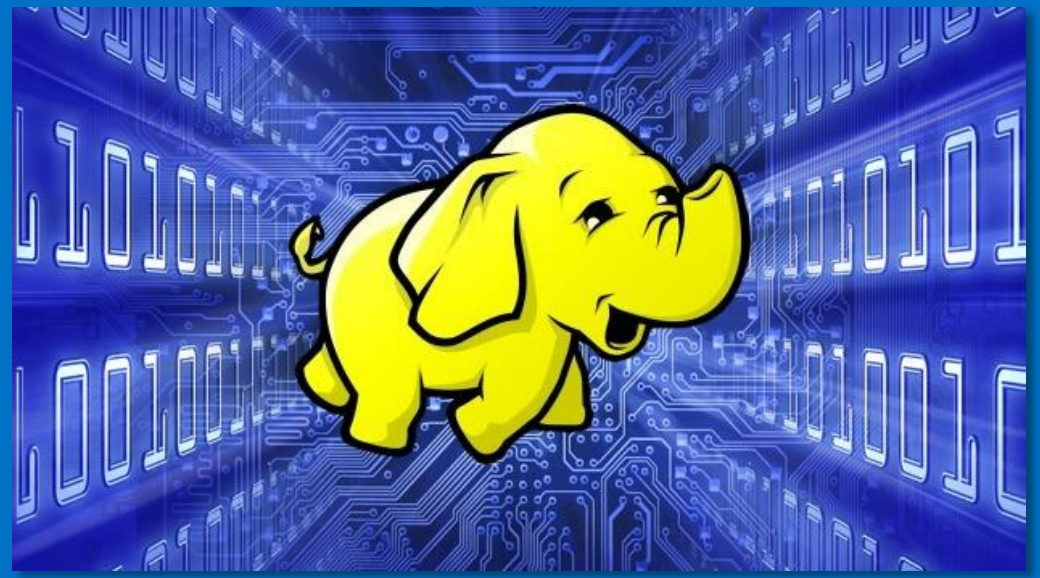


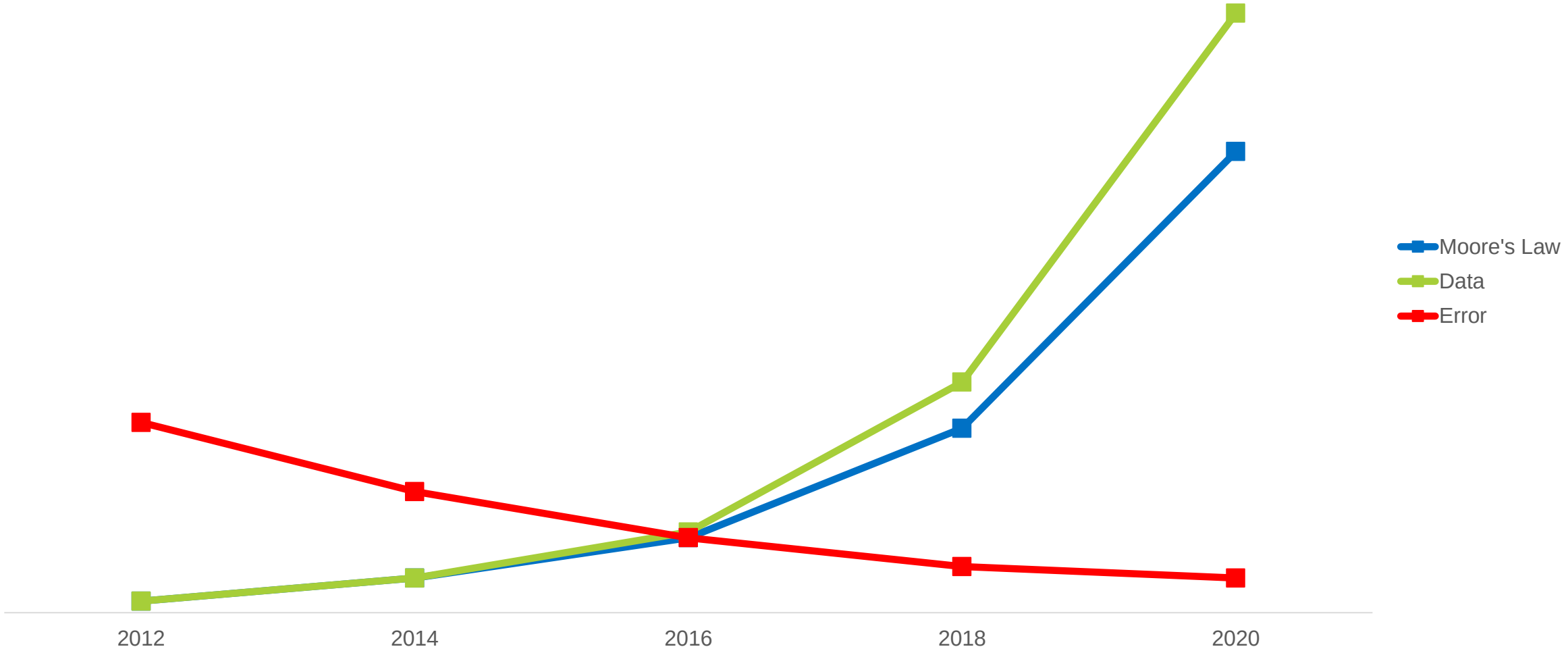
# Intel and Big Analytics

Richard Pilling
Director, Analytics and Big Data EMEA


FOLLOW
YOUR
DREAMS
CANCELLED

# Why does Intel Care about Big Data?

## Data is Growing Faster than Moore's Law




Source: https://amplab.cs.berkeley.edu/2013/02/07/for-big-data-moores-law-means-better-decisions/

# Great Partnerships Yield Great Results


# Faster Insights, Better Security, Less Complexity

cloudera
Ask Bigger Questions

+

intel



## Accelerate Hadoop Adoption

- Maintain an open horizontal platform for big data
- Continue to enhance Apache Hadoop and related projects

## Enable Cloudera to Run Best on Intel

- Optimize performance across compute, storage, & network
- Ensure platform performance, security, management

## Empower the Big Data Ecosystem

- Establish usage models and industry standard benchmarks
- Develop reference architectures and industry-wide solutions

