**WHEN?**

Most of scientific or industrial data $\rightarrow$ Multivariate data (huge size of data)
1) All data are useful?
2) If not, how do we **quickly** extract **useful** information only?
**WHY?**

When we use traditional techniques,

1. Not easy to extract useful information from the multivariate data
   1) Many bivariate plots are needed
   2) Bivariate plots, however, **mainly** represent correlations between variables (not samples).
   3) Not easy to visualize multivariate data
      - 1D: dot
      - 2D: Bivariate plot (i.e. X-Y plane)
      - 3D: X-Y-Z plot or ternary plot
      - 4D: X-Y-Z plot or ternary plot with a color code / Tetrahedron
      - 5D, 6D, etc. : ????

2. Principal Component Analysis (PCA) is useful when we need to **extract useful information** from multivariate data sets. This technique is based on the reduced dimensionality. Therefore, trends in multivariate data are easily visualized.
With the aid of Rapid Scientific Development

Combinatorial Materials Design & Synthesis

Various measuring techniques

Generations of lots of samples (observations)

Generations of many measured properties (variables)

Multivariate Data Table
General Concept of Principal Component Analysis

Multivariate Data Table

Multivariate Data Analysis (ex. Principal Component Analysis)

Information

+ Noise

Knowledge
Example: Elemental Properties

(For more detail explanations of PCA, use other sets of slides)
## PCA Example of Elemental Properties

### Multivariate Data Table

- **Columns** are 17 measured properties (i.e. multivariate data set)
- **Rows** are 58 elements

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Atomic Number</th>
<th>Atomic Weight</th>
<th>Valence electron number, $N_v$ (mod)</th>
<th>the Marynov-Bashirov electronegativity $X$ [(eV)$^{1/2}$]</th>
<th>Pseudopotential core radius $R_{C_p}$</th>
<th>Covalent Radius (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Na</td>
<td>11</td>
<td>22.989763</td>
<td>1</td>
<td>0.89</td>
<td>2.65</td>
<td>1.54</td>
</tr>
<tr>
<td>Mg</td>
<td>12</td>
<td>24.305</td>
<td>2</td>
<td>1.31</td>
<td>2.03</td>
<td>1.36</td>
</tr>
<tr>
<td>Al</td>
<td>13</td>
<td>26.981539</td>
<td>3</td>
<td>1.64</td>
<td>1.675</td>
<td>1.25</td>
</tr>
<tr>
<td>K</td>
<td>19</td>
<td>39.0983</td>
<td>1</td>
<td>0.8</td>
<td>3.69</td>
<td>2.03</td>
</tr>
<tr>
<td>α-Ca</td>
<td>20</td>
<td>40.078</td>
<td>2</td>
<td>1.17</td>
<td>3</td>
<td>1.74</td>
</tr>
<tr>
<td>Sc</td>
<td>21</td>
<td>44.95591</td>
<td>3</td>
<td>1.5</td>
<td>2.75</td>
<td>1.44</td>
</tr>
<tr>
<td>Ti</td>
<td>22</td>
<td>47.869</td>
<td>4</td>
<td>1.86</td>
<td>2.58</td>
<td>1.32</td>
</tr>
<tr>
<td>V</td>
<td>23</td>
<td>50.9415</td>
<td>5</td>
<td>2.22</td>
<td>2.43</td>
<td>1.22</td>
</tr>
<tr>
<td>Cr</td>
<td>24</td>
<td>51.9961</td>
<td>6</td>
<td>2</td>
<td>2.44</td>
<td>1.18</td>
</tr>
<tr>
<td>Mn</td>
<td>25</td>
<td>54.93805</td>
<td>7</td>
<td>2.04</td>
<td>2.22</td>
<td>1.17</td>
</tr>
<tr>
<td>Fe</td>
<td>26</td>
<td>55.847</td>
<td>6</td>
<td>1.67</td>
<td>2.11</td>
<td>1.16</td>
</tr>
<tr>
<td>Co</td>
<td>27</td>
<td>58.9332</td>
<td>3</td>
<td>1.72</td>
<td>2.02</td>
<td>1.16</td>
</tr>
<tr>
<td>Ni</td>
<td>28</td>
<td>58.6934</td>
<td>3</td>
<td>1.76</td>
<td>2.18</td>
<td>1.15</td>
</tr>
<tr>
<td>Cu</td>
<td>29</td>
<td>63.546</td>
<td>2</td>
<td>1.08</td>
<td>2.04</td>
<td>1.17</td>
</tr>
<tr>
<td>Zn</td>
<td>30</td>
<td>65.38</td>
<td>2</td>
<td>1.44</td>
<td>1.88</td>
<td>1.25</td>
</tr>
<tr>
<td>Ga</td>
<td>31</td>
<td>69.723</td>
<td>3</td>
<td>1.7</td>
<td>1.696</td>
<td>1.25</td>
</tr>
<tr>
<td>Rb</td>
<td>37</td>
<td>85.4678</td>
<td>1</td>
<td>0.8</td>
<td>4.1</td>
<td>2.16</td>
</tr>
<tr>
<td>Sr</td>
<td>38</td>
<td>87.62</td>
<td>2</td>
<td>1.13</td>
<td>3.31</td>
<td>1.99</td>
</tr>
</tbody>
</table>
How do we extract useful information from multivariate elemental properties?

