FIT1043 Introduction to Data Science

Module 3

Data Types and Storage

Lecture 5: characterising data and "big" data

Monash University

Discussion: SAS Visual Analytics

Advantages:

- visualisation with **no coding or programming**!
- handles large amounts of data easily
- flexible range of analytics
- capability of drilling down
- more accessible to broader range of users
- cross-platform so available almost anywhere

Disadvantages:

- vendor “lock-in”

Discussion: SAS Business Model

- not a "data-based" business
- a traditional business offering tools, Application as a Service (AaaS)
- more of a high-end, commodity analytics tool

  i.e., not aimed at the boutique client with unique or sophisticated requirements
  - greater capability that Excel

Unit Schedule: Modules

Module	Week	Content
1.	1	overview and look at projects
	2	(job) roles, and the impact
2.	3	data business models
	4	application areas and case studies
3.	**5**	**characterising data and "big" data**
	6	data sources and case studies
4.	7	resources and standards
	8	resources case studies
5.	9	data analysis theory
	10	data analysis process
6.	11	issues in data management
	12	data management frameworks

Video: How Big is Big Data?

Watch the video:

- *"Big Ideas: How Big is Big Data?"* by Patricia Florissa, VP at EMC
- According to the video, what aspect of big data makes it big?

Characterising Data (ePub section 3.1)

some general characteristics of data sets used to assess a project

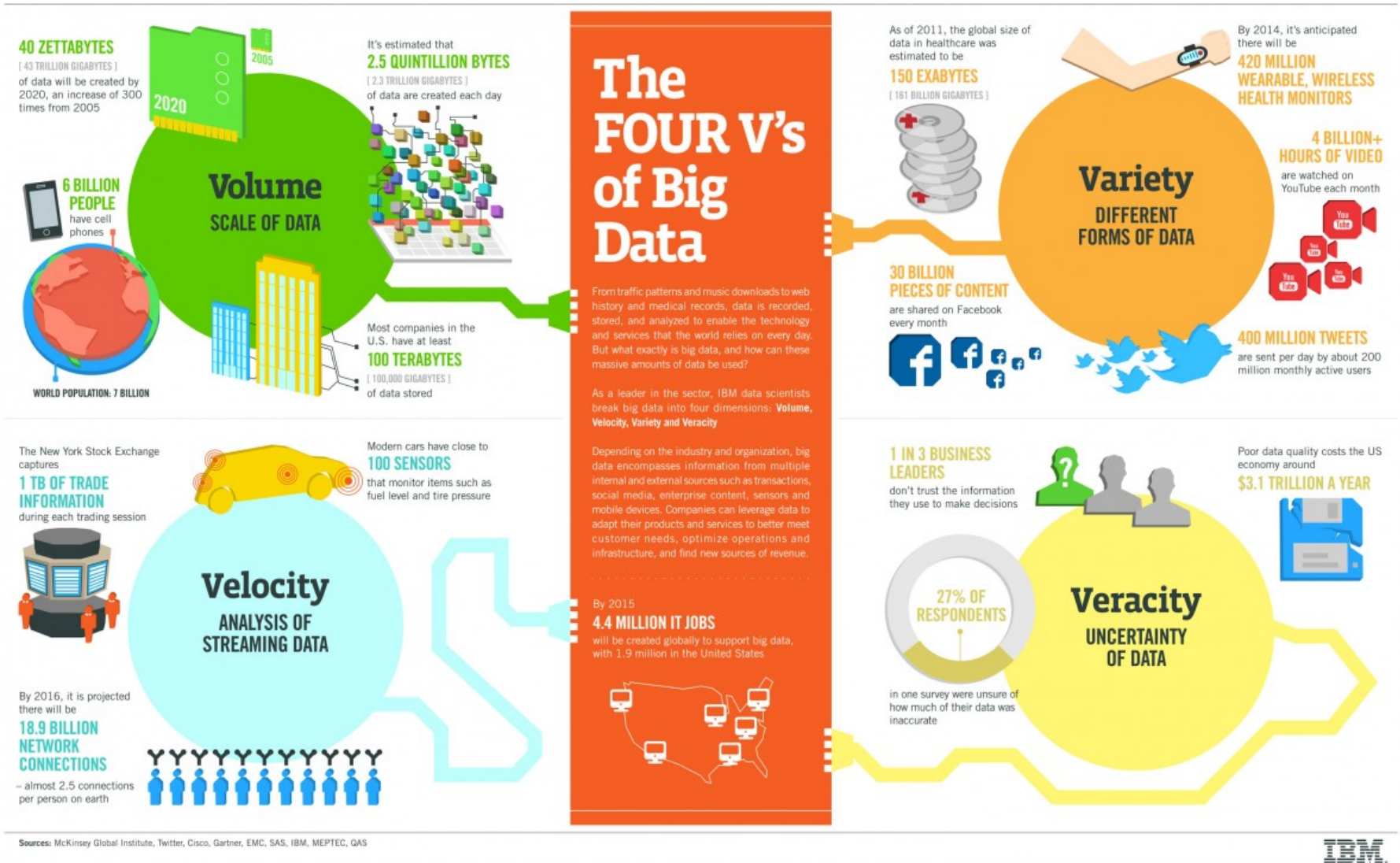
- the V's
  - the first charactisations by someone with a penchant for alliteration
- metadata
  - data about data is critical to understanding
- dimensions of data
  - infographics on data dimensions (how big is "big")
- growth laws
  - understanding the exponential growth

Characterising Data
The V's

the first charactisations 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!

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.)

