FIT1043 Introduction to Data Science

Week 9:Characterizing data and "big" data

Ts. Dr. Sicily Ting

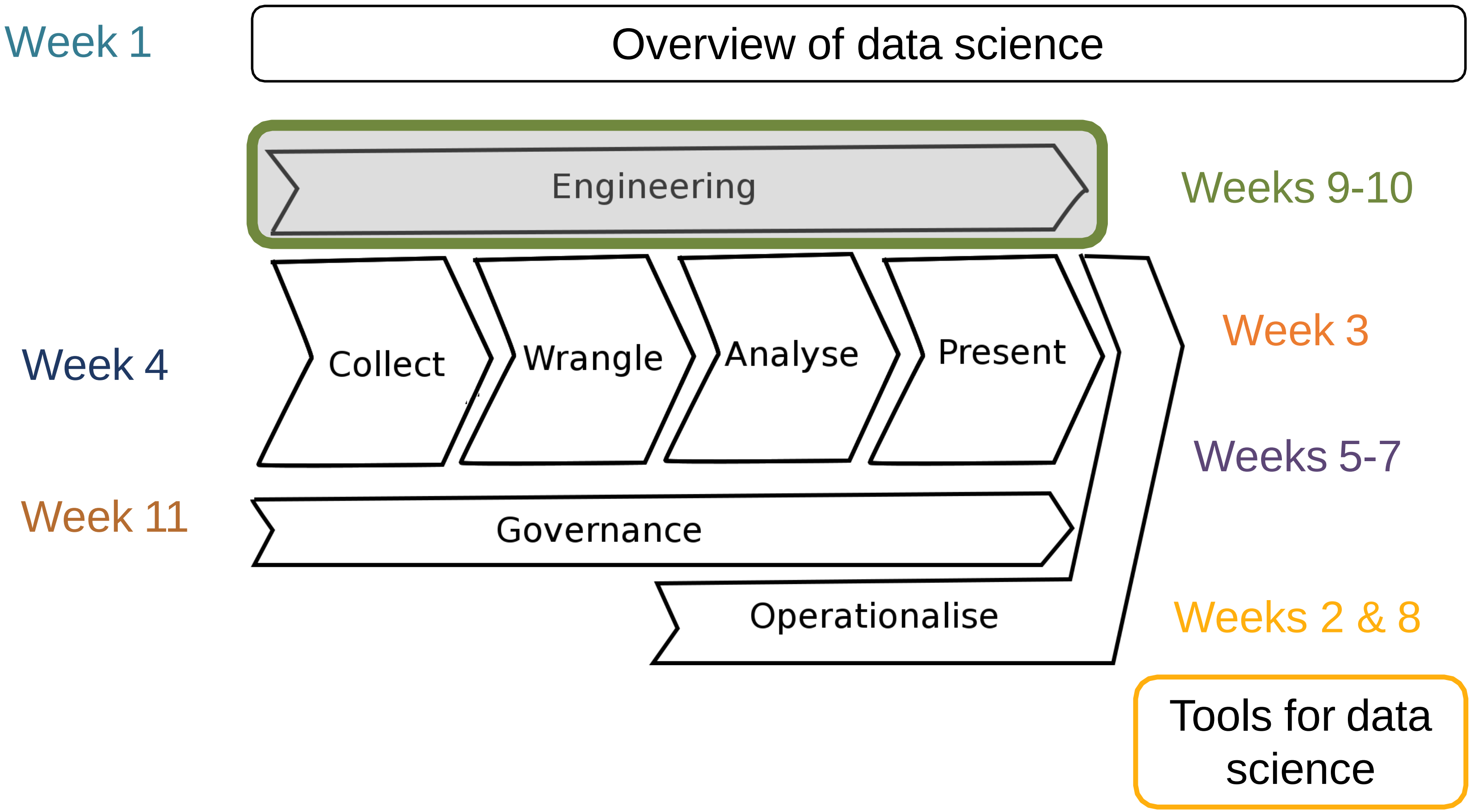
School of Information
Technology Monash
University Malaysia

With materials from Wray Buntine,
Mahsa Salehi

Week 8 Coverage

R

Week	Activities	Assignments
1	Overview of data science	
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	Assignment 2
9	Characterising data and "big" data	
10	Big data processing	
11	Issues in data management	Assignment 3
12	Industry guest lecture (tentative)	



