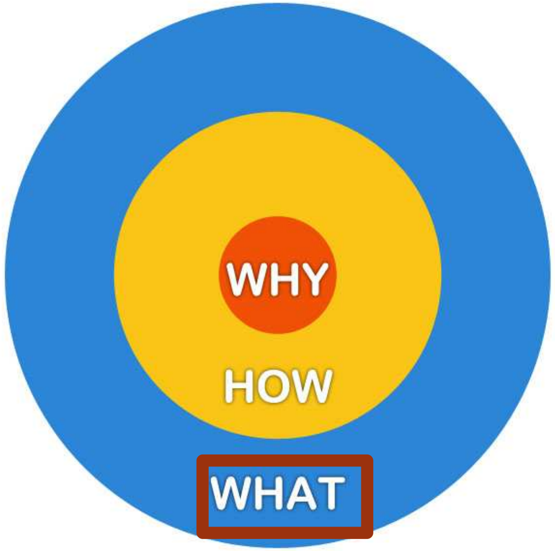
Introduction to Big Data

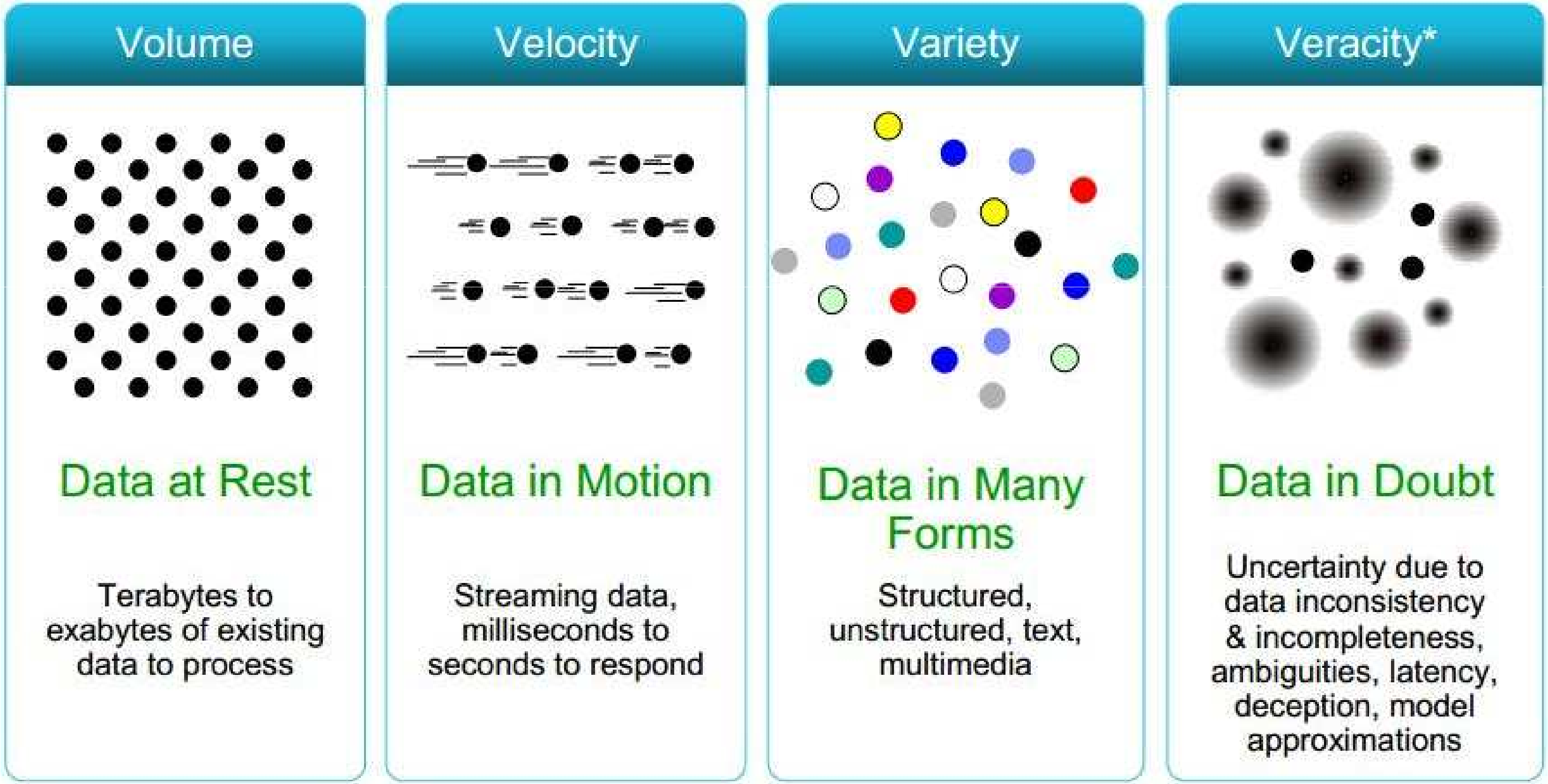
Rosaline Macharia

What is Big Data?



"**Big Data**" is data whose **scale**, **diversity**, and **complexity** require new architecture, techniques, algorithms, and analytics to manage it and extract **value** and **hidden knowledge** from it

Characteristics of Big Data



