MONASH University

FIT1043 Introduction to Data Science

Week 9

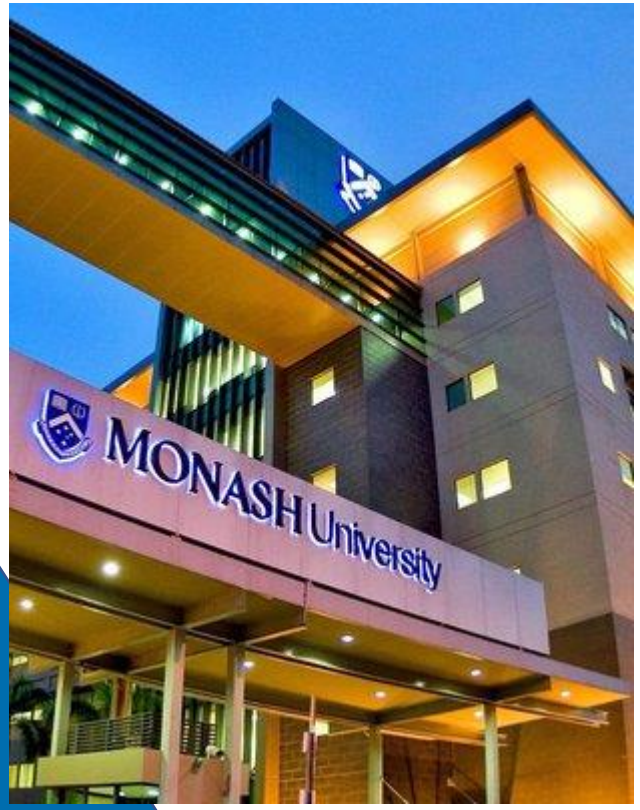
Ian K T Tan

School of Information Technology
Monash University Malaysia

With materials from Wray Buntine, Mahsa Salehi

Week 8 Coverage

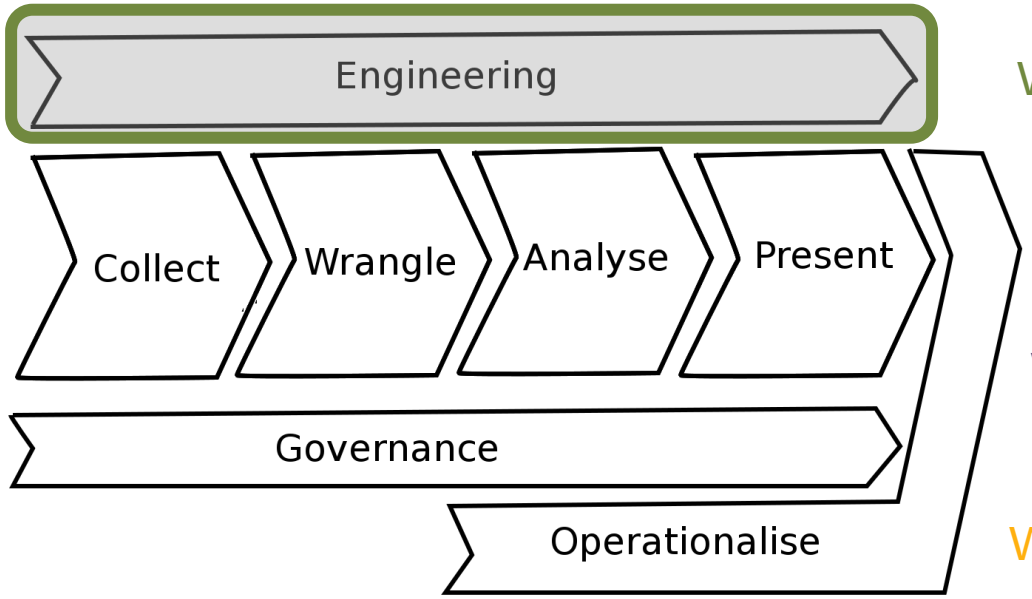
R



MONASH University

Week 1

Overview of data science



Weeks 9-10

Week 4

Week 3

Weeks 5-7

Week 11

