Data Analytics as a Strategic Capability

Why Data Analytics

Data Explosion

75% is Duplicate Data; 90% Generated in Last 2 Years

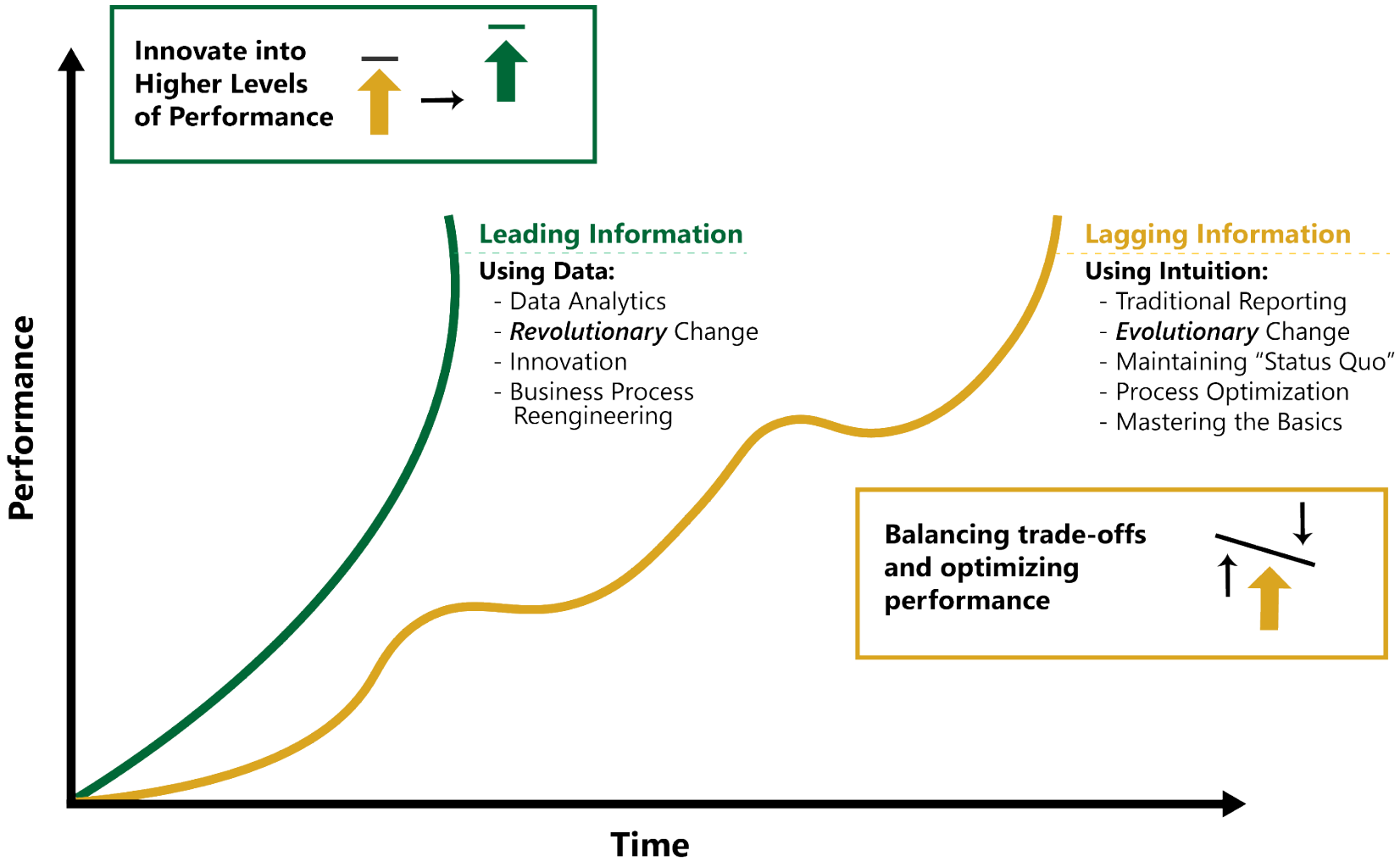
From dawn of time to 2003	5 billion Gigabytes of data created
In 2009	500 billion Gigabytes of data created
2012	2.5 billion Gigabytes PER DAY!
By 2020	44 trillion Gigabytes of data created

Medical Case Study

- 2.5 million peer reviewed medical research articles published per year
- Approx. 7,000 per day
- Only 14% ever make it into practice
- Takes 17 years to get into practice
- Only 50% adoption at 17 years
- Estimated to be One Century behind what has been researched, documented and validated

"If Only We Knew What We Know!"

Data Analytics Value Proposition



Aligning Data Analytics w/Org Reqts

- The Business Environment drives Information Requirements based on the Vision, Mission, Goals.
- The Technical Environment provides this Information as a Key Enabler to Data Analytics.

Truths

Truths

Business Truth

If the Stats truth is valid and there is an impact what types of decisions or implications does this have to the business and decision makers?

Stats Truth

The researchers truth. How significant are the numbers and how likely are the data going to provide a valid result versus just a coincidence.

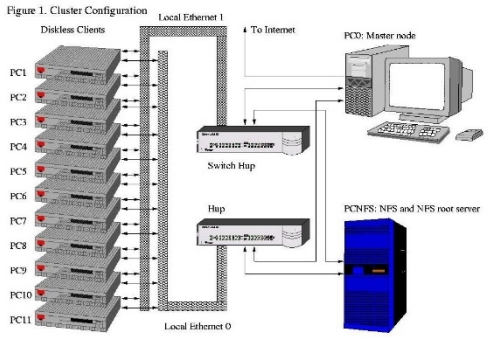
Data Truths

Verified, cleaned and ready for analysis.
Can be deceiving by itself.