1. Traditional (Classical) way: A lot of bivariate plots

2. New way: Principal Component Analysis by reducing the dimensionality of a data set while retaining most of original information in the data
Traditional way of handling multivariate data set
- Bivariate Plots of Elemental Properties

Some relationships can be found from the bivariate plots of raw data. Ex. Wiedemann-Franz Law

To identify the relationships for many properties, therefore, how many pair-wise plots do we need? We can use Principal Component Analysis to determine these relationships for many elements and properties simultaneously.
PCA of **Elemental Properties**

**Informative way of handling multivariate data set**

Columns are measured properties
- Order of properties does not matter for PCA calculations
- Certain organizations may be easier to understand

Rows are elements
- Order technically doesn’t matter, but atomic number is obvious metric
- Atomic number not used in calculations

Properties/elements chosen are limited by available data
Lack of property information for an element requires either property or element to be excluded from PCA
PCA Results: Scoring Plots - 1

1. Shows relationships between elements (samples)
2. Axes (PC-principal component) formed from linear combination of properties

Columns 11 and 12 in the periodic table are plotted off to one side, out of the usual order - indicating change in properties compared to rest of metals

17 variable problems are shown in 3D via PCA
PCA Results: Scoring Plots - 2

1. Previous scoring plot can be rotated by any axes to investigate any interesting patterns (see below).
2. Elements line in rows similar to standard periodic table

Each cluster: period (row) in the periodic table (see next slide)
### Periodic Table

![Periodic Table Image](http://www.efunda.com/materials/elements/periodic_table.cfm)

Note: X or XX = Solid, X or XX = Liquid, and X or XX = Gas around room temperature (25 °C; 78 °F) and 1 atmosphere.

If the axes are rotated to see other patterns, Elements line in columns similar to standard periodic table.

Each blue line: group (column) in the periodic table.
PCA Results: Loading Plots

1. Location of property in loading plot indicates influence of property on PC
   - Atomic weight has no influence on PC1 (i.e., ~0), high influence on PC2 and 3
2. Clustered properties indicate relationships
   - Melting and boiling points, heats of fusion and vaporization, and valence number
   - Pauling’s electronegativity and first ionization potential
   - Electrical and thermal conductivity - Wiedemann Franz Law
   - Molar volume and atomic radius
Conclusion:

Principal Component Analysis (PCA) is a technique to reduce the information dimensionality of a dataset consisting of a large number of interrelated variables obtained from a combinatorial experiment or from a well organized database by projection methods, in a way that minimizes the loss of information.


Then, we can easily (quickly) identify all the existing correlations between samples and variables in huge multivariate data.