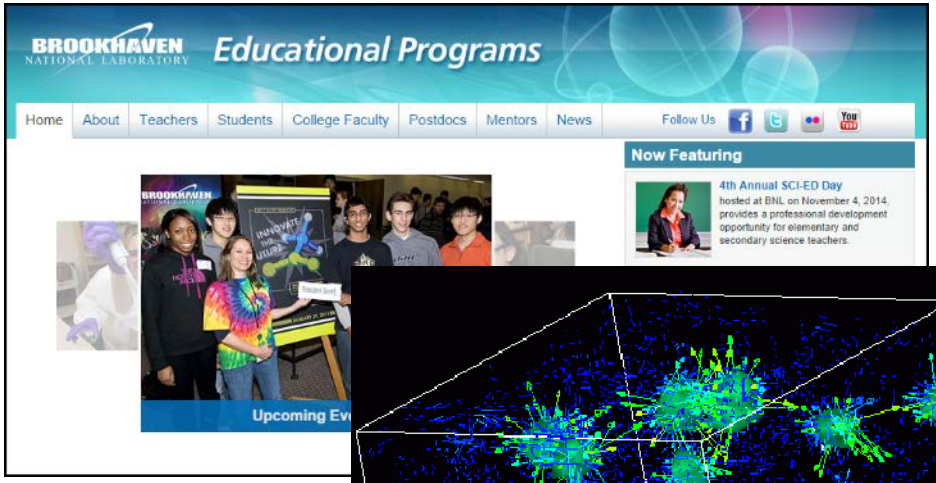
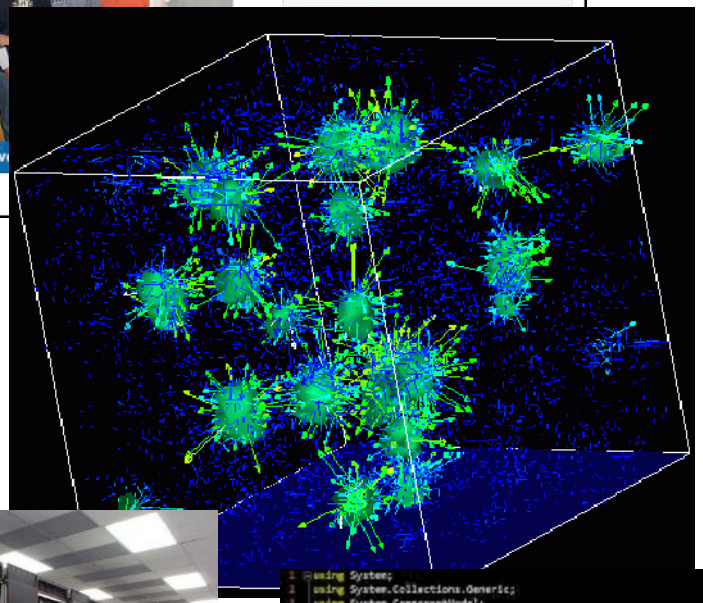
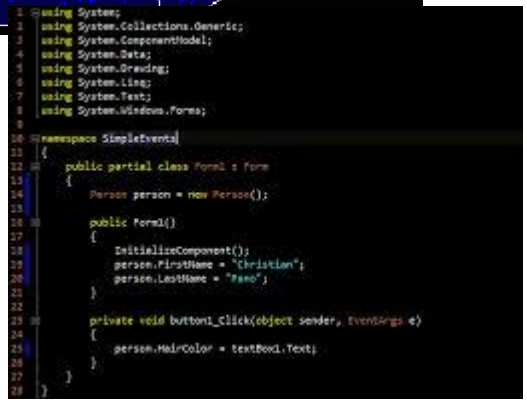






Survey of Scientific Computing (SciComp 301)

Dave Biersach
Brookhaven National Laboratory
dbiersach@bnl.gov

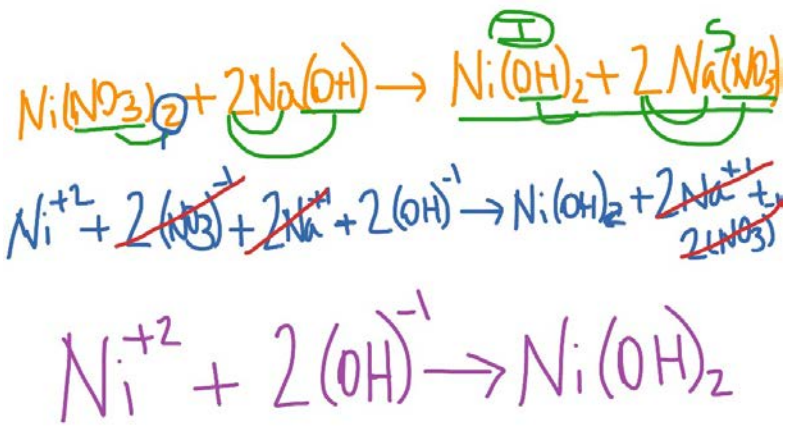
Session 17
Computational Chemistry, Clustering

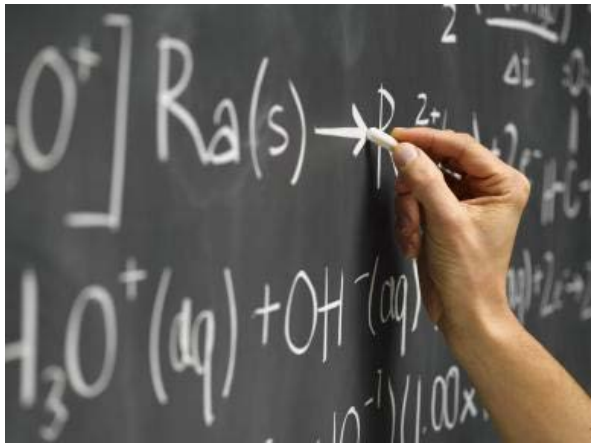
Session Goals

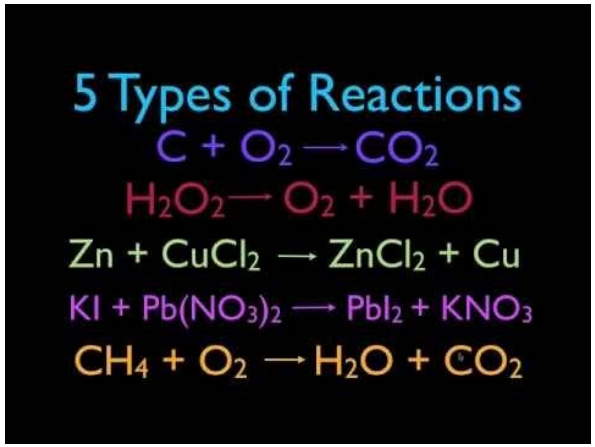
- Balance **ionic chemical equations** using linear programming
- Correlate "principle of atom conservation" (**POAC**) to the *minimization* of an objective function
- Represent molar & charge ratios as **linear constraints**
- Encode chemical equations as a matrix within a text file
- Implement **brute-force** searching using a **stack** data structure *instead* of using recursion or nested **for()** loops

Balancing Ionic Equations

- Can we write a program balance ionic equations?







Balancing Ionic Equations

Step 1: Write reactants and products

$$C_3H_8 + O_2 \rightarrow CO_2 + H_2O$$

Step 2: Find one atom that occurs only in one substance on both sides

$$C_3H_8 + O_2 \rightarrow CO_2 + H_2O$$

Step 3: Find coeffi

Step ther unbalanced atom which occu only one substance

$$1C_3H_8 + O_2 \rightarrow 3CO_2 + H_2O$$

Step 5: Find coefficient to balance this atom

$$1C_3H_8 + O_2 \rightarrow$$

Step 6: R

$$\rightarrow 3CO_2 + 4H_2O$$

$$1C_3H_8 + 5O_2 \rightarrow 3CO_2 + 4H_2O$$

Step 7: Make sure the coefficients are reduced to the smallest whole number values

$$1C_3H_8 + 5O_2 \rightarrow 3CO_2 + 4H_2O$$

Step 8: Check the balanced equation

$$1C_3H_8 + 5O_2 \rightarrow 3CO_2 + 4H_2O$$

3C, 8H, 10O 3C, 10O, 8H

Yikes - and this is not even ionic!

Chemistry is Optimization

- Consider a chemical formula as an **optimization** problem:

$$\mathrm{MnO_4^-} + \mathrm{H_2O_2} + \mathrm{H^+} \rightarrow \mathrm{Mn^{2+}} + \mathrm{O_2} + \mathrm{H_2O}$$

$$\boldsymbol{x_0}\mathrm{MnO_4^-} + \boldsymbol{x_1}\mathrm{H_2O_2} + \boldsymbol{x_2}\mathrm{H^+} \rightarrow \boldsymbol{x_3}\mathrm{Mn^{2+}} + \boldsymbol{x_4}\mathrm{O_2} + \boldsymbol{x_5}\mathrm{H_2O}$$

- Develop the **objective function** and the **constraints**:

$$obj\text{:}\ \ minimize \sum_{i=0}^{n} x_i$$

due to Principle of Atom Conservation (**POAC**)

Chemistry is Optimization

$$\boldsymbol{x_0}\mathrm{MnO_4^-} + \boldsymbol{x_1}\mathrm{H_2O_2} + \boldsymbol{x_2}\mathrm{H^+} \rightarrow \boldsymbol{x_3}\mathrm{Mn^{2+}} + \boldsymbol{x_4}\mathrm{O_2} + \boldsymbol{x_5}\mathrm{H_2O}$$

$$obj:\ \ minimize \sum_{i=0}^{n} x_i$$

$x_0 = x_3$	To balance the Manganese (Mn)
$4x_0 + 2x_1 = 2x_4 + x_5$	To balance the Oxygen (O)
$2x_1 + x_2 = 2x_5$	To balance the Hydrogen (H)
$-x_0 + x_2 = 2x_3$	To balance the charges

Chemistry is Optimization

- We have transformed solving a chemical equation into a solving a **linear optimization** problem
  - The **Simplex** method is a known good approach to solve linear programming problems in reasonable time
  - However managing **slack variables** and the **tableau** algorithm can be intimidating for novice programmers
- For small problems, we can just enumerate the ***search space*** (defined by all *permutations* of coefficients subject to the constraints) in order to find all of the possible solutions!

Encoding an Chemical Equation

- We can encode a chemical equation using a **2D matrix**
  - We need a row for every unique element (**atom**)
  - We (**may**) need another row to encode the **ionic charges**
  - The **last row** will always contain the **maximum values** we want to consider for each unknown coefficient
- We need a column for every term in the equation
  - There will be a column for each <u>molecule</u> or charged ion
  - Every column is another **unknown** we need to resolve
- The matrix elements will be stored in a text (.txt) file
  - The first line will describe the **# of cols** and **# of rows**

Encoding Chemical Equation 1

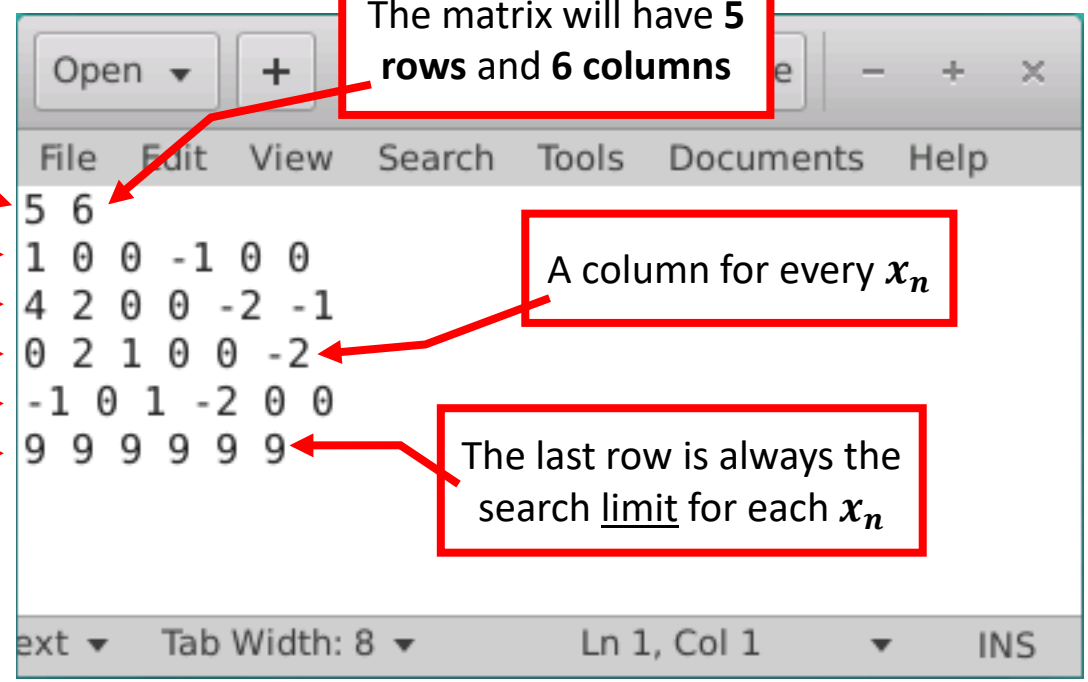
$$MnO_4^- + H_2O_2 + H^+ \rightarrow Mn^{2+} + O_2 + H_2O$$

$$\boldsymbol{x_0} MnO_4^- + \boldsymbol{x_1} H_2O_2 + \boldsymbol{x_2} H^+ \rightarrow \boldsymbol{x_3} Mn^{2+} + \boldsymbol{x_4} O_2 + \boldsymbol{x_5} H_2O$$

One row for every atom and a row for charges ***if needed***

The matrix will have **5 rows** and **6 columns**

Manganese (Mn)
Oxygen (O)
Hydrogen (H)
Charges (+/-)
Max Value



A column for every $\boldsymbol{x_n}$

The last row is always the search <u>limit</u> for each $\boldsymbol{x_n}$

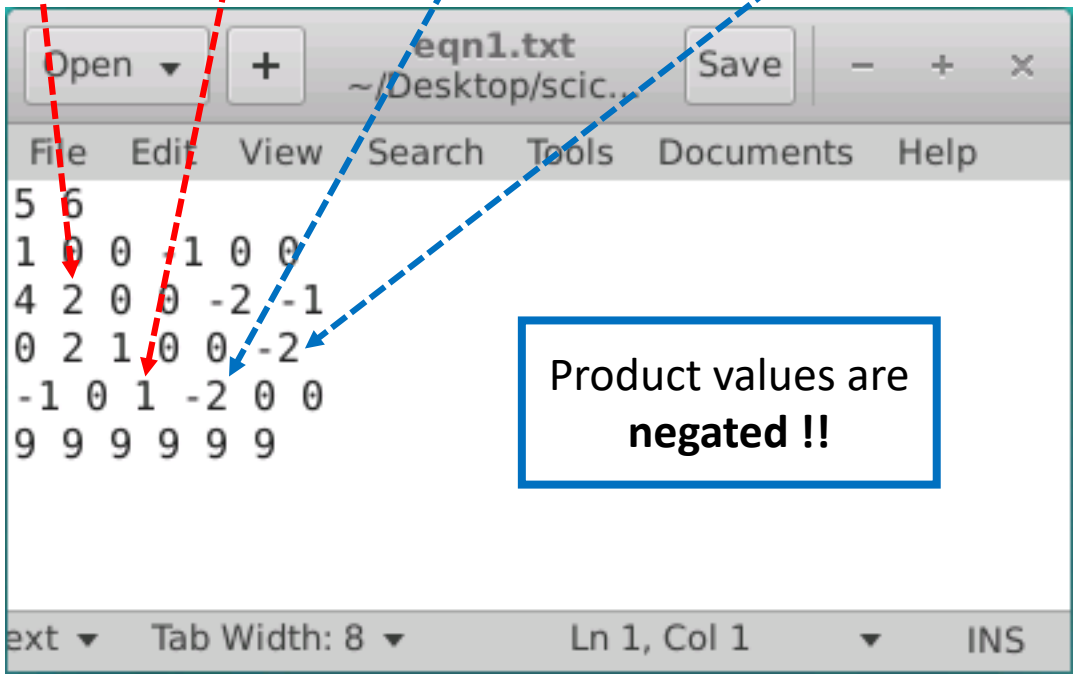
Encoding Chemical Equation 1

$$\mathrm{MnO_4^-} + \mathrm{H_2O_2} + \mathrm{H^+} \rightarrow \mathrm{Mn^{2+}} + \mathrm{O_2} + \mathrm{H_2O}$$

$$\boldsymbol{x_0}\mathrm{MnO_4^-} + \boldsymbol{x_1}\mathrm{H_2O_2} + \boldsymbol{x_2}\mathrm{H^+} \rightarrow \boldsymbol{x_3}\mathrm{Mn^{2+}} + \boldsymbol{x_4}\mathrm{O_2} + \boldsymbol{x_5}\mathrm{H_2O}$$

