# The Lecturing Team

<table>
<thead>
<tr>
<th>Role</th>
<th>Members</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coordinator</td>
<td>Sanjay Chawla, SIT</td>
</tr>
<tr>
<td>Lecturer</td>
<td>Sanjay Chawla and Wei Liu (NICTA)</td>
</tr>
<tr>
<td>Tutors</td>
<td>Didi Surian, Linsey Pang and Fei Wang (PhD Students)</td>
</tr>
</tbody>
</table>
Material and Lectures

- Lectures will be posted on http://www.it.usyd.edu.au/~comp5318

- We will mainly follow the textbook from Rajaram, Leskovic and Ullman from Stanford which is available online:

- However the ordering will be different
Assessment Package

<table>
<thead>
<tr>
<th>Assessment</th>
<th>Week</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-Class Test (15%)</td>
<td>Week 6</td>
</tr>
<tr>
<td>Group Assignment (20%)</td>
<td>Week 10</td>
</tr>
<tr>
<td>Research Paper Presentation</td>
<td>Week 11 - 12</td>
</tr>
<tr>
<td>(15%)</td>
<td></td>
</tr>
<tr>
<td>Final Exam (50%)</td>
<td>See Exam Calendar</td>
</tr>
</tbody>
</table>

To pass the class you must get at least 50% in the final exam; and at least 40% (combined) in other assessments.
So what is this class about....

1. Building large knowledge discovery pipelines to solve real-world problems (a.k.a "Data Analytic" pipelines)
2. Learning about techniques to analyze and algorithms to mine data
3. Learn how to read original research in data mining and machine learning
4. Learn how to solve large data problems in the cloud
Coordinates of Data Mining

- Algorithms/ Distributed & Parallel Computing
- Statistics and Linear Algebra
- Machine Learning
- Database Management Systems
- Computer Vision
- Information Retrieval
Abstract Tasks in Data Mining

- **Clustering and Segmentation**: how to automatically group objects into clusters
  - Take photographs from Flickr and automatically create categories
- **Classification and Regression**: how to make statistical models for prediction.
  - Predict whether an online user will click on a banner advertisement
  - Predict the currency exchange rate tomorrow (AUS/USD)
  - Predict who will win the NBA champion in 2013
Abstract Tasks in Data Mining....cont

● 3. Anomaly Detection: Identify entities which are different from the rest in the group
  ○ While galaxy is different in an astronomical database
  ○ Which area has an unusual flu rate
  ○ Is this credit card transaction fraudulent?
  ○ Identify cyber attacks: Denial of Service (DOS) and Portscan
  ○ Identify genes which are likely to cause a certain disease
Knowledge Discovery Pipeline

Data Source → Data Integration → Data Mining Task → Presentation of Results
Example: Large Scale Advertising Systems

Publisher Web Page

<user>

<uid, url>

<user>

<uid, url>

<ad, bid>

Ad Exchange

<ad, bid>

<ad, bid>

Demand Side Platform

<ad, bid>

<ad, bid>

<ad, bid>

Ad Server

<ad, bid>

<ad, bid>

<ad, bid>

Advertiser 1

Advertiser n
Let's do something tangible...

Underlying all data mining tasks...is the notion of similarity.

1. When are two images similar?
2. When are two documents similar?
3. When are two patients similar?
4. When are two shopping-baskets similar?
5. When are two job candidate similar?
6. When are two galaxies similar?
7. When is network traffic similar?
Data Vector

In Data Mining, data is often transformed to a vector of numbers. e.g.,

*D1: computer science and physics have a lot in common. In the former, we build models of computation and in the latter, models of the physical world.*

<table>
<thead>
<tr>
<th>a</th>
<th>the</th>
<th>brain</th>
<th>latter</th>
<th>of</th>
<th>world</th>
<th>cheese</th>
<th>in</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

What is the length of this vector?
Data Vector....cont

www.sydney.visitorsbureau.com.au

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>45</td>
<td>6</td>
<td>6</td>
<td>12</td>
<td>33</td>
<td>22</td>
<td>17</td>
</tr>
</tbody>
</table>

700 x 500

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>45</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>17</td>
<td>44</td>
<td></td>
</tr>
</tbody>
</table>
Data Vector...cont

<table>
<thead>
<tr>
<th>1.022</th>
<th>1.01</th>
<th>1.002</th>
<th>1.01</th>
<th>1.01</th>
<th>1.01</th>
<th>1.02</th>
<th>1.03</th>
<th>1.03</th>
<th>1.03</th>
</tr>
</thead>
</table>

![AUD/USD Chart](image)
Similarity

- Once we have data vectors, we can start the computation process...for example,..when are two data vectors similar?

