BIG DATA

The Age of Data Deluge

Internet: the unprecedented information collector

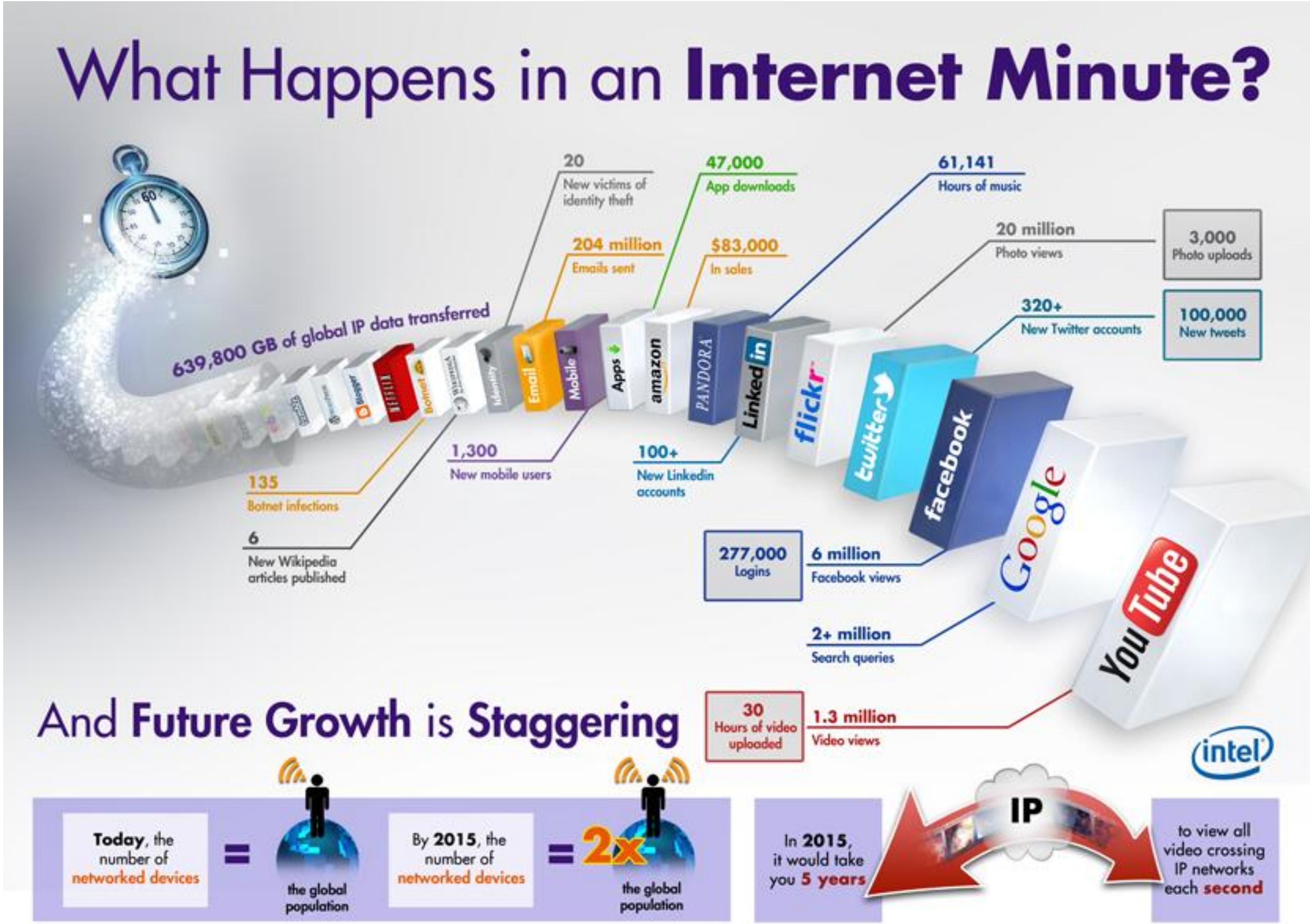
- May 2012: 200m Web servers [Yahoo]
- estd 50+b static pages [Yahoo]
- 40 b photos [Facebook]
- 2012: 31b searches/m [Google]

Typical Big Data:

- Business Intelligence
- Social networks:
  Facebook, Twitter, GPS, ...
- Life Science:
  patient data, imagery, genetics, ...
- Geo Sci: Satellite imagery, weather data, crowdsourcing, ...
  - *Petrol industry:*
    „more bytes than barrels“