Reactant values are copied verbatim

Manganese (Mn)
Oxygen (O)
Hydrogen (H)
Charges (+/-)
Max Value



Product values are **negated !!**

Encoding Chemical Equation 1

- A potential **valid** set of unknown values will yield a **sum of zero** when <u>convolved</u> with each row (constraint) in the matrix

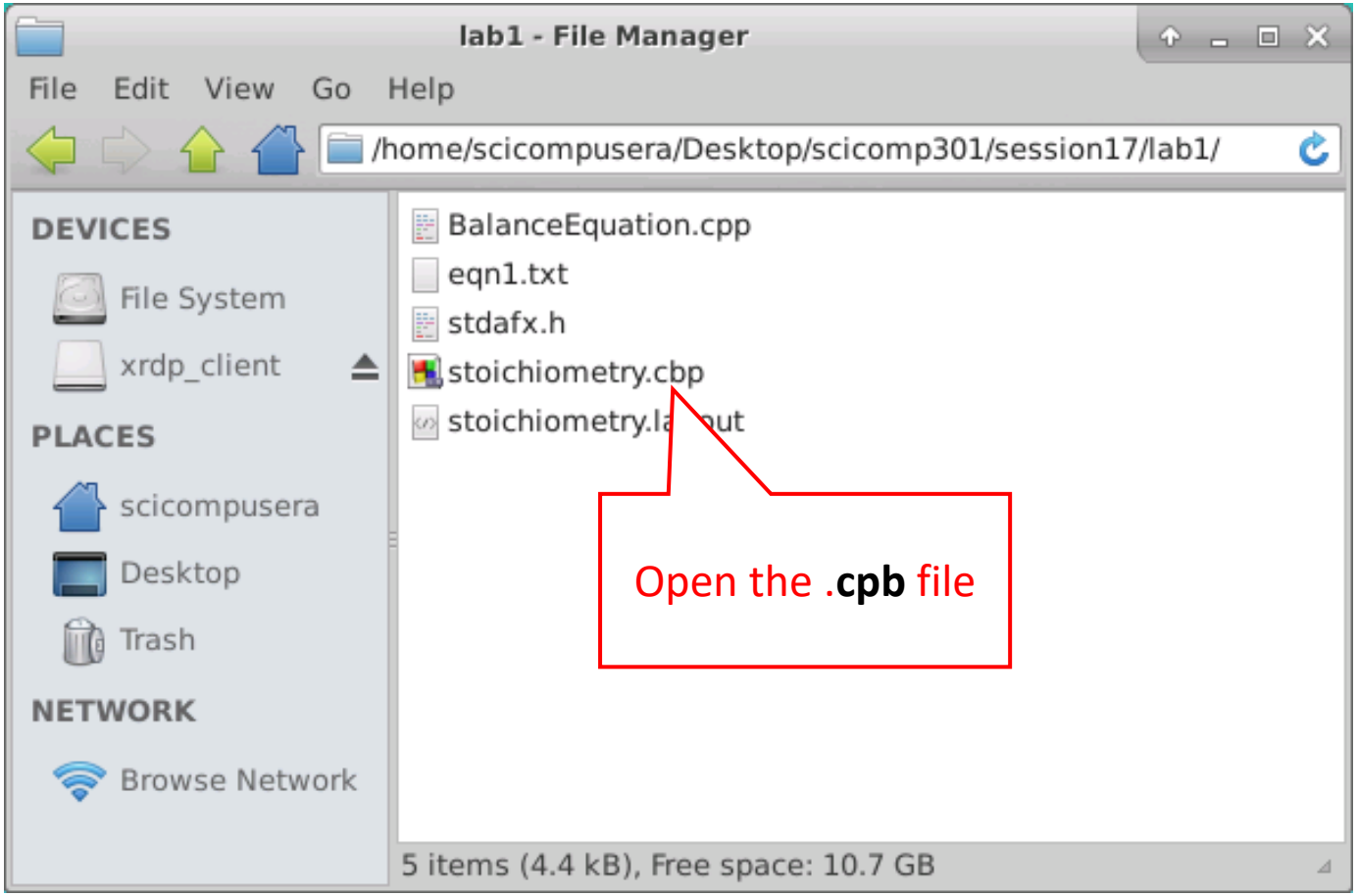
$$\boldsymbol{x_0}\mathrm{MnO_4^-} + \boldsymbol{x_1}\mathrm{H_2O_2} + \boldsymbol{x_2}\mathrm{H^+} \rightarrow \boldsymbol{x_3}\mathrm{Mn^{2+}} + \boldsymbol{x_4}\mathrm{O_2} + \boldsymbol{x_5}\mathrm{H_2O}$$

For the Oxygen atom these coefficients are taken from 3rd row in matrix

$$(\mathbf{4})\boldsymbol{x_0} + (\mathbf{2})\boldsymbol{x_1} + (\mathbf{0})\boldsymbol{x_2} + (\mathbf{0})\boldsymbol{x_3} + (\mathbf{-2})\boldsymbol{x_4} + (\mathbf{-1})\boldsymbol{x_5} = \boxed{\mathbf{0}}$$

- The current search values for each $\boldsymbol{x_n}$ are multiplied by each **column** coefficient in each **row** in the equation matrix – if the sum for **every row = 0**, then we have a *potential* solution!
- The final solution is where $\sum x_n$ has the <u>lowest</u> value (**POAC**)

Open Lab 1 – Stoichiometry



Build Lab 1 – Stoichiometry

BalanceEquation.cpp [stoichiometry] - Code::Blocks 16.01

File Edit View Search Project Build Debug Tools Plugins Settings Help

Debug

Management

Projects Symbols

- Workspace
  - **stoichiometry**
    - Sources
    - Headers
    - Others

BalanceEquation.

Build the solution

```
47                  size_t i{};i < eqn->front().size();++i)
48                                  1,eqn->back().at(i) + 1 });
49
50
51
52
53      for (size_t j{};j < eqn->size() - 1;++j) {
54          int sum = 0;
55          for (size_t i{};i < stack->size();++i)
56              sum += eqn->at(j).at(i) * stack->at(i).at(0);
57          if (sum != 0)
58              return false;
59      }
60      return true;
61  }
62
63  bool FindSolution()
64  {
65      int stackLevel = 0;
66      while (stackLevel >= 0) {
67          while (stack->at(stackLevel).at(0)
68              < stack->at(stackLevel).at(1))
69          {
70              if (stackLevel == (int)stack->size() - 1) break;
```

Project Folder & File Structure

```
scicompuser#/~/
    Desktop/scicomp301/
        session17/lab1/
            eqn1.txt
            eqn2.txt
            bin/
                Debug/
                    stoichiometry
                Release/
                    stoichiometry
```

In Linux only **folders** end with a forward slash **/**

A **parent** folder can be specified using the **..** wildcard

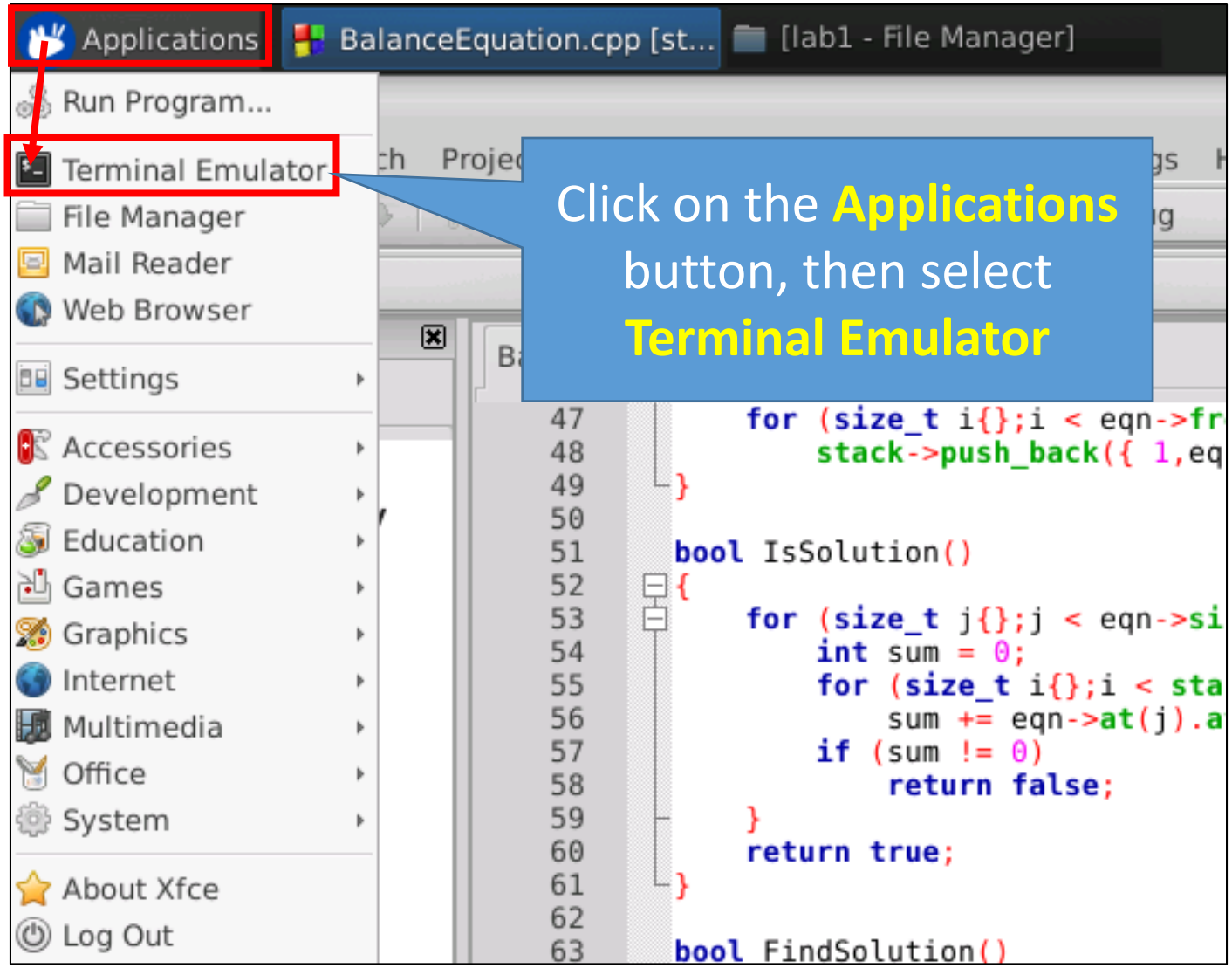
The program **stoichiometry** is in the Debug/ folder

The **current** folder can be specified using the **.** wildcard

app name *data file path*

```
./stoichiometry ../../eqn1.txt
```

Opening A Linux Terminal



Run Lab 1 – Stoichiometry

Terminal - scicompusera@ip-172-31-71-74: ~/Desktop/scicomp301/session17/lab1/bin/Debug

File Edit View Terminal Tabs Help

```
scicompusera@ip-172-31-71-74:~$ cd Desktop/scicomp301/session17/lab1/bin/Debug
scicompusera@ip-172-31-71-74:~/Desktop/scicomp301/session17/lab1/bin/Debug$ ./st
oichiometry ../../eqn1.txt
x0 = 2
x1 = 1
x2 = 6
x3 = 2
x4 = 3
x5 = 4
scicompusera@ip-172-31-71-74:~/Desktop/scicomp301/session17/lab1/bin/Debug$
```

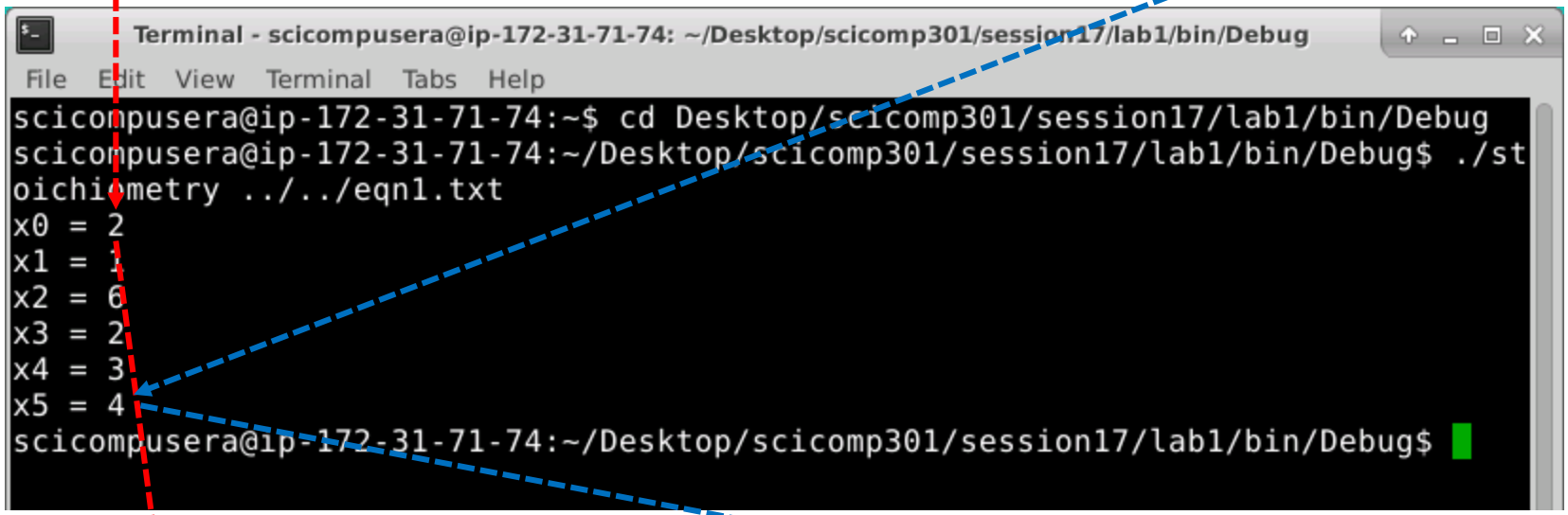
```
cd Desktop/scicomp301/session17/lab1/bin/Debug
./stoichiometry ../../eqn1.txt
```

Press **ENTER** after each line and **TAB** to auto-complete names

Remember Linux is case sensitive!