Volume

Data at Rest

Terabytes to exabytes of existing data to process

Velocity

Data in Motion

Streaming data, milliseconds to seconds to respond

Variety

Data in Many Forms

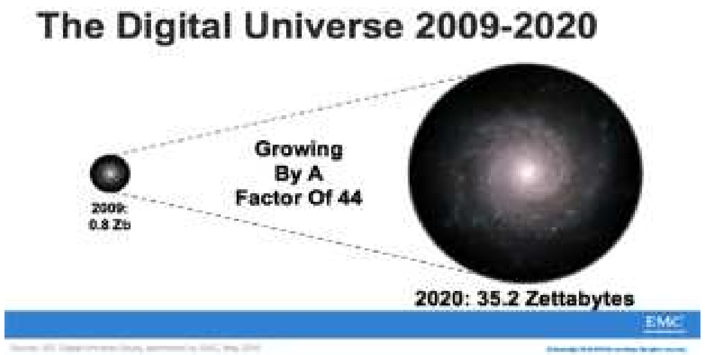
Structured, unstructured, text, multimedia

Veracity*

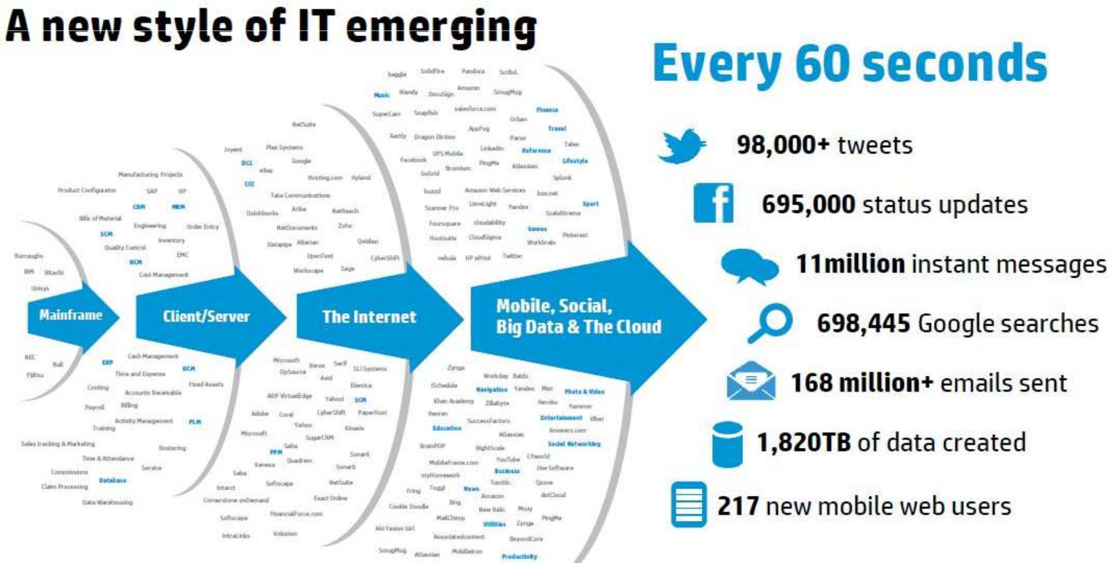
Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