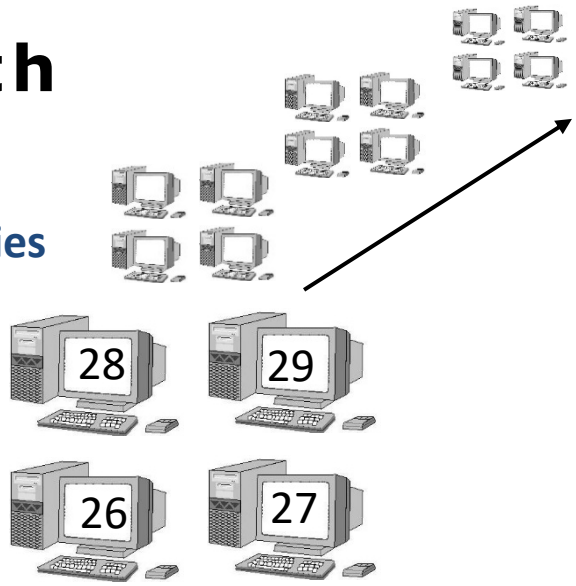
Versions of Truth

Localized Data Systems

Version 7-25



End users w/local copies of the data



Legacy Data Storage Centers

Version 4

Version 5

Version 6

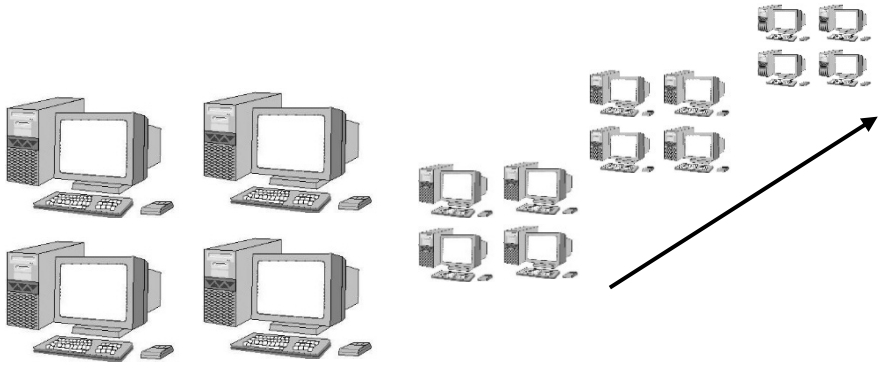
Source Systems

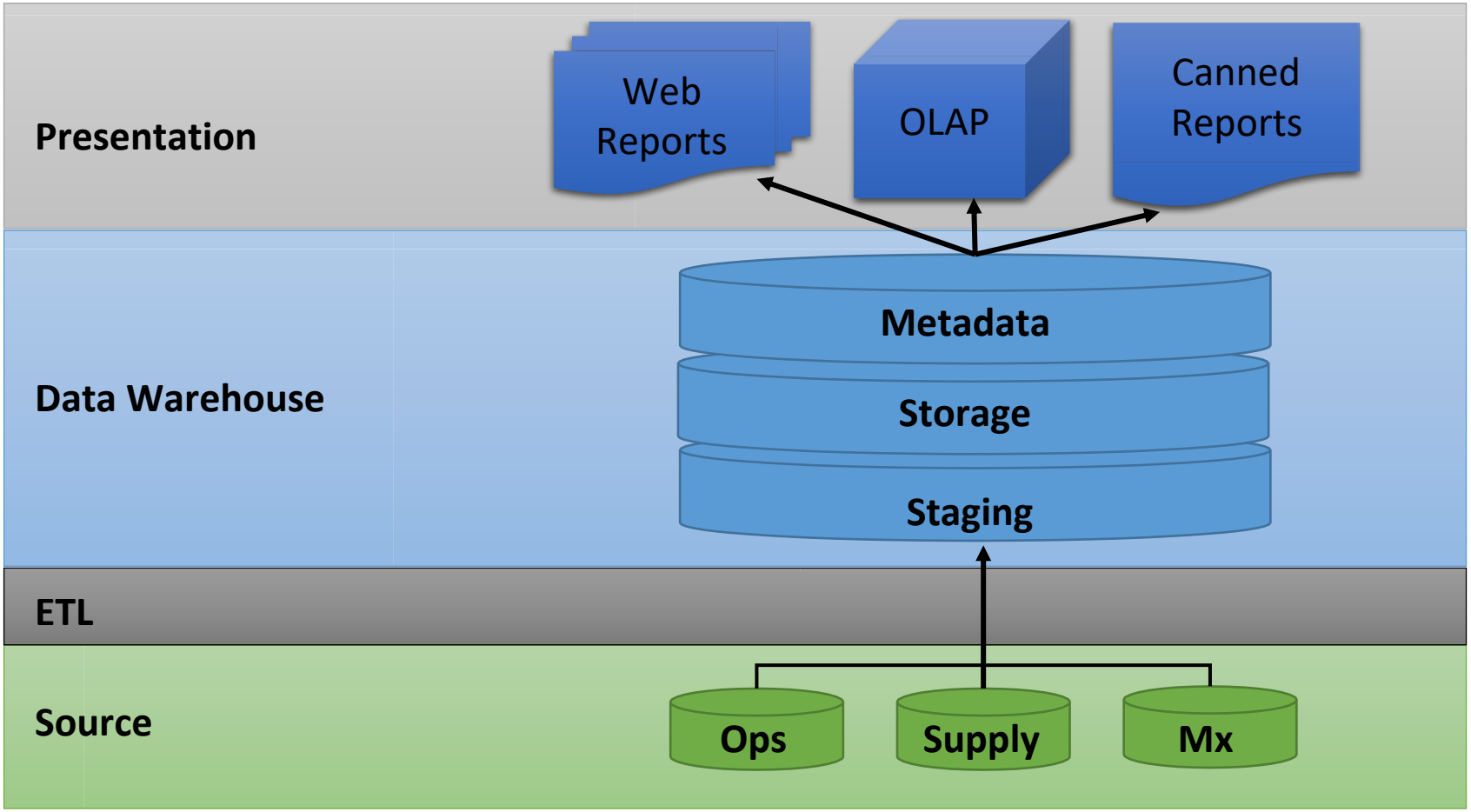
Version 1

Version 2

Version 3

Single Version of the Truth

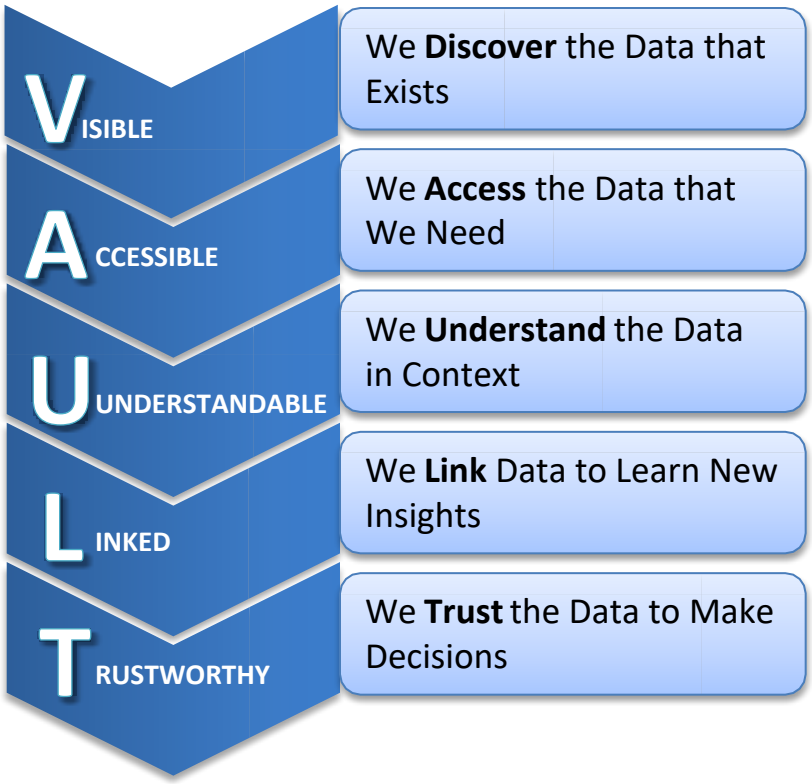




Air Force Data Panel

- Chaired by ~~A6 CTO~~ CDO ~~under MG~~ (SAF/CO)
- Has 4 Tiger Teams
  - Governance Team
  - **Authoritative Data Sources (ADS) Team**
  - Data Hub Team
  - **Data Analytics Team**
- Established~~ing~~ Chief Data Officer (CDO) Position and Staff (MajGen Crider >> Mrs. Eileen Vidrine)
  - IOC ~~October 2017~~ February 2018
  - FOC ~~August 2018~~ August 2018
- Working Closely with Whitehouse Data Cabinet

CDO Strategic Objectives



Data And Data Science

Other forms of data?

