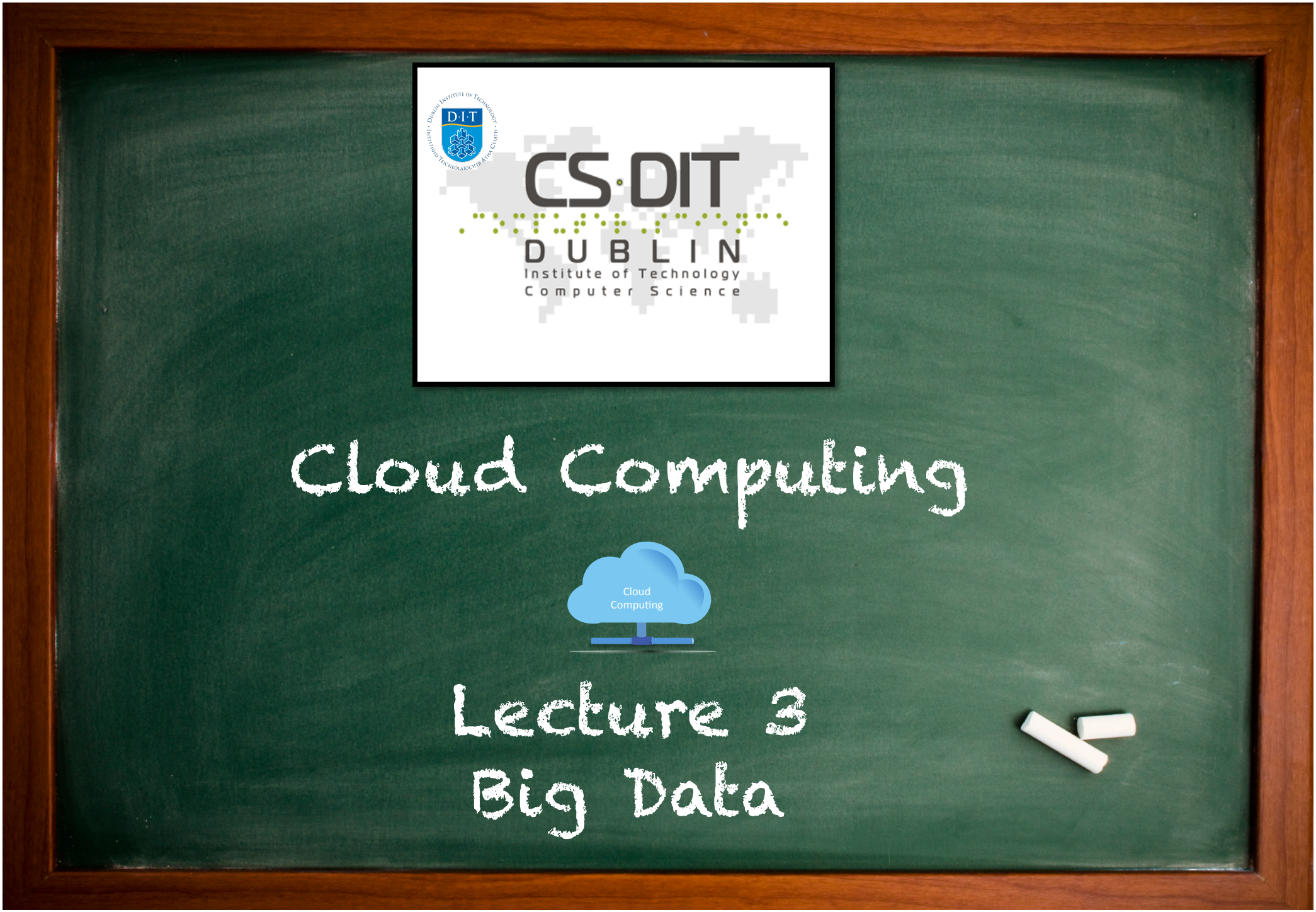

CS·DIT
DUBLIN
Institute of Technology
Computer Science
Cloud Computing
Cloud
Computing
Lecture 3
Big Data

Lecture Contents

- The “Problem” with Science
- Astronomy and Big Data
- Digital Processing process
- How Much Data?
- Visualizing big data
- Capture vs. Analysis
- Challenges

