

## Out-Googleing Google:

*Using the Grid to develop a Next-Generation Search Engine*



- *Introduction to Camtology*
- *Imense Image Searching*
- *iLexIR Ontological Searching*
- *The n-gram Web Spider*

Camtology is a joint venture between two Cambridge start-up companies, *Imense and iLexIR*

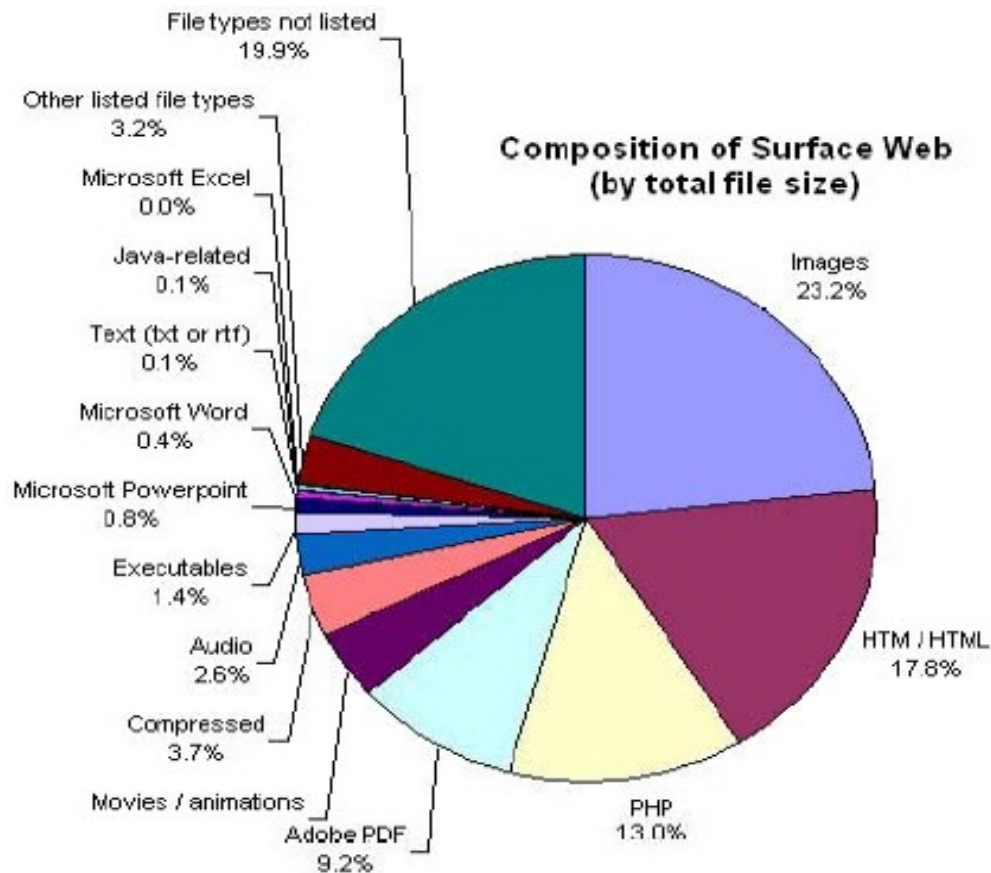
# Camtology



Both companies have benefited from the use of GridPP resources and PIPSS/mini-PIPSS funding to develop the *next generation of search engine*

Imense Ltd. is a spin off company from Cambridge University started by *David Sinclair and Chris Town*