Verify Chemical Equation 1

$$x_0 MnO_4^- + x_1 H_2O_2 + x_2 H^+ \rightarrow x_3 Mn^{2+} + x_4 O_2 + x_5 H_2O$$



$$2MnO_4^- + 1H_2O_2 + 6H^+ \rightarrow 2Mn^{2+} + 3O_2 + 4H_2O$$

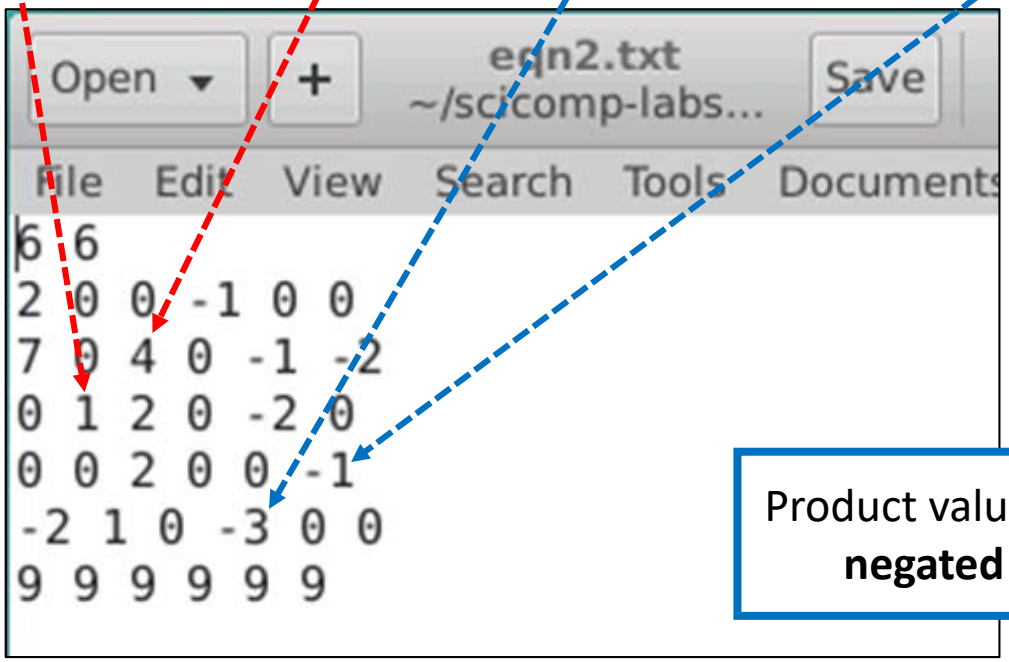
Encoding Chemical Equation 2

$$Cr_2O_7^{2-} + H^+ + H_2C_2O_4 \rightarrow Cr^{3+} + H_2O + CO_2$$

$$\boldsymbol{x_0}Cr_2O_7^{2-} + \boldsymbol{x_1}H^+ + \boldsymbol{x_2}H_2C_2O_4 \rightarrow \boldsymbol{x_3}Cr^{3+} + \boldsymbol{x_4}H_2O + \boldsymbol{x_5}CO_2$$

