Is BIG Necessarily Better?
A guest lecture for CS5344, 21/3/2014

Wong Limsoon
More data may not be better…
A few stories

- Discovering protein complexes from PPIN
- Identifying causal genes, Part 1
- Identifying causal genes, Part 2
- Finding interesting patterns
Protein-protein interaction networks

- Proteins come together & interact
- The collection of these interactions form a Protein Interaction Network or PPIN

Collection of such interactions in an organism

PPIN
Valuable source of knowledge

Protein Interaction Network
Detection & analysis of protein complexes in PPIN

PPIN derived from several high-throughput expt

Space-time info is lost

Identifying embedded complexes

Individual complexes (Some might share proteins)

Entire module might be involved in the same function/process

Space-time info is “recovered”

Embedded complexes identified from PPIN

Identifying embedded complexes
Difficulties

• Protein complexes are discovered from PPIN by, e.g., clustering approaches

• But success has been limited
  – Noise in PPI data
    • Spuriously-detected interactions (false positives), and missing interactions (false negatives)
  – Transient interactions
    • Many proteins that actually interact are not from the same complex, they bind temporarily to perform a function
  – Also, not all proteins in the same complex may actually interact with each other
Cytochrome BC1 Complex

- **Involved in electron-transport chain in mitochondrial inner membrane**

- **Discovery of BC1 from PPI data is difficult**
  - Sparseness of its PPI subnetwork
    - Only 19 out of 45 possible interactions were detected between the complex’s proteins
  - Extraneous interactions with other proteins outside the complex
    - E.g., UBI4 is involved in protein ubiquitination, and binds to many proteins to perform its function
Perhaps “big data” can help?

- **Composite network**
  - Vertices represent proteins, edges represent relationships between proteins. Put an edge between proteins u, v, if u and v are related according to any of the data sources.

<table>
<thead>
<tr>
<th>Data source</th>
<th>Database</th>
<th>Scoring method</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPI</td>
<td>BioGRID, IntACT, MINT</td>
<td>Iterative AdjustCD.</td>
</tr>
<tr>
<td>L2-PPI (indirect PPI)</td>
<td>BioGRID, IntACT, MINT</td>
<td>Iterative AdjustCD</td>
</tr>
<tr>
<td>Functional association</td>
<td>STRING</td>
<td>STRING</td>
</tr>
<tr>
<td>Literature co-occurrence</td>
<td>PubMed</td>
<td>Jaccard coefficient</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data source</th>
<th># Pairs</th>
<th>% co-complex</th>
<th>coverage</th>
<th># Pairs</th>
<th>% co-complex</th>
<th>coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPI</td>
<td>106328</td>
<td>5.8%</td>
<td>55%</td>
<td>48098</td>
<td>10%</td>
<td>14%</td>
</tr>
<tr>
<td>L2-PPI</td>
<td>181175</td>
<td>1.1%</td>
<td>18%</td>
<td>131705</td>
<td>5.5%</td>
<td>20%</td>
</tr>
<tr>
<td>STRING</td>
<td>175712</td>
<td>5.7%</td>
<td>89%</td>
<td>311435</td>
<td>3.1%</td>
<td>27%</td>
</tr>
<tr>
<td>PubMed</td>
<td>161213</td>
<td>4.9%</td>
<td>70%</td>
<td>91751</td>
<td>4.3%</td>
<td>11%</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td>531800</td>
<td><strong>2.1%</strong></td>
<td><strong>98%</strong></td>
<td>522668</td>
<td><strong>3.4%</strong></td>
<td><strong>49%</strong></td>
</tr>
</tbody>
</table>
More is not always better, unless...

While proteins in BC1 become fully connected in the composite network, there is a blow-up in extraneous proteins. So clustering won’t discover the complex, unless you know how to remove the extraneous proteins.
A few stories

• Discovering protein complexes from PPIN

• Identifying causal genes, Part 1

• Identifying causal genes, Part 2

• Finding interesting patterns
Microarray

Source: Affymetrix
Application: Disease subtype diagnosis

genes

samples

benign
benign
benign
benign
malign
malign
malign
malign

???
Application: Drug action detection

Which group of genes are the drug affecting on?
Typical analysis workflow

- Gene expression data collection
- DE gene selection by, e.g., t-statistic
- Classifier training based on selected DE genes
- Apply the classifier for diagnosis of future cases


Terminology: DE gene = differentially expressed gene
Hierarchical clustering

Percentage of overlapping genes

- Low % of overlapping genes from diff expt in general
  - Prostate cancer
    - Lapointe et al, 2004
    - Singh et al, 2002
  - Lung cancer
    - Garber et al, 2001
    - Bhattacharjee et al, 2001
  - DMD
    - Haslett et al, 2002
    - Pescatori et al, 2007

<table>
<thead>
<tr>
<th>Datasets</th>
<th>DEG</th>
<th>POG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate Cancer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>Top100</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>Lung Cancer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Top100</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>DMD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>Top100</td>
<td>0.54</td>
<td></td>
</tr>
</tbody>
</table>

“Most random gene expression signatures are significantly associated with breast cancer outcome”

Too many genes is bad, Need to eliminate irrelevant ones

• Suppose
  – Each gene has 50% chance to be high
  – You have 3 disease and 3 normal samples

• How many genes on a microarray are expected to perfectly correlate to these samples?

• Prob(a gene is correlated) = 1/2^6
• # genes on array = 30k
• E(# of correlated genes) = 469

⇒ Many false positives
• These cannot be eliminated based on pure statistics!
The situation is worse for people looking for genetic mutations that cause a disease.

- 10,000,000 SNPs, with 5% MAF
- 3 control vs 3 disease samples
- $\text{Prob}(\text{a SNP is correlated}) = 0.0001$
- $E(\text{# of correlated SNP}) = 1,071$
A few stories

• Discovering protein complexes from PPIN
• Identifying causal genes, Part 1
• Identifying causal genes, Part 2
• Finding interesting patterns
Biology to the rescue: Gene Regulatory Circuits

- Each disease phenotype has some underlying cause
- There is some unifying biological theme for genes that are truly associated with a disease subtype
- Uncertainty in selected genes can be reduced by considering biological processes of the genes
- The unifying biological theme is basis for inferring the underlying cause of disease subtype
<table>
<thead>
<tr>
<th>Database</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEGG</td>
<td>KEGG (<a href="http://www.genome.jp/kegg">http://www.genome.jp/kegg</a>) is one of the best known pathway databases (Kanehisa et al., 2010). It consists of 16 main databases, comprising different levels of biological information such as systems, genomic, etc. The data files are downloadable in XML format. At time of writing it has 392 pathways.</td>
</tr>
<tr>
<td>WikiPathways</td>
<td>WikiPathways (<a href="http://www.wikipathways.org">http://www.wikipathways.org</a>) is a Wikipedia-based collaborative effort among various labs (Kelder et al., 2009). It has 1,627 pathways of which 369 are human. The content is downloadable in GPML format.</td>
</tr>
<tr>
<td>Reactome</td>
<td>Reactome (<a href="http://www.reactome.org">http://www.reactome.org</a>) is also a collaborative effort like WikiPathways (Vastrik et al., 2007). It is one of the largest datasets, with over 4,166 human reactions organized into 1,131 pathways by December 2010. Reactome can be downloaded in BioPax and SBML among other formats.</td>
</tr>
<tr>
<td>Pathway Commons</td>
<td>Pathway Commons (<a href="http://www.pathwaycommons.com">http://www.pathwaycommons.com</a>) collects information from various databases but does not unify the data (Cerami et al., 2006). It contains 1,573 pathways across 564 organisms. The data is returned in BioPax format.</td>
</tr>
<tr>
<td>PathwayAPI</td>
<td>PathwayAPI (<a href="http://www.pathwayapi.com">http://www.pathwayapi.com</a>) contains over 450 unified human pathways obtained from a merge of KEGG, WikiPathways and Ingenuity® Knowledge Base (Soh et al., 2010). Data is downloadable as a SQL dump or as a csv file, and is also interfaceable in JSON format.</td>
</tr>
</tbody>
</table>
Human apoptosis pathway

<table>
<thead>
<tr>
<th>Apoptosis Pathway</th>
<th>Wiki x KEGG</th>
<th>Wiki x Ingenuity</th>
<th>KEGG x Ingenuity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Pair Count:</td>
<td>144 vs 172</td>
<td>144 vs 3557</td>
<td>172 vs 3557</td>
</tr>
<tr>
<td>Gene Count:</td>
<td>85 vs 80</td>
<td>85 vs 176</td>
<td>80 vs 176</td>
</tr>
<tr>
<td>Gene Overlap:</td>
<td>38</td>
<td>28</td>
<td>30</td>
</tr>
<tr>
<td>Gene % Overlap:</td>
<td>48%</td>
<td>33%</td>
<td>38%</td>
</tr>
<tr>
<td>Gene Pair Overlap:</td>
<td>23</td>
<td>14</td>
<td>24</td>
</tr>
<tr>
<td>Gene Pair % Overlap:</td>
<td>16%</td>
<td>10%</td>
<td>14%</td>
</tr>
</tbody>
</table>


- The various data sources have low overlap
  ⇒ Good to unify them to get more complete pathways, right?
A unified database of biological pathways

**IntPath**
Integrated pathway gene relationship database

**IDENTIFY PATHWAYS**
FIND THE MOST SIGNIFICANT ONES

The function of the "Identify Pathways" uses hyper-geometric test to find the most significant pathways of the input gene lists. Through this test, users can have a clear insight of which pathway is most related to the input gene list.