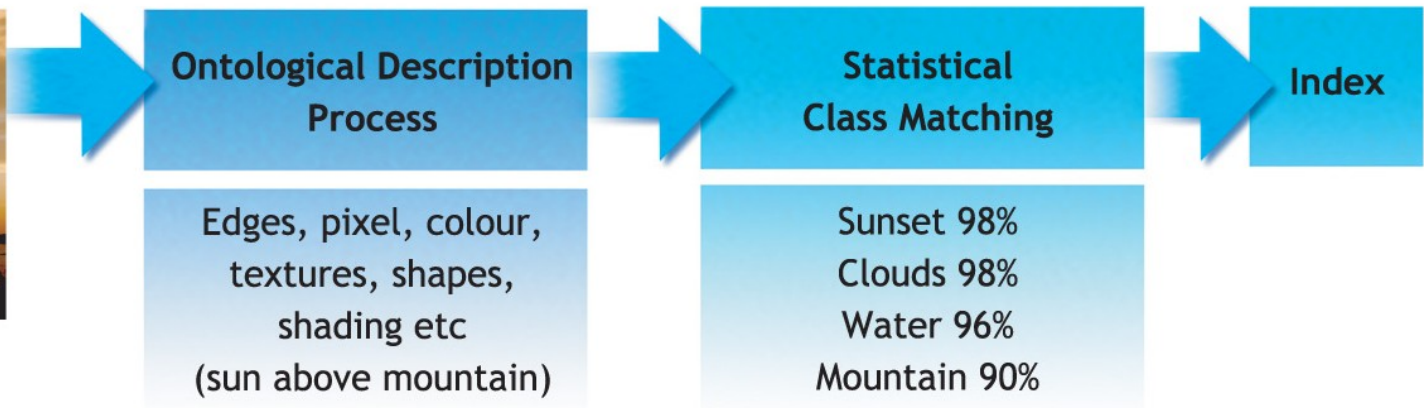
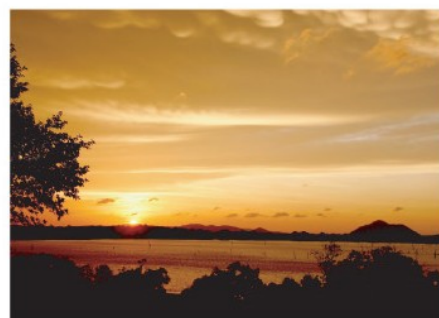
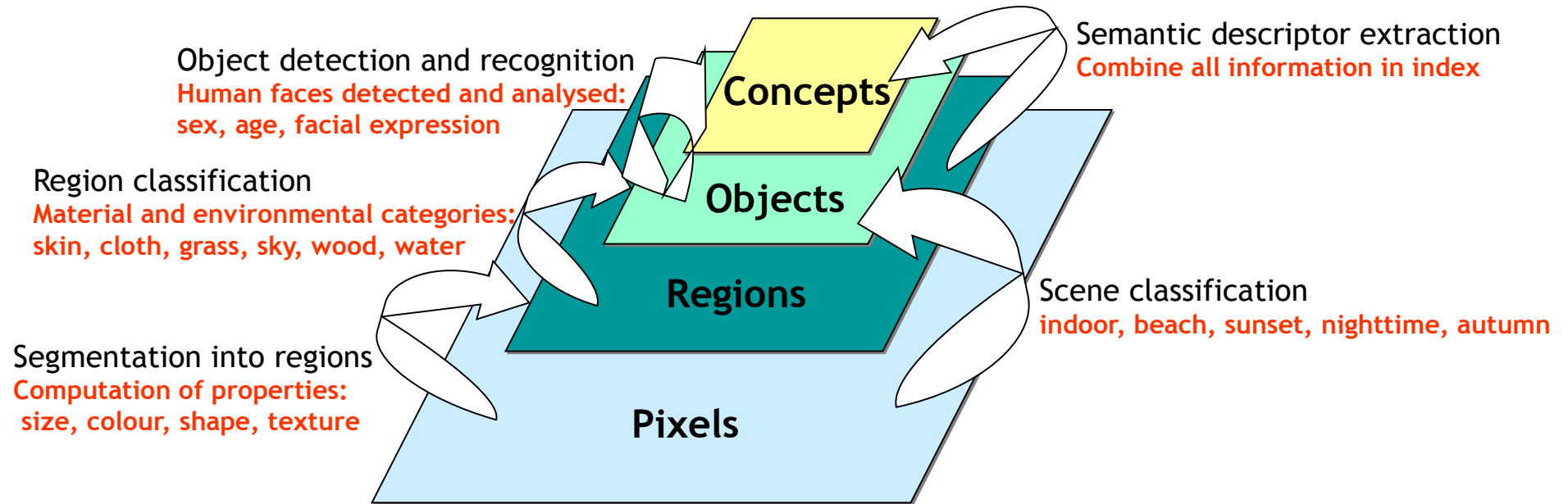
It's principal aim is to provide a context based image search engine:



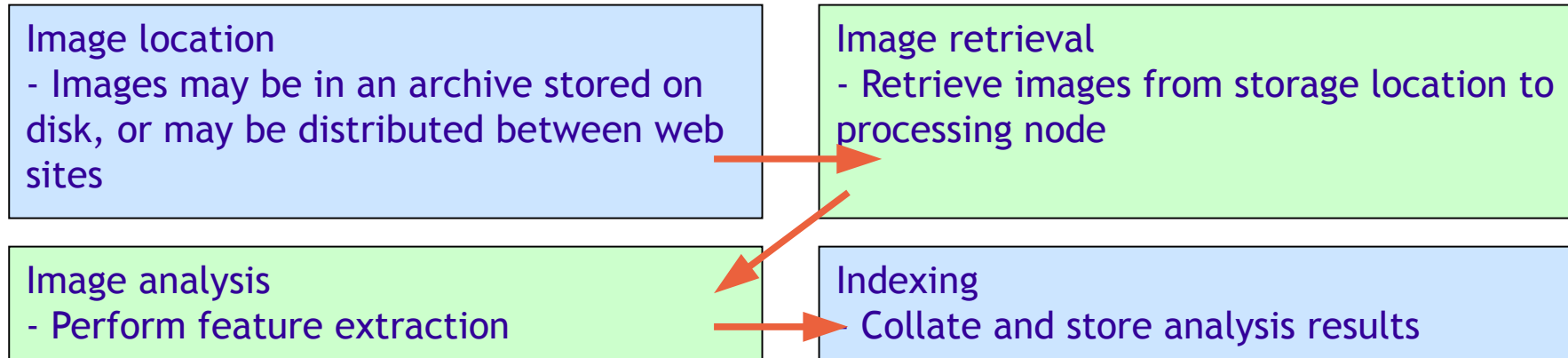
*Images make up a large fraction of the internet*

*Currently however, finding a picture requires accurate meta-data or surrounding text*

*Search for 'cloudy sky without people' in Google and you will not get what you wanted!*



Four basic steps to enabling searches based on image content



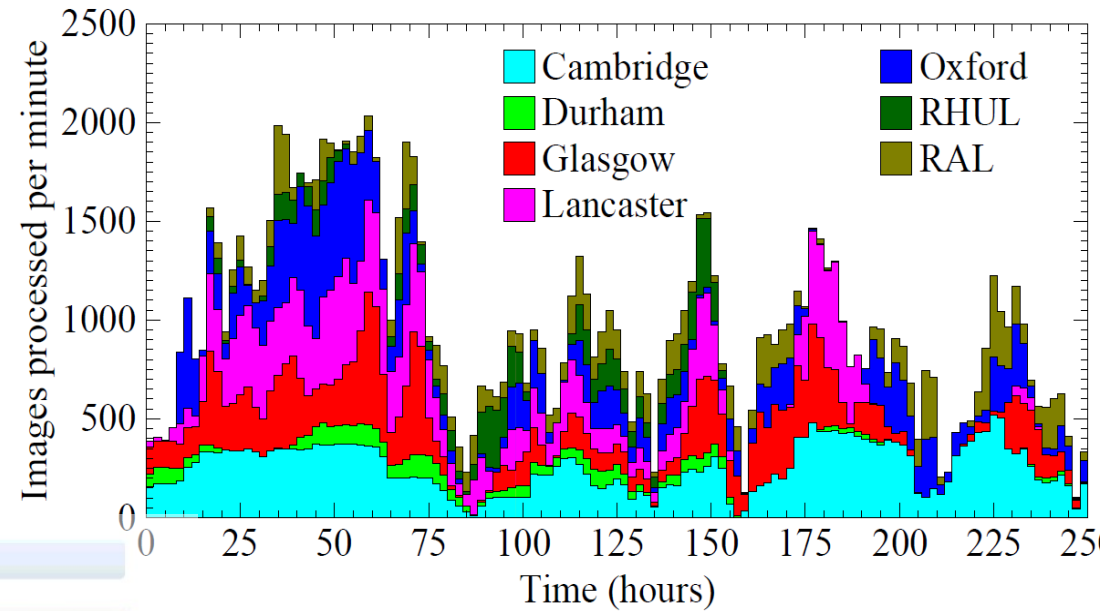
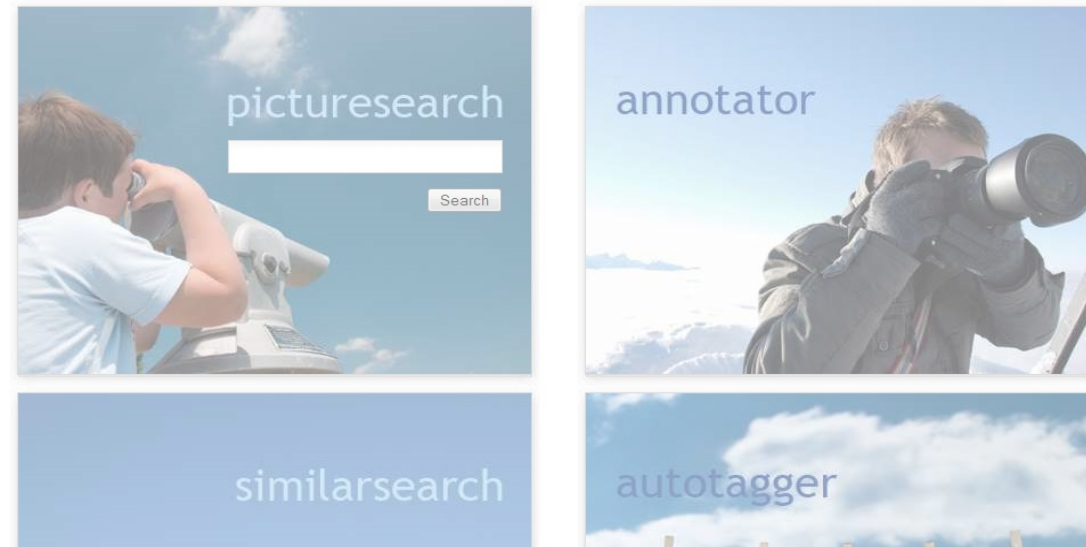
The majority of the computational time is spent *analysing the images* - typically a few seconds per image

