# Automatic Annotation of Gene Expression Patterns for Mouse Embryo

LIANGXIU HAN

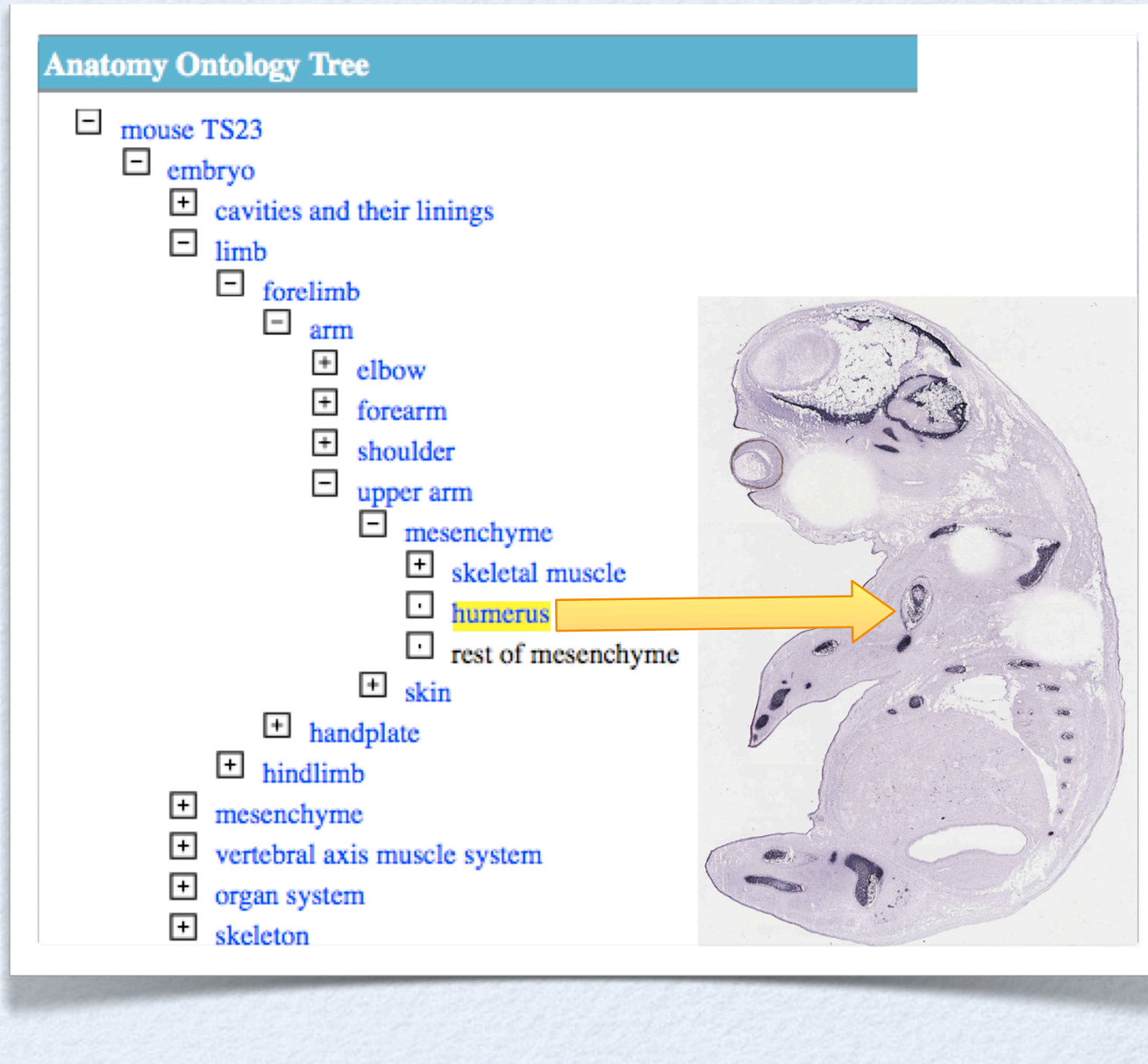
FUNDS Research Group - Future Networks and Distributed Systems
School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University
Group Webpage: http://www.scmdt.mmu.ac.uk/research/funds/