Week 1

Overview of data science

Engineering

Weeks 9-10

Week 4

Collect

Wrangle

Analyse

Present

Week 3

Weeks 5-7

Week 11

Governance

Operationalise

Weeks 2 & 8

Tools for data science

Week 9
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 Big Data

MONASH University

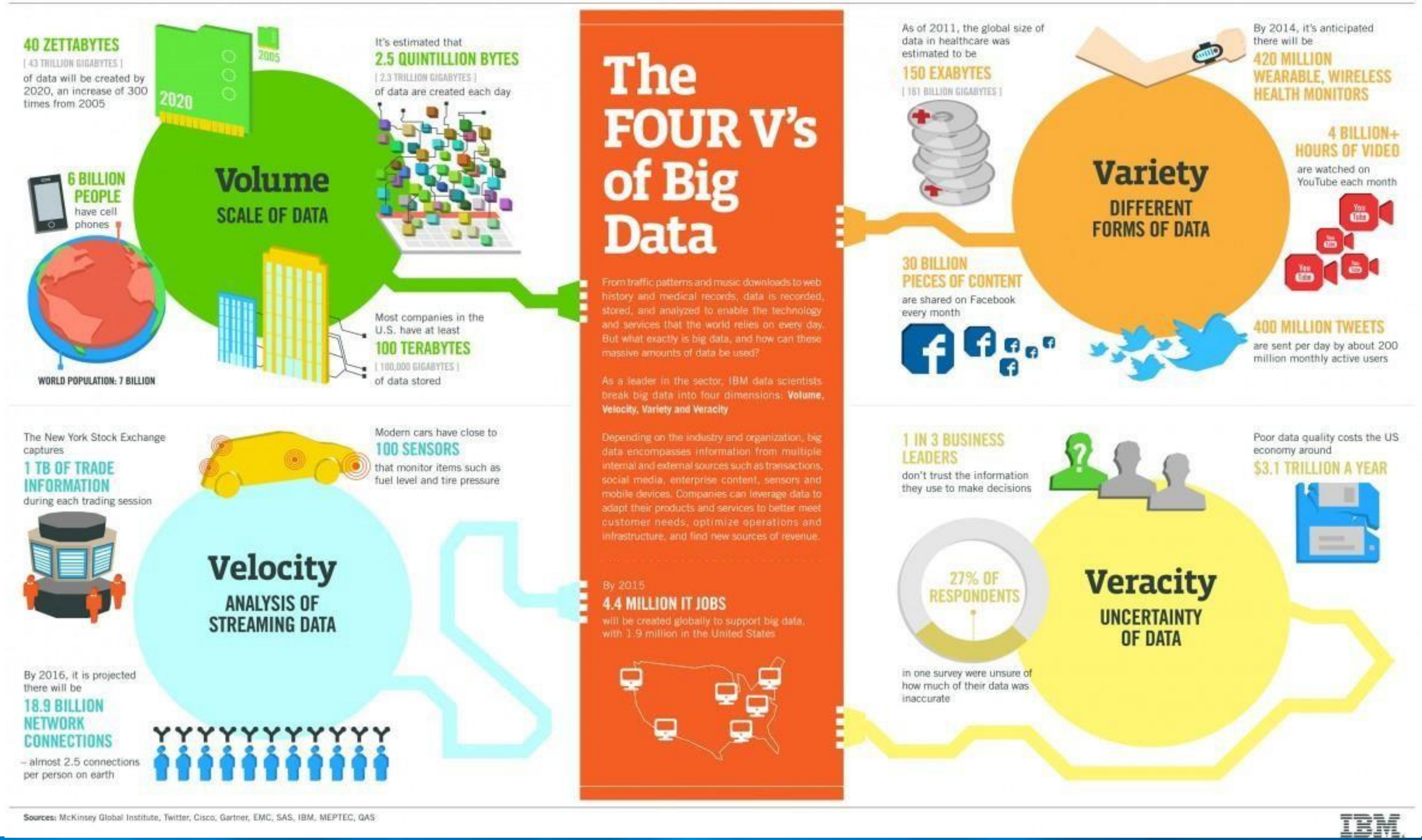
Characterising Data

Some general charactisations 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

The Four V’s of Big Data

“The Four V’s of Big Data,” by IBM (infographic)



40 ZETTABYTES [43 TRILLION GIGABYTES] of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE have cell phones

It’s estimated that 2.5 QUINTILLION BYTES [2.3 TRILLION GIGABYTES] of data are created each day

Most companies in the U.S. have at least 100 TERABYTES [100,000 GIGABYTES] of data stored

The New York Stock Exchange captures 1 TB OF TRADE INFORMATION during each trading session

Modern cars have close to 100 SENSORS that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be 18.9 BILLION NETWORK CONNECTIONS – almost 2.5 connections per person on earth

The FOUR V’s of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 4.4 MILLION IT JOBS will be created globally to support big data, with 1.9 million in the United States

As of 2011, the global size of data in healthcare was estimated to be 150 EXABYTES [161 BILLION GIGABYTES]

By 2014, it’s anticipated there will be 420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month

30 BILLION PIECES OF CONTENT are shared on Facebook every month

400 MILLION TWEETS are sent per day by about 200 million monthly active users

1 IN 3 BUSINESS LEADERS don’t trust the information they use to make decisions

27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around $3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

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, insights can come from detecting patterns
- A cost-free by product 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 adequately characterise “bigness”

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!

Summary

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

Characterising Data Metadata

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

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)