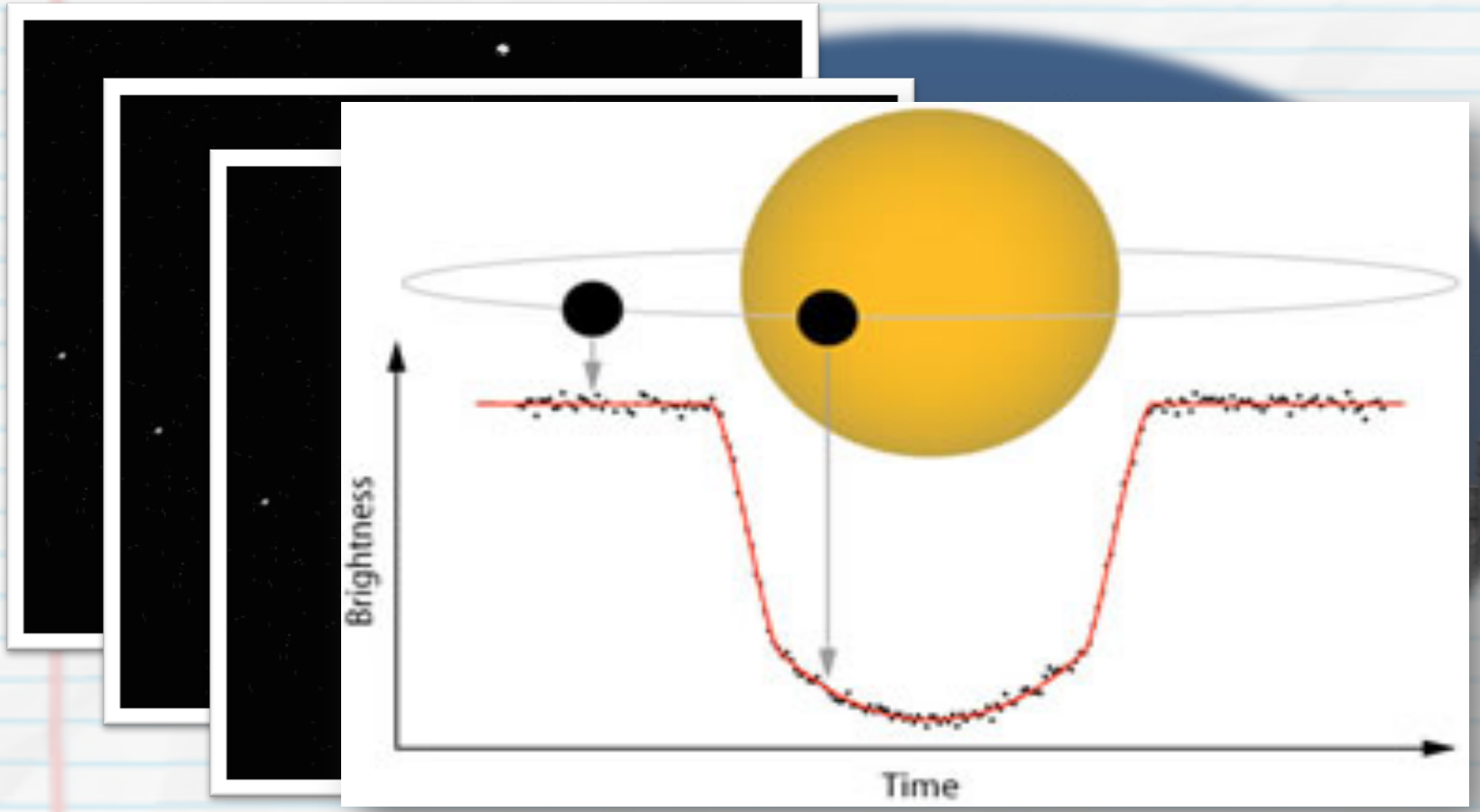
The "*Problem*" with Science

- Scientific Method:
  - Evidence Based, which requires data
  - Theory Based, which requires prediction
  - Model Based, which requires computation

The "*Problem*" with Science

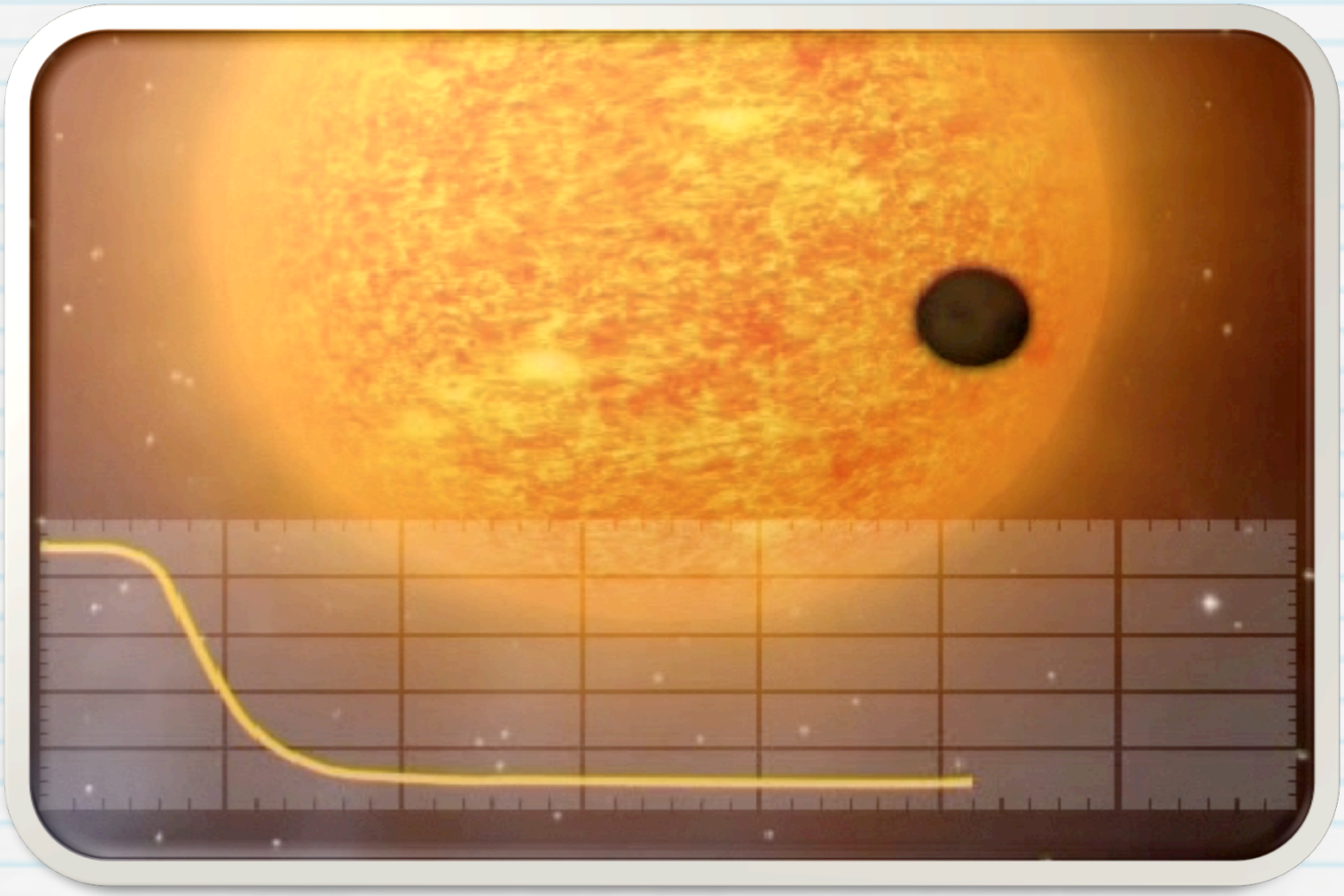
- Role of Computer Science:
  - Data Storage and retrieval
  - Data Management and transport
  - Data Processing and Data Modeling

