Chapter 7: The Future of Data Mining, Warehousing, and Visualization

Modern Data Warehousing, Mining, and Visualization: Core Concepts

by George M. Marakas

http://alainmaterials.webs.com/

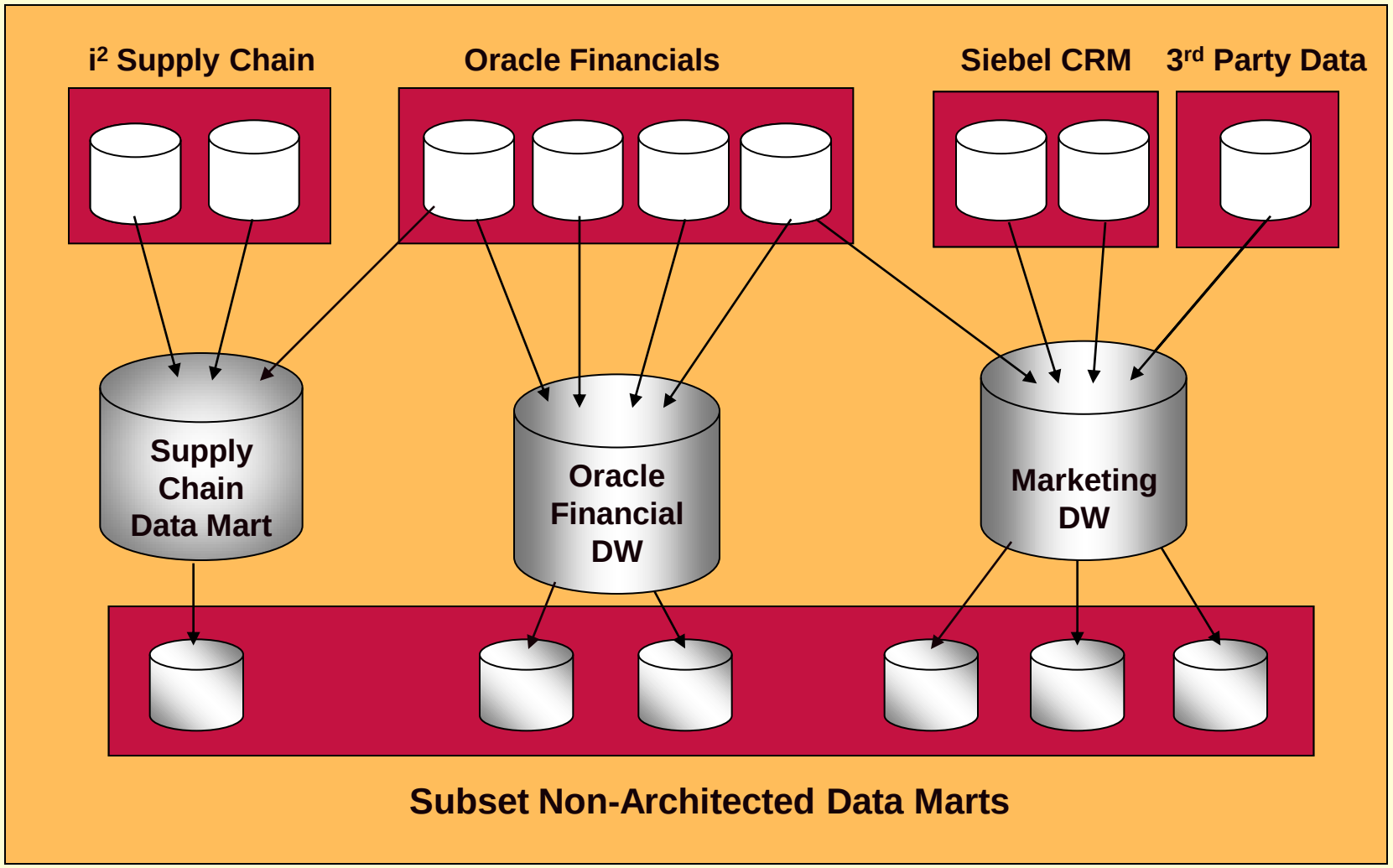
7-1: The Future of Data Warehousing

- As a DW becomes a mature part of an organization, it is likely that it will become as “anonymous” as any other part of the IS.
- One challenge to face is coming up with a workable set of rules that ensure privacy as well as facilitating the use of large data sets.
- Another is the need to store unstructured data such as multimedia, maps and sound.
- The growth of the Internet allows integration of external data into a DW, but its unstable quality is likely to lead to the evolution of third-party intermediaries whose purpose is to rate data quality.