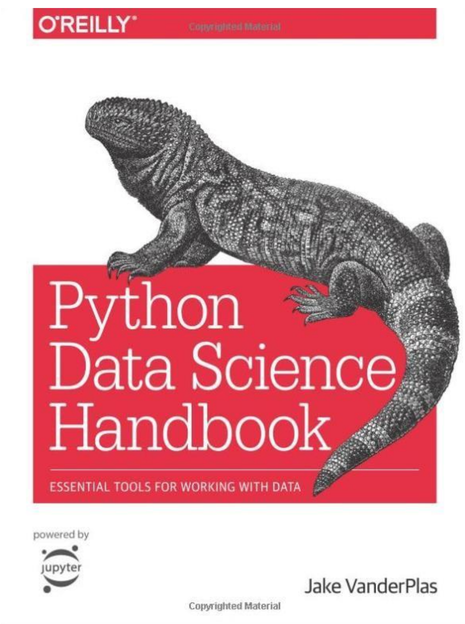
Metadata

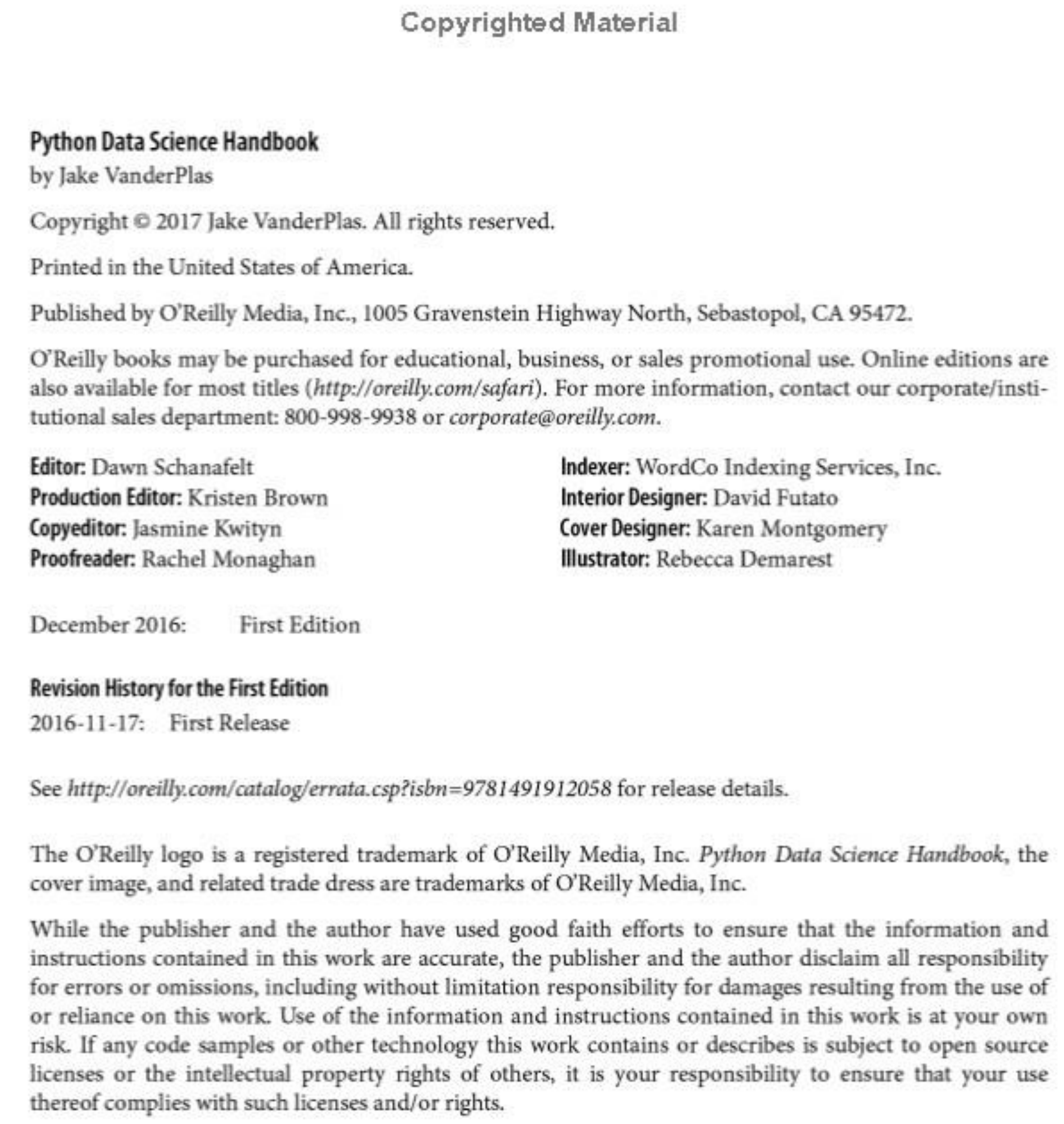
Why use Metadata?

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

Metadata of a Book

What are the Metadata of a Book?





Copyrighted Material

Python Data Science Handbook
by Jake VanderPlas

Copyright © 2017 Jake VanderPlas. All rights reserved.

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://oreilly.com/safari*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editor: Dawn Schanafelt
Production Editor: Kristen Brown
Copyeditor: Jasmine Kwityn
Proofreader: Rachel Monaghan

Indexer: WordCo Indexing Services, Inc.
Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Rebecca Demarest

December 2016: First Edition

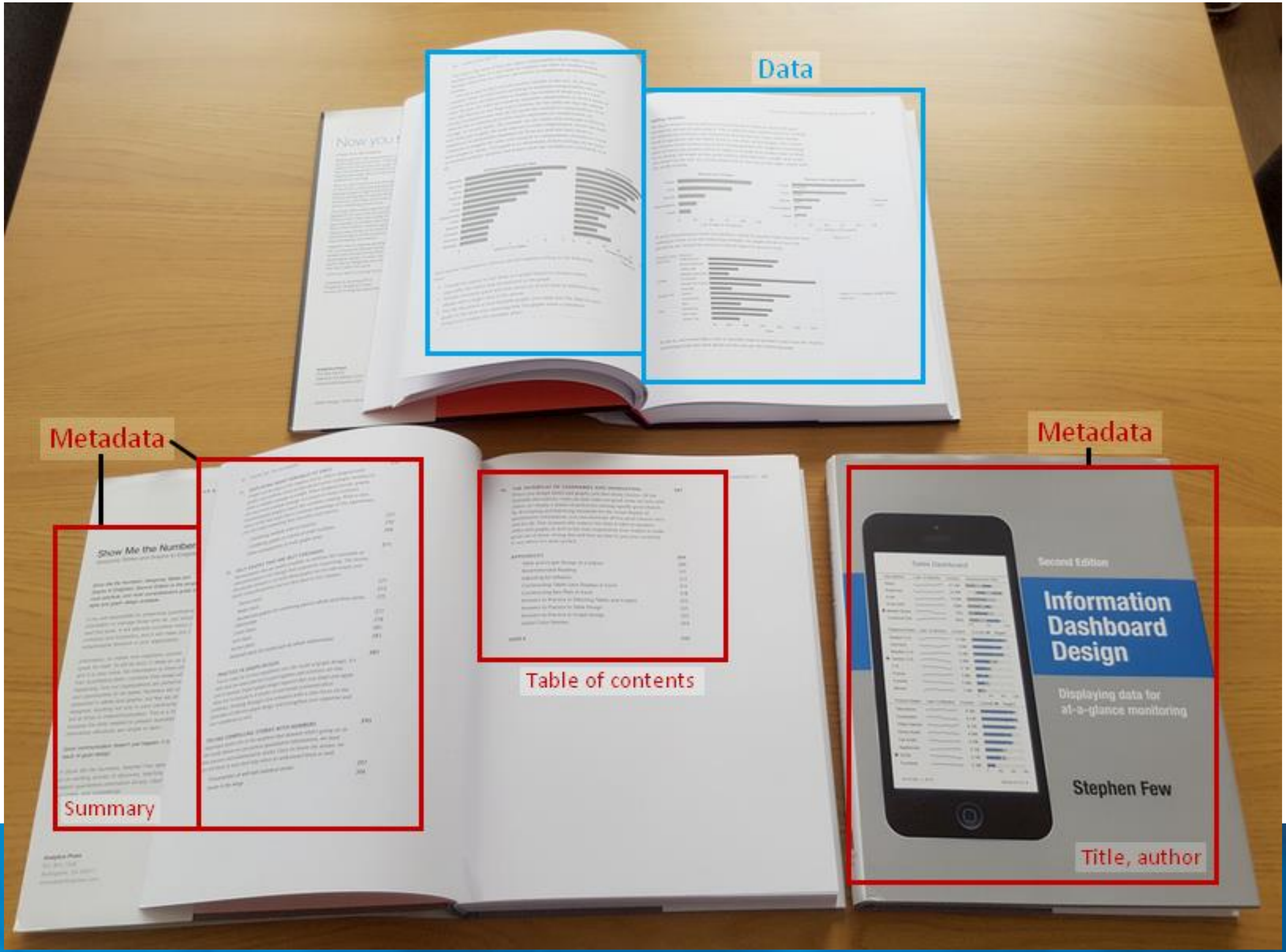
Revision History for the First Edition
2016-11-17: First Release