Astronomy Example:
But Why More Data?



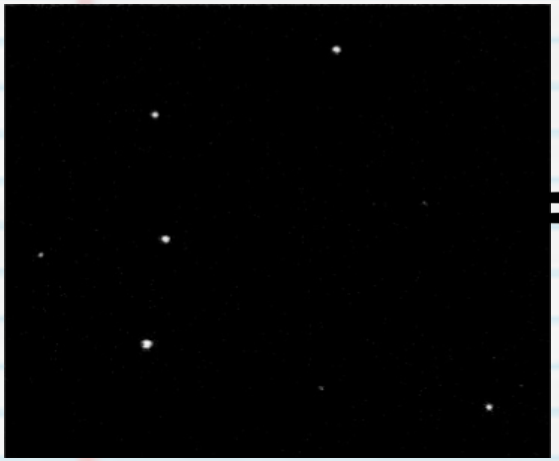
When you measure amount of light accurately
Enough you can find Exo-planets (transit method)

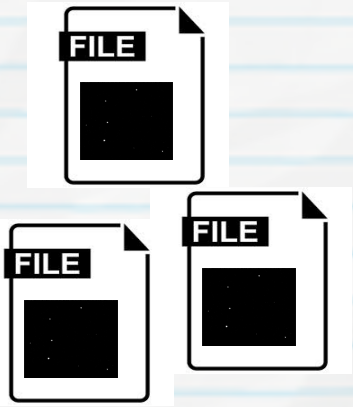
Extra solar planet hunting

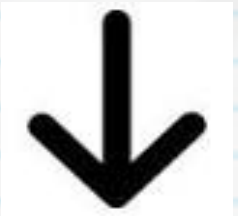


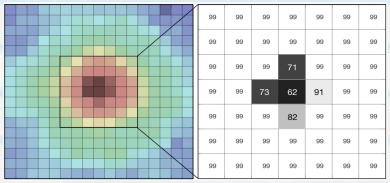
Digital Photometry: Basic Process

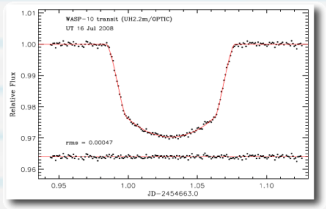
Typical flow, data capture to data processing

