Big Data

What's Big Data?

No single definition; here is from Wikipedia:

- **Big data** is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.
- The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."

Advantages of Big Data platform

- Real time Analytics
- Add-Hoc Reports and queries (B/I)
- Discover Hidden relations
- Deal with multiple data sources with deferent formats
- Co-relate Data Automatically
- Alerts
- Multiple configuration and policies

Deferent between the old and new way

OLD WAY:

- **Structure Data → Ingest Data → Analyze Data**
- Fixed Capacity
- Monolith

NEW WAY:

- **Ingest Data → Analyze Data → Structure Data**
- Dynamic Capacity
- Ecosystem

RECIPE:

- Cloud
- Data Lake
- Holistic Architecture & Framework

The 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

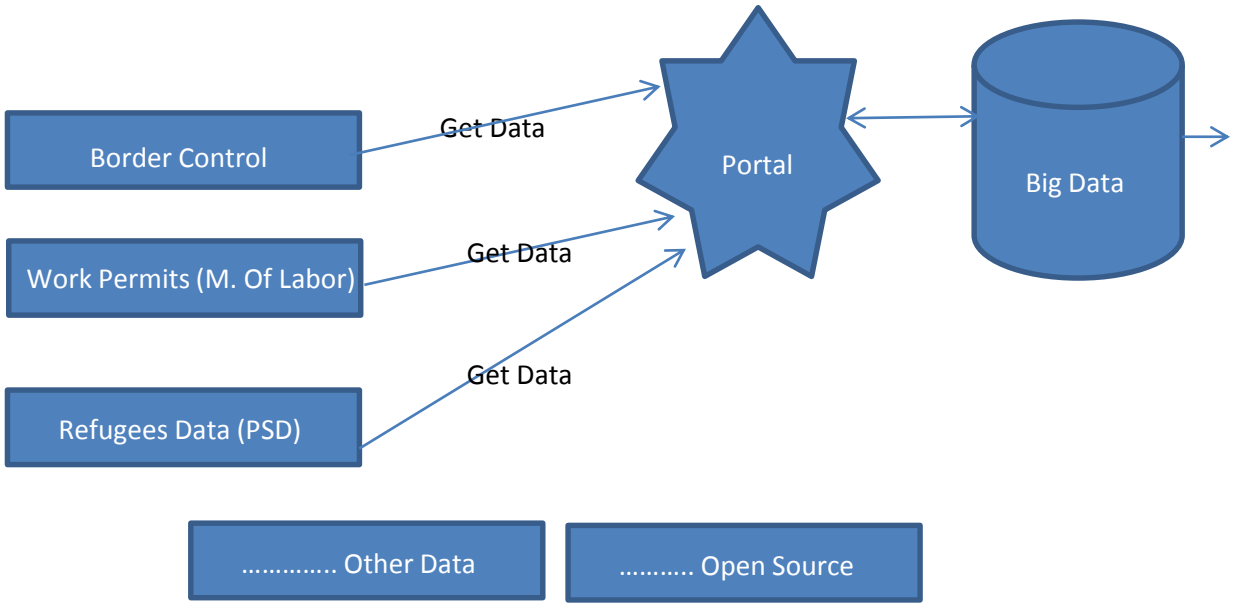
The Model Has Changed…

- **The Model of Generating/Consuming Data has Changed**

Old Model: Few Agencies are generating data, all others are consuming data

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

Jordan Case study



Border Statistics only

Border + Work statistics

Border + Work + Refugees

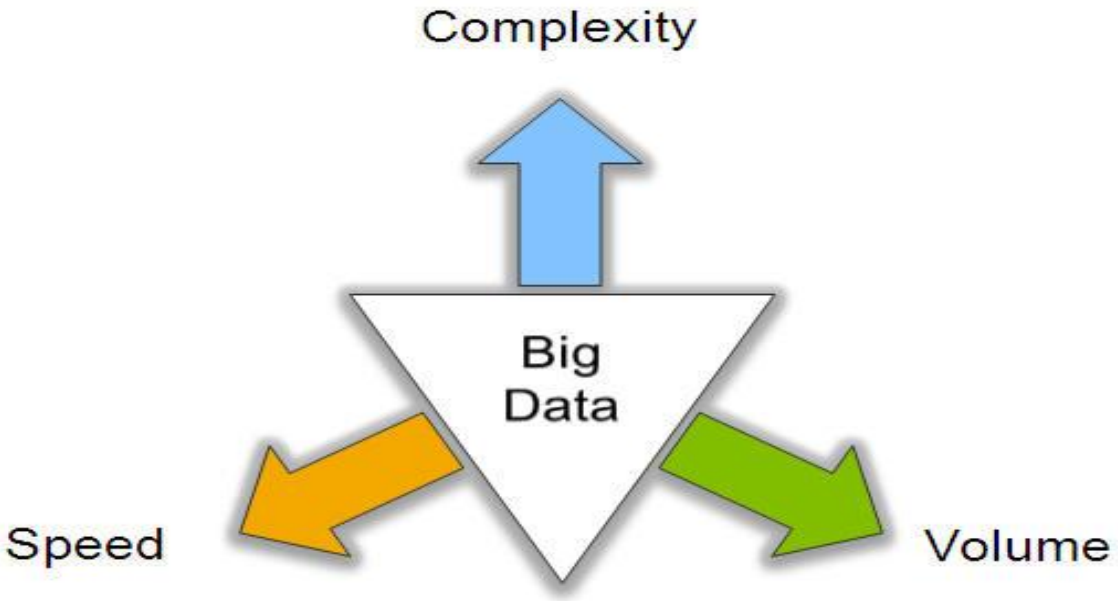
Now you can answer and create a lot of question like:

- How many Jordanians outside Jordan since more than X months
