

Introduction to Data Science Module 3

Data Types and Storage

FIT1043 2017 Lecture 4

Monash University

Reminder: Assessment

- ▶ all the Python you need to know is covered in tutorials
- ▶ **Assignment 1:** motion chart using Python code
 - ▶ Also: expect a question in exam like:
 - “compare Python as a data science tool with X” or
 - “interpret a given line of Python code”...
- ▶ **the exam** is based on the lectures
 - ▶ use Alexandria as readings to support understanding

Discussion: Python Language

- ▶ easy to learn
- ▶ flexible and multi-purpose
- ▶ great libraries
- ▶ good visualization for basic analysis

Unit Schedule: Modules

| Module | Week | Content |
|--------|------|---|
| 1. | 1 | Overview and look at projects (Job) roles, and the impact |
| | 2 | |
| 2. | 3 | Data business models / application areas |
| 3. | 4 | Characterising data and "big" data Data sources and case studies |
| | 5 | |
| 4. | 6 | Resources and standards Resources case studies |
| | 7 | |
| 5. | 8 | Data analysis theory Regression and decision trees Data analysis process |
| | 9 | |
| | 10 | |
| 6. | 11 | Issues in data management GUEST SPEAKER & EXAM INFO |
| | 12 | |

Characterising Data

(ePub section 3.1)

some general characteristics of data sets used to assess a data science project

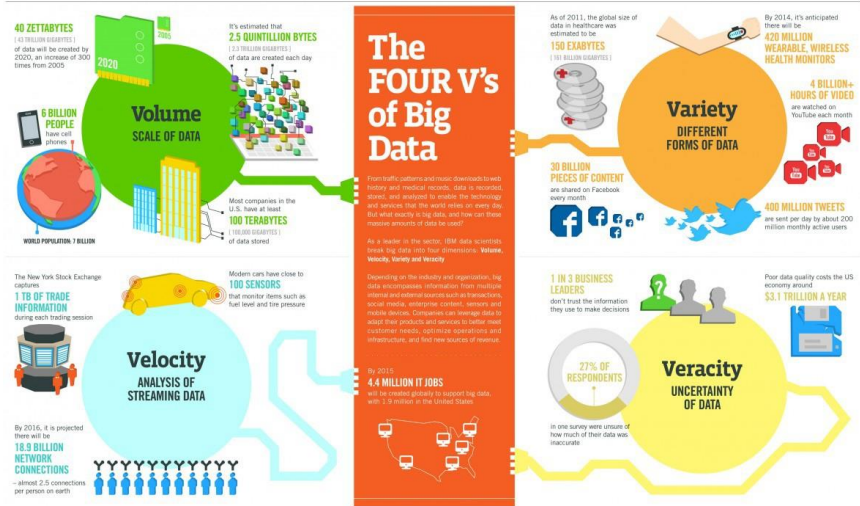
- the Four V's of big data
- metadata
 - data about data is critical to understanding
- growth laws
 - understanding the exponential growth

Characterising Data

The V's

The Four V's of Big Data

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



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

IBM

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, ...*

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
- ◆ think of any more? write a blog!

MARS 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.

MARS Question

video "[Big Ideas: How Big is Big Data?](#)" by Patricia Florissa

Which of the following is considered as big data?

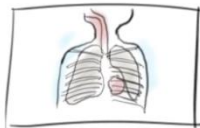


40MB



A. power point presentation

1TB



B. Healthcare image

1 PB



C. Movie

Summary

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

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 (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)