# Business

Compute Bound

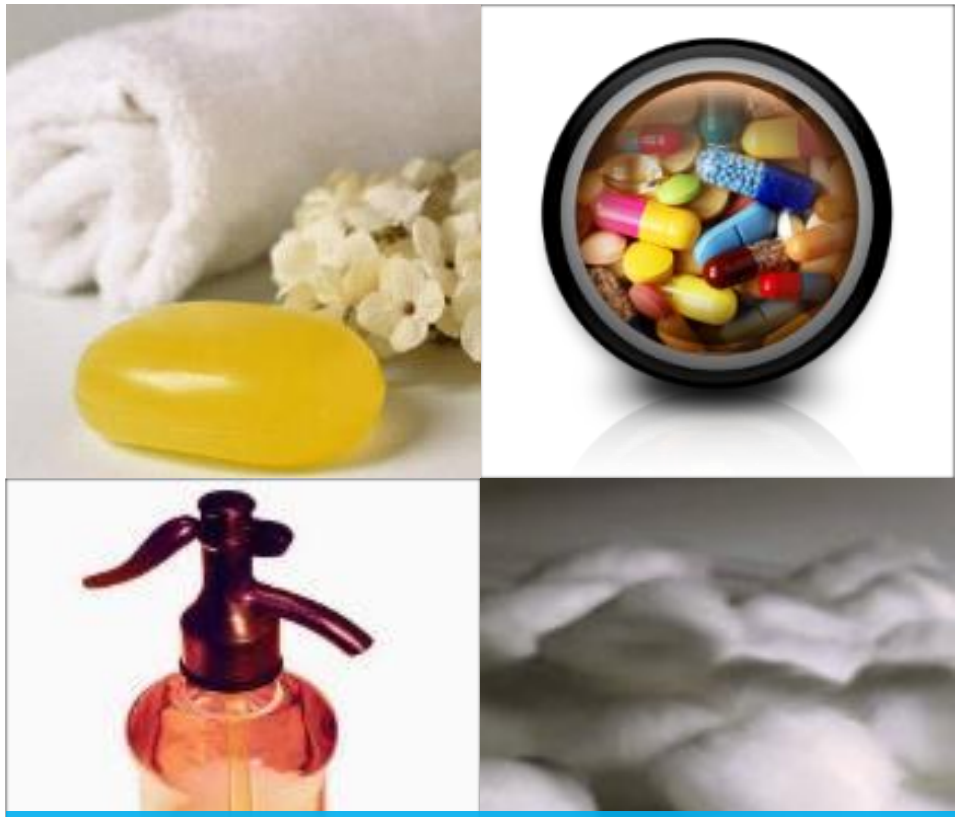
# Human

Storage Bound

# Machine

Network Bound

# Business Data



Retail – Customer Loyalty

What do the [illegible]cts have in common?

- large quantities of calc[illegible]esium and zinc
- unscented lotion [illegible]
- large bags of c[illegible]
- hand sanitize[illegible]

