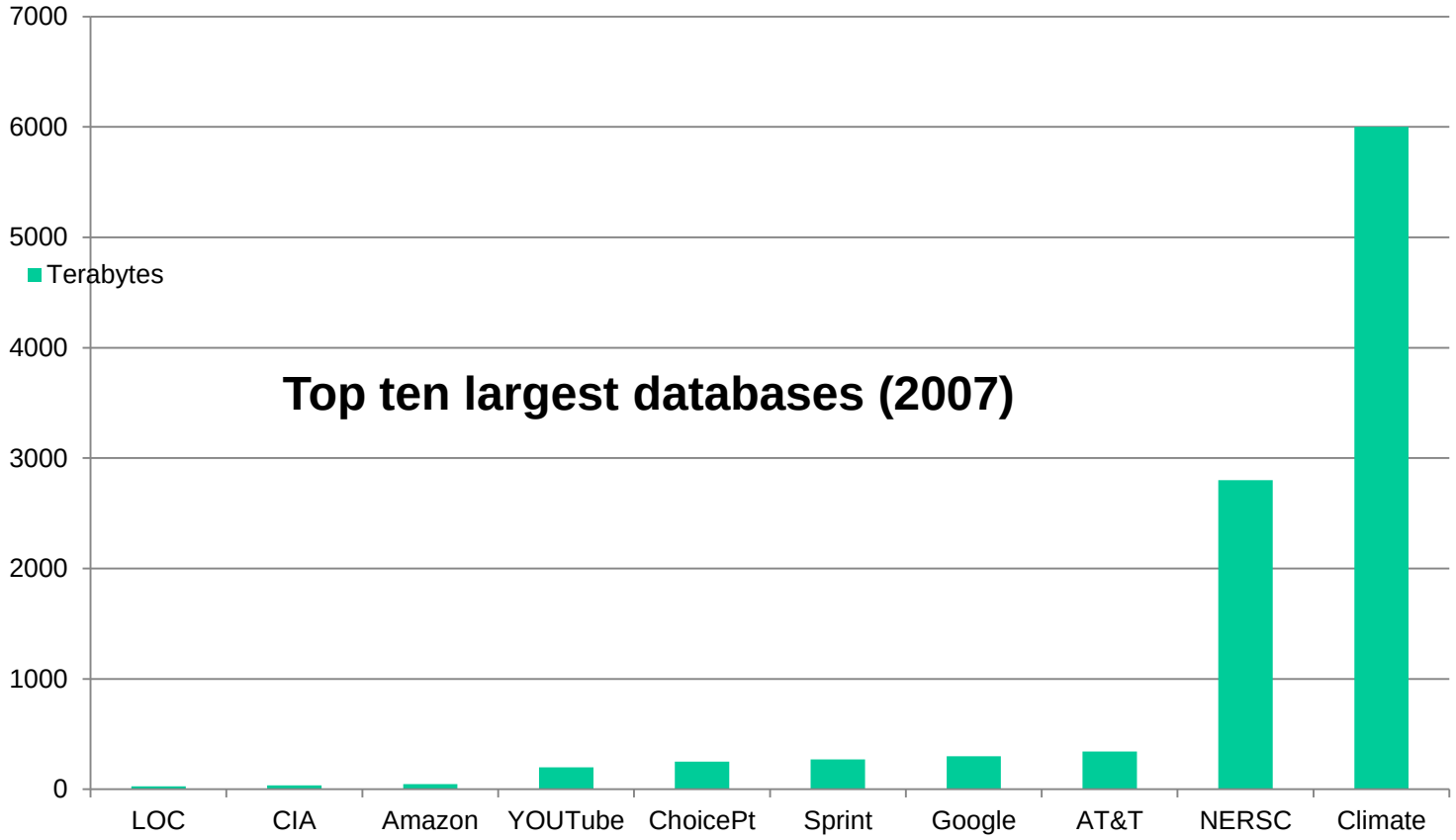
Defining Big-data

B. Ramamurthy

Big-data computing

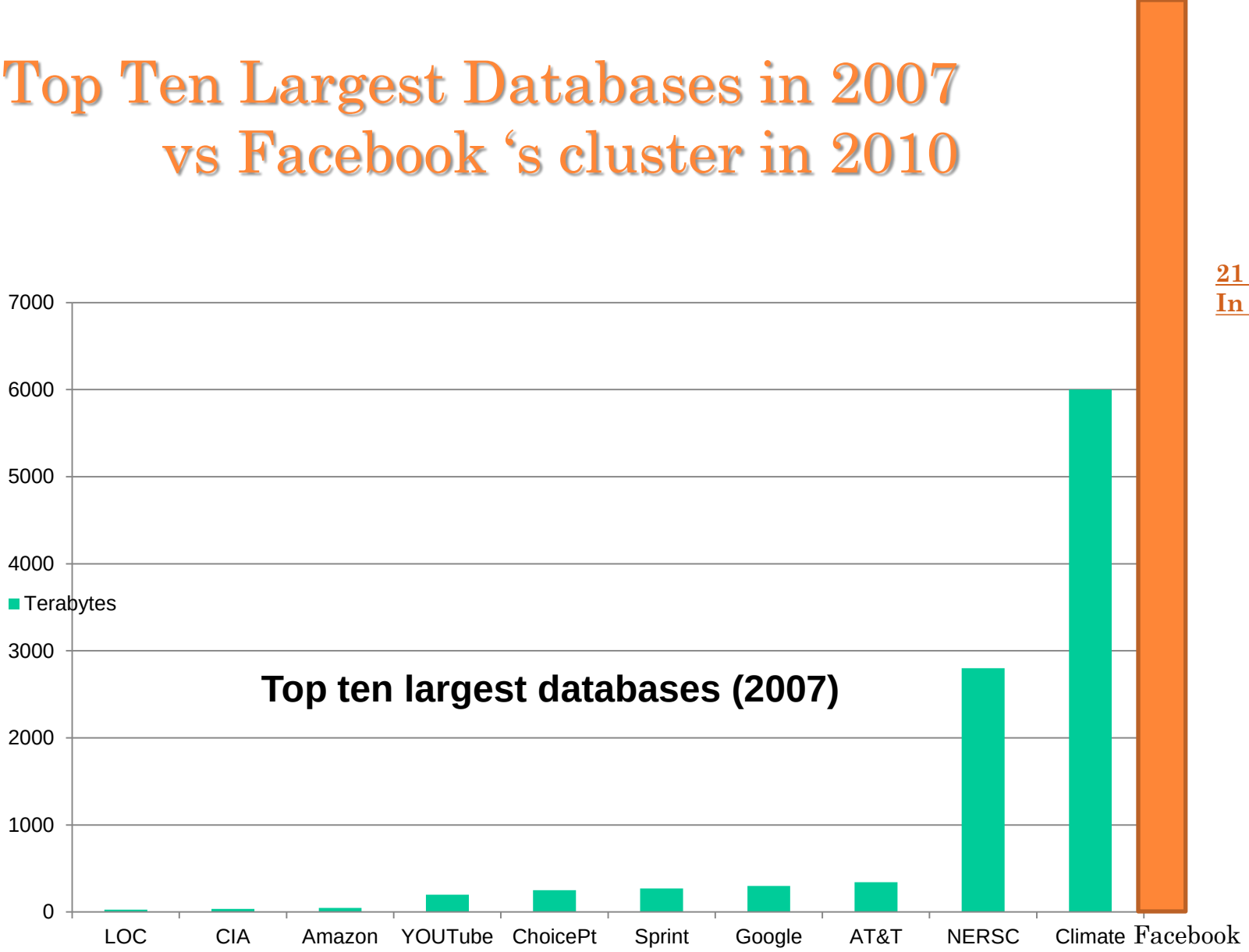
- What is it?
  - Volume, velocity, variety, veracity (uncertainity) (Gartner, IBM): slides 3,4
- How is it addressed? Slide 5
- Why now? Slide 6,7
- What do you expect to extract by processing this large data?
  - Intelligence for decision making: slides 8-15
- What is different?
  - Storage models, processing models: Slides 16
  - Big Data, analytics and cloud infrastructures: slides 17-19
- Summary