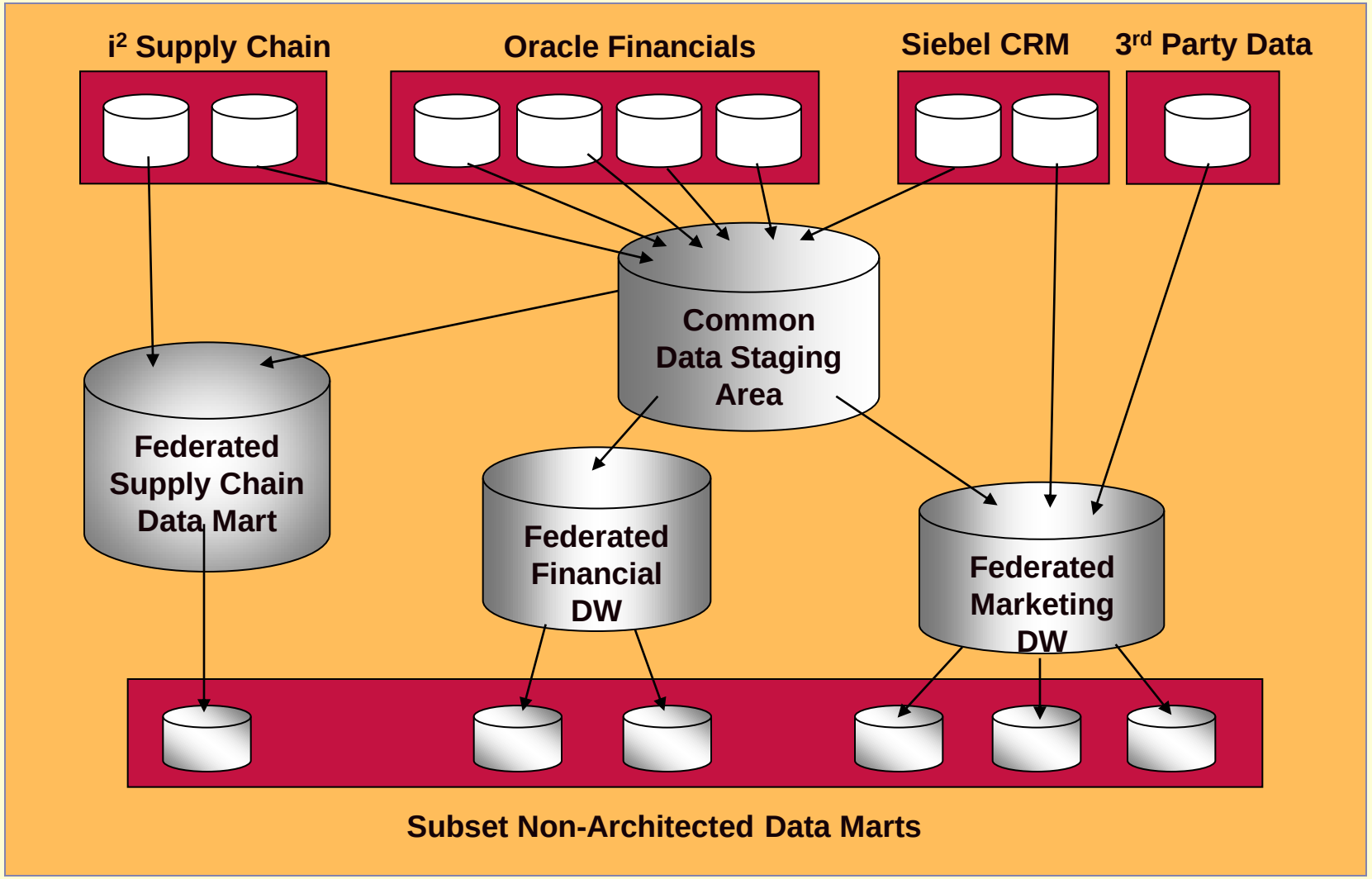
Integrated Architecture

- Historically, market and business forces have moved organizations toward ineffective nonintegrated DW systems (next slide).
- To survive in a future world of low-cost, turnkey application systems, the transition to a federated architecture (two slides ahead) must be made.