# OUTLINE

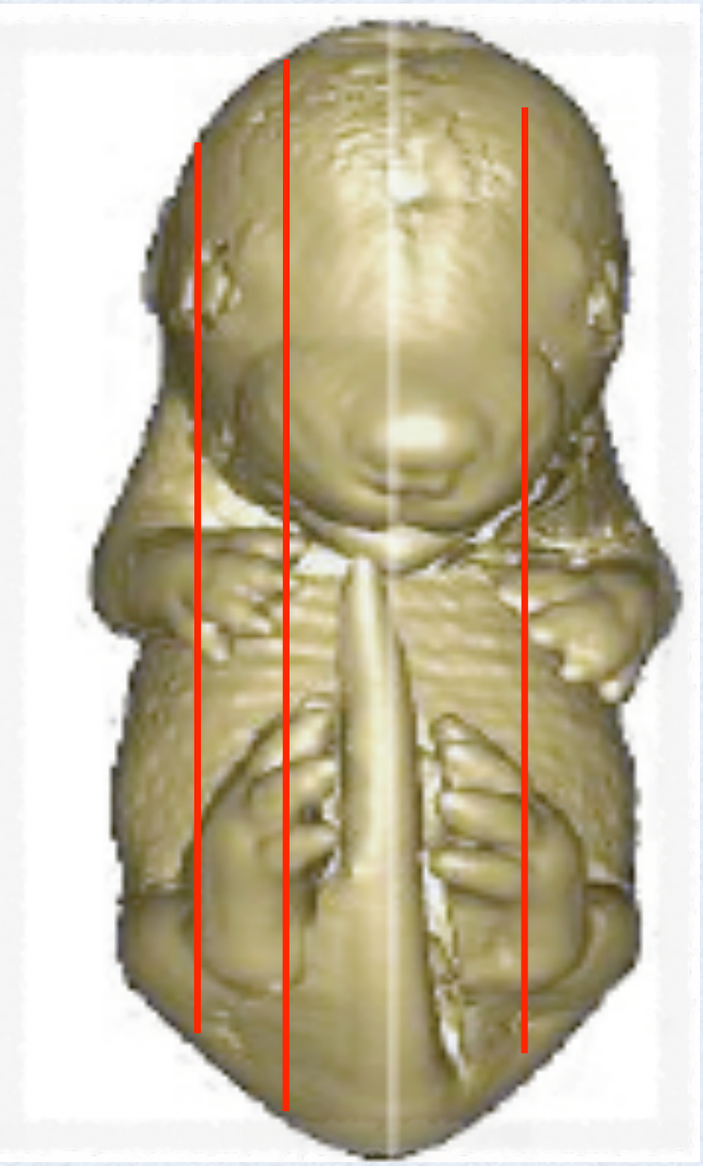
- What is Annotation?
- Motivation and Challenges
- Methodology
- Experimental and Evaluation
- Exploration of Parallelisation
- Ongoing Work

# What is annotation?-1

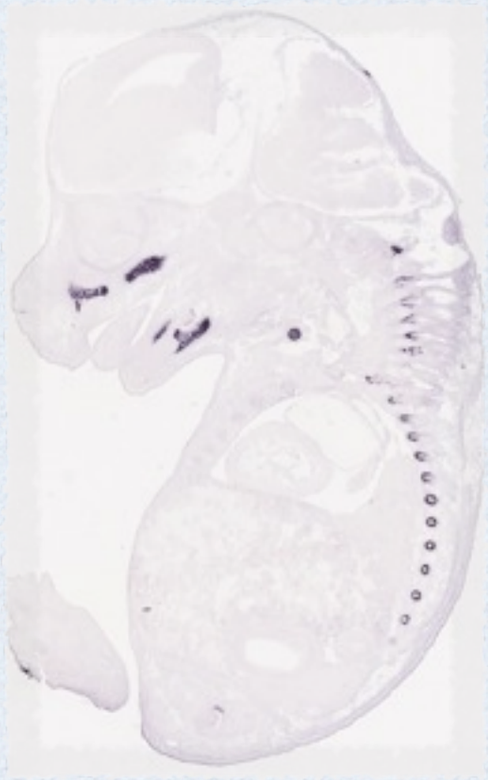
- Tagging an anatomical term from ontology with gene expression patterns of the anatomical component in images



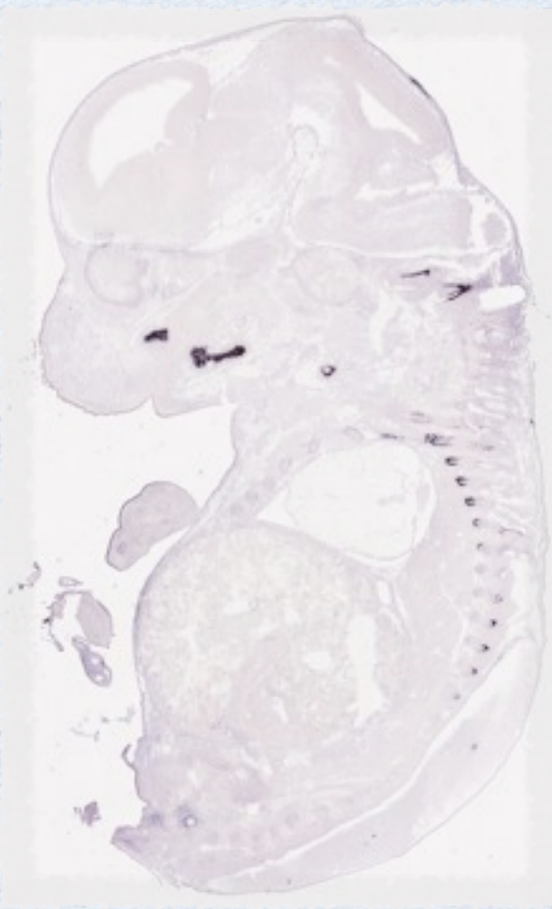
# WHAT IS ANNOTATION?-II



euxaxssay_007708_02.jpg



euxaxssay_007708_06.jpg



euxaxssay_007708_16.jpg

# OUTLINE

- What is Annotation?
- Motivation and Challenges
- Methodology
- Experimental and Evaluation
- Exploration of Parallelisation
- Ongoing Work

# Motivation & Challenges-1

- Gene expression patterns --- *a way to understand the interaction between genes*

- The availability of both ontological annotation and spatial gene pattern --- *a resource to identify the mechanism of embryo organisation*

- The current manual annotation --- costly and time consuming

- Massive amounts of data and complicated organism --- *necessity to automate the process of annotation*