Top Ten Largest Databases



Ref: http://www.comparebusinessproducts.com/fyi/10-largest-databases-in-the-world/

Top Ten Largest Databases in 2007 vs Facebook 's cluster in 2010

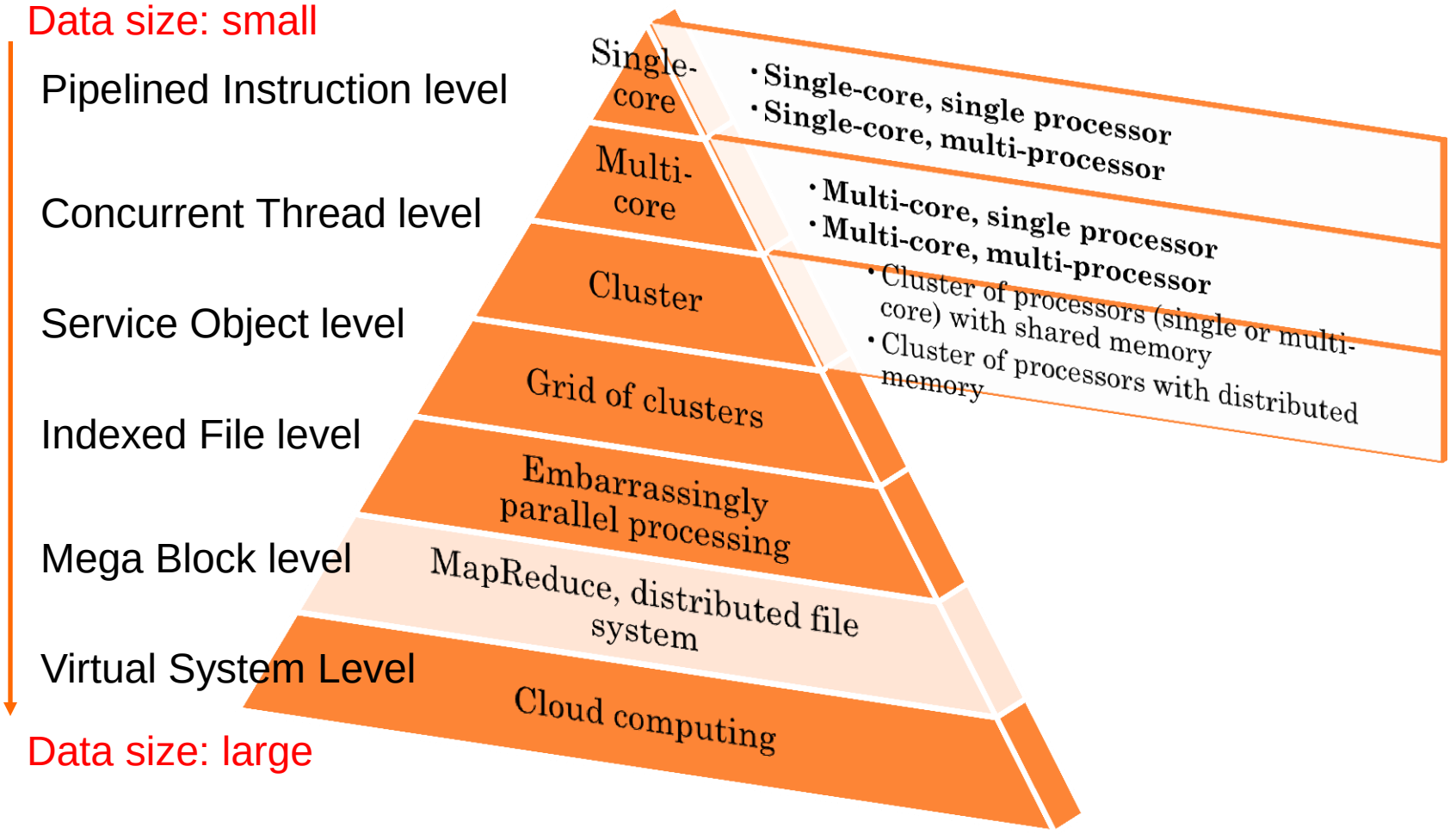


21 PetaByte In 2010

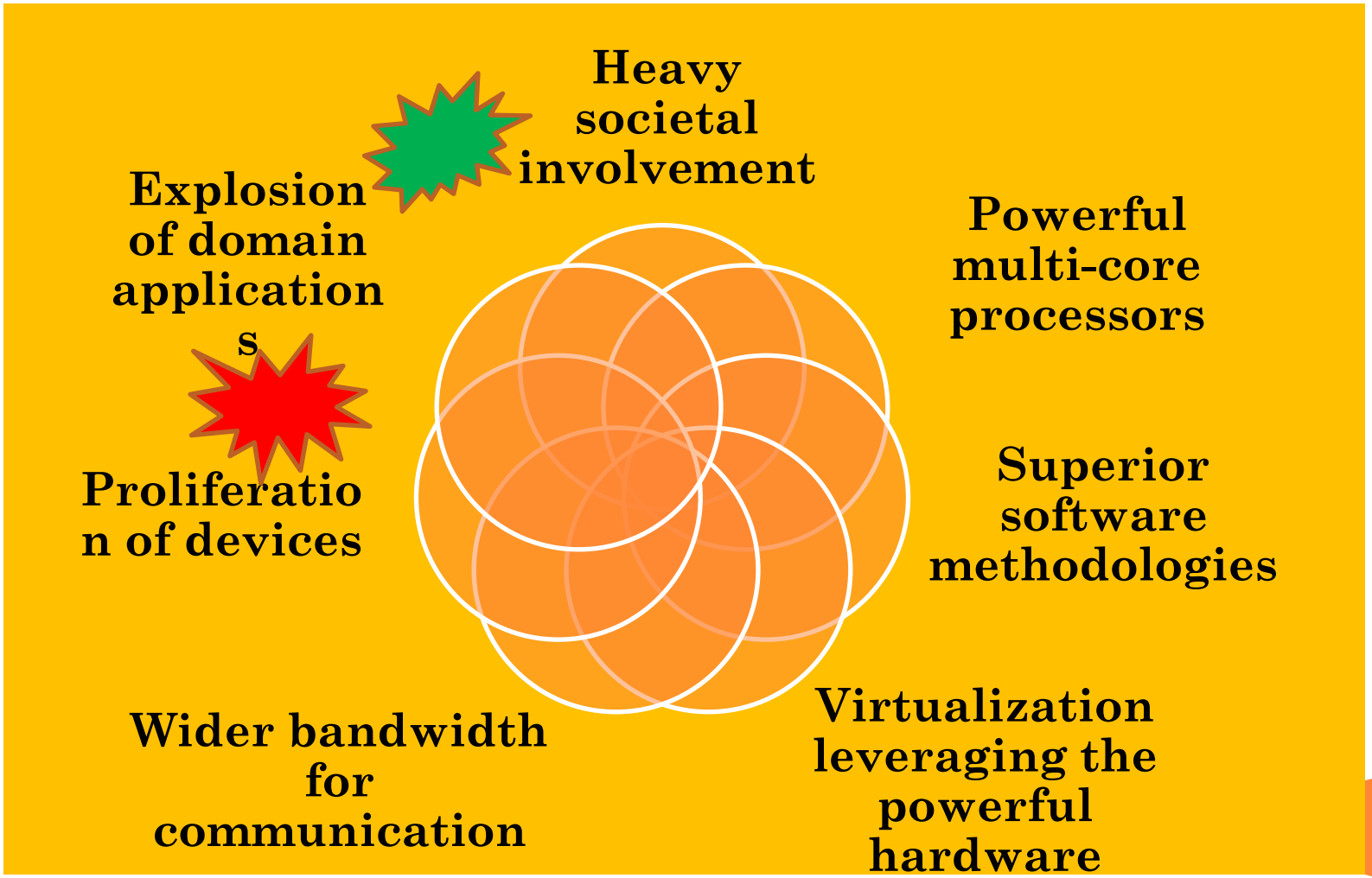
5/14/2013

Ref: http://www.comparebusinessproducts.com/fyi/10-largest-databases-in-the-world

