Data Analytics with HPC – L01

Introduction – What are data analytics, big data, data science,…?

Reusing this material

archer

epcc

THE UNIVERSITY OF EDINBURGH

DAwHPC Course Introduction

- An overview of data science and the analytical techniques that form its basis as well as exploring how HPC provides the power that has driven its adoption.
- Course team
  - Amy Krause, Data Architect, EPCC
  - Marc Sabate, Data Scientist, EPCC
  - Eilidh Troup, Applications Consultant, EPCC
  - Terry Sloan, Group Manager, EPCC

Course Content

The course will cover:

- Key data analytical techniques such as, classification and unsupervised learning
- Key parallel patterns for implementing analytical techniques

Learning Outcomes

- Understand what data analytics, data science and big data are.
- Understand the importance of data cleaning
- Have knowledge of the common, popular, important data analytics techniques.
- Understand relevance popular HPC infrastructures applicable to data analytics.

Timetable – Day 1

Time		Description
09:00 – 09:30		Arrival, set-up, welcome
09:30 – 10:30	L00	ARCHER/PATC Training courses
	L01	What are data analytics, big data, data science?
10:30 – 11:00		COFFEE
11:00 – 12:00	L02	Data Cleaning
12:00 – 12:30	P01	Practical: Data Cleaning
13:00 – 14:00		LUNCH
14:00 – 14:45	L03	Supervised learning, feature selection, trees, forests
14:45 – 15:30	L04	Naïve Bayes
15:30 – 16:00		COFFEE
16:00 – 17:00	P02	Practical: Naïve Bayes
17:00		CLOSE OF DAY 1

Timetable – Day 2

Time	L#/D#	Description
09:00 – 10:30	L05	MapReduce
	L06	Hadoop
10:30 – 11:00		COFFEE
11:00 – 11:30	D01	Hadoop demonstration
11:30 – 12:30	L07	Unsupervised learning
12:30 – 13:30		LUNCH
13:30 – 14:15	L08	Spark
14:15 – 15:00	L09	Data streaming
15:00 – 15:30		COFFEE
15:30 – 16:00	D02	Spark, data streaming demonstrations
16:00		CLOSE OF COURSE

Contents of Lecture 1

- Lecture Aim: An understanding of what data analytics, data science and big data are.
- Definitions - What are Data Analytics, Data Science, Big Data, etc?
- EPCC background in Data Analytics with HPC
- HPC ? Why, where and how does it fit ?

DATA ANALYTICS

Definition – Data Analytics

"Analytics ... a catch-all term for a variety of different business intelligence (BI)- and application-related initiatives.

For some, it is the process of analyzing information from a particular domain, such as website analytics.

For others, it is applying the breadth of BI capabilities to a specific content area (for example, sales, service, supply chain and so on).

…

Increasingly ... is used to describe statistical and mathematical data analysis that clusters, segments, scores and predicts what scenarios are most likely to happen. …

(It) has moved deeper into the business vernacular.

(It) has garnered a burgeoning interest from business and IT professionals looking to exploit huge mounds of internally generated and externally available data."

(See Gartner http://www.gartner.com/it-glossary/analytics)

BIG DATA

What is Big Data?

- "Big Data" is a buzzword.
  - It means different things to different people
  - One person's big is another person's normal

"big" is really a red herring. Oil companies, telecommunications companies, and other data-centric industries have had huge datasets for a long time. And as storage capacity continues to expand, today's "big" is certainly tomorrow's "medium" and next week's "small."

(See Loukides, Mike (2011-04-10). "What Is Data Science?" O'Reilly Media. Kindle Edition)

Big Data – the V's

"the three V's of volume, velocity and variety are commonly used to characterise different aspects of big data"

(See "Big Data Now: 2012 Edition", O'Reilly Media. Kindle Edition)

*"Big data is **high-volume, high-velocity and high-variety** information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."*

(See Gartner IT Glossary, http://www.gartner.com/it-glossary/big-data)

Big Data - Volume

- When the amount of data is too big for conventional IT infrastructure to process eg. your PC, your departmental computer cluster, …

- The remedy is
  - Scalable storage: ability to add more storage as required (eg. more disks)
  - Distributed queries: split the data processing up into tasks that process different parts of the data at the same time

- So major attraction is the ability to process large amounts of data

Big Data – Volume (cont.)

