COMP 499 Introduction to Data Analytics

Lecture 1 — Introduction

Greg Butler

Data Science Research Centre

and

Centre for Structural and Functional Genomics

and

Computer Science and Software Engineering
Concordia University, Montreal, Canada

gregb@cs.concordia.ca

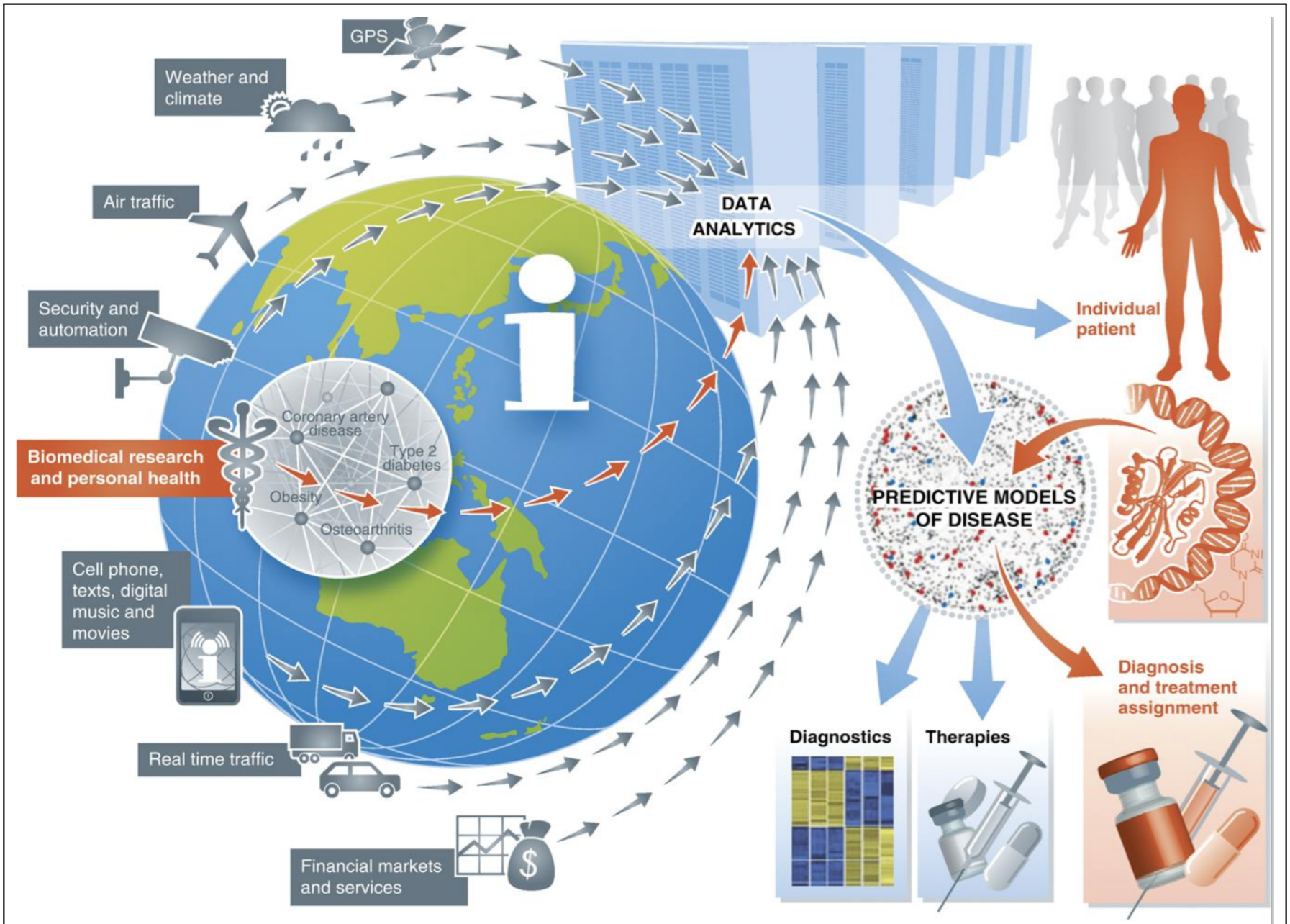
Overview of Lecture

1. Big Data
   - Actionable Data
   - History
   - Five V's
   - Types of Jobs
   - Privacy and Security
2. Data Analytics
   - Data Wrangling
   - Exploratory Data Analysis
   - Modeling
   - Story telling
3. Course Outline

Data & Feedback in Health ... Politics ... and Everything

Montreal Gazette 30 September 2016

Big Data



Eric E. Schadt, The Changing Privacy Landscape in the Era of Big Data, *Molecular Systems Biology* 8, 612 (2012).

