Big Data

Risks and Opportunities

Kenny Huang, Ph.D. 黃勝雄博士

Executive Council Member, APNIC
亞太網路資訊中心董事
huangksh@gmail.com
2015.06.09

Agenda

1	*Big Data Technologies*
2	*Big Data Market*
3	*Opportunities, Risks, & Capital Trends*
4	*Algorithmic Accountability & Privacy*

Internet of Things Definition

The
Economist

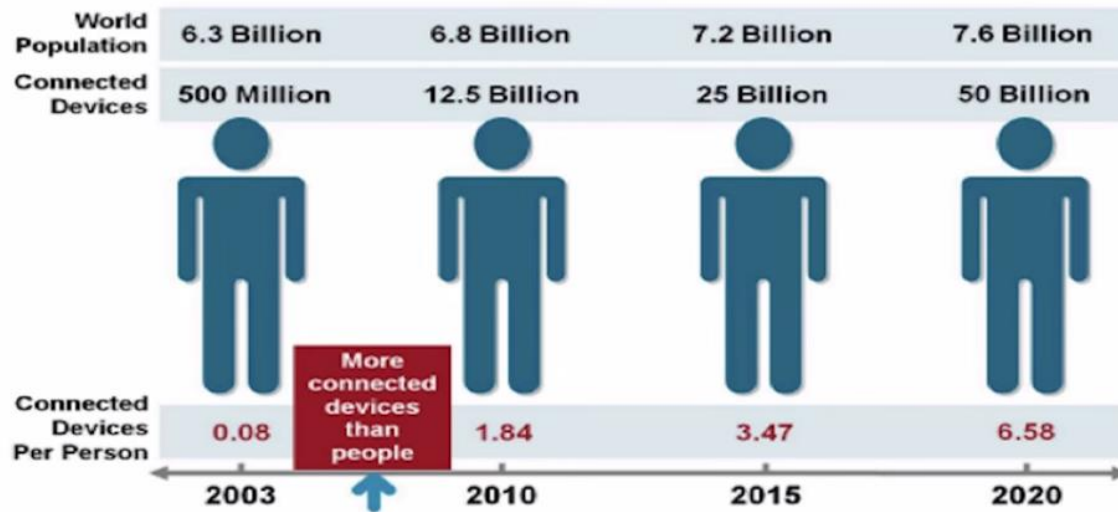"In the world of IoT, even cows will be connected"

Source : The Economist 2010

WIKIPEDIA
The Free Encyclopedia

The **Internet of Things** (**IoT**) is the network of physical objects or "things" embedded with electronics, software, sensors and connectivity to enable it to achieve greater value and service by exchanging data with the manufacturer, operator and/or other connected devices. Each thing is uniquely identifiable through its embedded computing system but is able to interoperate within the existing Internet infrastructure.

IoT = Devices (RFID tags, Sensors, ..) + Networks + Services + Data + Analytics

IoT Size & Potential

World Population	6.3 Billion	6.8 Billion	7.2 Billion	7.6 Billion
Connected Devices	500 Million	12.5 Billion	25 Billion	50 Billion

More connected devices than people

Connected Devices Per Person	0.08	1.84	3.47	6.58
	2003	2010	2015	2020

Source : CISCO

IoT Potential Applications

- ✓ Urban Planning
- ✓ Smart Cities
- ✓ Sustainable Environments
- ✓ Healthcare
- ✓ Emergency Response
- ✓ Waste Management
- ✓ Intelligent Shopping

- ✓ Smart Product Management
- ✓ Smart Meters
- ✓ Smart Homes
- ✓ Smart Automobiles
- ✓ Smart Agriculture (cows)
- ✓ Smart Grid
- ✓ Intelligent Business Decisions
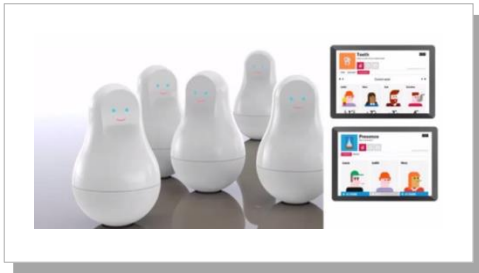
Examples of IoT

Google Glass
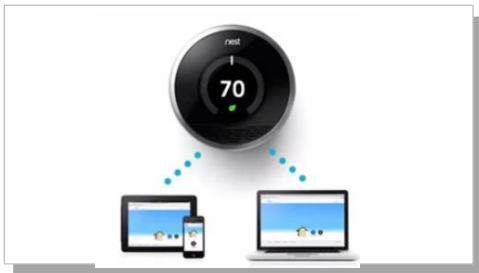
Weight Scale

SkyBell
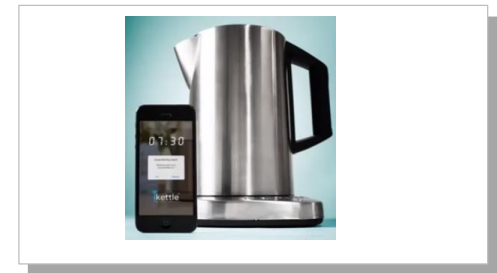
Sense Mother

Jawbone Up

Light bulb

Nest Thermostat

Belkin Wemo

Firefox iKettle

IoT Architectural Design

Question : How to build systems that work well ?

- ✓ Breaking them into tractable components.
- ✓ "Modularity based on abstraction is the way things get done." – Liskov
- ✓ If you can't manage, evolve, or understand a system, probably don't have the right abstraction.
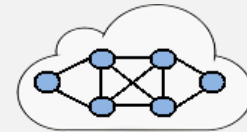
Cloud Data Centre
Web server hosting

Core Network
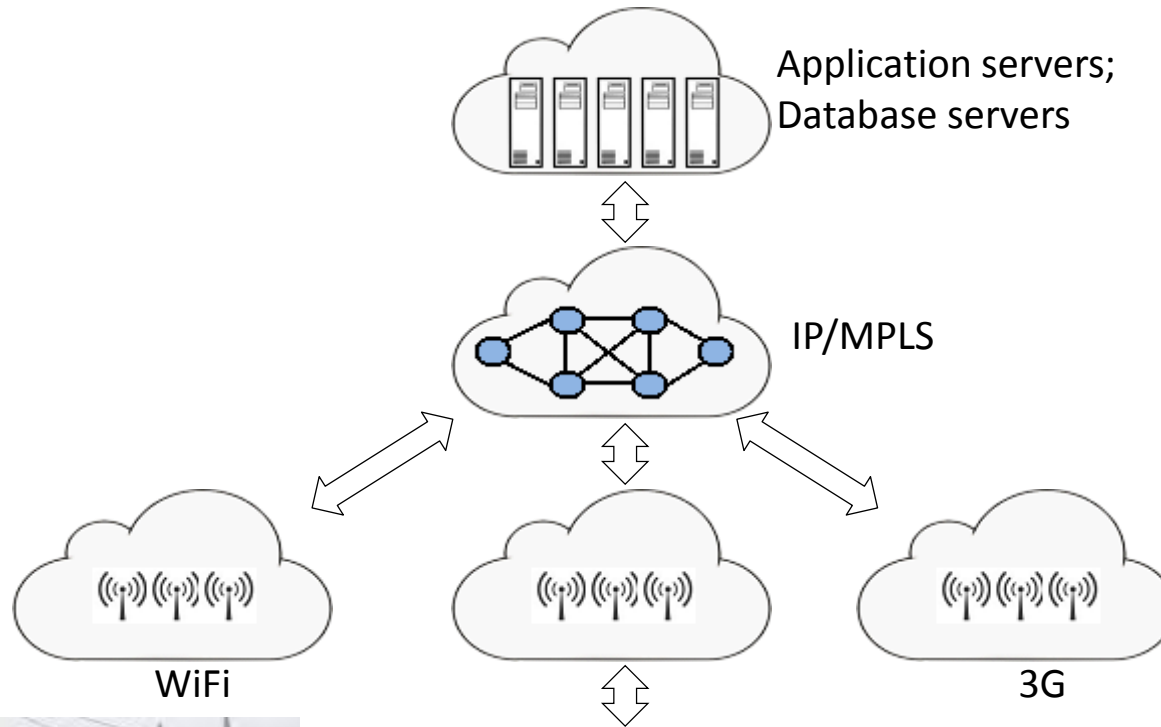IP/MPLS

Access Network
Ethernet; Mobile; WiFi

Things Network
RFID; NFC; Bluetooth

Energy Administration Architecture

Application servers;
Database servers

IP/MPLS

WiFi

3G

Reliability improvement

Water vapour

Furnace   Steam turbine   Cooling tower

Generator

Coal   Power lines

Water   Condenser

Condenser cooling water

Asset management

Smart metering

Water Administration Architecture

Application servers;
Database servers

IP/MPLS

WiFi

3G

Field devices

Smart metering

Quality monitoring and
Water control
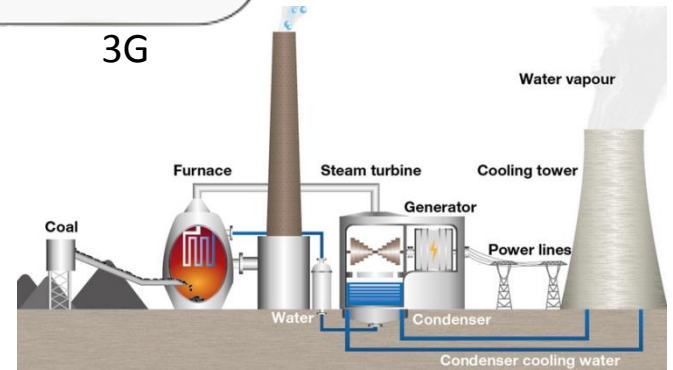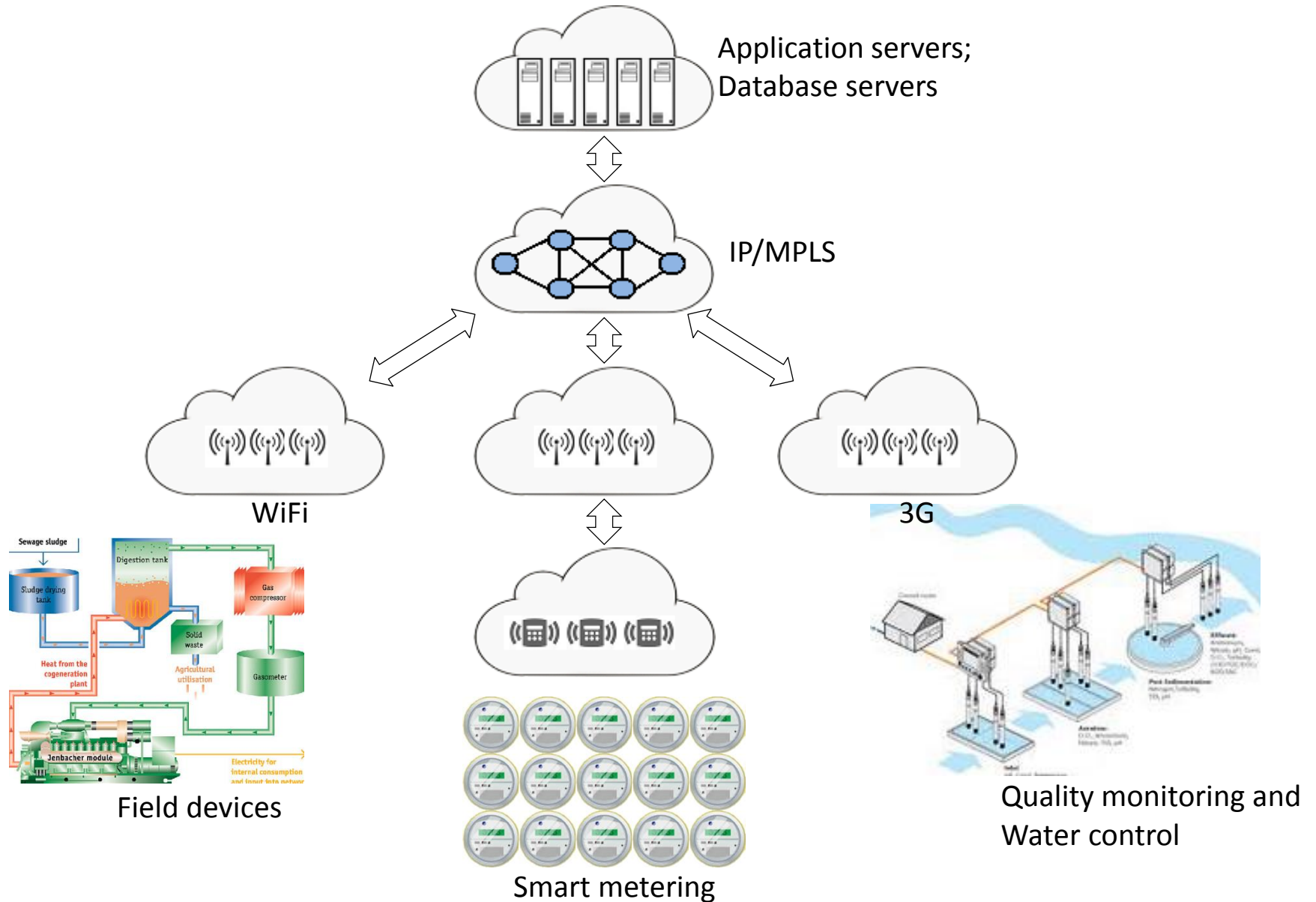
Smart City Architecture

Application servers;
Database servers

IP/MPLS

WiFi

3G

Public lighting

Intelligent transportation

Public safety management

IoT Technologies

Sensor Analytics

detect anomalies

Big Data Analytics

analyze a mix of structured, semi-structured and unstructured data

Stream Analytics

enables developers to combine streams of data with historic records to derive business insights

Machine Learning Analytics

building of the predictive models by using machine learning techniques

Real-time Analytics

enables business users to get up-to-the-minute data by directly accessing OS

Statistical Analytics

find patterns and identify trends

Big Data

ASA

2005 – Roger Magoulas uses the term "Big Data"

International Year of Statistics - 2013

McKinsey Global Institute

Big Data: The next frontier for innovation, Competition, and productivity.

