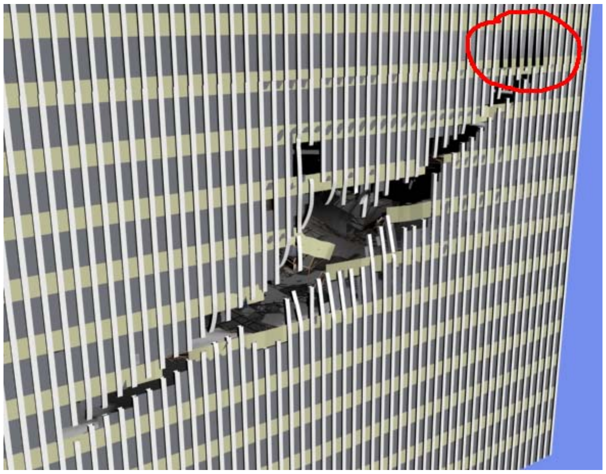
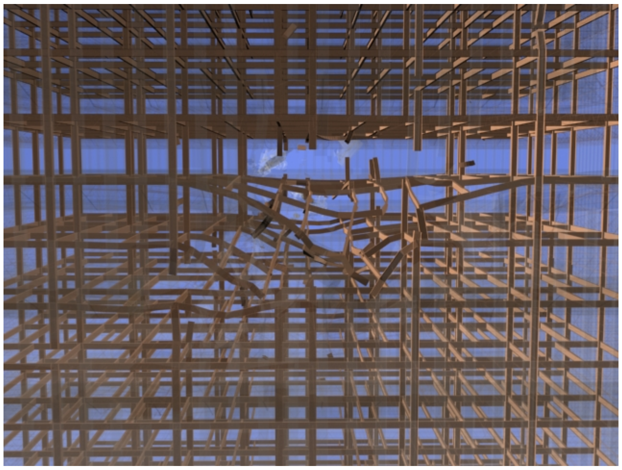
Visualization, Data Analysis and Management

Christoph Hoffmann, Computer Science, Purdue University, November 2007

Visualization is a powerful medium for communication as well as an efficient vehicle for apprehending data patterns and meaning. In the latter category an example is given by visualizing a large-scale mesh, here of the World-Trade Center (left picture). The section of the façade, in the upper right of the picture, appears to be bent. This turns out to be an orientation error of a few beam elements among several thousand. It is hard to imagine a simpler, more efficient way to locate such errors other than by visualization using sophisticated rendering techniques.





The persuasiveness of imagery entails the danger that a naïve viewer could assume that the image faithfully represents reality. In simulations that may not necessarily be the case, especially when the simulation involves explosions or material fractures. An example is shown here (right image), of a simulation of the impact of flight AA-11 on the North Tower. The core damage shown, and based on the simulation, should not be taken as reality. Fracture and explosions are chaotic events that the simulation illustrates only qualitatively.

Today's science research increasingly deals with data sets whose size is growing well beyond the ability to visualize them. Such data sets may originate from observation (single-point or distributed acquisition), as well as from computations. Several examples illustrate this trend.

Flow Cytometry is a technology for scanning blood cells and can be used to diagnose certain kinds of cancer. In flow cytometry it is possible to scan millions of cells, each generating a spectrum of 10 to 20 data values characterizing the cell. The high dimensionality of the resulting space precludes using visualization in a straightforward way, and a pre-filter is needed to cluster the data and reduce the number of attributes examined visually.

The **Compact Muon Solenoid (CMS) Experiment** is to start in the Spring of 2008 at CERN near Geneva, Switzerland. It will generate daily approximately 10 TB of data from high-energy particle collisions. The experiment is designed to search for the Higgs Boson, and on theoretical grounds it has been predicted that only 1 in about 10^6 measurements generated by the experiment is "interesting." Note that each such measurement is an image, yet the sheer number of images overwhelms and a globally distributed network of thousands of computers is ready to examine the images and reliably filter out “uninteresting” ones.

The **Large Synoptic Survey Telescope (LSST)** is an 8.4m telescope under construction that is to see first light in 2013. The telescope takes images with a resolution of approximately 3.6G pixels. After image correction compensating for CCD characteristics and normalization, the telescope is to generate 30TB of useful data nightly for 10 years. More than that, images are to be scanned for variable objects, such as super novae in other galaxies, and automatic alerts are to be generated within 60 seconds from taking the image. The data volume expected poses challenging problems storing and cataloguing the acquired data.

Other examples can be cited that illustrate the current trend towards huge data sets. Many are beyond the current state of the art in database technology. The trend suggests that future experimental and theoretical science research rests on a data pyramid whose massive base requires automated processing and filtering of the data into a sufficiently condensed form so that insights can be gleaned by humans. These data scanning and distillation techniques are expected to become an integral part of science research and argue for the need to teach these computational techniques in addition to a core set of visualization methods.