1-Volume



44x increase from 2009 - 2020

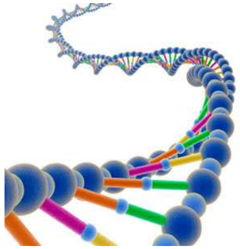


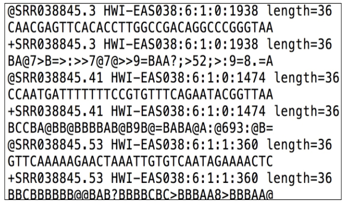
2- Variety

- Various formats, types, and structures

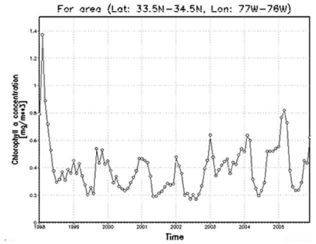
  Text, numerical, images, audio, video, **sequences,** time series, social media data, multi-dim arrays, etc…

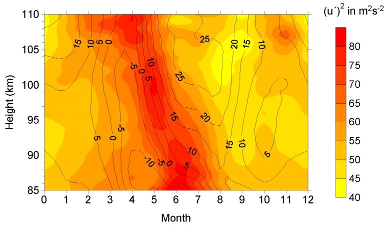
- Static data vs. streaming data



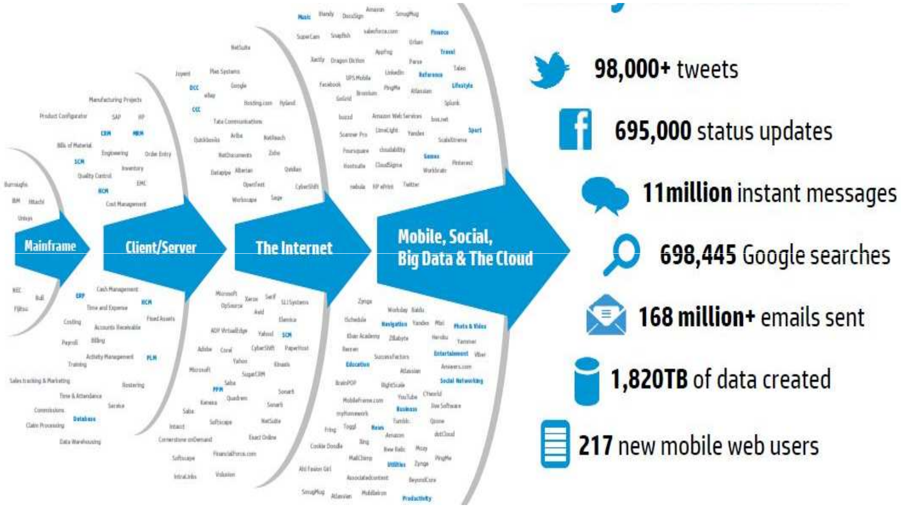


```
@SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
CAACGAGTTCACACCTTGGCCGACAGGCCCGGGTAA
+SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
BA@7>B=>:>>7@7@>>9=BAA?;>52;>:9=8.=A
@SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
CCAATGATTTTTTTCCGTGTTTCAGAATACGGTTAA
+SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
BCCBA@BB@BBBBAB@B9B@=BABA@A:@693:@B=
@SRR038845.53 HWI-EAS038:6:1:1:360 length=36
GTTCAAAAAGAACTAAATTGTGTCAATAGAAAACTC
+SRR038845.53 HWI-EAS038:6:1:1:360 length=36
BBCBBBBBB@@BAB?BBBBCBC>BBBAA8>BBBAA@
```