See *http://oreilly.com/catalog/errata.csp?isbn=9781491912058* for release details.

The O'Reilly logo is a registered trademark of O'Reilly Media, Inc. *Python Data Science Handbook*, the cover image, and related trade dress are trademarks of O'Reilly Media, Inc.

While the publisher and the author have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the author disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

Metadata of a Book



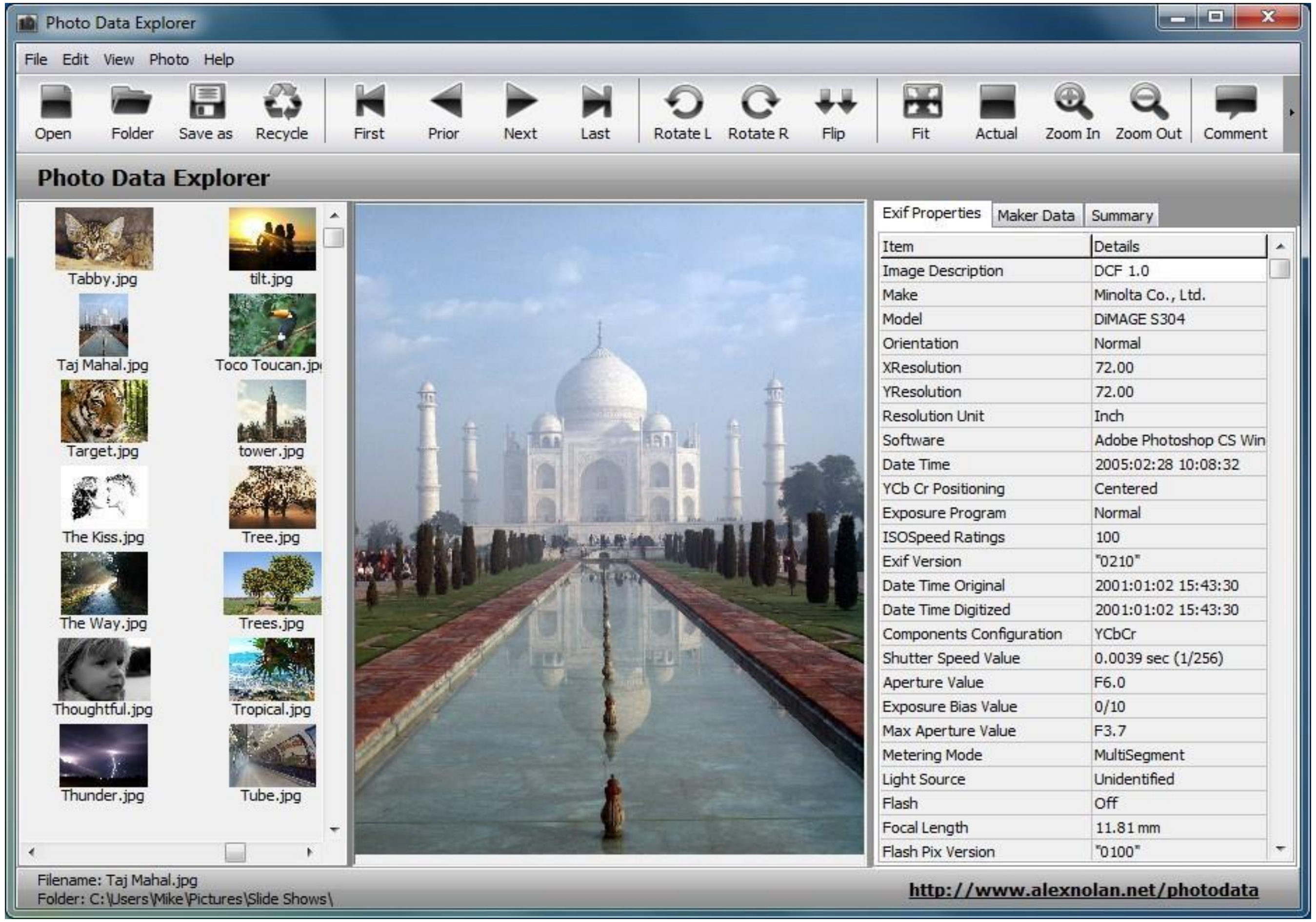
Other Examples of Metadata

- IPTC Photo Metadata User Guide
- USGS Metadata standards
- Medical bibliographic data in XML on PubMed,