- **Metadata** - or a description of other data
  - Example: Library Card Catalog is the metadata that provides a description of a books contents.
  - Can be both attribute and variable

Twitter Metadata


```
"actor":
{
   "objectType": "person",
   "id": "id:twitter.com:277184168",
   "link": "http:\/\/www.twitter.com\/KidCodo",
   "displayName": "Zach Codo",
   "postedTime": "2011-04-04T21:31:20.000Z",
   "image": "https:\/\/si0.twimg.com\/profile_images\/3664410292\/1c
normal.jpeg",
   "summary": null,
   "links":
   [
      {
         "href": null,
         "rel": "me"
      }
   ]
   "friendsCount": 64,
   "followersCount": 207,
   "listedCount": 1,
   "statusesCount": 11207,
   "twitterTimeZone": "Central Time (US & Canada)",
   "verified": false,
   "utcOffset": "-21600",
   "preferredUsername": "KidCodo",
   "languages":
   [
      "en"
   ],
   "location":
   {
      "objectType": "plac
      "displayName": "Nap
   },
      "favoritesCount": 1
}
```

```
"location":
    {
    "objectType": "place",
    "displayName": "Boulder, CO",
    "name": "Boulder",
    "country_code": "United States",
    "twitter_country_code": "US",
    "link": "https://api.twitter.com/1.1/geo/id/fd70c22040963ac7.json",
    "geo": {
        "type": "Polygon",
        "coordinates":
        [
            [
                [
                    -105.3017759,
                    39.953552
                ],
                [
                    -105.3017759,
                    40.094411
                ],
                [
                    -105.183597,
                    40.094411
                ],
                [
                    -105.183597,
                    39.953552
```

```
"object":
{
   "objectType": "note",
   "id": "object:search.twitter.com,2005:347769243409977344",
   "summary": "With Cardiff, Crystal Palace, and Hull City joining the EPL from the Championship it
will be a great relegation battle at the end.",
   "link": "http:\/\/twitter.com\/KidCodo\/statuses\/347769243409977344",
   "postedTime": "2013-06-20T17:33:43.000Z"
}
```

Structured Data

MICAP Period Sta	SOS RIC	ALC	SOS Code	Group	Org	PD
09/01/2016	880	Other	F01	Unrep	Other	Other
09/01/2016	880	Other	F01	Unrep	Other	Other

- Quantitative
- Has defined columns and rows
- Fits in a:
  - Relational Database Table
  - Database Matrix
    - Excel File (Flat File Database)

Unstructured Data

- Not in a relational or flat file database
- Data from the world around us
- Forms:
  - Meeting Minutes
  - Briefings
  - Discrepancy Write Ups (Maintenance or otherwise)
  - Empirical observations w/out a common data structure not captured in a table or matrix

Semi-Structured Data

- Unstructured Data Captured in a Markup Document (XML) to Provide Categorization, Context and Hierarchical Information
- Examples:
  - XML (Extensible Markup Language)
  - JSON (JavaScript Object Notation)

XML

```
... >
  <name>Barry & Associates, Inc.</name>
  <phone>612-321-8156</phone>
  <street1>14597 Summit Shores Dr</street1>
  <street2></street2>
  <city>Burnsville</city>
  <state>MN</state>
  <postalcode>55306</postalcode>
  <country>United States</country>
< ...
```

JSON

```
{
  "name"       : "Barry & Associates, Inc.",
  "phone"      : "612-321-8156",
  "street1"    : "14597 Summit Shores Dr",
  "street2"    : "",
  "city"       : "Burnsville",
  "state"      : "MN",
  "postalcode" : "55306",
  "country"    : "United States"
}
```

Data Analytics Knowledge

Data Analytics takes Business Knowledge, IT Knowledge, and Analytic Knowledge all working together to **gain insight** and **drive innovation**.

DATA ANALYTICS

Interdisciplinary field about methods, processes, and systems to extract knowledge or insights from data

BUSINESS KNOWLEDGE

Understanding business needs

Ability to help business managers set and balance priorities by analyzing consequences of choices and creating business cases

IT KNOWLEDGE

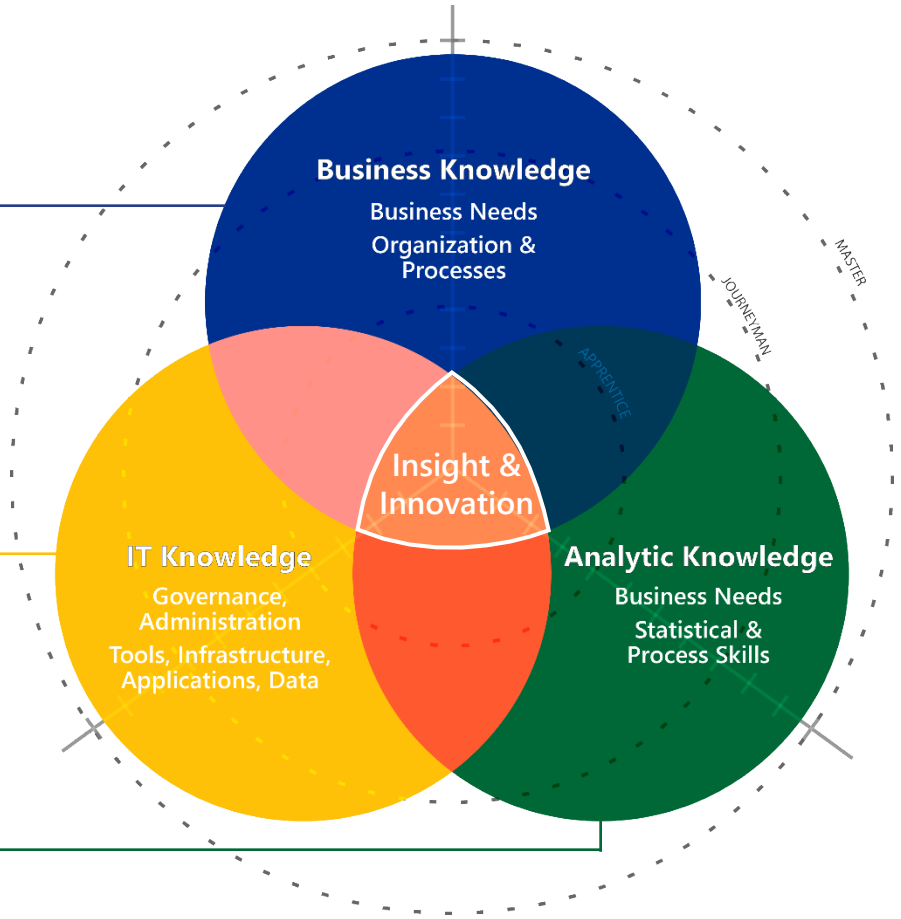
Ability to understand the business intelligence infrastructure implications of business and analytic requirements

Deep understanding of how to access and manage data required to support business and analysis requirements