# Motivation & Challenges-II

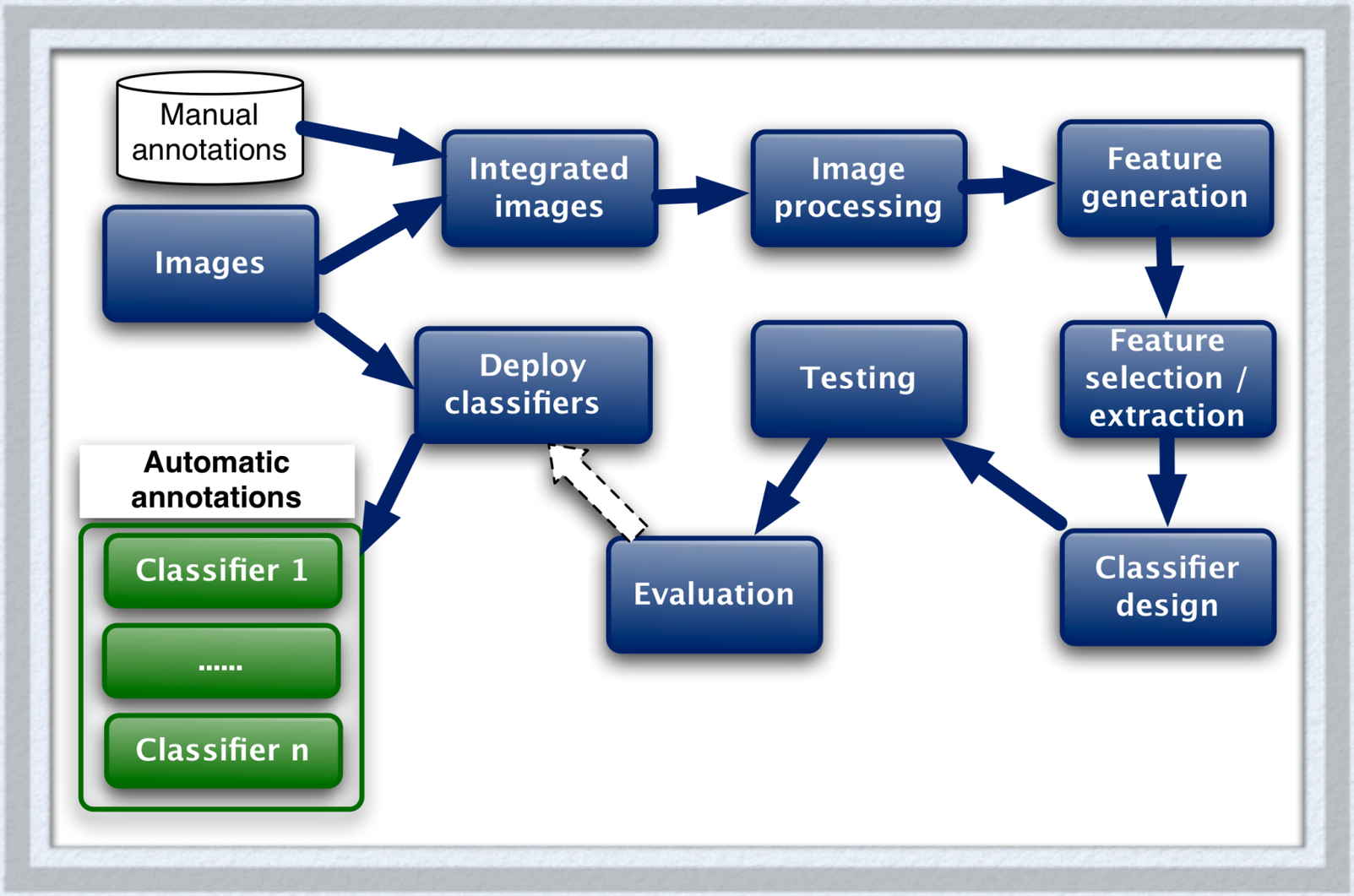
- Big data, now over 20 TB
- Multi-components coexisting in an image
- Variable shape, location and orientation of images
- The number of images associated with a certain gene is uneven
- The dimensionality of each image is high (3kx4k pixels)

# OUTLINE

- What is Annotation?
- Motivation and Challenges
- Methodology
- Experimental and Evaluation
- Exploration of Parallelisation
- Ongoing Work

# Methodology-1

- The Framework



# METHODOLOGY-II

- Image Processing
- Wavelet Transform
- Fishers'Ratio
- and LDA (SVM, ANN and LSVM)