Actionable Data in Data-Driven (Clinical) Healthcare

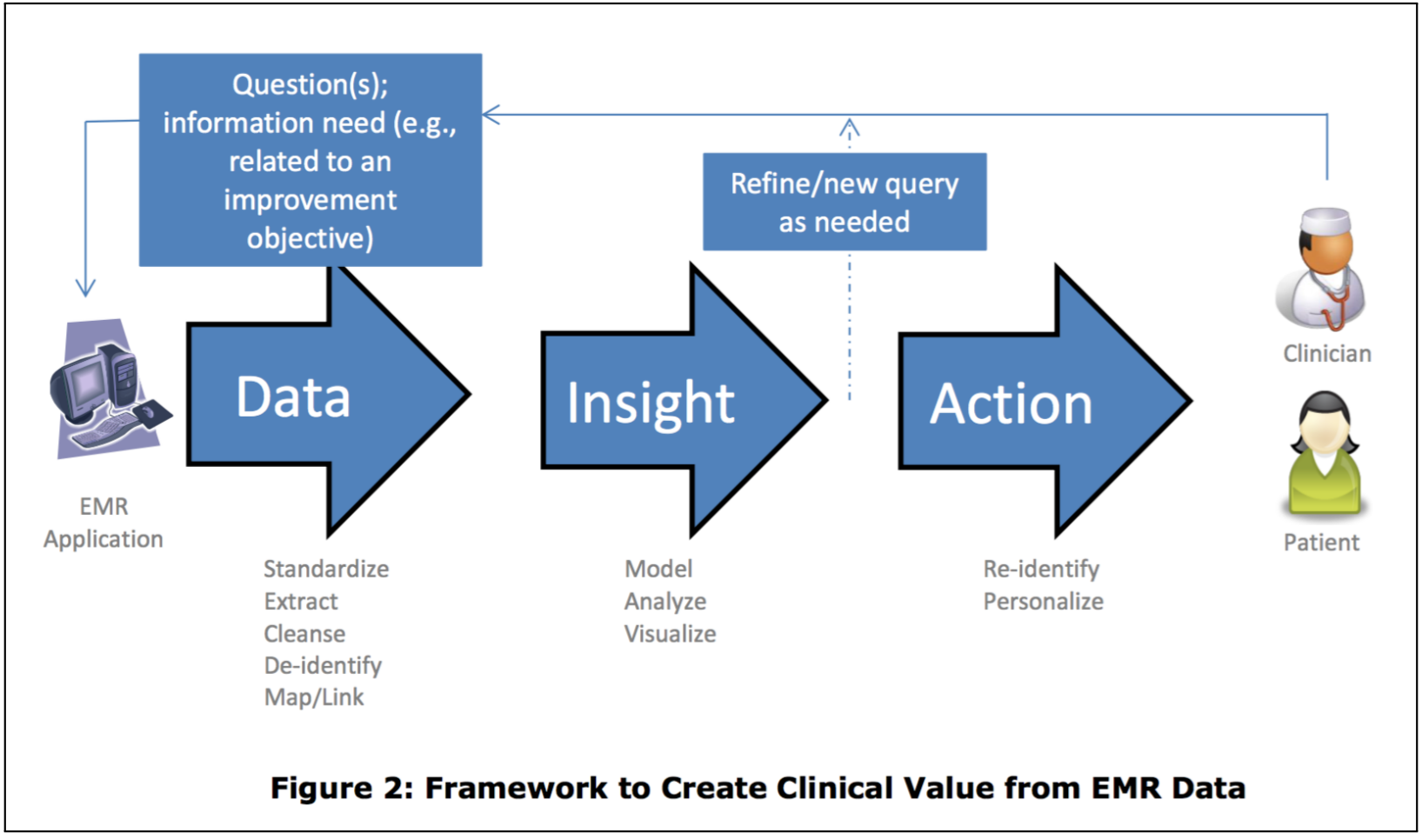


Figure 2: Framework to Create Clinical Value from EMR Data

(Infoway Health Canada 2016)

Big Data (http://dsrc.encs.concordia.ca/what-is-bigdata.html)

Big Data

Definition of *"Big"* has changed as we have become more advanced

History

Hollerith Cards 1890 (US population census)

Economic Data 1952 (GDP etc)

Computers 1959 — The First Digital Data Tsunami

World Wide Web 1990's — The Second Digital Data Tsunami

Social Media 1985 — The Third Digital Data Tsunami

Internet of Things 2000 — The Fourth Digital Data Tsunami

Big Science — 1960's onwards

Deep Knowledge — 2011 onwards

A key notion is **actionable data** that is useful in supporting decisions, determining actions, and adding value to an endeavour.

