Big Data:
How Will It Change Your Life?

Gregory Butler

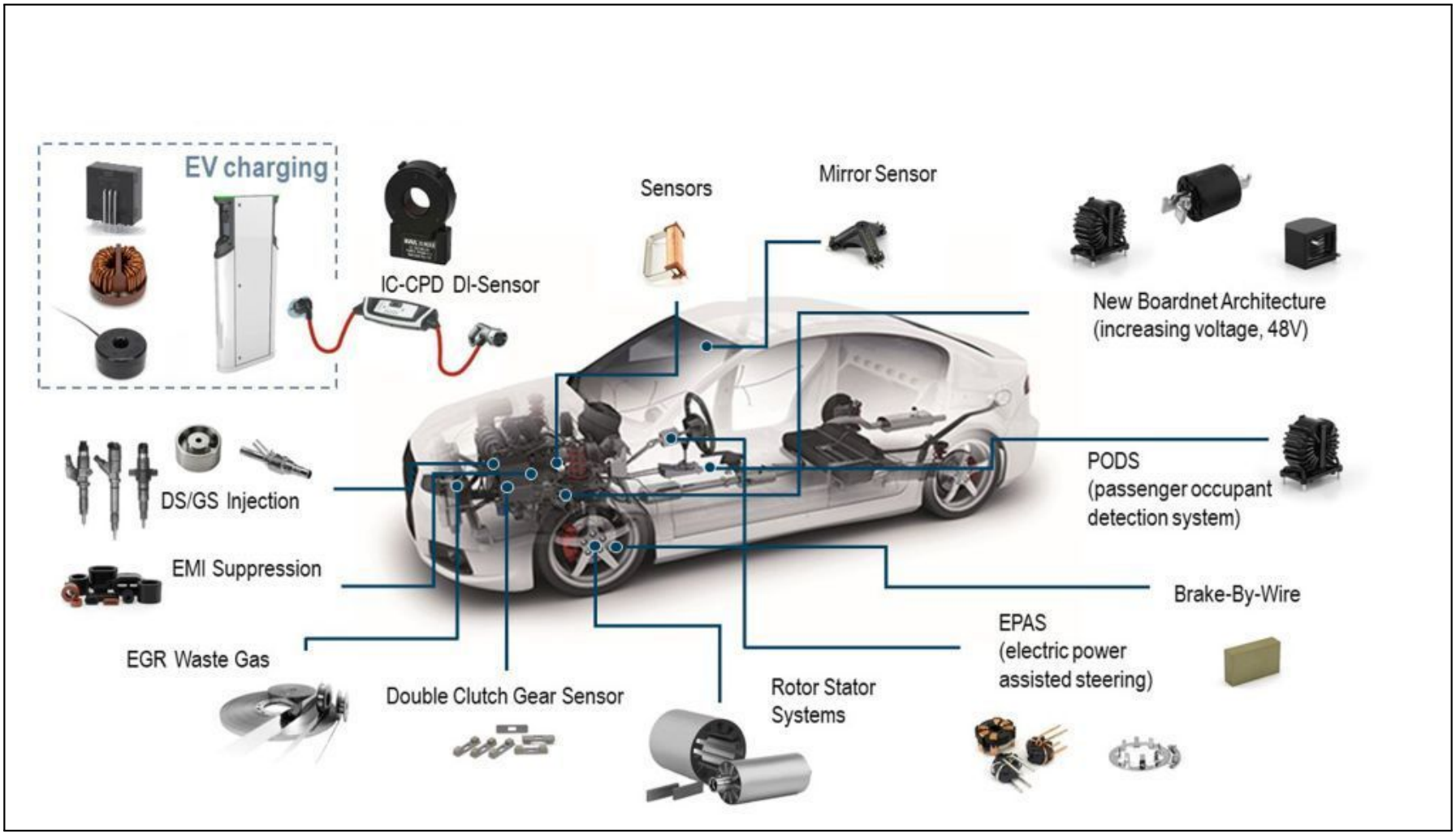
Department of Computer Science & Software Engineering
Data Science Research Centre
Centre for Structural and Functional Genomics
Concordia University, Montréal, Canada