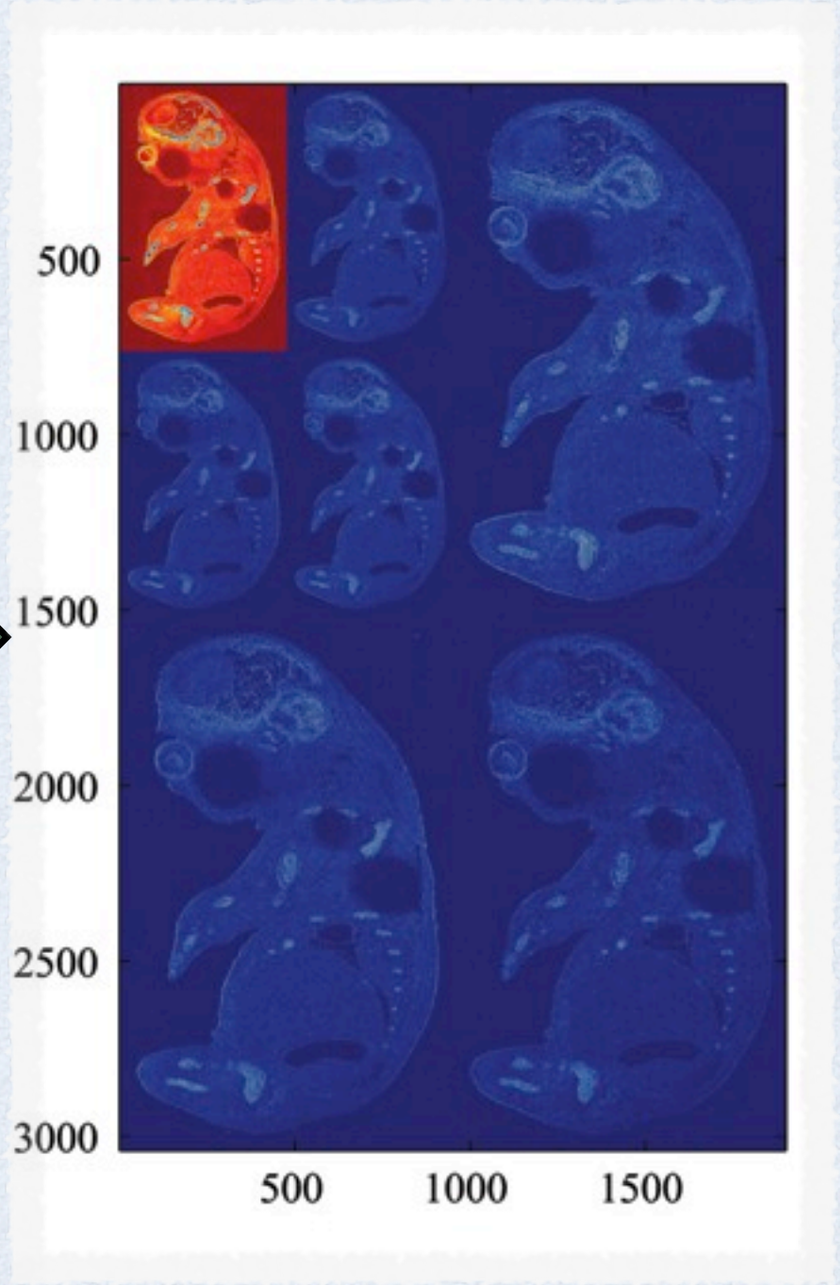
# METHODOLOGY-III

- Image Processing - Filtering

# Methodology-IV

- Wavelet transform

Wavelet decomposition



# METHODOLOGY-V

- Fishers'Ratio

$$Fisher\,Ratio = \frac{(m_{1,i} - m_{2,i})^2}{(v_{1,i}^2 + v_{2,i}^2)}$$

- LDA (Linear Discrimination Analysis)

Linear discriminant function:

$$f(X) = W^t X + w_0$$

Target function:

$$T(W) = \frac{|W^t S_B W|}{|W^t S_W W|}$$

Between-class scatter matrix

$$S_B = (m_1 - m_2)(m_1 - m_2)^t$$

Within-class scatter matrix

$$S_W = S_1 + S_2$$

where,
$S_1 = \sum_{x \in C_1} (X - m_1)(X - m_1)^t$ and $S_2 = \sum_{y \in C_2} (Y - m_2)(Y - m_2)^t$

# OUTLINE

- What is Annotation?
- Motivation and Challenges
- Methodology
- Experimental and Evaluation
- Exploration of Parallelisation
- Ongoing Work