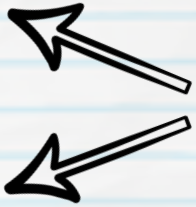




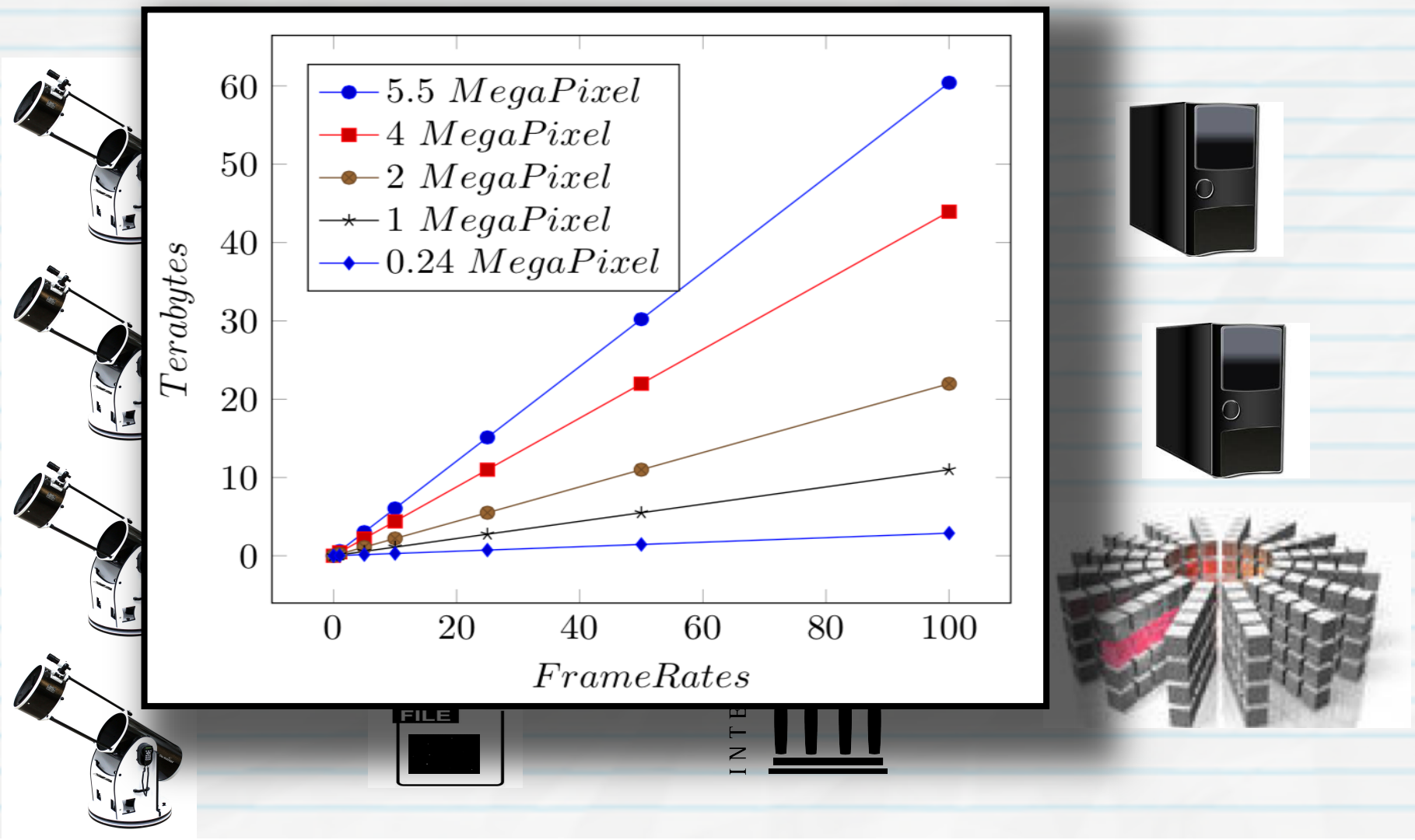








Victims of our own success



How much data?

Megabytes

Gigabytes

Terabytes

Petabytes

Exabytes

How much data?

- Google processes 100 PB a day
- Netflix has 3.14 PB Master copy of images)
- Facebook has 357 PB of user data
- eBay has 100 PB of user data
- CERN's Large Hadron Collider (LHC) generated 200 PB looking for the Higgs Boson
- German Climate Computing Centre stores 60 PB of climate data.

Can you visualise a large number?

1 Byte

Bits and Bytes

10001001

1,000 Pennies or 1 Kilobyte

bio - Notepad

File Edit Format View Help

Paul is currently lecturing and performing research in Dublin Institute of Technology in the area of desktop and server virtualisation deployments. He has been responsible for running multiple virtualisation and thin client pilot projects within the DIT. From 1993 to 2003 he worked at Sun Microsystems, and was responsible for the development of Thin Client and Blade Server technology.

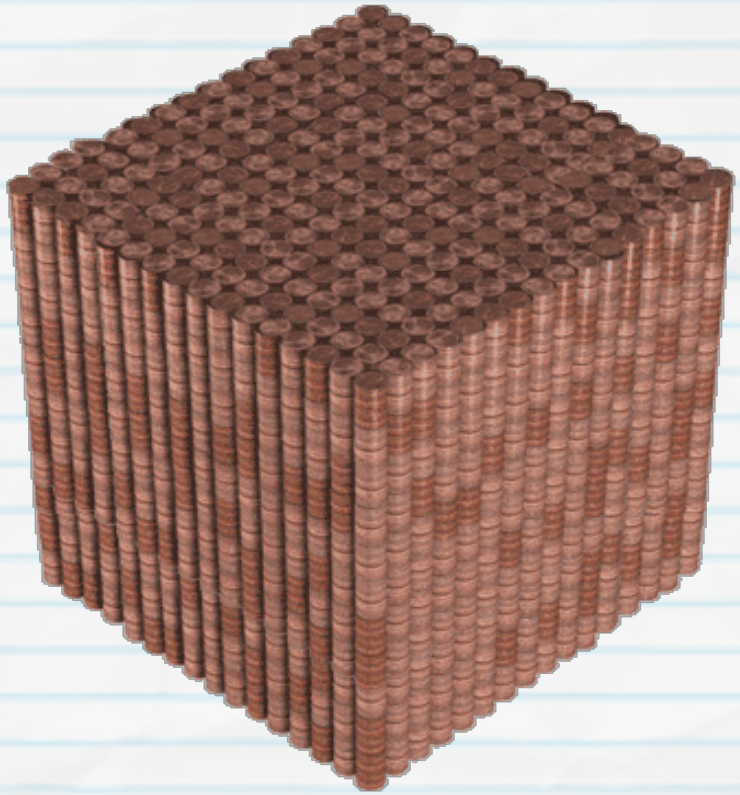
Talking about
With recent advances in virtualisation technology, educational institutes are starting to move towards server and desktop virtualisation based infrastructures. Virtualisation promises to provision a server's resources more efficiently, increase hardware utilisation and ultimately reduce costs.

