Introduction to Oracle on the NGS

Keir Hawker – DBA – Rutherford Laboratories

Simon Collins – Data Consultant – Manchester University

NGS
National Grid Service

Goals of today

- Introduce concepts of relational databases
  - Focus on Oracle as it is the NGS standard database
    - We can support other databases
- Oracle install on the NGS
  - Who's using the service now?
  - Benefits ?
  - How do I apply and get most the most of the service?
- Functionality specific to the life science community
- Get feedback on your specific requirements for data storage

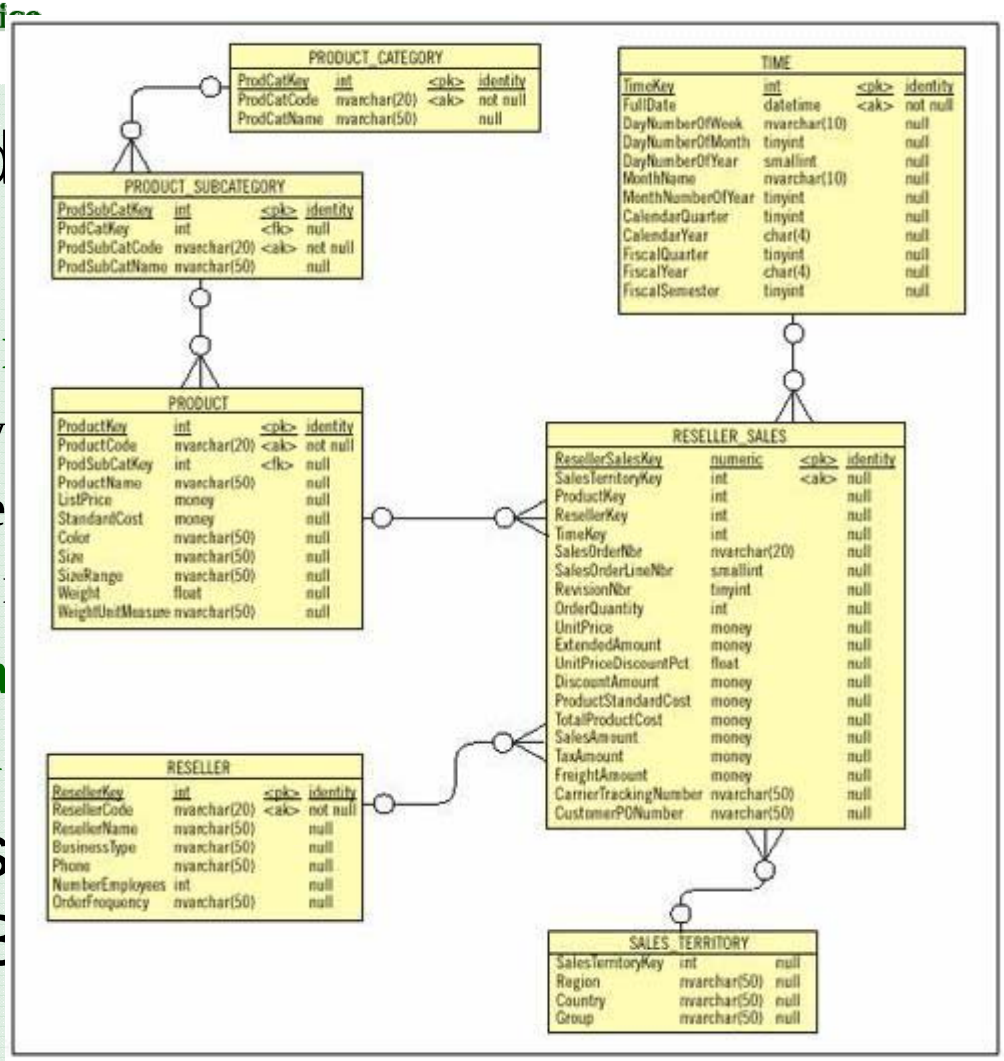
NGS
National Grid Service

Agenda

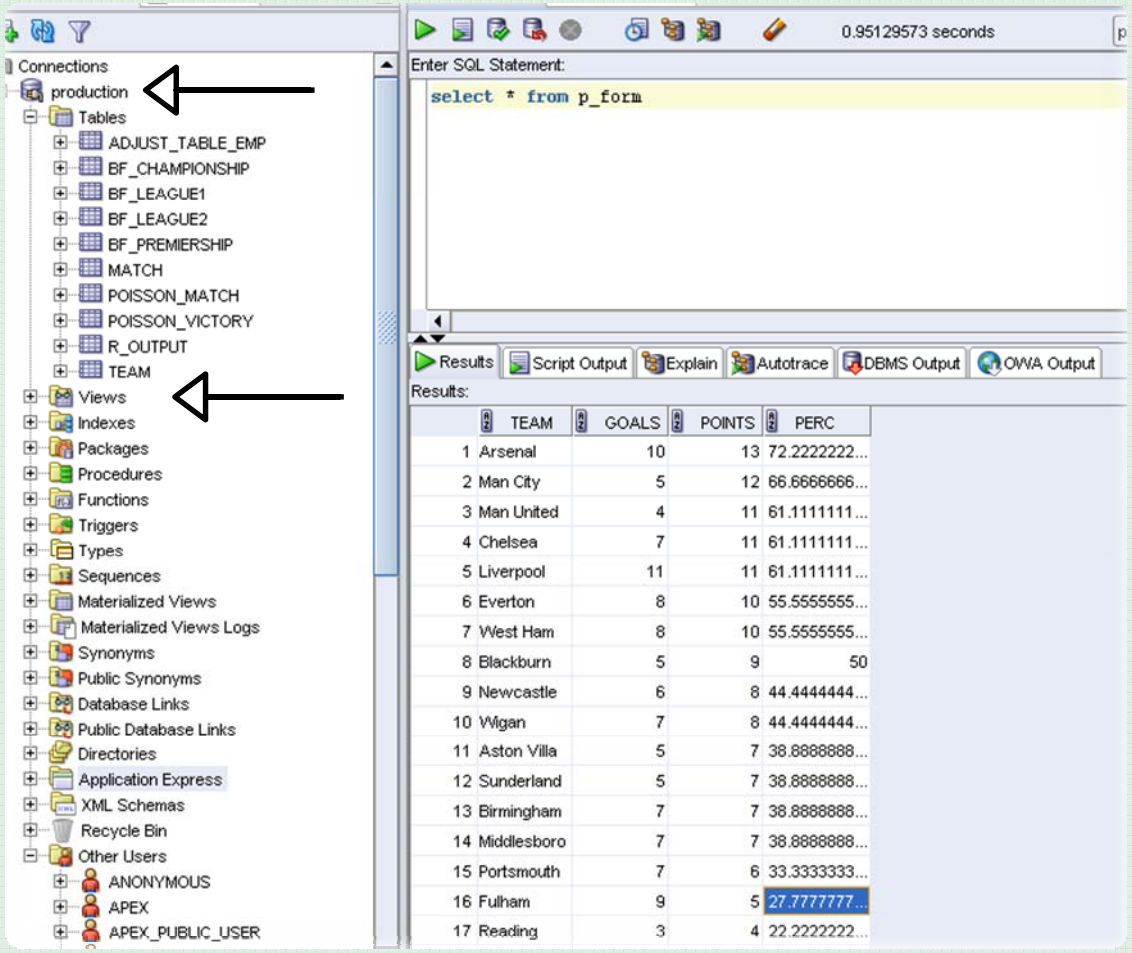
- Introduction to relational databases
- Oracle on the NGS
- Demonstration
  - Connectivity to Oracle on the NGS
- Oracle tools for life sciences
- Designing database systems
- Practical
  - BLAST queries within Oracle
- Q & A

Relational Databases

- Simply - d relational model
  - Collectio ta model
    - Data ty
    - Data re
    - Tables
  - Provide a e, retrieve, mine and
- Examples Server, Ingres , IE



Typical view of a database (shown through SQLDeveloper)