Ex: Facebook Graph

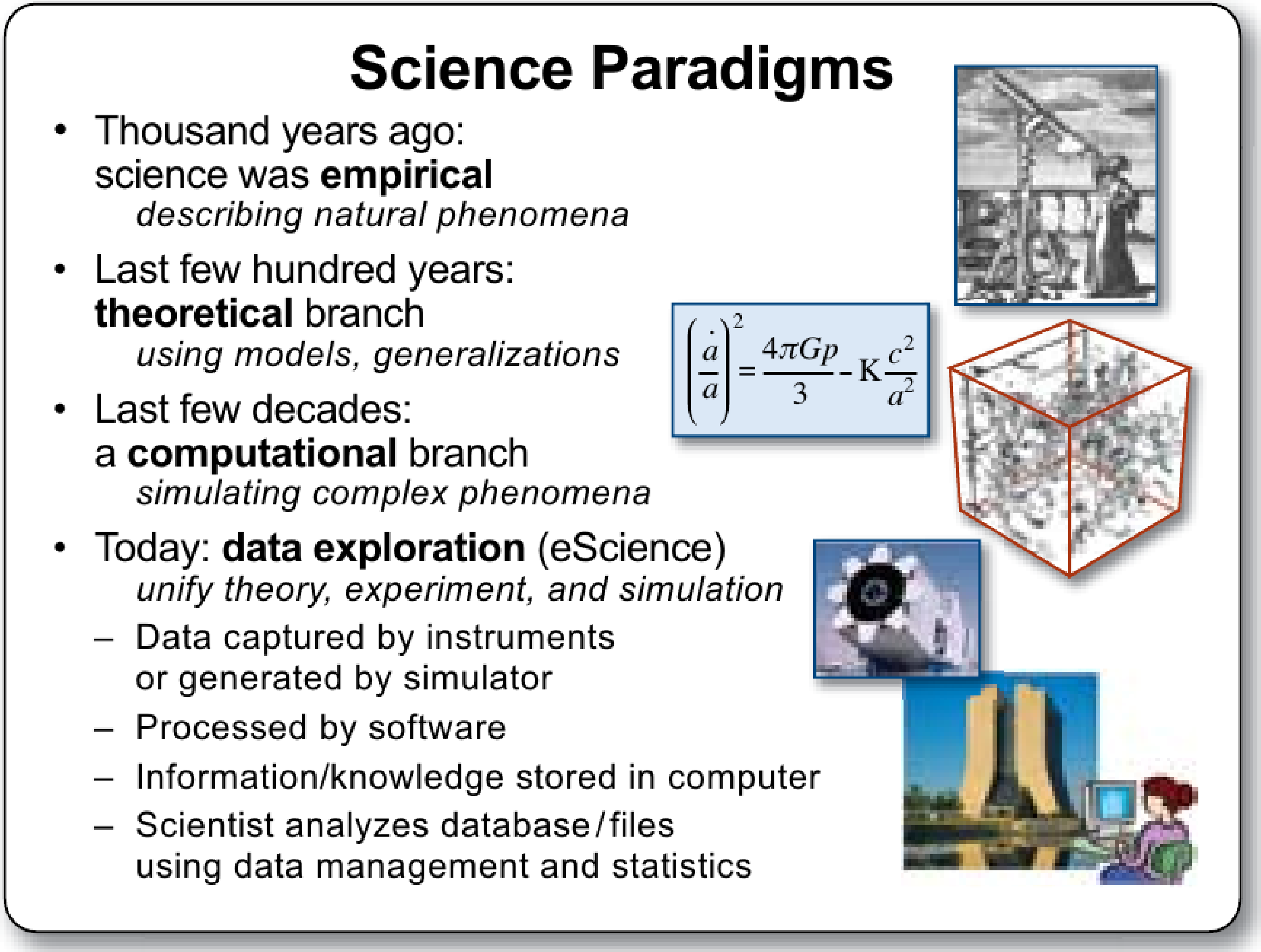
facebook
December 2010


What Happens in an Internet Minute?
639,800 GB of global IP data transferred
20
New victims of identity theft
204 million
Emails sent
47,000
App downloads
$83,000
In sales
61,141
Hours of music
20 million
Photo views
3,000
Photo uploads
320+
New Twitter accounts
100,000
New tweets
135
Botnet infections
6
New Wikipedia articles published
1,300
New mobile users
100+
New Linkedin accounts
277,000
Logins
6 million
Facebook views
2+ million
Search queries
30
Hours of video uploaded
1.3 million
Video views
Identity
Email
Mobile
Apps
amazon
PANDORA
Linked in
flickr
twitter
facebook
Google
You Tube
intel
And Future Growth is Staggering
Today, the number of networked devices
=
the global population
By 2015, the number of networked devices
= 2x
the global population
In 2015, it would take you 5 years
IP
to view all video crossing IP networks each second

[image: Intel]

„The 4th Paradigm“

Tony Hey, Stewart Tansley, Kristin Tolle (eds.)



Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  – Data captured by instruments or generated by simulator
  – Processed by software
  – Information/knowledge stored in computer
  – Scientist analyzes database/files using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi Gp}{3} - K\frac{c^2}{a^2}$$

„Big Data“: The 4+ Vs

- „data too big to transport“,
  but also „too complex to process“

- Volume - ngEO plannings: 10^12 images under ESA custody
- Velocity - NASA EOSDIS: 5 TB/d; LOFAR: 25 TB/h; phones: 1+ PB/d
- Variety - grids; point clouds; general meshes; vectors; text; graphs; ...
- Veracity - Quality, provenance, trust

- ...plus more in blogs: Value, Verisimilitude, Variability, Visualization, ...

Technology Responses

- Novel programming paradigms
  - Massive parallelization on distributed networks: MapReduce / Hadoop
    - *Fixed paradigm: map() input to different nodes, then reduce() to result*
  - Distribute algorithms over heterogeneous hw/sw: Apache Flink, Spark
- Database support for missing datatypes („NoSQL“)
  - Document DB (MongoDB), Graph DB (Neo4j), Array DB (rasdaman)
- Statistical & Machine Learning approaches

- *Big Data Analytics in a nutshell: Databases + Machine Learning*