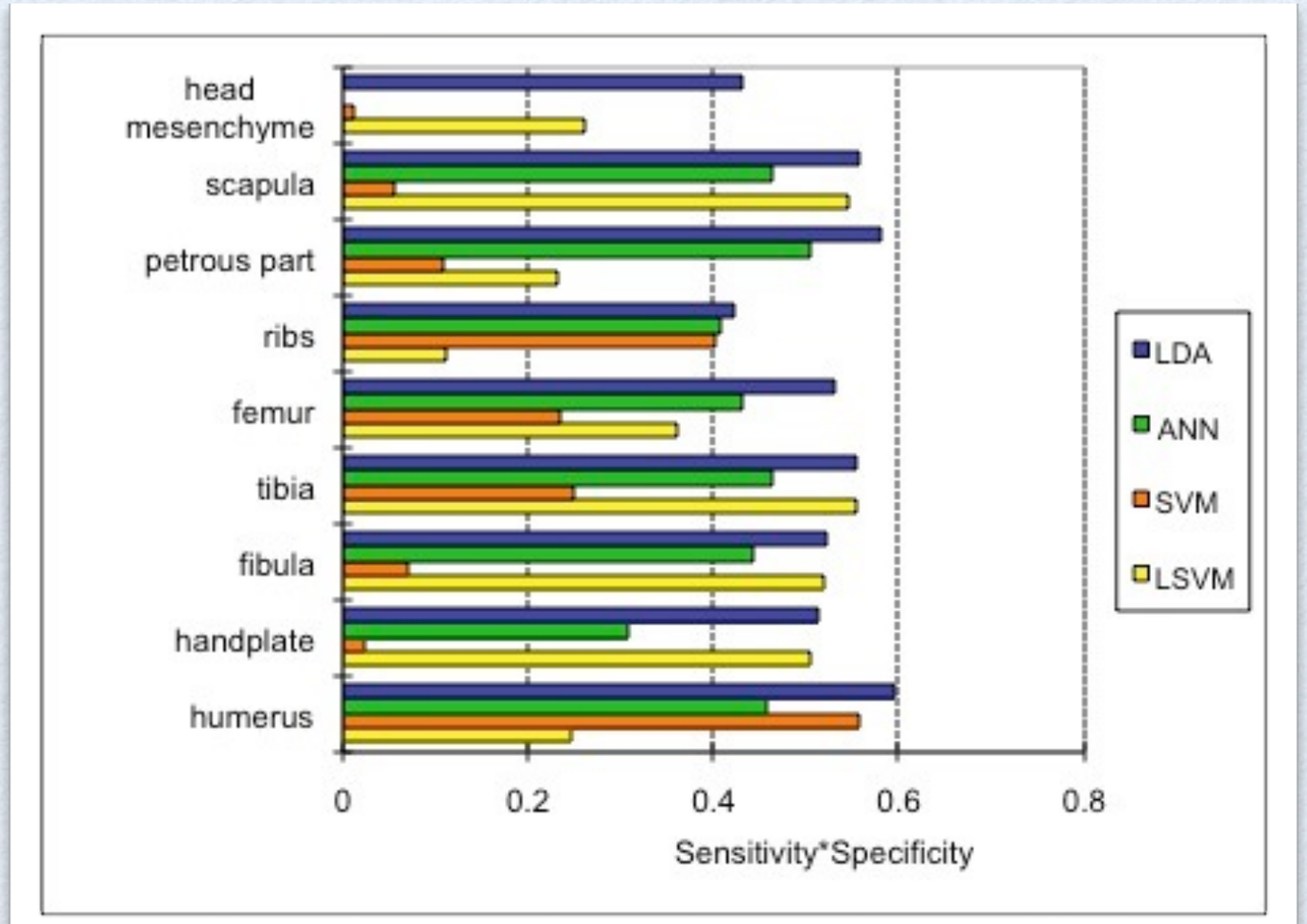
# EXPERIMENTAL EVALUATION-I

| Anatomic component | Our proposed method | | SVM | | LSVM | | ANN | |
|---|---|---|---|---|---|---|---|---|
| | Se | Sp | Se | Sp | Se | Sp | Se | Sp |
| Humerus | 0.7905 (0.0235) | 0.7521 (0.0529) | 0.838 | 0.6667 | 0.4992 | 0.4949 | 0.8985 | 0.5101 |
| Handplate | 0.7231 (0.0254) | 0.7125 (0.078) | 1.00 | 0.0263 | 0.6739 | 0.75 | 0.5744 | 0.5395 |
| Fibula | 0.7166 (0.0302) | 0.7256 (0.085) | 0.9738 | 0.0744 | 0.7253 | 0.719 | 0.7922 | 0.562 |
| Tibia | 0.7436 (0.02459) | 0.7467 (0.051) | 0.9439 | 0.2667 | 0.7511 | 0.74 | 0.9044 | 0.5133 |
| Femur | 0.7374 (0.0403) | 0.7235 (0.0607) | 0.9726 | 0.2414 | 0.5613 | 0.6466 | 0.8802 | 0.4914 |
| Ribs | 0.7519(0.0307) | 0.5608(0.0717) | 0.7939 | 0.5088 | 0.124 | 0.9018 | 0.7252 | 0.5649 |
| Petrous part | 0.737(0.0316) | 0.7885(0.094) | 0.9854 | 0.1129 | 0.2715 | 0.8629 | 0.8015 | 0.629 |
| Scapula | 0.710(0.03) | 0.789(0.0851) | 0.9945 | 0.0588 | 0.7265 | 0.7529 | 0.8218 | 0.5647 |
| Head mesenchyme | 0.5506(0.035) | 0.8286(0.109) | 1.00 | 0.0143 | 0.3045 | 0.8571 | 1.00 | 0.00 |

Sensitivity (true positive rate): the proportion of actual positives in the whole testing dataset

Specificity (true negative rate): the proportion of true negatives in the whole testing dataset

# EXPERIMENTAL EVALUATION-II




http://www.eurexpress.org/ee/databases/assay.jsp?assayID=euxassay_010351

This work has been published in Bioinformatics (Journal):

*Han, L., van Hemert, J., Baldock, R. "Automatically Identifying and Annotating Mouse Embryo Gene Expression Patterns", Bioinformatics 27(8),pp1101-11-07**,** Oxford Journals, Oxford University Press. DOI:10.1093/BIOINFORMATICS/BTR105, 2011

# OUTLINE

- What is Annotation?
- Motivation and Challenges
- Methodology
- Experimental and Evaluation
- Exploration of Parallelisation
- Ongoing Work

# PARALLELISATION EXPLORATION- 1

- Parallelisation is a sought after solution for speeding up an application, particularly for data intensive applications
- Three considerations for parallelising an application
  - How to distribute workloads or decompose an algorithm into parts
  - How to map the tasks onto various computing nodes and execute subtasks in parallel
  - How to coordinate and communicate subtasks on those computing nodes.