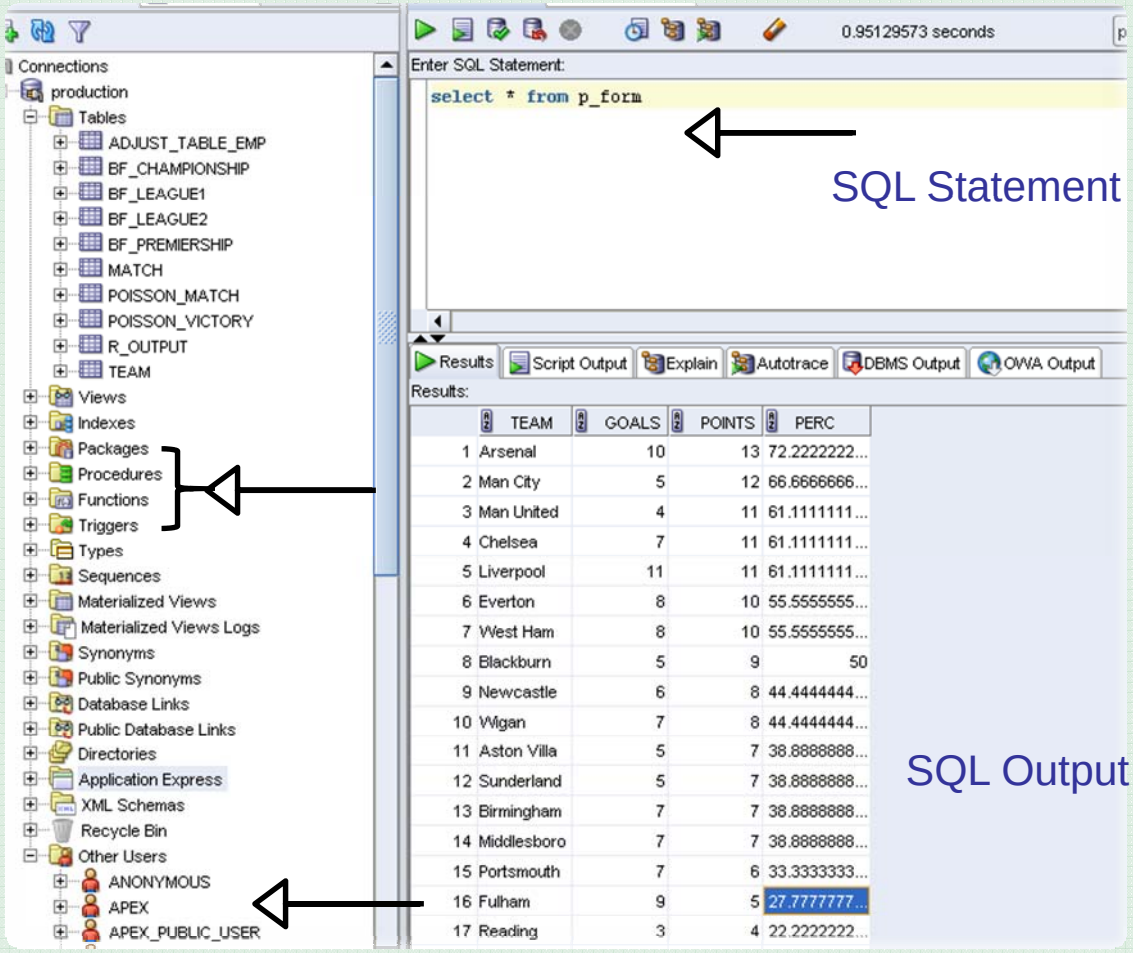
Schema – Synonymous to user

A design space for an application where tables / views etc are created

Tables – where base data is stored

Views – defined SQL queries on the table data to represent data in a more meaningful way

Typical view of a database (shown through SQLDeveloper)



Oracle PL /SQL

Types of code for manipulating data

Other Schemas in the database

Databases on the NGS

- NGS supports Oracle 10G database as standard
  - MySQL databases
- Databases are installed at RAL and Manchester
- Databases are administrated internally
- Any NGS user is entitled to an Oracle account
  - Full access to all Oracle functionality / tools
  - Unrestricted access to database development within their own partition (schema)

NGS National Grid Service

Who's using these databases?

- Portsmouth University – Astronomy data studies
  - Vast quantities of astronomy data is now being produced
    - Envisaged that such data in the future will have to be kept on heterogeneous data sources (multiple databases)
  - NGS Manchester has 3 Terabytes of SDSS digital data
  - Portsmouth conducting performance tests in querying distributed databases
  - Use OGSA-DAI and OGSA-DQP on the NGS to
    - Query the Oracle database
    - Query both the Oracle and SQL Server database located in Portsmouth

NGS
National Grid Service

Example databases on the NGS

- MIMAS sponsored projects to Grid enable data
  - GEMS Projects (Grid Enabled Mimas Services)
    - GEMS 1 project - access to 2001 census data via OGSA-DAI
    - GEMS 2 project – access to MIMAS satellite image data
  - ConvertGrid
    - Creation of large grid enabled geographic datasets
- Developed web front ends
  - Allow social scientists to filter data
  - Allow statistical analysis of the data through other resources on the Grid
  - Access through OGSA-DAI developed services to maintain Grid security

NGS
National Grid Service

Why put your data on the NGS?