October 2019 — Malaysia

Outline

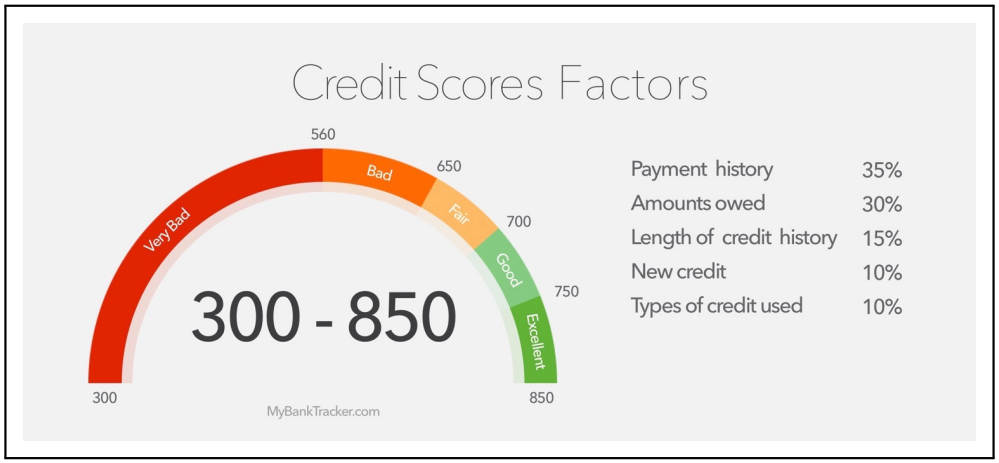
Outline

Transport and Mobility



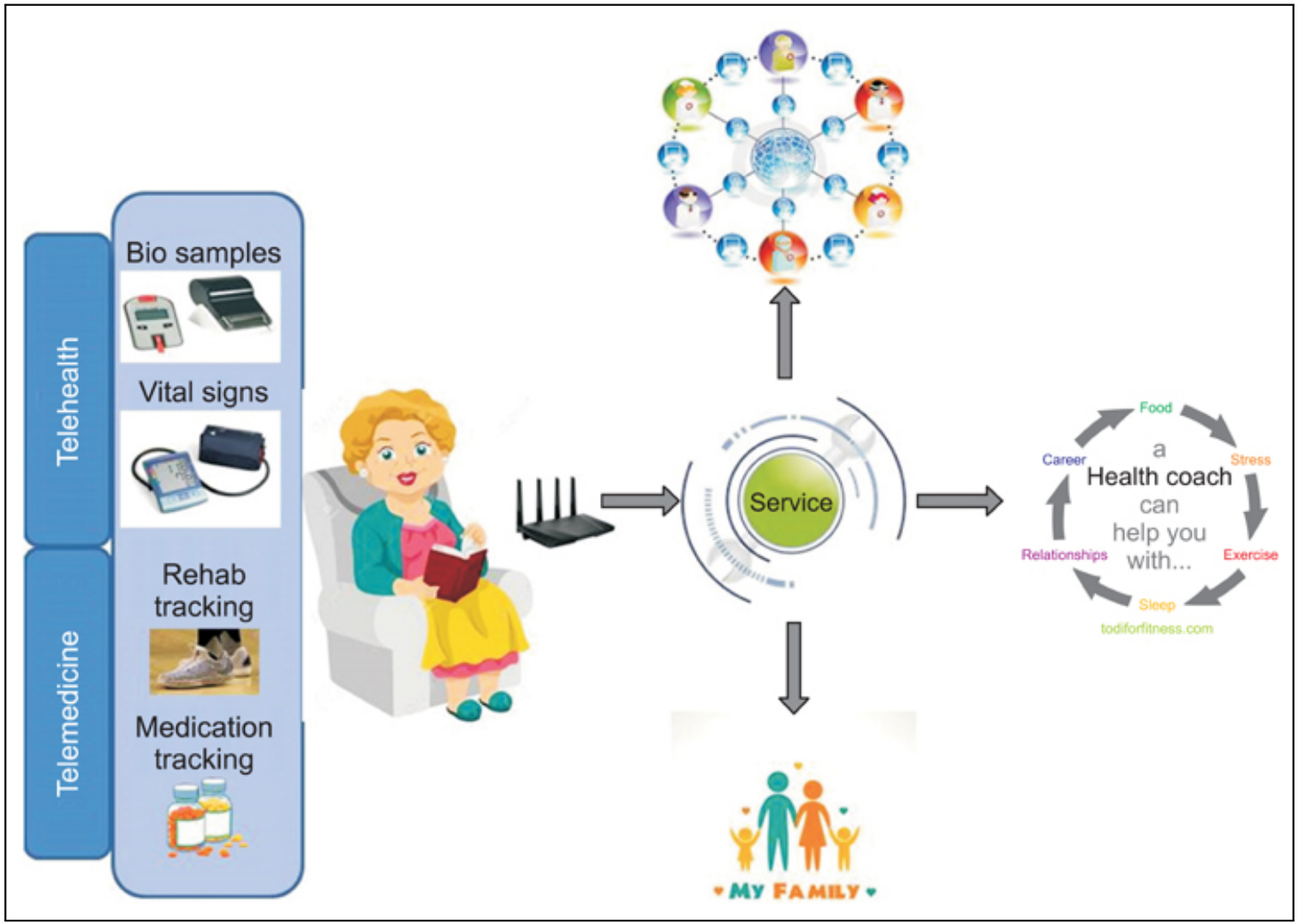
Money, Finance and Credit

Traditional Credit Score



Future Psychometric Credit Score

The Elderly or Remote Patient Perspective

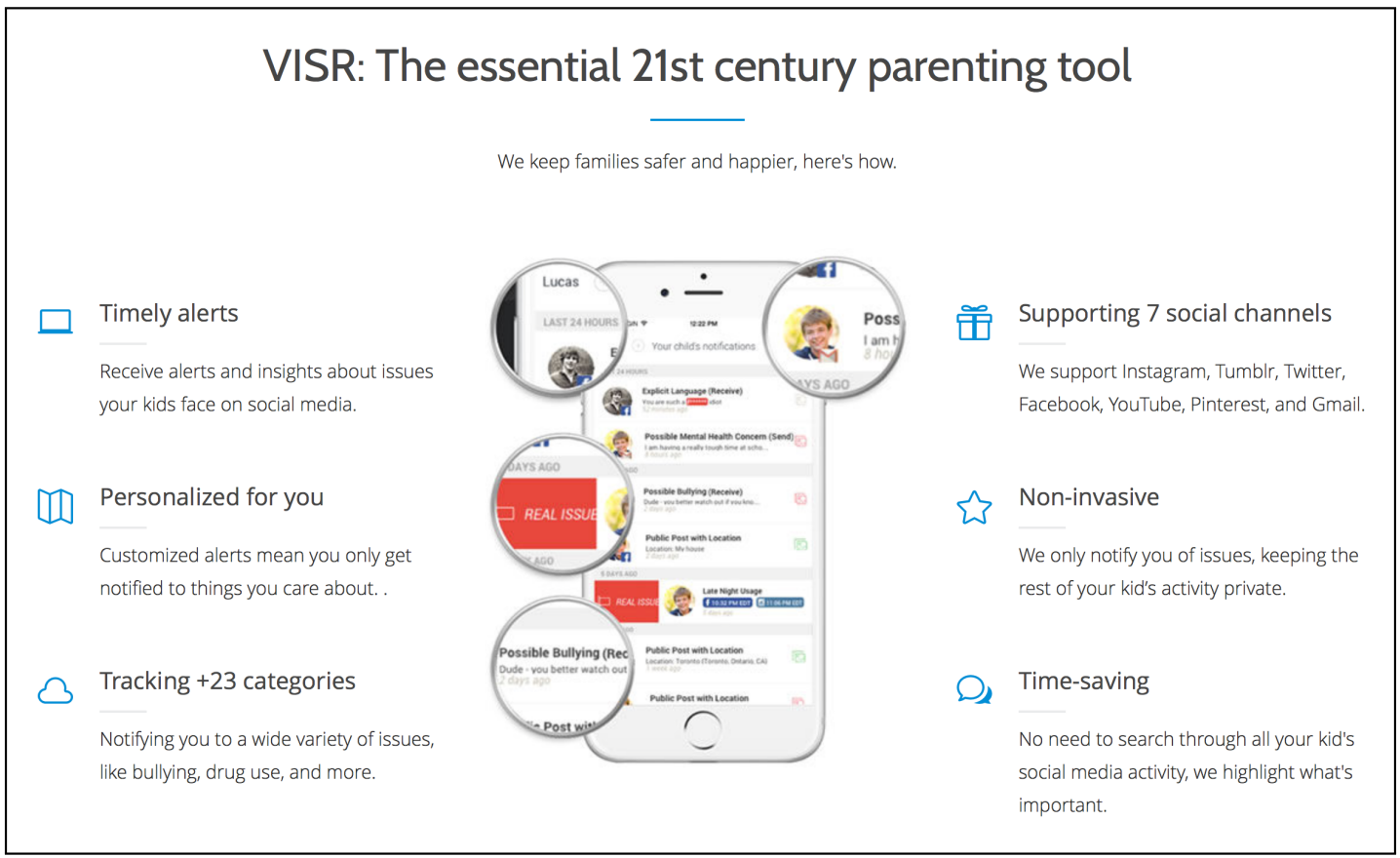