McKinsey Global Institute, May 2011

White House

Big Data Initiative : $200 Million in New R&D Investment on Big Data for scientific discovery, environmental and biomedical research, education, and national security

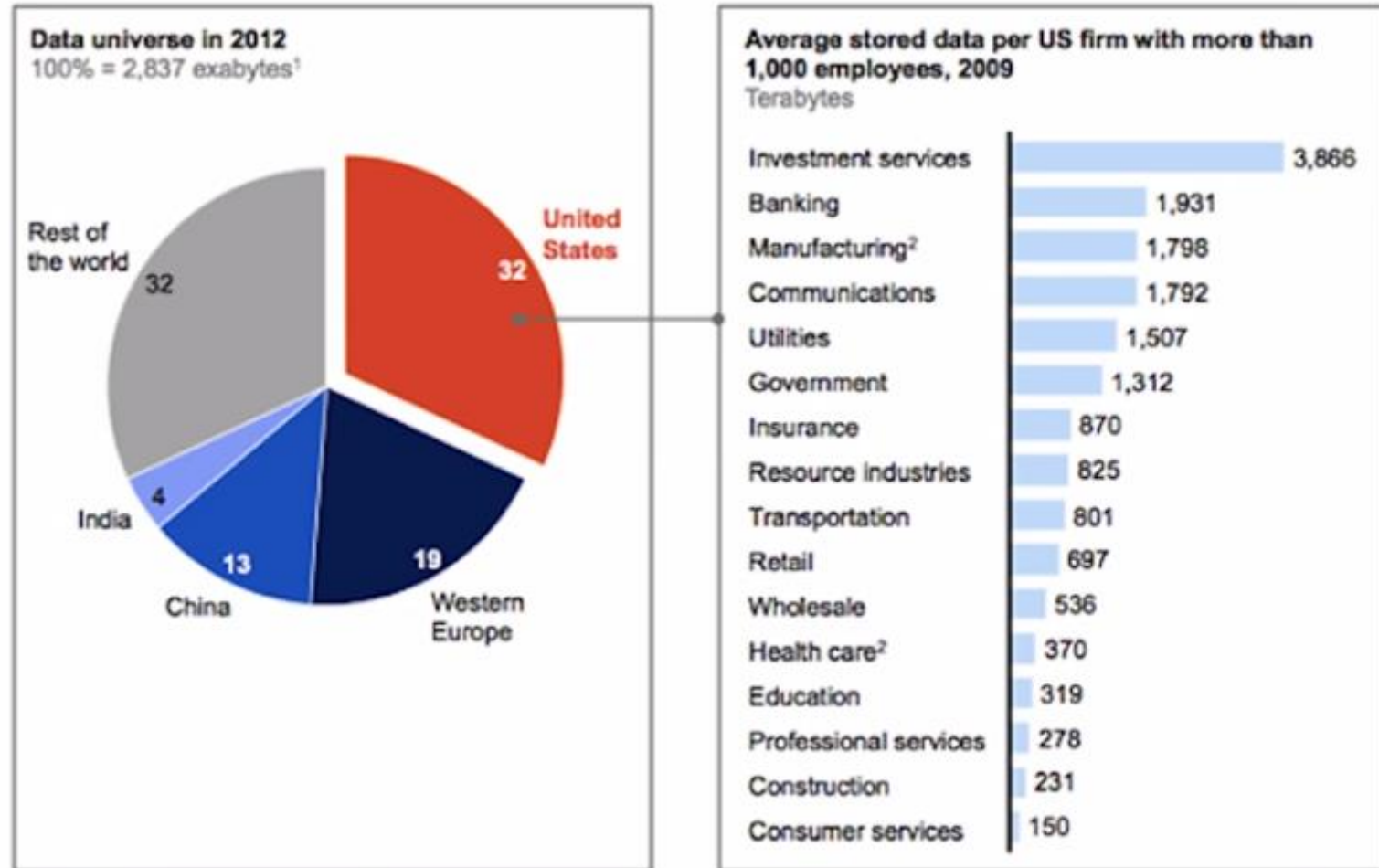*Press Release. White House of OSTP. March 29 , 2012*

The New Oil

As far back as 2006, market researcher Cliver Humby declared data "the new oil." Just as oil once fired dreams a century or more ago, data is today driving a vision of economic and technical innovation. If "crude" data can be extracted, refined, and piped to where it can impact decisions in real time, its value will soar.

CISCO ISBG, June 2012

Big Data : Size

The United States has one-third of the world's data

Data universe in 2012
100% = 2,837 exabytes[1]

Rest of the world: 32
United States: 32
India: 4
China: 13
Western Europe: 19

Average stored data per US firm with more than 1,000 employees, 2009
Terabytes

Industry	Value
Investment services	3,866
Banking	1,931
Manufacturing[2]	1,798
Communications	1,792
Utilities	1,507
Government	1,312
Insurance	870
Resource industries	825
Transportation	801
Retail	697
Wholesale	536
Health care[2]	370
Education	319
Professional services	278
Construction	231
Consumer services	150

1 One exabyte = 1,024 terabytes, nearly 2.5 times as large as US Library of Congress web archive (as of May 2013).
2 The large number of firms in the manufacturing and health-care sectors reduces the available storage per company.
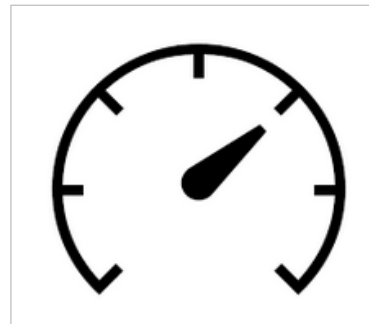SOURCE: IDC; US Bureau of Labor Statistics; US Library of Congress; McKinsey Global Institute analysis
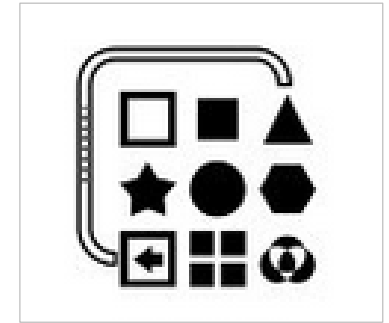
The 3Vs of Big Data

Volume

90% of the data in the world today was created within the last two years

People to People
People to Machine
Machine to Machine

Variety

Velocity

2.9 emails sent
every second
20 hours of video uploaded
every minute
50 million tweets per day

Big Data Ecosystem

Generation	Operational IT	Analytics	Usage
Data Class Types	**Store Access Prepare**	**Analyze Visualize**	**Analyze Business**

Data Types
- ✓ Structured (relational)
- ✓ Unstructured (adhoc)

Data Mgmt. & Storage
- ✓ Store
- ✓ Secure
- ✓ Access
- ✓ Network

Data Analytics
- ✓ Algorithmics
- ✓ automation
- ✓ In Real Time

Business Analysis
- ✓ Decision Support
- ✓ Just in Time Business Model

Data Classes
- ✓ Human
- ✓ Machine

Engines
- ✓ Hadoop MapReduce
- ✓ Apache Tools
- ✓ Cloudera/IBM/EMC
- ✓ Visualization

Business Analytics
- ✓ Visualization
- ✓ Interoperate with SQL -RDBMs
- ✓ BI/EDW

Business User
- ✓ Market Penetration Enhancement
- ✓ Cash Flow/ROI

Data Velocity
- ✓ Batch
- ✓ Streaming

Prepare Data For Analytics
- ✓ ETIL / Data Integration
- ✓ Workflow Scheduler
- ✓ System Tools

Source : Sybase

Analytics: Static Data vs. Streaming Data

Static Data	Streaming Data
Multiple Passes	Single Pass
Persistent	Inherently Temporal
Offline Analytics	Online as well as Offline Analytics
Analytics Based on All the Data	Analytics Based on a Subset of Data
Only the current state is relevant	Consideration of the order of the input
Relatively low update rate	Potentially high update rate
Little or no time requirements	Real-time requirements
Assumes exact data	Assumes outdated/inaccurate data
Plannable query processing	Variable data arrival and data characteristics

DBMS (Database Management System)	DSMS (Data Stream Management System)
Persistent relational data	Volatile transient data streams
Random access	Sequential access
One-time queries	Continuous queries
Unlimited secondary storage	Limited main memory
Only the current state is relevant	Consideration of the order of the input
Relatively low update rate	Potentially high update rate
Little or no time requirements	Real-time requirements
Assumes exact data	Assume outdated / inaccurate data
Standing queries	Ad-hoc queries

Big Data Challenges & Data Life of Cycle

- ✓ Sensor data brings numerous challenges with it in the context of data collection, storage and processing. This is because sensor data processing often requires efficient in-network and real-time data stream processing from massive volumes of possibly uncertain data from various sources. The data generated from these sensors arrives in the form of streams.
- ✓ At every phase of the big data life cycle, there are research issues along each steps
- ✓ To handle these streaming sensor data model-based techniques are employed, such as : statistical, signal processing, regression-based, machine learning, probabilistic, time series.

Input Raw Data

Collection

Cleaning, Validation and Serialization

Transformation & Augmentation

DB Storage & Management

Mining & Analytics

Interpretation & Presentation

Output
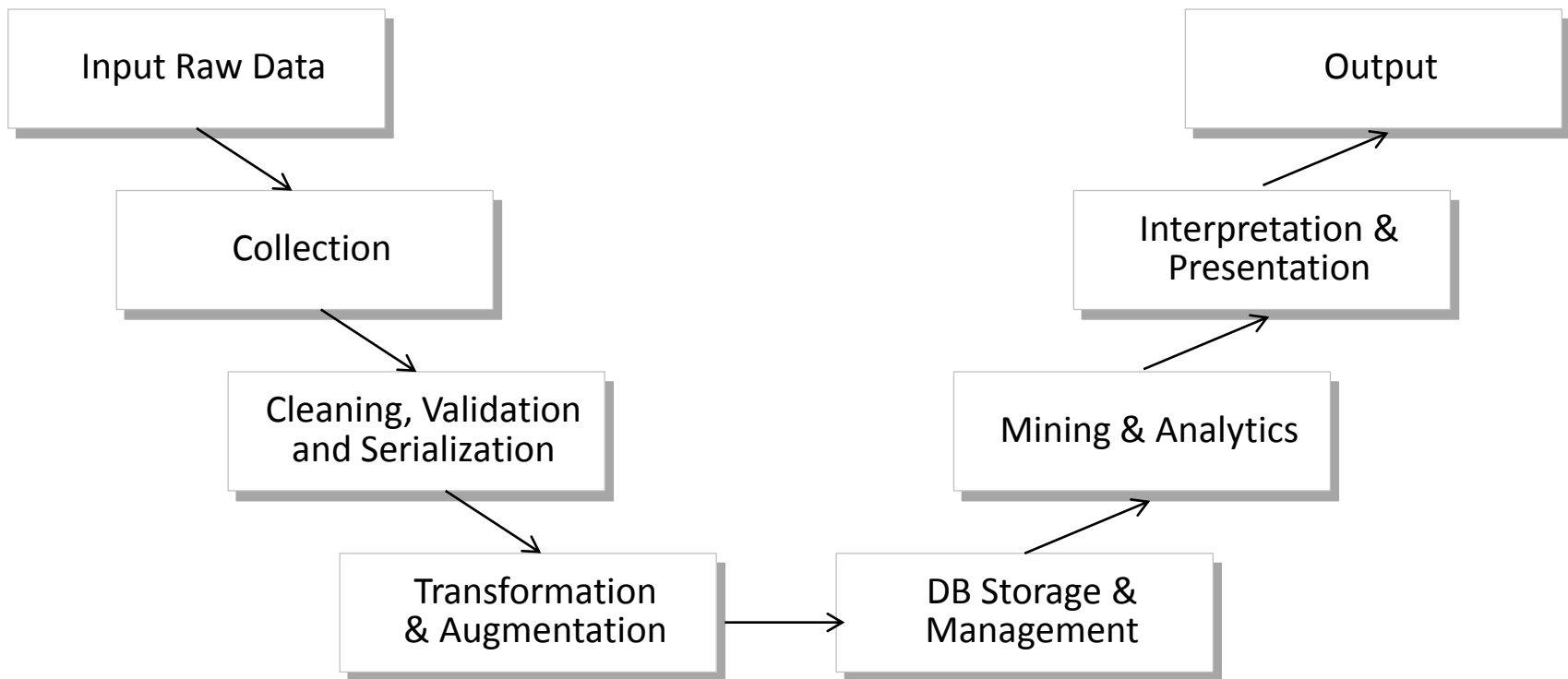
Sensors & Data

Let us denote a sensor network as S= {s_j | $1 \leq j \leq m$} = Sensor network consisting of sensors s_j, where j= (1... m).

s_j = Sensor identifier for a sensor in S.

v_{ij}= Sensor value observed by the sensor sj at time t_i, such as $v_{ij} \in R$, the real numbers.

v_i= Row vector of all sensor values observed at time t_i, such as $v_i \in R^m$.

V_{ij} = Random variable associated with the sensor value v_{ij}.

Data Streams

Let us denote a data stream as D_i = {(t_i, v_{ij})}, an ordered sequence of data tuples, where v_{ij} is the sensor value at time t_i.

A data stream is a structured tuple composed of time (implicit or explicit) and sensor values.

v_{ij}, defined above, is a data stream element.

v_i, defined above, is a data stream.

Data Streams

＊ The general data stream model can be defined as, an infinite tuple of time and values.

＊ In the example, time is defined implicitly by index $_i$ and explicitly by time t_i, sensor identifiers are either the sensor ID s_j or the sensor spatial coordinates x_j (i.e. longitude) and y_j (i.e. latitude), and the sensor reading value v_{ij}.

i	t_i	s_j	x_j	y_j	v_{ij}
1	10:00	1	0.5	3.5	1.8
1	10:00	2	2.0	2.5	0.5
1	10:00	3	4.0	0.5	2.5
1	10:00	4	3.5	3.0	2.9
2	10:15	1	0.5	3.5	1.9
2	10:15	2	2.0	2.5	5.9
-	-	-	-	-	-