- When volume too much for conventional relational DBMS, typically resort to
  - Massively parallel architectures for data warehouses or databases eg. shared nothing approaches like Greenplum
    - Usually means working with pre-determined schema
    - Suits a regular, slowly evolving dataset

  OR

  - Hadoop-based solutions
    - Places no restrictions on structure of data to process

Big Data - Velocity

This is concerned with both

- The rate at which data flows into an organisation

AND

- The speed at which decisions can be made based on that data. How quick is the feedback loop?
- Fast-moving data usually referred to as
  - Streaming, or
  - Complex Event processing
- Two reasons to consider processing live streamed data
  - Too much data, too quickly for storage capabilities eg. Large Hadron Collider discards most of its data
  - The application requires an immediate response to the data
- Streaming technologies
  - Apache Storm

Big Data - Variety

- Data is rarely perfectly ordered and ready for processing
- Data sources and formats are diverse
  - typically unstructured or semi-structured
  - Multiple content types eg. images, video, sensor data, tweets, …
- Conventional relational DBMS not always ideal
  - Static schemas can hinder analysis
  - Difficulty handling changing and diverse data structures
- NoSQL DBMS can help remedy this
  - MongoDB (document-based)
  - Casssandra (columnar)
  - Neo4j (graph)

Big Data - Value

There are other possible V's

- Value - the reason "Big Data" is popular
  - The large amounts of data now available mean it is now possible to do or at least consider doing things that could not have been done before
- Veracity
- Validity
- ….

DATA SCIENCE

Definition – Data Science

"… the study of the computational principles, methods, and systems for extracting knowledge from data. Although this is a new term, there is a sense in which this is not a new field, because it is the intersection of many existing areas, including: machine learning, databases, statistics, numerical optimization, algorithms, …."

(See The University of Edinburgh DTC for Data Science at http://www.inf.ed.ac.uk/student-services/data-science-cdt)

"The data scientist role is critical for organizations looking to extract insight from information assets for "big data" initiatives and requires a broad combination of skills that may be fulfilled better as a team, for example: Collaboration and team work is required for working with business stakeholders to understand business issues. Analytical and decision modeling skills are required for discovering relationships within data and detecting patterns. Data management skills are required to build the relevant dataset used for the analysis."

(See Gartner IT Glossary, http://www.gartner.com/it-glossary/data-scientist)

EPCC & Data Science?

1990 – EPCC set-up

1990-1994 – Information and Business Systems Group, Genetic Algorithms, Data Mining

1995 – Spin-out of Quadstone – customer analytics, now part of Pitney Bowes

1996-2000 – Sensor-based Condition Monitoring (Fishing, Mentor, Integriti), Customer behaviour data mining (Lloyds TSB Scotland), operational data analysis (Kwik-Fit, Arran Aromatics)

2001-2005 - Customer behaviour data mining (Cheltenham & Gloucester, First plc, INWA), eScience Data Integration (OGSA-DAI, eDIKT, ODDGenes)

2006- 2010 – Gene Analysis (SPRINT, Tenessa), Data Mining tooling (ADMIRE)

2010-2014 - customer behaviour data mining (BRIDGE), SPRINT, Tenessa cont., Seismics (VERCE)

2015 – 2016 EUDAT 2, Aviagen, Finance

2017 – Data Engineering Programme, ATI

A Data Science Team

- Data science teams rather than individuals
  - that is almost always what we have in EPCC projects.
  - Domain experts, data analysis experts, computational scientists.
    - EPCC usually provide computational expertise and some of the data analysis as well.
- Example – BRIDGE comprised
  - EPCC – computational and data analysis expertise
  - Business School – data analysis and domain expertise
  - China Retail Co – domain expertise