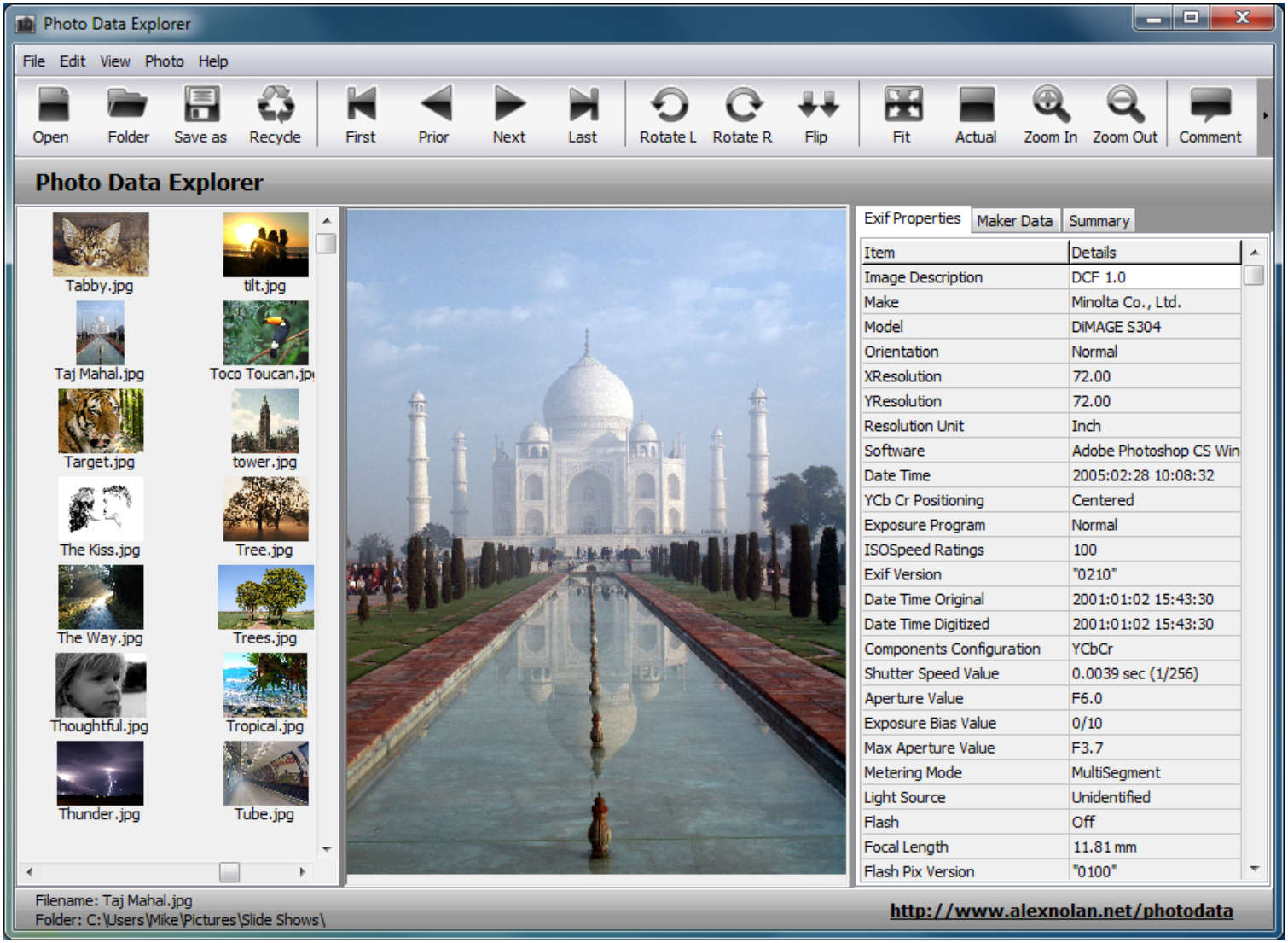
MetaData can be:

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

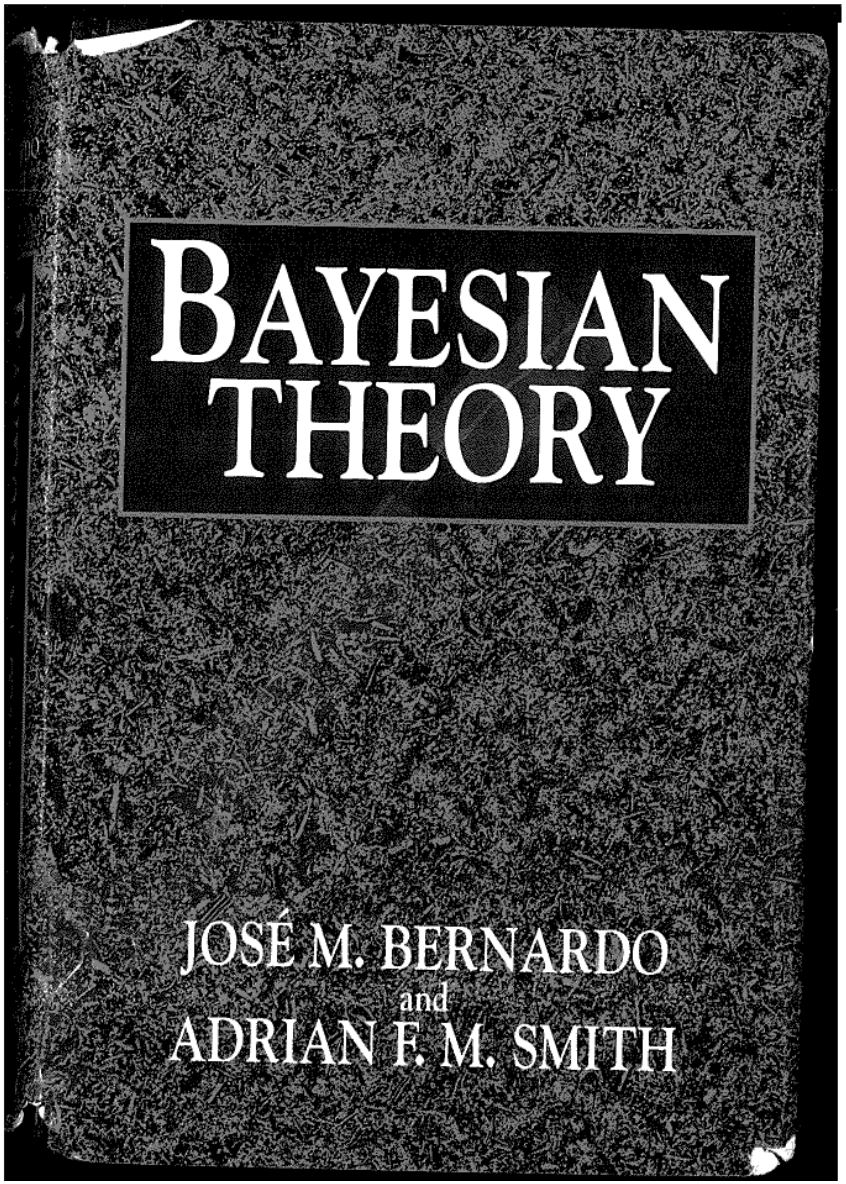
Structural: documents relationships and links
e.g. chapters in a book, elements in XML, containers in MPEG