While server virtualisation provides many advantages, desktop virtualisation is often complicated. Within the DIT, virtualisation has been in use for many years. This presentation discusses the issues of the larger deployments and reviews the lessons learned.

Paul Doyle BSc. MSc. is the Chairman of the Ubiquitous Computing Research Group and was born in Dublin, Ireland. He completed his undergraduate studies in Dublin City University, followed by an MSc. in Metrics for 4GLs. He has 20 years of industrial experience and has worked at Sun Microsystems as a senior manager for over 10 years, and was responsible for the research and development of the SunRay thin client appliance. Currently he is a member of the academic lecturing staff at Dublin Institute of Technology, in the School of Computing.

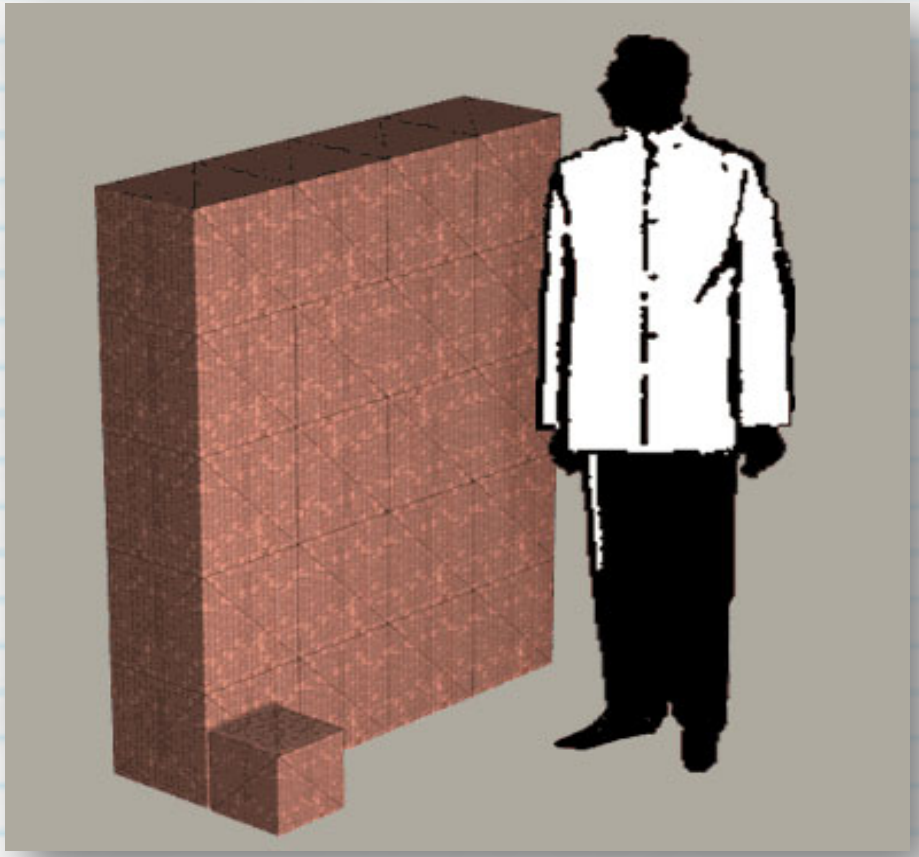
Currently pursuing a PhD in large scale thin client and virtualisation technologies. His research interests include predictive performance benchmarking of large scale virtualisation platforms, thin client performance and reliability, session mobility, and very small virtual machine (VSVM). Additional activities include supervision of PhD and MSc. projects in the area of virtualisation.