Dimitrov (Health Informatics Research, 2016)

VISR — A Canadian Company

Better mental and emotional health via social media data mining

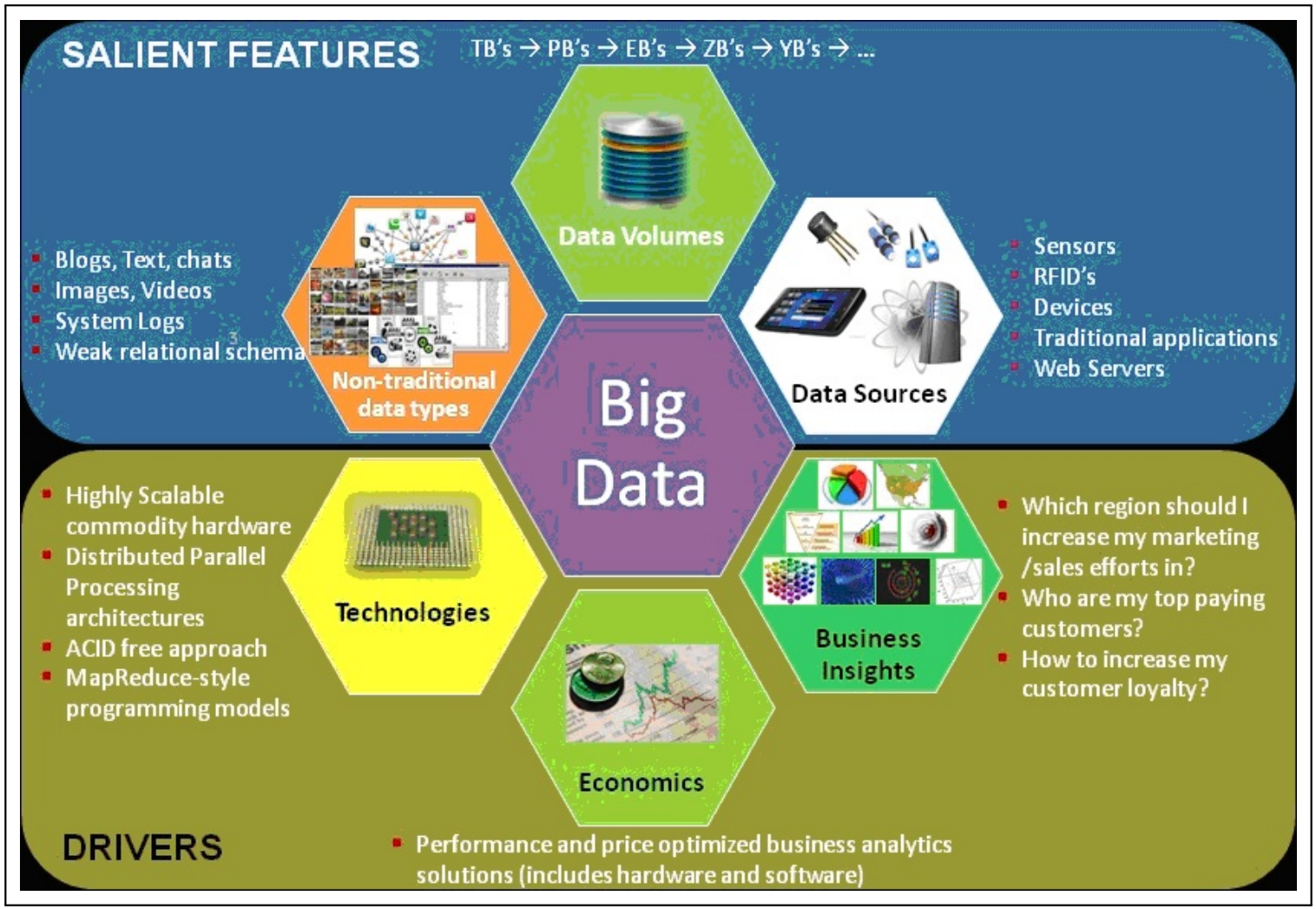
"On a mission to help families better navigate technology, by notifying parents about safety and wellness issues their kids face on social media"