Example of Model-based Technique : Kalman Filter

Probabilistic Models: In sensor data cleaning, inferring sensor values is perhaps the most import task, since systems can then detect and clean dirty sensor values by comparing raw sensor values with the corresponding inferred sensor values.
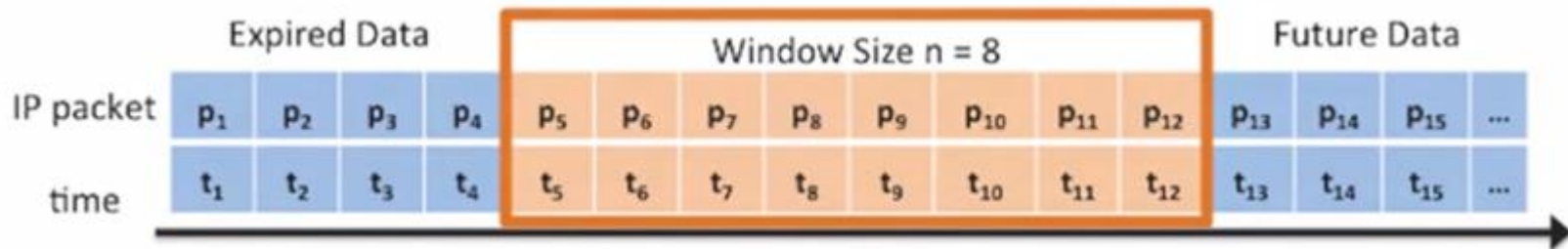**The Kalman filter** is perhaps on of the most common probabilistic models to compute inferred values corresponding to raw sensor values.
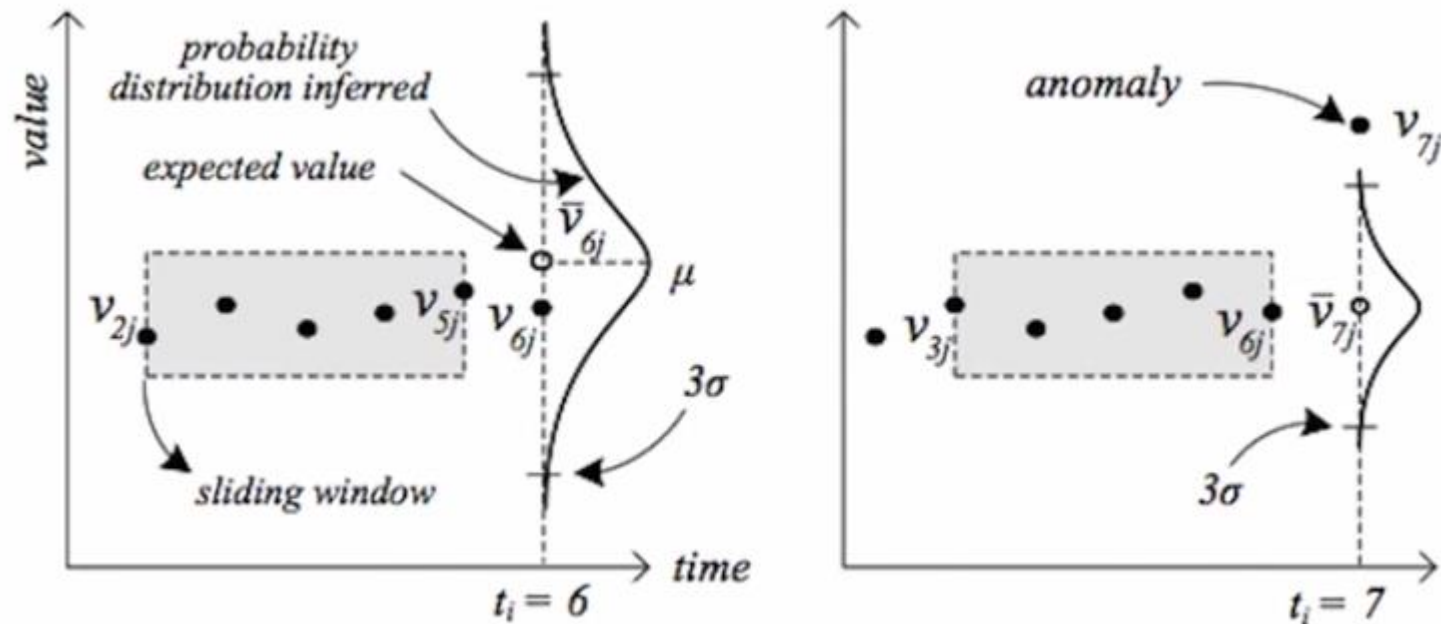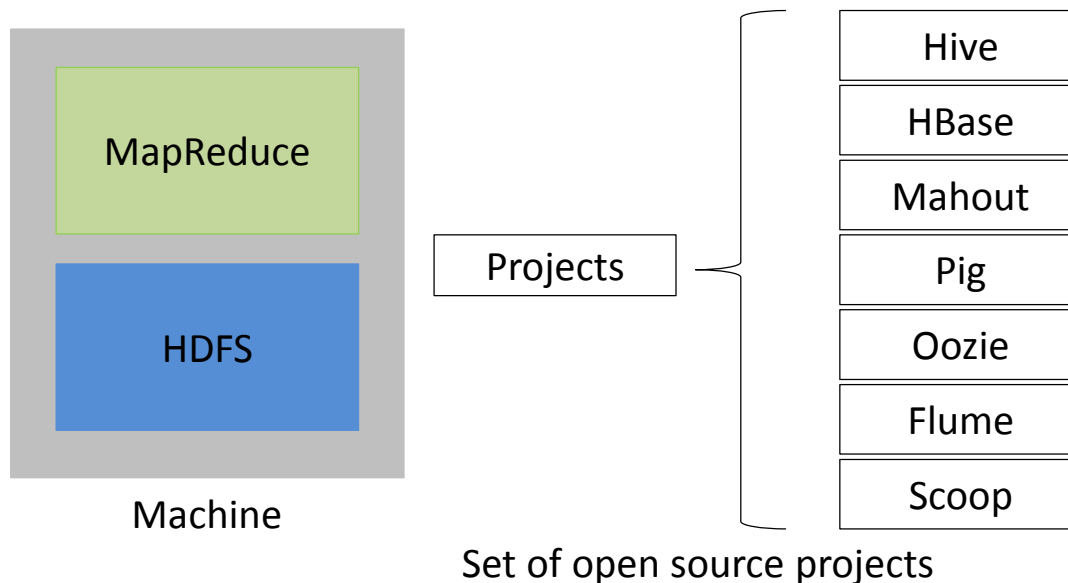
The Sliding Window Model

In the sliding window model, only the recent past is the objective concern of stream processing. The fundamental sliding windows are of fixed size, which are similar to first-in, first-out data structure.

✓ The input is still a stream of data values or elements.
✓ A data value arrives at each time instant; it later expires after a number of time stamps equal to the window size n
✓ The current window at any time instant is the set of data elements that have not yet expired.

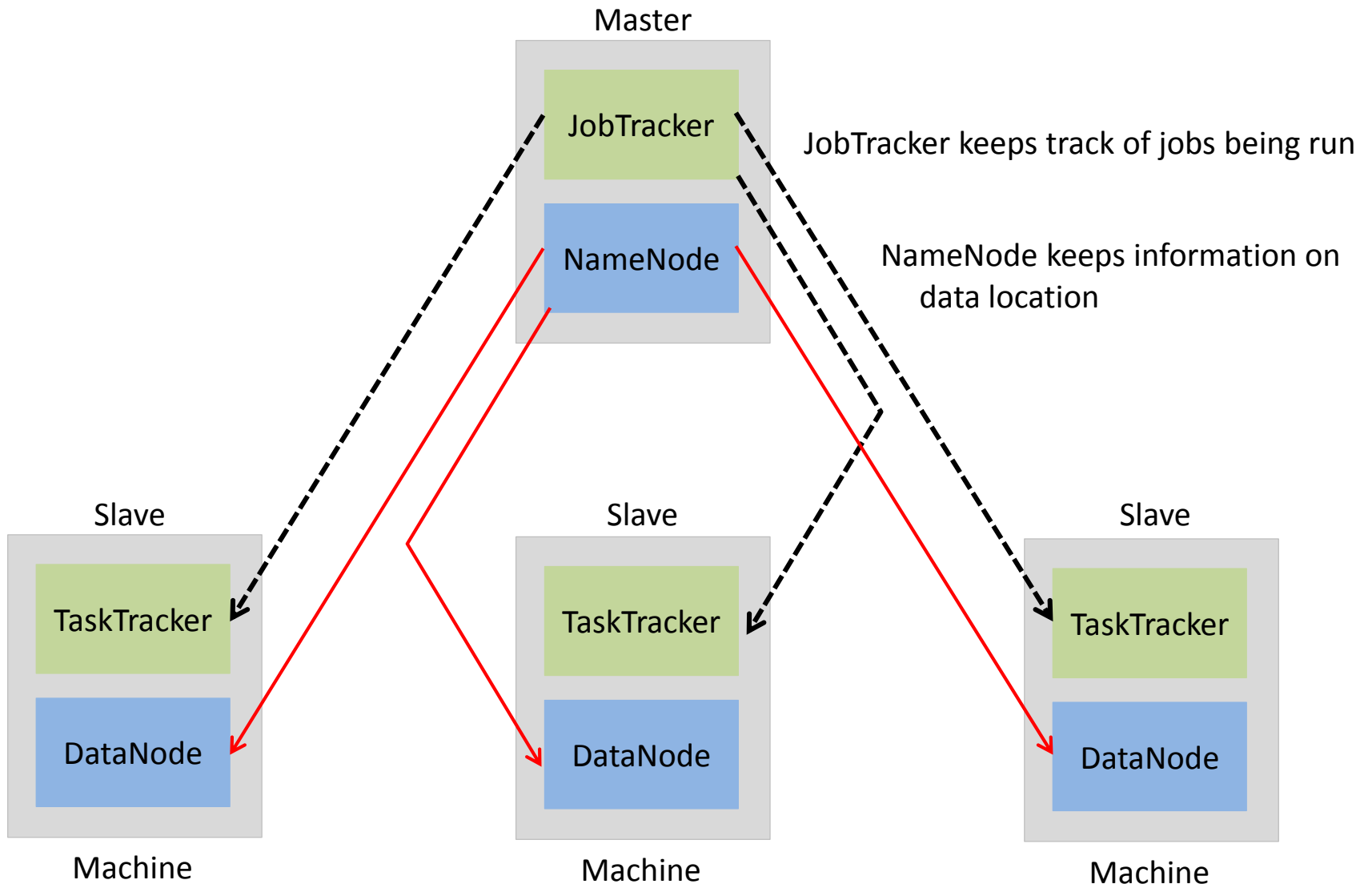|	Expired Data				Window Size n = 8								Future Data			
IP packet	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}	p_{12}	p_{13}	p_{14}	p_{15}	...
time	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}	...

Hadoop

- ✓ Processing Platform for Big Data Processing
- ✓ Using the "MapReduce" processing technique
  - ✓ MapReduce is the processing part of Hadoop
  - ✓ HDFS is the data part of Hadoop
- ✓ Attributes
  - ✓ Highly scalable
  - ✓ Commodity HW-based
  - ✓ Open source: low cost
  - ✓ Batch processing centric

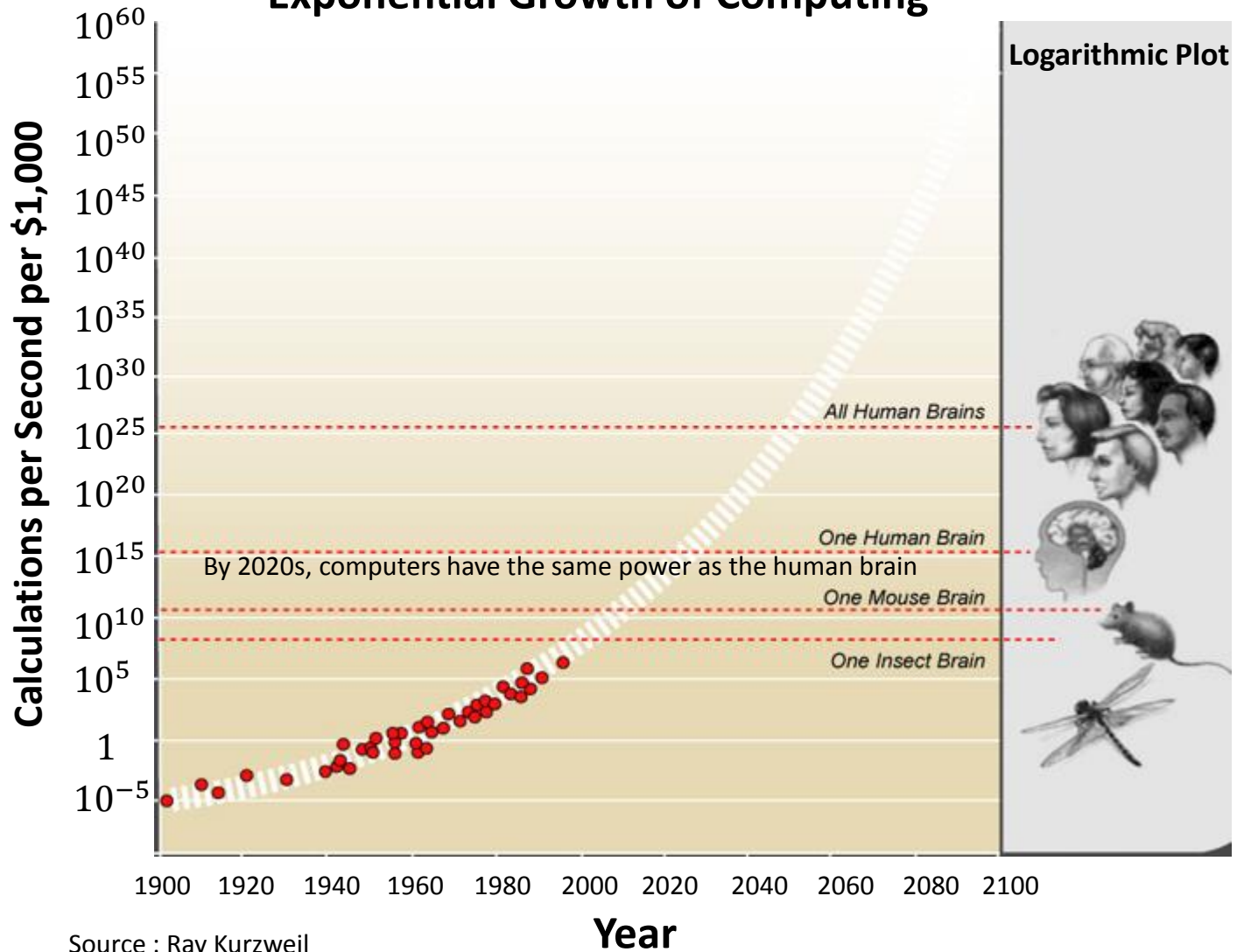| MapReduce |
|---|
| HDFS |

Machine

Projects

Hive
HBase
Mahout
Pig
Oozie
Flume
Scoop

Set of open source projects

Map->Reduce and HDFS Architecture

Master

JobTracker

JobTracker keeps track of jobs being run

NameNode

NameNode keeps information on data location

Slave

TaskTracker

DataNode

Machine

Slave

TaskTracker

DataNode

Machine

Slave

TaskTracker

DataNode

Machine

The Eight Fallacies of Distributed Computing

1. The network is reliable
2. Latency is zero
3. Bandwidth is infinite
4. The network is secure
5. Topology doesn't change
6. There is one administrator
7. Transport cost is zero
8. The network is homogeneous

Source: Peter Deutsch

Exponential Growth of Computing

Logarithmic Plot

Calculations per Second per $1,000

10^{60}
10^{55}
10^{50}
10^{45}
10^{40}
10^{35}
10^{30}
10^{25} — *All Human Brains*
10^{20}
10^{15} — *One Human Brain*

By 2020s, computers have the same power as the human brain

One Mouse Brain
10^{10}
One Insect Brain
10^{5}
1
10^{-5}

1900 1920 1940 1960 1980 2000 2020 2040 2060 2080 2100

Year

Source : Ray Kurzweil

Deep Learning

What is Deep Learning
✓ Iterative Algorithm
✓ Learning at different levels of abstraction
✓ Non-linear transforms
✓ Typically neural nets

Examples of Iterative Algorithm
✓ Genetic programming
✓ Neural networks
✓ Quantum computers
✓ Wisdom of Crowds

Google First Quantum Computer

"We actually think quantum machine learning may provide the most creative problem-solving process under the known laws of physics." – Google Blog

"In the future, every decision that mankind makes is going to be informed by a cognitive system like Watson."