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

Examples: EXIF Metadata



Examples: Book Metadata



Baffins Lane, Chichester
West Sussex PO19 1UD, England
National Chichester (0243) 779777
International (+44) 243 779777

Other Wiley Editorial Offices

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

Jacaranda 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

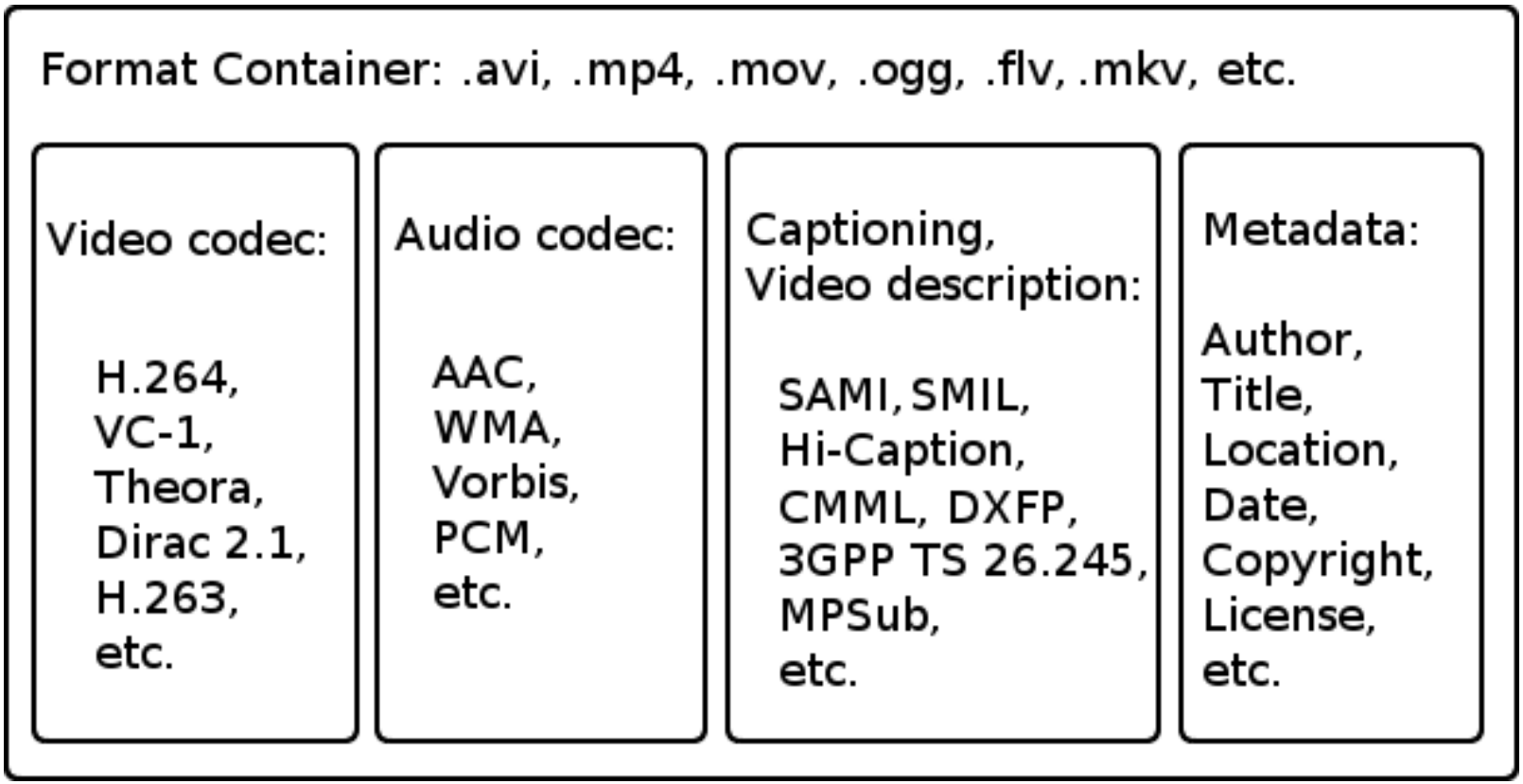
Library of Congress Cataloging-in-Publication Data
Bernardo, José M.
Bayesian theory / José M. Bernardo, Adrian F.M. Smith.
p. cm. — (Wiley series in probability and mathematical statistics)
Includes bibliographical references and indexes.
ISBN 0 471 92416 4
1. Bayesian statistical decision theory. I. 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