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

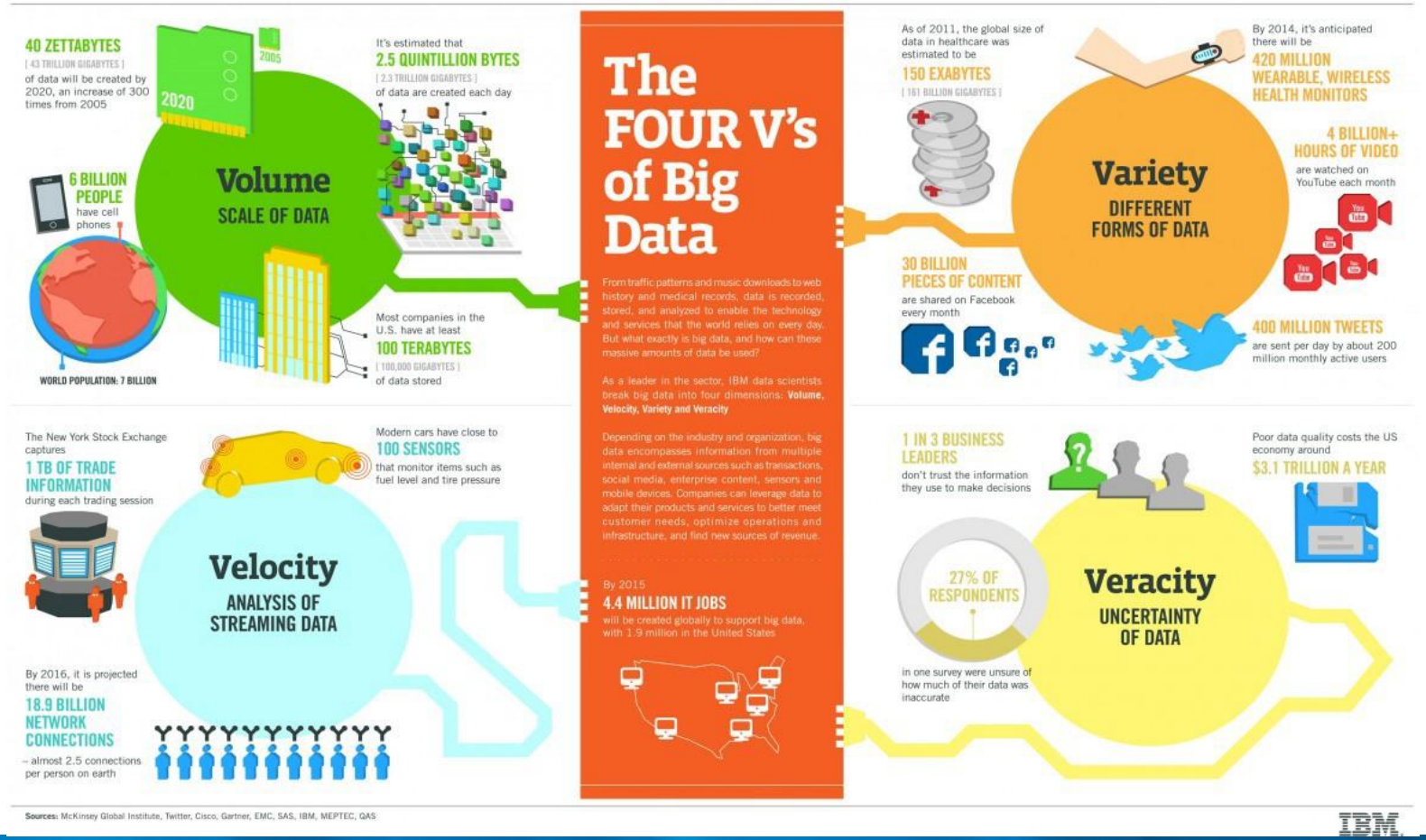
Characterising Data

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)



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

MONASH University

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)

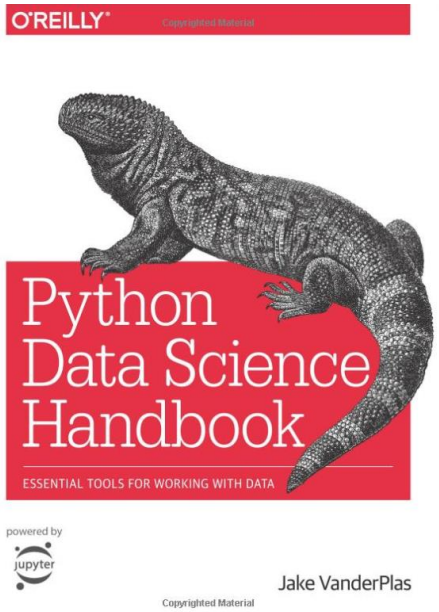
Metadata

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?