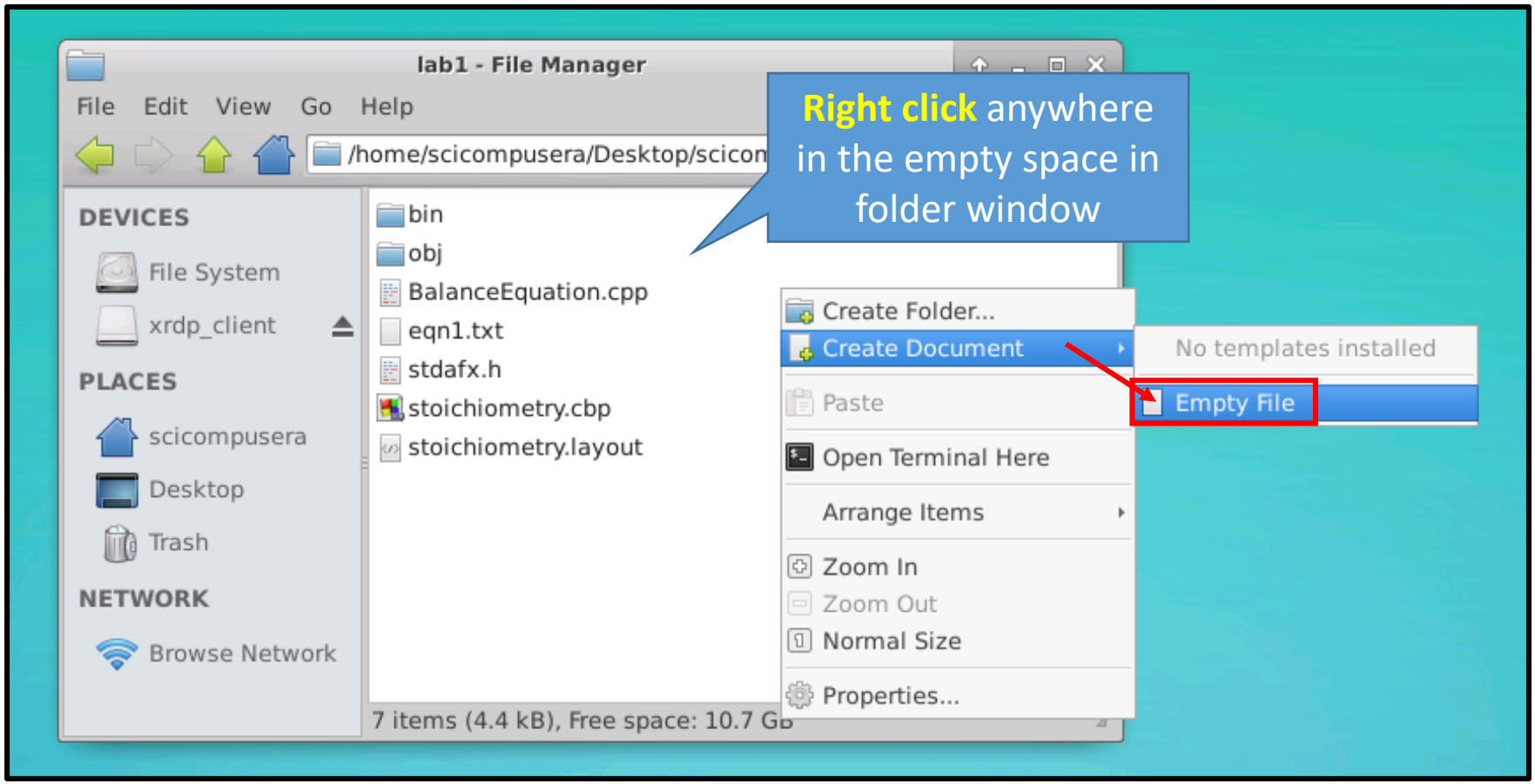
Reactant values are copied verbatim

Chromium (Cr)
Oxygen (O)
Hydrogen (H)
Carbon (C)
Charges (+/-)
Max Value

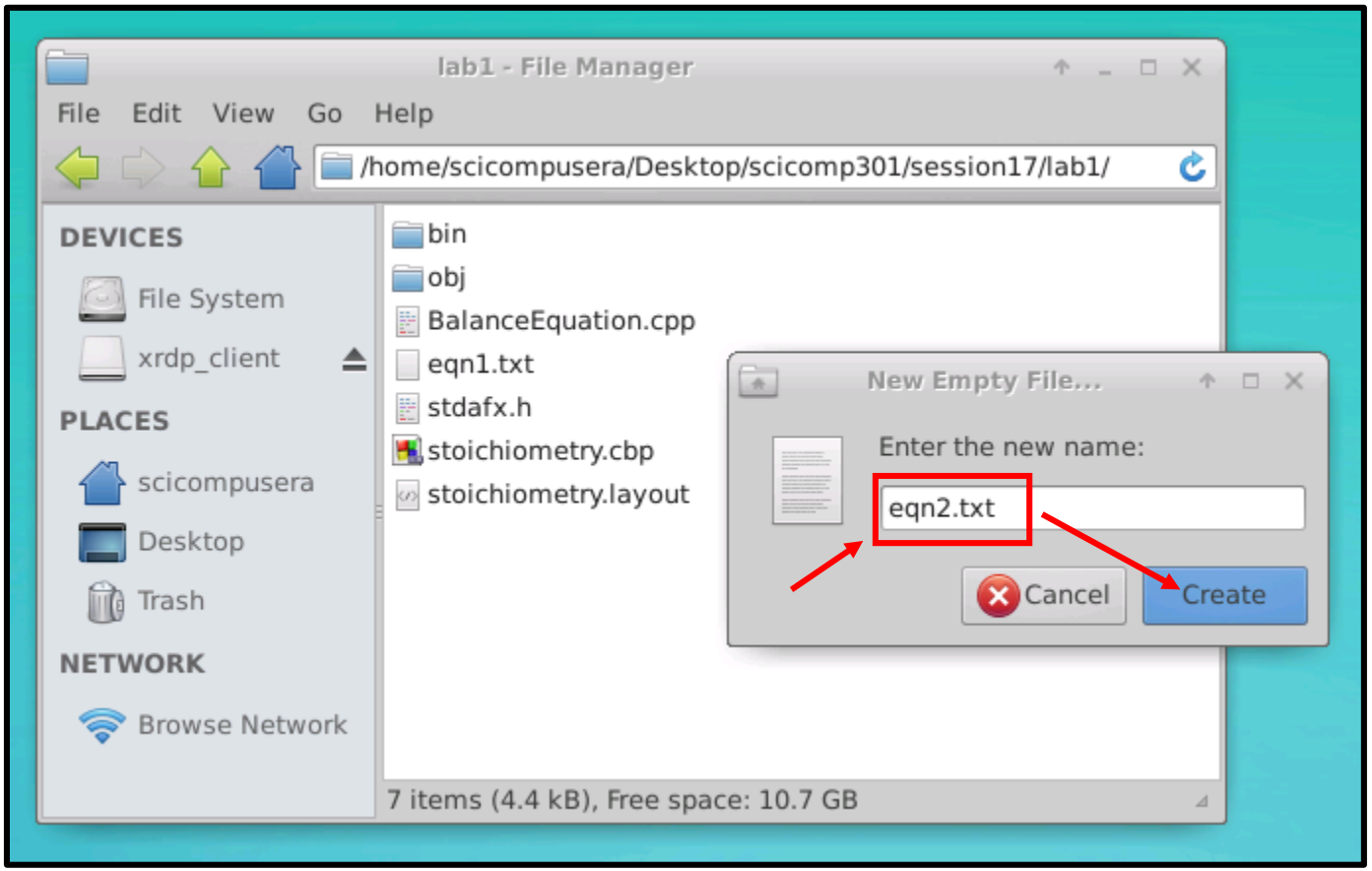


Product values are **negated !!**

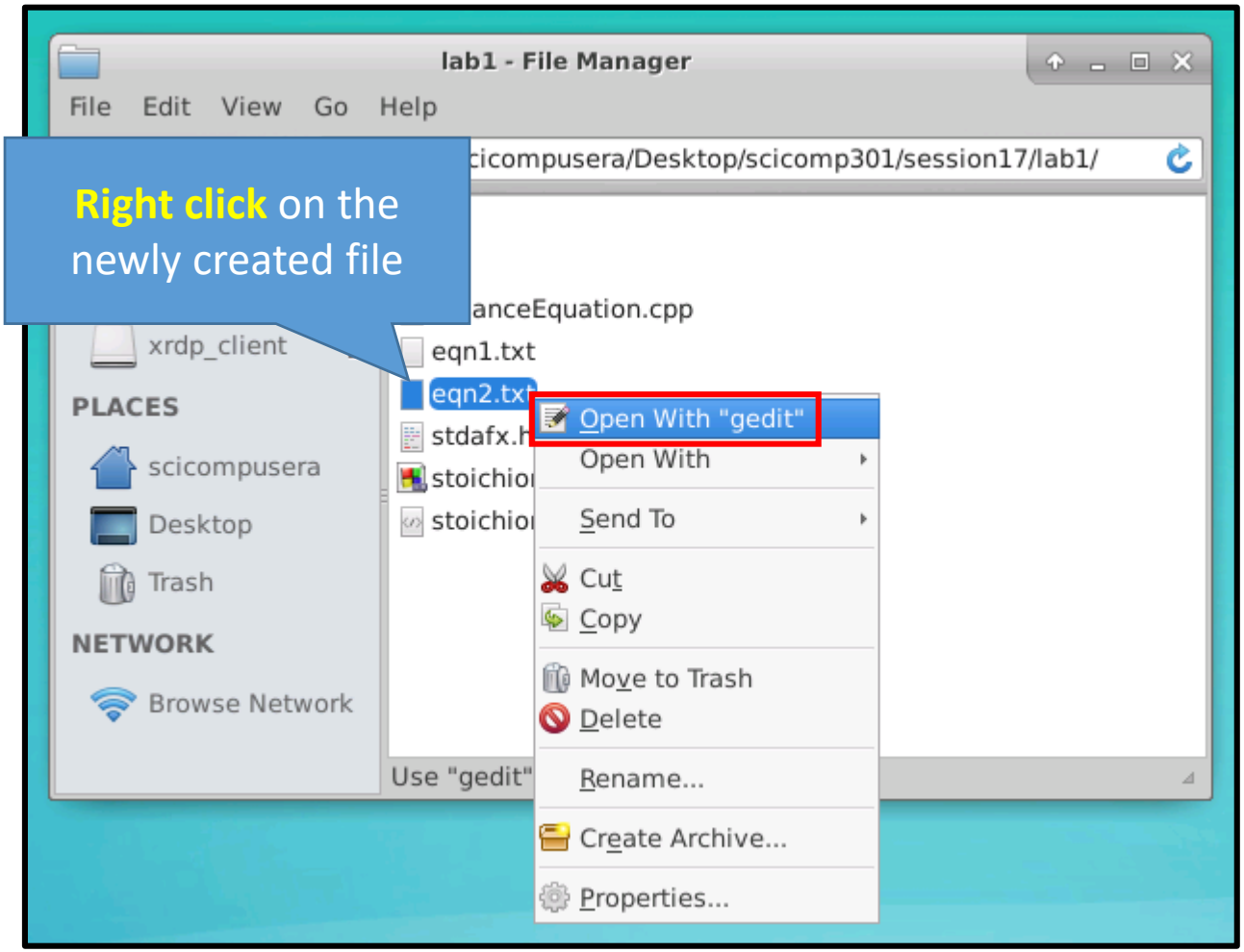
Encoding Chemical Equation 2



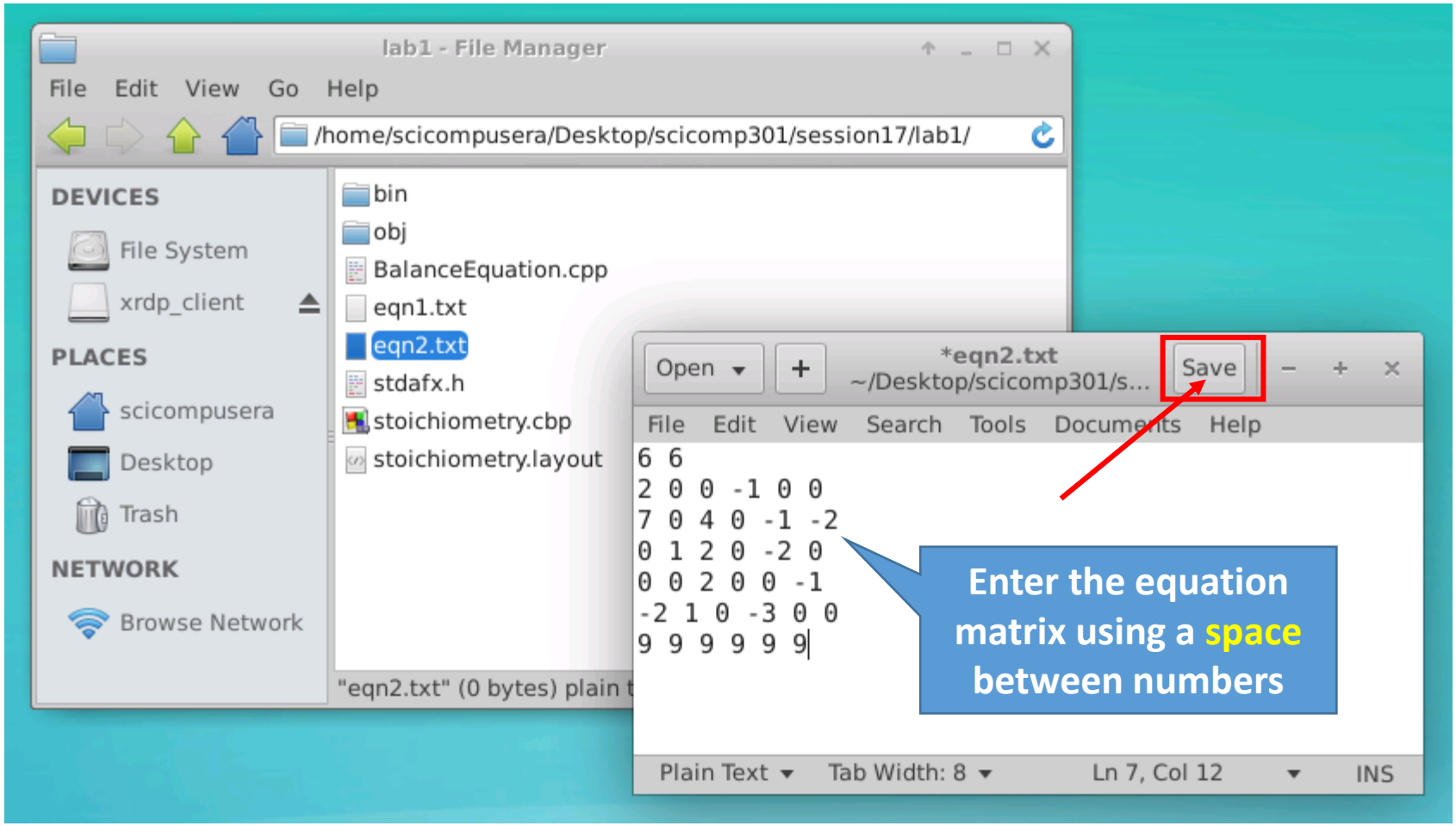
Encoding Chemical Equation 2



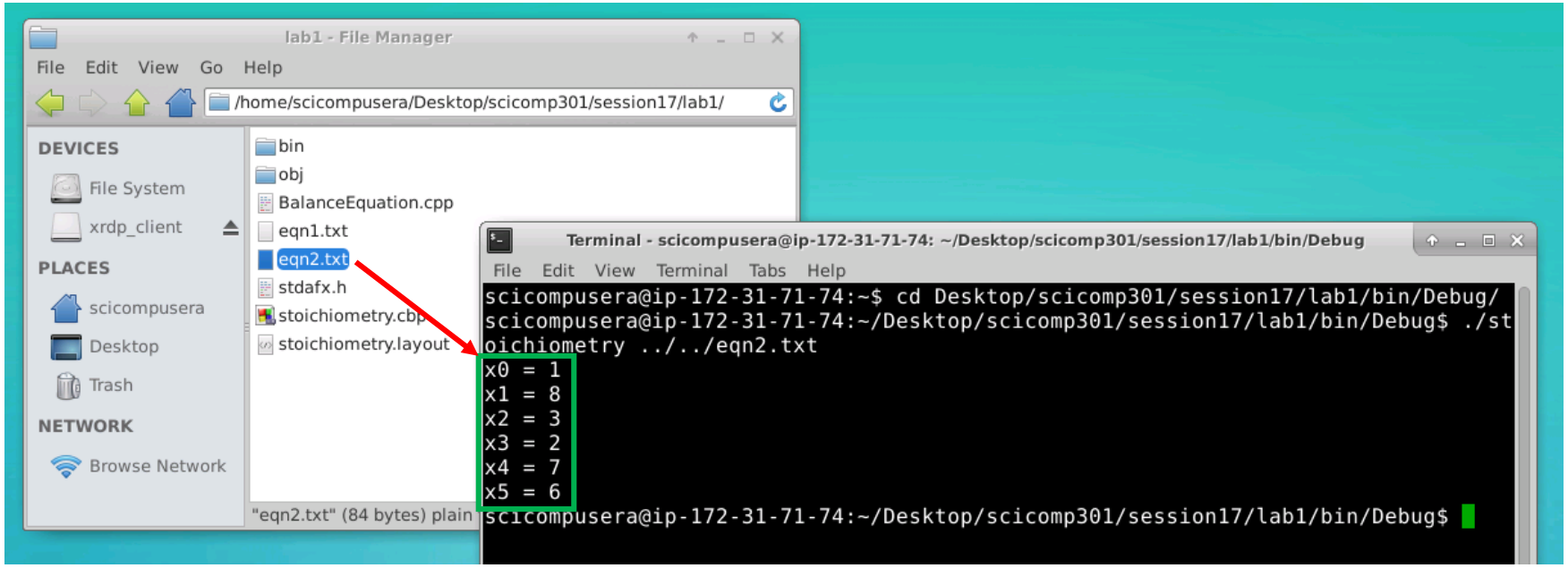
Encoding Chemical Equation 2



Encoding Chemical Equation 2



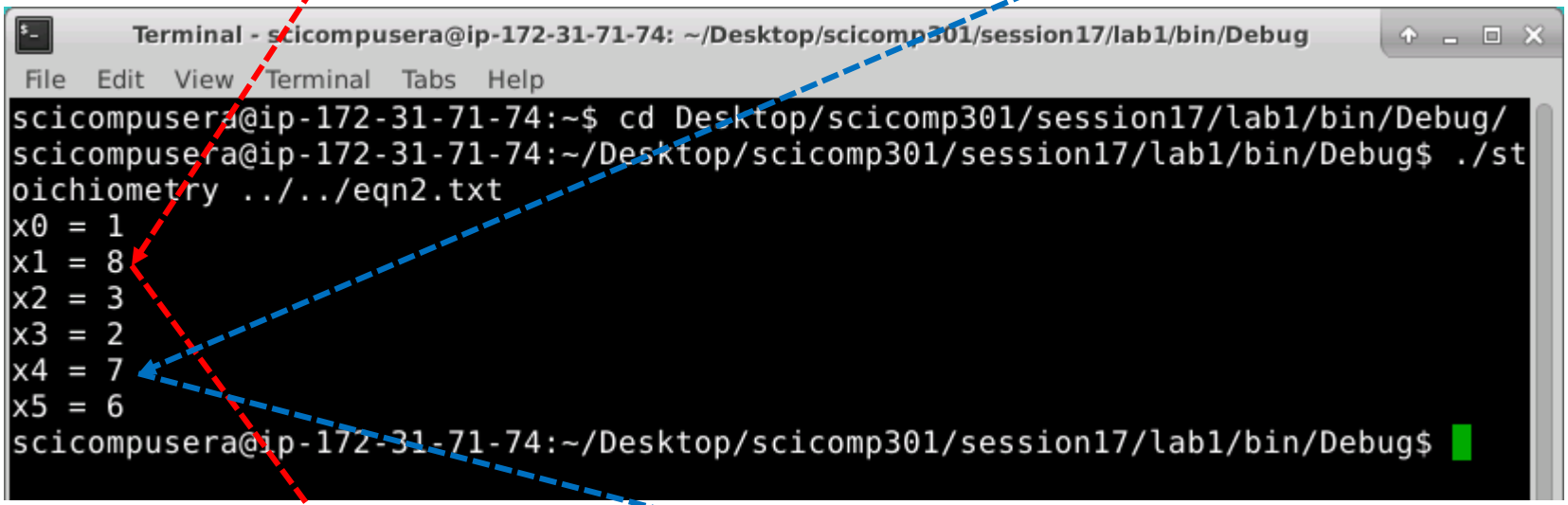
Balancing Chemical Equation 2




```
cd Desktop/scicomp301/session17/lab1/bin/Debug
./stoichiometry ../../eqn2.txt
```

Verify Chemical Equation 2

$$x_0 Cr_2O_7^{2-} + x_1 H^+ + x_2 H_2C_2O_4 \rightarrow x_3 Cr^{3+} + x_4 H_2O + x_5 CO_2$$



$$1 Cr_2O_7^{2-} + 8 H^+ + 3 H_2C_2O_4 \rightarrow 2 Cr^{3+} + 7 H_2O + 6 CO_2$$

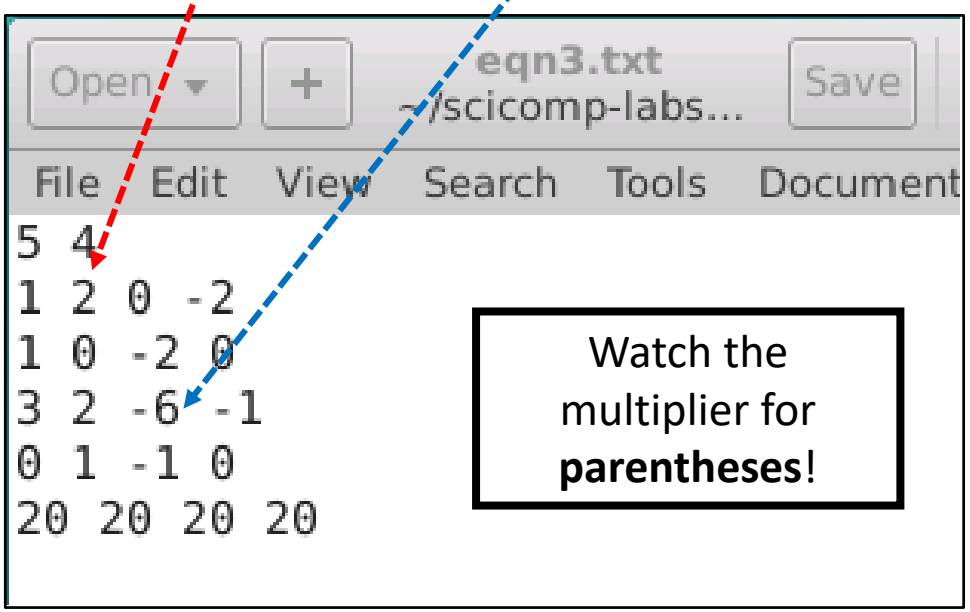
Encoding Chemical Equation 3

$$\mathrm{HNO_3} + \mathrm{Ca(OH)_2} \rightarrow \mathrm{Ca(NO_3)_2} + \mathrm{H_2O}$$

$$\boldsymbol{x_0}\mathrm{HNO_3} + \boldsymbol{x_1}\mathrm{Ca(OH)_2} \rightarrow \boldsymbol{x_2}\mathrm{Ca(NO_3)_2} + \boldsymbol{x_3}\mathrm{H_2O}$$

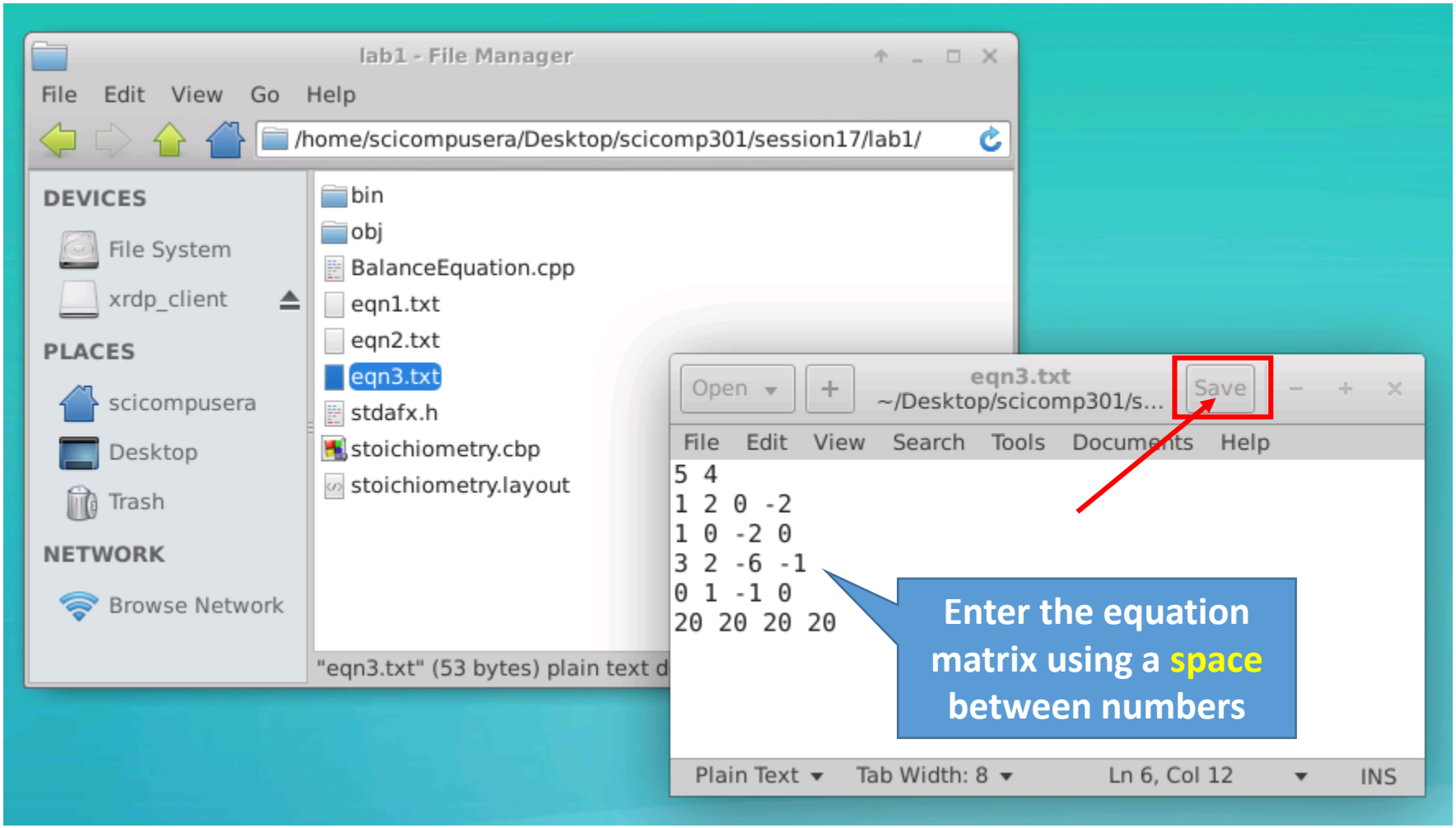
There is no row for **charges** in this equation

Hydrogen (H)
Nitrogen (N)
Oxygen (O)
Calcium (Ca)
Max Value

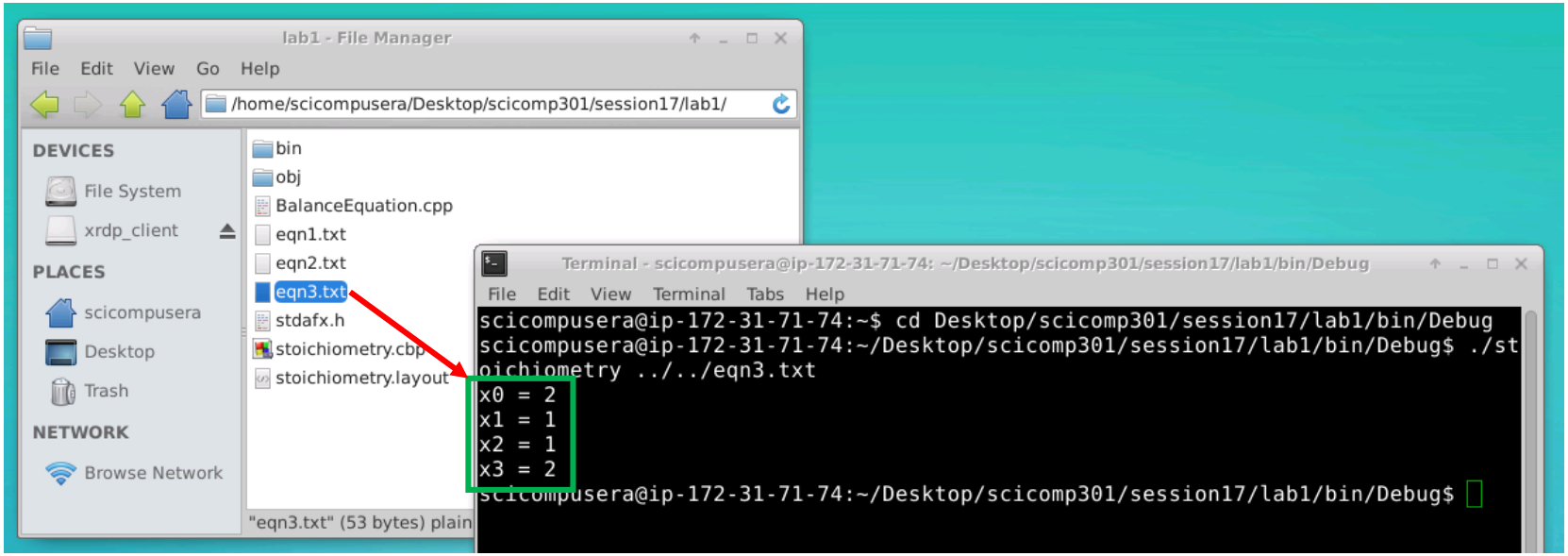


Watch the multiplier for **parentheses**!