Welcome to IntPath
IntPath is a pathway gene relationship database that integrates data from KEGG, WikiPathways, BioCyc. Currently, the following organisms are included: Homo sapiens, Mus musculus, Saccharomyces cerevisiae, and Mycobacterium tuberculosis H37Rv. A unified database of included organisms can be downloaded here, and Application Programming Interface (API) is also supported. IntPath also provides tools to "Identify Pathways" (single gene list analysis) and "Analyze Distances" (dual gene lists analysis) based on the methods described in Wilson Goh et al., and Donny Soh et al.

About us
IntPath database is developed by Computational Biology Lab in School of Computing of National University of Singapore. Principle Investigator: Professor Limsoon Wong, Database Administrator: Hufeng Zhou.
Computational Biology Lab
School of Computing
National University of Singapore
COM1, Room 51-10, NUS, Singapore
Email: ComBio.NUS@gmail.com

All Rights Reserved © 2011, Computational Biology Lab.

Using biology background: GSEA

• “Enrichment score”
  – The degree that the genes in gene set C are enriched in the extremes of ranked list of all genes
  – Measured by Komogorov-Smirnov statistic

• Null distribution to estimate the p-value of the scores above is by randomizing patient class labels

Subramanian et al., *PNAS*, 102(43):15545-15550, 2005
Unfortunately, it doesn’t always work

Table 2. Table showing the number and percentage of significant overlapping genes. $\gamma$ refers to the number of genes compared against and is the number of unique genes within all the significant subnetworks of the disease datasets. The percentages refer to the percentage gene overlap for the corresponding algorithms.

<table>
<thead>
<tr>
<th>Disease</th>
<th>$\gamma$</th>
<th>SNet</th>
<th>GSEA</th>
<th>SAM</th>
<th>t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leuk</td>
<td>84</td>
<td>91.3%</td>
<td>2.4%</td>
<td>22.6%</td>
<td>14.3%</td>
</tr>
<tr>
<td>Subtype</td>
<td>75</td>
<td>93.0%</td>
<td>4.0%</td>
<td>49.3%</td>
<td>57.3%</td>
</tr>
<tr>
<td>DMD</td>
<td>45</td>
<td>69.2%</td>
<td>28.9%</td>
<td>42.2%</td>
<td>20.0%</td>
</tr>
<tr>
<td>Lung</td>
<td>65</td>
<td>51.2%</td>
<td>4.0%</td>
<td>24.6%</td>
<td>26.2%</td>
</tr>
</tbody>
</table>

- Surprisingly, GSEA fails on large unified pathways!
More is not always better, unless ...

- Need to know how to capture the subnetwork branch within the pathway

A branch within pathway consisting of genes A, B, C, D and E are high in phenotype $X$

Genes C, D and E not high in phenotype $\neg X$

30 other genes not diff expressed

GSEA: Entire network is likely to be missed
A few stories

- Discovering protein complexes from PPIN
- Identifying causal genes, Part 1
- Identifying causal genes, Part 2
- Finding interesting patterns
Different rules may be produced,
Different drugs may be considered better

- **Mining only men’s data**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>lived</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>died</td>
<td>80</td>
<td>160</td>
</tr>
</tbody>
</table>

- You get rules like
  - “drug A → lived”
    - Supp = \(\frac{20}{310} = 6\%\)
    - Conf = \(\frac{20}{100} = 20\%\)
  - “drug B → live”
    - Supp = \(\frac{50}{310} = 16\%\)
    - Conf = \(\frac{50}{160} = 31\%\)

- **Mining combined data**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>lived</td>
<td>40</td>
<td>15</td>
</tr>
<tr>
<td>died</td>
<td>20</td>
<td>5</td>
</tr>
</tbody>
</table>

- You get rules like
  - “drug A → lived”
    - Supp = \(\frac{60}{390} = 2\%\)
    - Conf = \(\frac{60}{160} = 38\%\)
  - “drug B → lived”
    - Supp = \(\frac{65}{390} = 17\%\)
    - Conf = \(\frac{65}{180} = 36\%\)
Looks like treatment A is better

<table>
<thead>
<tr>
<th>Overall</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>lived</td>
<td>60</td>
<td>65</td>
</tr>
<tr>
<td>died</td>
<td>100</td>
<td>165</td>
</tr>
</tbody>
</table>

Looks like treatment B is better

<table>
<thead>
<tr>
<th>Women</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>lived</td>
<td>40</td>
<td>15</td>
</tr>
<tr>
<td>died</td>
<td>20</td>
<td>5</td>
</tr>
</tbody>
</table>

Looks like treatment A is better

<table>
<thead>
<tr>
<th>Men</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>lived</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>died</td>
<td>80</td>
<td>160</td>
</tr>
</tbody>
</table>

Challenge: Separating causal factors from confounding factors; making good inferences.
Rules are just rules, unless …

Data mining sensor & telemetry data in a factory may give you rules like …

- Fuse blow $\rightarrow$ Robot stop
- ... a thousand other rules ...
- Circuit overload $\rightarrow$ Fuse blow
- ... a thousand other rules ...
- Insufficient lubrication $\rightarrow$ Circuit overload
- ... a thousand other rules ...
- Oil pump clogged $\rightarrow$ Insufficient lubrication
- ... a thousand other rules ...
- Metal shavings $\rightarrow$ Oil pump clogged

Challenge: Asking “why” 5 levels deep, and getting to the root cause.
What have we learned?

• More data can offer a more complete picture, fill in gaps, etc.

• More data can also introduce noise into an analysis

• Unless you know how to tame this noise, more data may not lead to a better analysis
How we can get more out of big data
A few suggestions

• Look deeper into your $\chi^2$ test statistic

• Explore more stratifications of your data

• Know when to discard the 1$^{\text{st}}$ PC in PCA

• Make sure you get the null hypothesis right, and exploit domain knowledge properly
Comparison betw proportions

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Improvement</th>
<th>No improvement</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arthritic drug</td>
<td>18</td>
<td>6</td>
<td>24</td>
</tr>
<tr>
<td>placebo</td>
<td>9</td>
<td>11</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>27</td>
<td>17</td>
<td>44</td>
</tr>
</tbody>
</table>

- Proportion improved in drug group = $\frac{18}{24} = 75\%$
- Proportion improved in placebo group = $\frac{9}{20} = 45.0\%$

- Question: What is the probability that the observed difference of 30% is purely due to sampling error, i.e. chance in sampling?
- Use $\chi^2$ test
\( \chi^2 \) test for statistical association

<table>
<thead>
<tr>
<th>treatment</th>
<th>Improvement</th>
<th>No improvement</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arthritic drug</td>
<td>18 (a)</td>
<td>6 (b)</td>
<td>24</td>
</tr>
<tr>
<td>placebo</td>
<td>9 (c)</td>
<td>11 (d)</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>27</td>
<td>17</td>
<td>44</td>
</tr>
</tbody>
</table>

- Prob of selecting a person in drug group = \( \frac{24}{44} \)
- Prob of selecting a person with improvement = \( \frac{27}{44} \)
- Prob of selecting a person from drug group who had shown improvement = \( \frac{24}{44} \times \frac{27}{44} = 0.3347 \) (assuming two independent events)
- Expected value for cell (a) = \( 0.3347 \times 44 = 14.73 \)
**χ² test for statistical association**

<table>
<thead>
<tr>
<th>treatment</th>
<th>Improvement</th>
<th>No improvement</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arthritic drug</td>
<td>18 (14.73)</td>
<td>6 (9.27)</td>
<td>24</td>
</tr>
<tr>
<td>placebo</td>
<td>9 (12.27)</td>
<td>11 (7.73)</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>27</td>
<td>17</td>
<td>44</td>
</tr>
</tbody>
</table>

- **General formula for χ²**

\[
\chi^2 = \sum \frac{(obs - exp)^2}{exp}
\]

- **Note:** χ² test is always performed on categorical variables using absolute frequencies, never percentage or proportion
\( \chi^2 \) test for statistical association

- **For the given problem:**
  \[
  \sum \frac{(obs - exp)^2}{exp} = \frac{(18 - 14.73)^2}{14.73} + \frac{(6 - 9.27)^2}{9.27} + \frac{(9 - 12.27)^2}{12.27} + \frac{(11 - 7.73)^2}{7.73}
  \]
  \[
  = 4.14 \text{ with 1 degree of freedom}
  \]