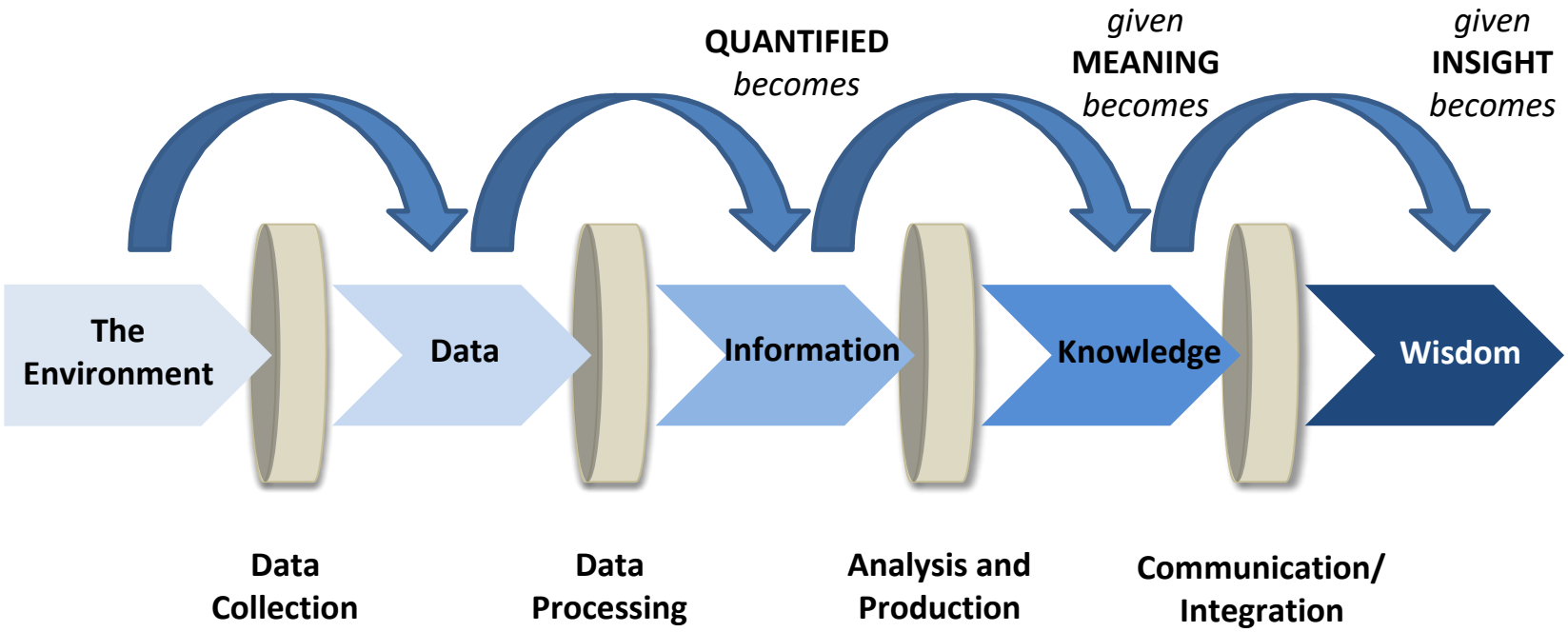
ANALYTICS KNOWLEDGE

Fluency with key analytic applications

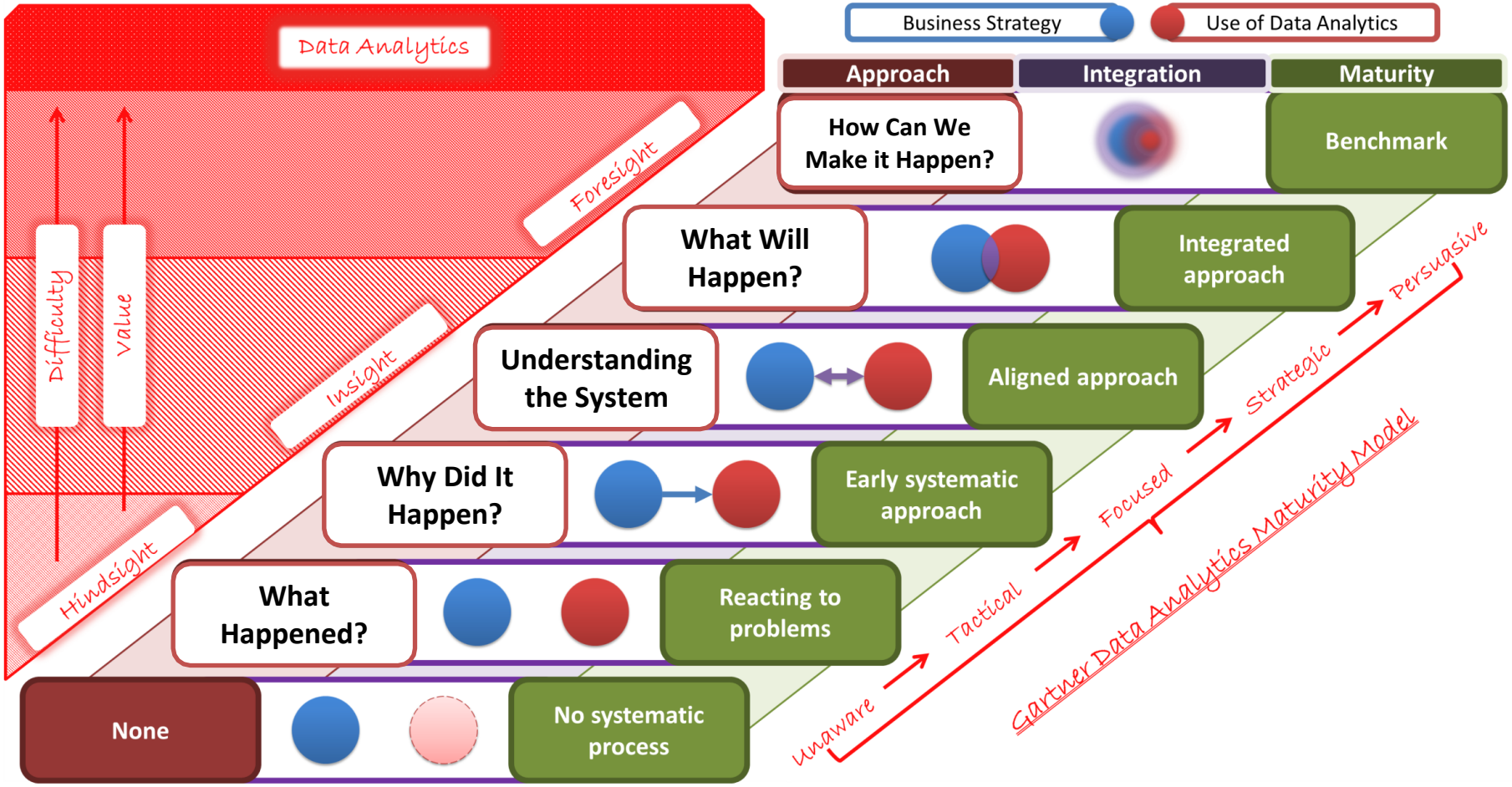
Researching business problems and creating models that help analyze these business problems