MARS Question



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



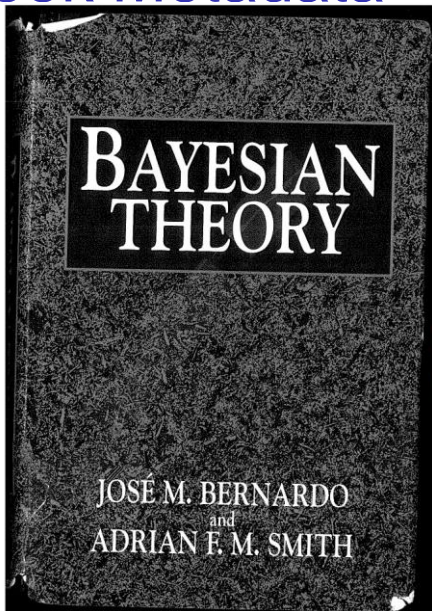
EXIF Metadata

The screenshot shows the Photo Data Explorer application window. The main window displays a large image of the Taj Mahal. To the left is a thumbnail grid of various images, including 'Tabby.jpg', 'Taj Mahal.jpg', 'Target.jpg', 'The Kiss.jpg', 'The Way.jpg', 'Thoughtful.jpg', 'Thunder.jpg', 'Tilt.jpg', 'Toco Toucan.jpg', 'Tower.jpg', 'Tree.jpg', 'Trees.jpg', 'Tropical.jpg', and 'Tube.jpg'. The 'Taj Mahal.jpg' thumbnail is selected. Below the thumbnails, the file name 'Filename: Taj Mahal.jpg' and folder path 'Folder: C:\Users\Mike\Pictures\Slide Shows\' are displayed. The right panel shows the EXIF Properties tab with a table of metadata.

| Item | Details |
|--------------------------|------------------------|
| Image Description | DCF 1.0 |
| Make | Minolta Co., Ltd. |
| Model | DIMAGE S30-4 |
| 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" |

<http://www.alexolan.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,
Reusdale, 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.S.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

1994/1/17/02416 4

Media Metadata

Format Container: .avi, .mp4, .mov, .ogg, .flv, .mkv, etc.

Video codec:

H.264,
VC-1,
Theora,
Dirac 2.1,
H.263,
etc.

Audio codec:

AAC,
WMA,
Vorbis,
PCM,
etc.

Captioning,
Video description:

SAMI, SMIL,
Hi-Caption,
CMML, DXFP,
3GPP TS 26.245,
MPSub,
etc.

Metadata:

Author,
Title,
Location,
Date,
Copyright,
License,
etc.

Examples: Call Data Record

Asterisk CDR Statistics

Select database server: Home PBX Quick search

Overview Month Report Year Report Advanced Search Help & Support Sunday 27 March 2011, 15:39

Search Results

Display 10 CDRs Search:

| Unique ID | Date | Time | Caller ID | Source | Destination | Context | Last app. | Duration | Billable | Disposition |
|---------------------------------|------------|----------|-----------|--------|-------------|-------------|-----------|----------|----------|-------------|
| 1294141101.3055 | 04/01/2011 | 11:38:21 | | | | from-sip | Dial | 00:00:28 | 00:00:02 | ANSWERED |
| 1294149884.3057 | 04/01/2011 | 14:04:44 | | | | gradwell-in | Queue | 00:06:36 | 00:06:30 | ANSWERED |
| 1294412324.3097 | 07/01/2011 | 14:58:44 | | | | gradwell-in | VoiceMail | 00:01:45 | 00:00:57 | ANSWERED |
| 1294693305.3115 | 10/01/2011 | 21:01:45 | | | | gradwell-in | VoiceMail | 00:01:12 | 00:00:25 | ANSWERED |
| 1294741395.3124 | 11/01/2011 | 10:23:15 | | | | gradwell-in | Queue | 00:00:59 | 00:00:52 | ANSWERED |
| 1294927333.3165 | 13/01/2011 | 14:02:13 | | | | gradwell-in | Queue | 00:04:13 | 00:04:01 | ANSWERED |
| 1294927997.3170 | 13/01/2011 | 14:13:17 | | | | gradwell-in | Queue | 00:05:49 | 00:05:32 | ANSWERED |
| 1295090917.3224 | 15/01/2011 | 11:28:37 | | | | from-sip | Dial | 00:04:43 | 00:04:32 | ANSWERED |
| 1295377331.3240 | 18/01/2011 | 19:02:11 | | | | gradwell-in | Queue | 00:01:35 | 00:01:32 | ANSWERED |
| 1295378544.3258 | 18/01/2011 | 19:22:24 | | | | gradwell-in | Queue | 00:00:23 | 00:00:21 | ANSWERED |

Showing 1 to 10 of 47 CDRs. First Previous 1 2 3 4 5 Next Last

You may alter your search parameters using the form opposite and re-submit your search.