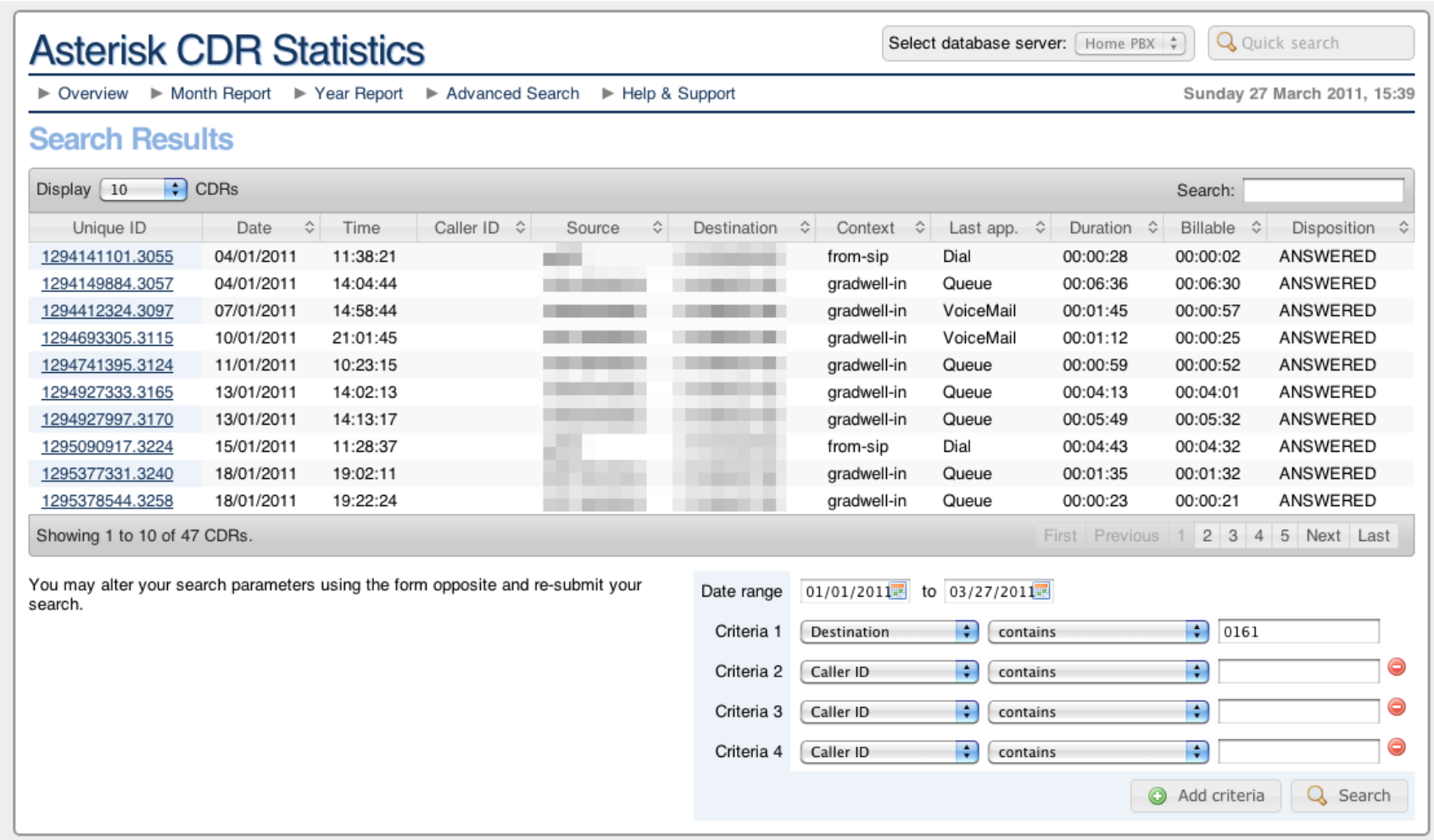
A catalogue record for this book is available from the British Library

ISBN 0 471 92416 4

Examples: Media Metadata



Examples: Call Data Record



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

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.

Other Metadata Examples

- Australian Government *Digital Service Standard* webpage
- *medical bibliographic data* in XML on PubMed,
- *Doxygen: documenting the code*

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

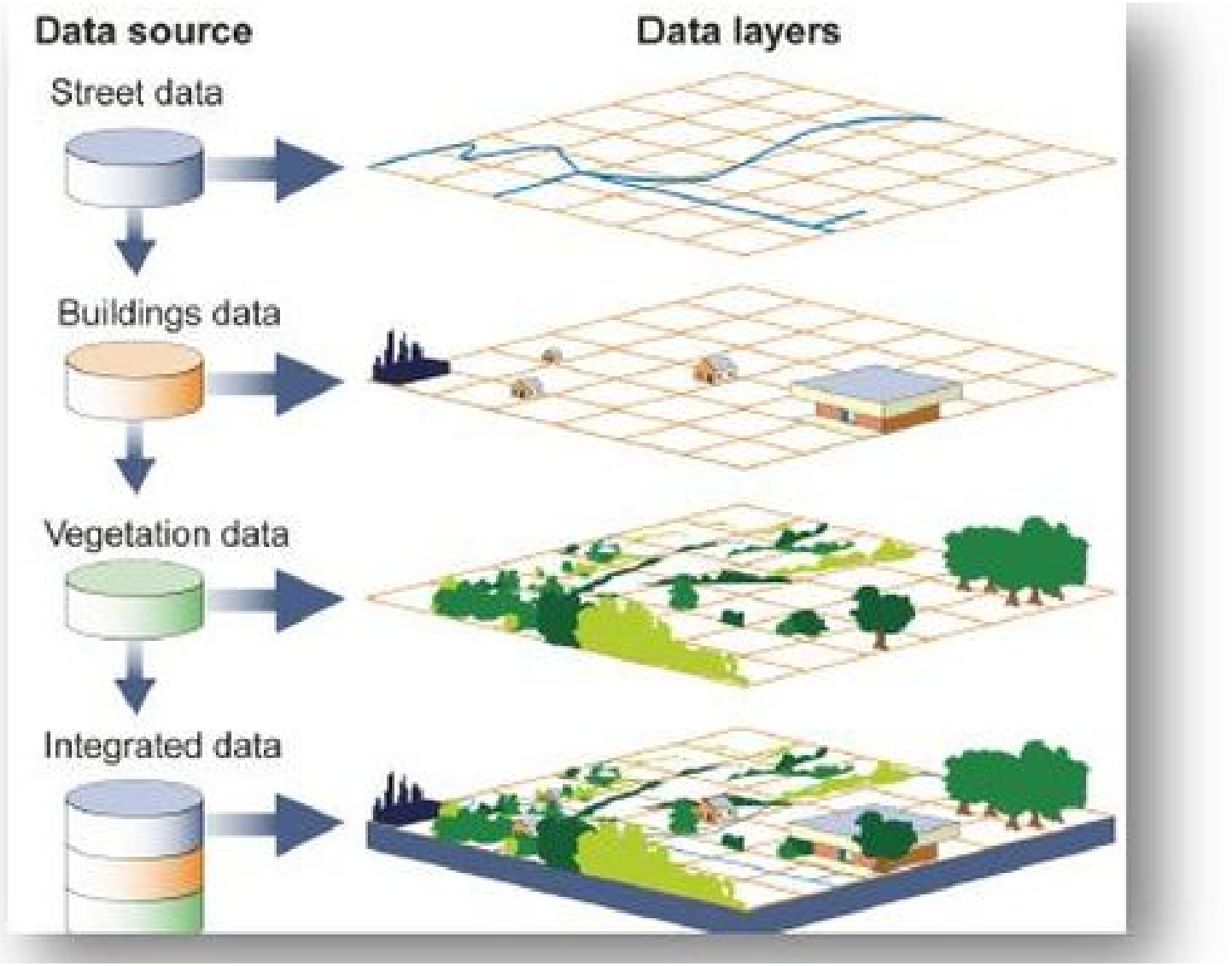
Markup language: system for annotating a document in a way that is syntactically distinguishable from the text
e.g., Markdown, Javadoc