Encoding Chemical Equation 3



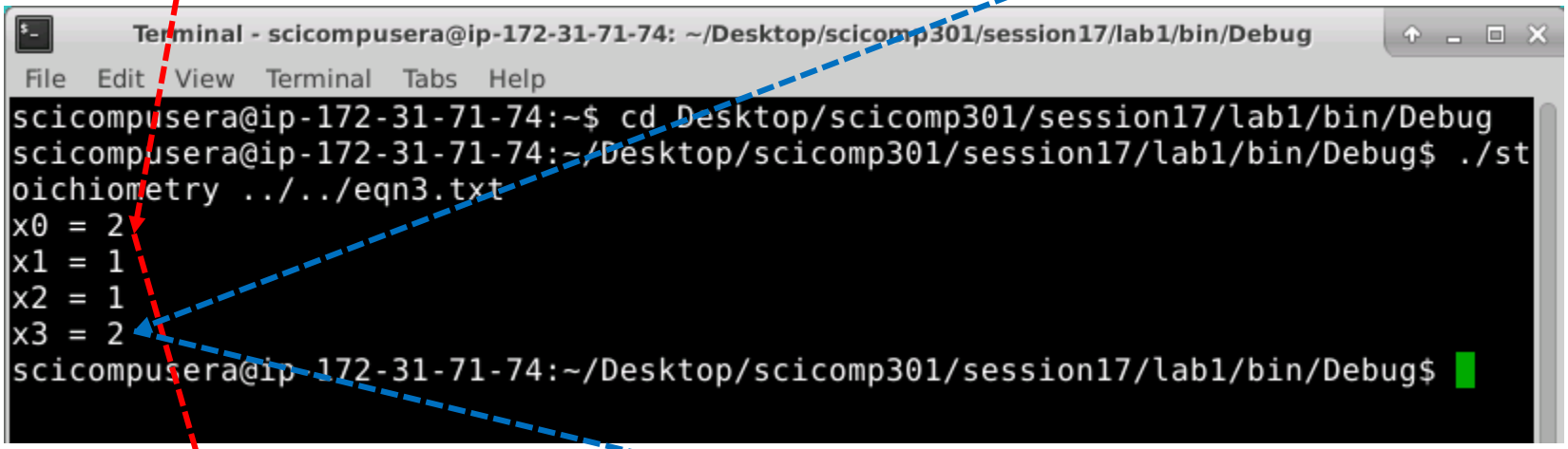
Balancing Chemical Equation 3




```
cd Desktop/scicomp301/session17/lab1/bin/Debug
./stoichiometry ../../eqn3.txt
```

Verify Chemical Equation 3

$$x_0 HNO_3 + x_1 Ca(OH)_2 \rightarrow x_2 Ca(NO_3)_2 + x_3 H_2O$$



$$2HNO_3 + 1Ca(OH)_2 \rightarrow 1Ca(NO_3)_2 + 2H_2O$$

Avoid Writing **Brittle** Code

```
static void NestedForLoops()
{
    for (int x0 = 0; x0 < 50; x0++)
    {
        for (int x1 = 0; x1 < 50; x1++)
        {
            for (int x2 = 0; x2 < 50; x2++)
            {
                for (int x3 = 0; x3 < 50; x3++)
                {
                    for (int x4 = 0; x4 < 50; x4++)
                    {
                        for (int x5 = 0; x5 < 50; x5++)
                        {
                            for (int x6 = 0; x6 < 50; x6++)
                            {
                                // Evaluate objective function
                            }
                        }
                    }
                }
            }
        }
    }
}
```

- We could use nested **for()** loops, but what if number of unknowns ($\boldsymbol{x_n}$) changes?
- Objective function would be evaluated $\mathbf{50^7}$ times =

$$\mathbf{781,250,000,000}$$

- No guarantee that molar coefficients $1 \leq x_n \leq 50$

$$P_2I_4 + P_4 + H_2O \rightarrow PH_4I + H_3PO_4$$

Use Data Structures vs. Brittle Code

```
vector<vector<int>>* stack;
```

Unknown	Index	Current Value	Max Value	
x_0	0	1	9	
x_1	1	1	9	
x_2	1	1	9	<--- stackLevel
x_3	2	1	9	
x_4	3	1	9	
x_5	4	1	9	

Our **stack** is a **vector of vectors**, with a row for **each unknown** and each row having **two** columns.
We also keep a pointer to the current ***stack level***

Use Data Structures vs. Brittle Code

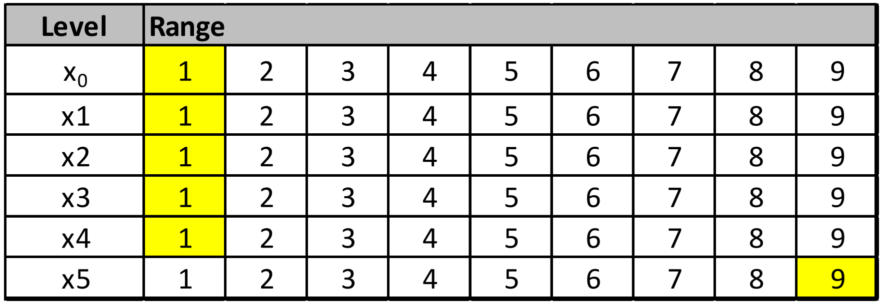
Level	Range								
x_0	1	2	3	4	5	6	7	8	9
x1	1	2	3	4	5	6	7	8	9
x2	1	2	3	4	5	6	7	8	9
x3	1	2	3	4	5	6	7	8	9
x4	1	2	3	4	5	6	7	8	9
x5	1	2	3	4	5	6	7	8	9

- At each iteration, the current level moves to the next position in its range, and checks the constraints

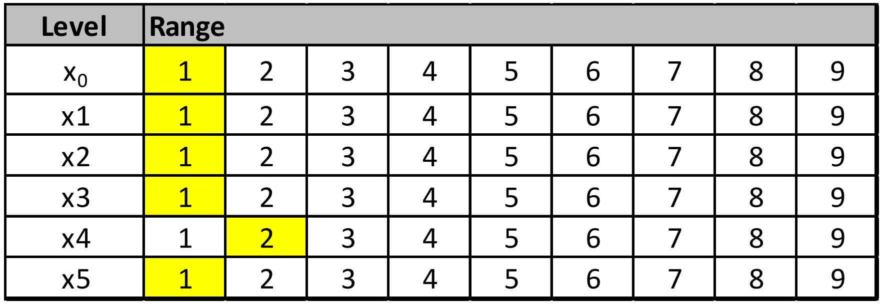
Level	Range								
x_0	1	2	3	4	5	6	7	8	9
x1	1	2	3	4	5	6	7	8	9
x2	1	2	3	4	5	6	7	8	9
x3	1	2	3	4	5	6	7	8	9
x4	1	2	3	4	5	6	7	8	9
x5	1	2	3	4	5	6	7	8	9

- If the equation balances, but a new **minimum** atom count (sum) is found, display the current value for each unknown, and that sum becomes the new minimum

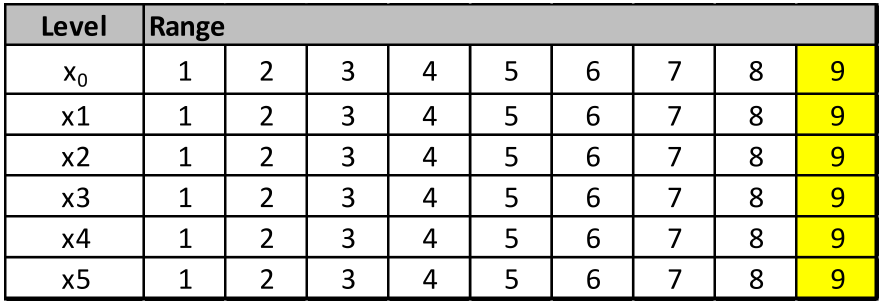
Use Data Structures vs. Brittle Code



Level	Range								
x_0	1	2	3	4	5	6	7	8	9
x1	1	2	3	4	5	6	7	8	9
x2	1	2	3	4	5	6	7	8	9
x3	1	2	3	4	5	6	7	8	9
x4	1	2	3	4	5	6	7	8	9
x5	1	2	3	4	5	6	7	8	9



Level	Range								
x_0	1	2	3	4	5	6	7	8	9
x1	1	2	3	4	5	6	7	8	9
x2	1	2	3	4	5	6	7	8	9
x3	1	2	3	4	5	6	7	8	9
x4	1	2	3	4	5	6	7	8	9
x5	1	2	3	4	5	6	7	8	9



Level	Range								
x_0	1	2	3	4	5	6	7	8	9
x1	1	2	3	4	5	6	7	8	9
x2	1	2	3	4	5	6	7	8	9
x3	1	2	3	4	5	6	7	8	9
x4	1	2	3	4	5	6	7	8	9
x5	1	2	3	4	5	6	7	8	9

- Just as a car odometer rolls over when a digit range reaches its limit, we increment the prior level’s position and reset the current level to 1
- The program ends when every level in the stack has reached the **limit** of its *individually* assigned range

Use Data Structures vs. Brittle Code

```
bool FindSolution()
{
    int stackLevel = 0;
    while (stackLevel >= 0) {
        while (stack->at(stackLevel).at(0)
            < stack->at(stackLevel).at(1))
        {
            if (stackLevel == (int)stack->size() - 1) break;
            stackLevel++;
            stack->at(stackLevel).at(0) = 1;
        }

        if (IsSolution())
            return true;

        stack->at(stackLevel).at(0)++;
        while (stack->at(stackLevel).at(0)
            == stack->at(stackLevel).at(1))
        {
            stackLevel--;
            if (stackLevel < 0) break;
            stack->at(stackLevel).at(0)++;
        }
    }
    return false;
}
```

Find the lowest level in the stack which has not yet enumerated its full range

Invoke the objective function

Use the next value in this level's range

Walk up the stack, skipping all parent levels that have already completed their range

Using Data Structures vs. Brittle Code

```
void LoadEquation()
{
    int constraints, unknowns;
    *infile >> constraints >> unknowns;

    eqn = new vector<vector<int>>(constraints);
    vector<int> constraint(unknowns);

    for (auto& e : *eqn) {
        for (auto& c : constraint)
            *infile >> c;
        e = constraint;
    }
}
```

To read from a file use the stream insertion >> operator

A solution is valid only if the convolution of every matrix row with every stack row **== 0**

```
bool IsSolution()
{
    for (size_t j{};j < eqn->size() - 1;++j) {
        int sum = 0;
        for (size_t i{};i < stack->size();++i)
            sum += eqn->at(j).at(i) * stack->at(i).at(0);
        if (sum != 0)
            return false;
    }
    return true;
}
```

Balancing Ionic Equations

- Pick **one** of the following equations to balance using the `stoichiometry` application
- In the Lab 1 folder, **create** a text file that encodes your chosen equation using the <u>corresponding</u> filename

`eqn4.txt:` $MnO_4^- + Fe^{2+} + H^+ \rightarrow Mn^{2+} + Fe^{3+} + H_2O$

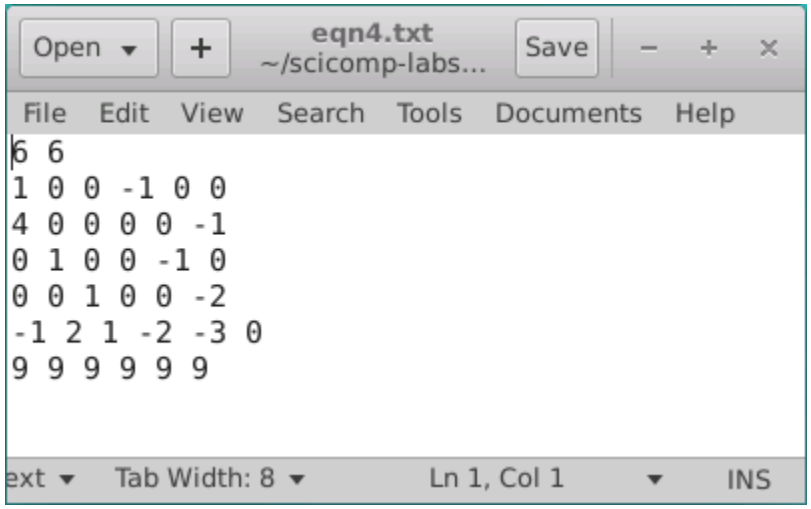
`eqn5.txt:` $C_8H_{18} + O_2 \rightarrow CO_2 + H_2O$

`eqn6.txt:` $C_7H_6O_2 + O_2 \rightarrow CO_2 + H_2O$

`eqn7.txt:` $P_2I_4 + P_4 + H_2O \rightarrow PH_4I + H_3PO_4$

Balancing Chemical Equation 4

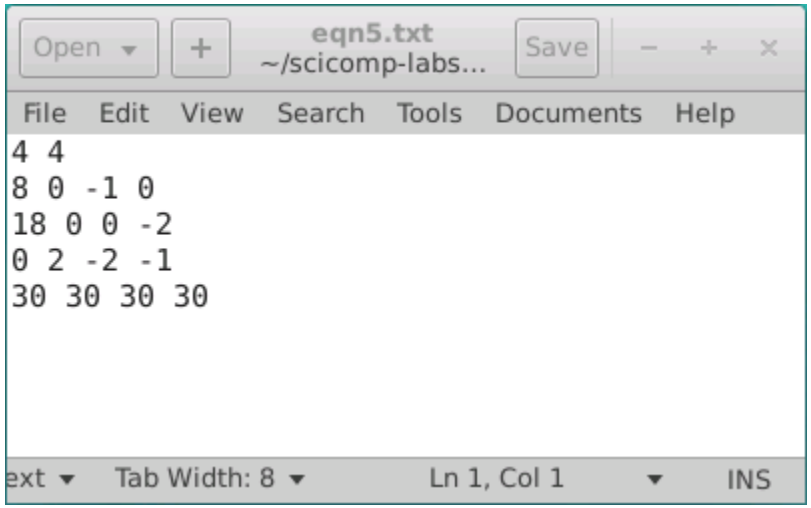
$$MnO_4^- + Fe^{2+} + H^+ \rightarrow Mn^{2+} + Fe^{3+} + H_2O$$



$$\mathbf{1}MnO_4^- + \mathbf{5}Fe^{2+} + \mathbf{8}H^+ \rightarrow \mathbf{1}Mn^{2+} + \mathbf{5}Fe^{3+} + \mathbf{4}H_2O$$