Processing Granularity



A Golden Era in Computing



Data Deluge: smallest to largest

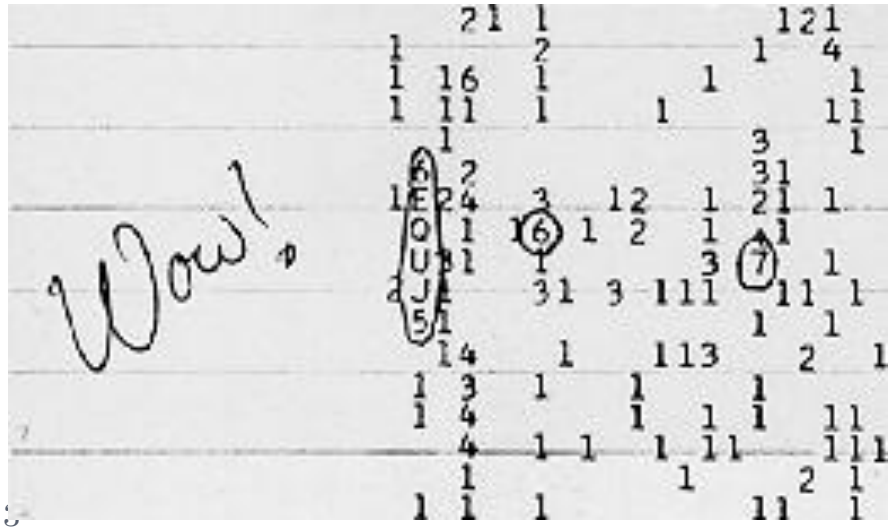
- Bioinformatics data: from about 3.3 billion base pairs in a human genome to huge number of sequences of proteins and the analysis of their behaviors
- The internet: web logs, facebook, twitter, maps, blogs, etc.: Analytics …
- Financial applications: that analyze volumes of data for trends and other deeper knowledge
- Health Care: huge amount of patient data, drug and treatment data
- The universe: The Hubble ultra deep telescope shows 100s of galaxies each with billions of stars

Intelligence and Scale of Data

- Intelligence is a set of discoveries made by federating/processing information collected from diverse sources.
- Information is a cleansed form of raw data.
- For statistically significant information we need reasonable amount of data.
- For gathering good intelligence we need large amount of information.
- As the Fourth Paradigm (FP) book points out enormous amount of data is generated by the millions of experiments and applications.
- Thus intelligence applications are invariably data-heavy, data-driven and data-intensive.
- Data is gathered from the web (public or private, covert or overt), generated by large number of domain applications.

Intelligence

- Search for Extra Terrestrial Intelligence (seti@home project)
- The Wow signal http://www.bigear.org/wow.htm