Ginni Rometty, CEO of IBM

Deep Learning Application Areas
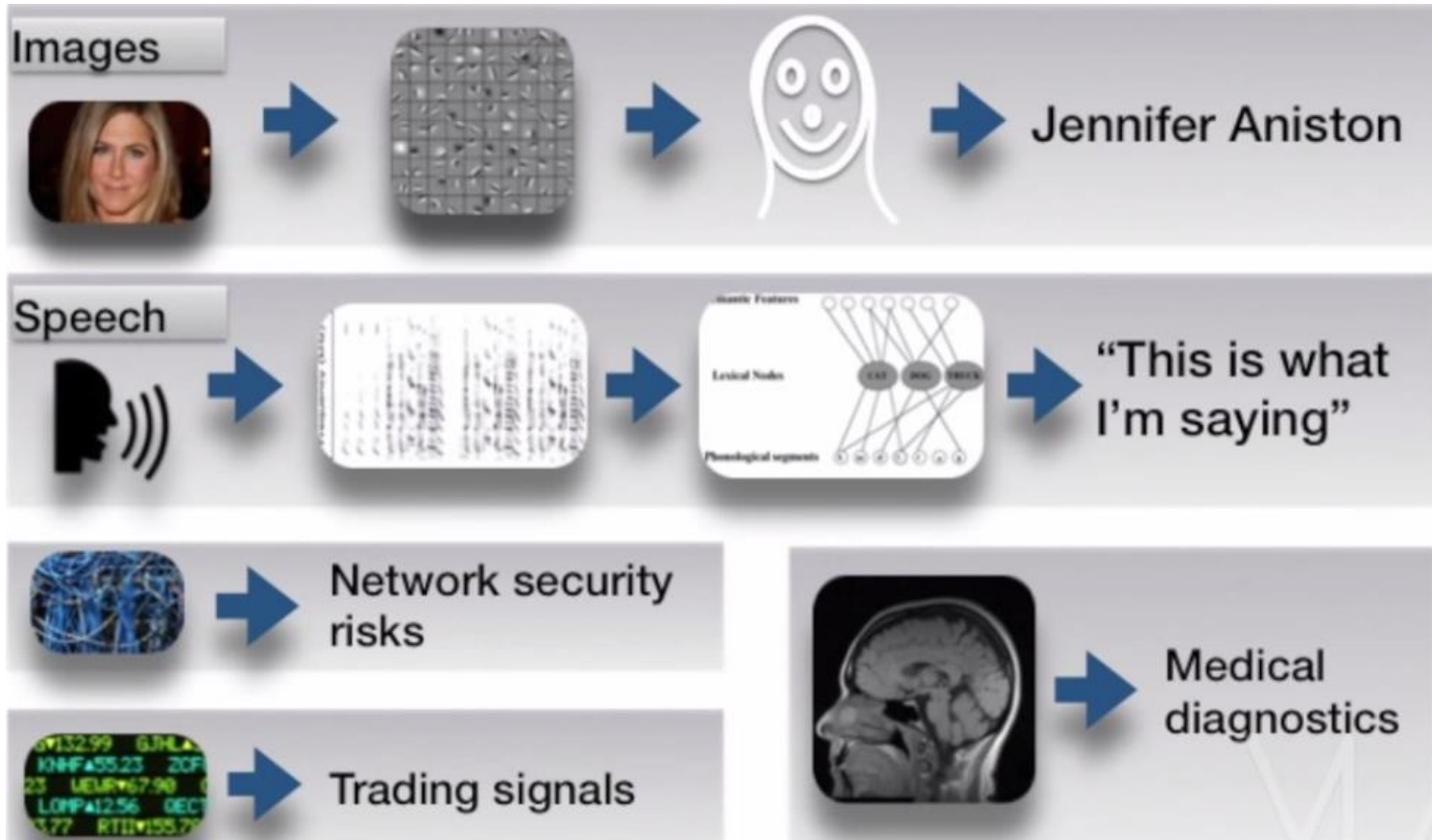
Images → → → Jennifer Aniston

Speech → → → "This is what I'm saying"

→ Network security risks

→ Trading signals

→ Medical diagnostics

1	*Big Data Technologies*
2	*Big Data Market*
3	*Opportunities, Risks, & Capital Trends*
4	*Algorithmic Accountability & Privacy*

Rainmaker I
Prophet

Eric Schmidt

Google
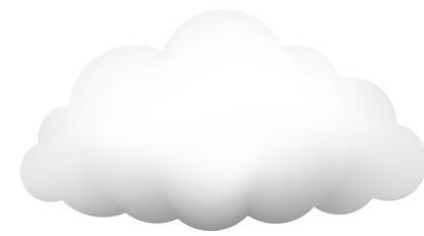
> "we now uncover as much data in 48 hours – 1.8 zettabytes – as humans gathered from the dawn of civilization to the year 2003"

Viktor Mayer-Schönberger

BELFER CENTER
for Science and International Affairs

UNIVERSITY·OF·OXFORD

BIG DATA

"No other book offers such an accessible and balanced tour of the many benefits and downsides of our continuing infatuation with data."

Viktor Mayer-Schönberger
Kenneth Cukier

LEARNING WITH BIG DATA
THE FUTURE OF EDUCATION
VIKTOR MAYER-SCHÖNBERGER
KENNETH CUKIER
authors of BIG DATA
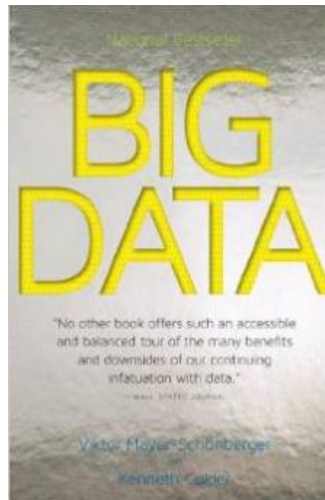
Rainmaker II
Knowledge Marketer

WALL STREET JOURNAL

HBR

MIT Technology Review

MITSloan

Gartner

Forbes

The Economist

McKinsey&Company

Big Data Hype Circle

visibility

Basic Education – Drive the Bandwagon
Reality Check – Manage Expectations
Customer Success Stories – Real Value

Tip o' the hat to H&K's Josh Reynolds for collaborating on how to approach the Hype Cycle from AR and marketing points-of-view.

Technology Trigger | Peak of Inflated Expectations | Trough of Disillusionment | Slope of Enlightenment | Plateau of Productivity

time

Years to mainstream adoption:
○ less than 2 years   ○ 2 to 5 years   ● 5 to 10 years   △ more than 10 years   ⊗ obsolete before plateau

Big Data Maturity Model

ADAPTIVE
ADOPTIVE
DEFINED
EXPLORATORY
UNAWARE

BUSINESS VALUE
STRATEGIC VALUE GAP
CAPABILITIES

Hype
Create Perception with Correlation

* correlation doesn't prove causation

IDC: 40% CAGR in digital information

Total info (EB)
Total storage (EB)

Exabyte Age
Transactions

Zettabyte Age
Interactions
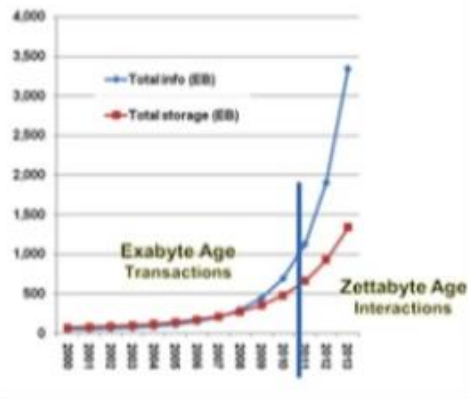
Twitter: Now over 50M tweets per day

34,722 tweets/min
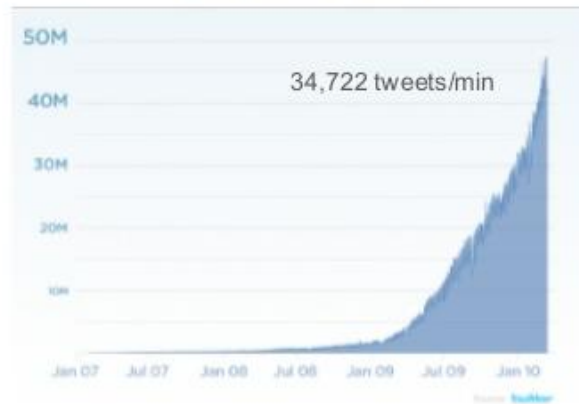
NCBI GenBank: Protein sequences double every 18 months
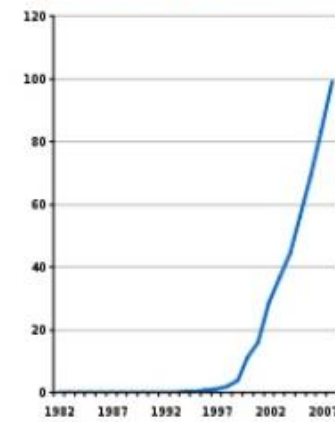
source: Google

... with new data sources and data types

Devices
phones, vehicles, set-top boxes, power meters, RFID, GPS

Statistics
open government / public transparency

Social
interactions, behaviors, user generated content

Enterprise
web analytics, customer database marketing, business operations

Formulate Illusion
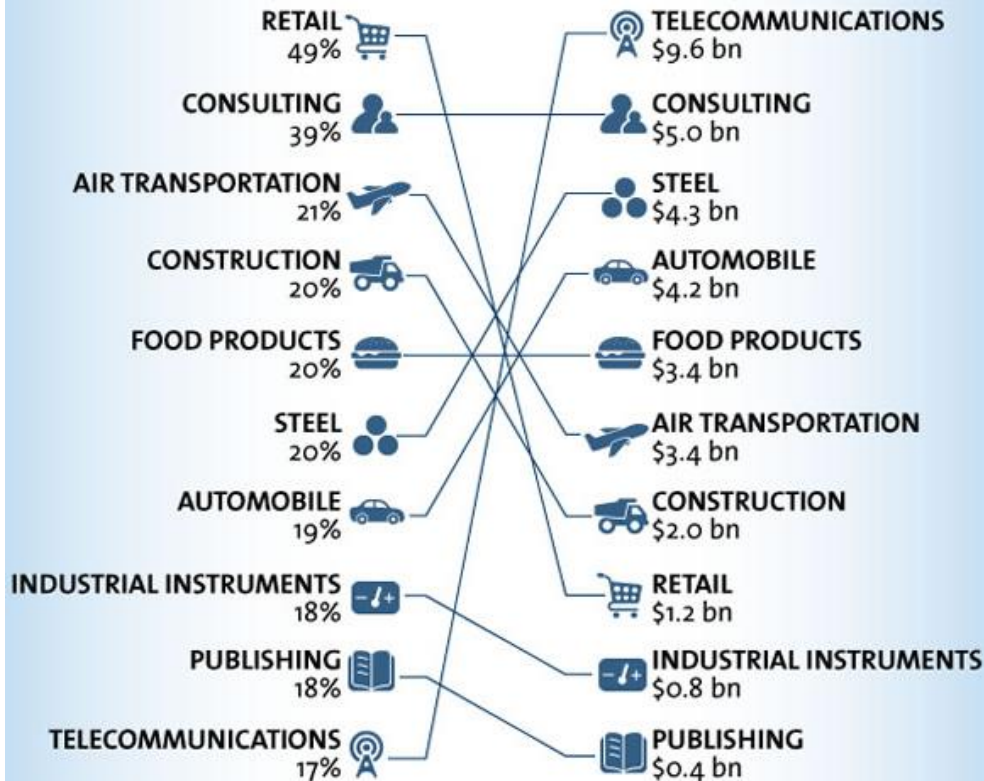
source: Sybase

Umbrella
Business Marketer

Vertical Apps
- PREDICTIVE POLICING
- bloomreach
- MYRRIX

Log Data Apps
- splunk
- loggly
- sumologic

Ad/Media Apps
- rocketfuel
- collective[i]
- bluefin
- Recorded Future
- LuckySort
- Media Science
- TURN
- DataXu

Data As A Service
- factual
- kaggle
- knoema
- GNIP
- DATASIFT
- Windows Azure
- INRIX
- LexisNexis
- LOQATE

Business Intelligence
- ORACLE | Hyperion
- SAP Business Objects
- RJMetrics
- Microsoft Business Intelligence
- IBM COGNOS
- birst
- MicroStrategy
- Autonomy
- QlikView
- bime
- DOMO
- Chart.io
- GoodData

Analytics and Visualization
- tableau
- Palantir
- OPERA
- metaLayer
- dataspora
- METAMARKETS
- centrifuge
- TERADATA ASTER
- SAS
- TIBCO
- KARMASPHERE
- panopticon
- pentaho
- Datameer
- platfora
- ClearStory
- CIRRO
- alteryx
- visual.ly
- AYATA

Analytics Infrastructure
- Hortonworks
- VERTICA
- MAPR
- cloudera
- INFOBRIGHT
- ParAccel
- EMC²
- GREENPLUM
- NETEZZA
- kognitio
- DATASTAX
- EXASOL

Operational Infrastructure
- COUCHBASE
- 10gen
- TERADATA
- HADAPT
- TERRACOTTA
- VoltDB
- MarkLogic
- INFORMATICA

Infrastructure As A Service
- amazon web services
- Windows Azure
- infochimps

Structured Databases
- ORACLE
- MySQL
- SQL Server
- PostgreSQL
- IBM DB2
- SYBASE
- memsql

Technologies
- hadoop
- hadoop MapReduce
- mahout
- APACHE HBASE
- Cassandra

Problem
Whose Problem ?

TBytes

Can't process all of this data on one machine
→ Distribute the processing to many machines

Chart axis: 250, 200, 150, 100, 50, 0

You | Facebook | Yahoo! Groups | German Climate Computing Centre

YOU ARE GOOD
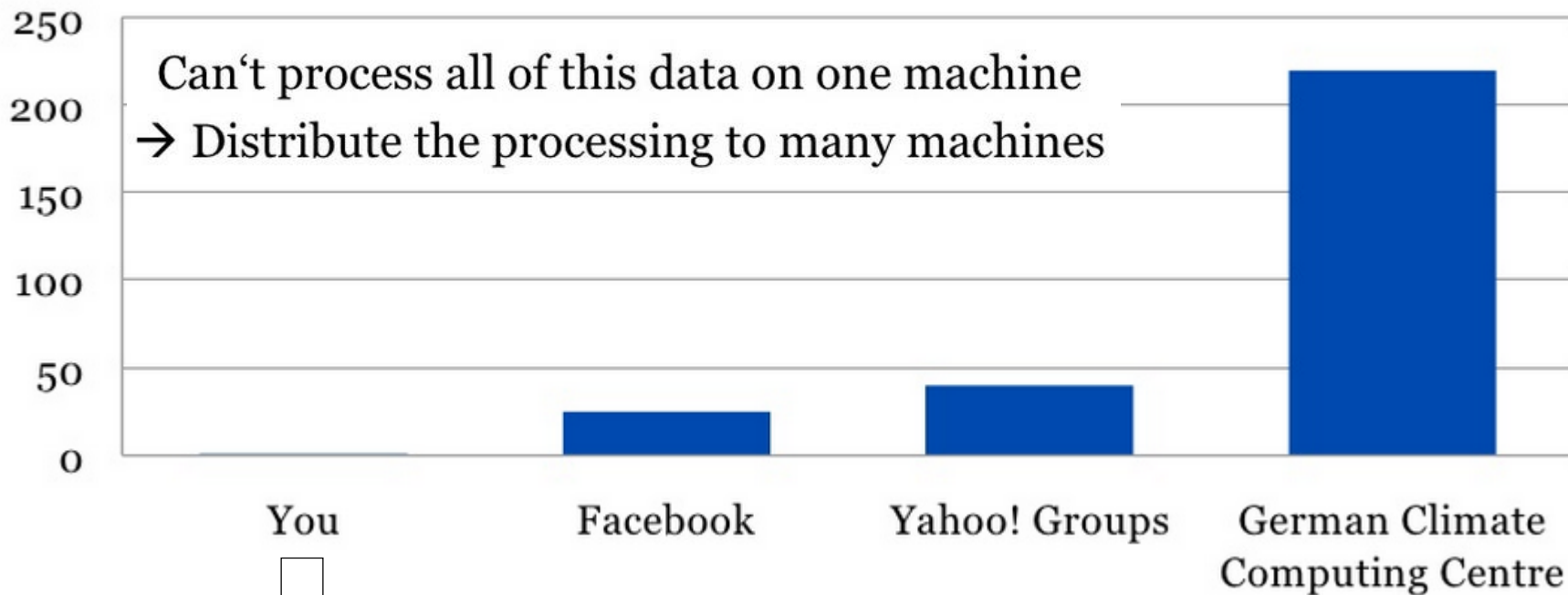
Questions :
1 You want to solve IT giants' (Google/FB) problems ?
2 You want to solve future problems with today's technologies and price ?
3 Forging illusive needs immediately to leverage technology trends ?