Enhancing Data



Evolution of Data Analytics



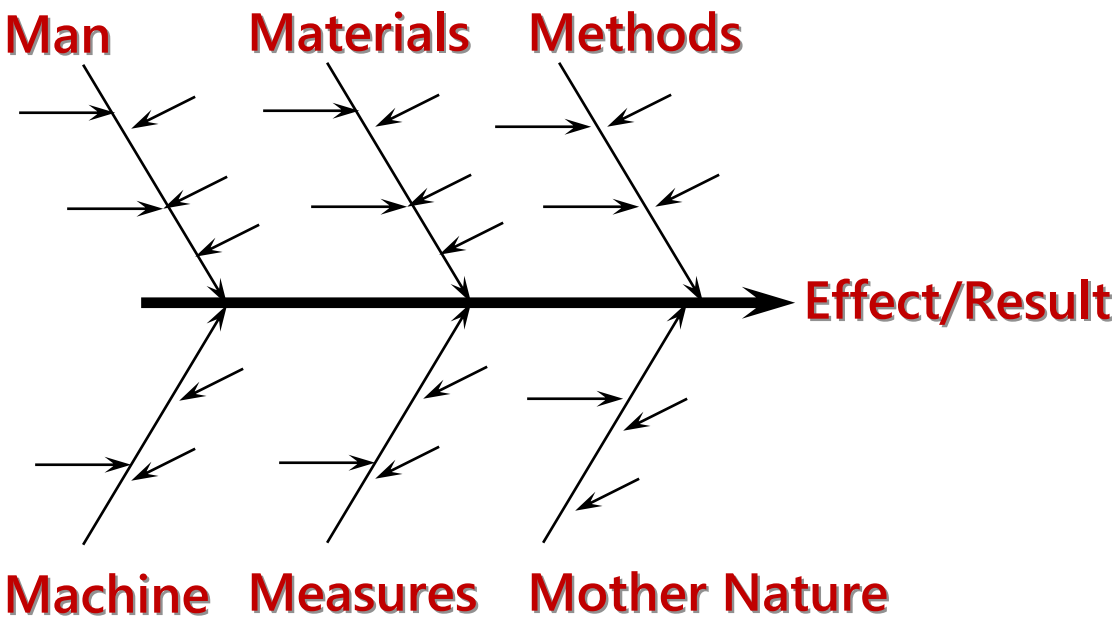
Data Analytics is Systematic

System Components

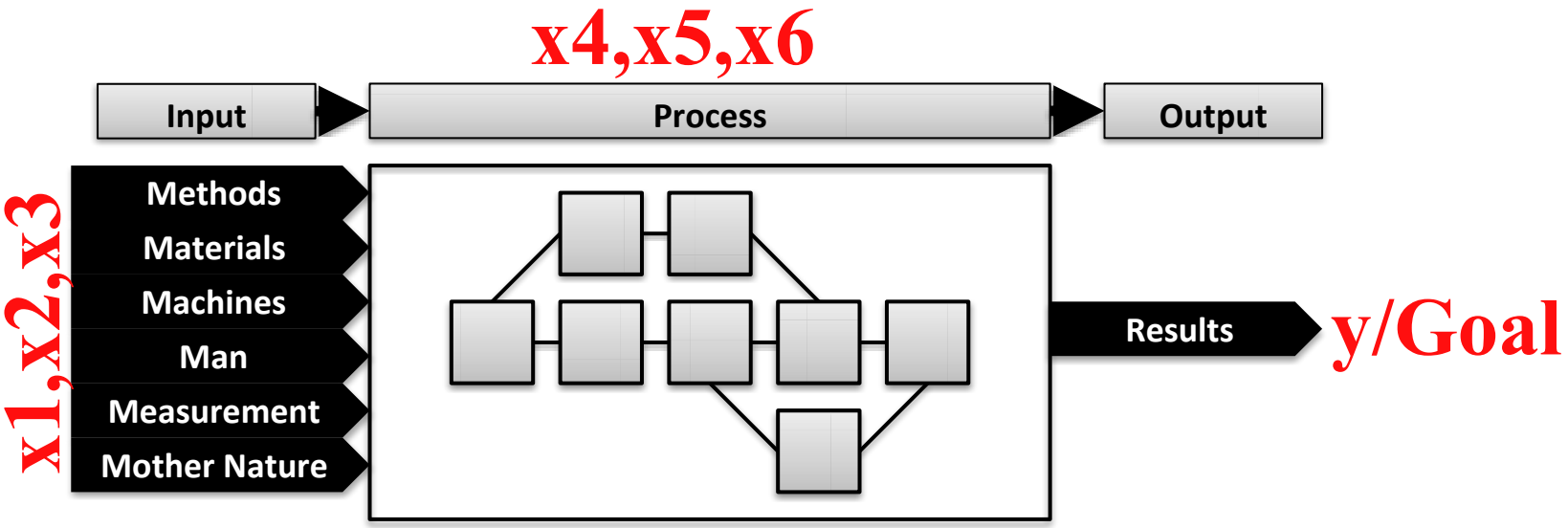
- **Has 4 Components:**
  1. Purpose or Function
  2. Goal (y)
  3. Processes to Achieve Goal (x's)
  4. Metric to Measure attainment of Goal