http://www.visr.co

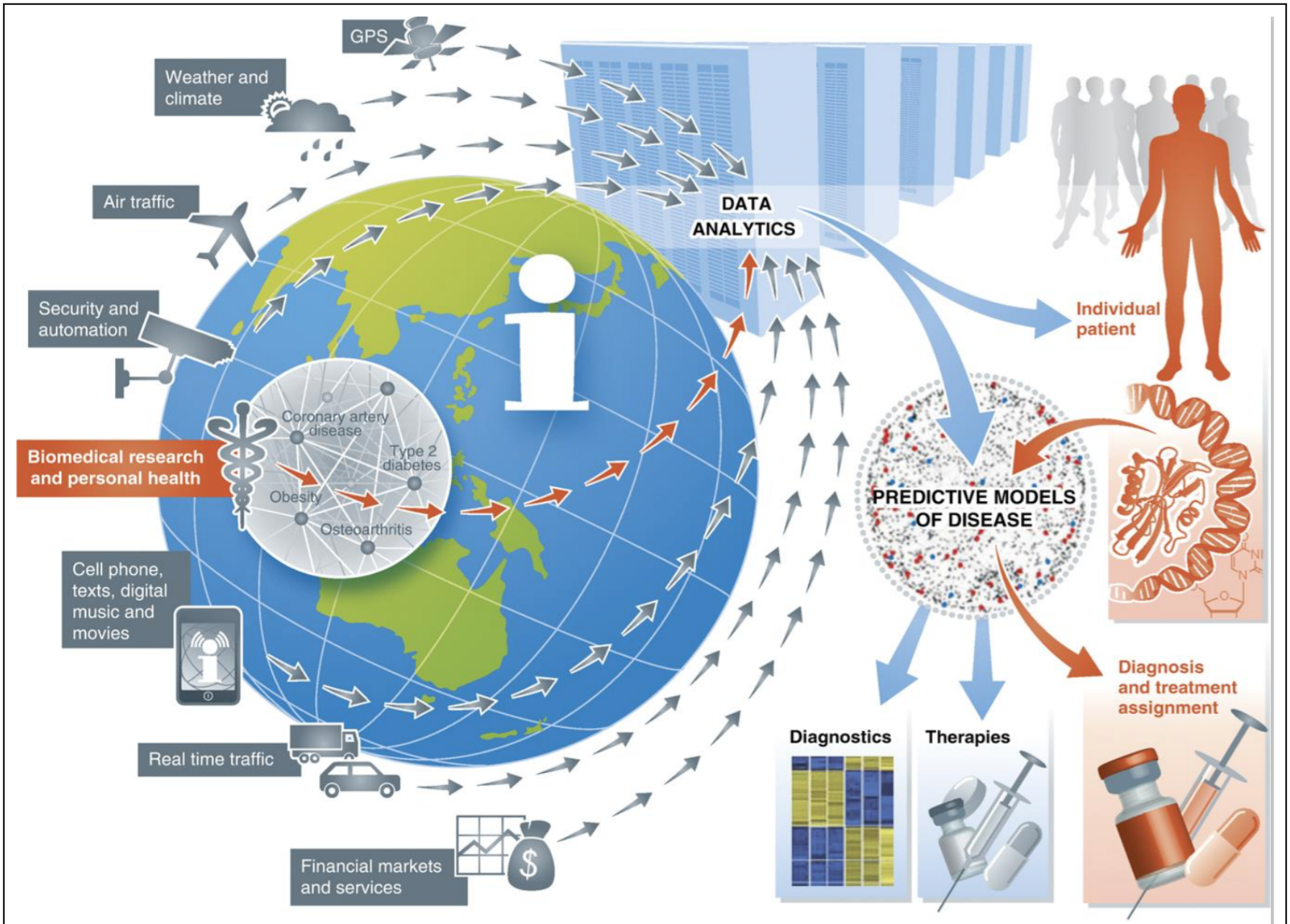
Outline

Big Data: 2,000,000,000,000,000,000 bytes (2EB)per day!