High Guest Pregnancy Prediction Score

www-nc.nytimes.com/2012/02/19/magazine/shopping-habits.html

# Machine Data
## Federal Government

Machine generated: **e.g. Video Analytics**

**Up to 65mile² image**

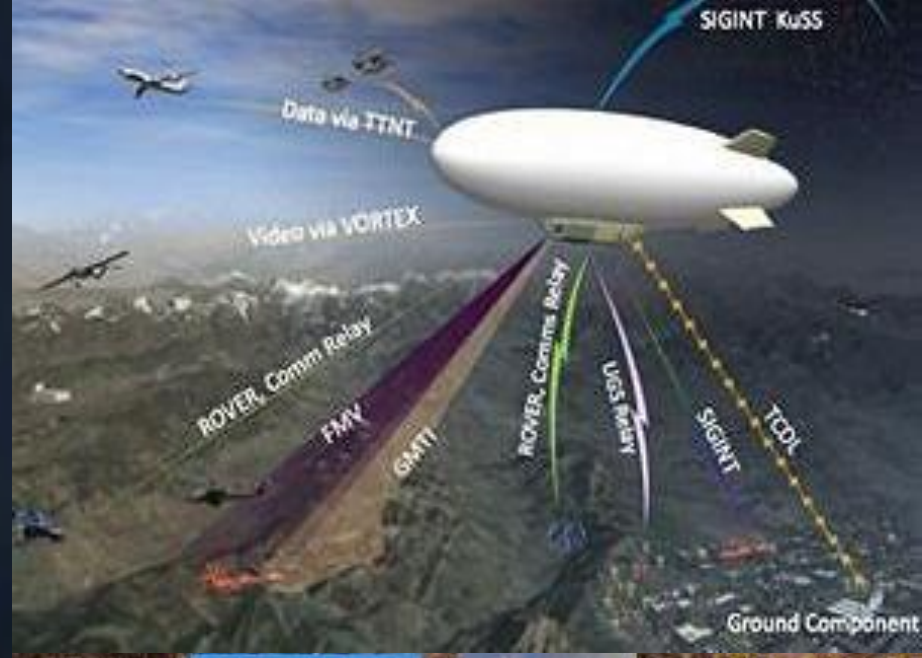