Typical Nonintegrated Information Architecture



Federated Integrated Information Architecture



7-2: Alternate Storage and the Data Warehouse

- Surprisingly, the future of data warehousing is not high-performance disk storage, but an array of alternative storage.
- This involves two forms of storage. Near-line storage involves an automated silo where tape cartridges are handled automatically.
- Secondary storage is slower and less expensive, such as CD-ROMs and floppy disks.

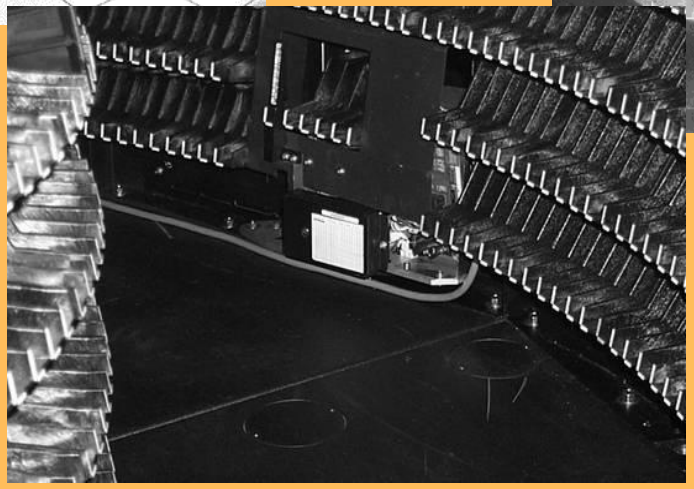
Speed and Capacity of Various Near-Line Storage Media

Device	Capacity	Data Access Speed	Media Lifetime	Write once or Write many
DAT DDS2	4-8 Gbyte	510 Kbyte/s	10-25 Yrs	WM
DAT DDS3	12-24 Gbyte	1 Mbyte/s	10-25 Yrs	WM
CD-ROM	640 Mbyte	X times 1.5 Mbits/s to Read	10 Yrs Plus	WO
CD-RW	640 Mbyte	X times 1.5 Mbits/s to Read	10 Yrs Plus	WM
Exabyte	20-40 Gbyte	3-6 Mbyte/s	10-25 Yrs	WM
DLT	35 Gbyte	5 MByte/s	30 Yrs	WM
DVD	up to 15Gbyte	Not Known	Not Known	WO
DTF	42Gbyte	12 Mbyte/s	10-25 Yrs	WM
Data D3	50 Gbyte	12 Mbyte/s	10-25 Yrs	WM
DVD-RAM	up to 3 Gbyte	Not Known	Not Known	WM
Magneto-optical	2.6-5.6 Gbyte	Not Known	Not Known	WM