- Fully administered
  - Databases are supported and maintained by NGS Staff
  - Databases are backed up daily
- Advice
  - Help on how to access and develop your database
- Data Storage
  - Significant storage space available on the NGS for large datasets.
- Data Integration
  - Oracle may be beneficially used in conjunction with other services / programs installed on the NGS

NGS
National Grid Service

Why put your data on the NGS?

- Computational Power
  - Potential for data analysis that may be prohibitive on other machines
- Data Protection
  - Databases configured on RAID 5/10 disks and fully backed up
- Oracle Functionality
  - NGS has fully licensed Oracle Enterprise suite
  - Lot of potentially useful functionality
  - It's all free

How does Oracle fit in with “The Grid”?

- Oracle software is not naturally part of the globus software stack
- “Grid enabling” data can be done through
  - OGSA-DAI
- Do I have to use OGSA-DAI to utilize the Oracle databases on the NGS?

NGS
National Grid Service

How do I apply for access?

- Through NGS website
  - http://www.ngs.ac.uk
  - http://www.grid-support.ac.uk/content/view/221/171/
- Provide following information
  - Storage Requirements
  - Type of access needed
  - Where you'd like the database to be hosted (RAL or Manchester)
  - What kind of help you need in setting it up

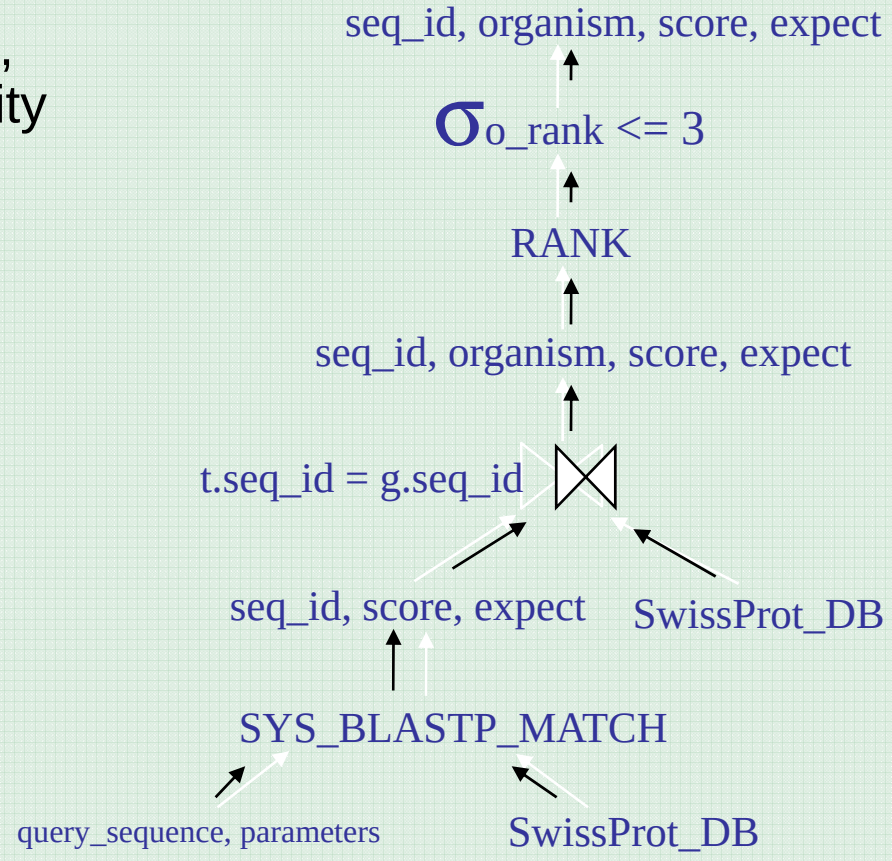
Demonstration – Connecting to Oracle

- Many tools for connecting to Oracle for administration or development
  - SQL*Plus (Shell or GUI)
  - SQLDeveloper (GUI)
  - Isqlplus (Web)
  - Oracle Enterprise Manager (Web)
- We will specifically look at
  - Sql Login (Script allowing SQL*Login on the NGS)