Big Data

The 5 V's

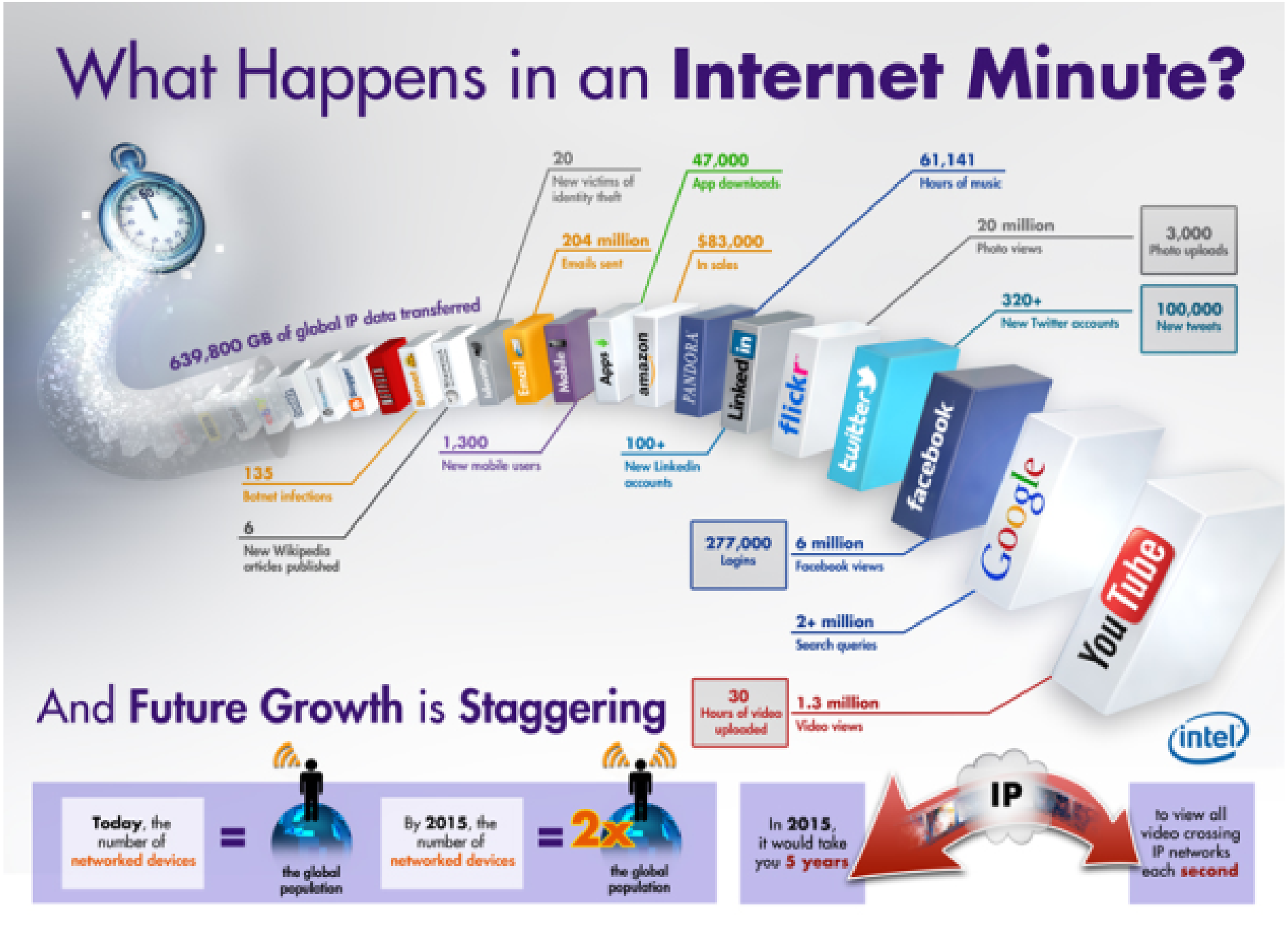
Volume: amount of data
Variety: different types of data
Velocity: rate at which data is generated
Veracity: trustworthiness, level of noise
Value: usefulness of data to a business

Drivers

Transactions
Mobile
Social Media
Internet of Things

MGI Report

McKinsey Global Institute, *Big data: The next frontier for innovation, competition, and productivity*, May 2011.