# PARALLELISATION EXPLORATION- II

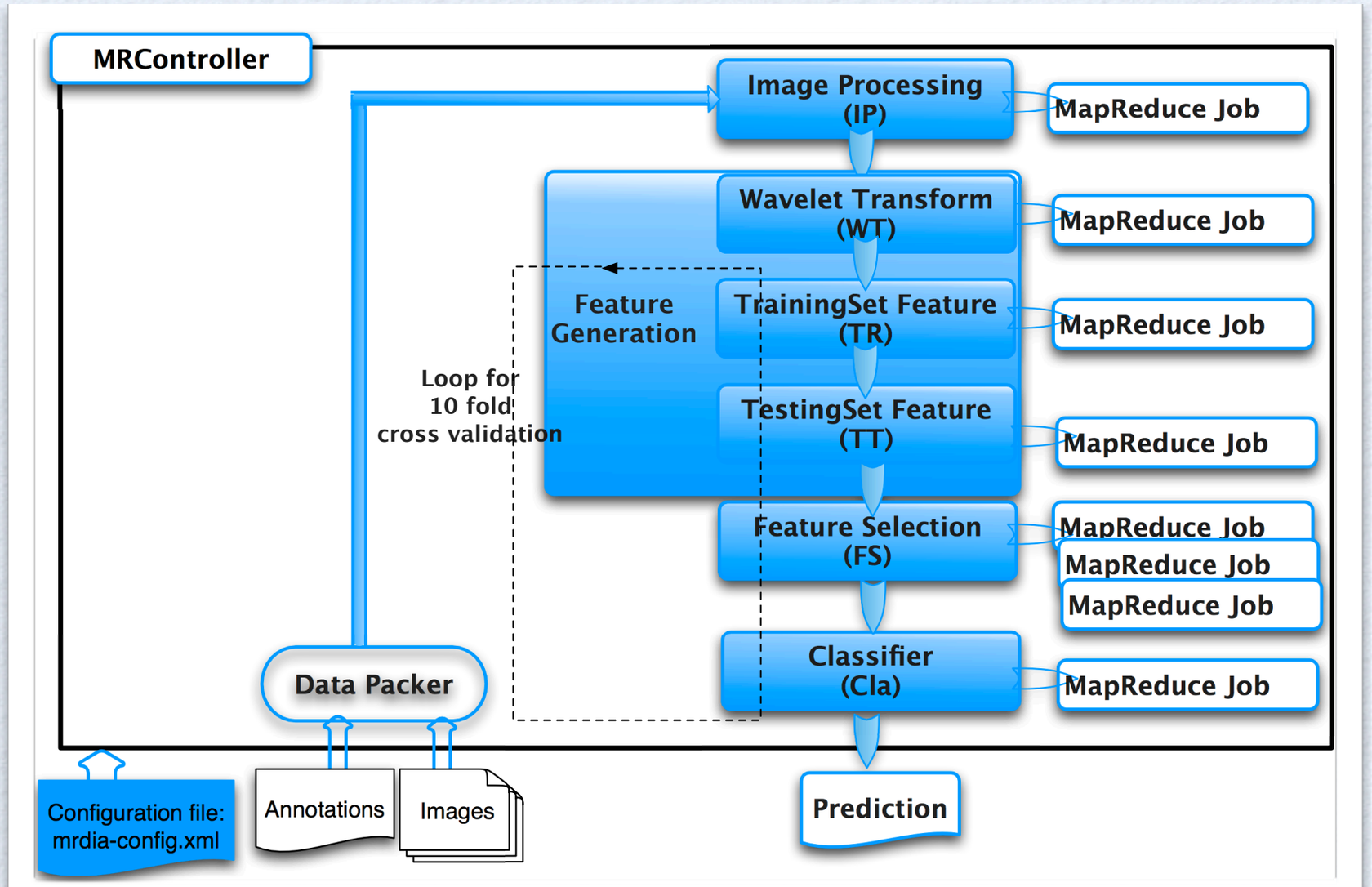
- Data Parallelism: workload are distributed into different computing nodes and the same task can be executed on different subsets of the data simultaneously
- Task Parallelism: tasks are independent and can be executed purely in parallel
- Pipelining: an iteration of a task consisting of many stages, where each stage in the task is chained and executed in order and the output of one stage is the input of the next one.

# PARALLELISATION EXPLORATION-III

- MapReduce Model and Cloud computing:
  - Adapt the data mining algorithms to MapReduce model
  - Evaluate it in the Cloud (based on IaaS model)
  - A paper accepted by IEEE/ACM Grid Computing 2012

# PARALLELISATION EXPLORATION-IV

- Adaptation of the data mining task to MapReduce Model



# PARALLELISATION EXPLORATION-IV

- Adaptation of the data mining task to MapReduce Model

  - MRC(MapReduce) Controller: specify data files and annotation file locations, as well as the output files of data mining results. It reads the system configuration file(mrdia-config.xml) about the runtime properties (e.g., nodes, locations of the intermediary results, the k value of the k-fold cross validation)