  “Lower respiratory tract disorder hospitalizations among children born via elective early-term delivery”

Image Metadata

EXIF



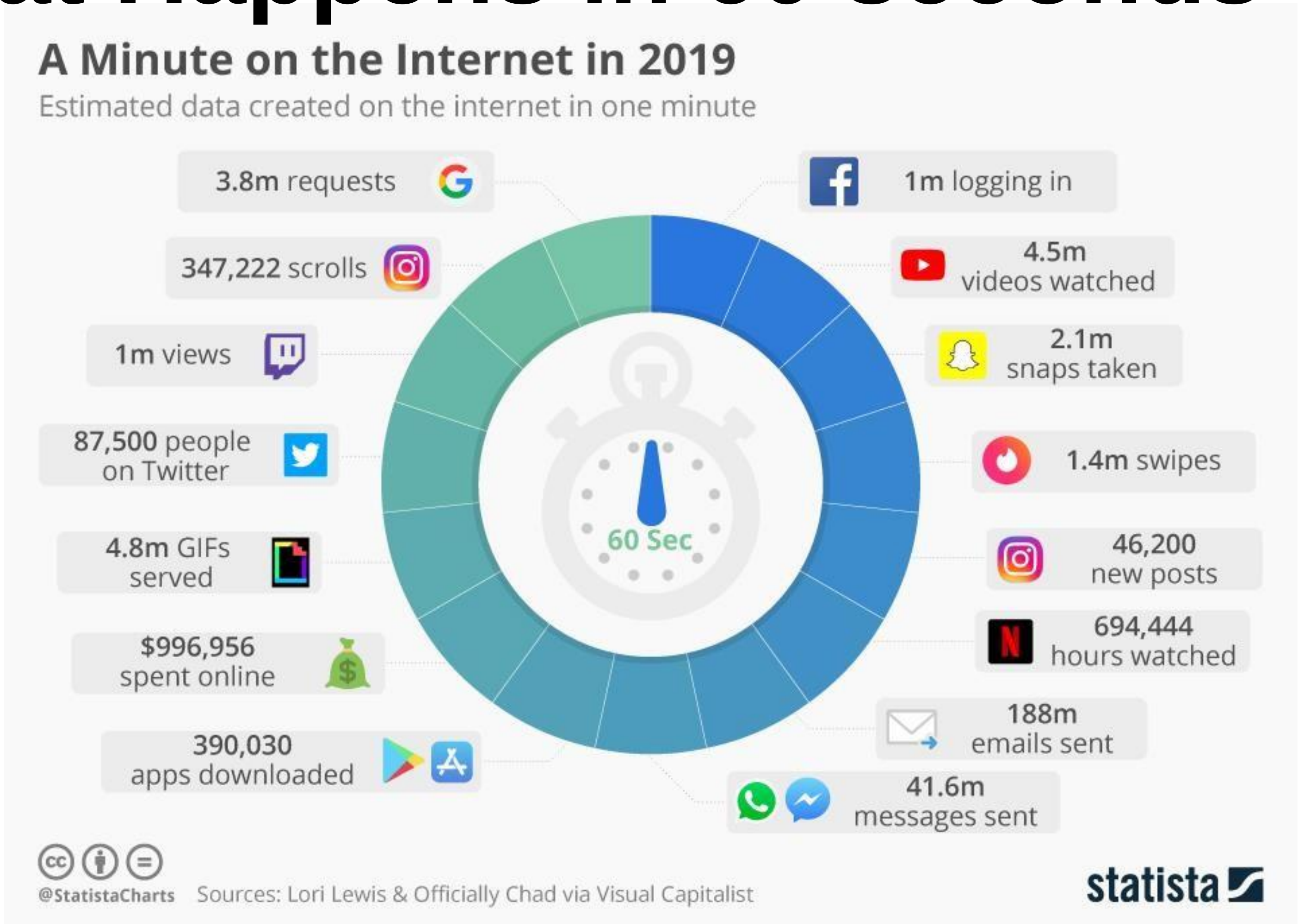
Item	Details
Image Description	DCF 1.0
Make	Minolta Co., Ltd.
Model	DiMAGE S304
Orientation	Normal
XResolution	72.00
YResolution	72.00
Resolution Unit	Inch
Software	Adobe Photoshop CS Win
Date Time	2005:02:28 10:08:32
YCb Cr Positioning	Centered
Exposure Program	Normal
ISOSpeed Ratings	100
Exif Version	"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

Characterising Data Dimensions of Data

Things that Happens in 60 seconds



Sources: Lori Lewis & Officially Chad via Visual Capitalist

The dimension is a data set composed of **individual, non-overlapping data elements**. The primary functions of dimensions are threefold: to provide filtering, grouping and labelling.

Infographics on Data

- “Data Science Matters” from the datascience@berkeley Blog
- Social Media Prisma from the Ethority.de site

Characterising Data Growth Laws

MONASH University

Moore's Law

Gordon Moore, Intel, 1965

Number of transistors per chip doubles every 2 years (starting from 1975)