Digital container: file format whose specification describes how different elements of data and metadata coexist in a computer file
e.g., MPEG

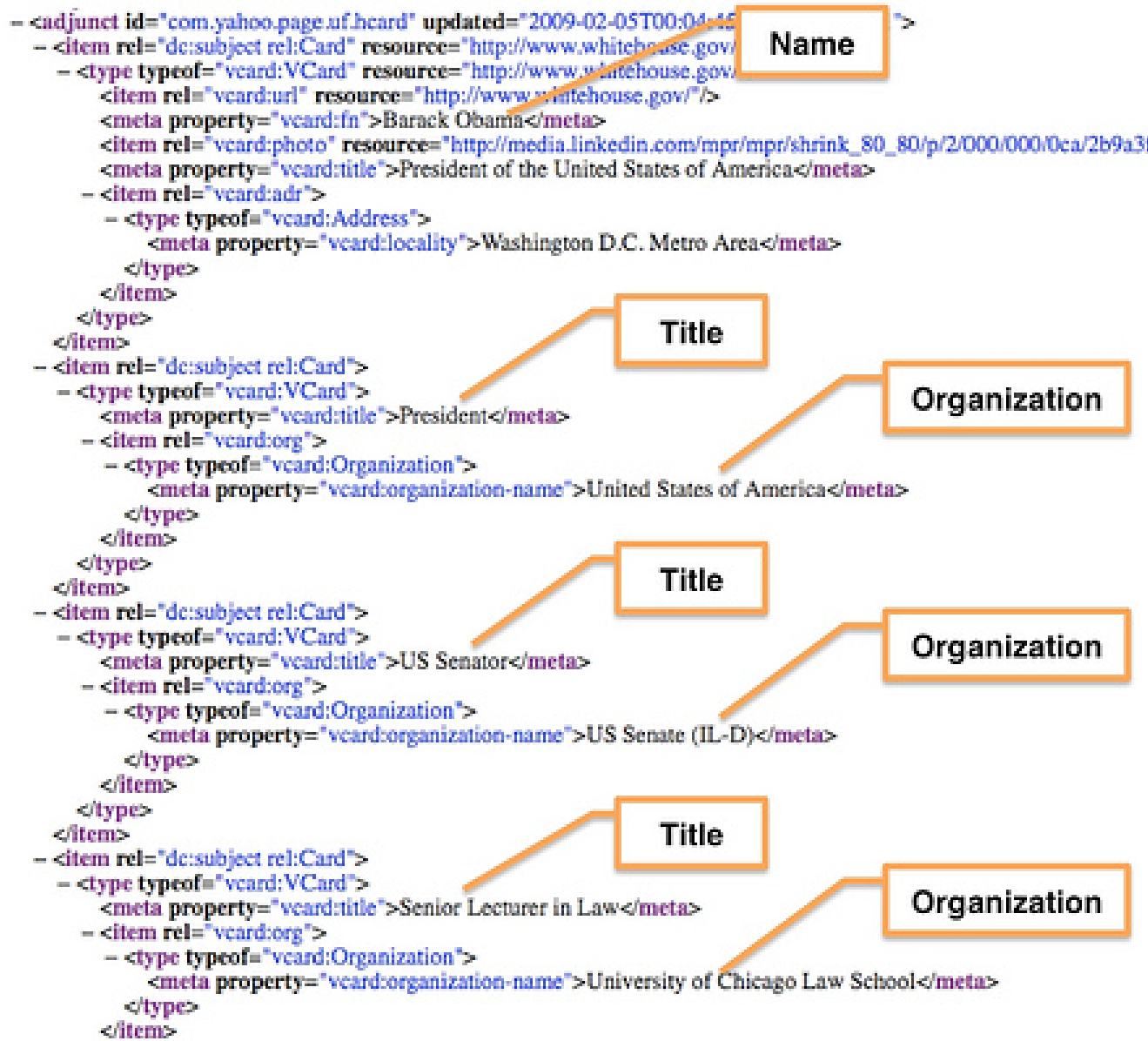
Characterising Data
Kinds of data

a quick walkthrough of different data types

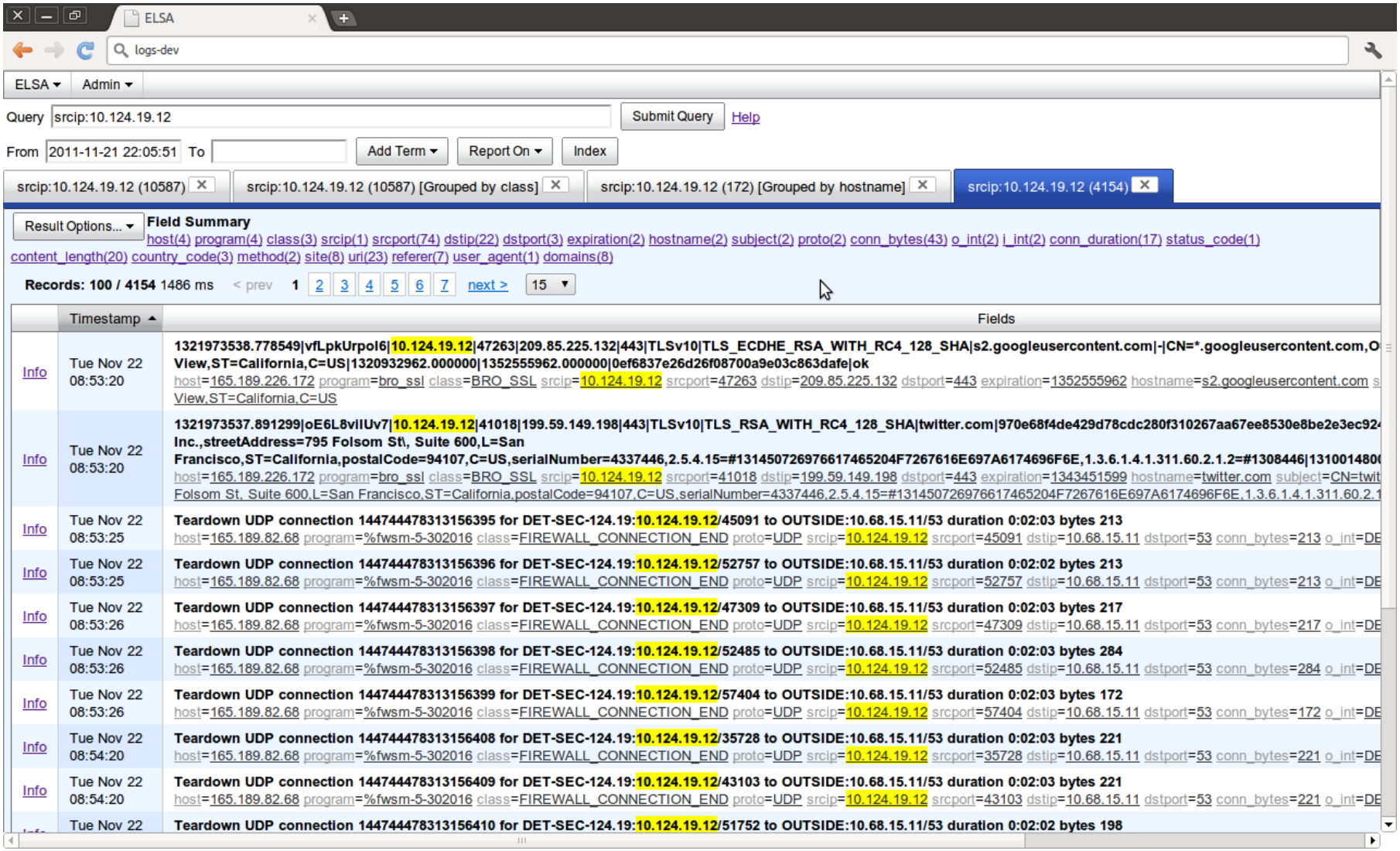
Geospatial Data