Oracle functionality for life sciences
Blast Searches

- For the query sequence “ATCGCGTT”, find the top 3 matches above a similarity threshold from each organism

```
select seq_id, organism, score, expect
from (select t.seq_id, t.score, t.expect, g.organism,
      RANK() OVER (PARTITION BY organism
      ORDER BY score DESC) as o_rank
   from SwissProt_DB g,
        Table(SYS_BLASTP_MATCH  (‘ATCGCGTT’,
      cursor (select seq_id, sequence from
           SwissProt_DB), 5)) t   /* expect_value */
   where t.seq_id = g.seq_id) where o_rank <= 3
```

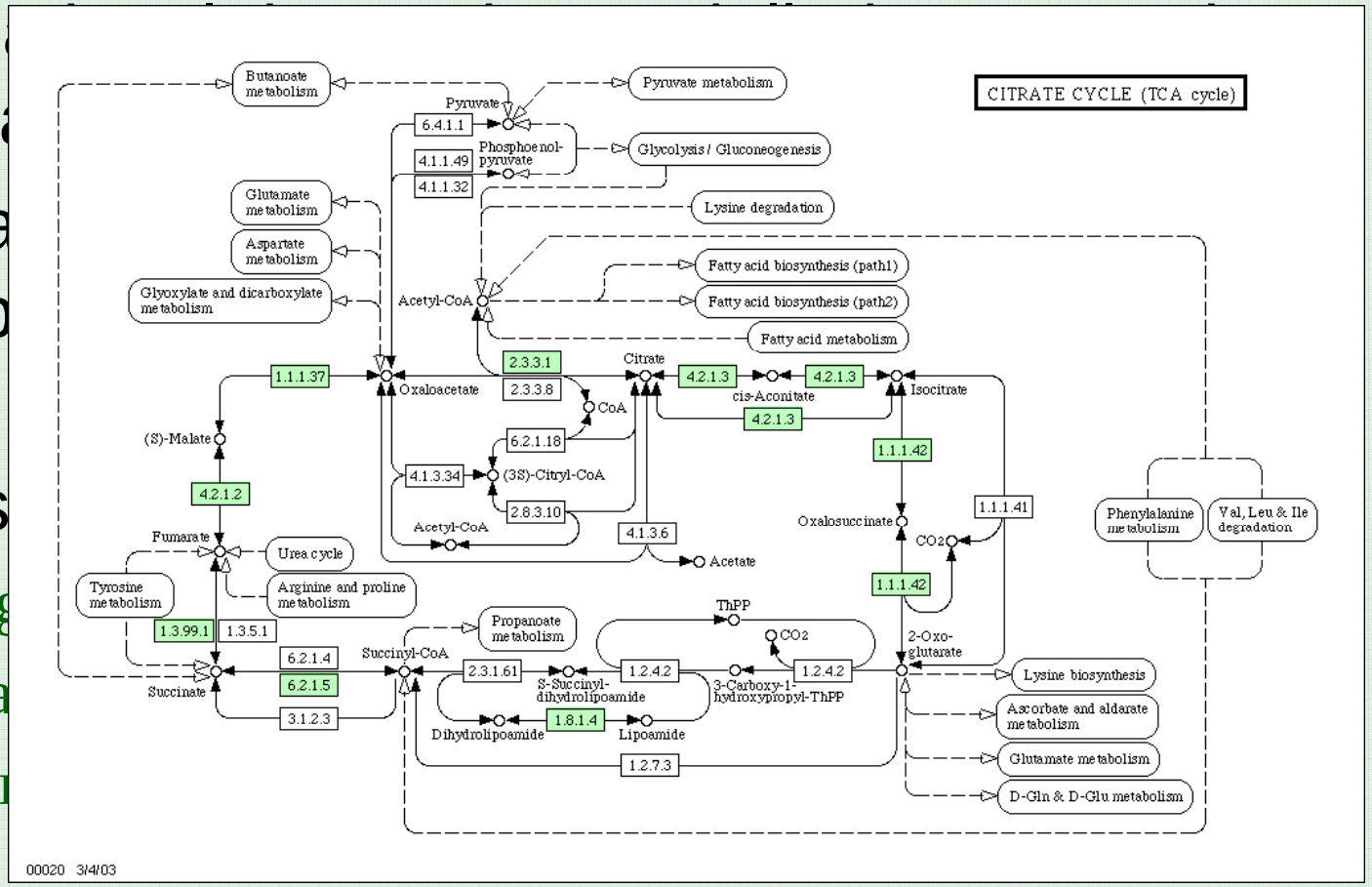


- BLAST “Delighters”
  - Queries performed in the database
  - Ability to perform combinatorial queries e.g. sequence similarity AND annotation contains “Lymphoma”