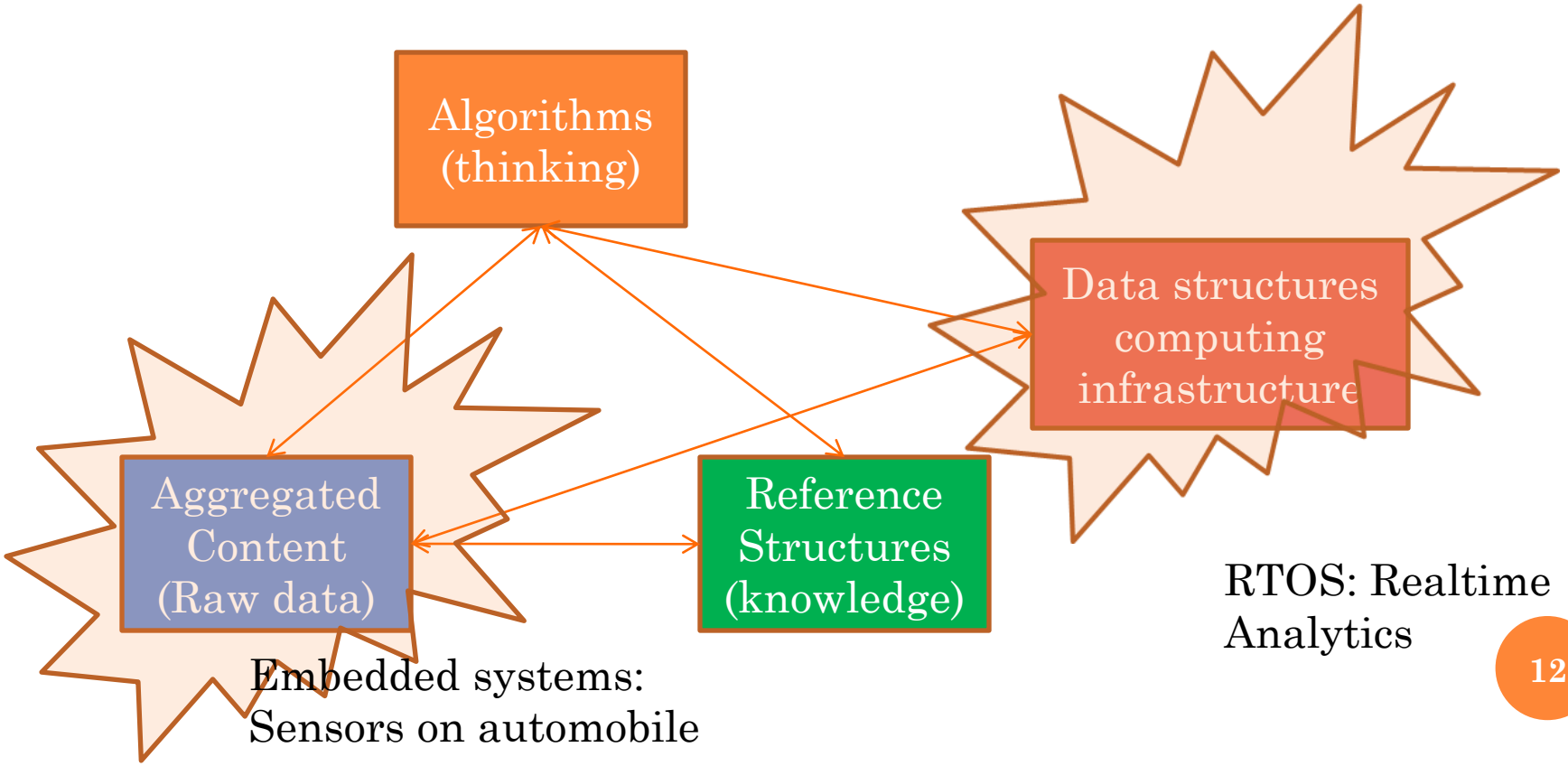
Characteristics of intelligent applications

- Google search: How is different from regular search in existence before it?
  - It took advantage of the fact the hyperlinks within web pages form an underlying structure that can be mined to determine the importance of various pages.
- Restaurant and Menu suggestions: instead of "Where would you like to go?" "Would you like to go to CityGrille"?
  - Learning capacity from previous data of habits, profiles, and other information gathered over time.
- Collaborative and interconnected world inference capable: facebook friend suggestion
- Large scale data requiring indexing
- …

Examples of data-intensive applications

- Search engines
- Recommendation systems:
  - CineMatch of Netflix Inc. movie recommendations
  - Amazon.com: book/product recommendations
- Biological systems: high throughput sequences (HTS)
  - Analysis: disease-gene match
  - Query/search for gene sequences
- Space exploration
- Financial analysis