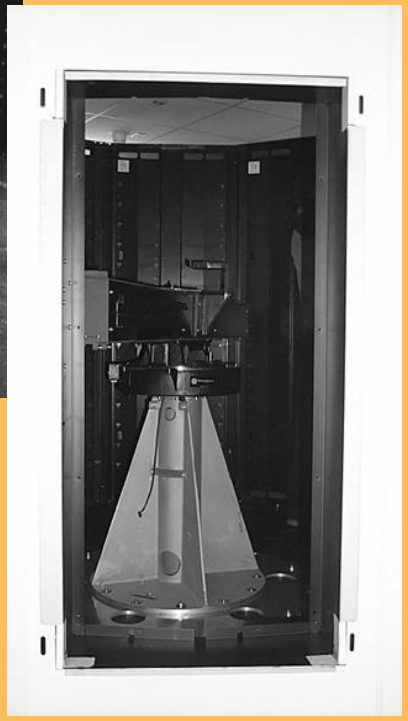
Typical Near-Line Tape Storage Silo

Main View of the Tape Silo

Robotic Tape Retrieval Arm



Close-up of tape storage carousel

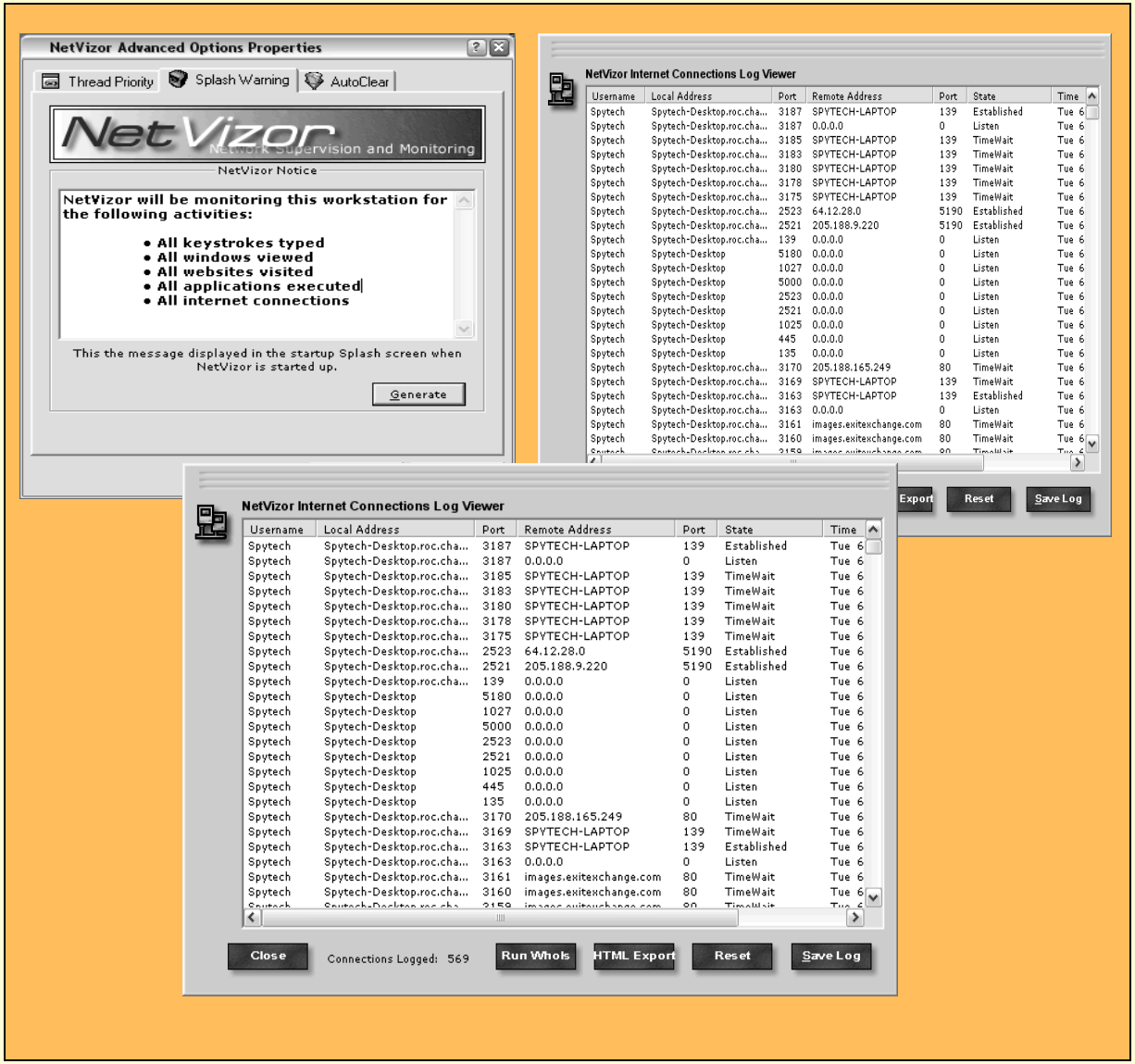


View Through Silo Entry Door

Why Use Alternative Storage?

1. The data in a DW are stable. They are placed there once and left alone, so do not need to be updated at high speed.
2. The queries that operate on the DW data often require long streams of data stored sequentially. Operational access requires different units of data from different storage areas.
3. The DW is of indeterminate size and is always increasing in volume, requiring flexible capacity.
4. When data gets accessed less often as it ages, it can be moved to secondary storage, making access to newer data more efficient.

To make this two-level storage work, we need both an activity monitor (shown here) and a cross media storage monitor.



7-3: Trends in Data Warehousing

- Customer interaction and learning relationships require capturing information “everywhere” and massive scalability.
- Enterprise applications generate data that is doubling very 9-12 months.
- The time available for working with data is shrinking and the need for 24×7 access is becoming the norm.
- Fast implementation and ease of management are becoming more and more important.
- In the future, more organizations will build Web applications that operate in conjunction with the DW.