NGS National Grid Service

Oracle functionality for life sciences
Oracle Spatial

- Allows rel
  represent
- Network a
  shortest-p
  minimum
- Examples
  - Managing
  - Signal tra
  - Protein pr



NGS Case Study: Soton University
Oracle RDF

- Southampton University has tens of thousands of statements on chemical resources.
- RDF [Resource Description Framework] is a way of storing statements about resources
- Examples of using subject-predicate-object
  - H2O boils at 100 Degrees Celcius
  - Salt is the compound NaCl
- Relationships can be defined explicitly AND implicitly meaning a more powerful and efficient search

NGS Case Study: Soton University
Oracle RDF

- A [very] Basic RDF example is:
  - Mr A is Mr B's Father
  - Mr B is Mr C's Father
  - Mr C is Mr B's Child
  - Statement: A grandfather is a father of a father of a child.
  - Therefore Implicitly: Mr A is Mr C's Grandfather.

- Southampton university use this kind of relationship to discover relationships between chemical groups that had previously not been linked.

Oracle functionality for life sciences
Data Mining

Process of sifting through massive amounts of data to find hidden patterns and discover new insights

Data Mining can provide valuable results

- Identify factors more associated with a business problem *(Attribute Importance)*
- Predict individual behavior *(Classification)*
- Predict or estimate a value *(Regression)*
- Find profiles of targeted people or items *(Decision Trees)*
- Segment a population *(Clustering)*
- Determine important relationships within the population *(Associations)*
- Find fraudulent or “rare events” *(Anomaly Detection)*

NGS
National Grid Service

Oracle functionality for life sciences
Data Mining

Life Sciences examples
- Leukemia AML/ALL Golub et al.
- NCI-60 ChemoSensitivity data

Database Marketing
- Target doctors likely to prescribe new drug(s)
- Target "best" patients

Discovery/Development
- Discover target genes and proteins
- Identify promising leads for new drugs
- Medline literature mining
- Pharmacovigilance

Health Care
- Predicting medical outcomes
  - Diabetes
  - Pneumonia
  - Respond to treatment
- Fraud detection

Oracle functionality for life sciences
Data Mining

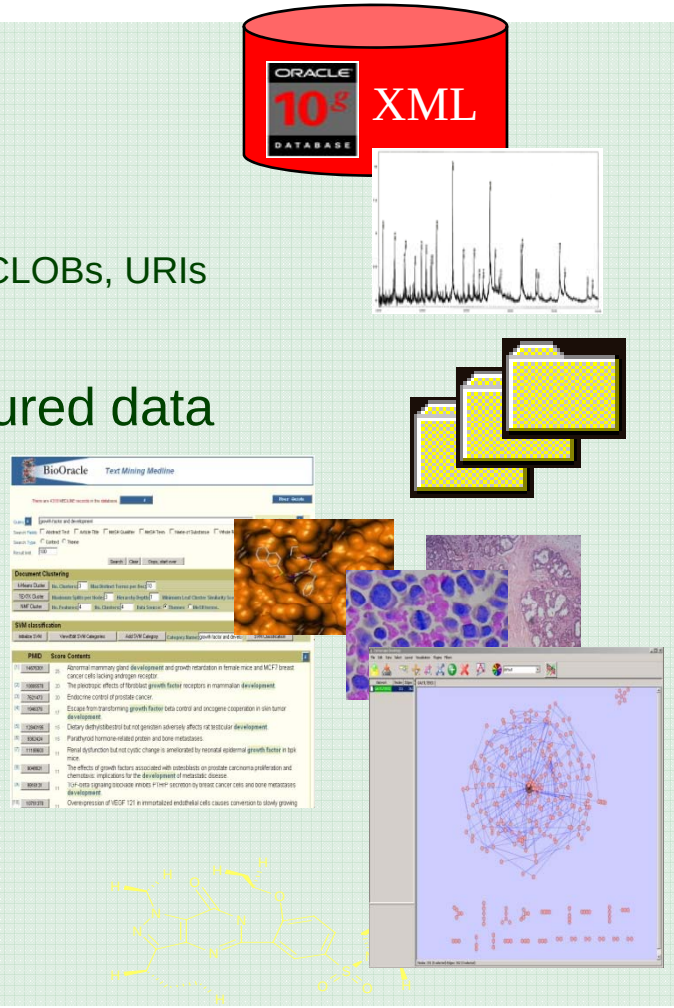
- Wide range of DM algorithms
  - Naïve Bayes, Adaptive Bayes Networks, Decision Trees, Attribute Importance, Association Rules, K-Means, O-Cluster, SVM, NMF algorithms
- Example use of data mining algorithms
  - Use gene expression data to distinguish between lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML)
    - Gene expression data from ~7000 Genes, 38 samples training set
    - Bayes Networks – supervised learning algorithms applied identify important gene expressions
    - Algorithm was able to distinguish between ALL and AML in 35 new test cases