Each image is analysed independently and so *the grid is perfect for this job!*

In November/December last year, **18 million stock images** were analysed and indexed

CEs from **7 UK sites** were used with the Cambridge SE used for data storage

imense



A working search engine is now available at:

<http://www.imense.com/>

iLexIR is another Cambridge start-up company and specialises in *text analytics, mining and classification*

The company have experience in many forms of text analysis; previous and current projects include:



- *Helping Corpora build a sentiment classifier*
- *Participation in the English Profile Project*
- *Helping Texperts develop efficient processing of SMS questions*
- *Development of a scientific paper text mining system*

iLexIR has developed two applications to help with textual analysis:

## RASP:

This analyses sentences for grammatical relations, e.g.

*The NYT today announced Google's acquisition of video hosting service, YouTube;*



*Possessive (acquisition, Google),  
Indirect-object (acquisition, of),  
Object (of, YouTube)*

## TAP:

This is a text classifier that, using RASP, can learn the relative importance of words without supervision

Current search engines are essentially only string matching and so a search such as:

*Bond films not featuring Roger Moore*

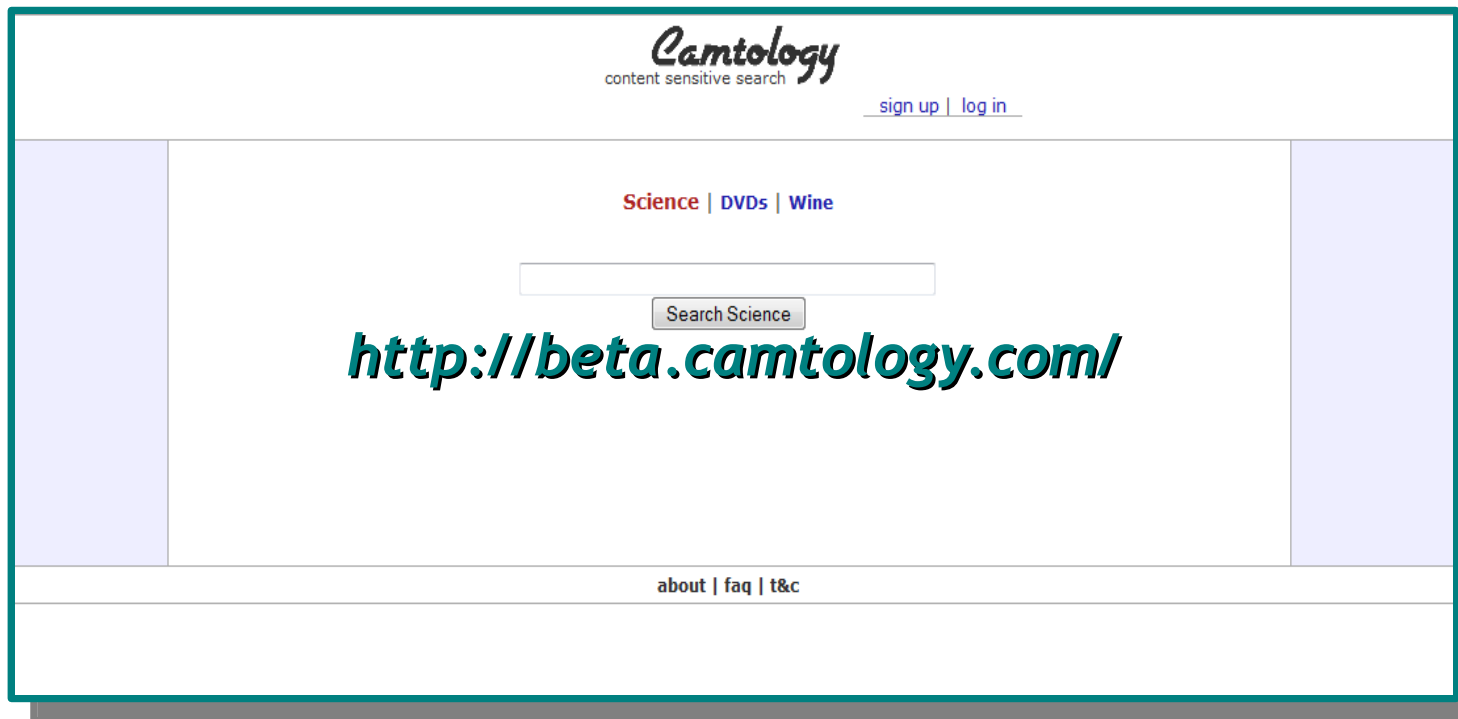
Will give you entirely the wrong results!