Balancing Chemical Equation 5

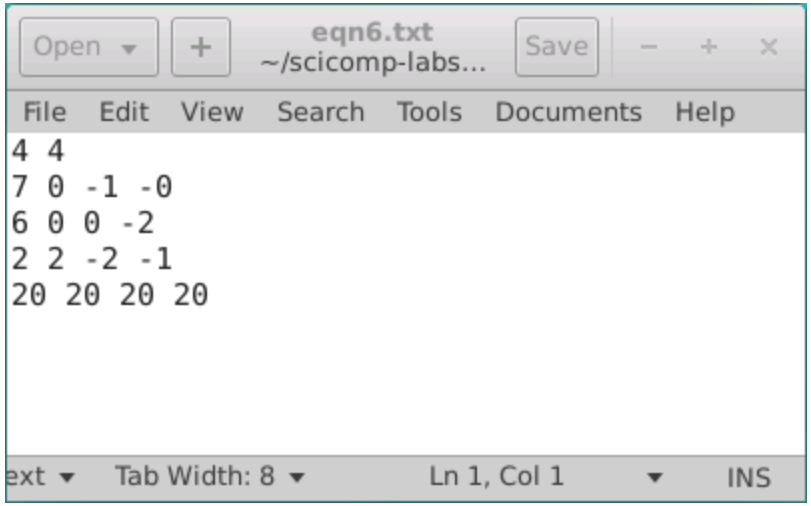
$$C_8H_{18} + O_2 \rightarrow CO_2 + H_2O$$



$$\mathbf{2}C_8H_{18} + \mathbf{25}O_2 \rightarrow \mathbf{16}CO_2 + \mathbf{18}H_2O$$

Balancing Chemical Equation 6

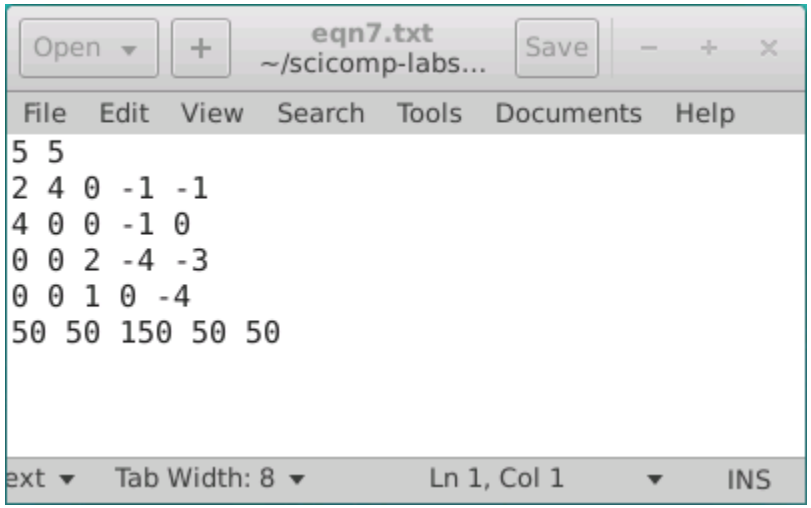
$$C_7H_6O_2 + O_2 \rightarrow CO_2 + H_2O$$



$$\mathbf{2}C_7H_6O_2 + \mathbf{15}O_2 \rightarrow \mathbf{14}CO_2 + \mathbf{6}H_2O$$

Balancing Chemical Equation 7

$$P_2I_4 + P_4 + H_2O \rightarrow PH_4I + H_3PO_4$$



$$\mathbf{10}P_2I_4 + \mathbf{13}P_4 + \mathbf{128}H_2O \rightarrow \mathbf{40}PH_4I + \mathbf{32}H_3PO_4$$

Diphosphorus tetraiodide

Now you know…

https://www.webqc.org/balance.php

- We can encode a chemical equation into a **matrix**, where the **rows** represent the element constraints for reactants and products and the **columns** represent the unknown coefficients of the molecules
- We can transform **balancing ionic equations** to a linear program, which can be solved using brute-force if the search space is sufficiently small
- Even relatively simple equations can exponentially increase the search space & using nested **for()** loops makes the code too **brittle**
- It is better to use the **Simplex** method, and normalize any rational solutions back to integers for the final answer

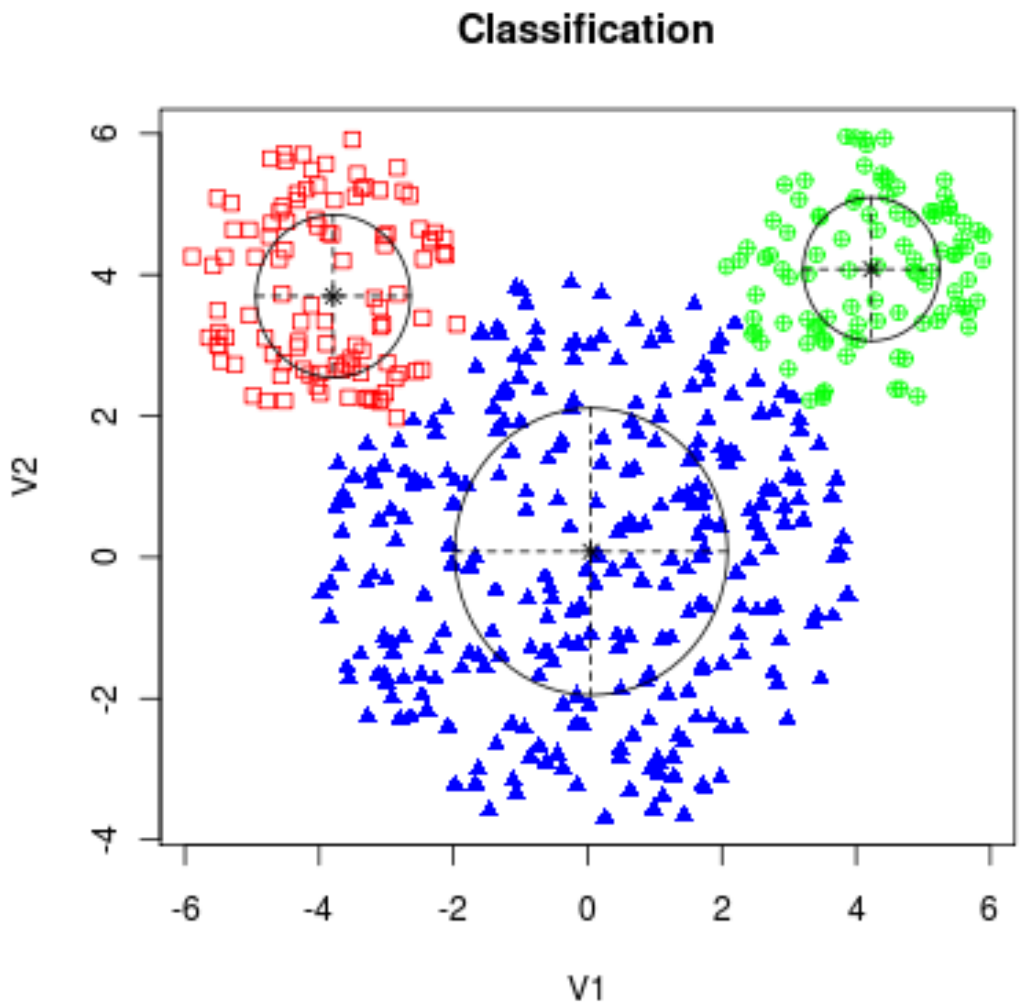
Session Goals

- Study an efficient algorithm for grouping similar **data sets**
- Gain an appreciation for **k-means Clustering** – an algorithm that iterates by **assigning data points to a cluster**
  - The goal is to programmatically assign data points that are "near" each other to the same **cluster** so we can perform group statistics on data that "belongs" together
  - Learn the limitations of k-means and visualize the dangers of "over fitting" data
  - Envision how we can use **variance** to programmatically identify **data outliers**

k-means Clustering

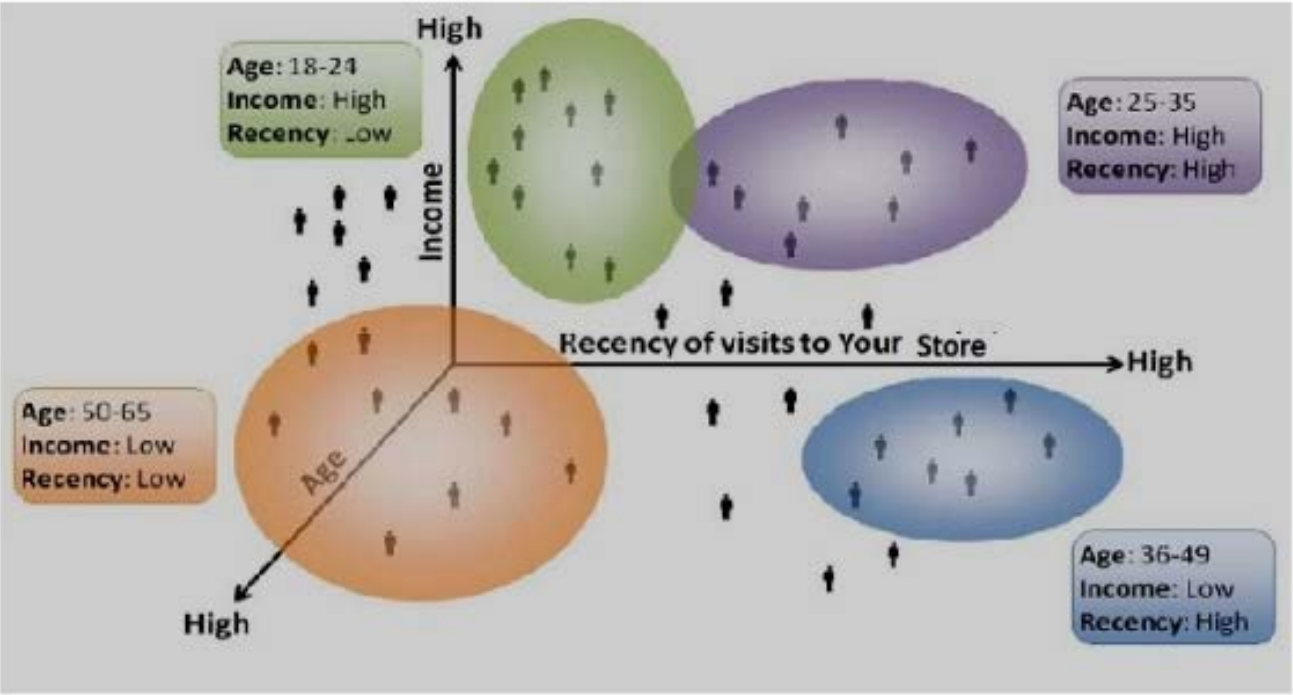
- *k*-means clustering aims to partition ***n*** observations into ***k*** clusters (where each cluster has its own $\boldsymbol{\mu}$ point)
- Each observation is assigned to the cluster whose $\boldsymbol{\mu}$ point **is closest** to that observation's point
- The problem is computationally difficult (NP-hard); however, there are efficient heuristic algorithms that are commonly employed and **converge quickly to a local optimum**
- The term "*k*-means" was first used by James MacQueen in 1967, though the idea goes back to Hugo Steinhaus in 1957
- A more efficient version was published Hartigan and Wong in **1979**

k-means Clustering



k-means Clustering

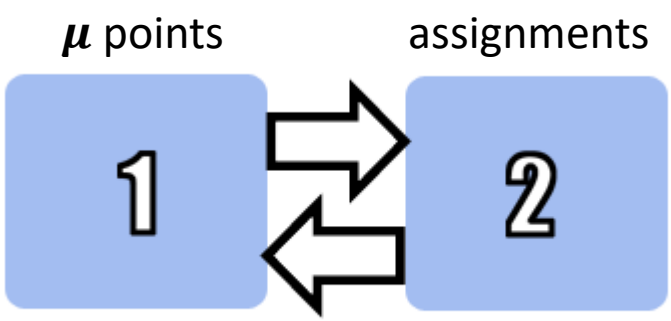
Example - Clusters using Age, Income & Recency



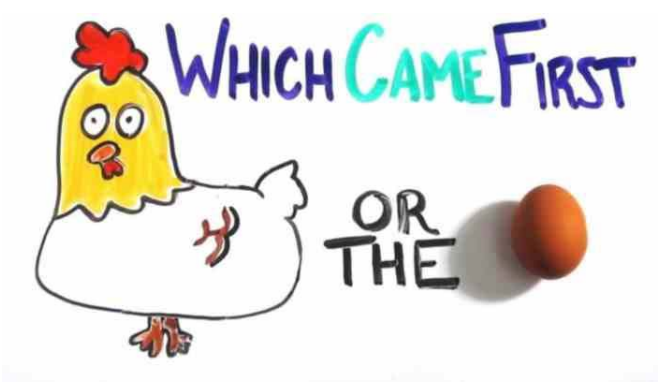
k-means Clustering

- A **one-to-many** mapping between **clusters** and **data points**
  - A **cluster** is a set of one or more unique data points
  - Every **data point** must be <u>assigned</u> to **only one** cluster at a time
  - Data points are initially assigned to clusters in a round-robin fashion
- The $\boldsymbol{\mu}$ point of each **cluster** is its geometric center (**centroid**) calculated by taking the average value (in each dimension) of ***<u>only</u>*** the **data points** assigned to that cluster
- Assignments *can* change the $\boldsymbol{\mu}$ points… which then *can* change the assignments… which then *can* change the $\boldsymbol{\mu}$ points… which then *can* change the assignments… which then *can*…

k-means Clustering



$$G_{\mu\nu} + \Lambda T_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}$$



Spacetime tells matter how to move;
matter tells spacetime how to curve.