Oracle functionality for life sciences
Statistics

- Ranking functions
  - rank, dense_rank, cume_dist, percent_rank, ntile
- Window Aggregate functions (moving and cumulative)
  - Avg, sum, min, max, count, variance, stddev, first_value, last_value
- LAG/LEAD functions
  - Direct inter-row reference using offsets
- Reporting Aggregate functions
  - Sum, avg, min, max, variance, stddev, count, ratio_to_report
- Statistical Aggregates
  - Correlation, linear regression family, covariance
- Linear regression
  - Fitting of an ordinary-least-squares regression line to a set of number pairs.
  - Frequently combined with the COVAR_POP, COVAR_SAMP, and CORR functions.
- Descriptive Statistics
  - average, standard deviation, variance, min, max, median (via percentile_count), mode, group-by & roll-up
  - DBMS_STAT_FUNCS: summarizes numerical columns of a table and returns count, min, max, range, mean, stats_mode, variance, standard deviation, median
- Correlations
  - Pearson’s correlation coefficients, Spearman's and Kendall's (both nonparametric).
- Cross Tabs
  - Enhanced with % statistics: chi squared, phi coefficient, Cramer's V, contingency coefficient, Cohen's kappa
- Hypothesis Testing
  - Student t-test , F-test, Binomial test, Wilcoxon Signed Ranks test, Chi-square, Mann Whitney test, Kolmogorov-Smirnov test, One-way ANOVA
- Distribution Fitting
  - Kolmogorov-Smirnov Test, Anderson-Darling Test, Chi-Squared Test, Normal, Uniform, Weibull, Exponential
- Pareto Analysis (documented)
  - 80:20 rule, cumulative results table

Oracle functionality for life sciences
Integrate variety of data types

- XML DB
  - Unite XML content & SQL/relational data
- LOBs
  - Manage unstructured data e.g. BFILES, BLOBs, CLOBs, URIs
- Files(Oracle9*i*FS)
  - Central repository for structured & unstructured data
- Text
  - Index & fast query of text content
- *inter*Media
  - Manage audio, video & image data
- Network Data Model (Oracle Spatial)
  - Graph (arc node) relationships
- Extensible indexing
  - Manage & index complex scientific data



Oracle functionality for life sciences
More information

- Oracle technology for bioinoformatics
  - http://crpit.com/confpapers/CRPITV19Blackwell.pdf
- Oracle life science offerings can be found at
  - http://www.oracle.com/technology/industries/life_sciences/index.html
- Good summary of functionality
  - http://www.accelrys.com/technologies/informatics/cheminformatics/user_group/files/Oracle_SF.pdf

NGS National Grid Service

Implementing a database solution

1) Database administration
   - Installation of the server software
   - Control access and security
   - Apply upgrades and patches
   - Define and implement back up strategies
   - Troubleshoot performance problems