Python Data Science Handbook
by Jake VanderPlas

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.

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”

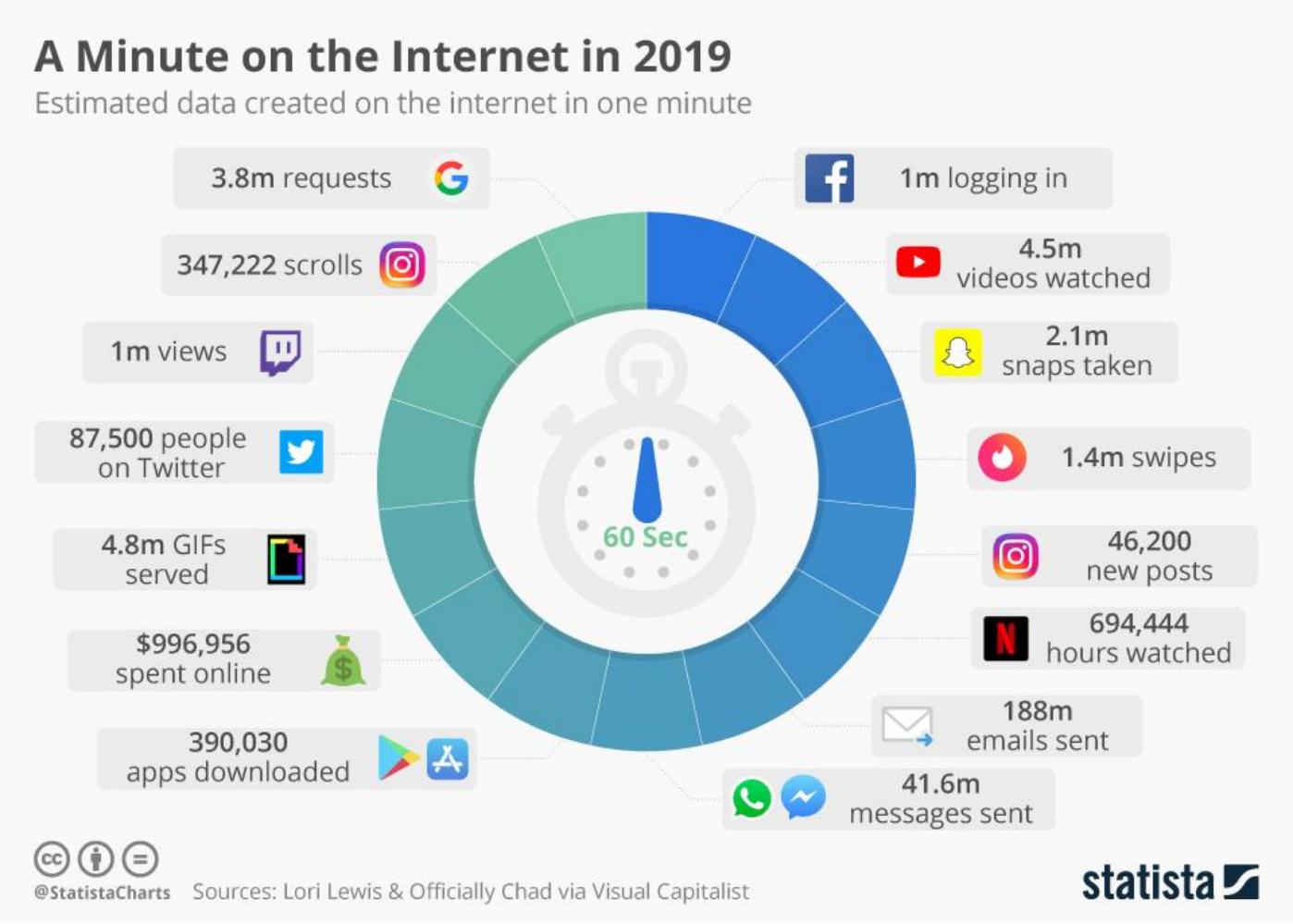
Image Metadata

EXIF

Characterising Data
Dimensions of Data

MONASH University

Things that Happens in 60 seconds



Infographics on Data

- "Data Science Matters" from the datascience@berkeley Blog
- "Intelligence by Variety – Where to Find and Access Big Data" from Kapow Software
- Social Media Prisma from the Ethority.de site