- **\( \chi^2 \) degree of freedom is given by:**
  \[
  (\text{no. of rows-1})*(\text{no. of cols-1})
  \]
  \[
  = (2-1)*(2-1) = 1
  \]

How many of these 4 cells are free to vary if we keep the row and column totals constant?
$\chi^2$ table

Critical values in the distributions of chi-squared for different degrees of freedom

<table>
<thead>
<tr>
<th>df</th>
<th>.05</th>
<th>.02</th>
<th>.01</th>
<th>.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.841</td>
<td>5.412</td>
<td>6.635</td>
<td>10.827</td>
</tr>
<tr>
<td>2</td>
<td>5.991</td>
<td>7.824</td>
<td>9.210</td>
<td>13.815</td>
</tr>
<tr>
<td>3</td>
<td>7.815</td>
<td>9.837</td>
<td>11.345</td>
<td>16.266</td>
</tr>
<tr>
<td>4</td>
<td>9.488</td>
<td>11.668</td>
<td>13.277</td>
<td>18.467</td>
</tr>
<tr>
<td>5</td>
<td>11.070</td>
<td>13.388</td>
<td>15.086</td>
<td>20.515</td>
</tr>
<tr>
<td>6</td>
<td>12.592</td>
<td>15.033</td>
<td>16.812</td>
<td>22.457</td>
</tr>
<tr>
<td>7</td>
<td>14.067</td>
<td>16.622</td>
<td>18.475</td>
<td>24.322</td>
</tr>
<tr>
<td>8</td>
<td>15.507</td>
<td>18.168</td>
<td>20.090</td>
<td>26.125</td>
</tr>
<tr>
<td>9</td>
<td>16.919</td>
<td>19.679</td>
<td>21.666</td>
<td>27.877</td>
</tr>
<tr>
<td>10</td>
<td>18.307</td>
<td>21.161</td>
<td>23.209</td>
<td>29.588</td>
</tr>
<tr>
<td>11</td>
<td>19.675</td>
<td>22.618</td>
<td>24.725</td>
<td>31.264</td>
</tr>
<tr>
<td>12</td>
<td>21.026</td>
<td>24.054</td>
<td>26.217</td>
<td>32.909</td>
</tr>
<tr>
<td>13</td>
<td>22.362</td>
<td>25.372</td>
<td>27.688</td>
<td>34.528</td>
</tr>
<tr>
<td>14</td>
<td>23.685</td>
<td>26.683</td>
<td>29.141</td>
<td>36.123</td>
</tr>
<tr>
<td>15</td>
<td>24.996</td>
<td>28.259</td>
<td>30.578</td>
<td>37.697</td>
</tr>
<tr>
<td>16</td>
<td>26.296</td>
<td>29.633</td>
<td>32.000</td>
<td>39.252</td>
</tr>
<tr>
<td>17</td>
<td>27.587</td>
<td>30.995</td>
<td>33.409</td>
<td>40.790</td>
</tr>
<tr>
<td>18</td>
<td>28.869</td>
<td>32.346</td>
<td>34.805</td>
<td>42.312</td>
</tr>
<tr>
<td>19</td>
<td>30.144</td>
<td>33.687</td>
<td>36.191</td>
<td>43.820</td>
</tr>
<tr>
<td>20</td>
<td>31.410</td>
<td>35.020</td>
<td>37.566</td>
<td>45.315</td>
</tr>
<tr>
<td>21</td>
<td>32.671</td>
<td>36.343</td>
<td>38.932</td>
<td>46.797</td>
</tr>
<tr>
<td>22</td>
<td>33.924</td>
<td>37.659</td>
<td>40.289</td>
<td>48.268</td>
</tr>
<tr>
<td>23</td>
<td>35.172</td>
<td>38.968</td>
<td>41.638</td>
<td>49.728</td>
</tr>
<tr>
<td>24</td>
<td>36.415</td>
<td>40.270</td>
<td>42.980</td>
<td>51.179</td>
</tr>
<tr>
<td>25</td>
<td>37.652</td>
<td>41.566</td>
<td>44.314</td>
<td>52.620</td>
</tr>
<tr>
<td>26</td>
<td>38.885</td>
<td>42.856</td>
<td>45.642</td>
<td>54.052</td>
</tr>
<tr>
<td>27</td>
<td>40.113</td>
<td>44.140</td>
<td>46.963</td>
<td>55.476</td>
</tr>
<tr>
<td>28</td>
<td>41.337</td>
<td>45.419</td>
<td>48.278</td>
<td>56.893</td>
</tr>
<tr>
<td>29</td>
<td>42.557</td>
<td>46.693</td>
<td>49.588</td>
<td>58.302</td>
</tr>
<tr>
<td>30</td>
<td>43.773</td>
<td>47.962</td>
<td>50.892</td>
<td>59.703</td>
</tr>
</tbody>
</table>

The observed $\chi^2$ value of 4.14 exceeds the critical value of 3.841 for $P=0.05$ but not the critical value of 5.412 for $P=0.02$ at 1 d.f.

i.e. $0.05 > P > 0.02$
\( \chi^2 \) test for statistical association

- Probability of getting an observed difference of 30% in improvement rates if the Null hypothesis of no association is correct is between 2% and 5%.

- Hence, there is some statistical evidence from this study to suggest that treatment of arthritic patient with the drug can significantly improve grip strength.
Extending to RxC tables

<table>
<thead>
<tr>
<th>Type of vaccines</th>
<th>Had flu</th>
<th>Avoided flu</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>43</td>
<td>237</td>
<td>280</td>
</tr>
<tr>
<td>II</td>
<td>52</td>
<td>198</td>
<td>250</td>
</tr>
<tr>
<td>III</td>
<td>25</td>
<td>245</td>
<td>270</td>
</tr>
<tr>
<td>IV</td>
<td>48</td>
<td>212</td>
<td>260</td>
</tr>
<tr>
<td>V</td>
<td>57</td>
<td>233</td>
<td>290</td>
</tr>
<tr>
<td>Total</td>
<td>225</td>
<td>1125</td>
<td>1350</td>
</tr>
</tbody>
</table>

- Null hypothesis assumes all vaccines tested had equal efficacy
Computation of the $\chi^2$

<table>
<thead>
<tr>
<th>Type of vaccines</th>
<th>Had flu</th>
<th>(O-E)$^2$/E</th>
<th>Avoided flu</th>
<th>(O-E)$^2$/E</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>43 (46.7)</td>
<td>0.293</td>
<td>237 (233.3)</td>
<td>0.059</td>
</tr>
<tr>
<td>II</td>
<td>52 (41.7)</td>
<td>2.544</td>
<td>198 (208.3)</td>
<td>0.509</td>
</tr>
<tr>
<td>III</td>
<td>25 (45.0)</td>
<td>8.889</td>
<td>245 (225.0)</td>
<td>1.778</td>
</tr>
<tr>
<td>IV</td>
<td>48 (43.3)</td>
<td>0.510</td>
<td>212 (216.7)</td>
<td>0.102</td>
</tr>
<tr>
<td>V</td>
<td>57 (48.3)</td>
<td>1.567</td>
<td>233 (241.7)</td>
<td>0.313</td>
</tr>
<tr>
<td>Total</td>
<td>225</td>
<td>13.803</td>
<td>1125</td>
<td>2.761</td>
</tr>
</tbody>
</table>

- $\chi^2 = 13.803 + 2.761 = 16.564$ with 4 d.f.
\[ \chi^2 \text{ table} \]

Critical values in the distributions of chi-squared for different degrees of freedom

