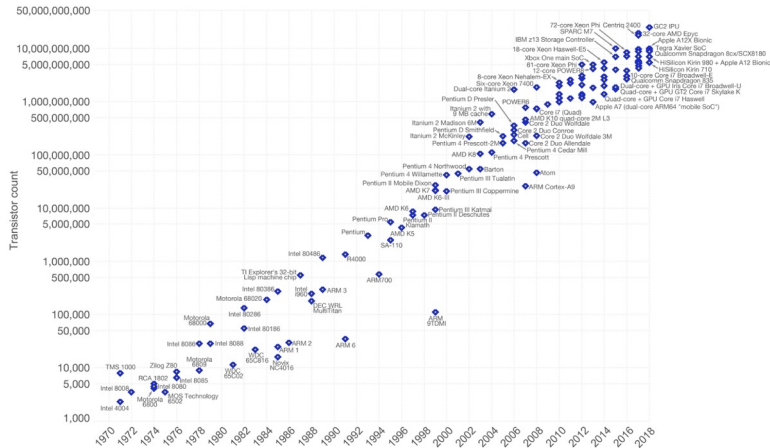
Understanding the exponential growth

Moore's Law

Moore's Law – The number of transistors on integrated circuit chips (1971-2018)

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are linked to Moore's law.

Our World
in Data



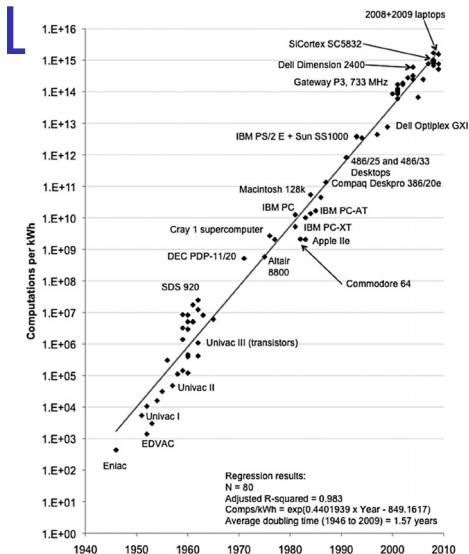
Data source: Wikipedia (https://en.wikipedia.org/wiki/Transistor_count)
The data visualization is available at [OurWorldinData.org](https://www.ourworldindata.org). There you find more visualizations and research on this topic.

Licensed under CC-BY-SA by the author Max Roser.

Moore's Law

- Number of transistors per chip doubles every 2 years (starting from 1975)
- Transistor count translates to:
 - More memory
 - Bigger CPUs
 - Faster memory, CPUs (smaller==faster)
- Pace currently slowing

Koomey's



By Dr Jon Koomey CC
 BY-SA 3.0, via Wikime-
 dia Commons

Koomey's Law

- Corollary of Moores Law
- Amount of battery needed will fall by a factor of 100 every decade
- Leads to ubiquitous computing

Bell's Law

Gordon Bell, Digital Equipment Corporation (DEC), 1972

- Corollary of Moore's Law and Koomey's Law
- "Roughly every decade a new, lower priced computer class forms based on a new programming platform, network, and interface resulting in new usage and the establishment of a new industry."

Yes: PCs, mobile computing, cloud, internet-of things

No: Java, big data, Hadoop, flash memory

Zimmerman's Law

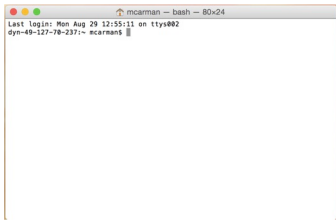
- Zimmerman is creator of Pretty Good Privacy (PGP), an early encryption system
- “Surveillance is constantly increasing”
- Privacy constantly decreasing

Introduction to Unix Shell for Data Science

This is a very brief introduction to shell

What is a Unix Shell?

- Command line interface to a Unix computer
 - Different shells have been around since the 70s
- Why are shells interesting for Data Scientists?
 - Provide powerful & easy way to **manipulate large data files**
 - And **move data** around a network
- Available on most Unix based operating systems
 - Linux
 - Mac OSX (BSD based)



What Shell Scripting?