Date range01/01/2011to03/27/2011

Criteria 1Destinationcontains0161

Criteria 2Caller IDcontains

Criteria 3Caller IDcontains

Criteria 4Caller IDcontains

Add criteriaSearch

Asterix Call Detail Record for an IP phone system

Examples: Javadoc

Self documenting code

```
/**
 * <h1>Function Description</h1> using <b>HTML Tags</b> and {@literal <b> JavaDoc </b> }
 * <ul><li>HTML list element 1</li><li>HTML list element 2</li></ul>
 * For more details: {@link http://www.dvteclipse.com/documentation/sv/Export_HTML_Documentation.html DVT Documentation}
 *
 * @param slave_name - first param
 * @param min_addr - second param
 * @param max_addr - third param
 * @return min_addr
 *
 * @see get_type
 * @see build_phase
 *
 * @author Author's name
 * @version 1.0
 */
function void set_slave_address_map(string slave_name,
int min_addr, int max_addr);
ubus_slave_monitor tmp_slave_monitor;
if( bus_monitor != null ) begin
    // Set slave address map for bus monitor
    bus_monitor.set_slave_configs(slave_name, min_addr, max_addr);
end
// Set slave address map for slave monitor
$cast(tmp_slave_monitor, lookup({slave_name, ".monitor"}));
tmp_slave_monitor.set_addr_range(min_addr, max_addr);
return min_addr;
endfunction : set_slave_address_map
```

public void

set_slave_address_map(string slave_name, int min_addr, int max_addr)

Function Description

using **HTML Tags** and ** JavaDoc **

- HTML list element 1
- HTML list element 2

For more details: [DVT Documentation](http://www.dvteclipse.com/documentation/sv/Export_HTML_Documentation.html)

Returns:

min_addr

Arguments:

slave_name - first param
min_addr - second param
max_addr - third param

See Also:

[get_type](#)
[build_phase](#)

Version:

1.0.

MetaData: Key Concepts

Machine-readable data: data (or metadata) which is in a format that can be understood by a computer

e.g., XML, JSON

Markup language: system for annotating a document in a way that is syntactically distinguishable from the text

e.g., Markdown, Javadoc

Digital container: file format whose specification describes how different elements of data and metadata coexist in a computer file

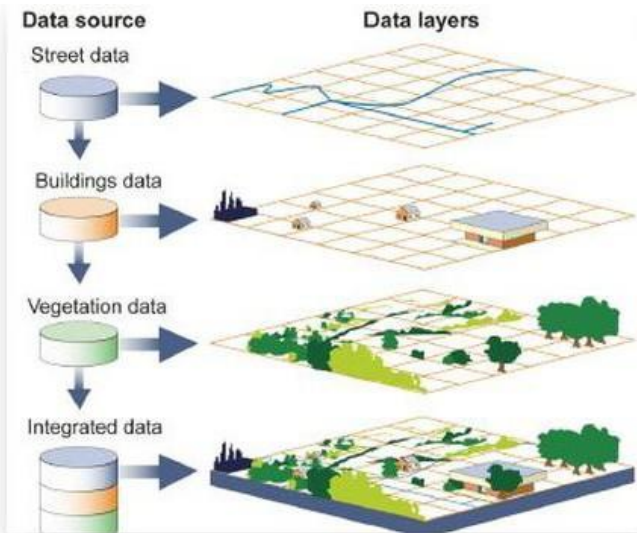
e.g., MPEG

Characterising Data

Kinds of data

a quick walkthrough of different data types

Geospatial Data



Linked Open Data: XML

```
- <adjunct id="com.yahoo.page.uf.hcard" updated="2009-02-05T00:04:45" >
- <item rel="dc:subject rel:Card" resource="http://www.whitehouse.gov/
- <type typeof="vcard:VCard" resource="http://www.whitehouse.gov/
- <item rel="vcard:url" resource="http://www.whitehouse.gov/" >
  <meta property="vcard:fn">Barack Obama</meta>
  <item rel="vcard:photo" resource="http://media.linkedin.com/mpr/mpr/shrink_80_80/p/2/000/000/0ca/2b9a3fb.jpg"/>
  <meta property="vcard:title">President of the United States of America</meta>
- <item rel="vcard:adr">
  <type typeof="vcard:Address">
    <meta property="vcard:locality">Washington D.C. Metro Area</meta>
  </type>
</item>
</type>
</item>
- <item rel="dc:subject rel:Card">
- <type typeof="vcard:VCard">
  <meta property="vcard:title">President</meta>
  <item rel="vcard:org">
    <type typeof="vcard:Organization">
      <meta property="vcard:organization-name">United States of America</meta>
    </type>
  </item>
</type>
</item>
- <item rel="dc:subject rel:Card">
- <type typeof="vcard:VCard">
  <meta property="vcard:title">US Senator</meta>
  <item rel="vcard:org">
    <type typeof="vcard:Organization">
      <meta property="vcard:organization-name">US Senate (IL-D)</meta>
    </type>
  </item>
</type>
</item>
- <item rel="dc:subject rel:Card">
- <type typeof="vcard:VCard">
  <meta property="vcard:title">Senior Lecturer in Law</meta>
  <item rel="vcard:org">
    <type typeof="vcard:Organization">
      <meta property="vcard:organization-name">University of Chicago Law School</meta>
    </type>
  </item>
</type>
</item>
```