Avoid Fallacy of Irrelevancy

"Excel is very powerful. The fact is that programmers generally don't realize this." (Jay, LinkedIn)

1	*Big Data Technologies*
2	*Big Data Market*
3	*Opportunities, Risks,& Capital Trends*
4	*Algorithmic Accountability & Privacy*

Big Data Business Model

Information based differentiation	Information based brokering	Information based delivery networks
Create new service offerings	Sell raw information	Foster marketplaces
Satisfy customers	Provide benchmarking	Drive deal making
Provide contextual relevance	Deliver analysis and insights	Enable advertising
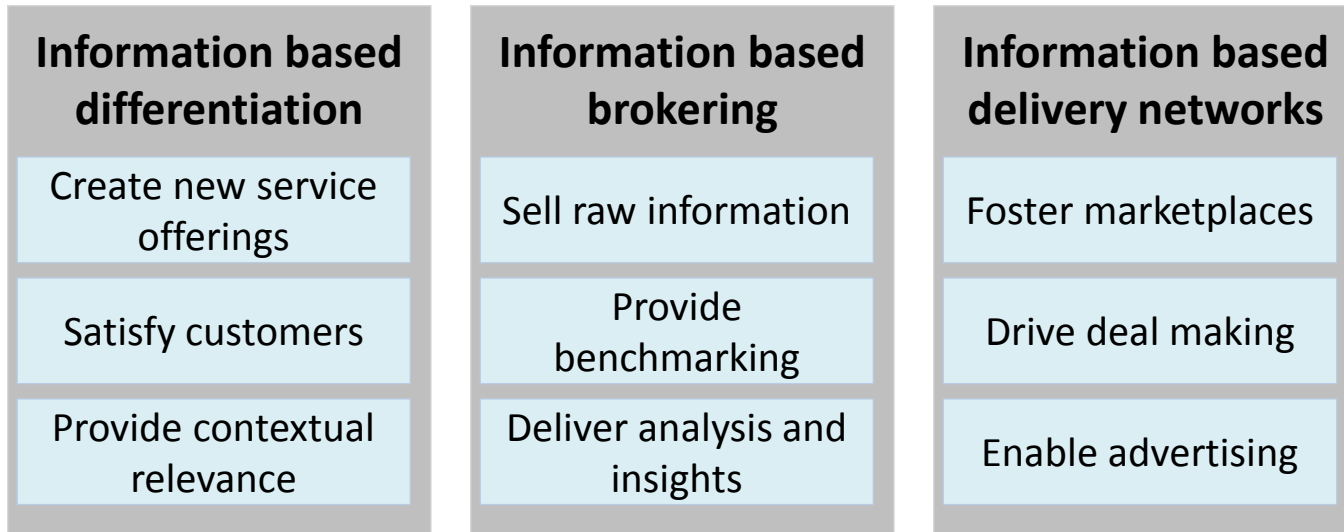
(HBR, 2012)

Questions Addressed by Data Analytics

	Past	Present	Future
Information	What happened ? (Reporting)	What is happening ? (Alert)	What will happen ? (Extrapolation)
Insight	How and why did it happen? (Modeling experimental design)	What's the next best action? (recommendation)	What's the best/worst than can happen? (prediction, optimization)

(Harris & Morrison)

Case Studies

Target used data mining to predict buying habits of customer going through major life events
- ✓ Target was able to identify 25 products that when analyzed together helped determine a "pregnancy prediction" score
- ✓ Sent baby-related promotions to women based score

Outcome
- ✓ Sales of Target's Mom and Baby products sharply increased soon after advertising campaigns
- ✓ Privacy concerns: Target had to adjust how it communicated the new promotions

General Electric using Big Data to optimize the service contracts & maintenance.

Netflix used Big Data to predict if a TV show will be successful – "House of Cards" series, Director & promotion.

LinkedIn used Big Data to develop "People You May Know" products - 30% higher click-thru-rates

Gartner Hype Cycle
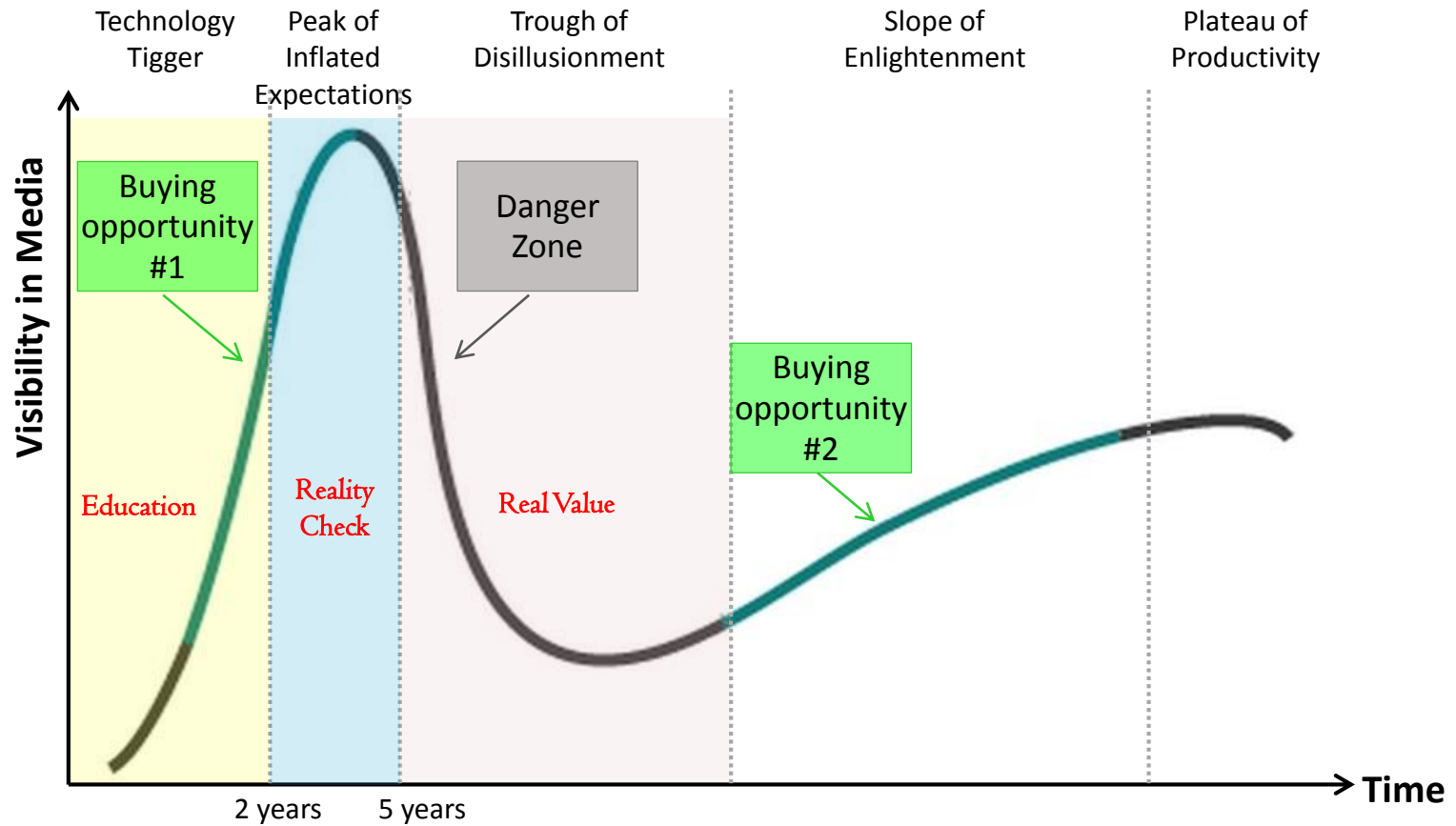
Source : Gartner; Dr. Kenny Huang Revised

Hype Cycle and Technology Adoption Cycle Plotted Together

Visibility in Media

Buying opportunity #1

Danger Zone

Buying opportunity #2

Technology Tigger

Peak of Inflated Expectations

Trough of Disillusionment

Slope of Enlightenment

Plateau of Productivity

Time

Chasm

Innovators
2.5%

Early Adapters
13.5%

Early Majority
34%

Late Majority
34%

Laggards
16%

2 years

5 years

Time

Source : Dr. Kenny Huang Revised
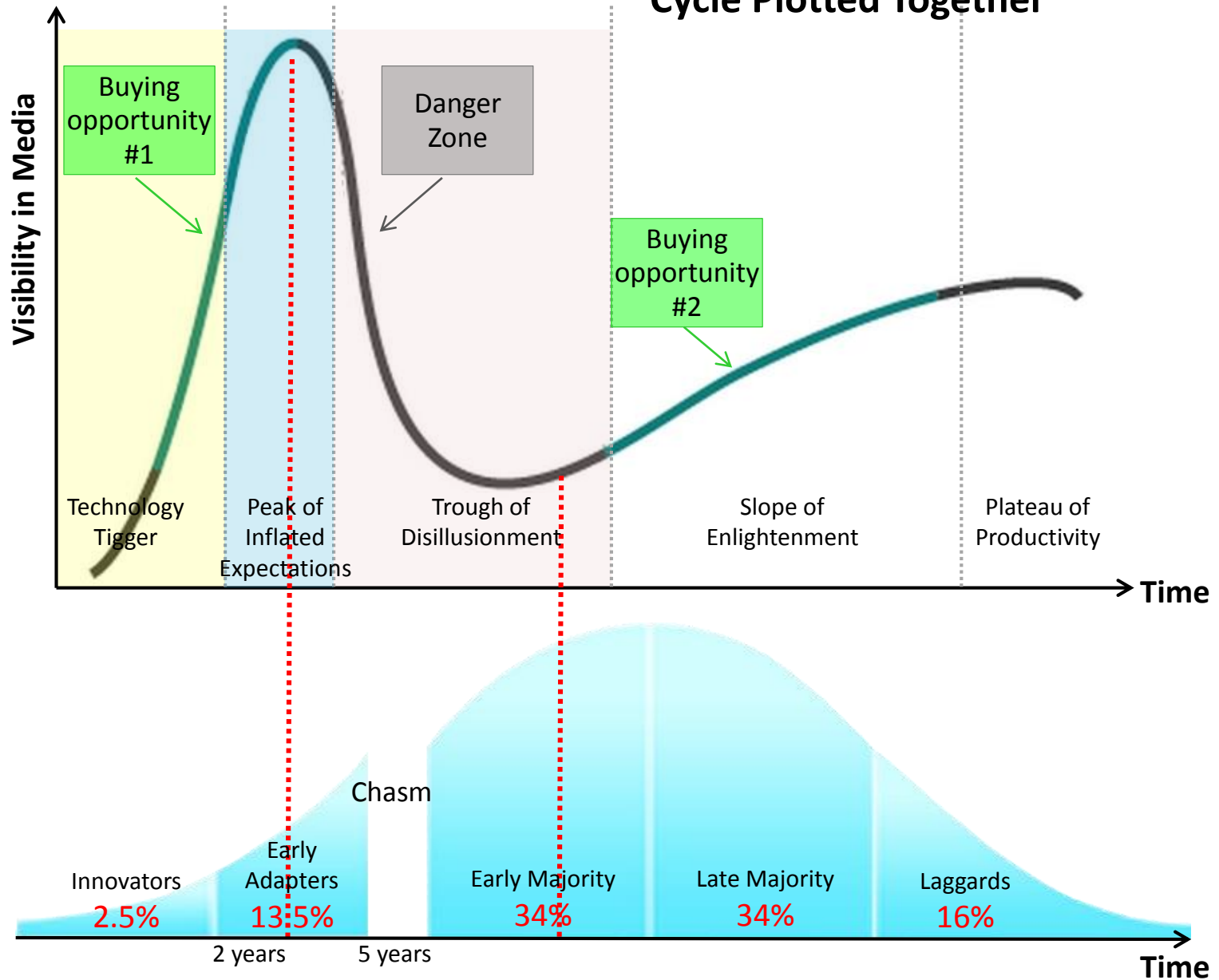
Big Data Visibility and Demand

"Big Data" Google Trends @2015.06.04

US | **TW**

2011 | 2013 | 2015

2015 Gartner research on adoption of Hadoop Technology

26% piloting
11% may invest in 1 year
7% may invest in 2 years

"Future demand for Hadoop looks fairly anemic over at least
 the next 24 months". Merv Adrian, Gartner Research. (2015)

Big Data Buying Opportunity for Taiwan

Big Data Visibility as of June 2015

Next Buying Opportunity

Danger Zone

Visibility in Media

Time

Source : Gartner; Dr. Kenny Huang Revised

2017 < Time *
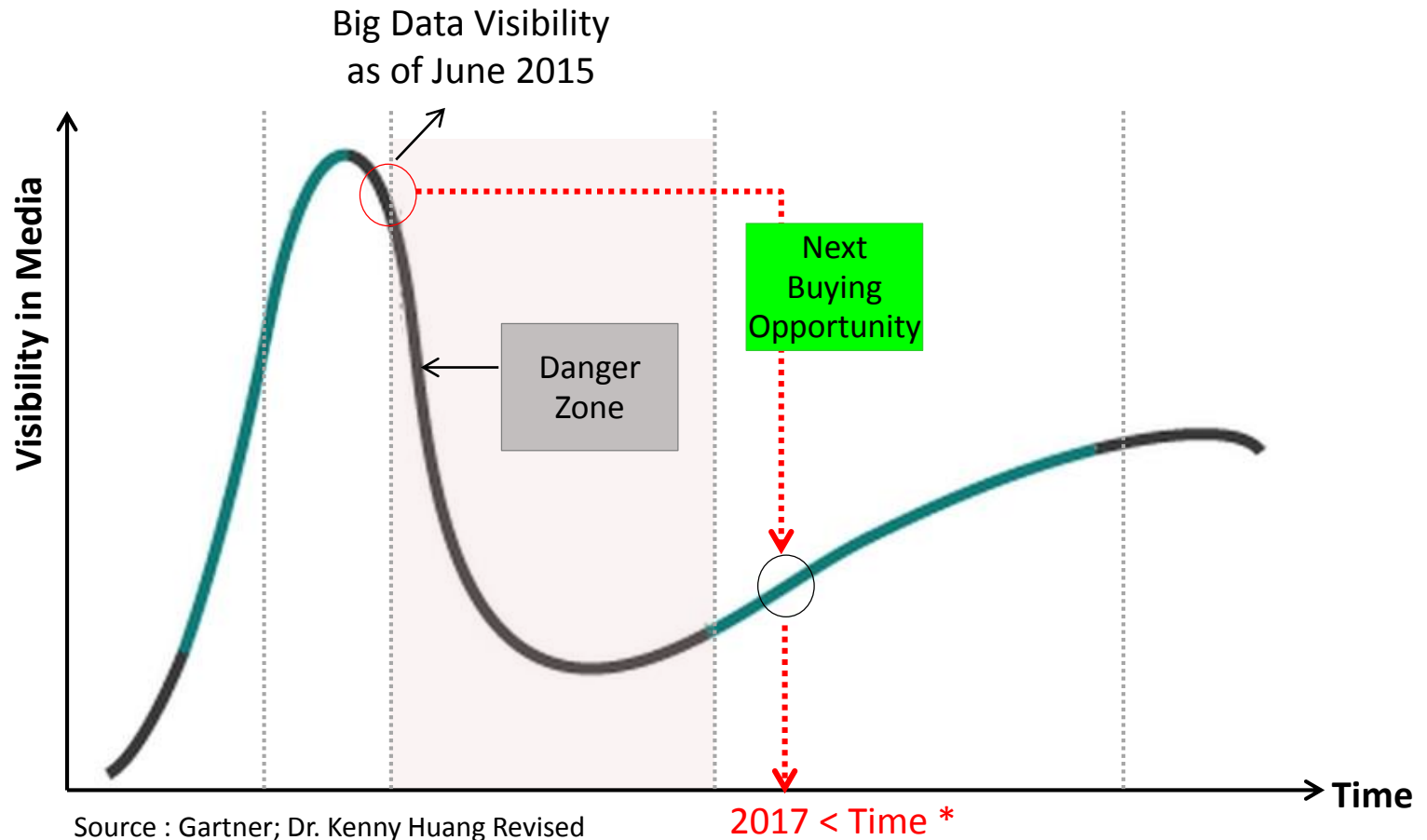
* Ref revised Hype cycle diagram, Google trends 2015, Gartner research 2015