$$Y=f(x1,x2,x3..)$$

Inputs to System and Processes

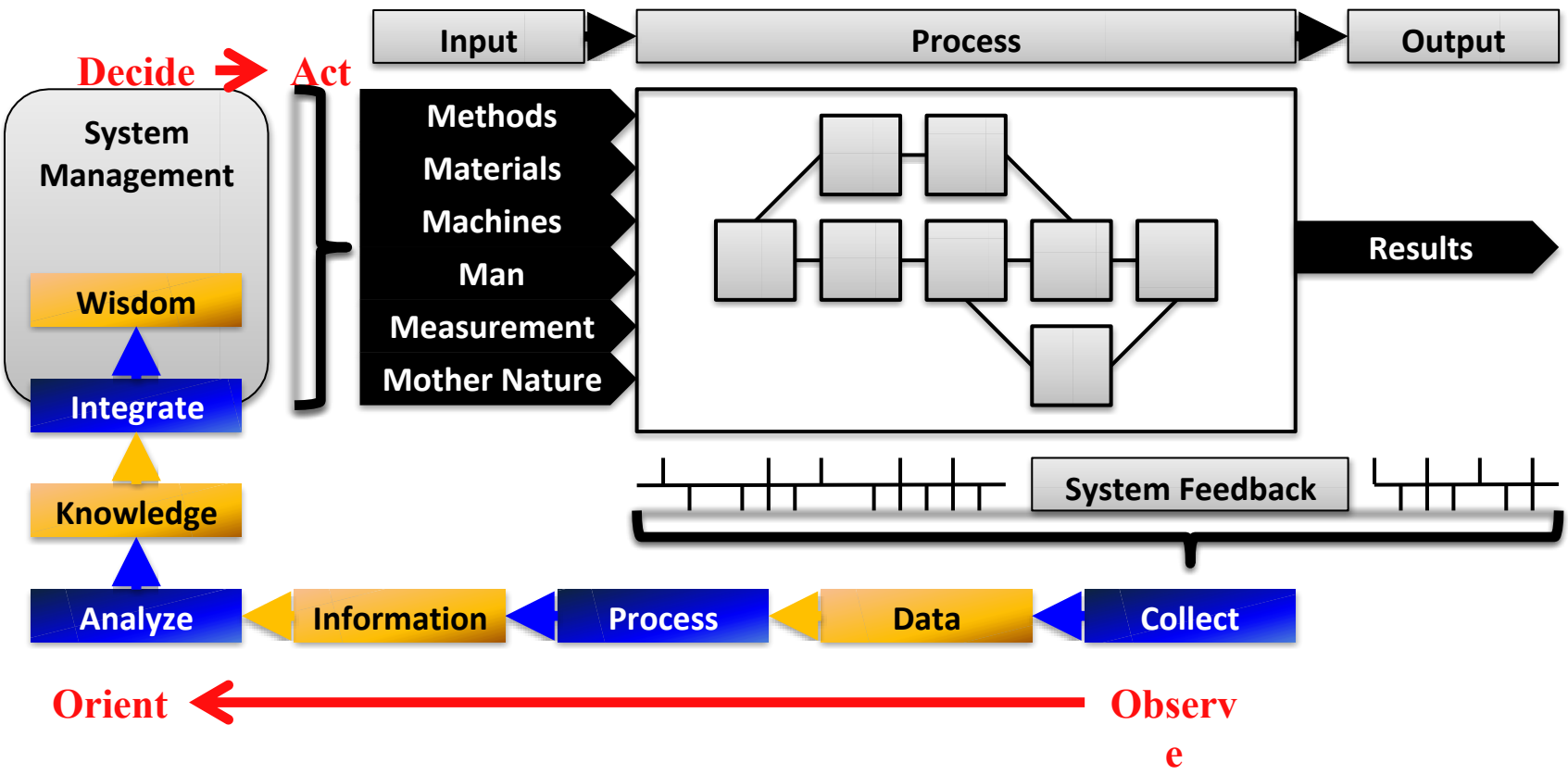


Things We Control

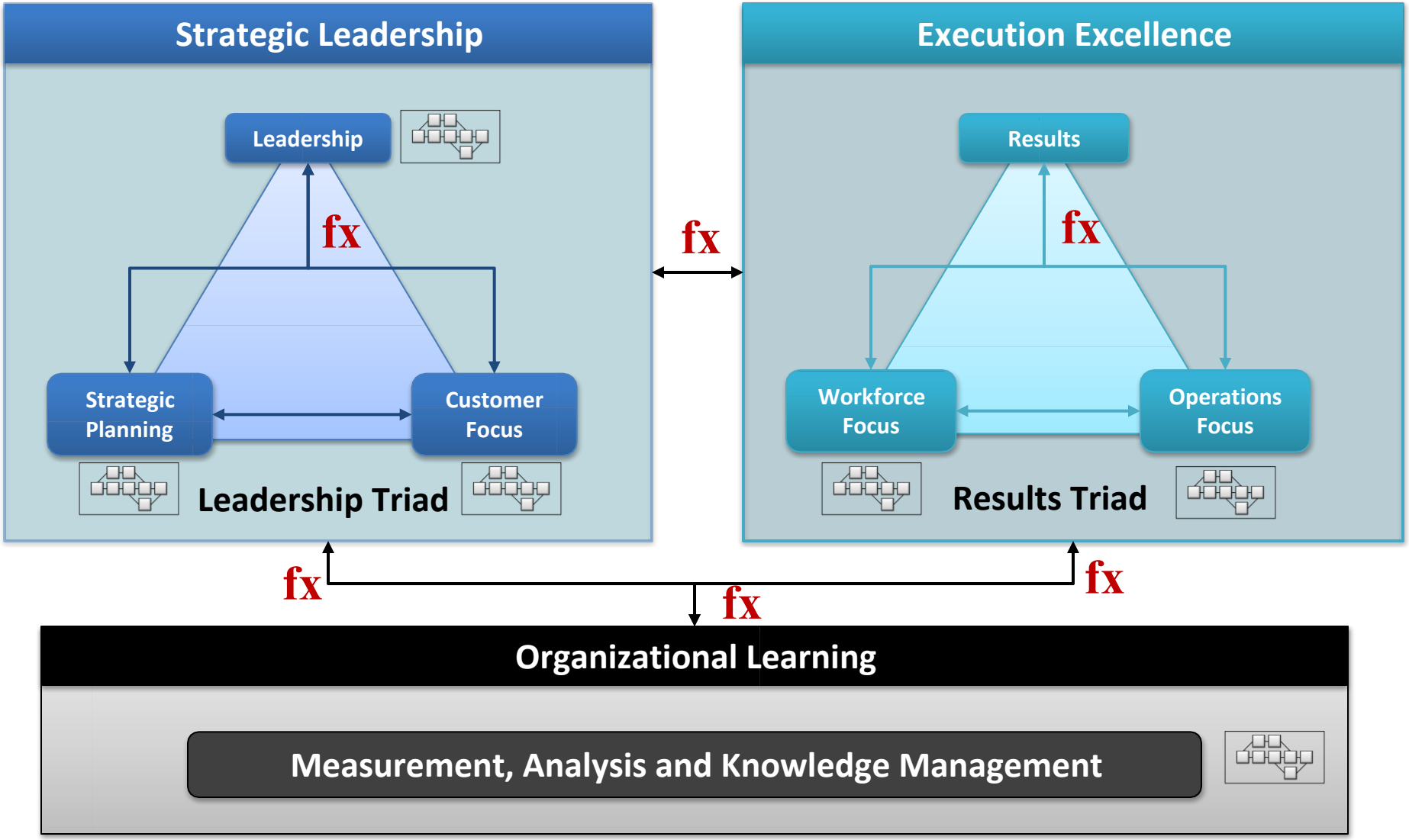


$$Y=f(x1,x2,x3..)$$

An Adaptive System

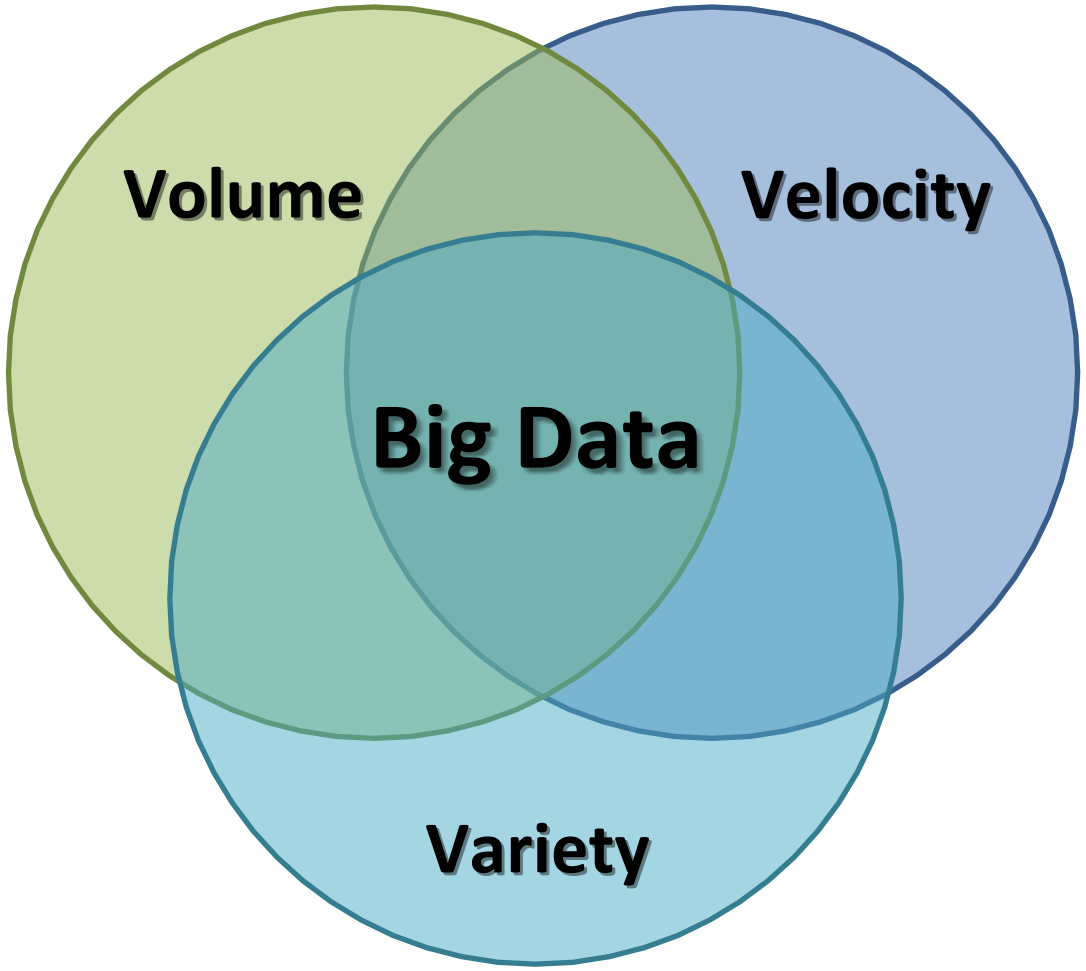


Organizational Systems