- NGS looks after all these tasks

NGS National Grid Service

Implementing a database solution

2) Design of the database schema

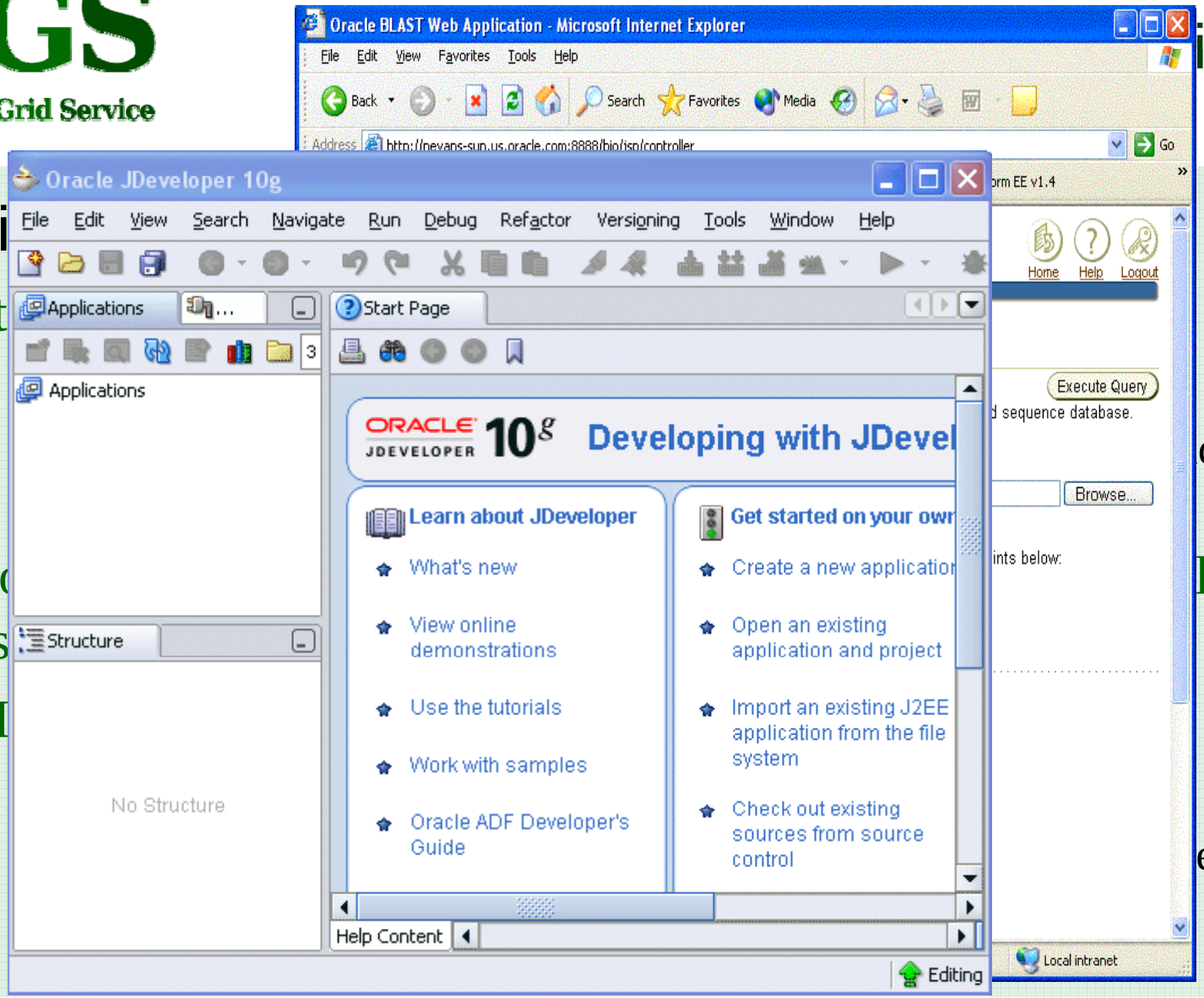
- Mapping data to objects within the database
  - Tables, views
- Creating internal logic perform calculations, manipulations data
- Load data
- Many tools available to help with this
  - SQL Plus, Sql Developer, Jdeveloper, Enterprise Manager

- NGS can give consultancy to help with these tasks.

NGS
National Grid Service

3) Appli…ion

- St…
  - …
  - …
- C…low us…
- M…
  - …
  - …ern



Practical 2 – Querying the database

- Purpose of practical
  - Use ISQPlus to connect to the database
  - Explore a schema (user area with pre built tables)
    - Schema contains small gene and protein databases
  - Use SQL to perform BLAST searches within Oracle
    - SQL (structured query language)
  - Show a simple developed application to perform the same query.
    - Developed in Oracle Application Express

NGS
National Grid Service

Contact Information

- K.C.Hawker@rl.ac.uk
- Simon.collins-2@manchester.ac.uk