50,000 Pennies or 50 Kilobytes



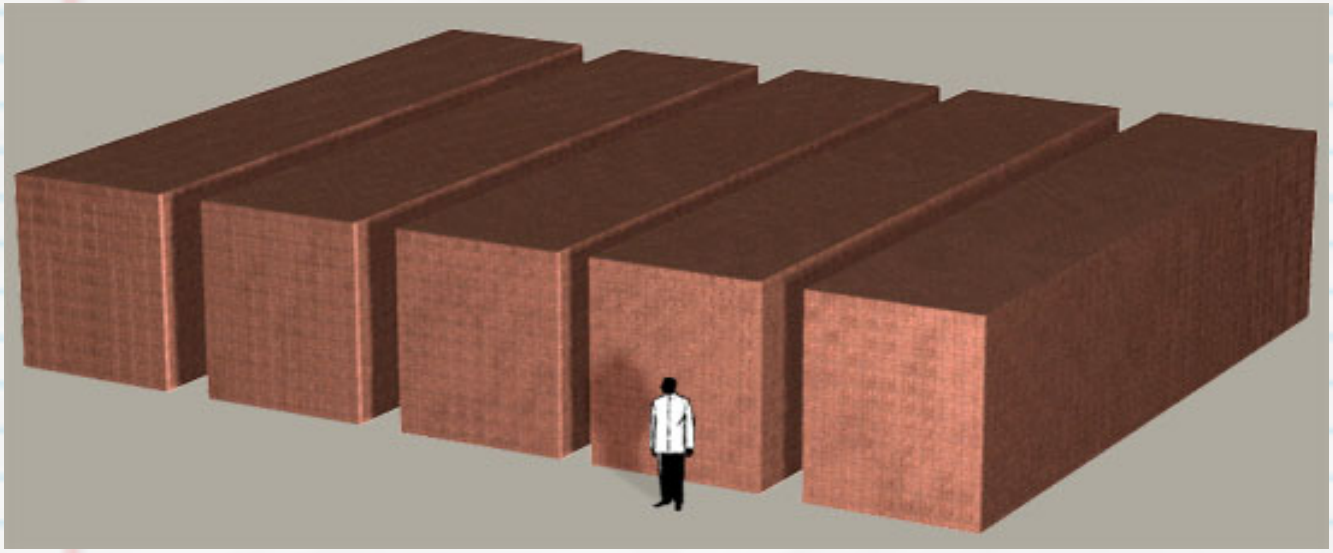
Low Resolution Image

1MB
1,000,000 Pennies

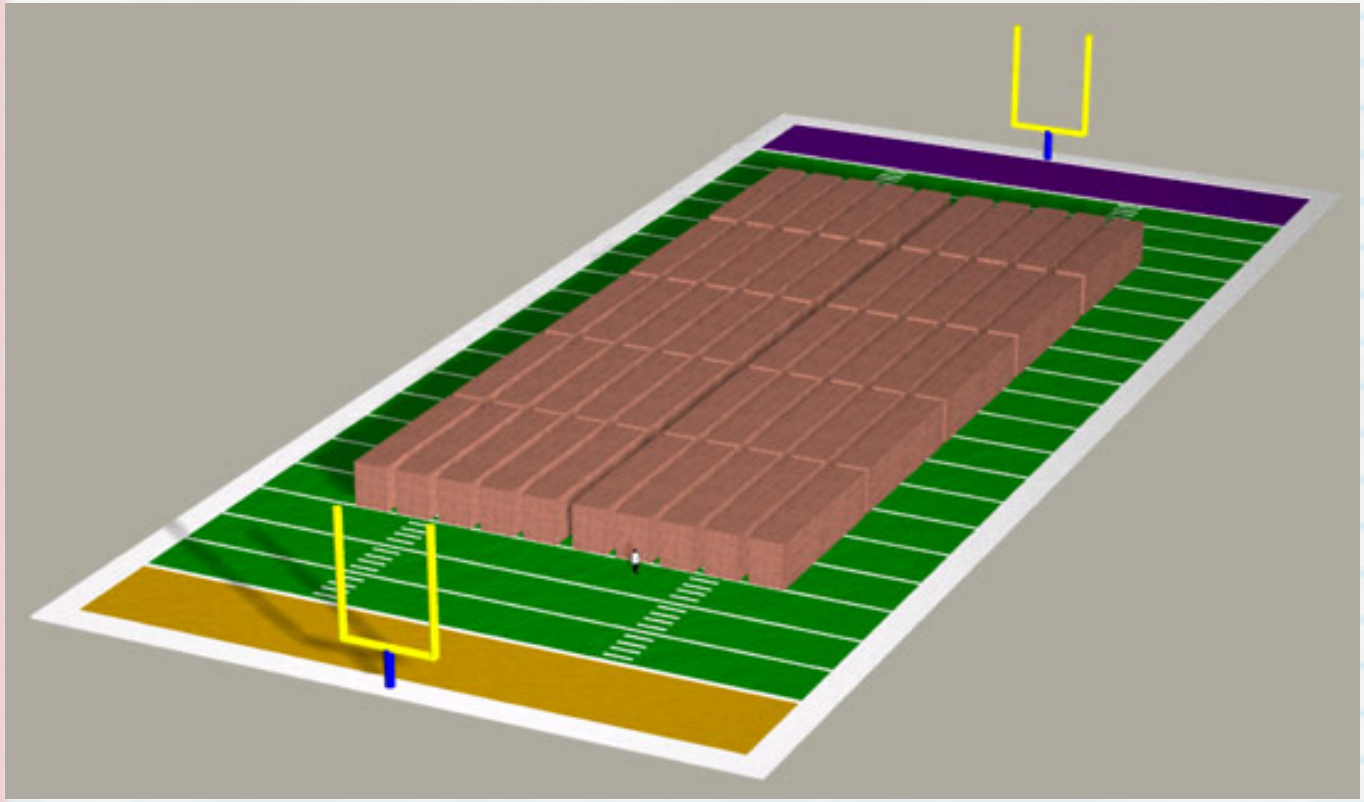