<table>
<thead>
<tr>
<th>df</th>
<th>Probability .05</th>
<th>Probability .02</th>
<th>Probability .01</th>
<th>Probability .001</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.841</td>
<td>5.412</td>
<td>6.635</td>
<td>10.827</td>
</tr>
<tr>
<td>2</td>
<td>5.991</td>
<td>7.824</td>
<td>9.210</td>
<td>13.815</td>
</tr>
<tr>
<td>3</td>
<td>7.815</td>
<td>9.837</td>
<td>11.345</td>
<td>16.266</td>
</tr>
<tr>
<td>4</td>
<td>9.488</td>
<td>11.668</td>
<td>13.277</td>
<td>18.467</td>
</tr>
<tr>
<td>5</td>
<td>11.070</td>
<td>13.388</td>
<td>15.086</td>
<td>20.515</td>
</tr>
<tr>
<td>6</td>
<td>12.592</td>
<td>15.033</td>
<td>16.812</td>
<td>22.457</td>
</tr>
<tr>
<td>7</td>
<td>14.067</td>
<td>16.622</td>
<td>18.475</td>
<td>24.322</td>
</tr>
<tr>
<td>8</td>
<td>15.507</td>
<td>18.168</td>
<td>20.090</td>
<td>26.125</td>
</tr>
<tr>
<td>9</td>
<td>16.919</td>
<td>19.679</td>
<td>21.666</td>
<td>27.877</td>
</tr>
<tr>
<td>10</td>
<td>18.307</td>
<td>21.161</td>
<td>23.209</td>
<td>29.588</td>
</tr>
<tr>
<td>11</td>
<td>19.675</td>
<td>22.618</td>
<td>24.725</td>
<td>31.264</td>
</tr>
<tr>
<td>12</td>
<td>21.026</td>
<td>24.054</td>
<td>26.217</td>
<td>32.909</td>
</tr>
<tr>
<td>13</td>
<td>22.362</td>
<td>25.372</td>
<td>27.688</td>
<td>34.528</td>
</tr>
<tr>
<td>14</td>
<td>23.585</td>
<td>26.873</td>
<td>29.141</td>
<td>36.123</td>
</tr>
<tr>
<td>15</td>
<td>24.996</td>
<td>28.259</td>
<td>30.578</td>
<td>37.697</td>
</tr>
<tr>
<td>16</td>
<td>26.296</td>
<td>29.633</td>
<td>32.000</td>
<td>39.252</td>
</tr>
<tr>
<td>17</td>
<td>27.587</td>
<td>30.995</td>
<td>33.409</td>
<td>40.790</td>
</tr>
<tr>
<td>18</td>
<td>28.869</td>
<td>32.346</td>
<td>34.805</td>
<td>42.312</td>
</tr>
<tr>
<td>19</td>
<td>30.144</td>
<td>33.687</td>
<td>36.191</td>
<td>43.820</td>
</tr>
<tr>
<td>20</td>
<td>31.410</td>
<td>35.020</td>
<td>37.566</td>
<td>45.315</td>
</tr>
<tr>
<td>21</td>
<td>32.671</td>
<td>36.343</td>
<td>38.932</td>
<td>46.797</td>
</tr>
<tr>
<td>22</td>
<td>33.924</td>
<td>37.659</td>
<td>40.289</td>
<td>48.268</td>
</tr>
<tr>
<td>23</td>
<td>35.172</td>
<td>38.968</td>
<td>41.638</td>
<td>49.728</td>
</tr>
<tr>
<td>24</td>
<td>36.415</td>
<td>40.270</td>
<td>42.980</td>
<td>51.179</td>
</tr>
<tr>
<td>25</td>
<td>37.652</td>
<td>41.566</td>
<td>44.314</td>
<td>52.620</td>
</tr>
<tr>
<td>26</td>
<td>38.885</td>
<td>42.856</td>
<td>45.642</td>
<td>54.052</td>
</tr>
<tr>
<td>27</td>
<td>40.113</td>
<td>44.140</td>
<td>46.963</td>
<td>55.476</td>
</tr>
<tr>
<td>28</td>
<td>41.337</td>
<td>45.419</td>
<td>48.278</td>
<td>56.893</td>
</tr>
<tr>
<td>29</td>
<td>42.557</td>
<td>46.693</td>
<td>49.588</td>
<td>58.302</td>
</tr>
<tr>
<td>30</td>
<td>43.773</td>
<td>47.962</td>
<td>50.892</td>
<td>59.703</td>
</tr>
</tbody>
</table>

Observed \( \chi^2 \) value of 16.564 with 4 d.f. exceeds critical value of 13.277 for \( P=0.01 \) but not critical value of 18.467 for \( P=0.001 \).

I.e. \( 0.01 > P > 0.001 \)
Digging deeper

<table>
<thead>
<tr>
<th>Type of vaccines</th>
<th>Had flu</th>
<th>(O-E)^2/E</th>
<th>Avoided flu</th>
<th>(O-E)^2/E</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>43 (46.7)</td>
<td>0.293</td>
<td>237 (233.3)</td>
<td>0.059</td>
</tr>
<tr>
<td>II</td>
<td>52 (41.7)</td>
<td>2.544</td>
<td>198 (208.3)</td>
<td>0.509</td>
</tr>
<tr>
<td>III</td>
<td>25 (45.0)</td>
<td>8.889</td>
<td>245 (225.0)</td>
<td>1.778</td>
</tr>
<tr>
<td>IV</td>
<td>48 (43.3)</td>
<td>0.510</td>
<td>212 (216.7)</td>
<td>0.102</td>
</tr>
<tr>
<td>V</td>
<td>57 (48.3)</td>
<td>1.567</td>
<td>233 (241.7)</td>
<td>0.313</td>
</tr>
<tr>
<td>Total</td>
<td>225</td>
<td>13.803</td>
<td>1125</td>
<td>2.761</td>
</tr>
</tbody>
</table>

- Vaccine III contributes to the overall $\chi^2 = (8.889 + 1.778)/16.564 = 64.4\%$
\( \chi^2 \) with Vaccine III removed

<table>
<thead>
<tr>
<th>Type of vaccines</th>
<th>Had flu</th>
<th>Avoided flu</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>43</td>
<td>237</td>
<td>280</td>
</tr>
<tr>
<td>II</td>
<td>52</td>
<td>198</td>
<td>250</td>
</tr>
<tr>
<td>IV</td>
<td>48</td>
<td>212</td>
<td>260</td>
</tr>
<tr>
<td>V</td>
<td>57</td>
<td>233</td>
<td>290</td>
</tr>
</tbody>
</table>

- \( \chi^2 = 2.983 \) with 3 d.f.
- \( 0.1 < p < 0.5 \), not statistically significant

i.e., vaccine III is necessary for significance
Vaccine III vs. rest

- $\chi^2 = 12.7$ with 1 d.f.
- $P < 0.001$
- There appear to be strong statistical evidence that the protective effect of vaccine III is significantly better than the other vaccines

i.e., vaccine III is sufficient for significance
A few suggestions

• Look deeper into your $\chi^2$ test statistic

• Explore more stratifications of your data

• Know when to discard the 1st PC in PCA

• Make sure you get the null hypothesis right, and exploit domain knowledge properly
Hypothesis testing

• A hypothesis compares two or more groups
  – Do smokers have higher cancer rates than non-smokers?
  – Are children more vulnerable to H1N1 flu than adults?

• Statistical hypothesis testing
  – Test whether a hypothesis is supported by data using statistical methods
Conventional hypothesis generation

- **Postulate a hypothesis**
  - Is drug A more effective than drug B?

- **How?**
  - Collect data and eye ball a pattern!

<table>
<thead>
<tr>
<th>PID</th>
<th>Race</th>
<th>Sex</th>
<th>Age</th>
<th>Smoke</th>
<th>Stage</th>
<th>Drug</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Caucasian</td>
<td>M</td>
<td>45</td>
<td>Yes</td>
<td>1</td>
<td>A</td>
<td>positive</td>
</tr>
<tr>
<td>2</td>
<td>Chinese</td>
<td>M</td>
<td>40</td>
<td>No</td>
<td>2</td>
<td>A</td>
<td>positive</td>
</tr>
<tr>
<td>3</td>
<td>African</td>
<td>F</td>
<td>50</td>
<td>Yes</td>
<td>2</td>
<td>B</td>
<td>negative</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>Caucasian</td>
<td>M</td>
<td>60</td>
<td>No</td>
<td>2</td>
<td>B</td>
<td>negative</td>
</tr>
</tbody>
</table>
P-value

• Use statistical methods to decide whether a hypothesis “Is drug A more effective than drug B? ” is supported by data
  
  – E.g., $\chi^2$-test

<table>
<thead>
<tr>
<th></th>
<th>Response= positive</th>
<th>Response= Negative</th>
<th>Proportion of positive responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug=A</td>
<td>890</td>
<td>110</td>
<td>89%</td>
</tr>
<tr>
<td>Drug=B</td>
<td>830</td>
<td>170</td>
<td>83%</td>
</tr>
</tbody>
</table>

• p-value = 0.0001
  
  – Prob of observed diff betw the two drugs given assumption that the they have same effect
Limitations of conventional approach

• **Hypothesis-driven**
  – Scientist has to think of a hypothesis first
  – Allow just a few hypotheses to be tested at a time

• **So much data have been collected …**
  – No clue on what to look for
  – Know something; but do not know all
  – Impossible to inspect so much data manually

⇒ **Exploratory hypothesis testing in a data-driven manner**
Exploratory hypothesis testing

- **Data-driven hypothesis testing**
  - Have a dataset but dunno what hypotheses to test
  - Use computational methods to automatically formulate and test hypotheses from data

- **Problems to be solved:**
  - How to formulate hypotheses?
  - How to automatically generate & test hypotheses?
Formulation of a hypothesis

• “For Chinese, is drug A better than drug B?”