**1.8GP image** via 92 – 5MP imagers

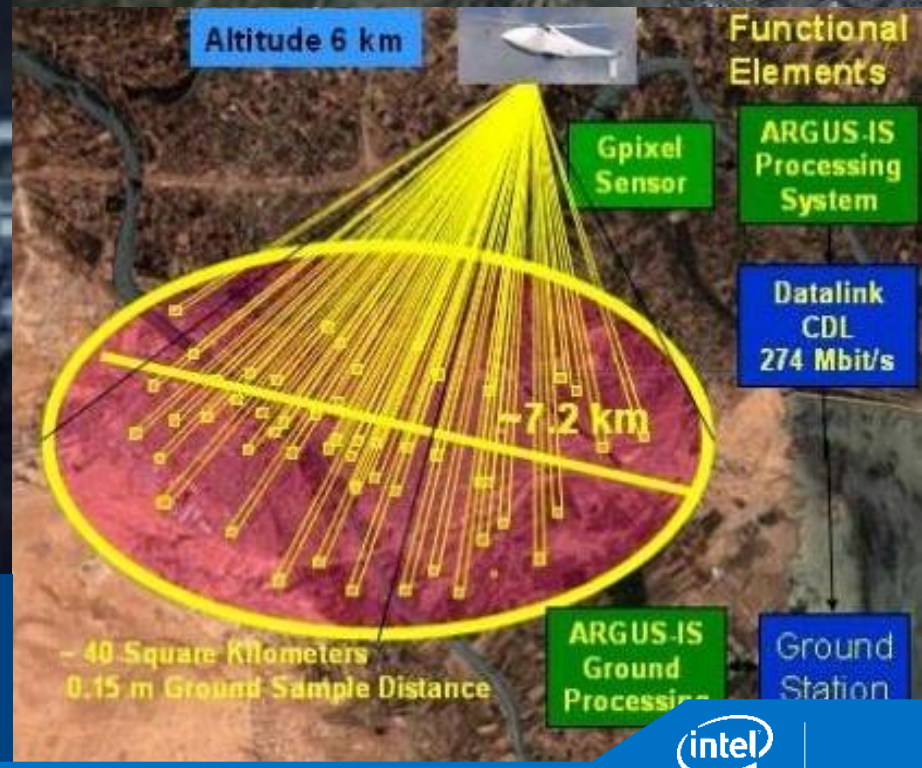
Tracks 65 targets real-time down to 6cms

12-hours flights @ ~20Kft

**6PB HDV per day**

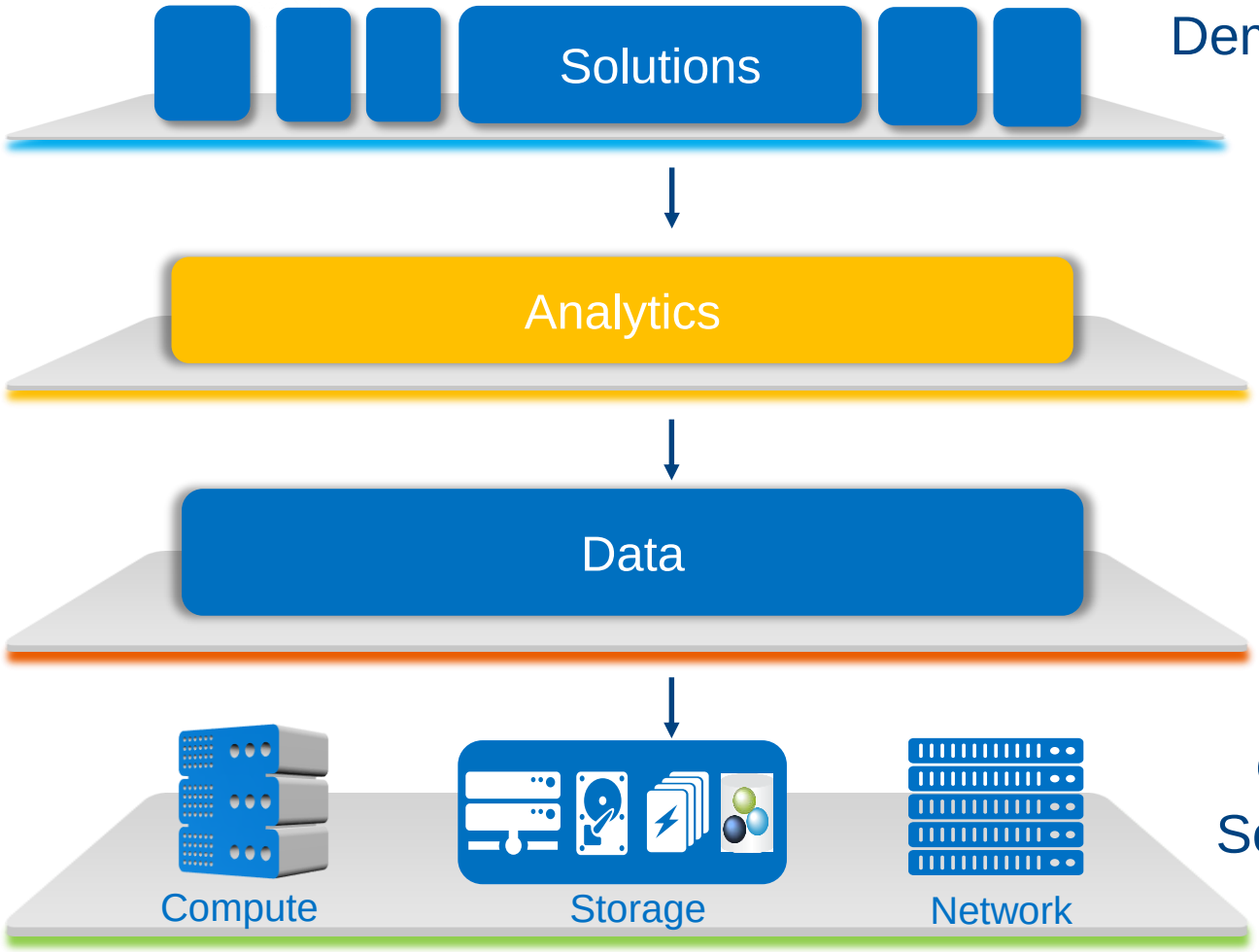
**x30K drones over next 10 years / 110 bases**