1MB transfer of 10 GBit/s link
< 1 Second

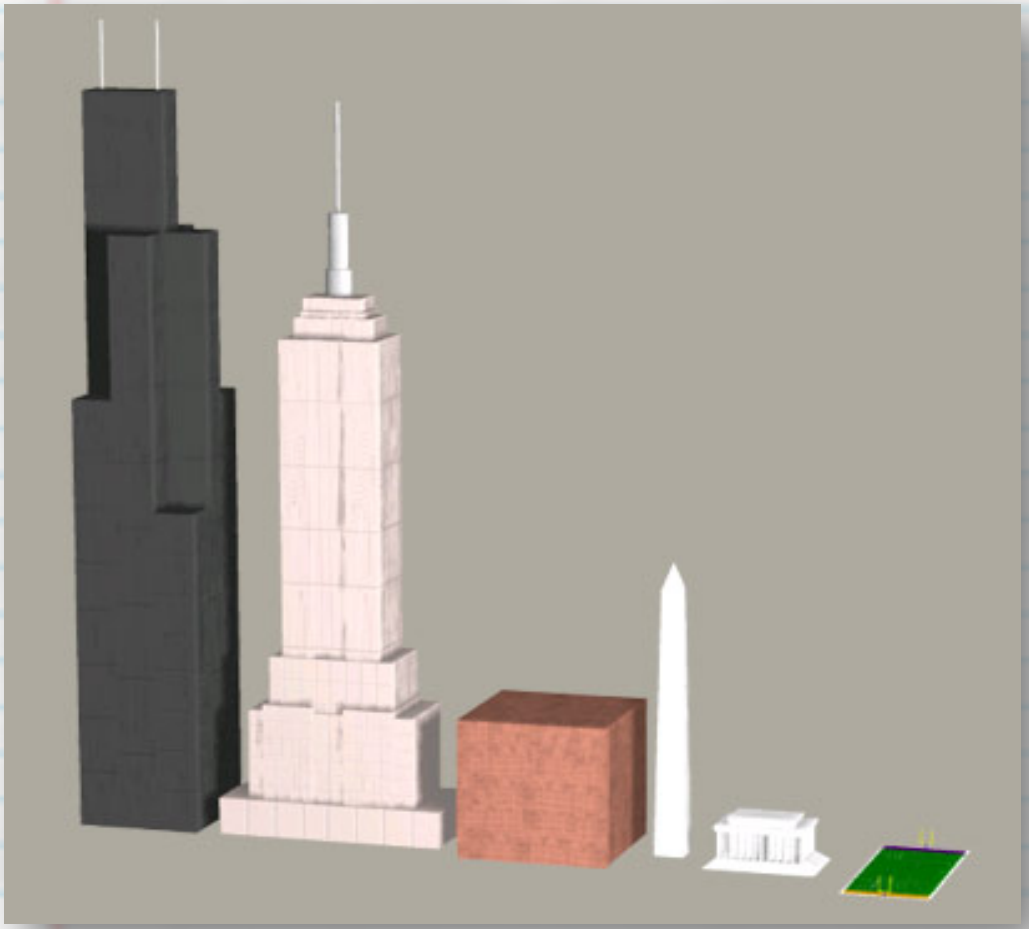
1,000,000,000 Pennies (billion)
1 Gigabyte