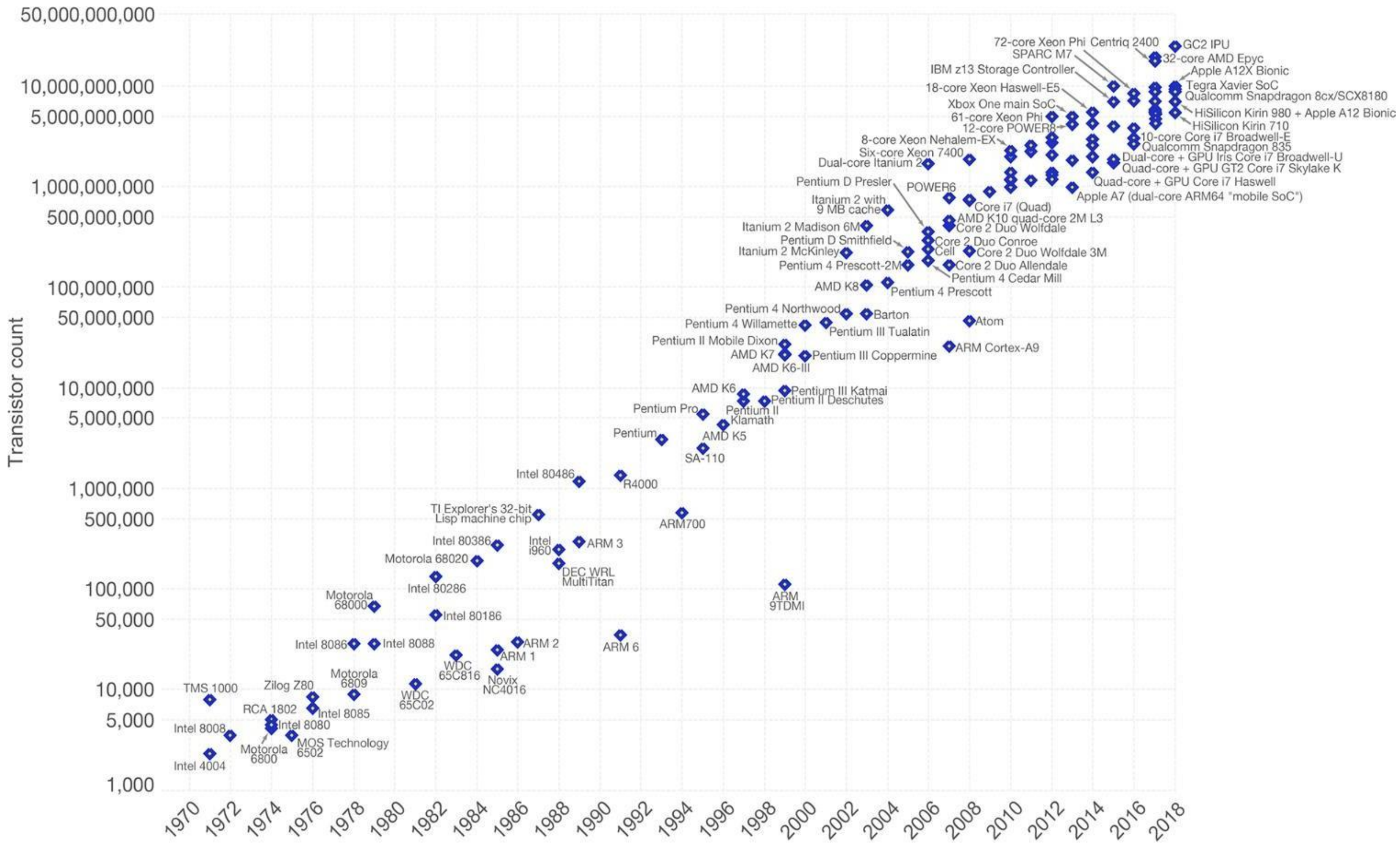
Transistor count translates to:

- More memory
- Bigger CPUs
- Faster memory, CPUs (smaller==faster)

Pace currently slowing

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.



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

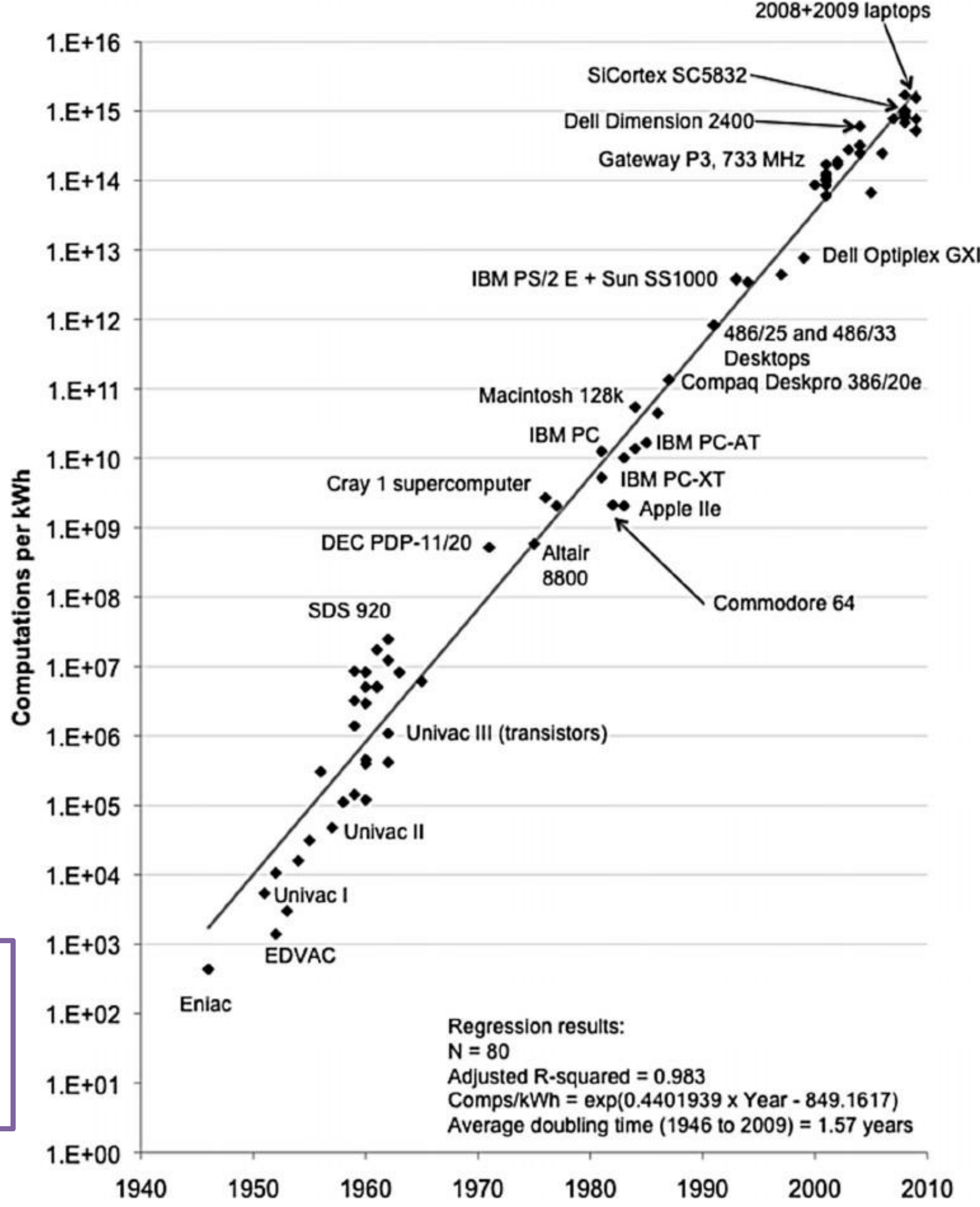
•Another tenet of Moore's Law says that the growth of microprocessors is exponential.