iLexIR, in combination with Imense are hoping to radically change this approach and offer:

- *Understanding of the query given*
- *Context based searching of text a variety of sources*
- *Search and classification of images based on content*
- *Linking both to allow a more accurate web search*

At present, Camtology have restricted their indexing and analysis to a few thousand scientific papers

However, a beta version of the site is up and running though is not available for public viewing as yet:



An n-gram corpus consists of the frequency of word combinations ('n-grams') and has many applications including *natural language processing and computational linguistics*

In 2006, Google made available an n-gram corpus for research projects of *just over  $10^{12}$  words* sampled from their internet indexing

Though this was very useful, there were certain limitations:

- It is limited to frequencies above 40
- Only up to 5 word n-grams are available
- No domain info was available



iLexIR hopes to at least *equal this size of corpus* but also overcome the shortcomings in Google's corpus

Over the last 6 months, they have *built on their experience and software* already used to search the internet for pdfs and images to develop a grid enabled web spider that can handle this task

After extensive testing, the spider should manage to *complete the search in a few weeks!*



To scan  $10^{12}$  words from the web takes quite a lot of effort!

First, assuming (from previous web search tests) an average of ~1000 words per web page, this will mean *scanning  $10^9$  web pages*

Each web page *takes ~1s to access and process*, so for one CPU, this would take 11500 days!