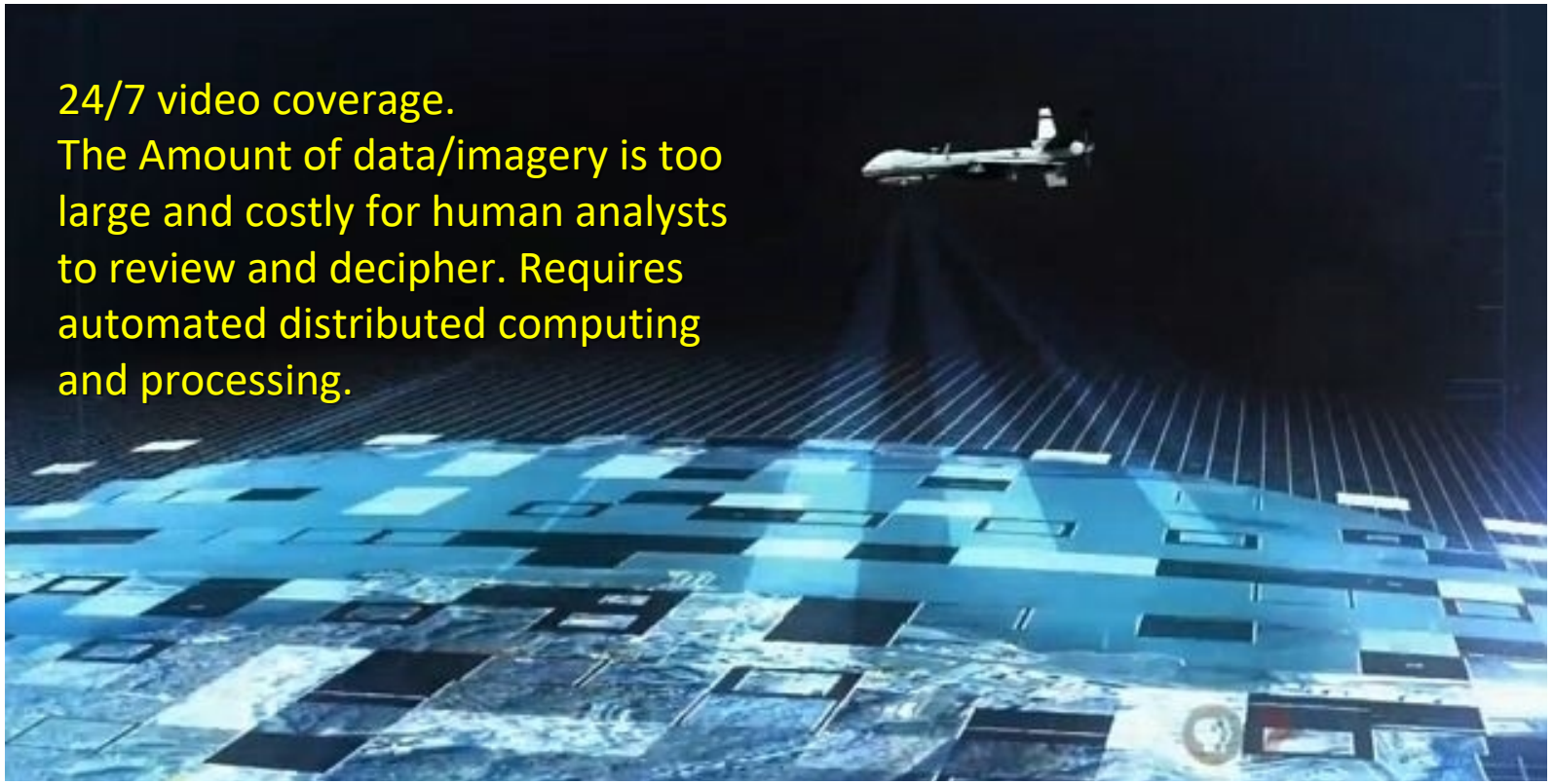
Big Data

3 V's of Big Data



Adapted from 2011 McKinsey Global Institute report

Volume



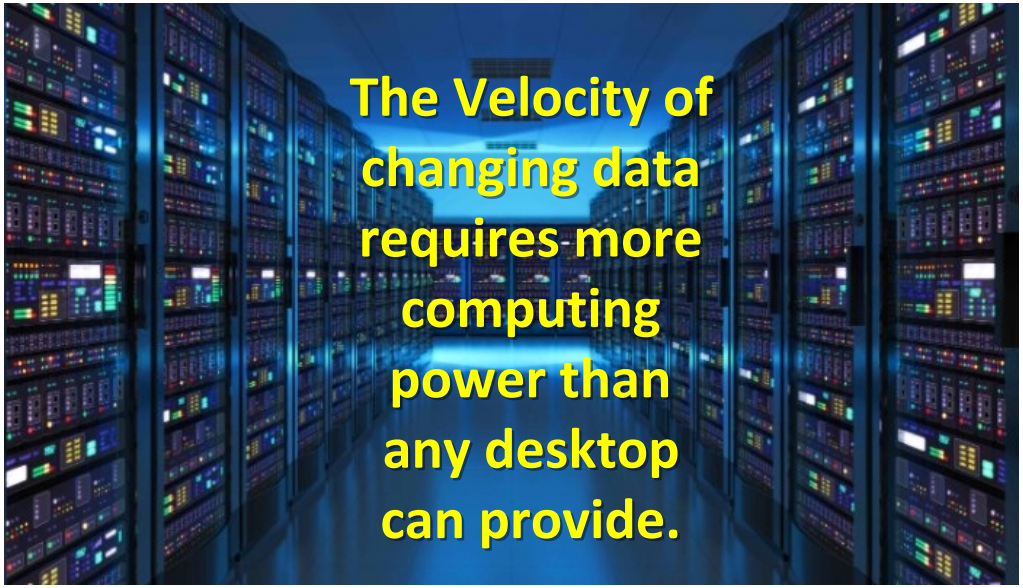
DARPA – Persistent Surveillance of large and densely populated urban environments with wide-area motion imagery sensors.