https://shhrota.com/2012/01/02/the-big-in-big-data/

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

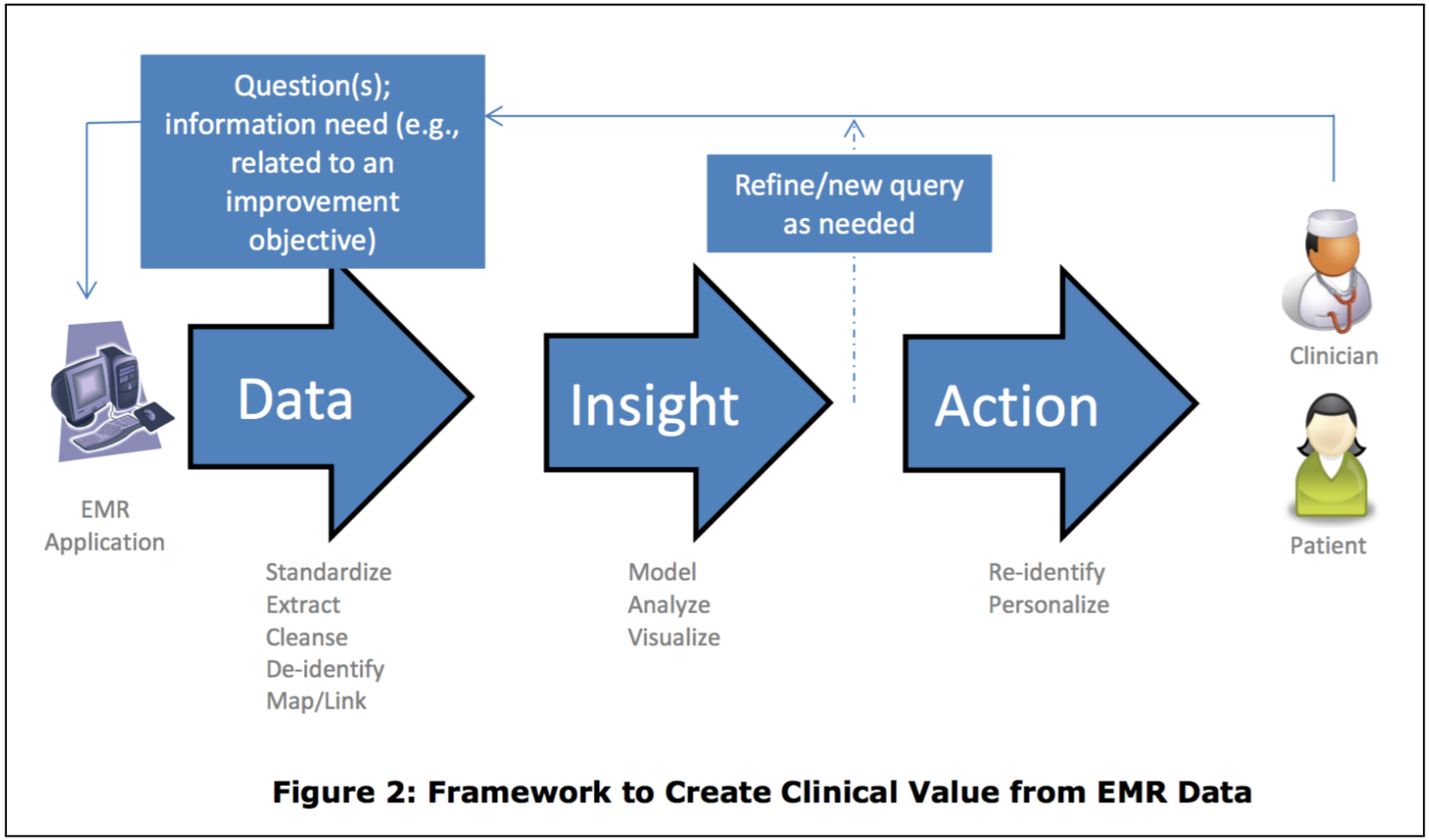


Figure 2: Framework to Create Clinical Value from EMR Data

(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
plus Visualization, Viscosity (sticky), Virality (convey a message)

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.

Types of Jobs in Big Data

Data Analyst

analyses data from a business perspective to put data to use

Data Scientist

expert at data analysis, mathematical modeling, machine learning and coding

Data Architect

design, create, deploy, manage an organization's

data architecture

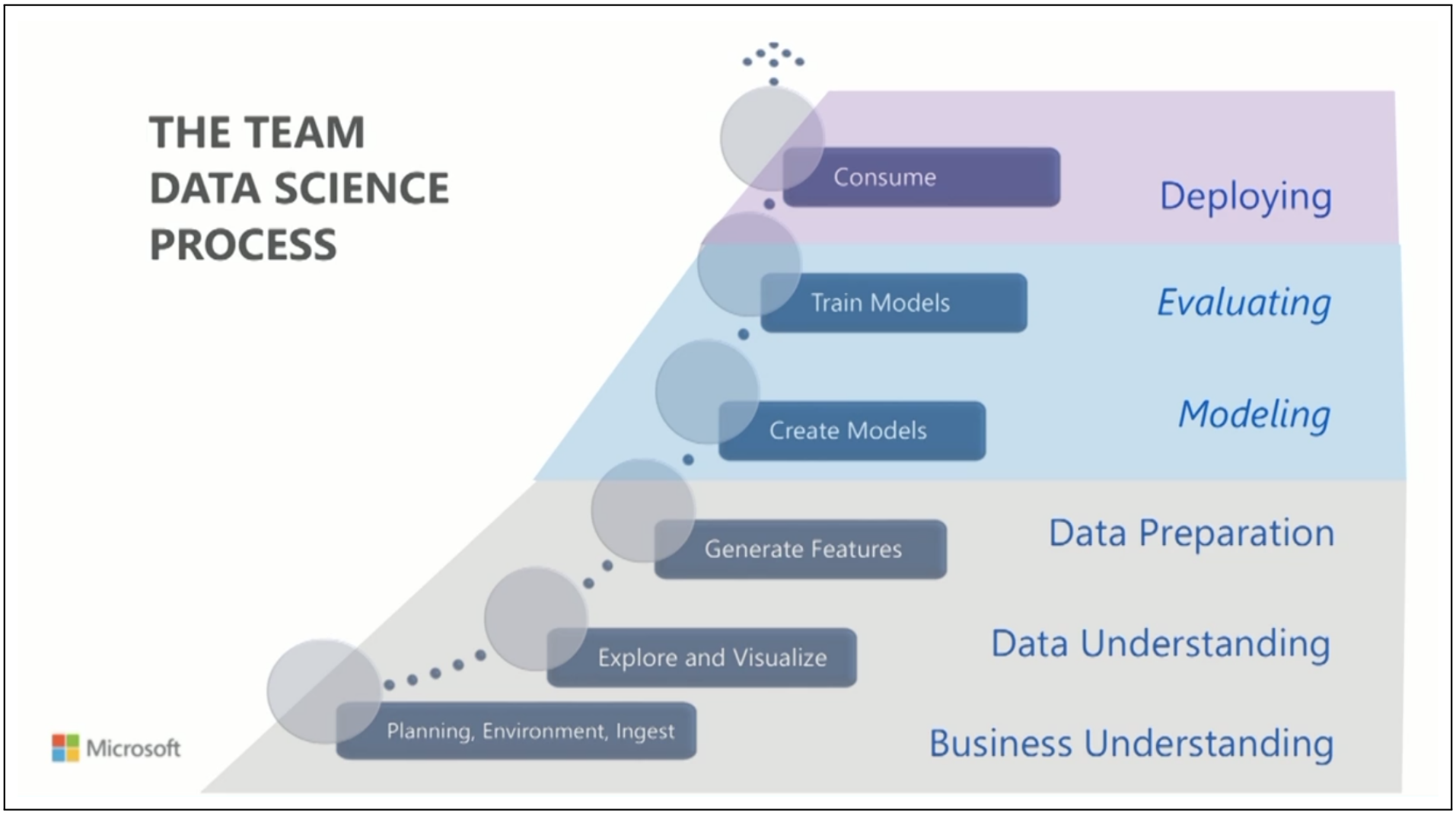
composed of models, policies, standards for data collection, storage, structure, integration, and use