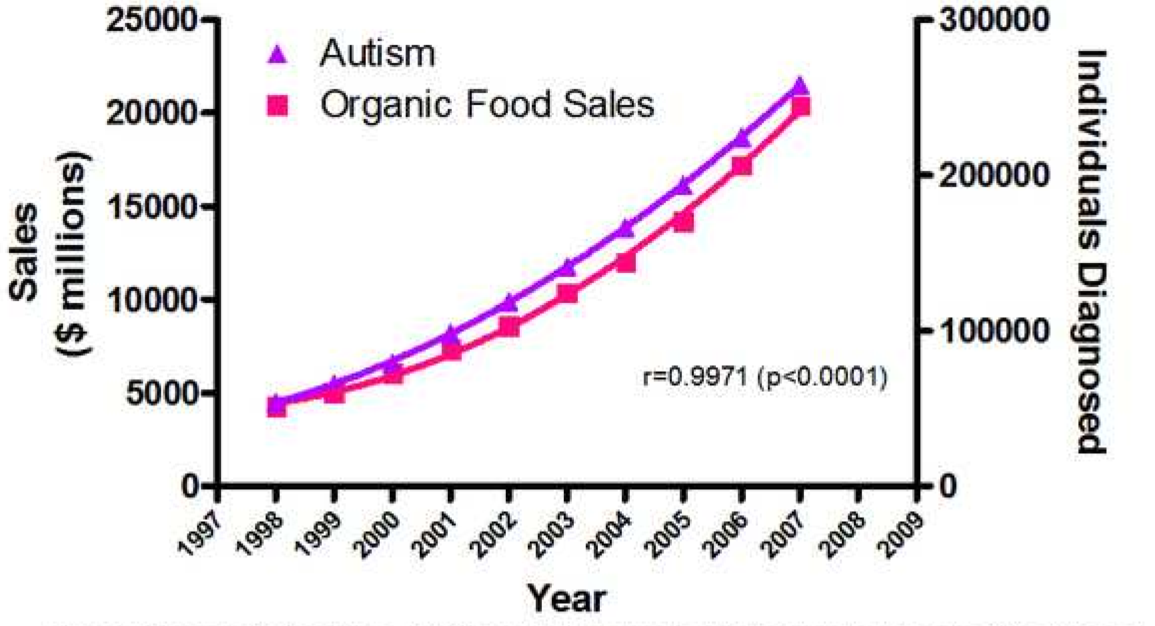
3-Velocity



Data is being generated fast and needs to be processed fast

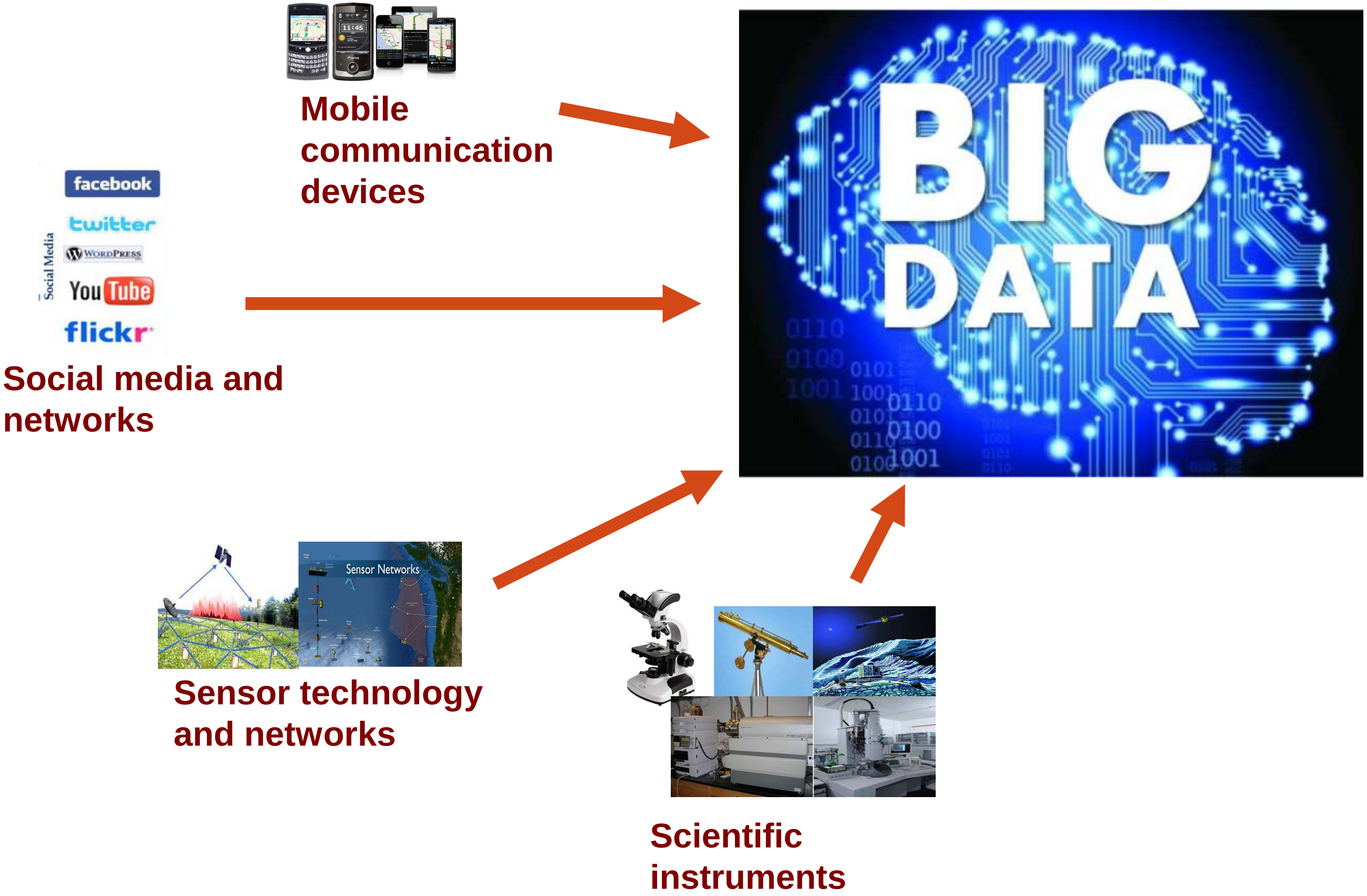
4- Veracity

Does organic food consumption contribute to autism?



Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals
with Disabilities Education Act

Where is it from?