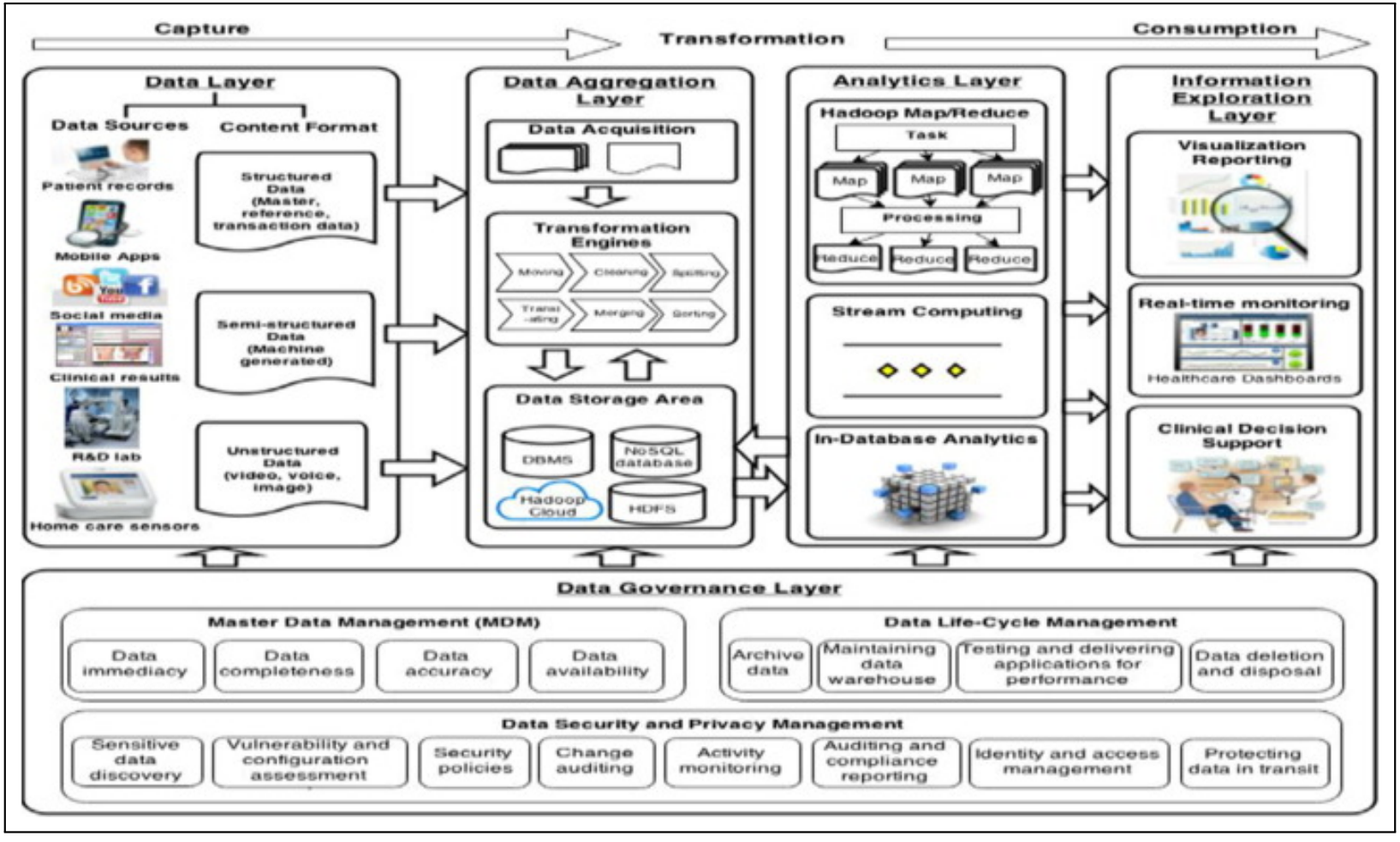


(Intel 2012, http://scoop.intel.com/what-happens-in-an-internet-minute/)