# 1/3 Zettabyte per week

# Intel focus on big data analytics from top-to-bottom



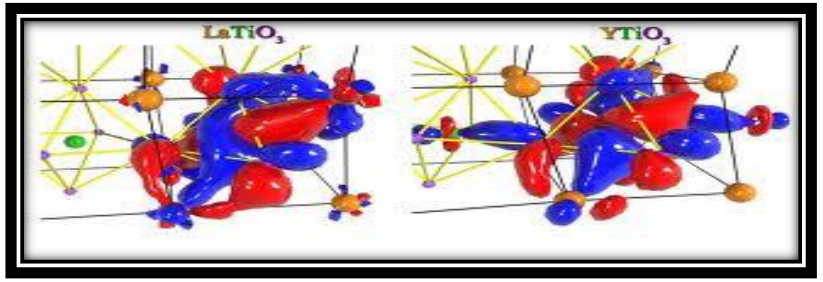

Demonstrate value of big data analytics with lighthouse solutions

Drive mainstream use of advanced analytics by reducing cost & complexity

Accelerate innovation in and adoption of Apache Hadoop as open data platform

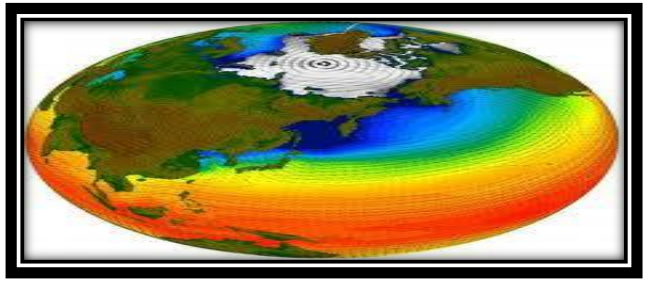
Catalyze industry growth by supporting Software Defined Infrastructure with open, standard building blocks