Characterising Data Growth Laws

MONASH University

Moore’s Law

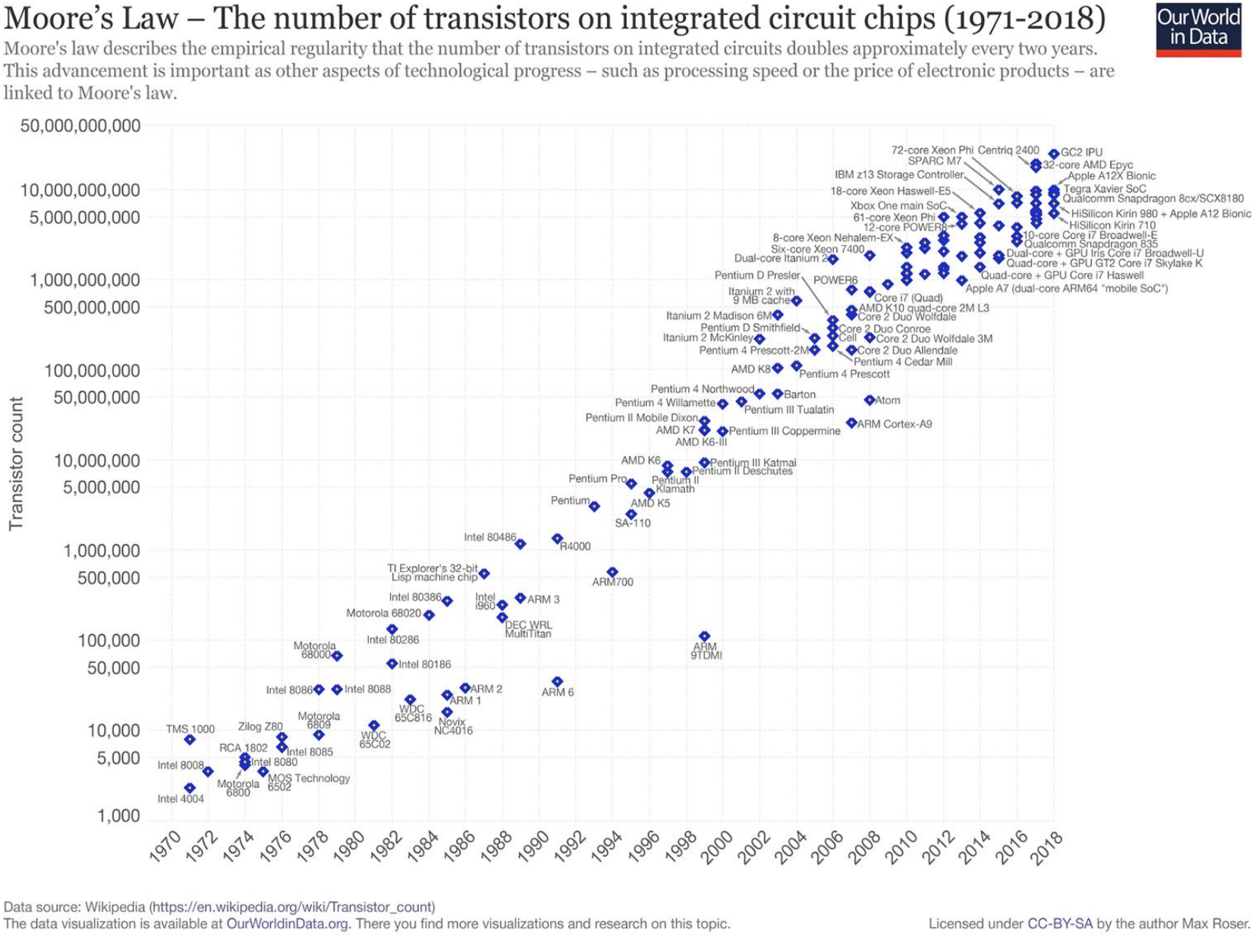
Gordon Moore, Intel, 1965

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



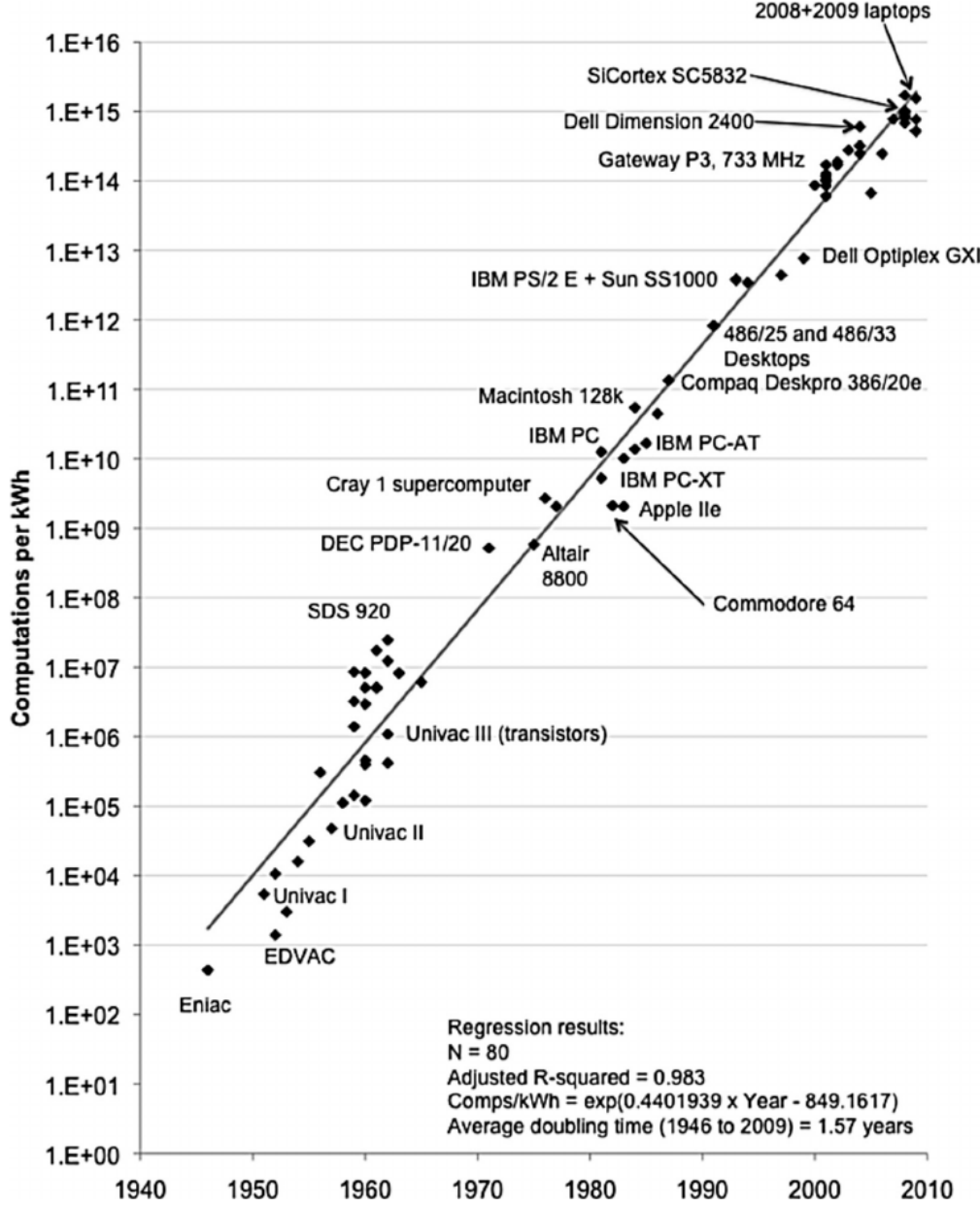
MONASH University

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



Bell's Law

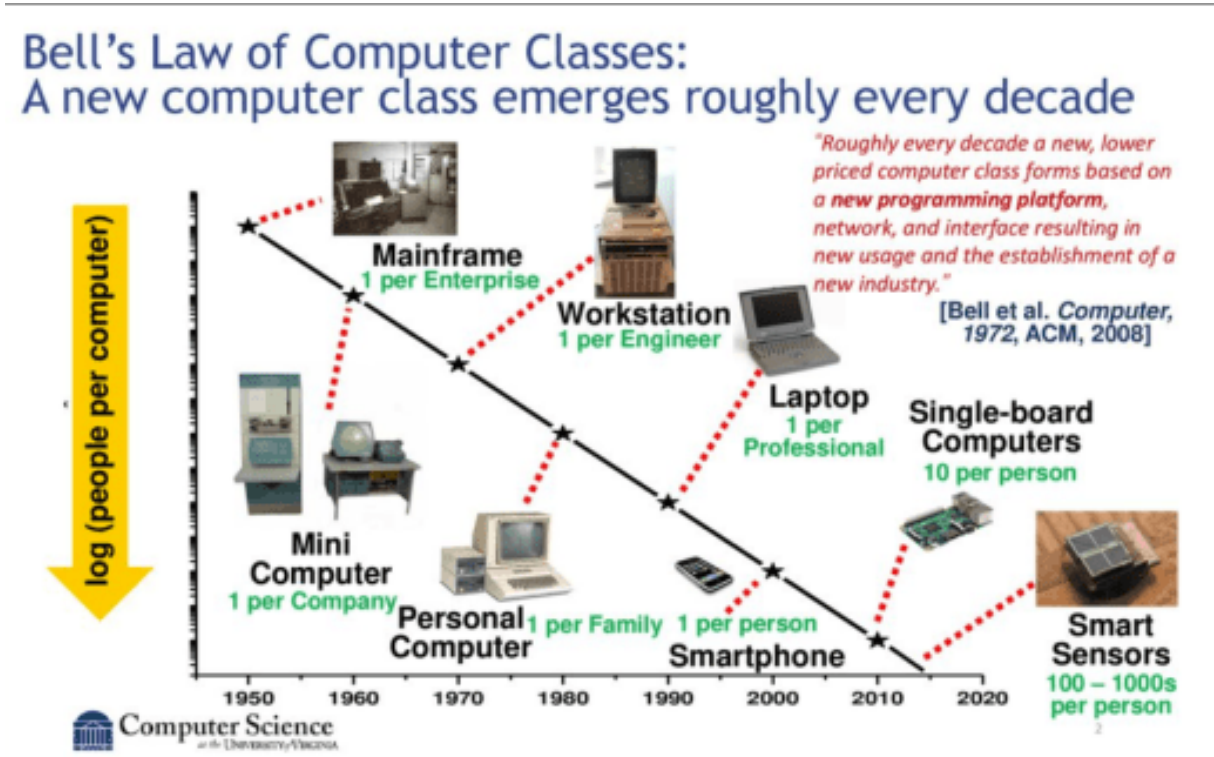