The diagram illustrates three VCard entries from the XML feed. Each entry is represented by a vertical line with two orange boxes branching off to the right, labeled 'Title' and 'Organization'. The first entry corresponds to Barack Obama, the President of the United States of America. The second entry corresponds to a US Senator from Illinois. The third entry corresponds to a Senior Lecturer in Law at the University of Chicago Law School.

IP Connection Data


The screenshot shows the ELISA web interface. At the top, there's a search bar with 'logs-dev' and a 'Submit Query' button. Below that, a query is entered: 'srcip:10.124.19.12'. The interface shows a list of records for this query. The 'Result Options' dropdown is set to 'Field Summary'. The 'Records' section shows 100 / 4154 1486 ms. The table below lists individual connection events with columns for 'Timestamp' and 'Fields'.


| | Timestamp | Fields |
|------|---------------------|--|
| Info | Tue Nov 22 08:53:20 | 1321973538.778549/vfl.pkUrp0l6[10.124.19.12]47263 209.85.225.132 443 TLSv10 TLS_ECDHE_RSA_WITH_RC4_128_SHA s2.googleusercontent.com -CN=s2.googleusercontent.com,View.ST=California,C=US 1320932962.000000 1352555962.000000 0ef6837e26d26f08700a3e03c863d4ef6jok host=165.189.226.172 program=bro_ssl class=BRO_SSL srcip=10.124.19.12 srcport=47263 dstip=209.85.225.132 dstport=443 expiration=1352555962 hostname=s2.googleusercontent.com s2.googleusercontent.com View.ST=California,C=US |
| Info | Tue Nov 22 08:53:20 | 1321973537.891299 oE6L8vUv7[10.124.19.12]41018 199.59.149.198 443 TLSv10 TLS_RSA_WITH_RC4_128_SHA twitter.com 970e68f4de429d78dc280f310267aa67ee8530e8be2e3ec92- Inc.streetAddress=795 Folsom St, Suite 600,L=San Francisco,ST=California,postalCode=94107,C=US,serialNumber=4337446.2.5.4.15=#131450726976617465204F7267616E697A6174696F6E,1.3.6.1.4.1.311.60.2.1.2=#1308446 1310014804 host=165.189.226.172 program=bro_ssl class=BRO_SSL srcip=10.124.19.12 srcport=41018 dstip=199.59.149.198 dstport=443 expiration=1343451599 hostname=twitter.com subject=CN=twi Folsom St, Suite 600,L=San Francisco,ST=California,postalCode=94107,C=US,serialNumber=4337446.2.5.4.15=#131450726976617465204F7267616E697A6174696F6E,1.3.6.1.4.1.311.60.2.1 |
| Info | Tue Nov 22 08:53:25 | Teardown UDP connection 144744478313156395 for DET-SEC-124.19:10.124.19.12/45091 to OUTSIDE:10.68.15.11/53 duration 0:02:03 bytes 213 host=165.189.82.68 program=%fwsm-5-302016 class=FIREWALL_CONNECTION_END proto=UDP srcip=10.124.19.12 srcport=45091 dstip=10.68.15.11 dstport=53 conn_bytes=213 o_int=DE |
| Info | Tue Nov 22 08:53:25 | Teardown UDP connection 144744478313156396 for DET-SEC-124.19:10.124.19.12/52757 to OUTSIDE:10.68.15.11/53 duration 0:02:02 bytes 213 host=165.189.82.68 program=%fwsm-5-302016 class=FIREWALL_CONNECTION_END proto=UDP srcip=10.124.19.12 srcport=52757 dstip=10.68.15.11 dstport=53 conn_bytes=213 o_int=DE |
| Info | Tue Nov 22 08:53:26 | Teardown UDP connection 144744478313156397 for DET-SEC-124.19:10.124.19.12/47309 to OUTSIDE:10.68.15.11/53 duration 0:02:03 bytes 217 host=165.189.82.68 program=%fwsm-5-302016 class=FIREWALL_CONNECTION_END proto=UDP srcip=10.124.19.12 srcport=47309 dstip=10.68.15.11 dstport=53 conn_bytes=217 o_int=DE |
| Info | Tue Nov 22 08:53:26 | Teardown UDP connection 144744478313156398 for DET-SEC-124.19:10.124.19.12/52485 to OUTSIDE:10.68.15.11/53 duration 0:02:03 bytes 284 host=165.189.82.68 program=%fwsm-5-302016 class=FIREWALL_CONNECTION_END proto=UDP srcip=10.124.19.12 srcport=52485 dstip=10.68.15.11 dstport=53 conn_bytes=284 o_int=DE |
| Info | Tue Nov 22 08:53:26 | Teardown UDP connection 144744478313156399 for DET-SEC-124.19:10.124.19.12/57404 to OUTSIDE:10.68.15.11/53 duration 0:02:03 bytes 172 host=165.189.82.68 program=%fwsm-5-302016 class=FIREWALL_CONNECTION_END proto=UDP srcip=10.124.19.12 srcport=57404 dstip=10.68.15.11 dstport=53 conn_bytes=172 o_int=DE |
| Info | Tue Nov 22 08:54:20 | Teardown UDP connection 144744478313156408 for DET-SEC-124.19:10.124.19.12/35728 to OUTSIDE:10.68.15.11/53 duration 0:02:03 bytes 221 host=165.189.82.68 program=%fwsm-5-302016 class=FIREWALL_CONNECTION_END proto=UDP srcip=10.124.19.12 srcport=35728 dstip=10.68.15.11 dstport=53 conn_bytes=221 o_int=DE |
| Info | Tue Nov 22 08:54:20 | Teardown UDP connection 144744478313156409 for DET-SEC-124.19:10.124.19.12/43103 to OUTSIDE:10.68.15.11/53 duration 0:02:03 bytes 221 host=165.189.82.68 program=%fwsm-5-302016 class=FIREWALL_CONNECTION_END proto=UDP srcip=10.124.19.12 srcport=43103 dstip=10.68.15.11 dstport=53 conn_bytes=221 o_int=DE |
| | Tue Nov 22 | Teardown UDP connection 144744478313156410 for DET-SEC-124.19:10.124.19.12/51752 to OUTSIDE:10.68.15.11/53 duration 0:02:02 bytes 198 |