# HPC's Current Frontier




**Molecular Structures**

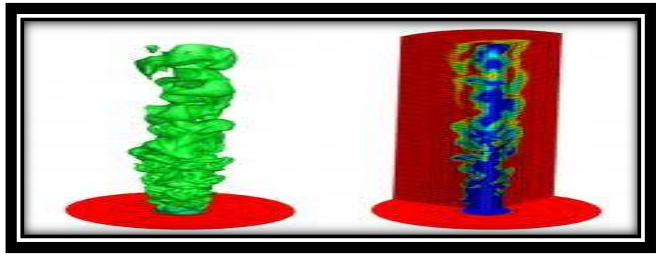
**Genomics**



**Weather & Climate**

**Manufacturing & Design**

**Financial Markets**



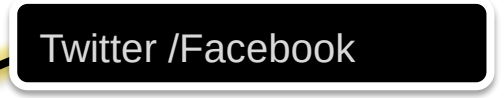
**Biofuels**

Large-Scale, Cumulative Computation
Distributed Over Multiple Connected Nodes
Intense Anything (Memory, I/O)…
Intense Floating-Point, Integer

Simulation, Technical
Batch, Highly Parallel
Scientific
Expensive

# Big Data – what is different?

Quickly anticipate demand for medication and health services




Social media updated with "feeling ill" messages


Have emergency room visits spiked suddenly?

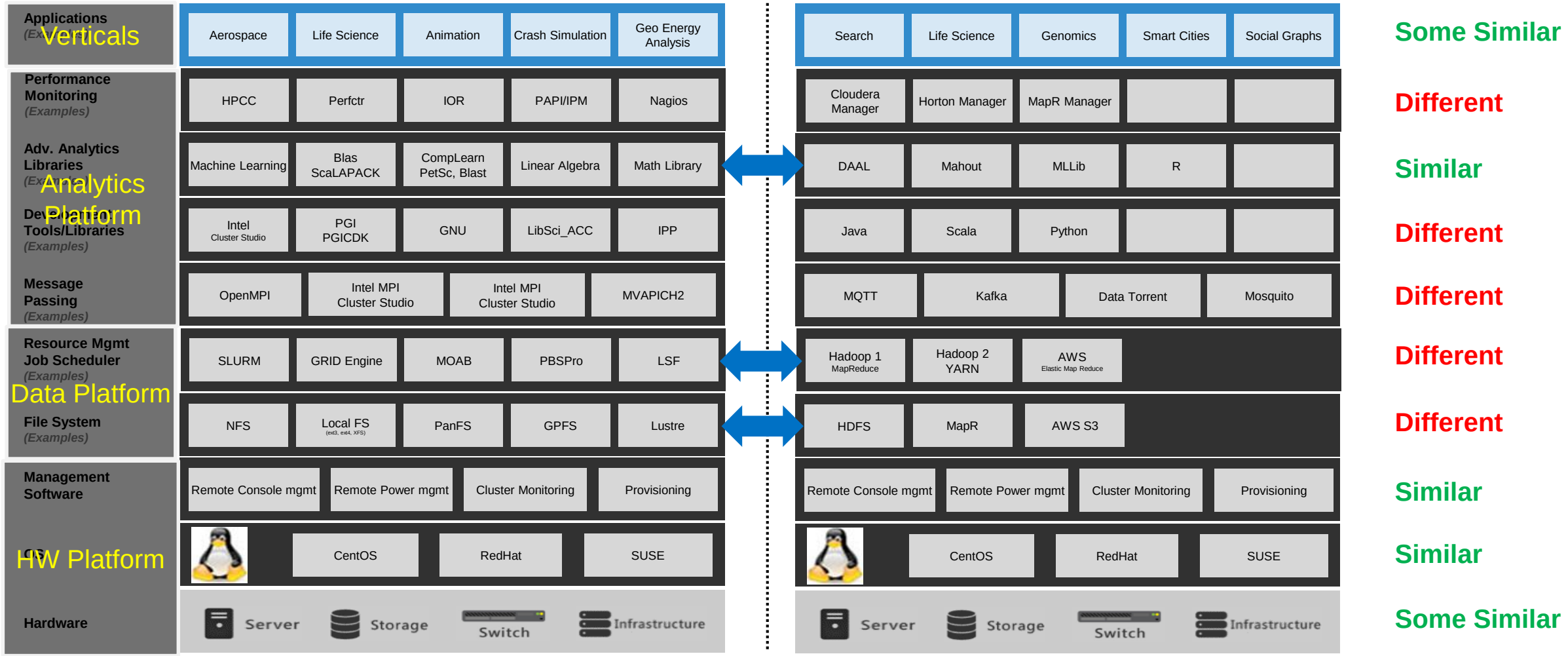

More than usual over-the-counter medications sold recently?