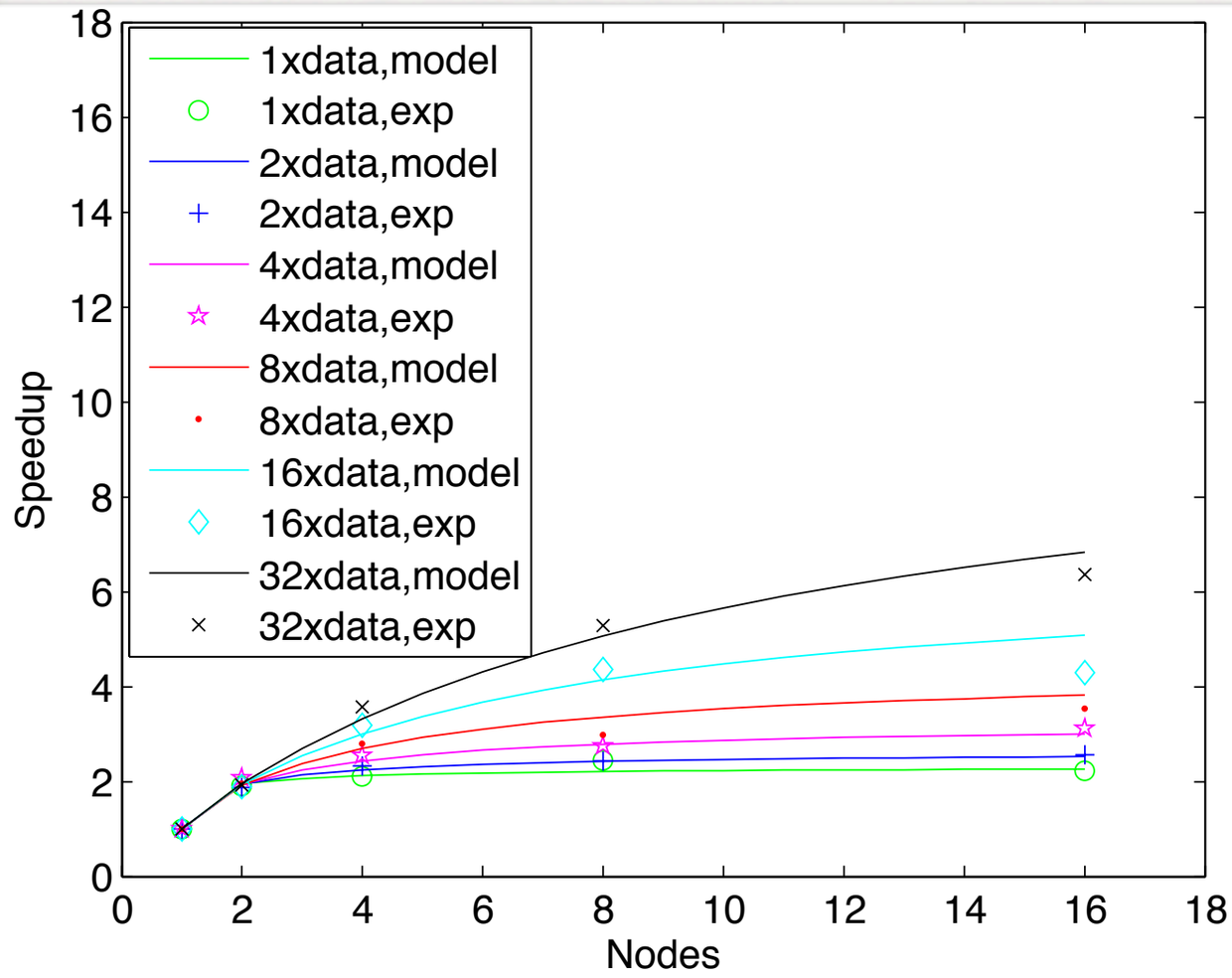
  - Data packer is to read the raw data and repack it into the desired data structure for MapReduce jobs. It acts as an initial workload distributor

# PARALLELISATION EXPLORATION-V

- Performance evaluation in the Cloud

Table I
A VIRTUAL INSTANCE CONFIGURATION

| Item | Configuration |
|---|---|
| CPU | 4 EC2 Compute units: 2 virtual cores with 2 EC 2compute Units. Note: one EC2 Compute unit provides the equivalent CPU capacity of a 1.0-1.2 GHZ 2007 Opteron or 2007 Xeon processor. |
| Memory | 7.5 GB memory |
| Storage | 850 GB instance storage (2x420GB plus 10 GB root partition) |
| I/O performance | High |
| Operating System | Federa Core 8 (2.6.21.7-2.ec2.v1.2.fc8xen Linux Kernel) |



# PARALLELISATION EXPLORATION-VI

- Performance evaluation in the Cloud

SPEEDUP OF EACH MAPREDUCE JOB

| | Node1 | Node2 | Node4 | Node8 | Node16 |
|---|---|---|---|---|---|
| IP | 1 | 1.97 | 3.86 | 7.16 | 10.95 |
| WT | 1 | 2.03 | 3.73 | 5.99 | 10.32 |
| TT | 1 | 2.39 | 3.54 | 5.58 | 6.97 |
| TR | 1 | 2.12 | 3.80 | 5.39 | 9.84 |
| MV | 1 | 1.76 | 2.86 | 2.71 | 2.05 |
| FR | 1 | 3.06 | 3.48 | 3.48 | 3.37 |
| TN | 1 | 1.89 | 3.96 | 2.12 | 2.12 |
| KNN | 1 | 1.75 | 3.64 | 6.39 | 9.50 |

# PARALLELISATION EXPLORATION-VII

- Cost estimation: considering storage cost and using large instances

| Item Description | Unit Cost (US$) | Unit Needed/Consumed | Total Cost (US$) |
|---|---|---|---|
| Amazon EC2 | | | *Total: 73.61* |
| • *Large Instance Type with Linux* | *0.38/hr* | *157 hr* | *59.66* |
| • *Storage Cost* | *0.11/GB-Month* | *126.794 GB-Month* | *13.95* |
| Privately-Owned Cluster* | | | |
| • *Dell Inspiron Zino HD* | *478.99/pc* | *12* | *5747.88* |

# PARALLELISATION EXPLORATION-VIII

- A cheaper solution than buying your big machines when performing experiments/tests (very suitable for pilot study)