• Three components of a hypothesis:
  – Context (under which the hypothesis is tested)
    • Race: Chinese
  – Comparing attribute
    • Drug: A or B
  – Target attribute/target value
    • Response: positive

• \( \{\text{Race=Chinese}, \text{Drug=}A\mid B, \text{Response=}\text{positive}\} \)
Testing a hypothesis

- \( \langle \{ \text{Race}=\text{Chinese} \}, \ \text{Drug}=A\mid B, \ \text{Response}=\text{positive} \rangle \)

<table>
<thead>
<tr>
<th>context</th>
<th>Comparing attribute</th>
<th>response=positive</th>
<th>response=negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>{Race=Chinese}</td>
<td>Drug=A</td>
<td>( N^A_{pos} )</td>
<td>( N^A - N^A_{pos} )</td>
</tr>
<tr>
<td></td>
<td>Drug=B</td>
<td>( N^B_{pos} )</td>
<td>( N^B - N^B_{pos} )</td>
</tr>
</tbody>
</table>

- To test this hypothesis we need info:
  - \( N^A \) = support(\{Race=Chinese, Drug=A\})
  - \( N^A_{pos} \) = support(\{Race=Chinese, Drug=A, Res=positive\})
  - \( N^B \) = support(\{Race=Chinese, Drug=B\})
  - \( N^B_{pos} \) = support(\{Race=Chinese, Drug=B, Res=positive\})

\[ \Rightarrow \text{Frequent pattern mining} \]
Significance of observed diff

• When a single hypothesis is tested, a $p$-value of 0.05 is recognized as low enough
  – If we test 1000 hypotheses, ~50 hypotheses will pass the 0.05 threshold by random chance

• Control false positives
  – Bonferroni’s correction
    • *Family-Wise Error Rate*: Prob of making one or more false discoveries
  – Benjamini and Hochberg’s method
    • *False Discovery Rate*: Proportion of false discoveries
  – Permutation method
Need for hypothesis analysis

- Exploration is not guided by domain knowledge
  ⇒ Spurious hypotheses has to be eliminated

- Reasons behind significant hypotheses
  - Find attribute-value pairs that change the diff a lot
    - **DiffLift**: How much diff between the two groups is lifted
    - **Contribution**: Freq of attribute-value pairs

**Definition 3 (DiffLift(A=v|H)).** Let \( H = \langle P, A_{diff} = \{v_1, v_2\}, A_{target}, v_{target} \rangle \) be a hypothesis, \( A_{target} \) be categorical, \( P_1 = P \cup \{A_{diff} = v_1\} \) and \( P_2 = P \cup \{A_{diff} = v_2\} \) be the two sub-populations of \( H \), \( A = v \) be an item not in \( H \), that is, \( A \neq A_{diff}, A \neq A_{target} \) and \( A = v \notin P \). After adding item \( A = v \) to \( H \), we get two new sub-populations: \( P'_1 = P_1 \cup \{A = v\} \) and \( P'_2 = P_2 \cup \{A = v\} \). The lift of difference after adding \( A = v \) to \( H \) is defined as \( \text{DiffLift}(A=v|H) = \frac{p'_1 - p'_2}{p_1 - p_2} \), where \( p_i \) is the proportion of \( v_{target} \) in sub-population \( P_i \), and \( p'_i \) is the proportion of \( v_{target} \) in sub-population \( P'_i \), \( i = 1, 2 \).

**Definition 6 (Contribution(A = v|H)).** Let \( H \) be a hypothesis, \( A = v \) be an attribute value not in \( H \), \( P_1 \) and \( P_2 \) be the two sub-populations of \( H \), \( P'_1 \) and \( P'_2 \) be the two sub-populations after adding \( A = v \) to \( H \). The contribution of \( A = v \) to \( H \) is defined as \( \text{Contribution}(A = v|H) = \frac{\frac{n'_1}{n_1}(p'_1 - p_1) - \frac{n'_2}{n_2}(p'_2 - p_2)}{p_1 - p_2} \), where \( p_i \) is the proportion of \( v_{target} \) in sub-population \( P_i \), and \( p'_i \) is the proportion of \( v_{target} \) in sub-population \( P'_i \), \( i = 1, 2 \).
Spurious hypotheses

<table>
<thead>
<tr>
<th>Drug</th>
<th>response=positive</th>
<th>response=negative</th>
<th>proportion of positive response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug=A</td>
<td>890</td>
<td>110</td>
<td>89.0%</td>
</tr>
<tr>
<td>Drug=B</td>
<td>830</td>
<td>170</td>
<td>83.0%</td>
</tr>
<tr>
<td>Drug=A, Stage=1</td>
<td>800</td>
<td>80</td>
<td>90.9%</td>
</tr>
<tr>
<td>Drug=B, Stage=1</td>
<td>190</td>
<td>10</td>
<td>95%</td>
</tr>
<tr>
<td>Drug=A, Stage=2</td>
<td>90</td>
<td>30</td>
<td>75%</td>
</tr>
<tr>
<td>Drug=B, Stage=2</td>
<td>640</td>
<td>160</td>
<td>80%</td>
</tr>
</tbody>
</table>

• **Simpson’s Paradox**
  – “Stage” has assoc w/ both “drug” & “response”:
    • Doc’s tend to give drug A to patients at stage 1, & drug B to patients at stage 2
    • Patients at stage 1 are easier to cure than patients at stage 2
  – Attribute “stage” is called a confounding factor
Reasons behind significant hypotheses

<table>
<thead>
<tr>
<th>Product</th>
<th>Failure rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product A</td>
<td>4%</td>
</tr>
<tr>
<td>Product B</td>
<td>2%</td>
</tr>
<tr>
<td>Product A, time-of-failure=loading</td>
<td>6.0%</td>
</tr>
<tr>
<td>Product B, time-of-failure=loading</td>
<td>1.9%</td>
</tr>
<tr>
<td>Product A, time-of-failure=in-operation</td>
<td>2.1%</td>
</tr>
<tr>
<td>Product B, time-of-failure=in-operation</td>
<td>2.1%</td>
</tr>
<tr>
<td>Product A, time-of-failure=output</td>
<td>2.0%</td>
</tr>
<tr>
<td>Product B, time-of-failure=output</td>
<td>1.9%</td>
</tr>
</tbody>
</table>

- **Problem is narrowed down**
  - Product A has exceptionally higher drop rate than product B only at the loading phase
Problem statement:
Exploratory hypothesis testing

• Given
  – Dataset D, min_sup, max_pvalue, min_diff
  – $A_{\text{target}} = v_{\text{target}}$
  – $\mathcal{A}_{\text{grouping}}$: context/comparing attributes

• Find all $H = \langle P, A_{\text{diff}} = v_1 | v_2, A_{\text{target}} = v_{\text{target}} \rangle$
  – $A_{\text{diff}} \in \mathcal{A}_{\text{grouping}}$ & $\forall (A=v)$ in $P$, $A \in \mathcal{A}_{\text{grouping}}$
  – $\text{sup}(P_i) \geq \text{min}_\text{sup}$, where $P_i = P \cup \{A_{\text{diff}} = v_i\}$, $i=1, 2$
  – p-value($H$) $\leq$ max_pvalue
  – $|p_1 - p_2| \geq \text{min}_\text{diff}$, where $p_i$ is proportion of $v_{\text{target}}$
    in sub-population $P_i$, $i=1, 2$
Problem statement:
Hypothesis analysis

• Given a significant hypothesis H, generate the following info for further analysis
  – Simpson’s Paradoxes formed by H with attributes not in H
  – List of attribute-value pairs not in H ranked in descending order of DiffLift(A=v|H) and Contribution(A=v|H)
  – List of attributes not in H ranked in descending order of DiffLift(A|H) and Contribution(A|H)
Algo for exploratory hypothesis testing

• A hypothesis is a comparison betw two or more sub-populations, and each sub-population is defined by a pattern

• Step 1: Use freq pattern mining to enumerate large sub-populations and collect their statistics
  – Stored in the CFP-tree structure, which supports efficient subset/superset/exact search

• Step 2: Pair sub-populations up to form hypotheses, and then calculate their p-values
  – Use each freq pattern as a context
  – Search for immediate supersets of the context patterns, and then pair these supersets up to form hypotheses
Algo for hypothesis analysis

• **Given a hypothesis H**
  - To check whether H forms a Simpson’s Paradox with an attribute A,
    * add values of A to context of H
    * re-calculate the diff betw the two sub-populations
  - To calculate DiffLift and Contribution of an attribute-value pair A=v,
    * add A=v to context of H
    * re-calculate the diff