Types of Jobs in Big Data

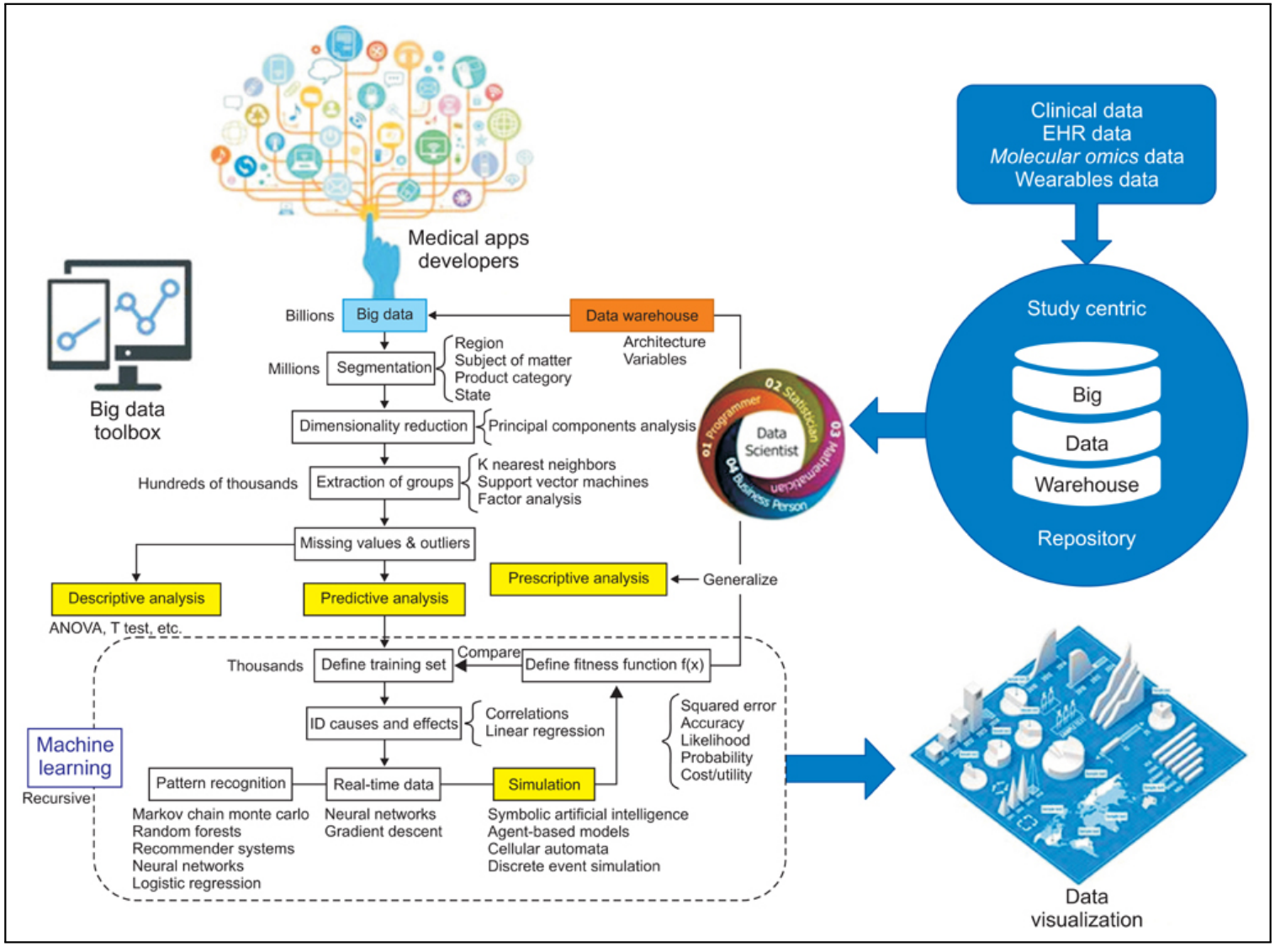
- Data Analyst
- Data Scientist
- Data Architect
- Chief Data Officer

The IT Perspective



Wang et al (Tech. Forecasting & Social Change, 2016)

The Big Data Analytics Perspective



Dimitrov (Health Informatics Research, 2016)

Privacy and Security

"**Privacy** *refers to an individuals right to control the collection, use, and disclosure of his/her personal health information (PHI) and/or personal information (PI) in a manner that allows health care providers to do their work.*

Security *is about ensuring the information gets to the right person in a secure manner."*

Ontario's Ehealth Blueprint `http://www.ehealthblueprint.com`

Privacy by Design 2009

Seven Foundational Principles

1) being proactive not reactive;
2) having privacy as the default setting;
3) having privacy embedded into design;
4) avoiding the pretence of false dichotomies,
   such as privacy vs. security;
5) providing full life-cycle management of data;
6) ensuring visibility and transparency of data; and
7) being user-centric

Prof. Ann Cavoukian, formerly Information and Privacy Commissioner of Ontario; now Ryerson University. `http://www.privacybydesign.ca`

Canada's Personal Information Protection and Electronic Documents Act (PIPEDA) 2000