7-4: The Future of Data Mining

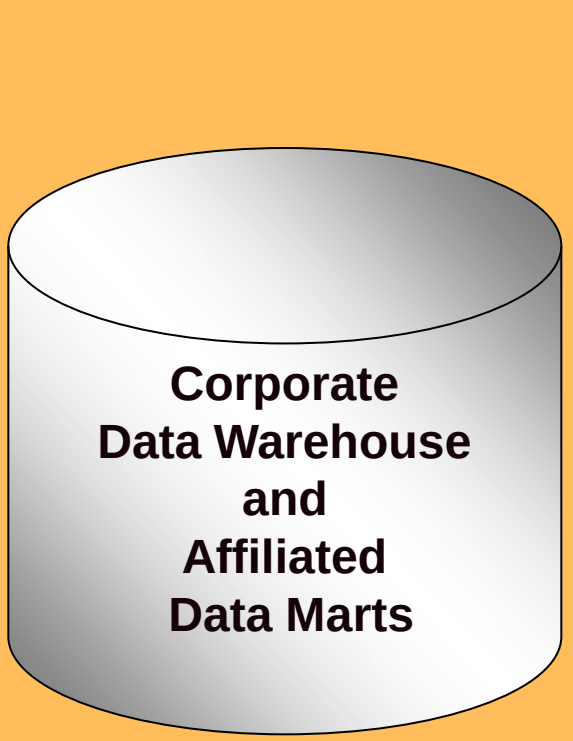
As promising as the field may be, it has difficulties:

- The quality of data can make or break the data mining effort.
- In order to mine the data, companies first have to integrate, transform and cleanse it.
- To obtain value from data mining, organizations must be able to change their mode of operation and maintain the effort.
- Finally, there are concerns about privacy.

Personalization versus Privacy

- Companies that use data mining for target marketing walk a tightrope between personalization and privacy.
- Further, technology appears to create new ways to acquire information faster than the legal system can handle the ethical and property issues.
- Nonetheless, many view information as a natural resource that should be managed as such.

Concept of Personal and Corporate Information as a National Resource



7-5: Using Data Mining to Protect Privacy

- While Internet use has grown, so have the problems of network intrusion.
- One current intrusion detection technique is misuse detection – the Intrusion Detection System (IDS) analyzes the information it gathers and compares it to large databases of attack signatures.
- Another technique is anomaly detection where there is an attempt to identify malicious activity based on deviations from norms. the system administrator defines the baseline, or normal, state of the network??s traffic load, breakdown, protocol, and typical packet size. The anomaly detector monitors network segments to compare their state to the normal baseline and look for anomalies.
- Most intrusion detection systems operate by the signature approach.

Shortfalls of Current Detection Schemes

- Variants – although signature lists are updated frequently, minor changes in the exploit code can produce a “new” intruder.
- False positives – a detection system may be too traditional and declare an intrusion when there is none.
- False negatives – an intrusion won’t be detected until a signature has been identified.
- Data overload – as traffic grows, the ability to find new hacks becomes harder and harder.

How Can Data Mining Help?

Data mining can help mainly by its ability to identify patterns of valid network activity.

- Variants – anomalies can be detected by comparing connection attempts to lists of know traffic.
- False positives – data mining can be used to identify returning patterns of false alarms.
- False negatives – if valid activity patterns are identified, invalid activity will be easier to spot.
- Data overload – data reduction is one of the major features of data mining.