— John Archibald Wheeler —

k-means Clustering

- The algorithm iterates in a variable number of two-step rounds:

  1. For each **cluster**, a new centroid ($\boldsymbol{\mu}$ point) is calculated using the ***current*** data point assignments

     If the new $\boldsymbol{\mu} \neq$ current $\boldsymbol{\mu}$ then **move cluster** so current $\boldsymbol{\mu}$ = new $\boldsymbol{\mu}$

  2. For each **data point**, find the cluster which has the closest $\boldsymbol{\mu}$ point using Pythagorean distance formula

     If necessary, **reassign the data point** to the new closest cluster - subject to constraint that every cluster must have one data point

- The algorithm repeats steps #1 and #2 until no cluster moves or data point reassignments are needed ⇒ "**convergence**"

kMeansClustering.h

```
class DataPoint {
public:
    DataPoint();
    DataPoint(double x, double y);
    ~DataPoint();
    double x;
    double y;
    Cluster* c;
};
```

```
class Cluster {
public:
    Cluster();
    Cluster(int index);
    ~Cluster();
    double x;
    double y;
    string clr;
    int population;
};
```

k-means Clustering

```
void InitDataPoints(bool includeOutlier) {

    // Populate data point vector
    dataPoints = new vector<DataPoint*>();
    dataPoints->push_back(new DataPoint(23, 35));
    dataPoints->push_back(new DataPoint(35, 18));
    dataPoints->push_back(new DataPoint(14, 3));
    dataPoints->push_back(new DataPoint(17, 6));
    dataPoints->push_back(new DataPoint(38, 15));
    dataPoints->push_back(new DataPoint(26, 41));
    dataPoints->push_back(new DataPoint(27, 38));
    dataPoints->push_back(new DataPoint(30, 35));
    dataPoints->push_back(new DataPoint(19, 12));
    dataPoints->push_back(new DataPoint(24, 32));
    dataPoints->push_back(new DataPoint(41, 26));
    dataPoints->push_back(new DataPoint(38, 24));
    dataPoints->push_back(new DataPoint(36, 21));
    dataPoints->push_back(new DataPoint(30, 32));
    dataPoints->push_back(new DataPoint(17, 3));
    dataPoints->push_back(new DataPoint(12, 5));
    dataPoints->push_back(new DataPoint(24, 38));
    dataPoints->push_back(new DataPoint(14, 6));
    dataPoints->push_back(new DataPoint(27, 32));
    dataPoints->push_back(new DataPoint(17, 9));

    if (includeOutlier) {
        // A data outlier
        dataPoints->push_back(new DataPoint(5, 40));
    }
}
```

(x, y) coordinates of each data point

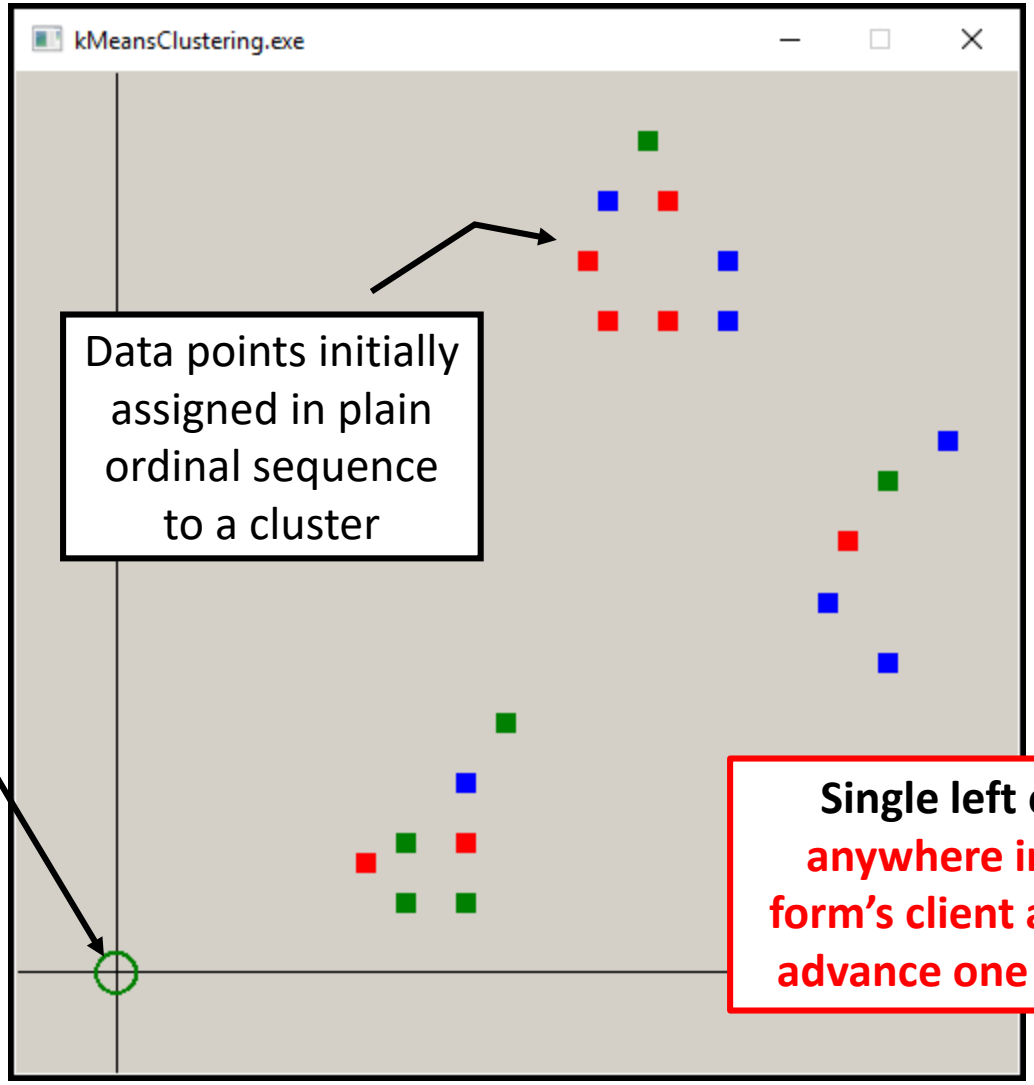
k-means Clustering

```
void InitClusters(int numClusters) {
    // Populate cluster vector
    clusters = new vector<Cluster*>();
    for (int i{}; i < numClusters; ++i) {
        clusters->push_back(new Cluster(i));
    }

    // Assign each data point an initial cluster
    for (size_t i{}; i < dataPoints->size(); ++i) {
        Cluster* c = clusters->at(i % clusters->size());
        dataPoints->at(i)->c = c;
        c->population++;
    }
}
```

k-means Clustering

```
int main()
{
    ss = new SimpleScreen();
    ss->SetWorldRect(-5, -5, 45, 45);

    InitDataPoints(false);

    InitClusters(3);

    DrawClusters();

    while (true){
        ALLEGRO_EVENT ev = ss->Wait();
        if (ev.type == ALLEGRO_EVENT_DISPLAY_CLOSE)
            break;

        if (ev.type == ALLEGRO_EVENT_MOUSE_BUTTON_DOWN) {
            if (UpdateClusters())
                cout << "Cluster has converged!" << endl;
            DrawClusters();
        }
    }
```

k-means Clustering – Round 1

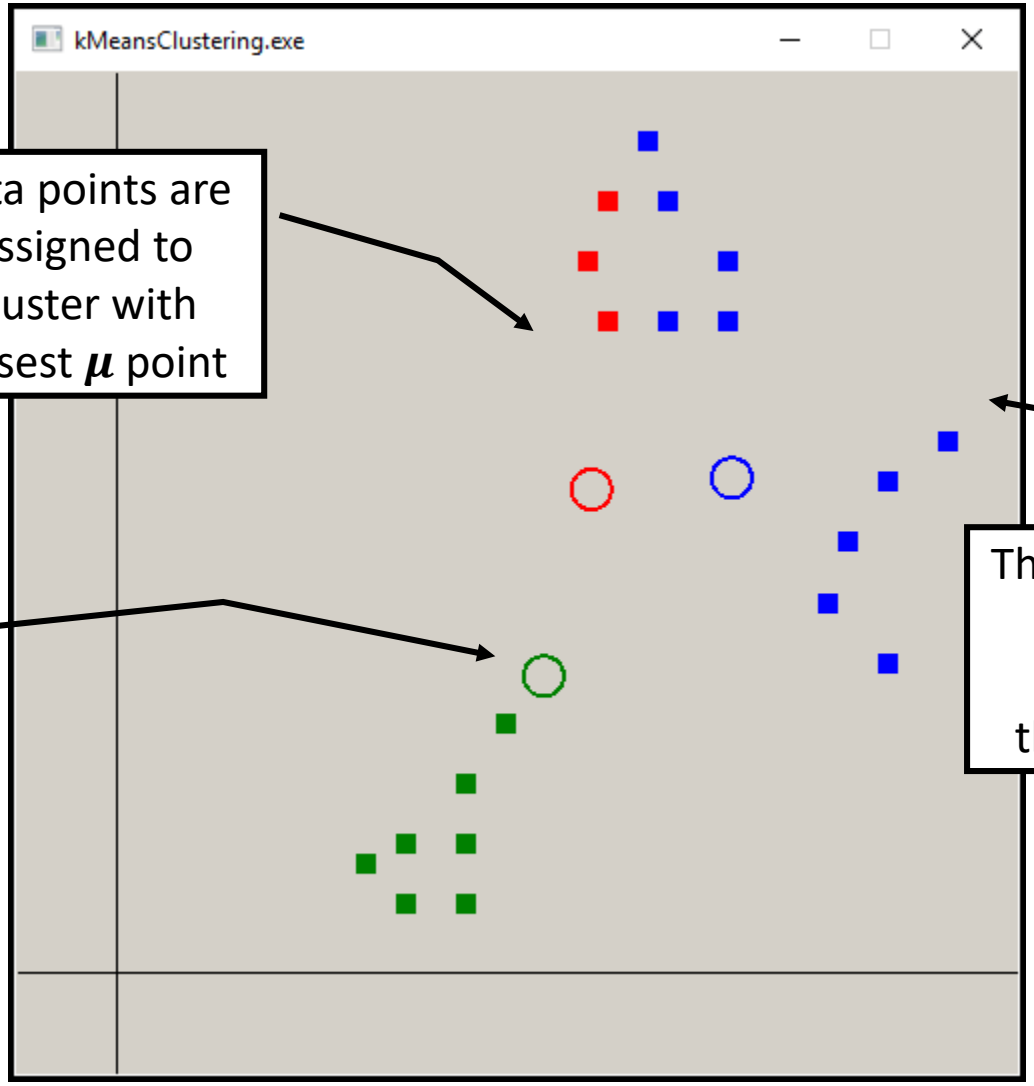


Data points initially assigned in plain ordinal sequence to a cluster

All cluster $\boldsymbol{\mu}$ points start at (0,0)

Single left click anywhere in the form's client area to advance one round

k-means Clustering – Round 2

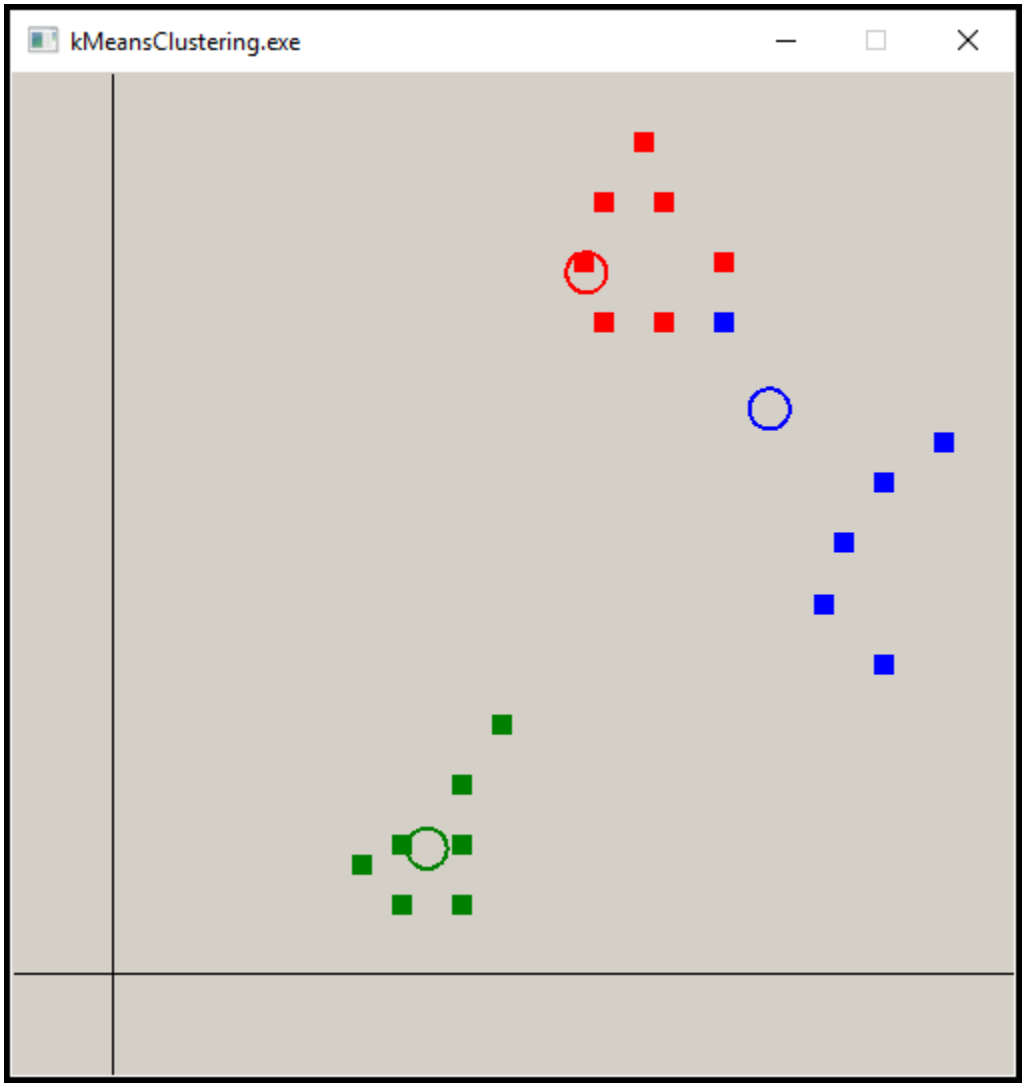


Data points are assigned to cluster with closest $\boldsymbol{\mu}$ point

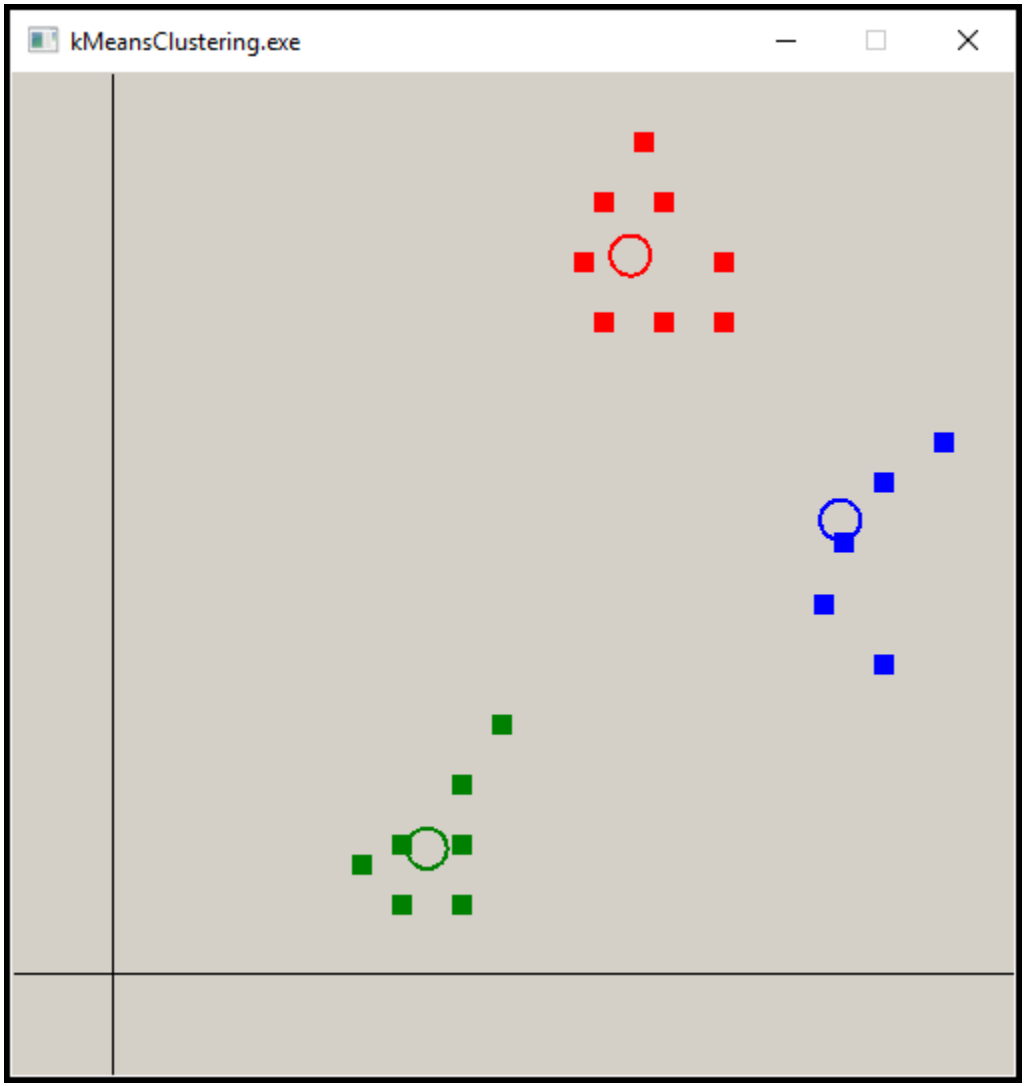
Cluster $\boldsymbol{\mu}$ points have moved