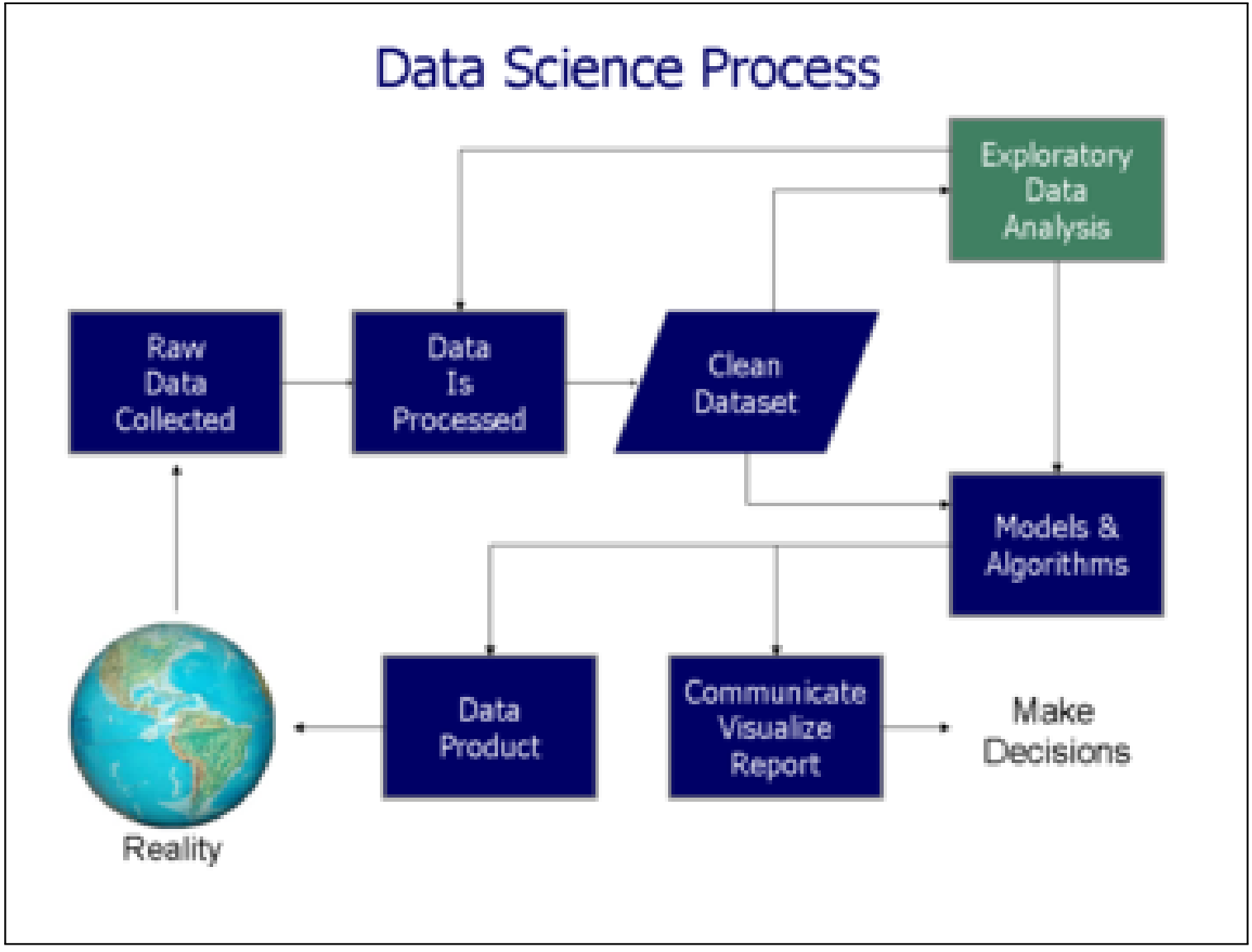
Ten Privacy Principles

Accountability: An organization is responsible for personal information under its control and shall designate an individual or individuals who are accountable for the organization's compliance with the following principles.
Identifying Purposes: The purposes for which personal information is collected shall be identified by the organization at or before the time the information is collected.
Consent: The knowledge and consent of the individual are required for the collection, use or disclosure of personal information, except when inappropriate.
Limiting Collection: The collection of personal information shall be limited to that which is necessary for the purposes identified by the organization. Information shall be collected by fair and lawful means.
Limiting Use, Disclosure, and Retention: Personal information shall not be used or disclosed for purposes other than those for which it was collected, except with the consent of the individual or as required by the law. Personal information shall be retained only as long as necessary for fulfilment of those purposes.
Accuracy: Personal information shall be as accurate, complete, and up-to-date as is necessary for the purposes for which it is to be used.
Safeguards: Personal information shall be protected by security safeguards appropriate to the sensitivity of the information.
Openness: An organization shall make readily available to individuals specific information about its policies and practices relating to the management of personal information.
Individual Access: Upon request, an individual shall be informed of the existence, use and disclosure of his or her personal information and shall be given access to that information. An individual shall be able to challenge the accuracy and completeness of the information and have it amended as appropriate.
Challenging Compliance: An individual shall be able to address a challenge concerning compliance with the above principles to the designated individual or individuals for the organization's compliance.