Chief Data Officer

responsible for enterprise-wide governance and utilization of data as an asset; integrate data-driven business process

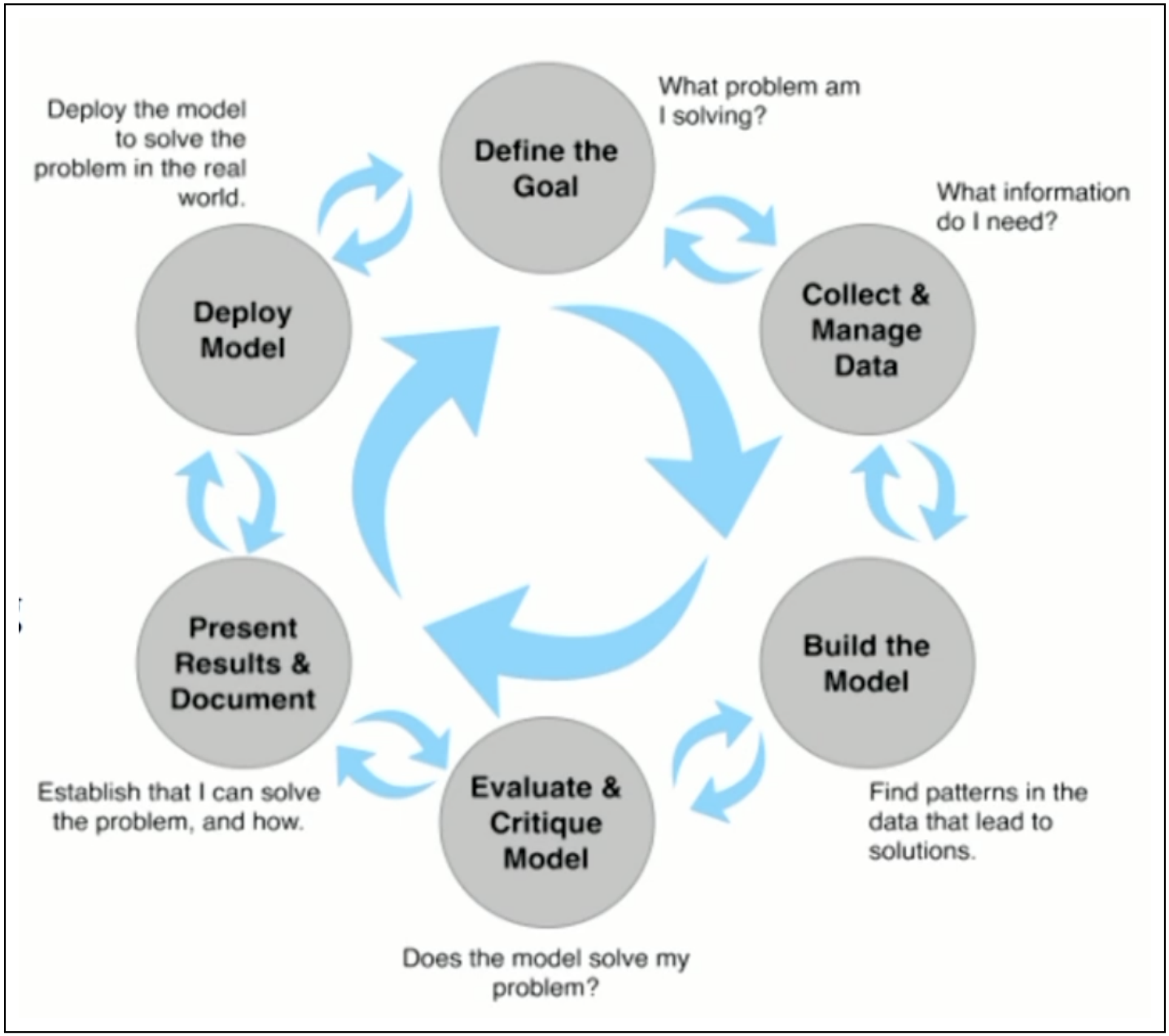
Outline

Data Analytics



(Microsoft)

Data Analytics



(Microsoft)

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

Data Analytics: Data Wrangling

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: Exploratory Data Analysis

Exploratory Data Analysis

Learn about the properties of the data

Steps

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

Data Analytics: Modeling

Modeling

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
- Compare results against other techniques or sources

Data Analytics: Modeling

The choice of a model affects (and is affected by)

- Whether the model meets the business goal
- How much pre-processing the model needs
- How accurate the model is
- How explainable the model is
- How fast the model is (in making predictions)
- How scalable the model is (building and predicting)

(Microsoft)

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 results to make decisions

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

Outline

Big Data Analytics — Compute Clusters & the Cloud

Map Reduce Approach

Hadoop, Spark
Distributed database support (HBase)

Knowledge Graphs

Linked data, ontologies & semantic web

Cloud

Flexible, distributed computing, as needed

noSQL Databases

Modern technology for varieties of data

Canada's Big Data Consortium

Who

Established by Ryerson University in mid 2014
Bring together Govt, Industry, and Universities
Four academic founding partners: Ryerson, SFU, Dal, Concordia

Lessons Learned

Govt keen to exploit Big Data job growth & economic growth
and improve efficiency of govt operations

Some schools recognise importance of numeracy & analytics
Many universities introducing courses & programmes

Enormous talent gap between demand and supply

4 Job Types Identified:
Chief Data Officer, Data Scientist, Data Analyst, Data Architect

Montreal International Panel of Experts

Who — Dec 2014 to Dec 2015

MI profile of Big Data industry in Montreal and Quebec

All the "movers and shakers"

- quebec govt: industry, trade, research, labour
- open montreal director
- montreal smart city initiative
- industry, SME, university, CRIM

Lessons Learned

Montreal has what it takes!

Big Data is ideal space for innovative small start-ups!