Transactional Data


| Trans | | | | | | |
|--|----------------|----------------|---------------------|--------------------------------|------------------------|------------|
| Entity: All Entity Time: Default member: All Time Calendar Account: All Account | | | | | | |
| | Credit | Debit | Account | Entity | Transaction | Txn Date |
| All Transactions | \$2,441,364.68 | \$1,402,410.62 | | | | |
| # 501, BillPaymentCheck | \$0.00 | \$625.00 | Checking | Wheeler's Tile Etc. | BillPaymentCheck # 501 | 12/15/2012 |
| # none, Transfer | \$0.00 | \$500.00 | Savings | None | Transfer # none | 12/15/2012 |
| # none, ReceivePayment | \$440.00 | \$0.00 | Undeposited Funds | Roche, Diarmuid Garage repairs | ReceivePayment # none | 12/15/2012 |
| # none, Bill | \$0.00 | \$670.00 | Accounts Payable | Keswick Insulation | Bill # none | 12/15/2012 |
| # 6236, PurchaseOrder | \$0.00 | \$65.00 | Purchase Orders | Daigle Lighting | PurchaseOrder # 6236 | 12/15/2012 |
| # 502, BillPaymentCheck | \$0.00 | \$640.92 | Checking | Daigle Lighting | BillPaymentCheck # 502 | 12/15/2012 |
| # 503, BillPaymentCheck | \$0.00 | \$754.50 | Checking | Palton Hardware Supplies | BillPaymentCheck # 503 | 12/15/2012 |
| # 1097, Invoice | \$12,420.98 | \$0.00 | Accounts Receivable | Robson, Darci Robson Clinic | Invoice # 1097 | 12/15/2012 |
| # 504, BillPaymentCheck | \$0.00 | \$6,935.75 | Checking | Perry Windows & Doors | BillPaymentCheck # 504 | 12/15/2012 |
| # 505, BillPaymentCheck | \$0.00 | \$45.00 | Checking | Lew Plumbing | BillPaymentCheck # 505 | 12/15/2012 |
| # 12/03, Bill | \$0.00 | \$122.68 | Accounts Payable | Cal Gas & Electric | Bill # 12/03 | 12/15/2012 |
| # 506, BillPaymentCheck | \$0.00 | \$1,631.52 | Checking | East Bayshore Tool & Supply | BillPaymentCheck # 506 | 12/15/2012 |
| # 507, BillPaymentCheck | \$0.00 | \$1,358.00 | Checking | Timberloft Lumber | BillPaymentCheck # 507 | 12/15/2012 |
| # 508, BillPaymentCheck | \$0.00 | \$1,476.23 | Checking | East Bayshore Tool & Supply | BillPaymentCheck # 508 | 12/15/2012 |
| # 509, BillPaymentCheck | \$0.00 | \$450.00 | Checking | Hopkins Construction Rentals | BillPaymentCheck # 509 | 12/15/2012 |
| # 510, BillPaymentCheck | \$0.00 | \$896.00 | Checking | Timberloft Lumber | BillPaymentCheck # 510 | 12/15/2012 |
| # 511, BillPaymentCheck | \$0.00 | \$696.52 | Checking | East Bayshore Tool & Supply | BillPaymentCheck # 511 | 12/15/2012 |
| # 512, BillPaymentCheck | \$0.00 | \$400.00 | Checking | Palton Hardware Supplies | BillPaymentCheck # 512 | 12/15/2012 |
| # 513, BillPaymentCheck | \$0.00 | \$1,610.00 | Checking | Timberloft Lumber | BillPaymentCheck # 513 | 12/15/2012 |