`https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/`

`the-personal-information-protection-and-electronic-documents-act-pipeda/p_principle/`

Data Analytics



wikipedia

Data Analytics — Example from wikipedia

Find the variables which best predict the tip given to the waiter.

The variables available in the data collected:

- the tip amount,
- total bill,
- payer gender,
- smoking/non-smoking section,
- time of day,
- day of the week, and
- size of the party

The approach is to fit a regression model to predict the tip rate. The fitted model is

- $tip\_rate = 0.18 - 0.01 \times party\_size$

if size of the dining party increases by one (leading to a higher bill), the tip rate will decrease by 1%.

Data Analytics

Steps in Data Analytics

- Setting Questions
- Data Wrangling
- Exploratory Data Analysis
- Modeling
- Story telling

Iterative process

The process is rarely linear.
Each step can push data scientist to revisit methods, techniques ... or reconsider whether the original question was the right one?
And the final answer simply sparks more questions!

Data Analytics

Hypothesis-driven Experimental Design and Analysis

Not exploratory.

You have a single, specific hypothesis to accept or reject.

Steps

- Set null hypothesis H_0 and alternative hypothesis H_1
- Design experiment to collect data, and
- Design analysis of experimental data to accept/reject hypothesis
- Determine *statistical power* of experiment
  Do you have enough data points?
- Do experiment, do analysis, accept/reject hypothesis

Data Analytics: Setting Questions

Ask an Interesting Question

Steps

- Is there a business goal to achieve?
- Some object of scientific interest that would be helpful to discover?
- What parameters would the ideal answer fulfill?

Data Analytics: Data Wrangling

Design a Data Collection Program

- Establish whether or not the data exists in the real world and is relevant to the question
- Devise a collection scheme to acquire it
  Logistical considerations? Cost? Privacy issues?
- Coordinate with departments or agencies needed for collection program liaison

Collect and Review the Data

- Store the incoming data to allow modeling and reporting
- Join data from multiple sources in relevant & logical manner
- Check for anomalies or unusual patterns
  - Caused by the collection process?
  - Inherent to topic of investigation?
  - Correct them, or develop new collection scheme?

Data Analytics: Data Wrangling

Data Wrangling or Data Munging

Bring skills and intuition to bear ...
to take messy, incoherent information ...
and shuffle it into clean, accessible sets

"Munging" the Data

- Select your tools to comb through raw
- Store the munged data as a fresh data set, **or**
- use programmatic pre-processing for each subsequent query

Data Analytics: Exploratory Data Analysis

Exploratory Data Analysis

Learn about the properties of the data

Steps

- Descriptive statistics: mean/median and variance, quantiles, outliers
- Correlation
- Fitting curves and distributions
- Dimension reduction
- Clustering

Data Analytics: Modeling

Modeling

the fun stuff of getting *"meaning"* from a clean data set

Steps

- Build a data model to fit the question
- Validate the model against the actual collected data
- Perform the necessary statistical analyses
- Machine-learning or recursive analysis
- Regression testing and other classical statistical analysis techniques
- Compare results against other techniques or sources

Data Analytics: Story telling

Visualize and Communicate the Results

The most challenging part of the data scientist's job is taking the results of the investigation and presenting them to the public or internal consumers of information in a way that makes sense and can be easily communicated.

Steps

- Graph or chart the information
- Tell a story to fit the results: Interpret the data to describe the real-world sources in a plausible manner
- Assist decision-makers in using the results to drive their decisions

COMP 499 — Course Summary

Course Web Site:
http://users.encs.concordia.ca/~gregb/home/comp499-s2018.html

Instructor: Greg Butler, EV-3.219, gregb@encs
`http://users.encs.concordia.ca/~gregb`

Lectures: Tuesdays & Thursdays 13:30 – 16:00 H-403

Labs: Tuesdays & Thursdays 11:20 – 13:20 H-907
Labs are Mandatory.

Office Hours: By appointment in EV 3.219

- Ask questions at lectures!

Recommended Books

- *Data Science from Scratch: First Principles with Python*, by Joel Grus, O'Reilly, 2015.
- *Data Crunching: Solve Everyday Problems using Java, Python and More*, by Greg Wilson, The Pragmatic Bookshelf, 2005. This book is out of print, but can be found in the Library.
- *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, by Hadley Wickham, Garrett Grolemund, O'Reilly Media, 2016.
- *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, by William McKinney, O'Reilly Media, 2012.
- Material on available data resources and data analysis problems from a range of disciplines will be provided.