- How many Foreigners inside Jordan For More Than X Months (According to Sex, Age, Nationality ….)
- How many workers in Jordan and they are refugees (By any criteria you want)
- ……

Why is it important for statistics to be one of the key disciplines for Big Data?

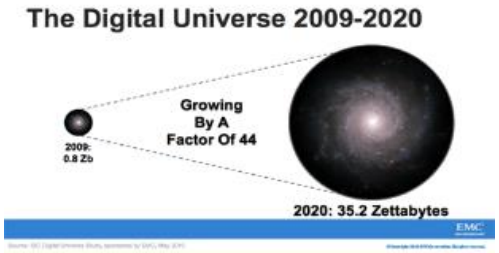
- Statistics is fundamental to ensuring meaningful, accurate information is extracted from Big Data. The following issues are crucial and are only exacerbated by Big Data:
- Data quality and missing data
- Observational nature of data, so that causal questions such as the comparison of interventions may be subject to confounding.
- Quantification of the uncertainty of predictions, forecasts and models
- The scientific discipline of statistics brings sophisticated techniques and models to bear on these issues
- Statisticians help translate the scientific question into a statistical question, which includes carefully describing data structure; the underlying system that generated the data (the model); and what we are trying to assess (the parameter or parameters we wish to estimate) or predict

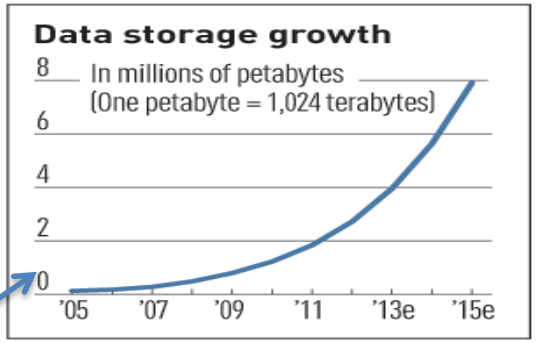
Every thing increases without stop

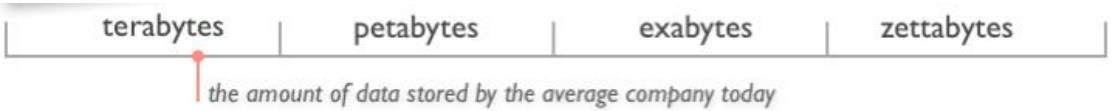


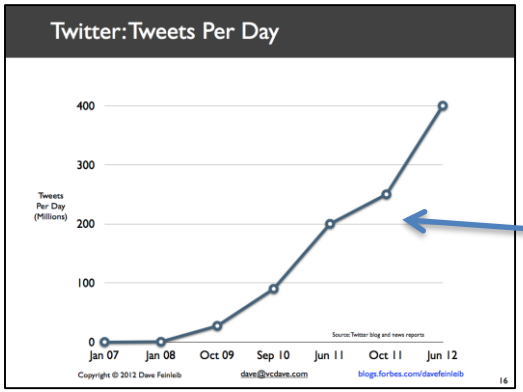
Volume (Scale)