Koomey's Law

Jonathan Koomey, Standford University, 2010

Corollary of Moores Law

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



Regression results:
N = 80
Adjusted R-squared = 0.983
Comps/kWh = exp(0.4401939 x Year - 849.1617)
Average doubling time (1946 to 2009) = 1.57 years

In short, if Moore's Law talks about computing power and the number of transistors, Koomey's tells us that **the efficiency of processors and computing devices doubles approximately every 1.57 years** .

Koomey's Law describes a *trend that dictates the number of computations per joule of dissipated energy, which doubles every 1.57 years*

Bell’s Law

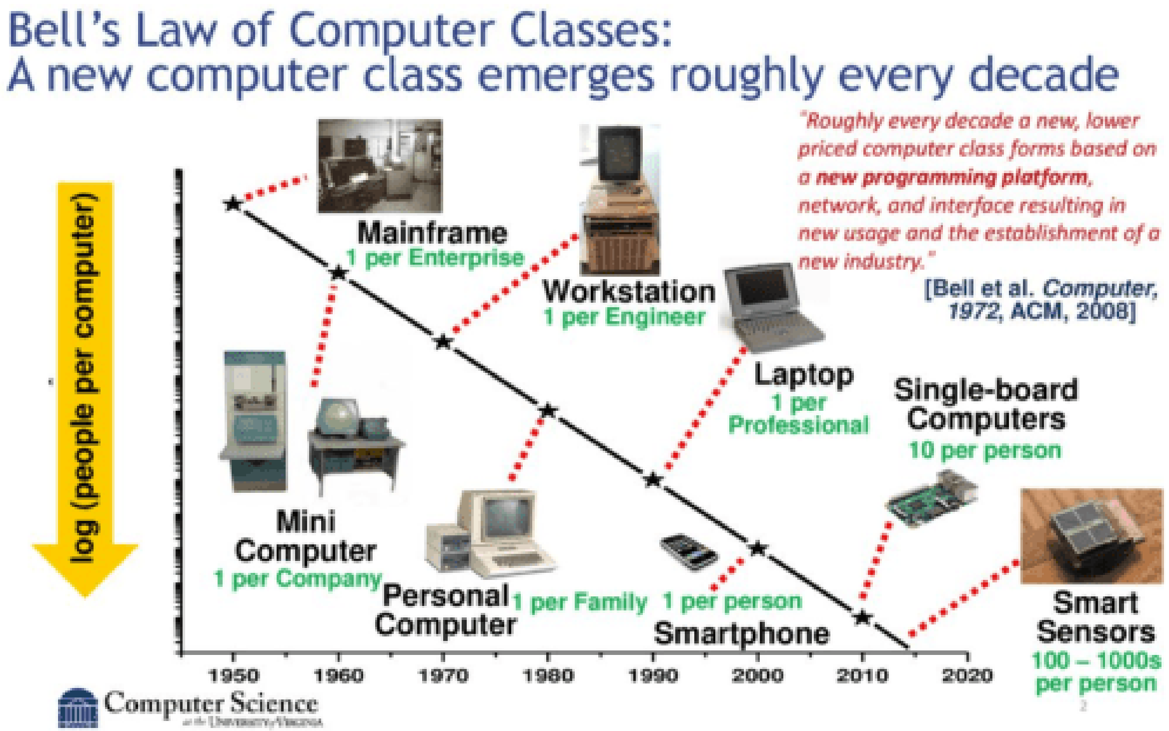
Gordon Bell, Digital Equipment Corporation (DEC), 1972

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



Bell’s Law of Computer Classes:
A new computer class emerges roughly every decade