Evaluation

Assignments $\times$ 2	20%
Project — Individual	30%
Midterm Exam	15%
Final Exam	35%
Total	100%

You must pass both project and exam components of the course.

Assignments

Submit to eas electronic submission system as Jupyter notebooks.

Marked in Labs

Assignment 1 — Data Wrangling

Perform Lab 2 data wrangling task in Python for Movie dataset as Jupyter notebook

Assignment 2 — Exploratory Data Analysis

Part A — Lab 4 using Python to explore a dataset
Part B — Lab 5 using R to explore a dataset

Project

Select a dataset — analyse it

What are the questions?
Data Wrangling
Exploratory Data Analysis
Model Construction
Story Telling

Project Report

Do project as a Jupyter notebook.
Prepare presentation for class to tell your story.
Submit pdf version of report to eas electronic submission system as Jupyter notebooks.

Examinations

Midterm

60 minute examination
true-false, multiple choice, short answer
focus on terminology and basic technical matters

Final

three-hour examination
true-false, multiple choice, short answer
focus on process of data annlytics: data wrangling, exploratory data analysis, modeling, story telling, validation (maybe)

COMP 499 is a slot course!

First time COMP 499 is being taught

Be prepared to be flexible.
Course schedule and content is subject to change.
My organization may not be the best.

Accelerated summer schedule!

COMP 499 moves at twice the normal pace.
Expect to spend 20 hours/week on COMP 499.
Be proactive in watching videos, reading, and doing labs.
Labs are essential — this is a DOING subject!
My organization may not be the best.

Context — Data to Knowledge

Data

raw, calibrated, normalized, validated

derived, aggregated, interpreted

Metadata describes source and properties of the data

Information

newsworthy

actionable

Claude Shannon's information theory

Knowledge

applicable wisdom, organized information

concepts, relations, constraints, taxonomy/ontology

axioms, rules, plans

Application — aka Knowledge Translation

Context — Data Level of Interpretation

Raw Data

raw values obtained directly from the measurement device

Calibrated Data

raw physical values, corrected with calibration operators

Validated Data

calibrated data that has been filtered through quality assurance procedures
(most commonly used data for scientific purposes)

Derived Data

frequently aggregated data, such as gridded or averaged data

Interpreted Data

derived data that is related to other data sets, or to the literature of the field

Context — Syntax to Pragmatics

Lexical = atomic units

Defined by regular expressions

Represented as enum's

Syntax = structure

Defined by grammars

Represented as Abstract Syntax Trees (AST)

Semantics = meaning

Defined by interpretation mappings

Represented as actions (procedural) in compiling

Pragmatics = goals

Context — Approaches to Data Analysis

Scripting

Unix tools, eg
text files, csv files for inputs, outputs, intermediate steps
stepwise development of analysis
script captures steps, parameters
easy to replay

Notebooks

Jupyter, eg
interactive scripting with "literate programming"
keep track of thought processes during analysis
work with files to replay analysis

"Spreadsheet" Environments

OpenRefine, eg
lots of tools, little guidance
need macros, histories, to capture/replay work
often proprietary

References

Infoway Health Canada: Big Data Analytics in Health White Paper, May 2013.
Infoway Health Canada: Clinical Analytics in Primary Care White Paper, February 2016.

McKinsey Global Institute: Big data: The next frontier for innovation, competition, and productivity, June 2011.
McKinsey: The big-data revolution in US health care: Accelerating value and innovation January 2013.

The Catalyst towards an Ontario Health Innovation Strategy, Ontario Health Innovation Council (ohic.ca), December 2014.

R. Fang et al, Computational Health Informatics in the Big Data Age: A Survey, ACM Computing Survey July 2016 http://dl.acm.org/citation.cfm?id=2932707

Y. Wang et al, Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations, Technological Forecasting and Social Change Available online 26 February 2016 http://www.sciencedirect.com/science/article/pii/S0040162516000500

D.V. Dimitrov, Medical Internet of Things and Big Data in Healthcare, Healthc Inform Res. 2016 Jul;22(3):156-163. English. http://synapse.koreamed.org/search.php?where=aview&id=10.4258/hir.2016.22.3.156&code=1088HIR&vmode=FULL

C. Auffray et al, Making sense of big data in health research: Towards an EU action plan, Genome Medicine June 2016 http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-016-0323-y

Guillaume Taglang, David B. Jackson, Use of "big data" in drug discovery and clinical trials, Gynecologic Oncology 141 (2016) 17–23. http://www.sciencedirect.com/science/article/pii/S0090825816300464

Michael KK Leung, Andrew Delong, Babak Alipanahi, Brendan J Frey, Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets, Proceedings of the IEEE 104:1 (2016) 176–197. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7347331