- **Data Volume**
  - 44x increase from 2009 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially









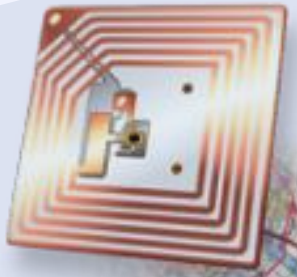
Exponential increase in collected/generated data

12+ TBs
of tweet data
every day

? TBs of
data every day

25+ TBs **of**
log data
every day

30 billion RFID
tags today
(1.3B in 2005)



4.6
billion
camera
phones
world wide

100s of
millions
of GPS
enabled
devices sold
annually

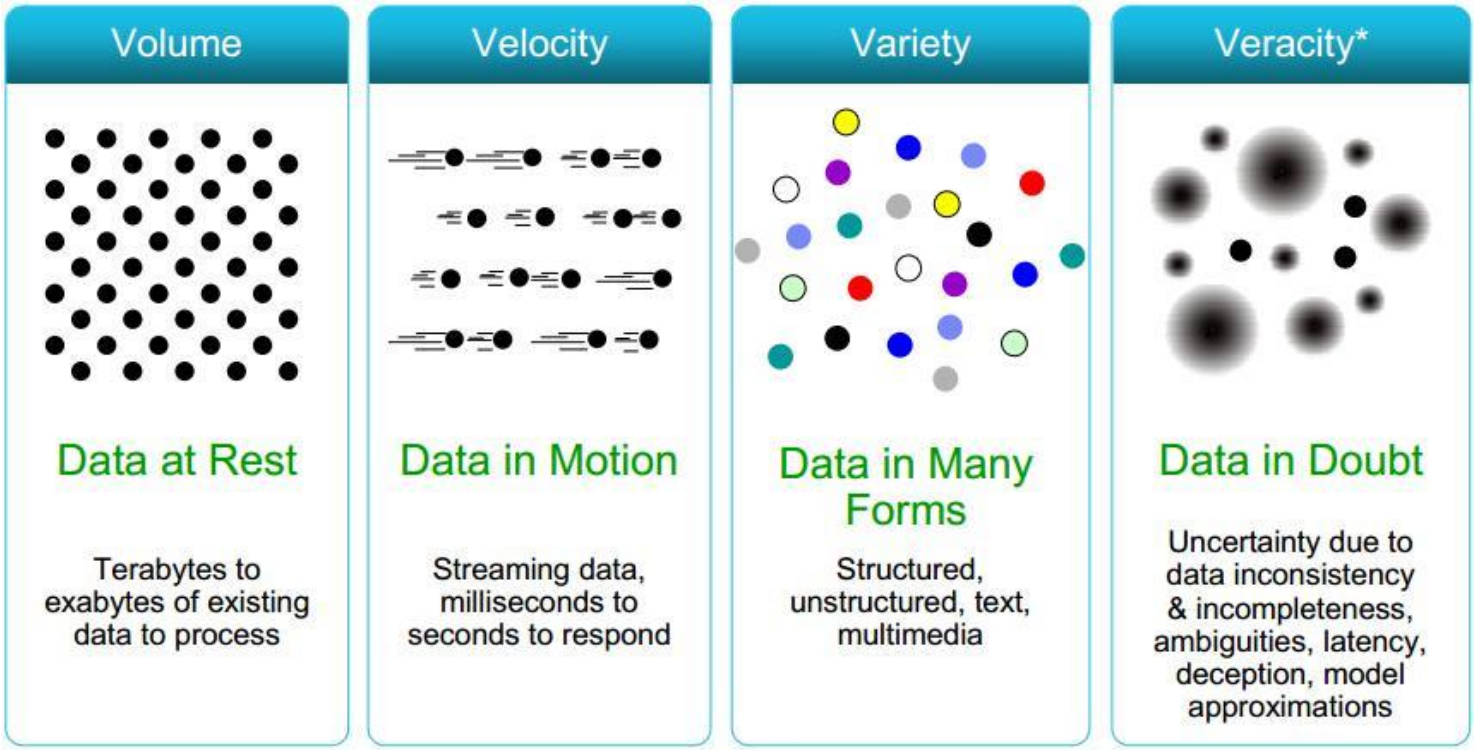
76 million smart meters
in 2009…
200M by 2014