Mobile communication devices

Social media and networks

Sensor technology and networks

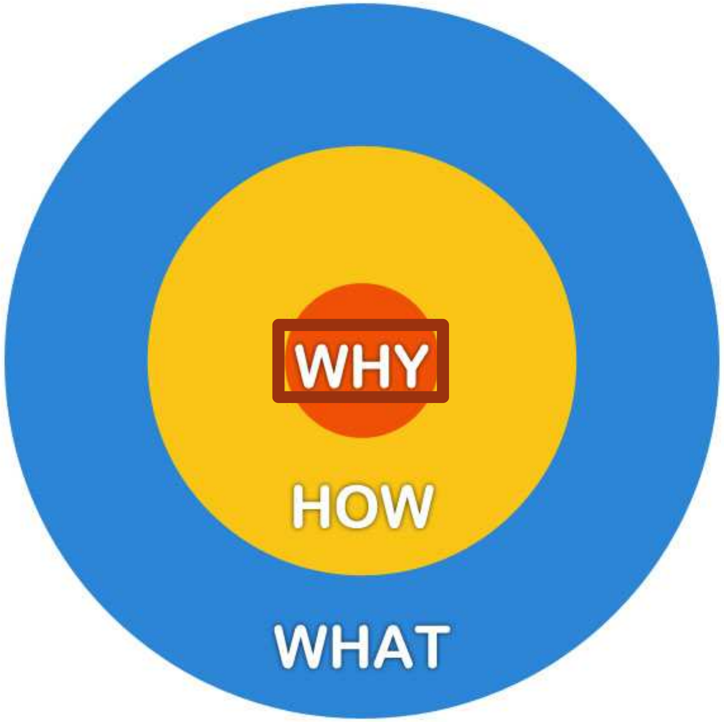
Scientific instruments

Data Generation & Consumption

Old Model: Few companies were generating data, while all others consume it

New Model: All of us are generating data, and all of us are consuming data

Why Big data



Progress and innovation is no longer hindered by the ability to collect data…

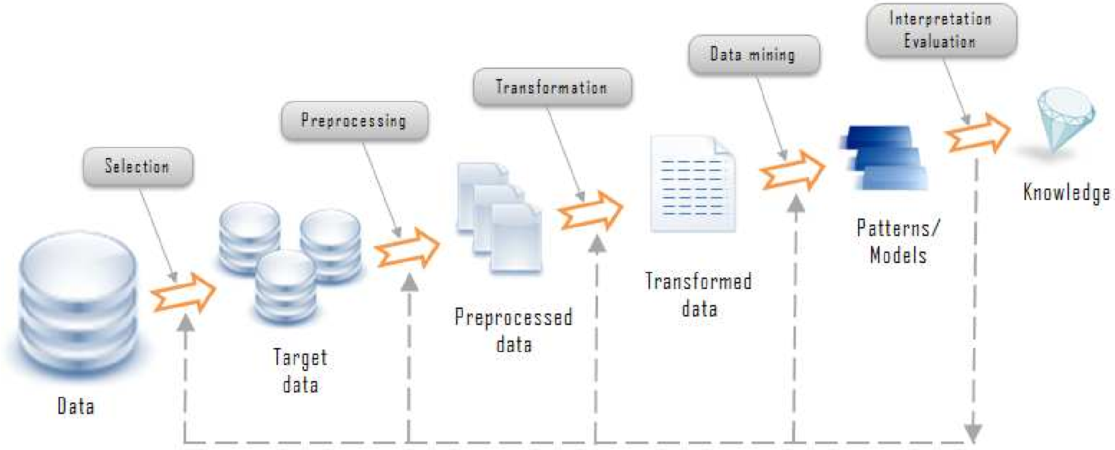
…but, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

Data to Knowledge

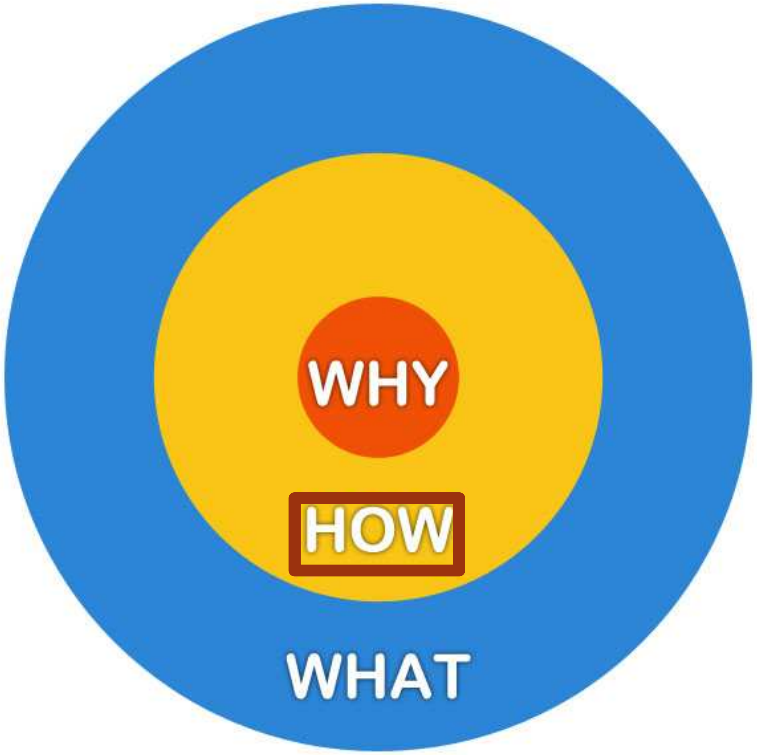
What drives big data?

Technology – “Big data: Only for the privileged”

Value – “Data is the new oil/ currency”



Big data in Biology



How is it generated?

A historical perspective of Bioinformatics data

- The 1960s: the birth of bioinformatics
  - computer languages
  - Academic access to computers
  - First protein database