“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.”*

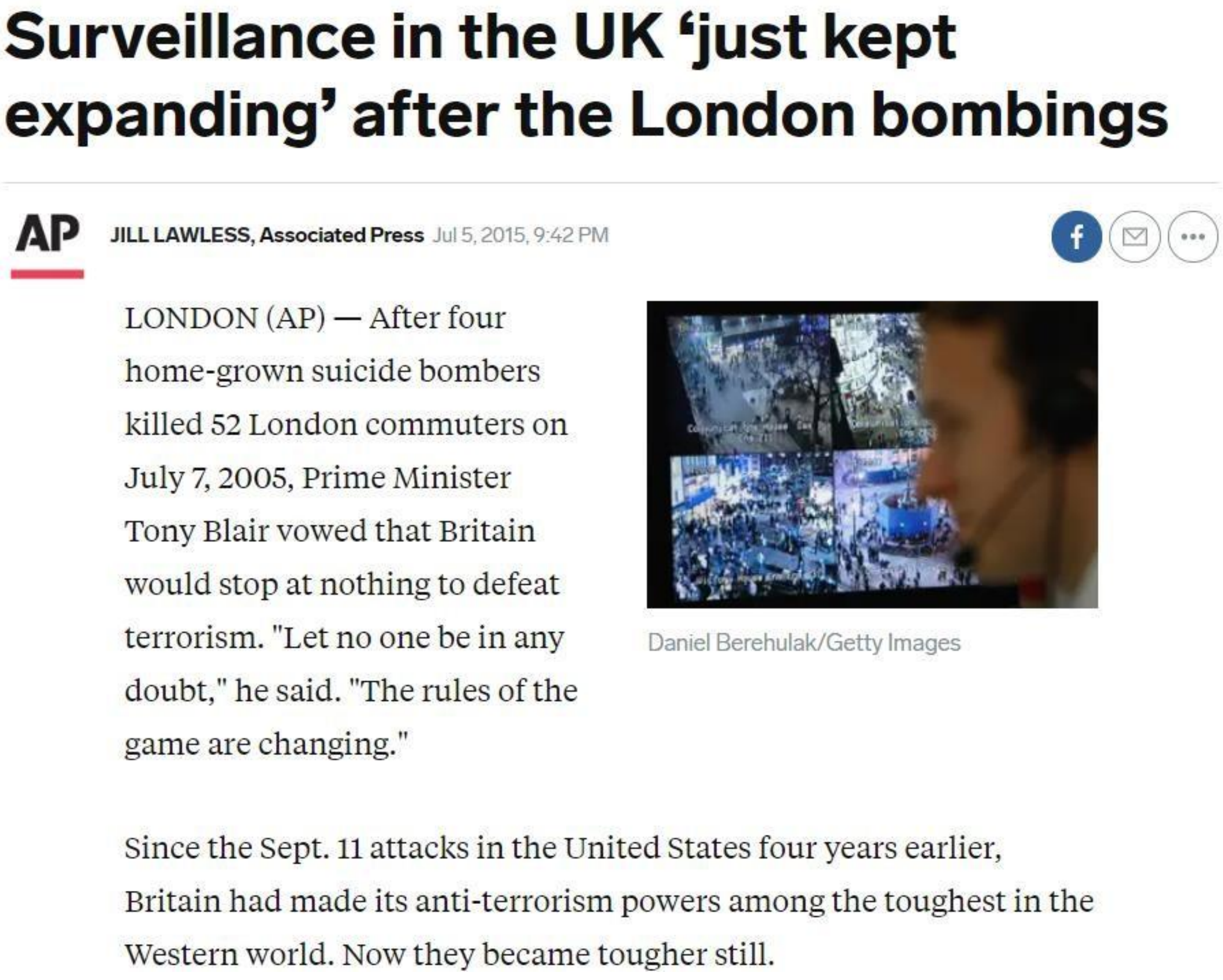
[Bell et al. *Computer, 1972*, ACM, 2008]

Zimmerman's Law

Phil Zimmermann, 2013

Zimmerman is creator of Pretty Good Privacy (PGP), an early encryption system

- "Surveillance is constantly increasing"
- Privacy constantly decreasing



Surveillance in the UK 'just kept expanding' after the London bombings

AP JILL LAWLESS, Associated Press Jul 5, 2015, 9:42 PM

LONDON (AP) — After four home-grown suicide bombers killed 52 London commuters on July 7, 2005, Prime Minister Tony Blair vowed that Britain would stop at nothing to defeat terrorism. "Let no one be in any doubt," he said. "The rules of the game are changing."

Daniel Berehulak/Getty Images

Since the Sept. 11 attacks in the United States four years earlier, Britain had made its anti-terrorism powers among the toughest in the Western world. Now they became tougher still.

Recap: Learning Outcomes

Week 9

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

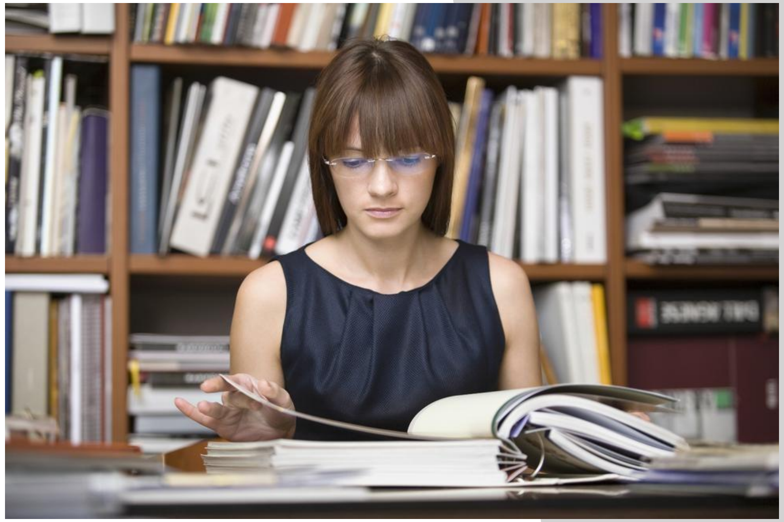
- Characterize data sets used to assess a data science project
- Explain what Big data is
- Understand the V’s in Big data
- Understand and analyse the growth laws: Moore’s Law, Koomey’s Law, Bell’s Law and Zimmerman’s Law
- *Analyze and use shell commands to read and manipulate big data*

Home Activities

Suggested Activities for the week

Online Materials

Watch
https://www.youtube.com/watch?time_continue=90&v=AWPrOvzzqZk



Books (Articles)

Go through the links provided in the lecture slides.

- https://www.digitalinformationworld.com/2019/04/what-happens-online-in-60-seconds.html

MONASH University

Tutorials
Week 9

Introduction to Shell Scripting

MONASH
University