Twitter Data

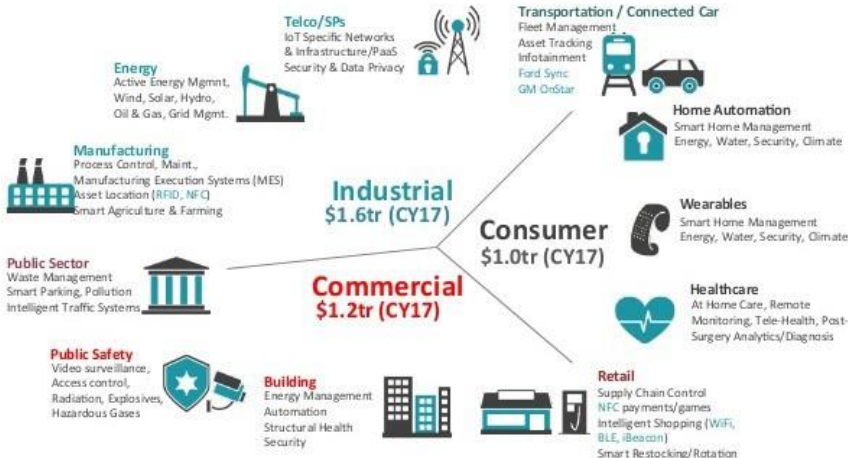
**Brian D. Earp** @briandavidearp · 3h
Major publisher retracts 64 scientific papers in fake **peer review** outbreak - The Washington Post [washingtonpost.com/news/morning-m...](https://www.washingtonpost.com/news/morning-m...)
6 4 [View summary](#)

**Christina Larson** @larsonchristina · 4h
Scientific publisher Springer retracts 64 papers - mostly by Chinese academics - for fake **peer review**: blogs.wsj.com/chinarealtime/... @feliciasonmez
3 4 [View summary](#)

**Felicia Sonmez** @feliciasonmez · 4h
A publisher has retracted 64 articles for fake peer reviews. Nearly all were from China. By me for @ChinaRealTime: on.wsj.com/1NQDVez
9 7 [View summary](#)

**Academic Life in EM** @ALiEMteam · 8h
CAPSULES module 2 is out! Pressors & Inotropes
By: @iEMPharmD & @DougEDPharm
Peer-review: @EMPharm & @DavidJuurlink
aliemu.com/courses/presso...
9 10