Data Science Skills Diagram

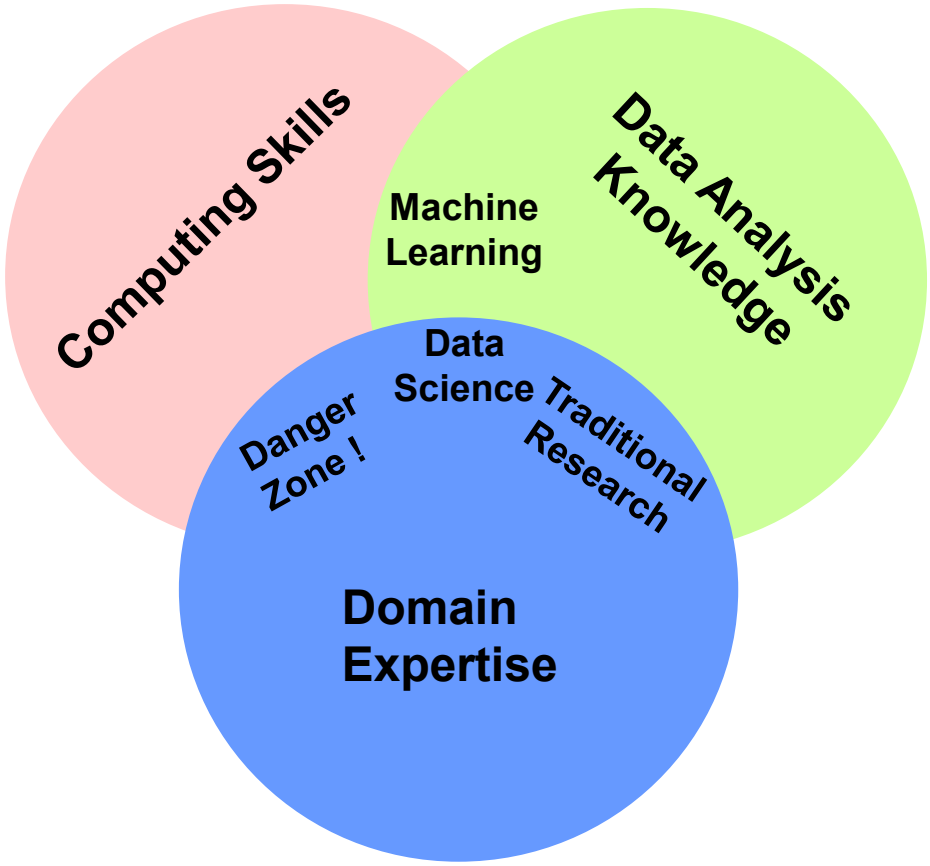
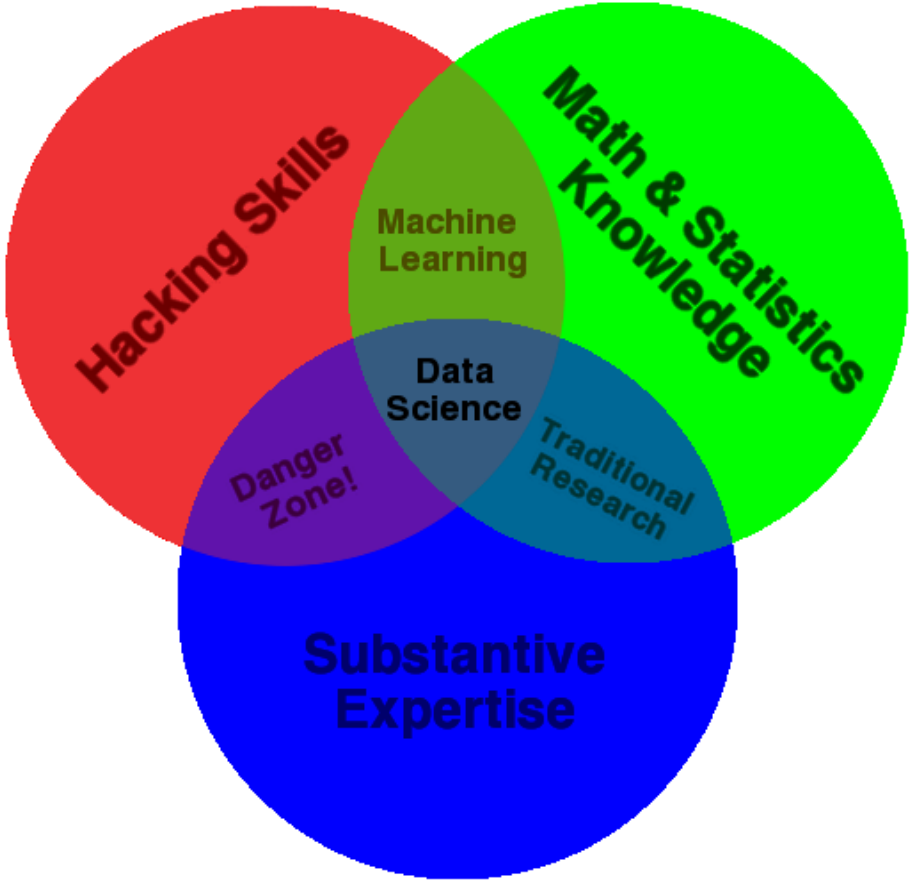


Fig1-1 from O'Neil and Shutt, "Doing Data Science". This is based on Drew Conway's Venn Diagram of Data Science

Drew Conway Venn Diagram

Drew Conway's Venn Diagram of Data Science



Hacking = Computer Science, Substantive = Domain expertise

(See http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram)

The Danger Zone!