Velocity

6,000 tweets per second

500 million tweets per day

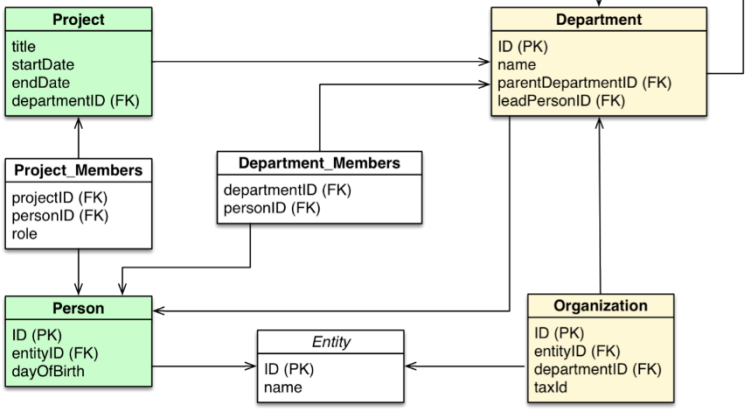
200 billion tweets per year

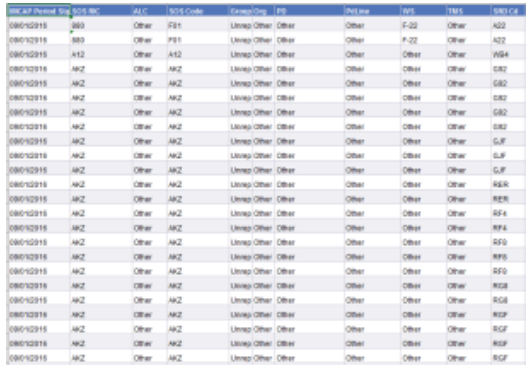


Distributed Computing, Processing

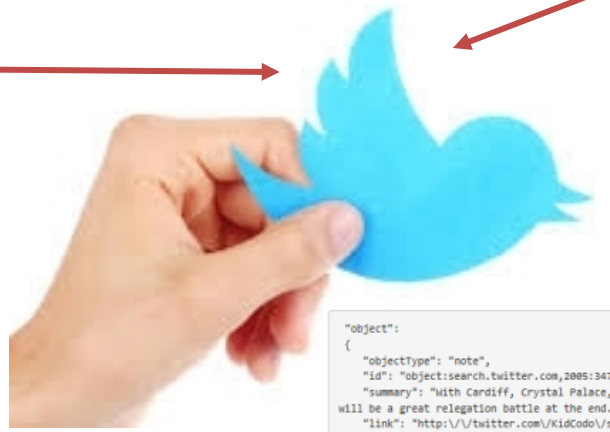
Strategic Capability

Variety






```
<Sensor>
  <name>Sensor 193</name>
  <attributes>
    <Attribute>
      <name>Alpha</name>
      <x>101</x>
      <y>20031</y>
    </Attribute>
    <Attribute>
      <name>Beta</name>
      <x>243</x>
      <y>3037</y>
    </Attribute>
  </attributes>
  <scale>1</scale>
</Sensor>
```

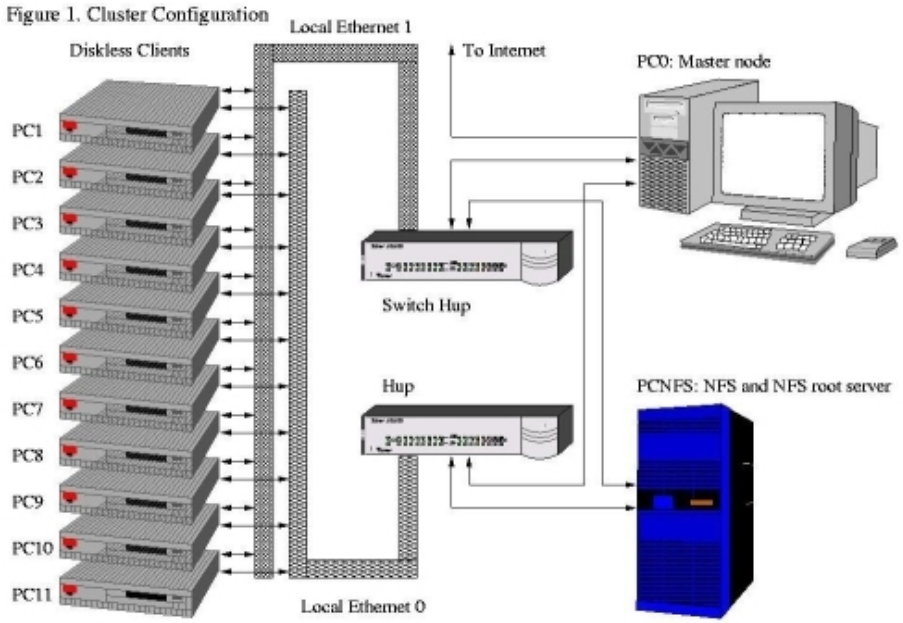


```
"object":
{
   "objectType": "note",
   "id": "object:search.twitter.com,2005:347769243409977344",
   "summary": "With Cardiff, Crystal Palace, and Hull City joining the EPL from the Championship it
will be a great relegation battle at the end.",
   "link": "http:\/\/twitter.com\/KidCodo\/statuses\/347769243409977344",
   "postedTime": "2013-06-20T17:33:43.000Z"
}
```

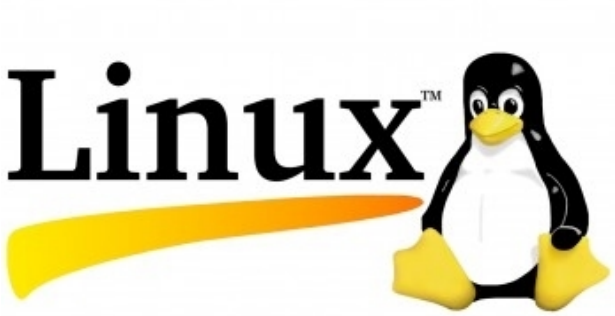
Big Data Infrastructure
Machine Learning/AI

Hardware for Big Data/Machine Learning

- Multiple Servers Running in Parallel
- Each Server having Multiple Processor Cores making independent calculations
- Requires Programming to task each core



An Example



- Open Source Software
- Used as Operating System for Servers
- Supports the ability to Cluster Computers for Distributed Computing

A Database by Any Other Name (Store)

- Open Source Software
- Used for Distributed Storage and Processing of Big Data
- Utilizes Computer Clusters for Distributed Computing
- Parallel File System and Processing

An Example

Getting the Stuff I Need (Coding)

- Open Source Software
- Programming Language used to both access and run continuous routines on various types of data (including Hadoop)
- Facilitates Distributed Computer and Parallel Processing across clustered servers

An Example

```
# For loop on a list
>>> numbers = [2, 4, 6, 8]
>>> product = 1
>>> for number in numbers:
...     product = product * number
...
>>> print('The product is:', product)
The product is: 384
```

Questions?