Big Data essential for biotech, pharma industries and research
- and Quebec is lagging in such data-driven life sciences

Importance of open source and open data

Working Group on Predictive Health

Who — WG of Canada's Big Data Consortium

Academia, govt, consultants, industry, hospital, insurance, start-ups

Data Access is Key!

Open data by default will be the future.
Data secure in the cloud and on personal smartcards
Privacy and sharing of data is in control of patient!
Anonymised data using *"id servers"* for data integration
Future is linked open data using RDF, ontologies, semantic web

Incubators need Technology + Business + Healthcare

Physicians must buy-in ... or ... it's a no-go!

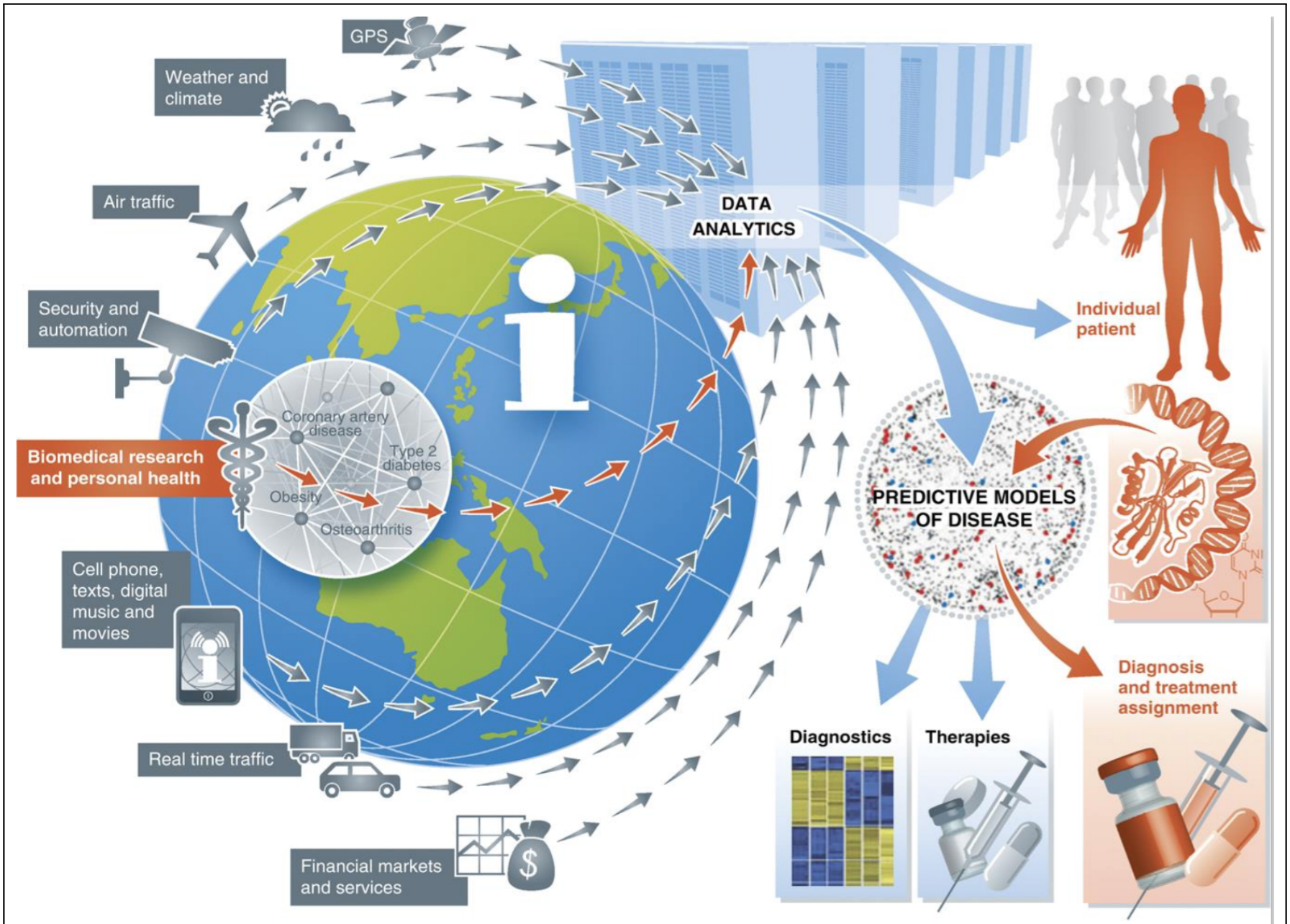
Scale Out beyond Incubators

Need community engagement ...

- for acceptance, use, and data sharing
- to establish agreed terminology for data sharing
- to reap benefits of innovation from Big Data

Outline

Big Data



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

What Data is Gathered?

GPS in your watch and car

Your location at all times

Your web browser

Which sites you visit
How long you stay
What search queries you ask
When, from where, how often

Amazon and Shops

What you look at
What you buy
Are you influenced by recommendations

Email

What you say
Who you know

Social Media

What you say
Who you know
What you like

Your Fitness Tracker

Your health
Your exercise regime

Credit Card and Bank

What you buy
When, from where, how often

Business Benefits of Big Data

Better decision-making

Increased productivity

Reduced costs

Understanding your customers

Fraud detection

Applications for Big Data in Healthcare

Diagnostics
Data mining and analysis to identify causes of illness

Preventative medicine
Predictive analytics and data analysis of genetic, lifestyle, and social circumstances to prevent disease