- People who know enough to be dangerous
- Capable of extracting and structuring data
- Know quite a bit about the field and can even run a linear regression

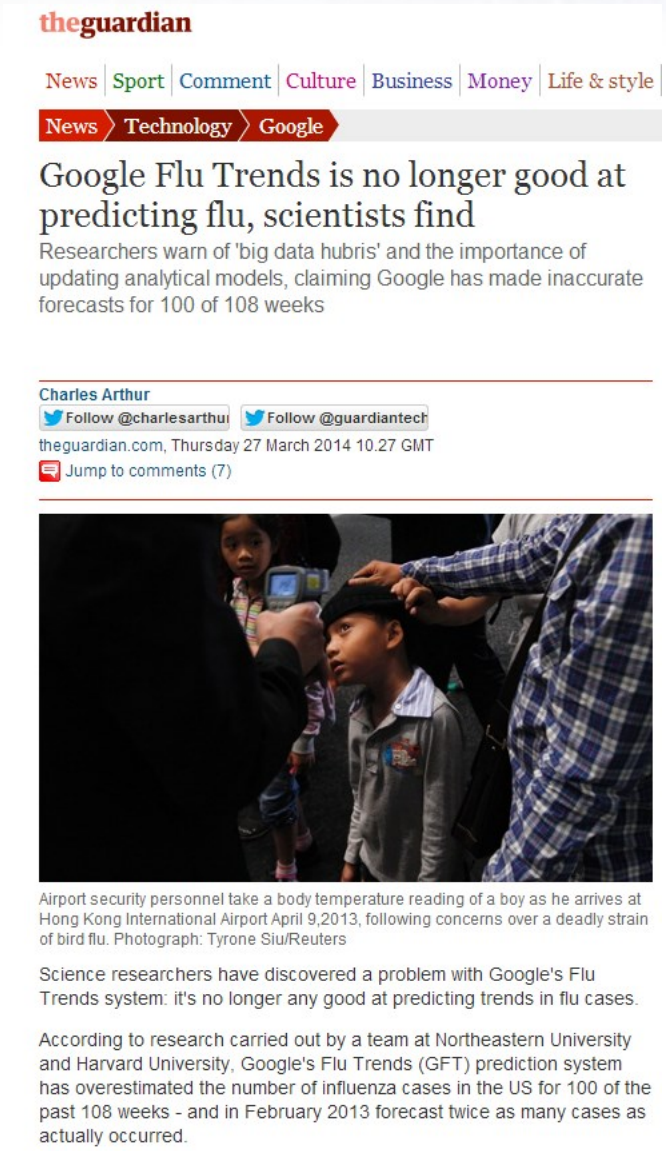
But…

- Lack understanding of what the regression coefficients mean

So…

- Have ability to create what appears to be legitimate analysis without understanding how they got there or what they have created

The Danger Zone!- Google Flu trends



theguardian

News | Sport | Comment | Culture | Business | Money | Life & style

News › Technology › Google

Google Flu Trends is no longer good at predicting flu, scientists find

Researchers warn of 'big data hubris' and the importance of updating analytical models, claiming Google has made inaccurate forecasts for 100 of 108 weeks

Charles Arthur

Follow @charlesarthur Follow @guardiantech

theguardian.com, Thursday 27 March 2014 10.27 GMT

Jump to comments (7)

Airport security personnel take a body temperature reading of a boy as he arrives at Hong Kong International Airport April 9,2013, following concerns over a deadly strain of bird flu. Photograph: Tyrone Siu/Reuters

Science researchers have discovered a problem with Google's Flu Trends system: it's no longer any good at predicting trends in flu cases.

According to research carried out by a team at Northeastern University and Harvard University, Google's Flu Trends (GFT) prediction system has overestimated the number of influenza cases in the US for 100 of the past 108 weeks - and in February 2013 forecast twice as many cases as actually occurred.

- More on Google flu
  - http://www.bbc.co.uk/news/business-27683581
- "Why big data is in trouble: they forgot about applied statistics"
  - http://simplystatistics.org/2014/05/07/why-big-data-is-in-trouble-they-forgot-about-applied-statistics/
- Care.data debacle:
  - "NHS Care.data information scheme 'mishandled'"
  - http://www.bbc.co.uk/news/health-27069553

http://www.theguardian.com/technology/2014/mar/27/google-flu-trends-predicting-flu

- Big Data requires thinking differently about how you work with data

but…

- it can provide new insight and offer more value

WHY, WHERE & HOW DOES HPC FIT?

Why HPC?

- Volume
- Velocity
- Variety
- Performance

Where HPC?

Data

- Collection
- Cleaning
- Analysis

Knowledge Exploitation

How HPC?

- Computational Architectures
  - Multicore, clusters, Cloud, supercomputing, …
- Storage architectures
  - Databases (Relational, NoSQL, Shared-nothing,…)
  - OLAP (on-line analytical processing)
  - Hadoop Distributed File System (HDFS)
  - …
- Parallel computing paradigms
  - Map-Reduce, MPI, …

WRAP-UP