IBM 7090 computer

Data Processing Developments

1960

IBM 7090 computer

32 Kbytes RAM

2.18 µHz

$2,900,000 in 1960

2015

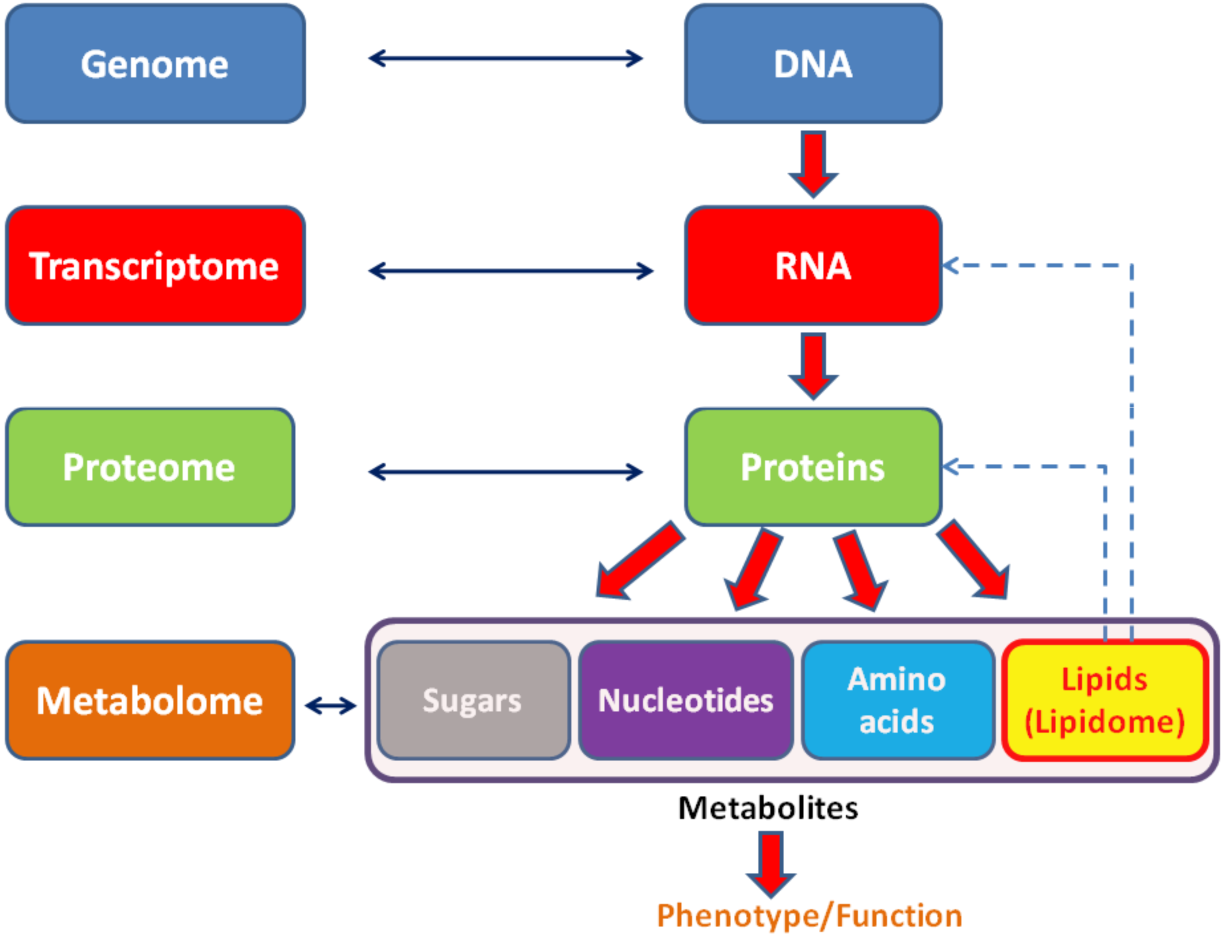
21.5” Apple iMac

8 GB RAM

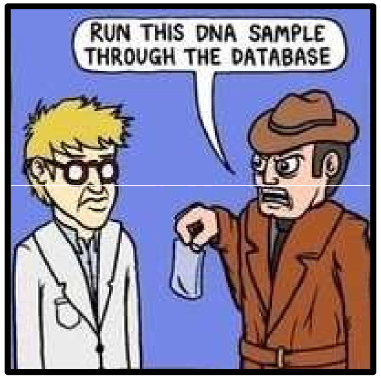
2.7 GHz

$1,149 in 2015

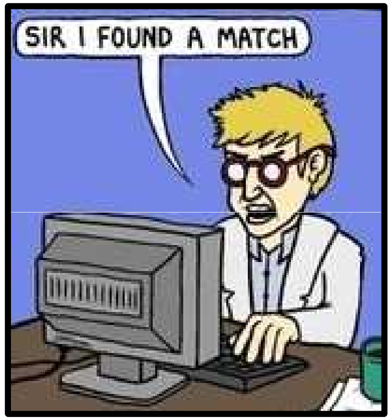
BIG “Omics” DATA



Generation of OMICS DATA



illumina
Roche
life technologies
PACIFIC BIOSCIENCES
Helicos BioSciences Corporation
ion torrent
Oxford NANOPORE Technologies