10,000,000,000 Pennies (billion)
10 Gigabytes



1TB
1,000,000,000,000 Pennies

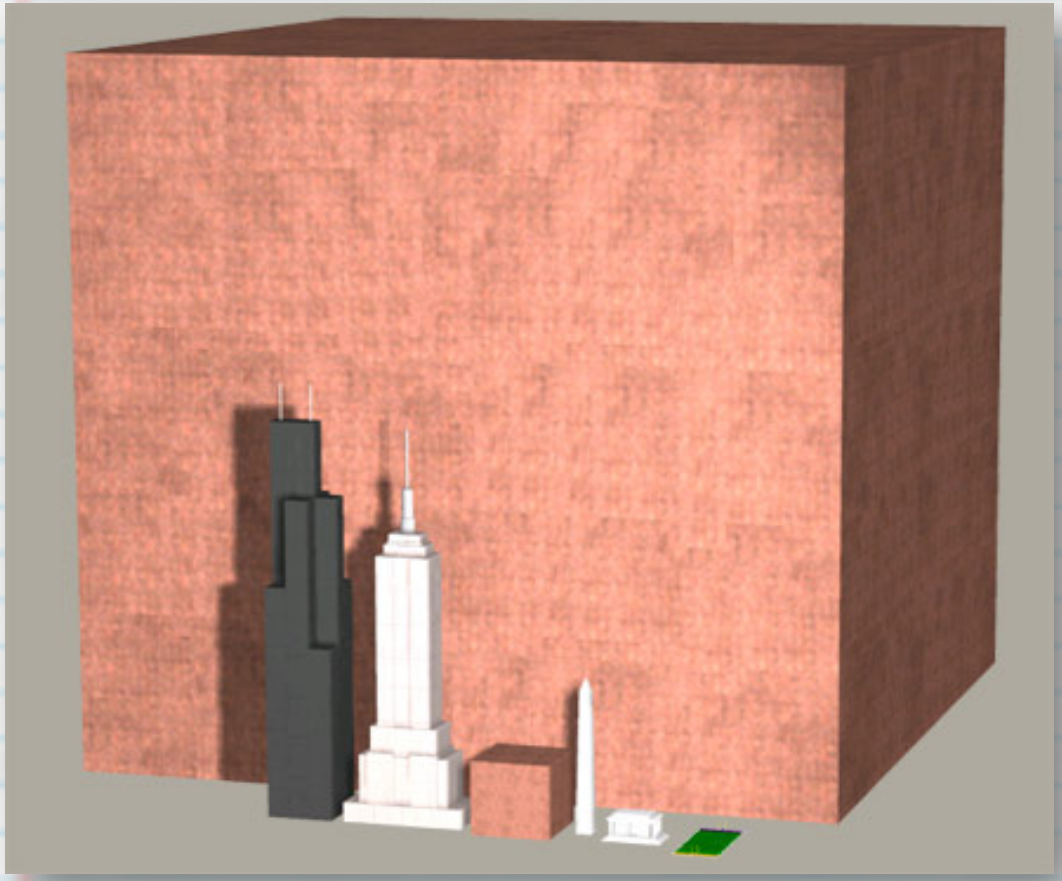


HD

400Hrs

1TB transfer of 10GBit/s link
~ 20 minutes

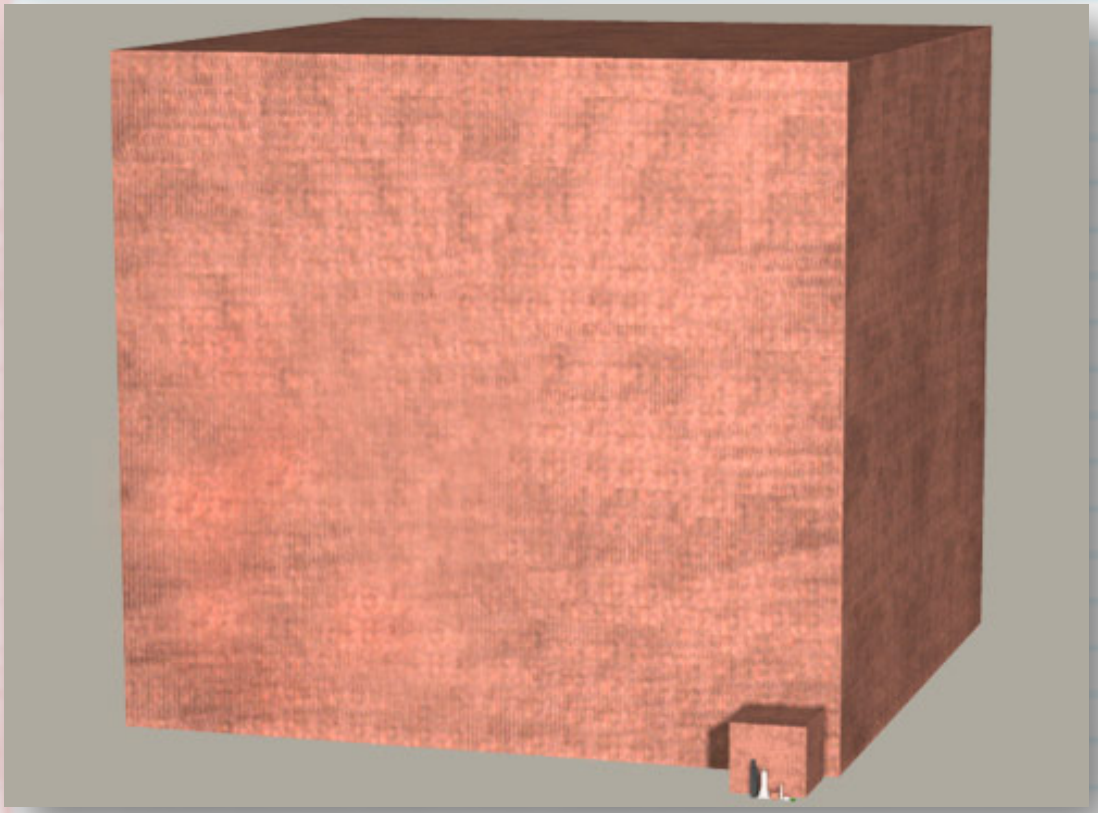
1PetaByte
1,000,000,000,000,000 Pennies
(1 Quadrillion)



5 Billion
Photos

1PB transfer of 10 GBit/s link
~ 11 days

1ExaByte
1,000,000,000,000,000,000 Pennies
(1 Quintrillion)



All words ever spoken
(~ 100 Billion People ever lived)

1 ExaByte transfer of 10GBit/s link
~ 25 years

Capture rate vs Analysis rate

1. Speed to read/write to a disk
2. Speed to process data
3. Speed to analysis data
4. Speed to transfer data on a network
5. Ability to store large volumes of data
6. Ability to Search large volumes of data
7. Elastic capacity
8. Keeping things secure

Big Data

- Data is generated by taking lots of measurements & recording events
- Sensors such as
  - Images (Pictures, Telescopes (CCD))
  - Temperature
  - Location
  - Time
  - Speed
  - Context

Challenges

- Large Data will be measured in Exabytes no Gigabytes
- Data won't fit on a single device or datastore
  - Seagate in June 2014 announced an 8 TB HD
- Data needs to be processed to become useful
- Data will naturally be more distributed
  - Movement of data
  - Processing of data
  - Storage of data
  - Security of data
  - Searching Data

Course notes, video lecture and podcast version available at
www.dit.ie/webcourses