Linked Open Data: XML



IP Connection Data



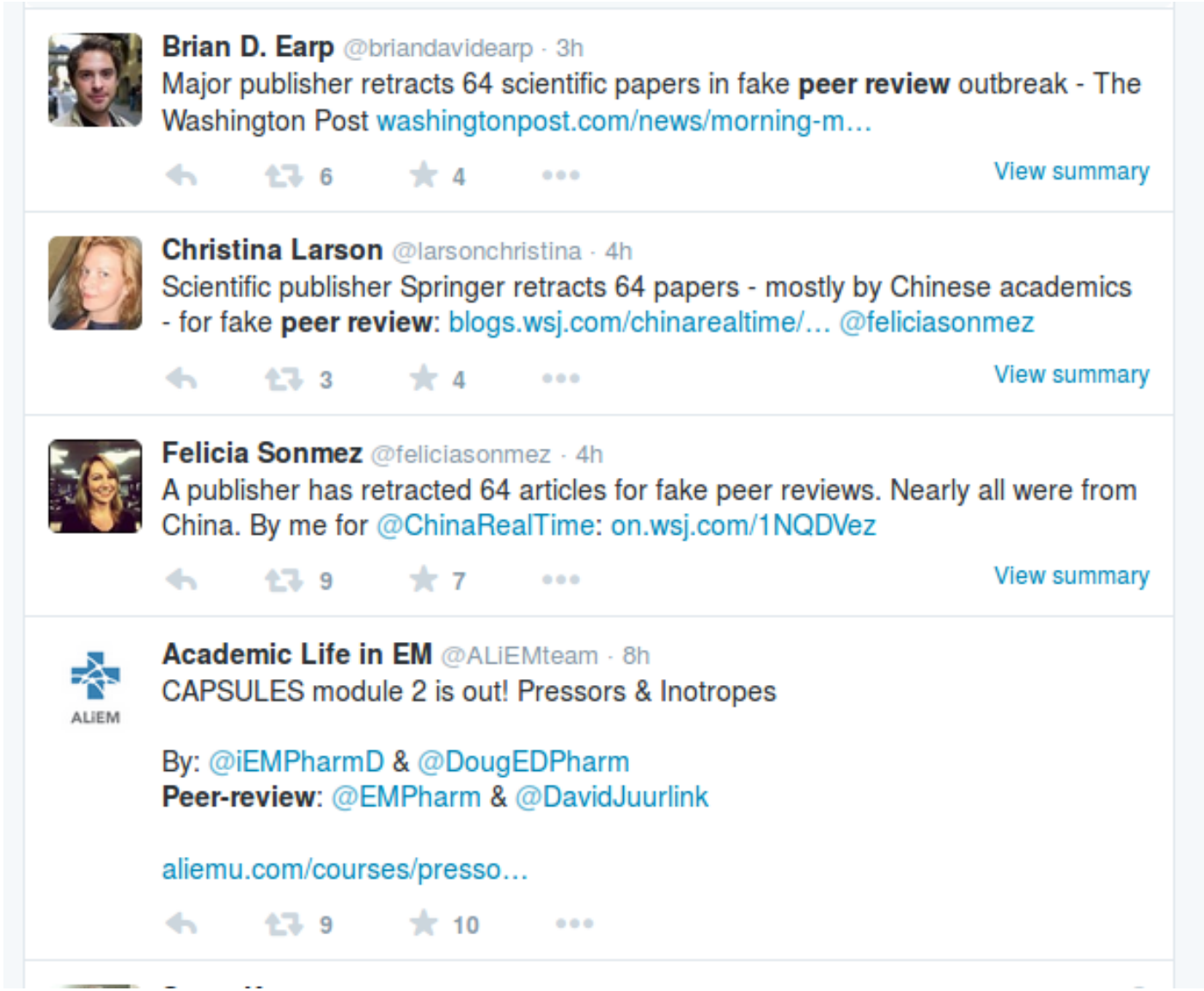
Transactional Data

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	Patton 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	Patton 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…
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