Risk-Return Tradeoff
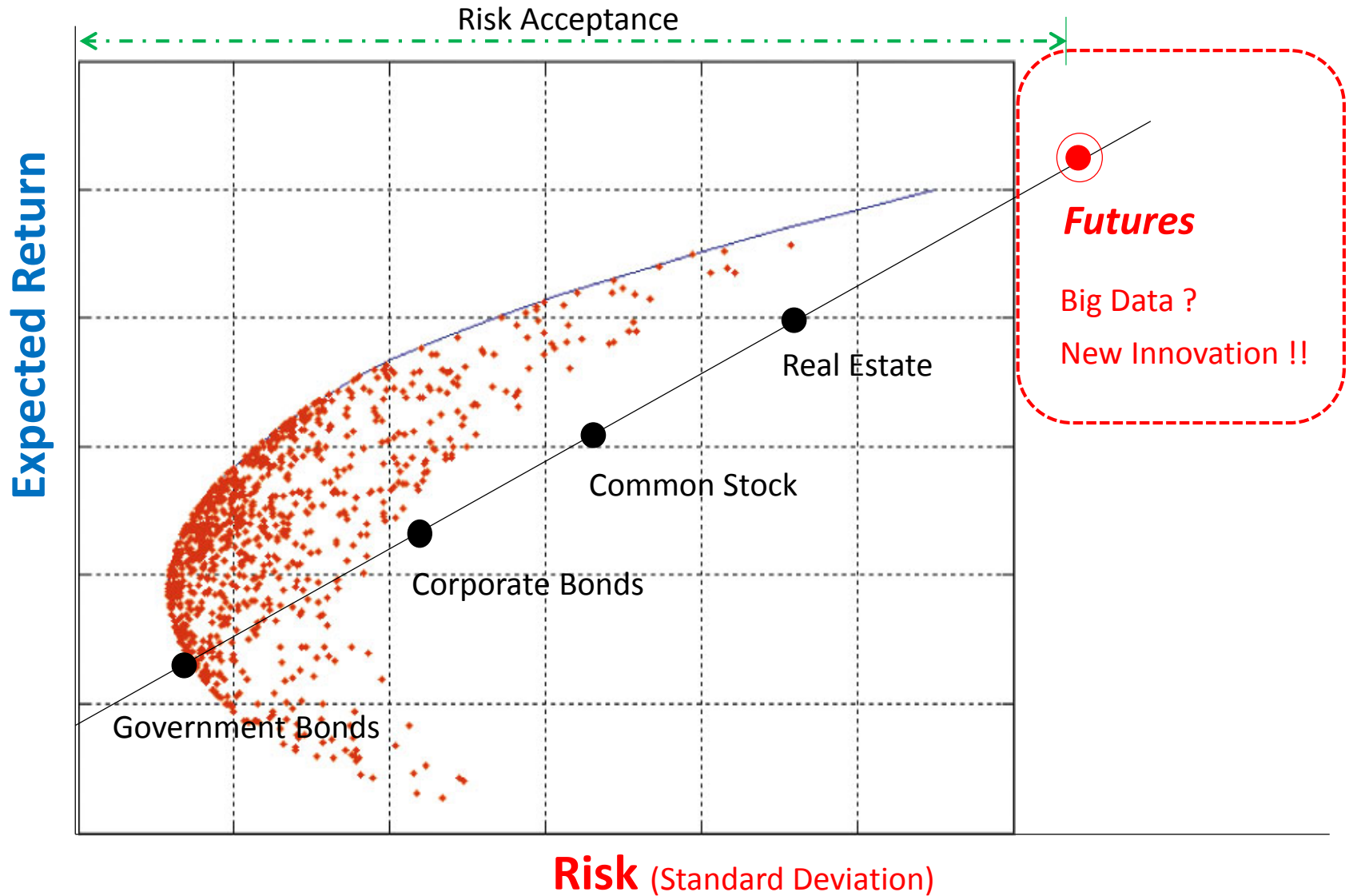
Source : Dr. Kenny Huang Illustration

Investment Risking Model

Business Entity

Risk Acceptance	Startup; Series A
Risk Mitigation	Due Diligence
Risk Avoidance	Change Investment Objects

Government Institution

Risk Acceptance	**Don't Use Taxpayers' Money**
Risk Mitigation	Pilot Projects; Research
Risk Avoidance	Change Technology Policy

Big Data Adoption Strategy

Ability To Response

High
- Focus on your own business

- Adopt and separate; or
- Adopt keep internal; or
- Attack back and disrupt the disruption

Low
- Focus on your own business

- Attack back and disrupt the disruption; or
- Embrace the innovation and scale it up

Low   High

Motivation To Response
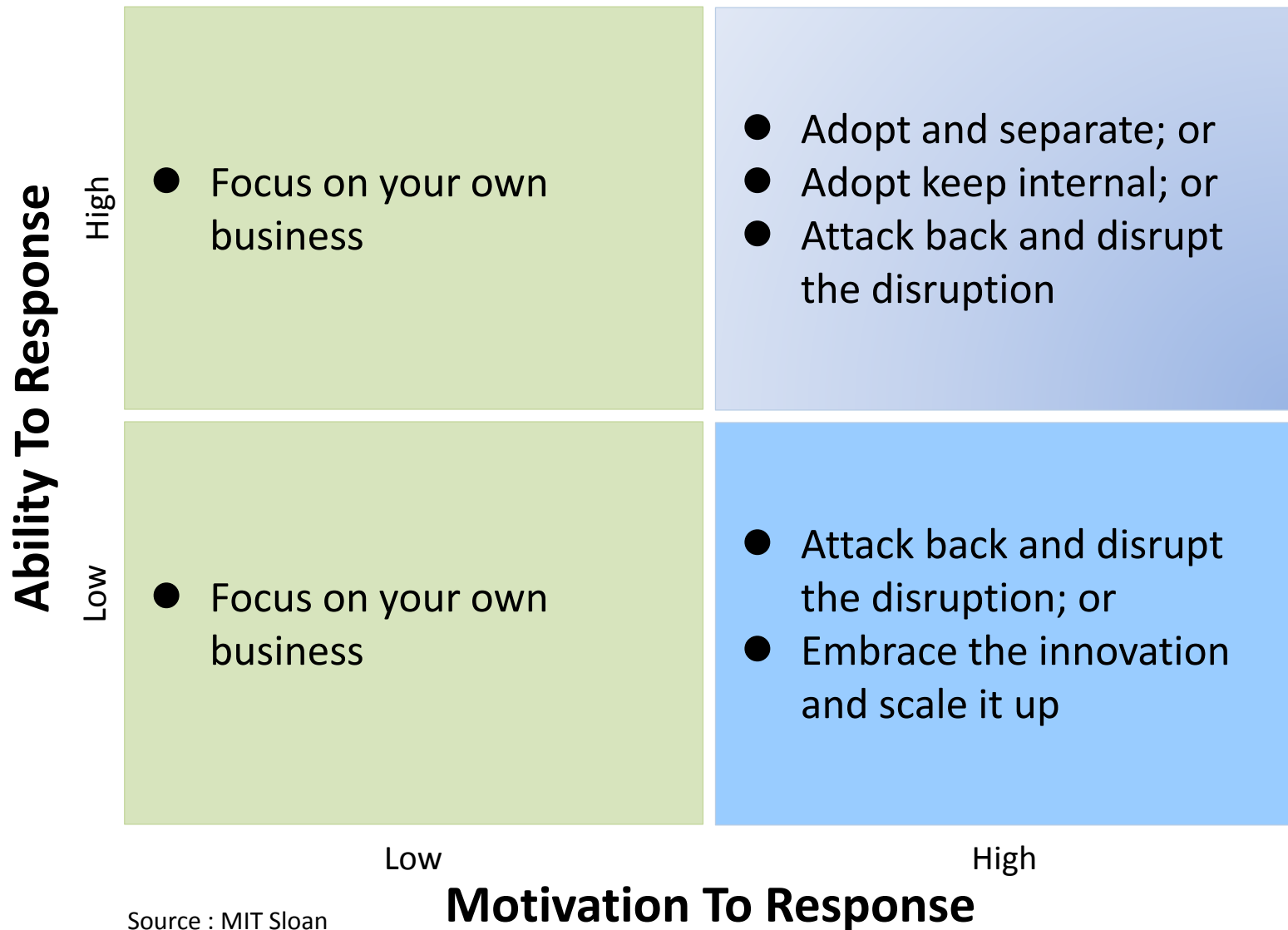
Source : MIT Sloan

Financial Model Quizzes

A

B

C

[   ]
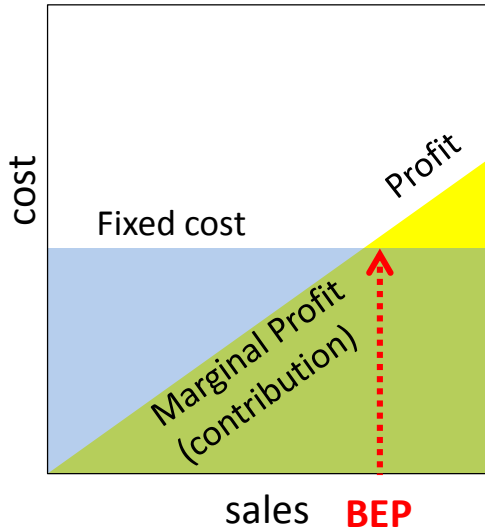
Big Data
Technology
Provider

[   ]
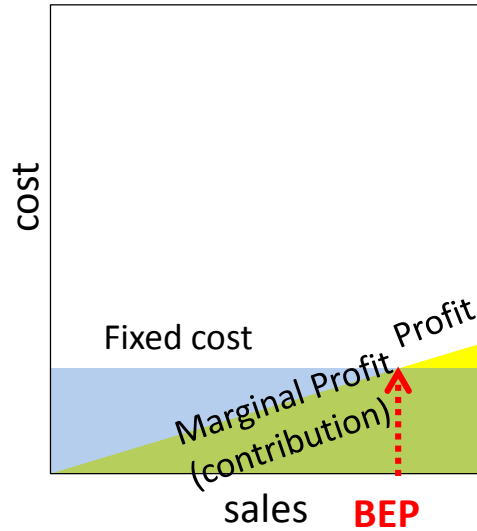
Big Data
As A Service

[   ]
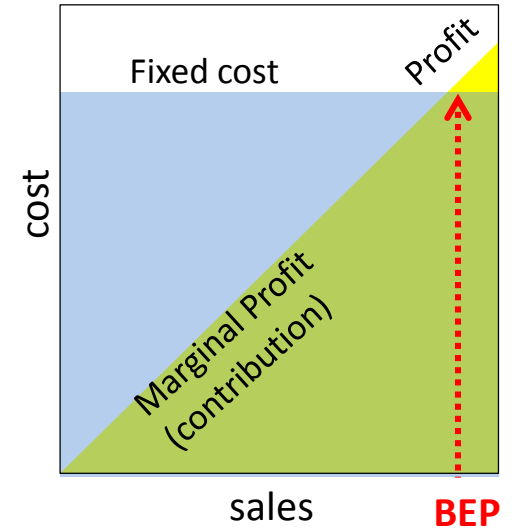
Big Data
Solution
Integration

*BEP : Breakeven Point

Global Capital Market Trends

MM USD

外國直接投資流入金額 (百萬美元)

1239 億美元/中國

746 億美元/香港

- CN
- HK
- TW
- VN
- KR

122 億美元/韓國

89 億美元/越南

37 億美元/台灣
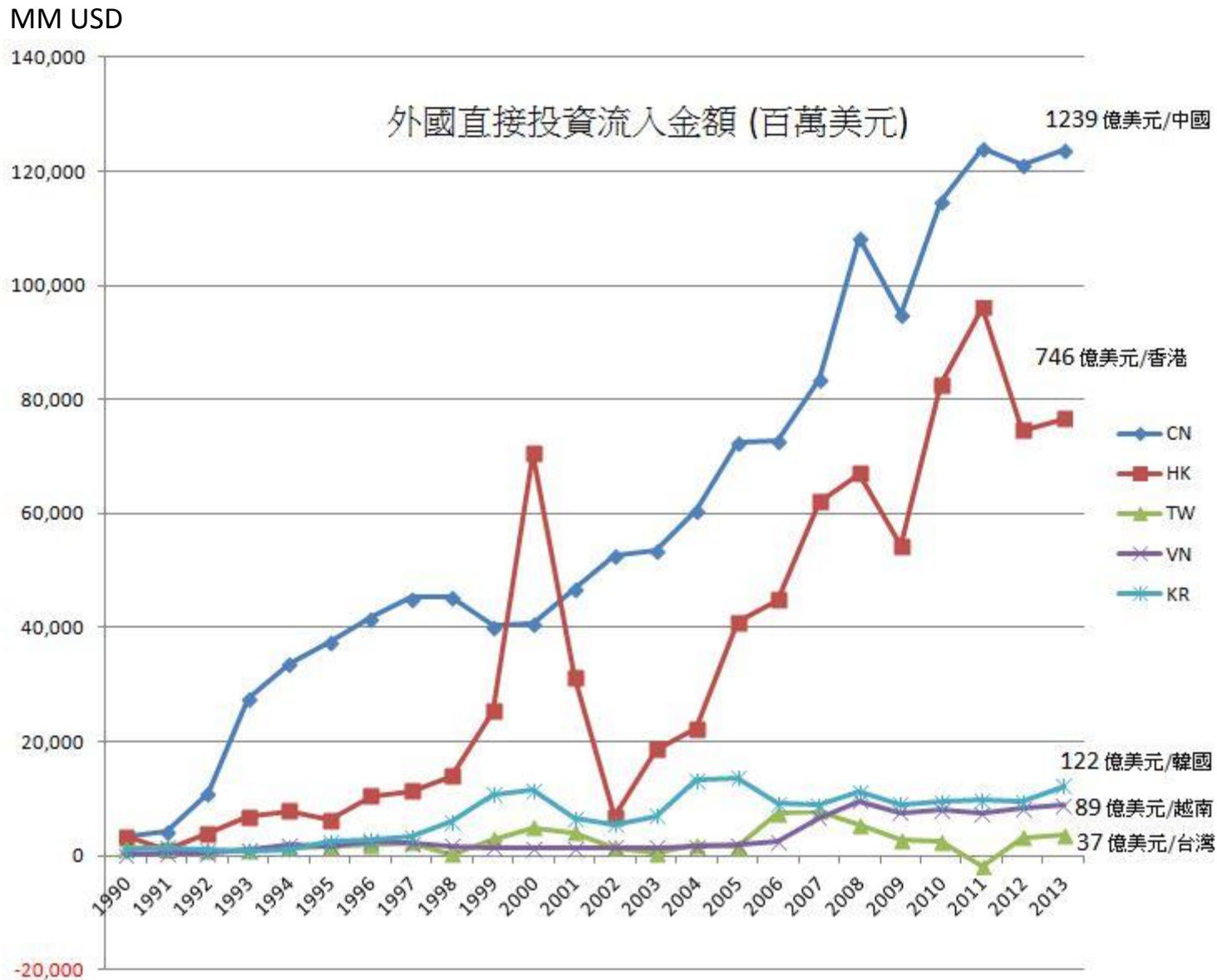
Source : Designed by Dr. Kenny Huang