• **All can be done via immediate superset search**
Experiment settings

• **PC configurations**
  – 2.33Ghz CPU, 3.25GB memory, Windows XP

• **Datasets:**
  – mushroom, adult: UCI repository
  – DrugTestI, DrugTestII: study assoc betw SNPs in several genes & drug responses.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>#instances</th>
<th>#continuous attributes</th>
<th>#categorical attributes</th>
<th>$A_{\text{target}}$/$v_{\text{target}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>adult</td>
<td>48842</td>
<td>6</td>
<td>9</td>
<td>class=&gt;$50K$ (nominal)</td>
</tr>
<tr>
<td>mushroom</td>
<td>8124</td>
<td>0</td>
<td>23</td>
<td>class=poisonous (nominal)</td>
</tr>
<tr>
<td>DrugTestI</td>
<td>141</td>
<td>13</td>
<td>74</td>
<td>logAUUCT (continuous)</td>
</tr>
<tr>
<td>DrugTestII</td>
<td>138</td>
<td>13</td>
<td>74</td>
<td>logAUUCT (continuous)</td>
</tr>
</tbody>
</table>
Running time

- **Three phases**
  - Frequent pattern mining
  - Hypothesis generation
  - Hypothesis analysis

<table>
<thead>
<tr>
<th>Datasets</th>
<th>min_sup</th>
<th>min_diff</th>
<th>GenH</th>
<th>AnalyzeH</th>
<th>AvgAnalyzeT</th>
<th>#tests</th>
<th>#signH</th>
</tr>
</thead>
<tbody>
<tr>
<td>adult</td>
<td>500</td>
<td>0.05</td>
<td>0.42 s</td>
<td>6.30 s</td>
<td>0.0015 s</td>
<td>5593</td>
<td>4258</td>
</tr>
<tr>
<td>adult</td>
<td>100</td>
<td>0.05</td>
<td>2.69 s</td>
<td>37.39 s</td>
<td>0.0014 s</td>
<td>41738</td>
<td>26095</td>
</tr>
<tr>
<td>mushroom</td>
<td>500</td>
<td>0.1</td>
<td>0.67 s</td>
<td>19.00 s</td>
<td>0.0020 s</td>
<td>16400</td>
<td>9323</td>
</tr>
<tr>
<td>mushroom</td>
<td>200</td>
<td>0.1</td>
<td>5.45 s</td>
<td>123.47 s</td>
<td>0.0020 s</td>
<td>103025</td>
<td>61429</td>
</tr>
<tr>
<td>DrugTestI</td>
<td>20</td>
<td>0.5</td>
<td>0.06 s</td>
<td>0.06 s</td>
<td>0.0031 s</td>
<td>3627</td>
<td>20</td>
</tr>
<tr>
<td>DrugTestII</td>
<td>20</td>
<td>0.5</td>
<td>0.08 s</td>
<td>0.30 s</td>
<td>0.0031 s</td>
<td>4441</td>
<td>97</td>
</tr>
</tbody>
</table>

max_pvalue = 0.05
Case study: Adult dataset

- **Simpson’s paradox**

<table>
<thead>
<tr>
<th>Context</th>
<th>Comparing Groups</th>
<th>sup</th>
<th>(P_{\text{class=&gt;50K}})</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race = White</td>
<td>Occupation = Craft-repair</td>
<td>3694</td>
<td>22.84%</td>
<td>1.00 (\times) 10^{-19}</td>
</tr>
<tr>
<td></td>
<td>Occupation = Adm-clerical</td>
<td>3084</td>
<td>14.23%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sex = Male</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Race = White</td>
<td>Occupation = Craft-repair</td>
<td>3524</td>
<td>23.5%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Occupation = Adm-clerical</td>
<td>1038</td>
<td>24.2%</td>
<td></td>
</tr>
<tr>
<td>Sex = Female</td>
<td>Occupation = Craft-repair</td>
<td>107</td>
<td>8.8%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Occupation = Adm-clerical</td>
<td>2046</td>
<td>9.2%</td>
<td></td>
</tr>
</tbody>
</table>
A few suggestions

• Look deeper into your $\chi^2$ test statistic

• Explore more stratifications of your data

• **Know when to discard the 1st PC in PCA**

• Make sure you get the null hypothesis right, and exploit domain knowledge properly
Uses of PCA

• Dimension reduction
  – Summarize the data with a smaller number of variables, losing as little info as possible
  – Graphical representations of data

• Input for regression analysis
  – Highly correlated explanatory variables are problematic in regression analysis
  – One can replace them by their principal components, which are uncorrelated by definition

Credit: Marloes Maathuis
Principal component analysis

Credit: Alessandro Giuliani
PCA, a la Pearson (1901)

For example:—Let \( P_1, P_2, \ldots, P_n \) be the system of points with coordinates \( x_1, y_1; x_2, y_2; \ldots; x_n, y_n \), and perpendicular distances \( p_1, p_2, \ldots, p_n \) from a line \( \text{AB} \). Then we shall make
\[
U = S(y^2) = \text{a minimum.}
\]
If \( y \) were the dependent variable, we should have made
\[
S(y' - y)^2 = \text{a minimum}
\]

Credit: Alessandro Giuliani
PCA, in modern English 😊

Introduction
- Technique quite old: Pearson (1901) and Hotelling (1933), but still one of the most used multivariate techniques today
- Main idea:
  - Start with variables $X_1, \ldots, X_p$
  - Find a rotation of these variables, say $Y_1, \ldots, Y_p$ (called principal components), so that:
    - $Y_1, \ldots, Y_p$ are uncorrelated. Idea: they measure different dimensions of the data.
    - $\text{Var}(Y_1) \geq \text{Var}(Y_2) \geq \ldots \geq \text{Var}(Y_p)$. Idea: $Y_1$ is most important, then $Y_2$, etc.

Definition of PCA
- Given $X = (X_1, \ldots, X_p)'$
- We call $a'X$ a standard linear combination (SLC) if $\sum a_i^2 = 1$
- Find the SLC $a^{(1)} = (a_1, \ldots, a_p)$ so that $Y_1 = a^{(1)}_i X$ has maximal variance
- Find the SLC $a^{(2)} = (a_2, \ldots, a_p)$ so that $Y_2 = a^{(2)}_i X$ has maximal variance, subject to the constraint that $Y_2$ is uncorrelated to $Y_1$.
- Find the SLC $a^{(3)} = (a_3, \ldots, a_p)$ so that $Y_3 = a^{(3)}_i X$ has maximal variance, subject to the constraint that $Y_3$ is uncorrelated to $Y_1$ and $Y_2$
- Etc...
1st principal component

• How to combine the scores on 5 different exams to a total score? One could simply take the average. But it may be better to use the first principal component

• How to combine different cost factors into a cost of living index? Use first principal component

• The first principal component maximizes the variance, it spreads out the scores as much as possible

Credit: Marloes Maathuis
2nd and other principal components

• When all measurements are positively correlated, the 1st principal component is often some kind of average of the measurements
  – Size of birds
  – Severity index of psychiatric symptoms, …

• The 2nd and other principal components give important info about the remaining pattern
  – Shape of birds
  – Pattern of psychiatric symptoms, …

Credit: Marloes Maathuis

SIZE AND SHAPE VARIATION IN THE PAINTED TURTLE.¹
A PRINCIPAL COMPONENT ANALYSIS

Pierre Jolicoeur and James E. Mosimann²

Walker Museum, University of Chicago
and
Institut de Biologie, Université de Montréal

(Received for publication July 11, 1960)

Credit: Alessandro Giuliani
<table>
<thead>
<tr>
<th></th>
<th>length</th>
<th>width</th>
<th>height</th>
<th></th>
<th>length</th>
<th>width</th>
<th>height</th>
</tr>
</thead>
<tbody>
<tr>
<td>24 Males</td>
<td>93</td>
<td>74</td>
<td>37</td>
<td></td>
<td>98</td>
<td>81</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>94</td>
<td>78</td>
<td>35</td>
<td></td>
<td>103</td>
<td>84</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>96</td>
<td>80</td>
<td>35</td>
<td></td>
<td>103</td>
<td>86</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>101</td>
<td>84</td>
<td>39</td>
<td></td>
<td>105</td>
<td>86</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>102</td>
<td>85</td>
<td>38</td>
<td></td>
<td>109</td>
<td>88</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>103</td>
<td>81</td>
<td>37</td>
<td></td>
<td>123</td>
<td>92</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>104</td>
<td>83</td>
<td>39</td>
<td></td>
<td>123</td>
<td>95</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>106</td>
<td>83</td>
<td>39</td>
<td></td>
<td>133</td>
<td>99</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>107</td>
<td>82</td>
<td>38</td>
<td></td>
<td>133</td>
<td>102</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>112</td>
<td>89</td>
<td>40</td>
<td></td>
<td>133</td>
<td>102</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>113</td>
<td>88</td>
<td>40</td>
<td></td>
<td>134</td>
<td>100</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>114</td>
<td>86</td>
<td>40</td>
<td></td>
<td>136</td>
<td>102</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>116</td>
<td>90</td>
<td>43</td>
<td></td>
<td>137</td>
<td>98</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>117</td>
<td>90</td>
<td>41</td>
<td></td>
<td>138</td>
<td>99</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>117</td>
<td>91</td>
<td>41</td>
<td></td>
<td>141</td>
<td>105</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>119</td>
<td>93</td>
<td>41</td>
<td></td>
<td>147</td>
<td>108</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>120</td>
<td>89</td>
<td>40</td>
<td></td>
<td>149</td>
<td>107</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>120</td>
<td>93</td>
<td>44</td>
<td></td>
<td>153</td>
<td>107</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>121</td>
<td>95</td>
<td>42</td>
<td></td>
<td>155</td>
<td>115</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td>93</td>
<td>45</td>
<td></td>
<td>155</td>
<td>117</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>127</td>
<td>96</td>
<td>45</td>
<td></td>
<td>158</td>
<td>115</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>128</td>
<td>95</td>
<td>45</td>
<td></td>
<td>159</td>
<td>118</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>131</td>
<td>95</td>
<td>46</td>
<td></td>
<td>162</td>
<td>124</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>135</td>
<td>106</td>
<td>47</td>
<td></td>
<td>177</td>
<td>132</td>
<td>67</td>
</tr>
</tbody>
</table>