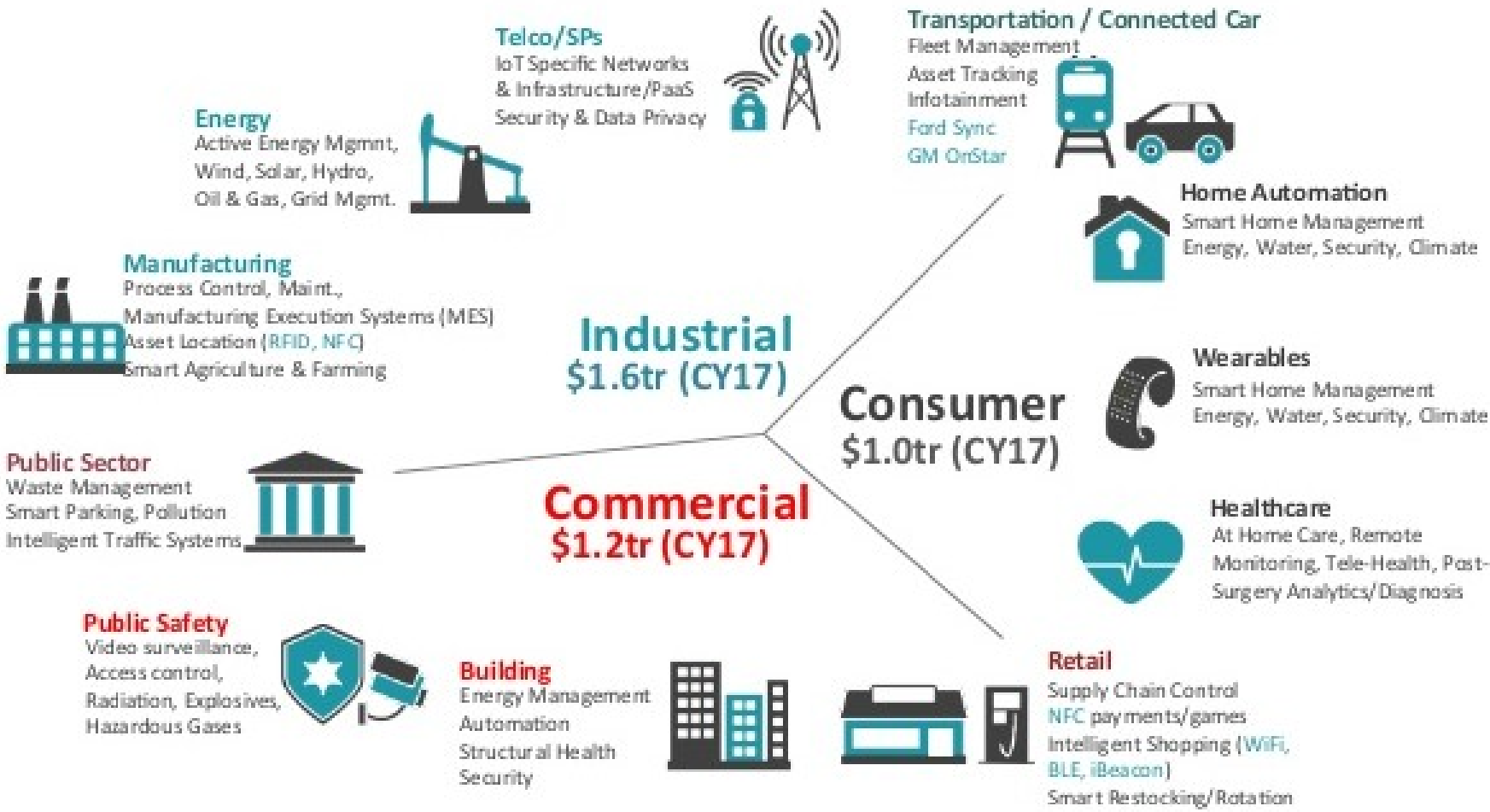
Felicia Sonmez @feliciasonmez · 4h
A publisher has retracted 64 articles for fake peer reviews. Nearly all were from China. By me for @ChinaRealTime: on.wsj.com/1NQDVez
9 7 View summary