Game Rule : You Pick The Valuation, I Pick The Terms
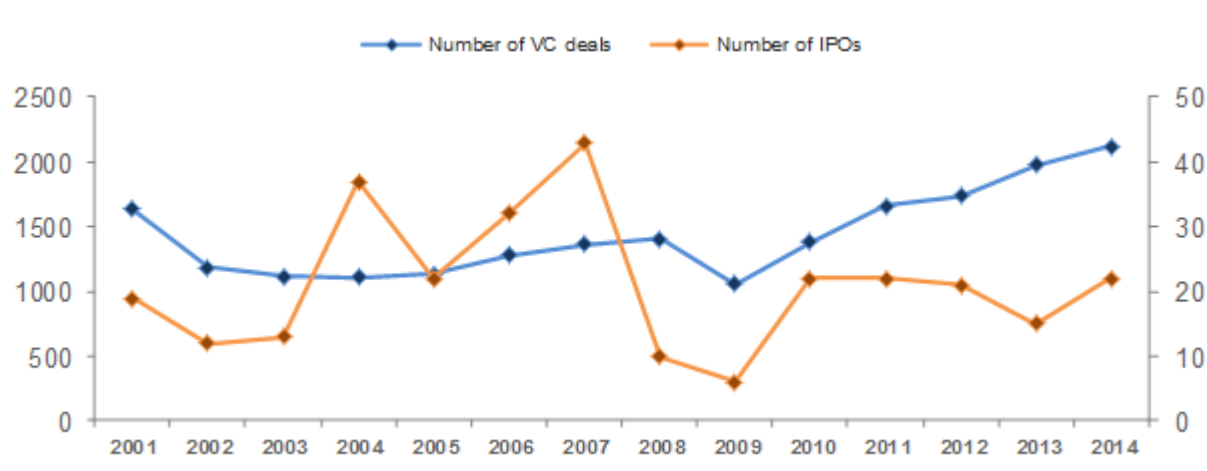
Source: PwC

IPOs and Private Financing Deals in the Tech Sector since 2000 (United States)

Company	Round	Investment ($MM)	Stake Diluted	Implied Valuation ($MM)	Liquidation Preference	Participation Rights
Uber	Series E	1,200	3%	44,271	1x	-
AirBnB	Series D	475	5%	10,521	1x	-
Dropbox	Series C	325	3%	9,379	1x	-
Square	Series E	150	3%	5,265	1x	-
Pinterest	Series F	160	4%	4,558	1x	-
Cloudera	Series F	370	9%	4,288	1x	-
DocuSign	Series E	30	2%	1,583	1x	1.5x
Lookout	Series F	87	6%	1,381	1x	-
Jasper	Series F	50	4%	1,313	1x	-
AppNexus	Series E	60	5%	1,194	1x	-
Good Technology	Series C	101	9%	1,130	1x	2.5x

Source: Techcrunch

If there is a bubble, investors would recover their investment and perhaps walk away with positive return, the biggest losers for sure would be the employees and founders.

WHERE BIG DATA INVESTMENT GOES
Real-time data takes a backseat to historical data.

VELOCITY OF DATA

COST SAVINGS AND EFFICIENCY · · · · · · · · · · 8

3

VOLUME OF · · · · · · · · · · 16
DATA

47% · · · · · · · · · · PROCESS
IMPROVEMENT

VARIETY OF · · · · · · · · · · 26
DATA
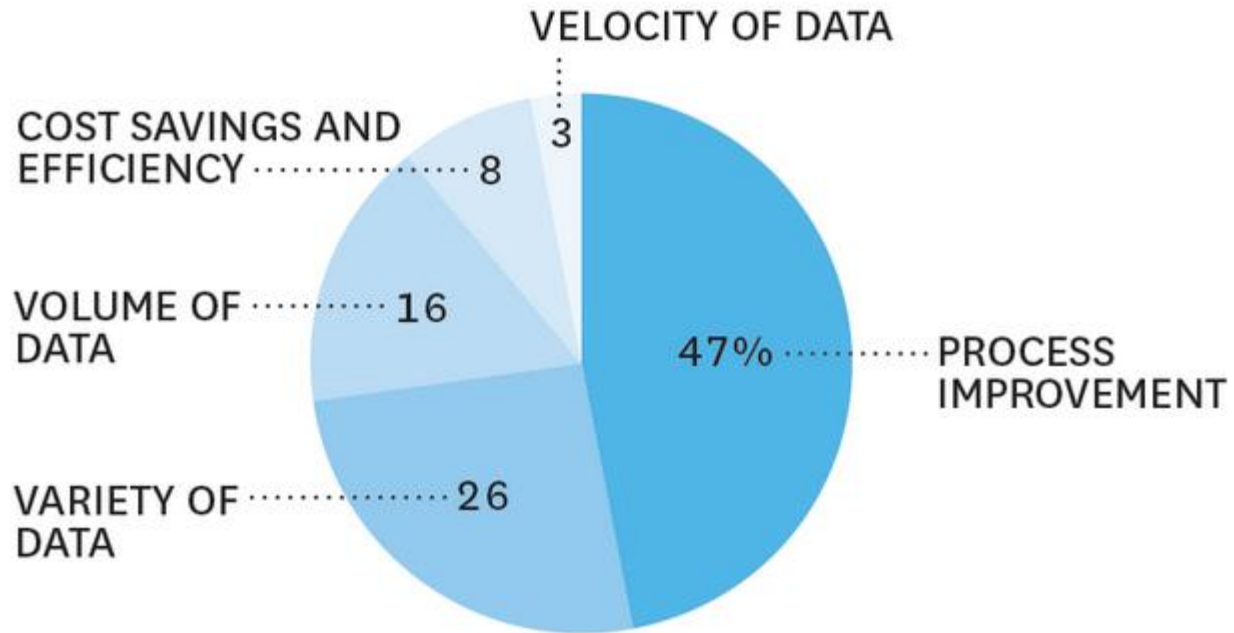
SOURCE NEWVANTAGE PARTNERS 2014 BIG DATA
EXECUTIVE SURVEY

HBR.ORG

47

1	*Big Data Technologies*
2	*Big Data Market*
3	*Opportunities, Risks,& Capital Trends*
4	*Algorithmic Accountability & Privacy*

Algorithms Rule The World

> We should interrogate the architecture of cyberspace as we interrogate the code of Congress.
>
> - Lawrence Lessig, Code is Law, 2000

BBC

Technology

When algorithms control the world

theguardian
Winner of the Pulitzer prize 2014

Mathematics

How algorithms rule the world

CHRISTOPHER STEINER
HOW ALGORITHMS CAME
TO RULE OUR WORLD

WIRED
ALGORITHMS TAKE CONTROL
OF WALL STREET

Algorithmic Accountability

Algorithms Are Everywhere

Political messaging
Romantic matching
Policing   Advertising systems
Predictive bots   Welfare management
Credit scores Teacher scoring Copyright infringement detection
Insider tradingiPhone censorship Recommendation engines
IRS audit flags   Medicare management
RedistrictingStudent matchingGovernment Censorship
Twitter trends   News personalization
Yelp reviews College rankings
TSA   Amazon pricing

Algorithmic Confusing

✓ Algorithms are not transparent
✓ Technical complexity is a barrier

Algorithmic Accountability

✓ How can we characterize the bias or power of an algorithm?
✓ When might algorithms be wronging us, or making consequential decisions?
✓ What role should be involved in holding algorithmic power to account ?

(Nick Diakopoulos)

Algorithmic Power : Decisions

Prioritization (3) (2) (1)

Classification

Association →

Filtering
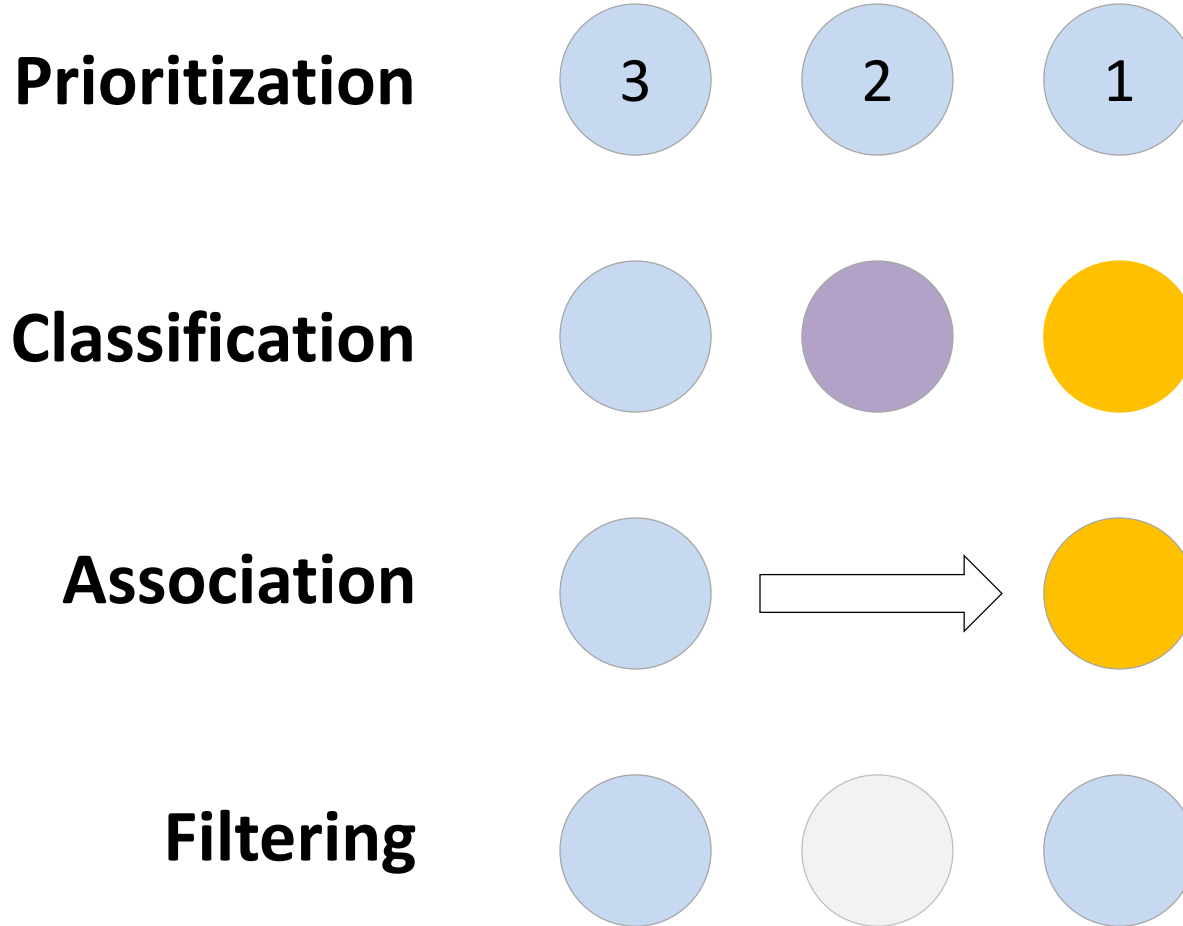
Input / Output of An Algorithm

Input → **Algorithm** → **Output**

WSJ Price Discrimination

Do different people pay different prices depending on their geography or browser history ? Yes

Staples.com

Input → **Algorithm** → **Output**

Source: WSJ, Dec 2012

Likelihood of receiving higher prices, by ZIP code
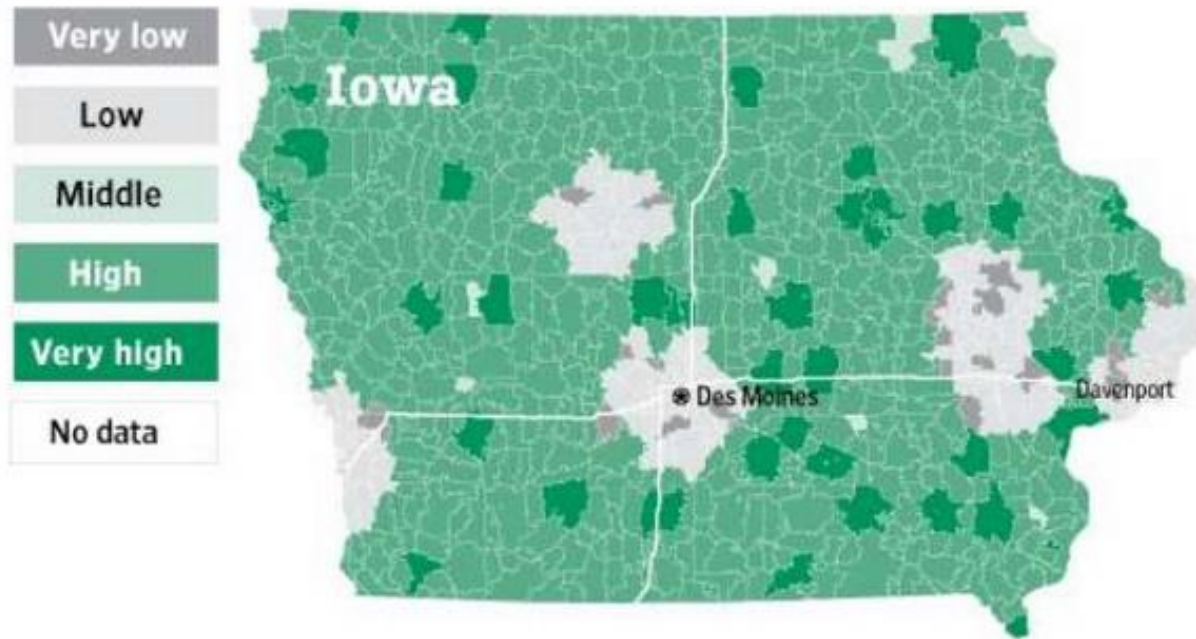
Transparency

✓ Voluntary incentives for self-disclosure about algorithms

✓ Trade secrets

✓ Gaming / manipulation
  ➢ Goodhart's Law: " When a measure becomes a target, it ceases to be a good measure."