- MapReduce model and Cloud computing can improve the performance. The speedup depends on the parallelisation of algorithms

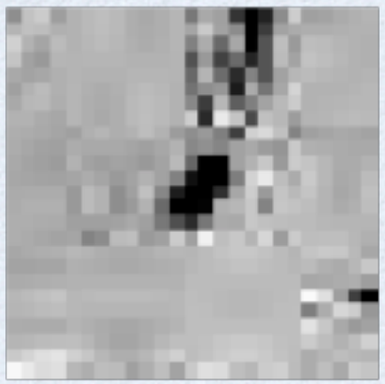
- Not just press a button, you need a precooked prototype (especially when adopting IaaS model)

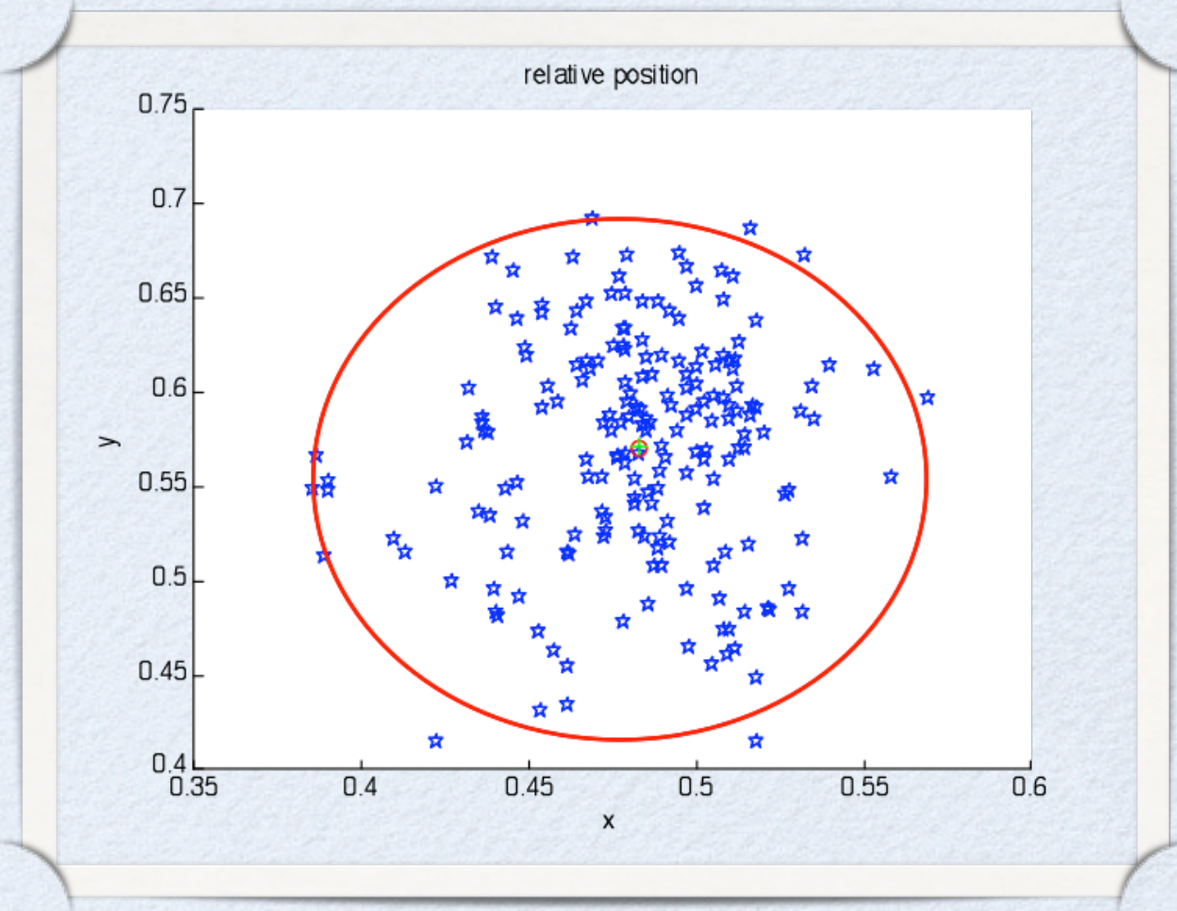
# OUTLINE

- What is Annotation?
- Motivation and Challenges
- Methodology
- Experimental and Evaluation
- Exploration of Parallelisation
- Ongoing Work

# ONGOING WORK- 1

- Data mining side
  - Feature extraction ( locate the region, process the region of the image, and reduce computing time)





# ONGOING WORK- II

- Parallelisation and cloud computing side
  - Development of parallelisation algorithms
  - Development of data-reuse mechanism for cost-effective and optimisation of the data intensive applications running in the Cloud

- Funding support from BBSRC ( in collaboration with MRC, Prof. Richard Baldock)

# Acknowledgement and Collaborators


- Prof. Malcolm Atkinson: University of Edinburgh
- Dr. Jano van Hemert: Optos PLC
- Prof. Richard Baldock: Medical Research Council, Human Genetics Unit, Edinburgh
- ADMIRE project (EU-FP7),http://www.admire-project.eu

# FUNDS Research Group

- Group members: Liangxiu Han, Thar Baker, Mohammad Hammoudeh, Andy Nisbet, Maybin Muybe and a RA to be recruited
- Associate member: Darren Dancey
- Students:
  - 7 PhDs enrolled
  - 4 Msc students

THANK YOU