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
Dimensions of data

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

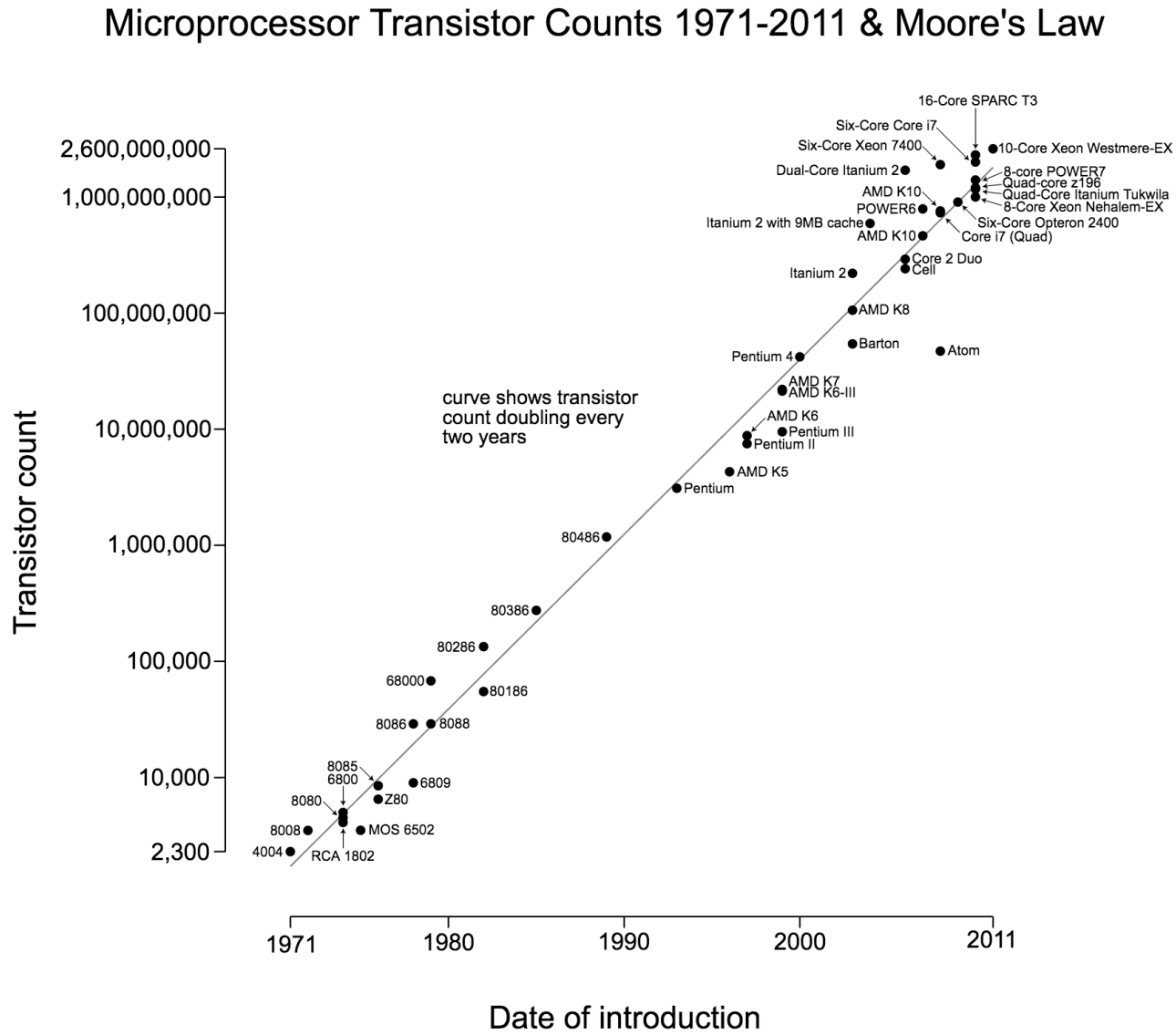
Infographics on Data

- *"Data Science Matters"* from the datascience@berkeley Blog
- *"Intelligence by Variety – Where to Find and Access Big Data"* from Kapow Software
- *"60 Seconds – Things That Happen On Internet Every 60 secs"* from GO-Gulf
- *"60 Seconds – Things That Happen Every 60 secs Part 2"* again
- Social Media Prisma from the *Ethority.net site*

Characterising Data
Growth laws

understanding the exponential growth

Moore's Law

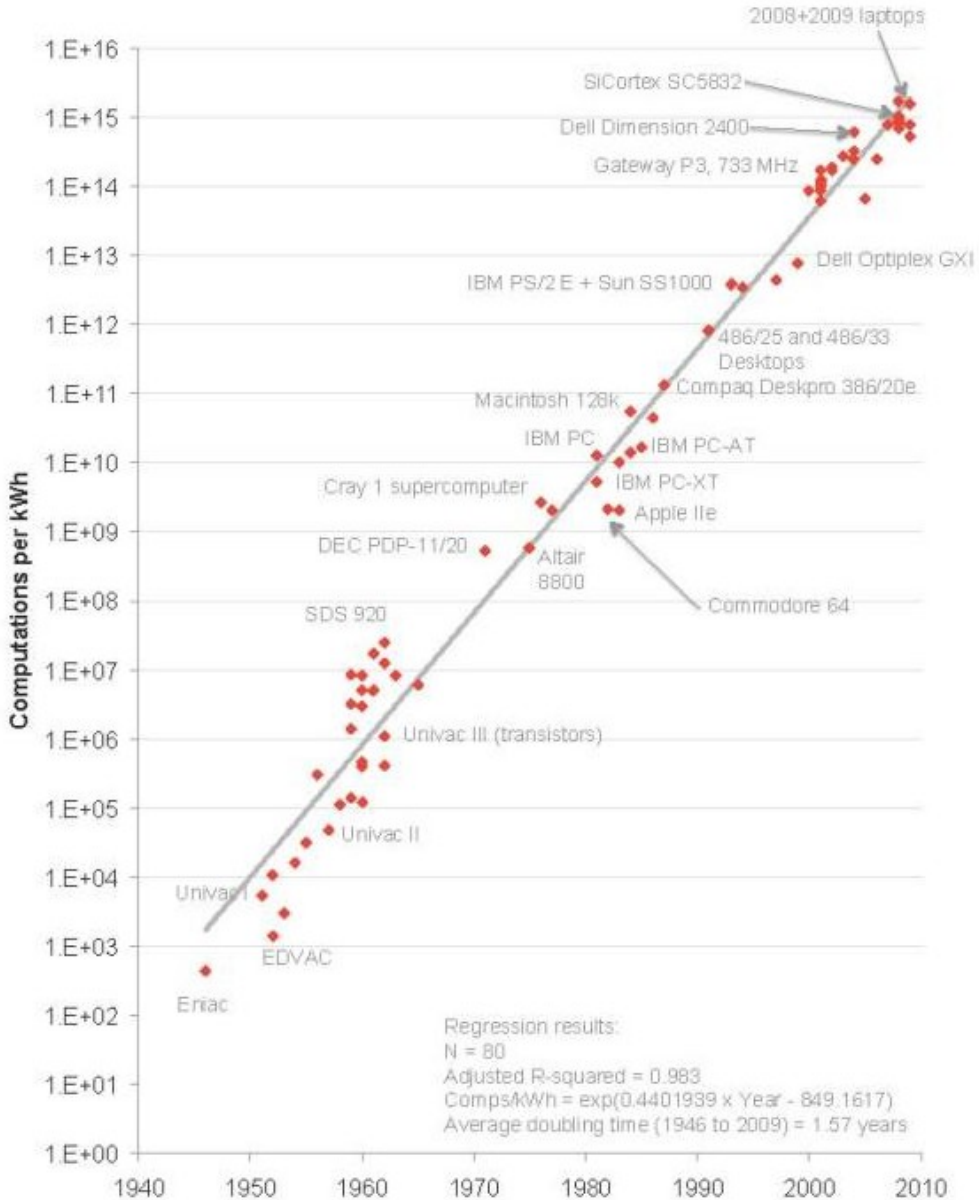


By Wgsimon (Own work) CC BY-SA 3.0, via Wikimedia Commons

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 Law



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

Koomey's Law

- corollary of Moores Law
- amount of battery power 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 Goood Privacy (PGP), an early encription system
- "surveillance is constantly increasing"
- privacy constantly decreasing

Next Week: Hadoop & Data Case Studies

(ePub sections 3.3)