Relevance of Embedded Systems



Embedded systems:
Sensors on automobile

RTOS: Realtime
Analytics

Basic Elements

- **Aggregated content**: large amount of data pertinent to the specific application; each piece of information is typically connected to many other pieces. Ex:
- **Reference structures**: Structures that provide one or more structural and semantic interpretations of the content. Reference structure about specific domain of knowledge come in three flavors: dictionaries, knowledge bases, and ontologies
- **Algorithms**: modules that allows the application to harness the information which is hidden in the data. Applied on aggregated content and some times require reference structure Ex: MapReduce
- **Data Structures**: newer data structures to leverage the scale and the WORM characteristics; ex: MS Azure, Apache Hadoop, Google BigTable

More intelligent data-intensive applications

- Social networking sites
- Mashups : applications that draw upon content retrieved from external sources to create entirely new innovative services.
- Portals
- Wikis: content aggregators; linked data; excellent data and fertile ground for applying concepts discussed in the text
- Media-sharing sites
- Online gaming
- Biological analysis
- Space exploration

Algorithms

- Machine learning is the capability of the software system to generalize based on past experience and the use of these generalization to provide answers to questions related old, new and future data.
- Data mining
- Soft computing
- We also need algorithms that are specially designed for the emerging storage models and data characteristics.

Different Type of Storage

- Internet introduced a new challenge in the form web logs, web crawler's data: large scale "peta scale"
- But observe that this type of data has an uniquely different characteristic than your transactional or the "customer order" data, or "bank account data" :
- The data type is "write once read many (WORM)" ;
  - **Privacy protected healthcare and patient information;**
  - **Historical financial data;**
  - **Other historical data**
- Relational file system and tables are insufficient.
- Large <key, value> stores (files) and storage management system.
- Built-in features for fault-tolerance, load balancing, data-transfer and aggregation,…
- Clusters of distributed nodes for storage and computing.
- Computing is inherently parallel

Big-data Concepts

- Originated from the Google File System (GFS) is the special <key, value> store
- Hadoop Distributed file system (HDFS) is the open source version of this. (Currently an Apache project)
- Parallel processing of the data using MapReduce (MR) programming model
- Challenges
  - Formulation of MR algorithms
  - Proper use of the features of infrastructure (Ex: sort)
  - Best practices in using MR and HDFS
- An extensive ecosystem consisting of other components such as column-based store (Hbase, BigTable), big data warehousing (Hive), workflow languages, etc.

Data & Analytics

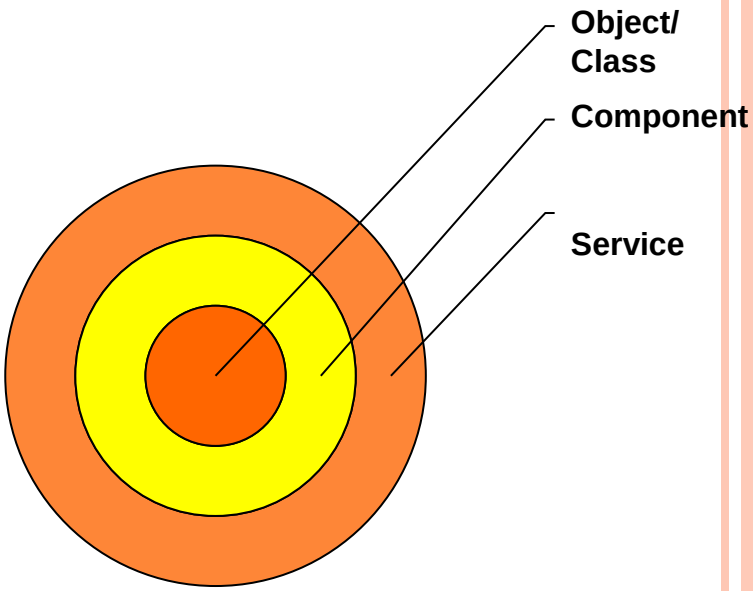
- We have witnessed explosion in algorithmic solutions.
- "In pioneer days they used oxen for heavy pulling, when one couldn't budge a log they didn't try to grow a larger ox. We shouldn't be trying for bigger computers, but for more systems of computers." Grace Hopper
- What you cannot achieve by an algorithm can be achieved by more data.
- Big data if analyzed right gives you better answers: Center for disease control prediction of flu vs. prediction of flu through "search" data 2 full weeks before the onset of flu season! http://www.google.org/flutrends/