Internet of Things Data



Source: IDC Internet of Things Spending Guide by Vertical Market 2014

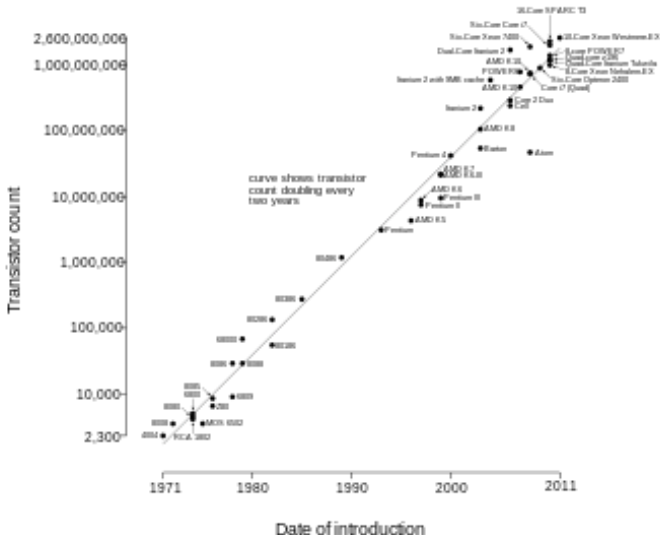
Characterising Data

Growth laws

understanding the exponential growth

Moore's Law

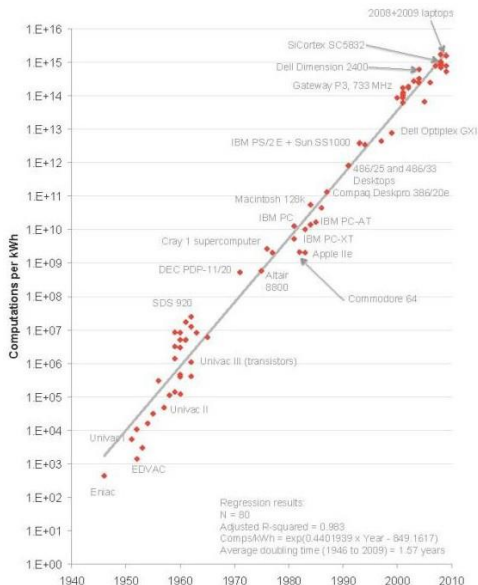
Microprocessor Transistor Counts 1971-2011 & Moore's Law



Moore's Law

- ◆ stated to **double every 2 years** starting 1975
- ◆ transistor count translates to:
 - more memory
 - bigger CPUs
 - faster memory, CPUs (smaller==faster)
- ◆ pace currently slowing

Koomey's Law



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

- ◆ corollary of Moore's Law and Koomey's Law
- ◆ at a broad level, whole new computing classes will emerge every decade

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

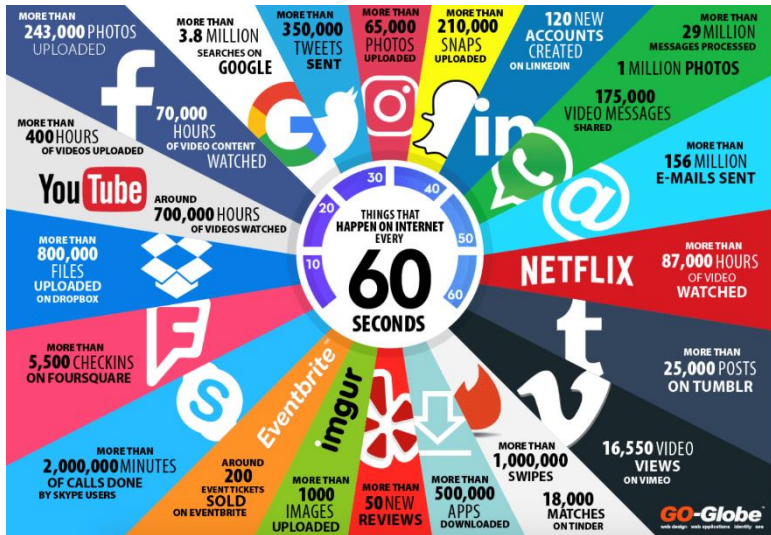
No: Java, big data, Hadoop, flash memory

Zimmerman's Law

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

Things that happen in 60secs

from [GO-Gulf](#)



Introduction to R for Data Science

MARS Question



Which of the following R commands is not a summary statistics of my_table?

- A.** `min(my_table$ages)`
- B.** `mean(my_table$height)`
- C.** `sd(my_table$height)`
- D.** `str(my_table)`

Next Week:

Distributed Processing and Data Case Studies (ePub sections 3.3, 3.4)

Homework:

- ◆ follow up on some of the case studies in 3.3 yourself!