Credit: Alessandro Giuliani
Pearson Correlation Coefficients, length width height

<table>
<thead>
<tr>
<th></th>
<th>length</th>
<th>width</th>
<th>height</th>
</tr>
</thead>
<tbody>
<tr>
<td>length</td>
<td>1.00000</td>
<td>0.97831</td>
<td>0.96469</td>
</tr>
<tr>
<td>width</td>
<td>0.97831</td>
<td>1.00000</td>
<td>0.96057</td>
</tr>
<tr>
<td>height</td>
<td>0.96469</td>
<td>0.96057</td>
<td>1.00000</td>
</tr>
</tbody>
</table>

Width = 19.94 + 0.605*Length

Credit: Alessandro Giuliani
Interesting info are often in the 2\textsuperscript{nd} principal component

<table>
<thead>
<tr>
<th></th>
<th>PC1 (98%)</th>
<th>PC2 (1.4%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>0.992</td>
<td>-0.067</td>
</tr>
<tr>
<td>Width</td>
<td>0.990</td>
<td>-0.100</td>
</tr>
<tr>
<td>Height</td>
<td>0.986</td>
<td>0.168</td>
</tr>
</tbody>
</table>

PC1 = 33.78*Length + 33.73*Width + 33.57*Height

PC2 = -1.57*Length – 2.33*Width + 3.93*Height

- Presence of an overwhelming size component explaining system variance comes from the presence of a ‘typical’ common shape
- Displacement along pc = size variation (all positive terms)
- Displacement along pc2 = shape deformation (both positive and negative terms)
Female turtles are larger and have more exaggerated height 😊

<table>
<thead>
<tr>
<th>unit</th>
<th>sex</th>
<th>Length</th>
<th>Width</th>
<th>Height</th>
<th>PC1(size)</th>
<th>PC2(shape)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T25</td>
<td>F</td>
<td>98</td>
<td>81</td>
<td>38</td>
<td>-1.15774</td>
<td>0.80754832</td>
</tr>
<tr>
<td>T26</td>
<td>F</td>
<td>103</td>
<td>84</td>
<td>38</td>
<td>-0.99544</td>
<td>-0.1285916</td>
</tr>
<tr>
<td>T27</td>
<td>F</td>
<td>103</td>
<td>86</td>
<td>42</td>
<td>-0.7822</td>
<td>1.37433475</td>
</tr>
<tr>
<td>T28</td>
<td>F</td>
<td>105</td>
<td>86</td>
<td>40</td>
<td>-0.82922</td>
<td>0.28526912</td>
</tr>
<tr>
<td>T29</td>
<td>F</td>
<td>109</td>
<td>88</td>
<td>44</td>
<td>-0.55001</td>
<td>1.4815252</td>
</tr>
<tr>
<td>T30</td>
<td>F</td>
<td>123</td>
<td>92</td>
<td>50</td>
<td>0.027368</td>
<td>2.47830153</td>
</tr>
<tr>
<td>T31</td>
<td>F</td>
<td>123</td>
<td>95</td>
<td>46</td>
<td>-0.05281</td>
<td>0.05403829</td>
</tr>
<tr>
<td>T32</td>
<td>F</td>
<td>133</td>
<td>99</td>
<td>51</td>
<td>0.418589</td>
<td>0.88691967</td>
</tr>
<tr>
<td>T33</td>
<td>F</td>
<td>133</td>
<td>102</td>
<td>51</td>
<td>0.498425</td>
<td>0.33681756</td>
</tr>
<tr>
<td>T34</td>
<td>F</td>
<td>133</td>
<td>102</td>
<td>51</td>
<td>0.498425</td>
<td>0.33681756</td>
</tr>
<tr>
<td>T35</td>
<td>F</td>
<td>134</td>
<td>100</td>
<td>48</td>
<td>0.341664</td>
<td>-0.774911</td>
</tr>
<tr>
<td>T36</td>
<td>F</td>
<td>136</td>
<td>102</td>
<td>49</td>
<td>0.467898</td>
<td>-0.8299156</td>
</tr>
<tr>
<td>T37</td>
<td>F</td>
<td>137</td>
<td>98</td>
<td>51</td>
<td>0.457894</td>
<td>0.76721682</td>
</tr>
<tr>
<td>T38</td>
<td>F</td>
<td>138</td>
<td>99</td>
<td>51</td>
<td>0.501055</td>
<td>0.50628189</td>
</tr>
<tr>
<td>T39</td>
<td>F</td>
<td>141</td>
<td>105</td>
<td>53</td>
<td>0.790215</td>
<td>0.10640554</td>
</tr>
<tr>
<td>T40</td>
<td>F</td>
<td>147</td>
<td>108</td>
<td>57</td>
<td>1.129025</td>
<td>0.96505915</td>
</tr>
<tr>
<td>T41</td>
<td>F</td>
<td>149</td>
<td>107</td>
<td>55</td>
<td>1.055392</td>
<td>0.06026089</td>
</tr>
<tr>
<td>T42</td>
<td>F</td>
<td>153</td>
<td>107</td>
<td>56</td>
<td>1.161368</td>
<td>0.22155593</td>
</tr>
<tr>
<td>T43</td>
<td>F</td>
<td>155</td>
<td>115</td>
<td>63</td>
<td>1.687277</td>
<td>1.86938869</td>
</tr>
<tr>
<td>T44</td>
<td>F</td>
<td>158</td>
<td>115</td>
<td>62</td>
<td>1.696753</td>
<td>1.17117077</td>
</tr>
<tr>
<td>T45</td>
<td>F</td>
<td>159</td>
<td>118</td>
<td>63</td>
<td>1.833086</td>
<td>1.00956637</td>
</tr>
<tr>
<td>T46</td>
<td>F</td>
<td>162</td>
<td>124</td>
<td>61</td>
<td>1.962332</td>
<td>-1.261771</td>
</tr>
<tr>
<td>T47</td>
<td>F</td>
<td>177</td>
<td>132</td>
<td>67</td>
<td>2.662548</td>
<td>-1.0787317</td>
</tr>
<tr>
<td>T48</td>
<td>F</td>
<td>155</td>
<td>117</td>
<td>60</td>
<td>1.620491</td>
<td>0.09690818</td>
</tr>
<tr>
<td>T1</td>
<td>M</td>
<td>93</td>
<td>74</td>
<td>37</td>
<td>-1.46649</td>
<td>2.01289241</td>
</tr>
<tr>
<td>T2</td>
<td>M</td>
<td>94</td>
<td>78</td>
<td>35</td>
<td>-1.42356</td>
<td>0.26342486</td>
</tr>
<tr>
<td>T3</td>
<td>M</td>
<td>96</td>
<td>80</td>
<td>35</td>
<td>-1.33735</td>
<td>-0.258445</td>
</tr>
<tr>
<td>T4</td>
<td>M</td>
<td>101</td>
<td>84</td>
<td>39</td>
<td>-0.98842</td>
<td>0.4920808</td>
</tr>
<tr>
<td>T5</td>
<td>M</td>
<td>102</td>
<td>85</td>
<td>38</td>
<td>-0.98532</td>
<td>-0.2361914</td>
</tr>
<tr>
<td>T6</td>
<td>M</td>
<td>103</td>
<td>81</td>
<td>37</td>
<td>-1.11528</td>
<td>-0.436547</td>
</tr>
<tr>
<td>T7</td>
<td>M</td>
<td>104</td>
<td>83</td>
<td>39</td>
<td>-0.95555</td>
<td>0.44687352</td>
</tr>
<tr>
<td>T8</td>
<td>M</td>
<td>106</td>
<td>83</td>
<td>39</td>
<td>-0.93257</td>
<td>0.29305841</td>
</tr>
<tr>
<td>T9</td>
<td>M</td>
<td>107</td>
<td>82</td>
<td>38</td>
<td>-0.98269</td>
<td>-0.066727</td>
</tr>
<tr>
<td>T10</td>
<td>M</td>
<td>112</td>
<td>89</td>
<td>40</td>
<td>-0.63393</td>
<td>-0.8042059</td>
</tr>
<tr>
<td>T11</td>
<td>M</td>
<td>113</td>
<td>88</td>
<td>40</td>
<td>-0.64405</td>
<td>-0.6966061</td>
</tr>
<tr>
<td>T12</td>
<td>M</td>
<td>114</td>
<td>86</td>
<td>40</td>
<td>-0.68078</td>
<td>-0.4047389</td>
</tr>
<tr>
<td>T13</td>
<td>M</td>
<td>116</td>
<td>90</td>
<td>43</td>
<td>-0.42133</td>
<td>0.10845233</td>
</tr>
<tr>
<td>T14</td>
<td>M</td>
<td>117</td>
<td>90</td>
<td>41</td>
<td>-0.48485</td>
<td>-0.9039457</td>
</tr>
<tr>
<td>T15</td>
<td>M</td>
<td>117</td>
<td>91</td>
<td>41</td>
<td>-0.45824</td>
<td>-1.0882131</td>
</tr>
<tr>
<td>T16</td>
<td>M</td>
<td>119</td>
<td>93</td>
<td>41</td>
<td>-0.37202</td>
<td>-1.610083</td>
</tr>
<tr>
<td>T17</td>
<td>M</td>
<td>120</td>
<td>89</td>
<td>40</td>
<td>-0.50198</td>
<td>-1.4175463</td>
</tr>
<tr>
<td>T18</td>
<td>M</td>
<td>120</td>
<td>93</td>
<td>44</td>
<td>-0.23552</td>
<td>-0.2831547</td>
</tr>
<tr>
<td>T19</td>
<td>M</td>
<td>121</td>
<td>95</td>
<td>42</td>
<td>-0.24581</td>
<td>-1.640875</td>
</tr>
<tr>
<td>T20</td>
<td>M</td>
<td>125</td>
<td>93</td>
<td>45</td>
<td>-0.11305</td>
<td>-0.1986272</td>
</tr>
<tr>
<td>T21</td>
<td>M</td>
<td>127</td>
<td>96</td>
<td>45</td>
<td>-0.00223</td>
<td>-0.9047645</td>
</tr>
<tr>
<td>T22</td>
<td>M</td>
<td>128</td>
<td>95</td>
<td>45</td>
<td>-0.01035</td>
<td>-0.7971646</td>
</tr>
<tr>
<td>T23</td>
<td>M</td>
<td>131</td>
<td>95</td>
<td>46</td>
<td>0.079136</td>
<td>-0.559302</td>
</tr>
<tr>
<td>T24</td>
<td>M</td>
<td>135</td>
<td>106</td>
<td>47</td>
<td>0.477846</td>
<td>-2.4250481</td>
</tr>
</tbody>
</table>
Caution: PCA is not scale invariant