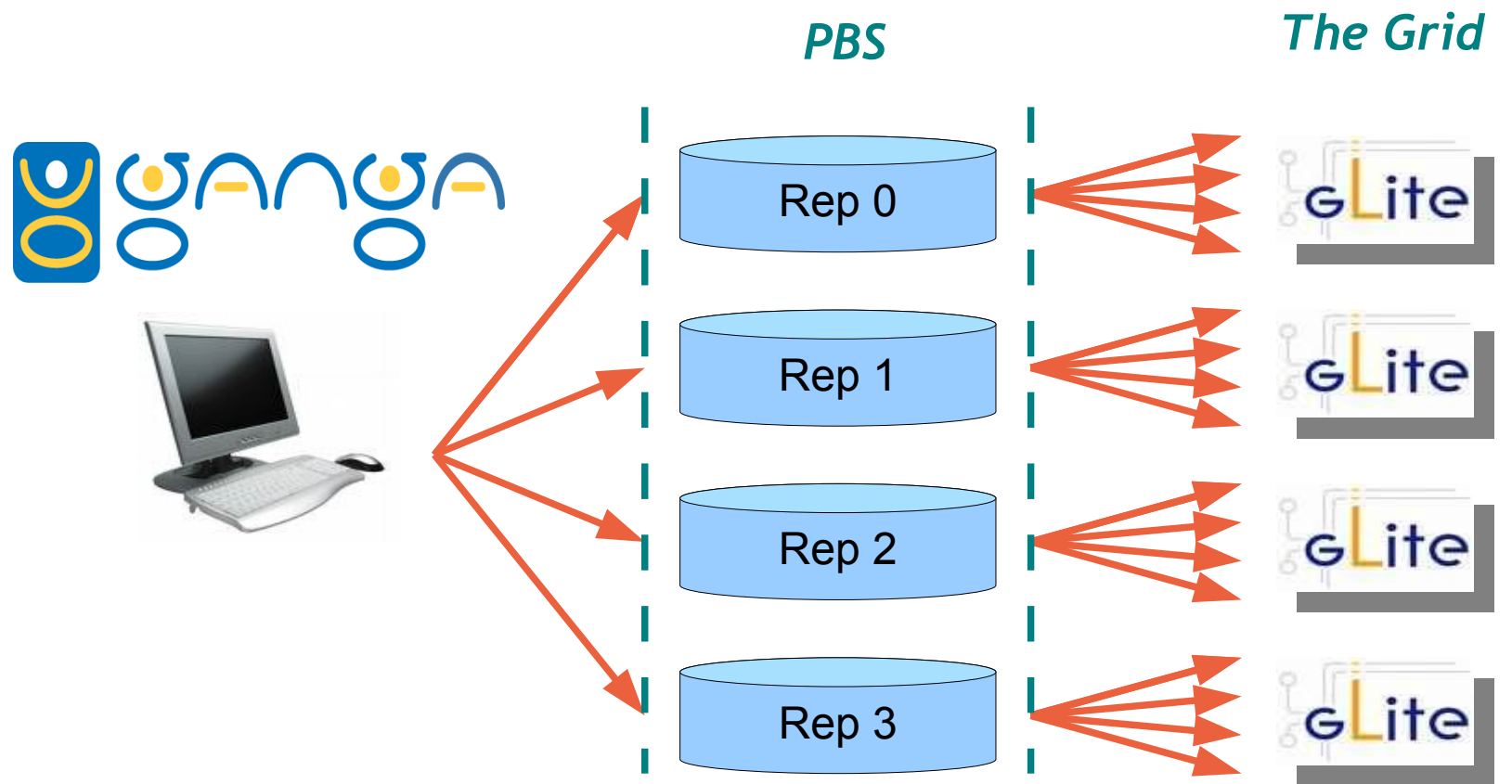
Using the grid though, this can immediately be cut down significantly. Assuming 2000 cores, the same scan *would only take less than a week!*

However, this is only part of the story:

- Submission time and job management* ●
- Keeping track of the domains/links scanned* ●
- Ensuring that Denial of Service attacks aren't launched accidentally!* ●

Each submission takes ~5s and we can't run very long jobs due to the input of new jobs being dependent on the output of others

We have therefore developed *a two tier submission scheme* using Ganga so we could keep 2000+ jobs running simultaneously



To ensure that all the jobs are not hitting the same domains many times a second, a job is given a unique list of domains and grabs the list of queued links from a repository kept on the submit machine

It is ensured that *no two active jobs are scanning the same domain*

As we will go over 1000s of domains, we use md5 sub-directories, e.g:

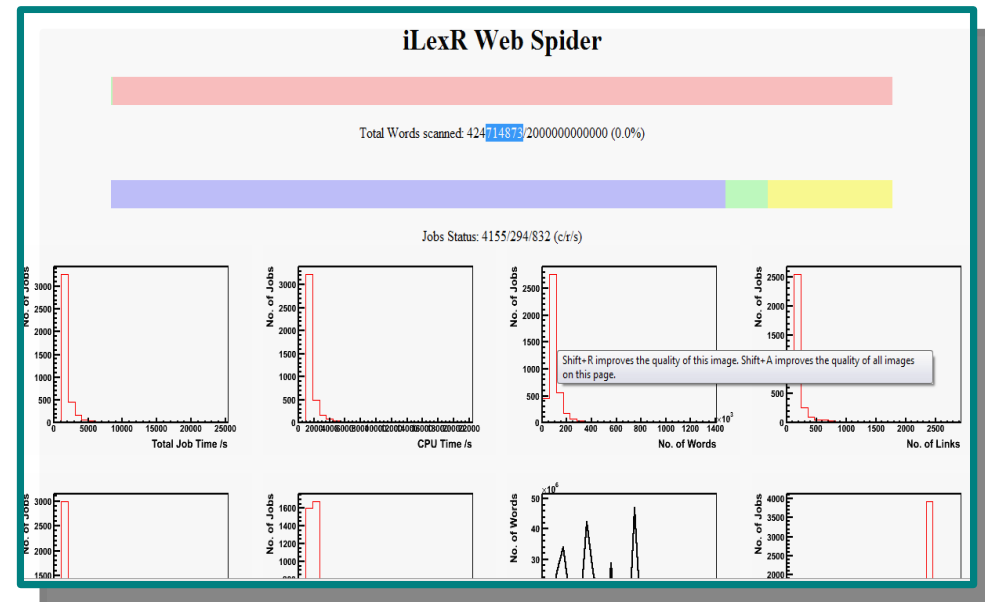
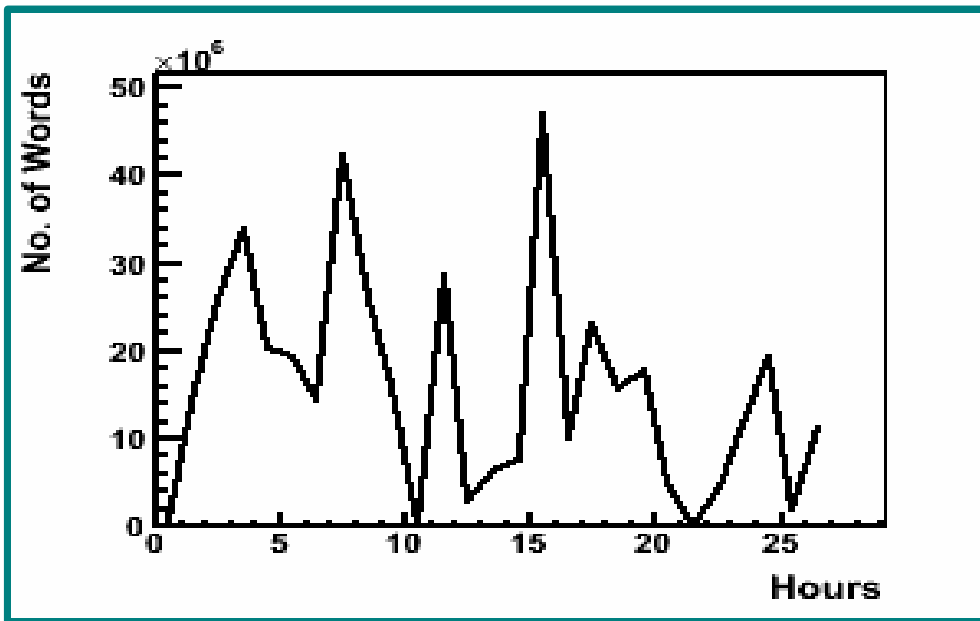
*/0ae/f25/<domain\_info>*

The domain info is a *pickled and zipped python dictionary* that stores both the viewed and queued links associated with the domain

In addition to the domain info, there are 10 job repositories linked to each submission job that's sent to PBS that keep track of the grid jobs each repository has sent

During this time we have improved the code significantly to the point where in the last test, over 24 hours we managed:

- *Nearly 0.5 billion words found*
- *Over 750000 links spidered*
- *5000 jobs submitted in total*
- *~250 jobs running simultaneously*
- *0.035 links per second*



Significant improvement has been made after this test and so we now have:

- Improved Anti-DOS logic* ●
- 20 download threads per job* ●

Initial 4 hour tests indicate that this gives us a speed increase of >10 meaning we could *complete the corpus in a few weeks* using 2000 cores!

Also, in order to address concerns raised by sys-admins as regards this work:

*There is no breach of IPR or copyright as no information is collected that could reproduce the scanned text*

*As regards 'banned' sites, no images or content is downloaded*

*As we obey robots.txt, any site that charges cannot have legal comeback if they don't restrict robots via this method*

*We are only doing what Google and other projects in the UK have done!*

Camtology and their associated companies, Imense and iLexIR are progressing well with their goal of *producing the next generation search engine*