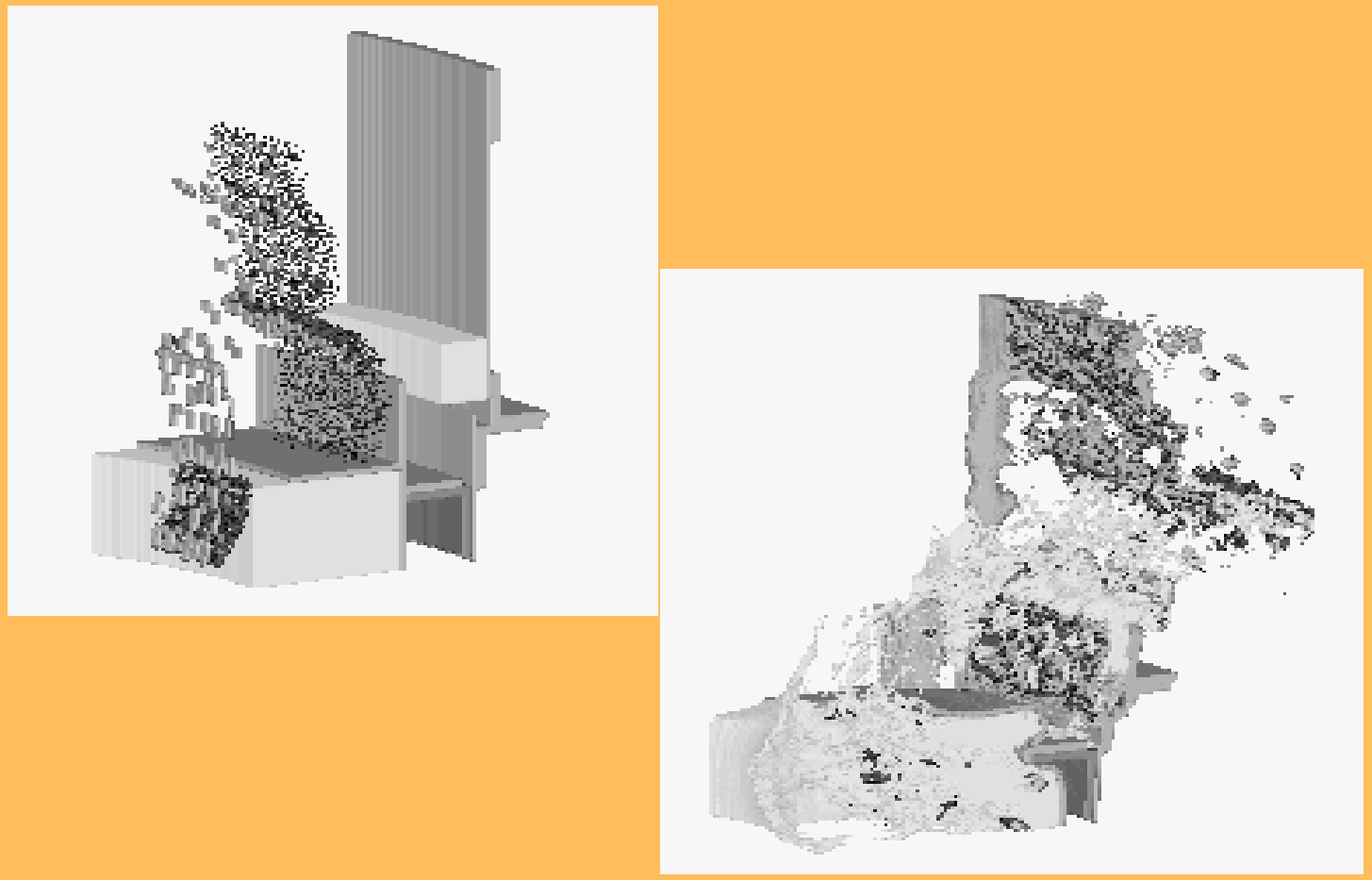
7-6: Trends Affecting the Future of Data Mining

- While the available data increases exponentially, the number of new data analysts graduating each year has been fairly constant. Either of lot of data will go unanalyzed or automatic procedures will be needed.
- Increases in hardware speed and capacity makes it possible to analyze data sets that were too large just a few years ago.
- The next generation Internet will connect sites 100 times faster than current speeds.
- To be more profitable, businesses will need to react more quickly and offer better service, and do it all with fewer people and at a lower cost.

7-7: The Future of Data Visualization

- Weapons performance and safety – data visualization coupled with simulation models can show how weapons perform under typical conditions and the effect of weapons aging.
- Medical trauma treatment – today's surgeons use computer vision to assist in surgery. In the future this trend suggests that local medical personnel can also be assisted from afar by specialists through telepresence.

Visualization of a Simulated Warhead Impact



Augmented-reality Headset Worn by Surgeon

Surgery Being Conducted Via Telepresence

7-8: Components of Future Visualization Applications

- The data visualization environment links the critical components and enables the smooth flow of information among the components.
- In the future, the bounds between computers, graphics and human knowledge will become more blurred.
- Many advances in technology will be need to handle the visualization environment of the future. Intelligent file systems and data management software will contend with thousands of coupled storage devices.

Conceptual Mapping of an Information Architecture