Gordon Bell, Digital Equipment Corporation (DEC), 1972

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

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:

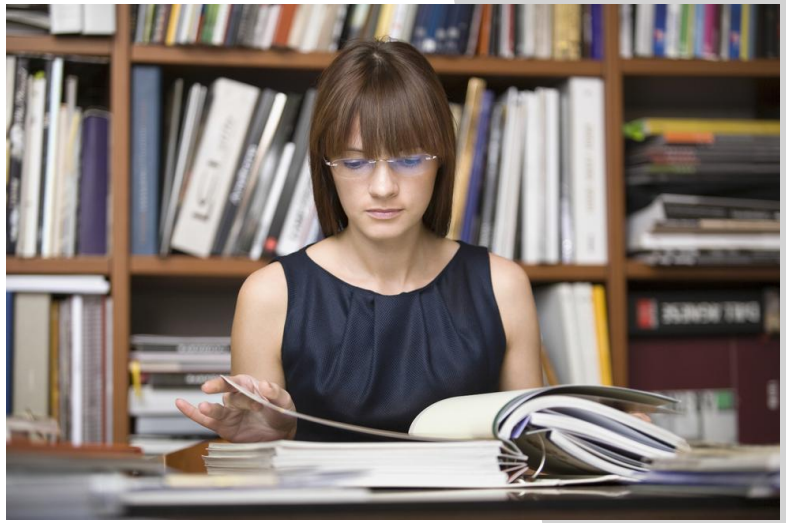
- Characterize data sets used to assess a data science project
- Explain what Big data is
- Understand the V's in Big data
- Understand and analyse the growth laws: Moore's Law, Koomey's Law, Bell's Law and Zimmerman's Law
- *Analyze and use shell commands to read and manipulate big data*

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.ibmbigdatahub.com/infographic/four-vs-big-data
- https://www.digitalinformationworld.com/2019/04/what-happens-online-in-60-seconds.html

MONASH University

Tutorials Week 9

Introduction to Shell Scripting