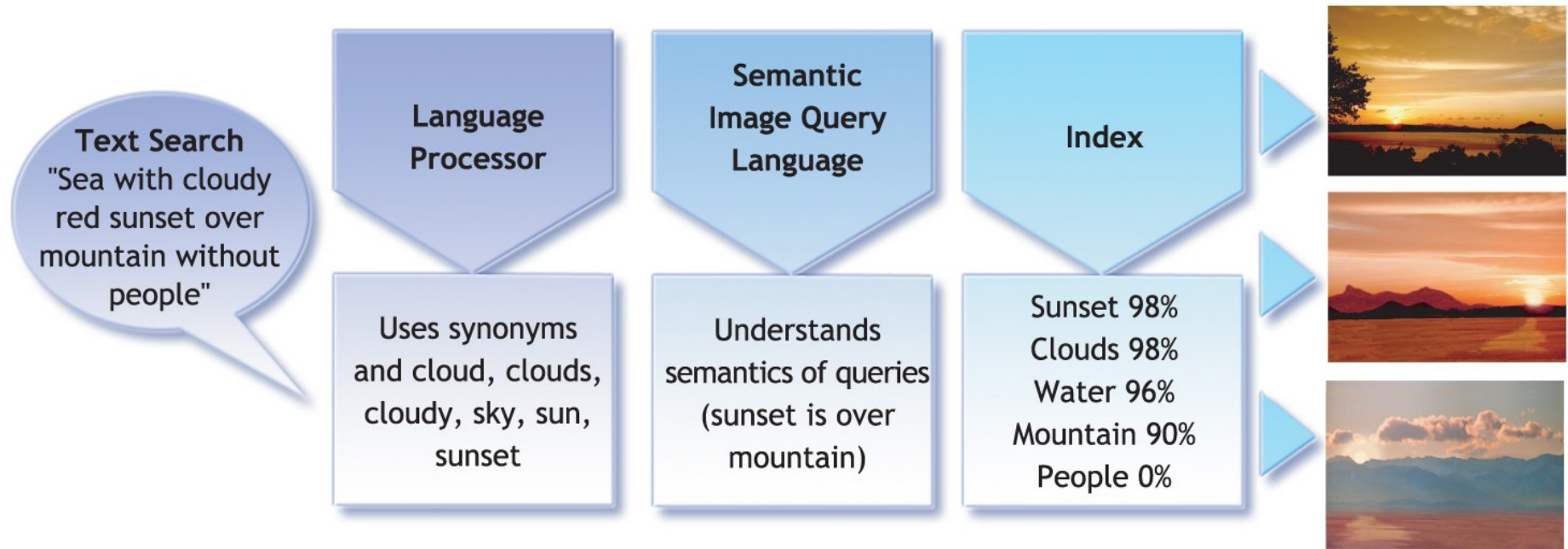
They have already managed to produce examples of this for *both images and a small selection of Scientific papers*

Coupled with this, their goal of using the grid to *produce a  $> 10^{12}$  word n-gram corpus using grid based spidering* seems to be possible on the timescale of ~1 month

All that is needed now is:

- A large scale test of 1000 job slots for 24 hours* ●
- The finalisation of the n-gram code* ●
- Making sure storage and processing resources are available* ●

## Backup Slides

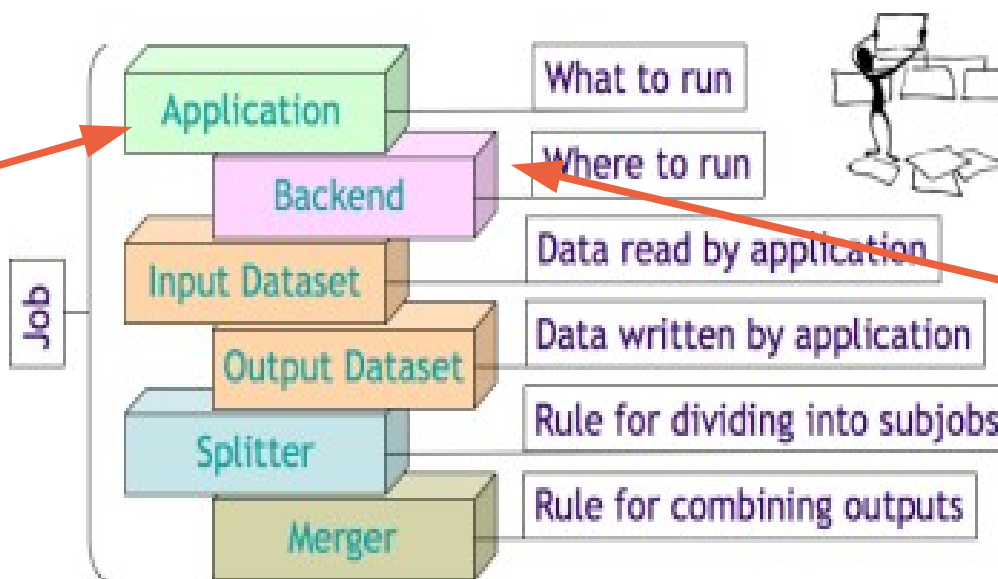


The backbone of the entire system is the *Ganga job management tool*

This is used to submit both to the *PBS system and the grid* using the Python scripting interface



We have developed a new 'Spider' application for Ganga that takes domain names and a repository path and sends jobs to the grid that scan the available links



We submit spider jobs to the LCG backend, but also 'Ganga' jobs to the PBS backend