2+
billion
people on
the Web
by end
2011

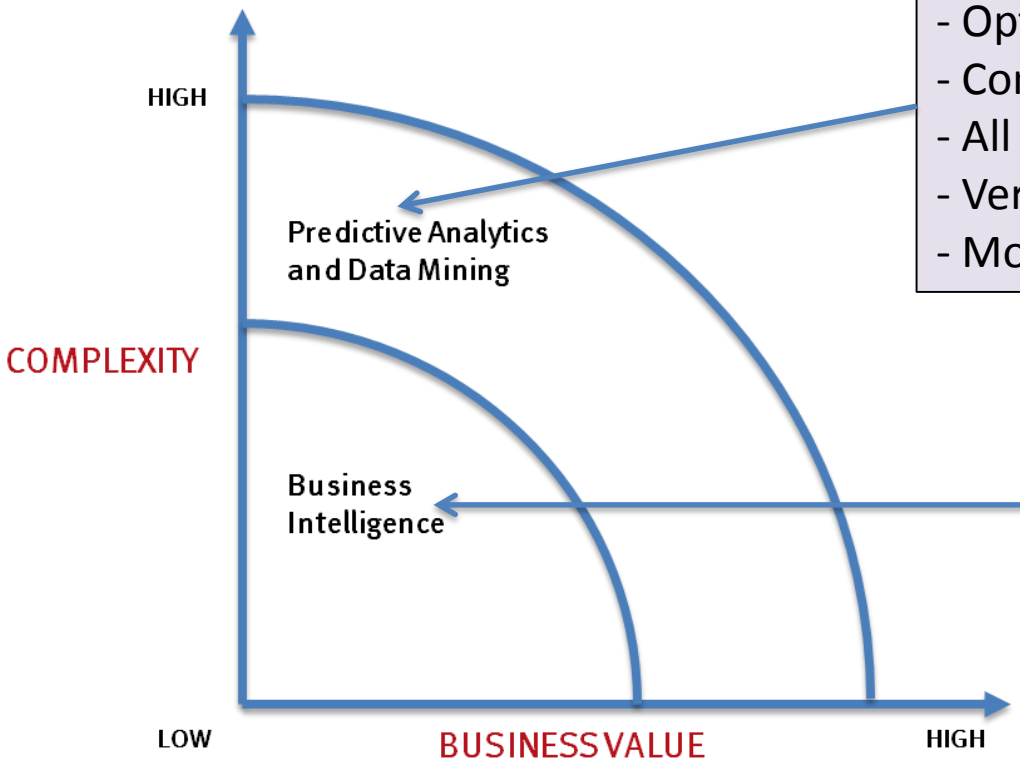
The Earthscope

- The Earthscope is the world's largest science project. Designed to track North America's geological evolution, this observatory records data over 3.8 million square miles, amassing 67 terabytes of data. It analyzes seismic slips in the San Andreas fault, sure, but also the plume of magma underneath Yellowstone and much, much more. (http://www.msnbc.msn.com/id/44363598/ns/technology_and_science-future_of_technology/#.TmetOdQ--ul)

Some Make it 4V's



What's driving Big Data



- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

How to start

1. Get a clever IT staff
2. Define your Data sources and define sizes
3. Buy a Big Data Model (Like Oracle-Hadoop , IBM-Watson …)
4. Create a Portal.
5. Connect the model to data sources (use internet if you haven’t)
6. Start working.