Cloud Computing

- Cloud is a facilitator for Big Data computing and is an indispensable in this context
- Cloud provides processor, software, operating systems, storage, monitoring, load balancing, clusters and other requirements as a service
- *Cloud offers accessibility to Big Data computing*
- Cloud computing models:
  - platform (PaaS), Microsoft Azure 
  - software (SaaS), Google App Engine (GAE) 
  - infrastructure (IaaS), Amazon web services (AWS) 
  - Services-based application programming interface (API)

Evolution of the service concept

- A service is a meaningful activity that a computer program performs on request of another computer program.
- Technical definition: A service a remotely accessible, self-contained application module.
- From IBM,



Class, Component and Service

- Class is a core concept is object-oriented architectures. An object is instantiated form a class.
  - Focus on client side, single address space programs.
- Then came the component/container concept to improve scalability and deployability. Ex: EJBs.
  - Focus on server side business objects and separation of resources from code.
- Service came into use when publishing, discoverability, on-demand operation among interacting enterprise became necessity.
  - Focus of enterprise level activities, contracts, negotiations, reservations, audits, etc.

Web Services and the Cloud

- *Web Service* is a technology that allows for applications to communicate with each other in a standard format.
- A *Web Service* exposes an interface that can be accessed through XML messaging.
- A Web service uses XML based protocol to describe an operation or the data exchange with another web service. Ex: SOAP
- A group of web services collaborating accomplish the tasks of an application. The architecture of such an application is called Service-Oriented Architecture (SOA).
- Web service is an important enabling technology of cloud computing: software-as-a-service (SaaS), platform-as-a-service(PaaS), infrastructure-as-a-service (IaaS)

XML to SOAP

- Simple xml can facilitate sending message to receive information.
- The message could be operations to be performed on objects.
- Simple Object Access Protocol (SOAP) or REST

SOAP Request

```
<soap:Envelope
xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/">
  <soap:Body>
    <getProductDetails xmlns="http://warehouse.example.com/ws">
      <productId>827635</productId>
    </getProductDetails>
  </soap:Body>
 </soap:Envelope>
```


SOAP Reply

```
<soap:Envelope xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/">
  <soap:Body>
    <getProductDetailsResponse xmlns="http://warehouse.example.com/ws">
      <getProductDetailsResult>
        <productName>Toptimate 3-Piece Set</productName>
        <productId>827635</productId>
        <description>3-Piece luggage set.  Black Polyester.</description>
        <price>96.50</price>
        <inStock>true</inStock>
      </getProductDetailsResult>
    </getProductDetailsResponse>
  </soap:Body>
</soap:Envelope>
```


SOAP→Web Services (WS)

- Read this paper:

http://www.w3.org/DesignIssues/WebServices.html

- Lets look at some WScode:

REST-based Web services

- Representational State Transfer (REST)
- Ph.D. thesis by Roy Fielding, who was the chairman of the Apache Software Foundation (not anymore)
- We will use REST-based WS for our projects.
- We will discuss basics of REST next classes.
- 8/30/2011: Reading Assignment: Front material, and up to page 44 of the Fourth Paradigm text.
- Review your Java skills, install your favorite IDE (Eclipse, netbeans etc.), J2EE (helios), any design tool.

Summary

- We are entering a watershed moment in the internet era.
- This involves in its core and center, big data analytics and tools that provide intelligence in a timely manner to support decision making.
- Newer storage models, processing models, and approaches have emerged.
- Embedded systems and RTOS play a critical role supporting big data analytics.
- The cloud is indispensable for Embedded Systems and RTOS.