- Suppose we have measurements in kg and meters, and we want to have principal components expressed in grams and hectometers

- Option 1: multiply measurements in kg by 1000, multiply measurements in meters by 1/100, and then apply PCA

- Option 2: apply PCA on original measurements, and then re-scale to the appropriate units

- These two options generally give different results!

Credit: Marloes Maathuis
Caution: PCA is sensitive to outliers

- PCA is sensitive to outliers, since it is based on the sample covariance matrix $\Sigma$ which is sensitive to outliers

Credit: Marloes Maathuis
A few suggestions

• Look deeper into your $\chi^2$ test statistic

• Explore more stratifications of your data

• Know when to discard the 1$^{\text{st}}$ PC in PCA

• **Make sure you get the null hypothesis right, and exploit domain knowledge properly**
Percentage of overlapping genes

- Low % of overlapping genes from diff expt in general
  - Prostate cancer
    - Lapointe et al, 2004
    - Singh et al, 2002
  - Lung cancer
    - Garber et al, 2001
    - Bhattacharjee et al, 2001
  - DMD
    - Haslett et al, 2002
    - Pescatori et al, 2007

<table>
<thead>
<tr>
<th>Datasets</th>
<th>DEG</th>
<th>POG</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prostate Cancer</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>Top 100</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td><strong>Lung Cancer</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Top 100</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td><strong>DMD</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>Top 100</td>
<td>0.54</td>
<td></td>
</tr>
</tbody>
</table>

Gene regulatory circuits

- Each disease phenotype has some underlying cause
- There is some unifying biological theme for genes that are truly associated with a disease subtype
- Uncertainty in selected genes can be reduced by considering biological processes of the genes
- The unifying biological theme is basis for inferring the underlying cause of disease subtype
Overlap analysis: ORA

ORA tests whether a pathway is significant by intersecting the genes in the pathway with a pre-determined list of DE genes (we use all genes whose t-statistic meets the 5% significance threshold), and checking the significance of the size of the intersection using the hypergeometric test.

Disappointing performance

upregulated in DMD

DMD gene expression data
- Pescatori et al., 2007
- Haslett et al., 2002

Pathway data
- PathwayAPI, Soh et al., 2010
Issue #1 with ORA

- Its null hypothesis basically says “Genes in the given pathway behaves no differently from randomly chosen gene sets of the same size”

- This may lead to lots of false positives

- A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in a cell. Thus necessarily the behaviour of genes in a pathway is more coordinated than random ones
Issue #2 with ORA

• It relies on a pre-determined list of DE genes

• This list is sensitive to the test statistic used and to the significance threshold used

• This list is unstable regardless of the threshold used when sample size is small
Issue #3 with ORA

- It tests whether the entire pathway is significantly differentially expressed

- If only a branch of the pathway is relevant to the phenotypes, the noise from the large irrelevant part of the pathways can dilute the signal from that branch
GSEA

- **Issue #2 is mostly solved**
  - Does not need pre-determined list of DE genes
  - But gene ranking (based on t-test p-value) is still unstable when sample size is small


Note: Class label permutation mode cannot be used when sample size is small
Better performance

upregulated in DMD

subnetwork agreement

sample size (N)

GSEA

ORA
ORA-Paired: Paired test and new null hypothesis

- Let \( g_i \) be genes in a given pathway \( P \)
- Let \( p_j \) be patients
- Let \( q_k \) be normals

- Let \( \Delta_{i,j,k} = \text{Expr}(g_i, p_j) - \text{Expr}(g_i, q_k) \)

- Test whether \( \Delta_{i,j,k} \) is a distribution with mean 0

**Issue #1 is solved**
- The null hypothesis is now “If a pathway \( P \) is irrelevant to the difference between patients and normals, then the genes in \( P \) are expected to behave similarly in patients and normals”

**Issue #2 is solved**
- No longer need a pre-determined list of DE genes
- Sample size is now much larger
  - \( \# \text{patients} + \# \text{normals} \)
  - \( \# \text{patients} \times \# \text{normals} \times \# \text{genes in } P \)
Much better performance

**upregulated in DMD**

![Graph showing subnetwork agreement vs. sample size (N)](image)

- **ORA-Paired**
- **PFSNet**
- **GSEA**
- **ORA**
NEA-Paired: Paired test on subnetworks

- Given a pathway $P$
- Let each node and its immediate neighbourhood in $P$ be a subnetwork
- Apply ORA-Paired on each subnetwork individually

- Issues #1 & #2 are solved as per ORA-Paired

- Issue #3 is partly solved
  - Testing subnetworks instead of whole pathways
  - But subnetworks derived in fragmented way
Even better performance

upregulated in DMD

subnetwork agreement

sample size (N)

NEA-Paired
ORA-Paired
PFSNet
GSEA
ORA
ESSNet: Larger subnetworks

- Compute the average rank of a gene based on its expression level in patients
- Use the top $\alpha\%$ to extract large connected components in pathways
- Test each component using ORA-Paired

- Gene rank is very stable
- Issues #1 - #3 solved
Fantastic performance

upregulated in DMD

- ESSNet
- NEA-Paired
- ORA-Paired
- PFSNet
- GSEA
- ORA

subnetwork agreement

sample size (N)
For the Leukemia dataset (in which patients are either classified to have acute lymphoblastic leukemia or acute myeloid leukemia), one of the significant subnetworks that is biologically relevant is part of the Interleukin-4 signaling pathway, see figure 6b (supplementary material). The binding of Interleukin-4 to its receptor (Cardoso et al., 2008) causes a cascade of protein activation involving JAK1 and STAT6 phosphorylation. STAT6 dimerizes upon activation and is transported to the nucleus and interacts with the RELA/NFkB1 transcription factors, known to promote the proliferation of T-cells (Rayet and Gelinas, 1999). In contrast, acute myeloid leukemia does not have genes in this subnetwork up-regulated and are known to be unrelated to lymphocytes.
What have we learned?

• Mechanical application of statistical and data mining techniques often does not work

• Must understand statistical and data mining tools

• Must understand the problem domain

• Must know how to logically exploit both