While pair of currency trades are more similar?
Similarity Computation

Suppose want to compute similarity between two vectors:

\[ x = <3,4,1,2>; \quad y = <1,2,3,1> \]

**Step 1:** compute the length of each vector:

\[ ||x|| = \sqrt{3^2 + 4^2 + 1^2 + 2^2} = \sqrt{9 + 16 + 1 + 4} = 5.48 \]

\[ ||y|| = \sqrt{1^2 + 2^2 + 3^2 + 1^2} = \sqrt{1 + 4 + 9 + 1} = 3.87 \]

**Step 2:** compute the dot product:

\[ x \cdot y = 3.1 + 4.2 + 1.3 + 2.1 = 3 + 8 + 3 + 2 = 16 \]

**Step 3:**

\[ \frac{x \cdot y}{||x|| \cdot ||y||} = \frac{16}{(5.48)(3.87)} = 0.75 \]
Cosine Similarity

1. Thus similarity \( \text{sim}(x, y) \) between two data vectors \( x \) and \( y \) is given by \( x \cdot y / (\|x\| \cdot \|y\|) \)

2. This is called cosine similarity (Why ?)

3. This is a very general concept and underpins much of data-driven computation

4. We will be coming back to it over and over again
More examples

\[ x = \langle 1,0,1,0 \rangle; \quad y = \langle 0,1,0,1 \rangle \]
\[ \text{sim}(x,y) = 0 \]

\[ x = \langle 1,3,2,1 \rangle; \quad y = \langle 1,3,2,1 \rangle \]
\[ \text{sim}(x,y) = 1 \]

If all elements of data vector are non-negative, then: \[ 0 \leq \text{sim}(x,y) \leq 1 \]
Cost of Computation

\[ \mathbf{x} = <3, 4, 1, 2>; \mathbf{y} = <1, 2, 3, 1> \]

**Step 1:** compute the length of each vector

\[
\|\mathbf{x}\| = 3^2 + 4^2 + 1^2 + 2^2 = 9 + 16 + 1 + 4 = 30 \quad [4 \text{ mult}; 3 \text{ adds} + 1 \text{ sqrt}]
\]

\[
\|\mathbf{y}\| = 1^2 + 2^2 + 3^2 + 1^2 = 1 + 4 + 9 + 1 = 15 \quad [4 \text{ mult}; 3 \text{ adds} + 1 \text{ sqrt}]
\]

**Step 2:** compute the dot product:

\[
\mathbf{x} \cdot \mathbf{y} = 3 \cdot 1 + 4 \cdot 2 + 1 \cdot 3 + 2 \cdot 1 = 3 + 8 + 3 + 2 = 16 \quad [4 \text{ mult}; \text{ adds}]
\]

**Step 3:** \((\mathbf{x} \cdot \mathbf{y}) / (\|\mathbf{x}\| \|\mathbf{y}\|) = (16/30.15) = 0.036 \quad [1 \text{ mult}, 1 \text{ divide}]
\]

Total FLOPS (assuming 1 FLOP per operation) = \(d + d + d + d + (d-1) + 1 + 1 = 6d + 1\)
Cost of Computation

Cost of Similarity between two vectors of length d, is 6d+1 or ~ 6d.

Suppose want to find the similarity between all Wikipedia documents.  
  Number of Articles: ~4,000,000  
Length of data vector: ~100,000 [# of words in dictionary]  
Number of pairwise combinations: ~ 8 x 10^{12}  
Number of flops: ~ 8 x 10^{12} x 6 x 10^5 = 48 x 10^{17} ~ 10^{18}  
World's fastest computer (2012); Titan at Oak Ridge Labs: 27 peta flops (27 thousand, trillion flops; 10^{16}) [100 seconds]  
Current desktop: 3 Ghz; ~10^9 flops per second. Thus 10^9 seconds ~ 33 yrs.
Summary

We use data mining to build knowledge discovery pipelines.

Data Mining is the process of applying algorithms to data.

A key concept is that of defining similarity between entities.