NGS technologies

Biological Databases & Tools

UniProt
Universal Protein Resource
prosite
Fasta
Similarity and Homology Searching
WU-Blast2
Basic Local Alignment Search Tool
IntEnz
SeqVISTA
Pfam
PRINTS
Protein Fingerprints Database
InterPro
CAST
GENOMES
ClustalW
Multiple Sequence Alignment Tool
emboss
ProDom
SMART
Alternative Splicing Database
Integr8
CSA
Catalytic Site Atlas
PDBj
Protein Data Bank Japan
TIGR
tigr fams
PANTHER
Classification System
caBIG
SeqHound
PIRSF
Gene3D
BioLayout
NCBI
LION
Superfamily
SCOP
DALI
GeneQuiz
PubMed.gov
U.S. National Library of Medicine
National Institutes of Health
caBIO
MOUNT SINAI HOSPITAL
CATH
Protein Structure Classification
MSD
e!
Ensembl
Parasite Blast
DDBJ
DNA Data Bank of Japan
wellcome trust
sanger
institute
SWISS-MODEL
ModBase
Genome Review
IUPHAR
KEGG
Kyoto Encyclopedia of Genes and Genomes
PathPort
The Pathogen Portal Web Project

Challenges in Handling Big Data

- **The Bottleneck in technology**

  New architecture, algorithms, techniques are needed

- **Technical skills**

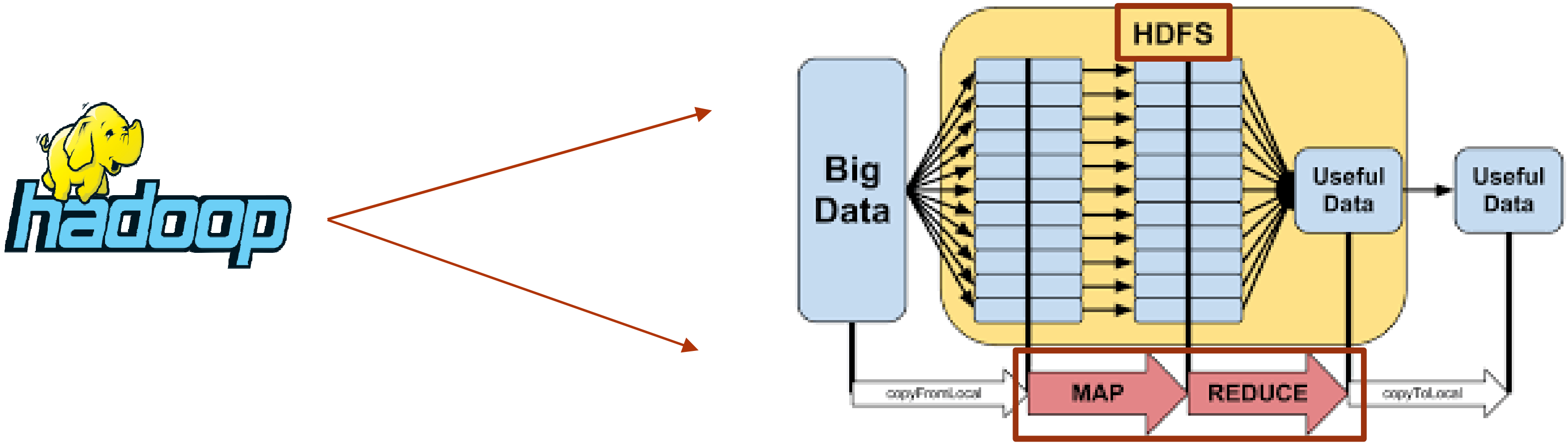
  Experts in using the new technology and dealing with big data

Big Data Analytics

Challenges with existing DBMS tools

- Fixed Schema
- Cost
- Saving/Accessing huge files
- Performing Analysis

Examining data to uncover hidden patterns



Cloud computing

“computing in which dynamically scalable and virtualized resources are provided as a service over the internet”

Pros	Cons
Scale	Security
Cost is relative on scale	Lack of control
Choice	Reliability
Access to NGS architecture	