✓ Cognitive complexity
  ➢ Transparency information needs to be accessible and understandable

Other Stories from Algorithms

- ✓ Discriminatory / Unfair
- ✓ Mistake that denies a service
- ✓ Censorship
- ✓ Breaks law or social norm
- ✓ False Prediction

Next Step

- ✓ Teaching algorithmic accountability
    - ✓ It will be messy and hard
- ✓ Legal issues
    - ✓ Computer Fraud and Abuse Act
- ✓ Ethical implications of publishing more information
- ✓ Transparency policy
    - ✓ What factors to expose, frequency, format of disclosure

Critical Considerations for Big Data Practices

Privacy	Transparency	Ownership
Customers will want to know ➢that you are collecting data ➢why and what you are collecting ➢that their confidentiality is preserved ➢that their data is accessible	Customers will want ➢an unique URL where they can see what you've collected ➢to know what sensors you are using ➢that an API is interrogating the data	Customers will expect ➢to be the owner of the data & be the copyright holder. ➢To decide who they allow access to (might not even be you)

Concern with Big Data Practices
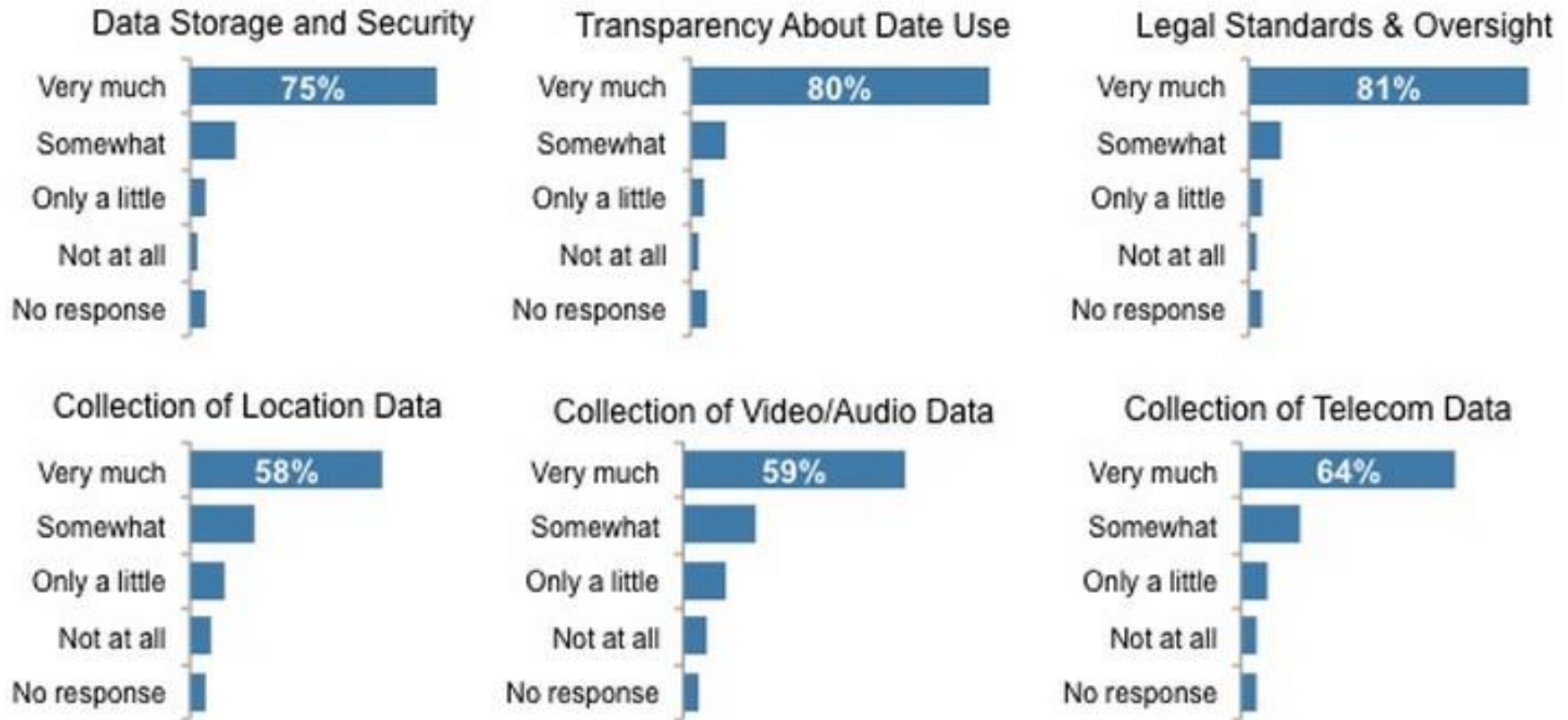
Source : Whitehouse Big Data Review

PRISM Tasking Process
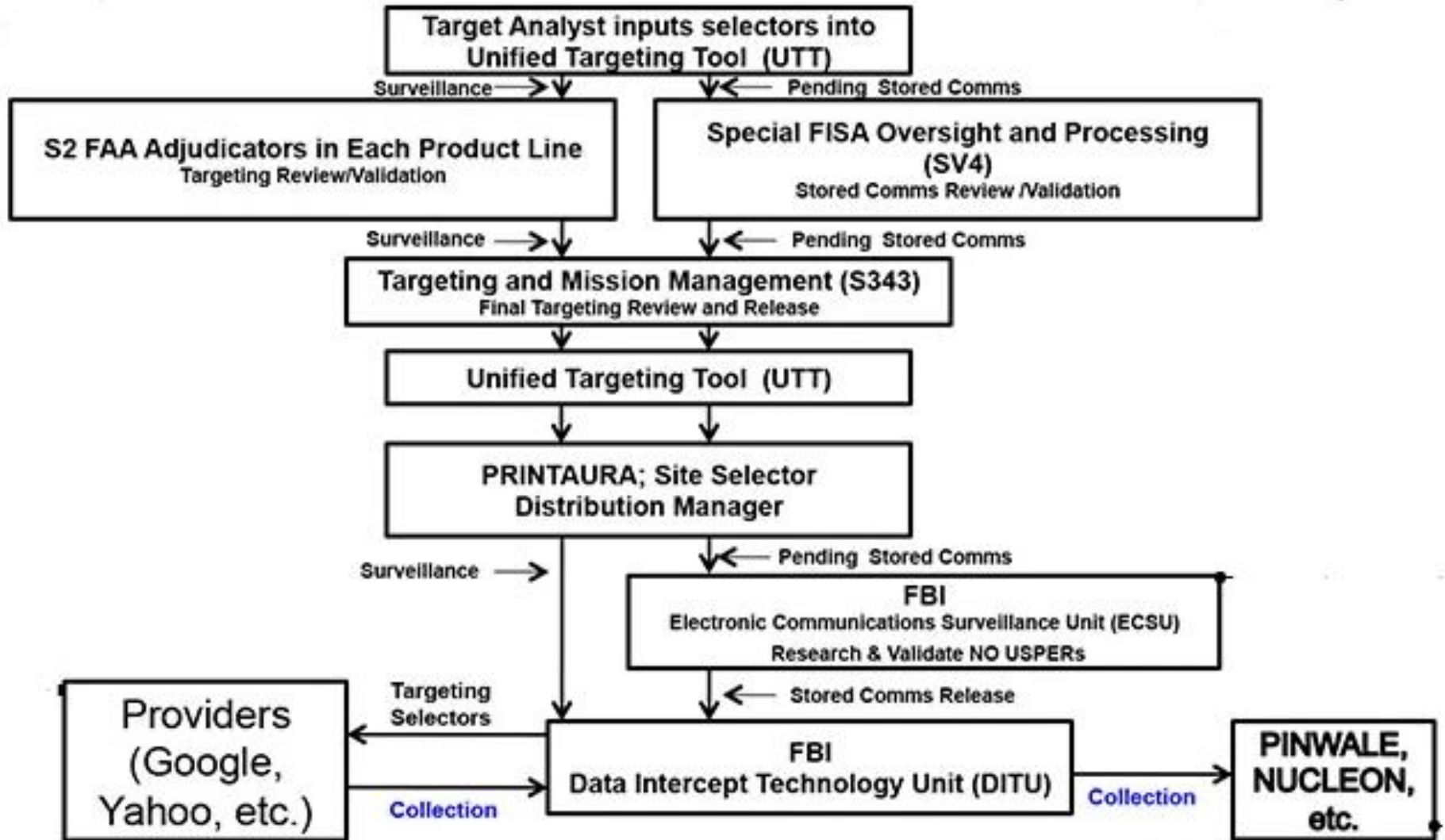
Massive Surveillance vs. Human Rights

Current Providers

- Microsoft (Hotmail, etc.)
- Google
- Yahoo!
- Facebook
- PalTalk
- YouTube
- Skype
- AOL
- Apple

What Will You Receive in Collection (Surveillance and Stored Comms)? It varies by provider. In general:

- E-mail
- Chat – video, voice
- Videos
- Photos
- Stored data
- VoIP
- File transfers
- Video Conferencing
- Notifications of target activity – logins, etc.
- Online Social Networking details
- **Special Requests**

The Universal Declaration of Human Rights

Article 12:
No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation.
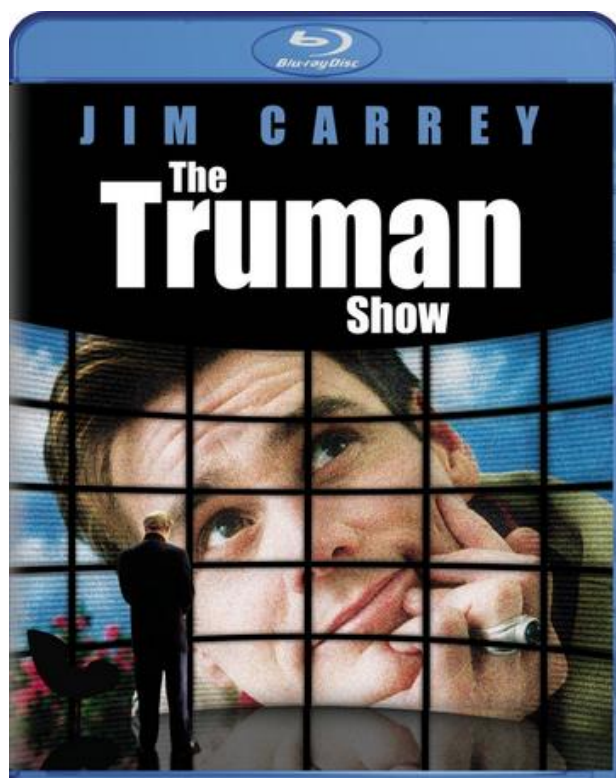
Source : The Truman Show

議員反監視器抓違停 柯：腦袋裝大便

罵死守法條者 民反酸市府搶錢

2015年04月25日　　f 讚 2,744　　g+1 2　　Pin it

台北市長柯文哲要求交通局研議，利用監視器取締違規停車。 杭大鵬攝

【綜合報導】為遏阻違規停車，台北市長柯文哲拋出以監視器抓違停構想，但遭議員和里長批評搶錢及違法，柯昨主持交通會報堅持己見，痛罵死守法條者是「腦袋裝大便」。市議員昨再批，北市靠1.4萬支監視器治市，形同把人民當犯人監看、沒人權也侵犯隱私、擾民。有民眾支持嚴抓違停，也有民眾批市府有那麼缺錢嗎？

Source : appledaily, 2015.04.25

USA Today
2015.06.03

The New York Times
2015.06.03

網路治理領域與參與機構建議
黃勝雄博士 整理 2015.05.02

功能領域	項目	主要參與機構
1.Control of "Critical Internet Resources" 網路關鍵資源管制	網路名稱與 IP 位址監理機構	ICANN, IANA, US DoC; 交通部; 通傳會
	IP 位址技術規範	IETF;
	頂級網域名稱授權	ICANN; 交通部; 通傳會
	網域名稱代理註冊	授權代理註冊商
	根域檔案監理(Root Zone File)	US DoC/NTIA;
	IP 位址發放	IANA, APNIC, 交通部;通傳會; TWNIC; 網路服務商
	根域檔案管理(Root Zone File)	IANA;
	網路自治號碼 ASN 發放	IANA, APNIC, TWNIC
	根伺服器管理	Versign, Cogent, 及其他
	DNS 查詢解析	Versign, TWNIC; .taipei 註冊管理局
2.Setting Internet Standards 訂定網路標準	網路協定號碼分配	IANA
	網路技術標準	IETF
	網頁技術標準	W3C
	其他通信標準	ITU, IEEE, MPEG, JPEG, ISO,通傳會
3.Access and Interconnection coordination 網路互連與接取協調	多方互連協調	網路交換中心; 通傳會批發價格管制
	網路互連及轉接合約	網路服務商, 內容服務商
	網路互連標準(如. BGP)	IETF
	網路管理 (服務品質)	網路服務商
	用戶接取政策	網路服務商
	網路接取法規 (如網路中立 Net Neutrality)	通傳會
4.Cybersecurity Governance 網路安全治理	維護網路基礎設施安全	網路服務商,機構網管部門
	加密標準	相關標準制訂機關, 通傳會, 國安局
	網路安全法規/執法	軟體法規, 多國合作協議
	修正軟體資安漏洞	軟體公司
	軟體更新管理	使用者
	維護路由安全, IP 位址安全, DNS 安全	網路服務商, IETF, APNIC, TWNIC
	資安通報	CERTs/CSIRTs, 技服中心
	網站信任憑證	Certificate Authorities (CAs)國家憑證管理中心
5.Information Intermediation 資訊中介	協助商業交易	電子商務網站,第三方金流機構
	協助政府內容審查與移除	搜尋引擎, 社群網路, 內容彙整網站
	App 調解 (規範, 執法)	智慧手機製造商 (如. Apple, HTC)
	隱私政策	社群網路, 廣告中介商, Email 服務商,網路服務商
	網路霸凌與誹謗	內容中介商
	隱私權法規	立法院
	政府協調, 個人資訊申請	內容中介商, 網路服務商
6.Architecture-Based Intellectual Property Rights Enforcements 智財權執法	網域名稱爭議處理	ICANN UDRP, TWNIC, .taipei 註冊管理局
	侵權內容移除	內容中介商
	演算法執法 (如搜尋排序)	搜尋引擎
	侵權使用者禁止接取	網路服務商; 經濟部智慧財產局
	DNS 侵權執法	TWNIC, .taipei 管理局, 代理註冊商
	網路智財權法規	立法院
	標準形式專利政策	經濟部標準檢驗局
	內容商業機密制訂	搜尋引擎

(Laura DeNardis, 2013; revised by Dr. Kenny Huang)
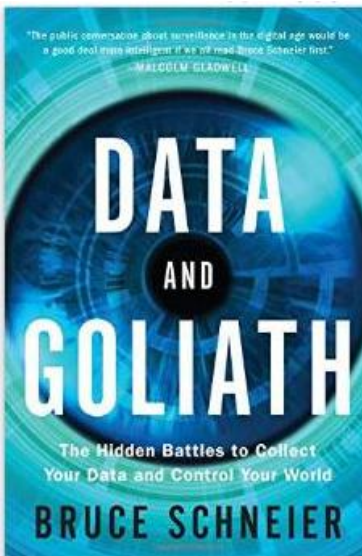
".... The big question is this: how do we design systems that make use of our data collectively to benefit society as a whole, while at the same time protecting people individual? Or..... how do we find a "Nash equilibrium" for data collection..........."

Thank You Question?