Major event in the city with large number of travelers?

Sudden change in the weather?

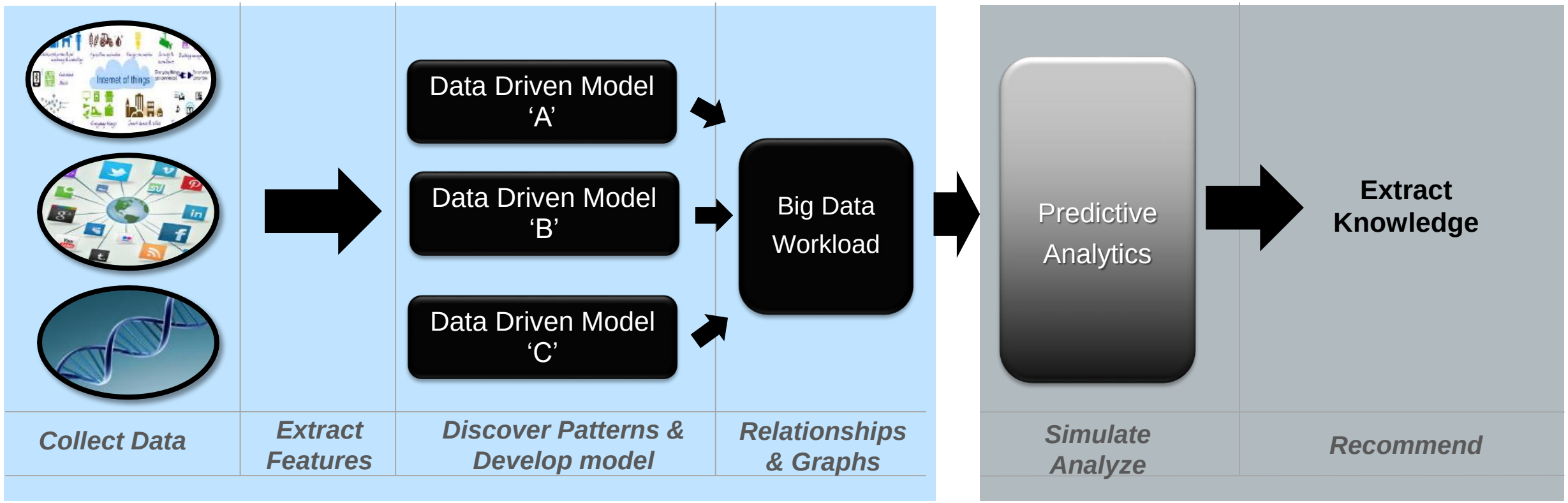
## Transform Data Into Knowledge (vs. today's Knowledge into Data)

# HPC and Big Data Analytics Stacks




## Emergence of High Performance Analytics

# High Performance Analytics Workflow




Compute On/With
Dependencies & Relationships

# The Big Data Paradigm Change

Codes Based on Analytic Models → Codes Based on Data Driven Models

**Well Known HPC Workloads**

## HPC Today

- ✓ Compute Focused
- ✓ Minimizes Data Movement
- ✓ I/O predominantly For Checkpoints
- ✓ Datasets are ~Petabytes
- ✓ Data for compute is sampled or generated

**New Big Data Workload**

## HPC Tomorrow

- ✓ I/O Focused
- ✓ Lots of Data Movement
- ✓ I/O predominantly for storing & retrieving data
- ✓ Datasets are ~100's of Petabytes
- ✓ All data is needed all the time

System Design Points Will Change!