Precision medicine
Leveraging aggregate data to drive hyper-personalized care

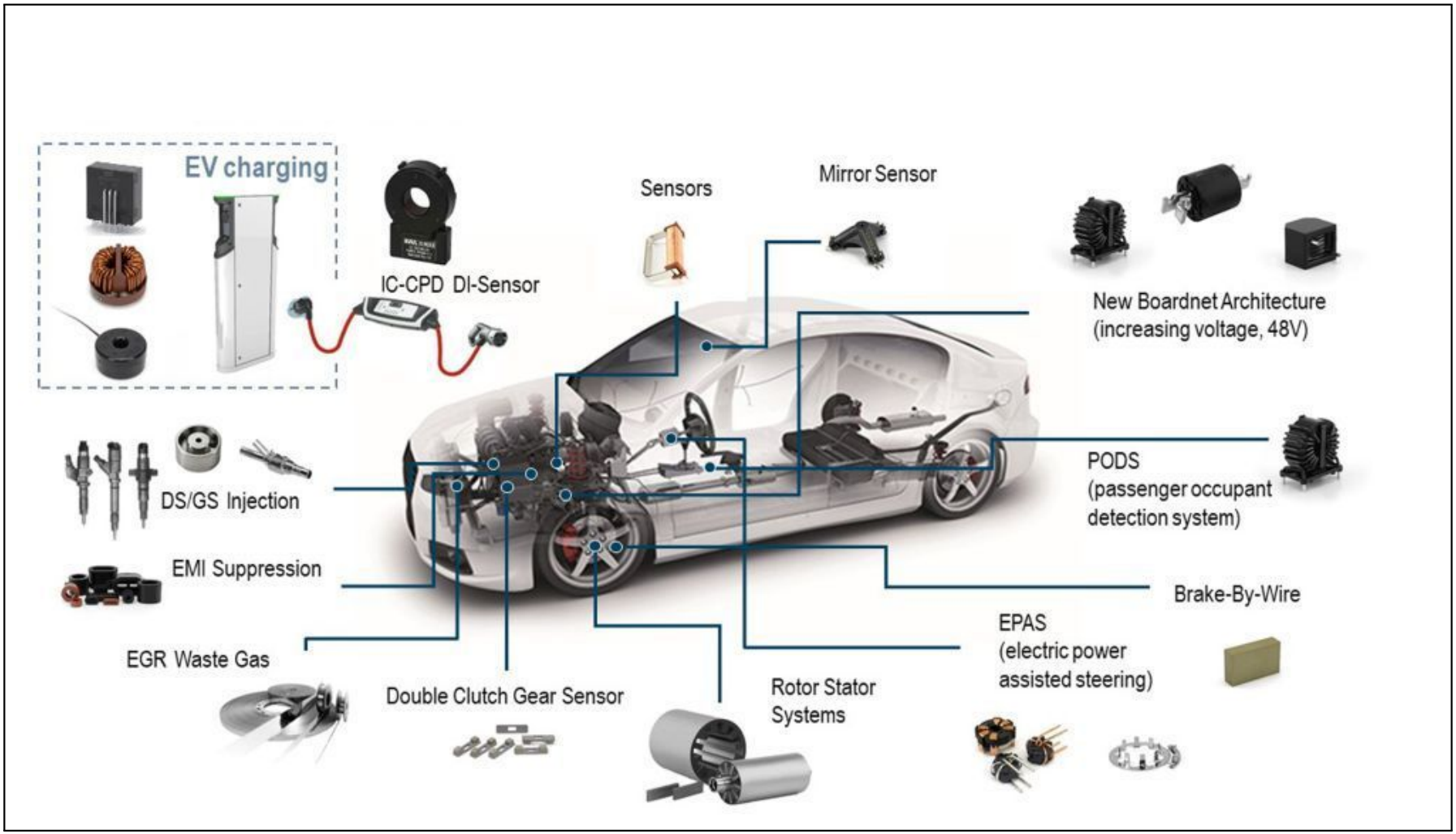
Medical research
Data-driven medical and pharmacological research to cure disease and discover new treatments and medicines

Reduction of adverse medication events
Harnessing of big data to spot medication errors and flag potential adverse reactions

Cost reduction
Identificaton of value that drives better patient outcomes for longterm savings

Population health
Monitor big data to identify disease trends and health strategies based on demographics, geography, and socio-economics

Transport and Mobility



Transport and Mobility

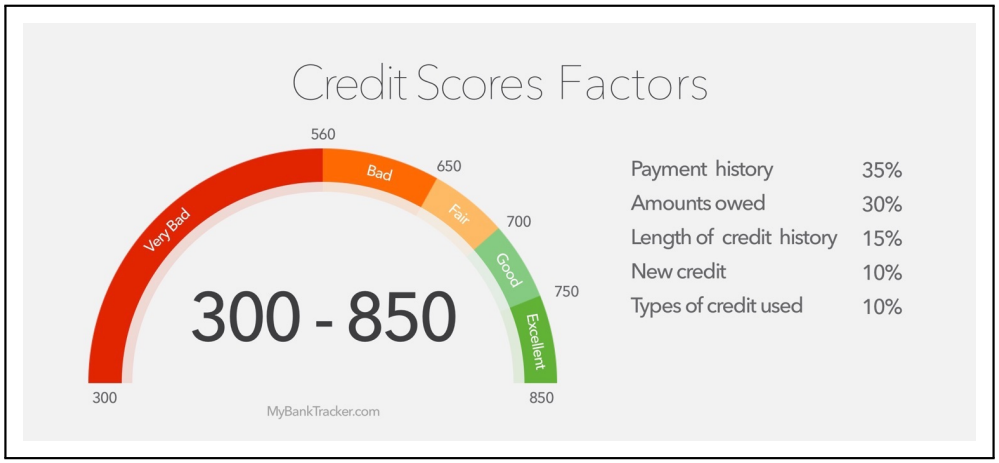
Walker for the Blind

Sport and Leisure

Badminton

Money, Finance and Credit

Traditional Credit Score



Future Psychometric Credit Score

Outline

Conclusion — Challenges

Data Ownership and Control

Will you control your data?
Control ... Who can access it? How they can use it?
Will you benefit ($$$) from others using your data?

Balance of Control between Computer & Human

smart homes, smart cars, smart buildings, smart cities

Safeguards for Disasters

for example, loss of electrical power
for example, loss of communication
for example, loss of satellites & GPS

when everything depends on computers

Thank You!

Questions, Please?

Privacy and Security

*"**Privacy** refers to an individual's 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`