Super-computers are typically UNIX based

- Explore data before you use it in Python or R
- Easier to manipulate and wrangle Big Data
 - Simple and easy to learn.
 - Ideal for textual data, e.g. unstructured data for social networks, life sciences, system logs, etc.
 - Quick to sort, search, match, replace and clean your data.



Getting Started

Installing a Shell on Windows:

- <https://www.cygwin.com/>
- For Windows 10, you can enable the Linux subsystem under (Control Panel → Programs → Turn Windows features on or off) after you have enabled the Developer mode on Windows.

Running a Shell:

- In Linux, click on the black square at the top left of the screen.
- In MacOSX, go to Applications -> Utilities -> Terminal.

Navigating the Filesystem

- Working directory:
`pwd`
- Change directory:
`cd [destination]`
`cd /my/favourite/place`
- Special cases:
`cd ..` <- Takes you up one directory
`cd` <- Without argument, takes you to home directory
- List files in the current directory:
`ls`
- Copy files from one location to another:
`cp [source] [destination]`
`cp Desktop/myfile .`

Reading a Text File

- Open a text file for reading using less:
`less myfile.txt`
- Navigate within the text file using
 - `[up/down]` <- move one line the file
 - `[space]` <- move down a whole page
 - `q` <- quit
 - `[shift]+g` <- go to the end of file
 - `g` <- go to the start of file
 - `/keyword` <- search for the first occurrence of
"keyword"
 - `/` <- find the next occurrence of keyword

Some Useful Commands

- Count the number of words/lines in a file (-l for line)
`wc myfile.txt`
- Find lines in a file containing a keyword
`grep "elephant" myfile.txt`
- Print the first/last few lines of a file
`head myfile.txt`
`tail myfile.txt`

Some Useful Commands

- Print the contents of a file to the screen

`cat myfile.txt`

- Read manual for a particular command

`man wc`

Then hit `q` for quit

Note: If you ever get into trouble on the command line (for example you get trapped in some program and don't know how to quit, just type `[control]+c` to kill the program.

FLUX Question

Unix shell commands like “less” and “grep”:

- A. can be used to manipulate large data files easily
- B. are poorly documented
- C. are examples of technology that is too old to be useful to a modern data scientist
- D. are used to fit regression tree models

Flags and Arguments

Many programs take flags and command line arguments that modify their behaviour, for example:

- Sort the contents of a file lexicographically (alphabetically)
`sort myfile.txt`
- Sort the data by column one, then column two and finally column three:
`sort -k1,3 myfile.txt`

Pipes

Sometimes we'd like the output of one program to be used as the input to another.

- Doing this is super easy in the shell. We just use the pipe operator “|”

```
program1 | program2
```

- We can chain as many programs together as we want, for example:

```
cat hourly_44201_2014-06.csv.gz | gunzip | less
```

Pipes(Cont.)

The pipe is buffered

- Each program in the list only generates data **as it is needed** by the next stage in the pipeline.
- Thus memory requirement for processing the data is limited
- Crucial for **scaling up processing to enormous data files.**

Redirects

If we want to save the results in a file rather than pipe them to a new program:

- Just change the pipe operator “|” to be a greater than symbol “>” and provide a filename:

```
cat hourly_44201_2014-06.csv.gz | gunzip > newFile.txt
```

Wildcards

- Some unix commands can take multiple files as input, for example:

```
cat myfile1.txt myfile2.txt
```

- In order to avoid listing large number of files, we can use the wildcard syntax to specify all files in a directory with a certain pattern, e.g.:

```
cat myfile*.txt
```

awk

In the tutorial, we'll have a look at a powerful command for **processing text files one line at a time** called awk

- awk syntax:

```
awk '[select line?] {do something}'
```

- Example

```
awk 'rand()<1/100 {print $6,$7,$14}'
```

- Since awk processes data one line a time, **it can scale up to massive datasets!**

End of Introduction

- We'll be experimenting with the Unix shell in this week's tutorial
- There are MANY excellent shell tutorials online if you'd like to learn more!