The algorithm is trying to converge to three clusters

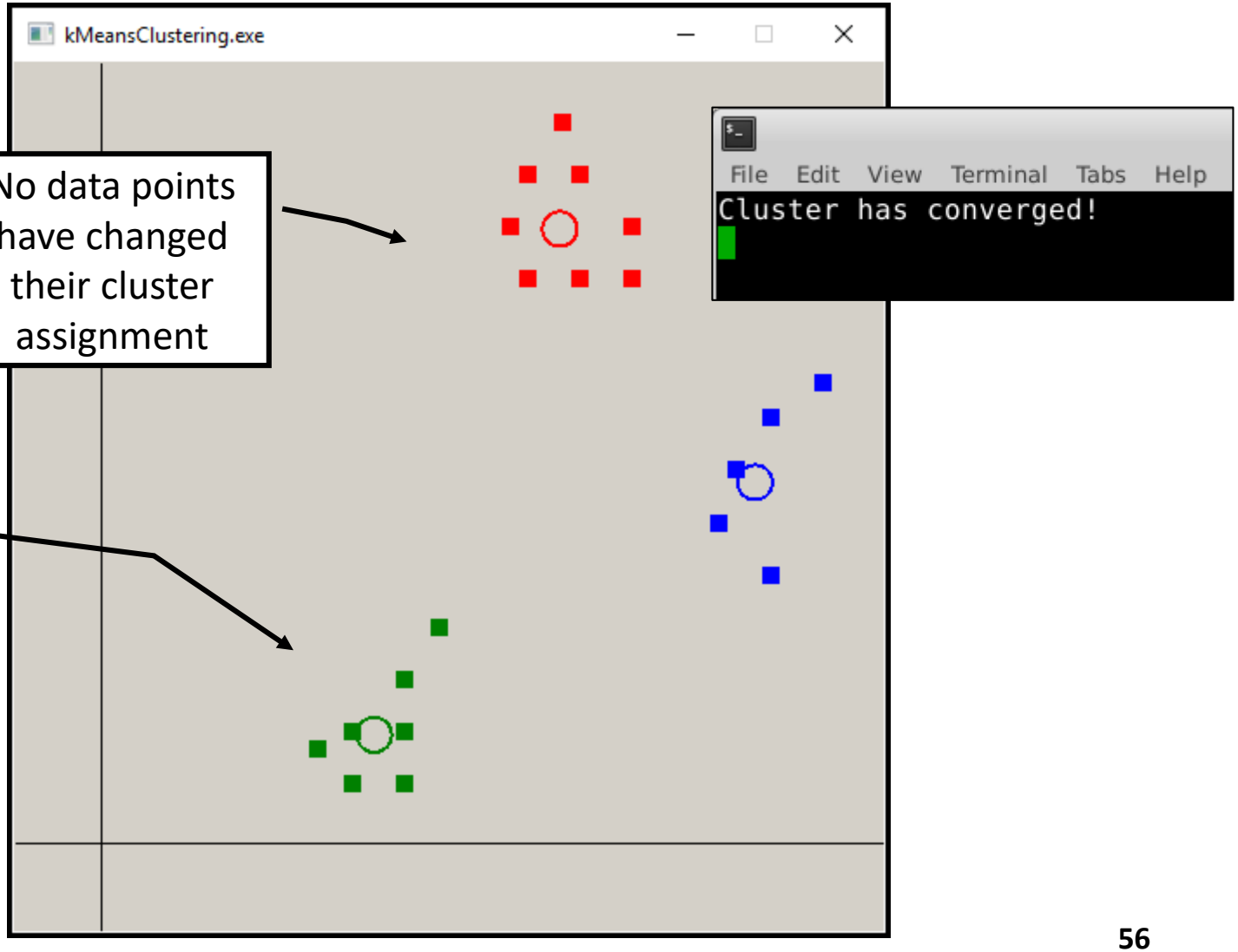
k-means Clustering – Round 3



k-means Clustering – Round 4



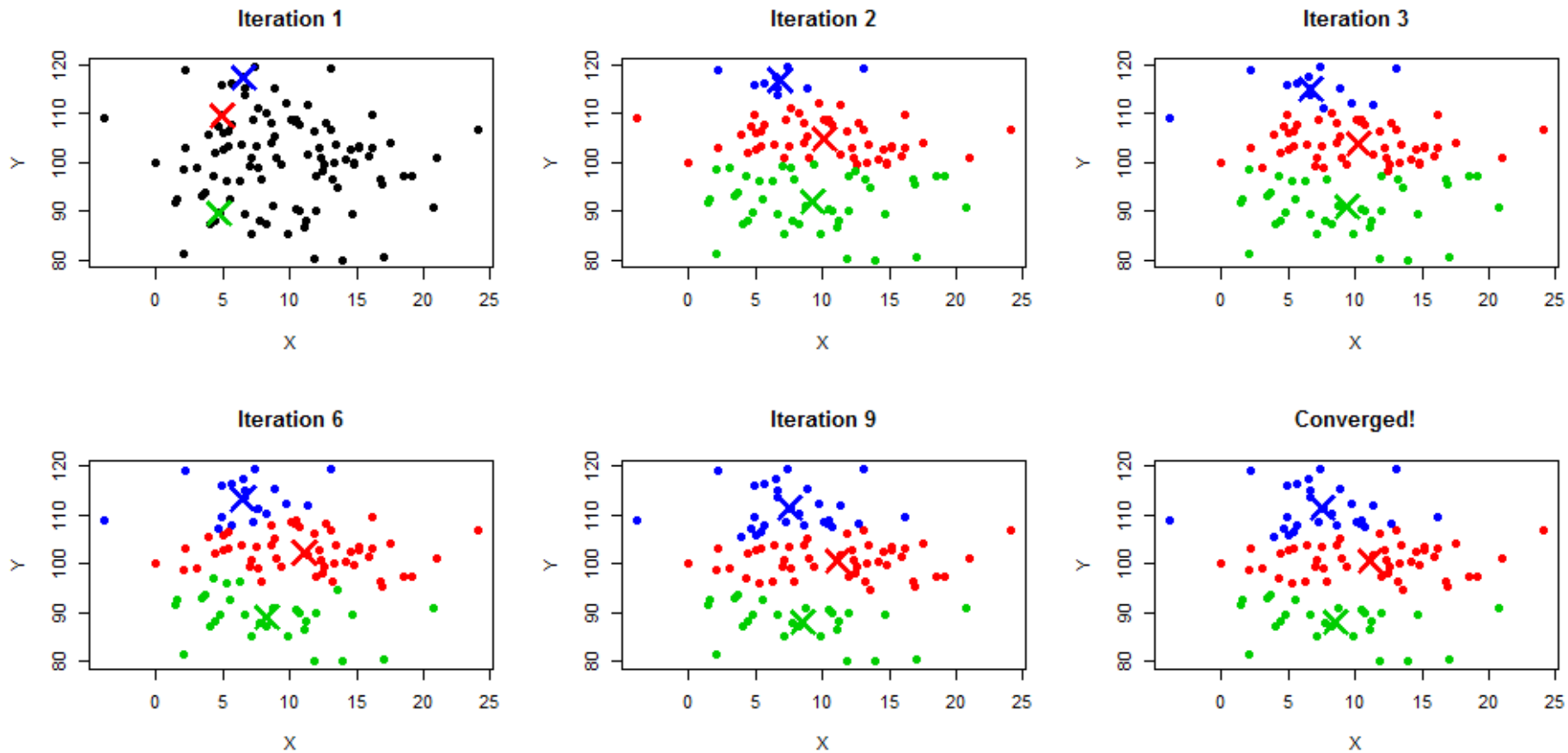
k-means Clustering – Converged



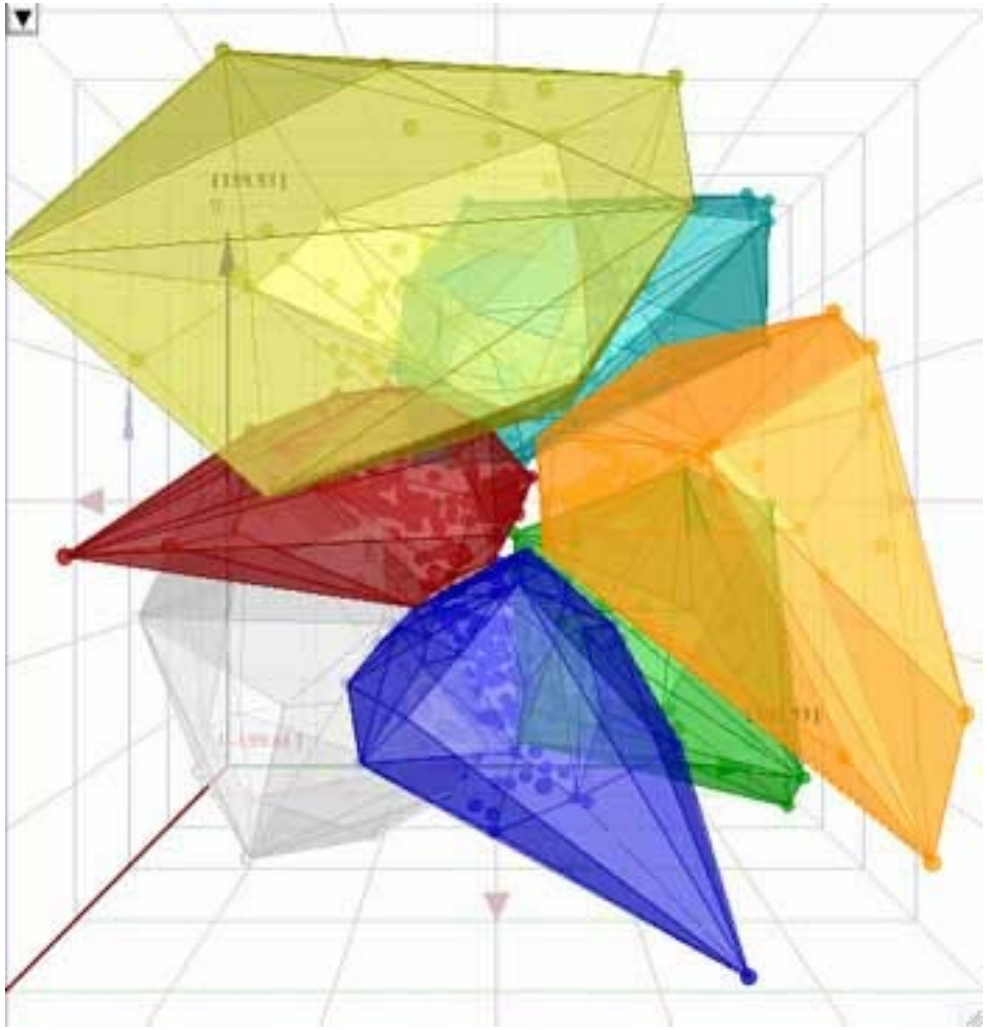
No data points have changed their cluster assignment

No cluster $\boldsymbol{\mu}$ points have moved

k-means Clustering



n-Dimensional k-means Clustering



k-means Clustering

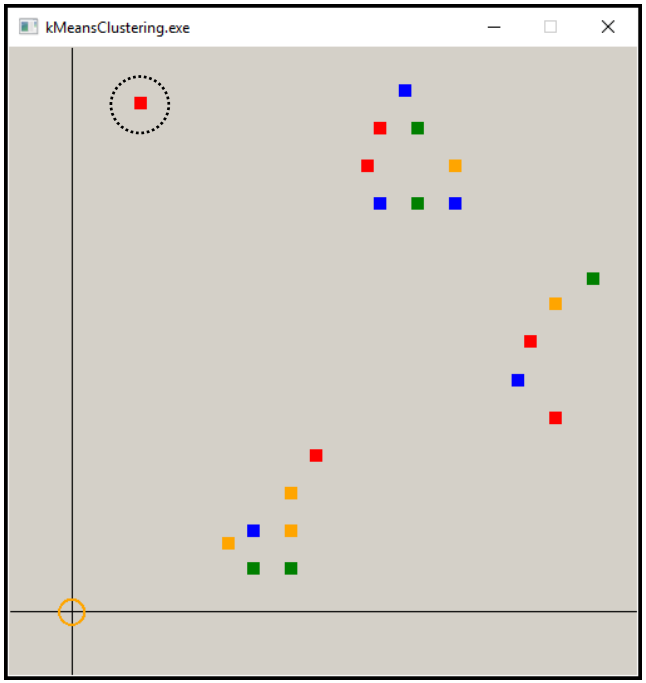
```
bool UpdateClusters() {
    // Assume the k-means clustering is now stable until
    // proven otherwise by either moving means or reassignments
    bool converged = true;

    // Phase I:
    // Calculate new geometric mean of each cluster
    // based upon current data point assignments
    for (auto c : *clusters) {
        DataPoint pt;
        size_t count{};
        for (auto dp : *dataPoints) {
            if (dp->c == c) {
                pt.x += dp->x;
                pt.y += dp->y;
                count++;
            }
        }
        pt.x /= count;
        pt.y /= count;

        // Move cluster center (mean) if necessary
        if (pt.x != c->x || pt.y != c->y) {
            c->x = pt.x;
            c->y = pt.y;
            converged = false;
        }
    }
```

```
// Phase II:
// Assign data points to nearest cluster
for (auto dp : *dataPoints) {
    double distMin = numeric_limits<double>::max();
    Cluster* cNearest = nullptr;
    for (auto c : *clusters) {
        double dist = sqrt(pow(dp->x - c->x, 2) +
            pow(dp->y - c->y, 2));
        if (dist < distMin) {
            distMin = dist;
            cNearest = c;
        }
    }
    if (cNearest != dp->c) {
        // Reassign data point to the new cluster but only
        // if this is not the last point the current cluster.
        // Reassinging should never result in an empty cluster
        if (dp->c->population > 1) {
            dp->c->population--;
            dp->c = cNearest;
            dp->c->population++;
            converged = false;
        }
    }
}

return converged;
```

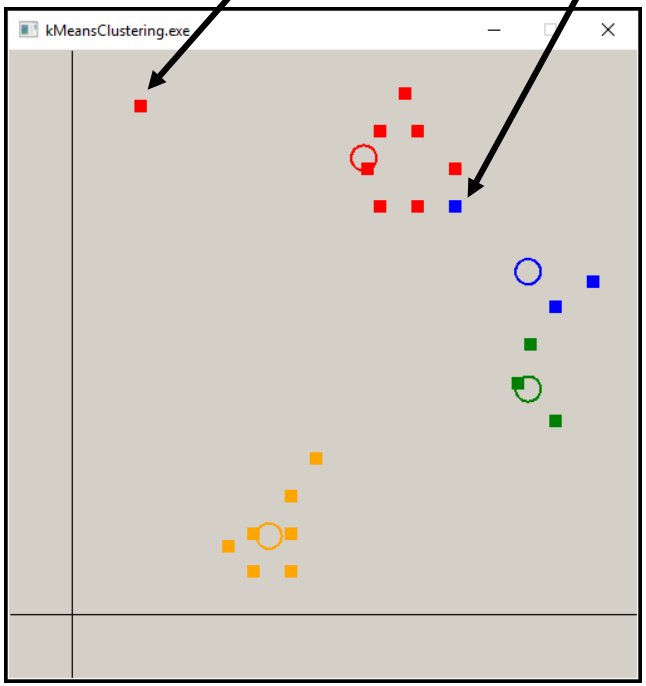
Identifying Data Outliers

main() in "kMeansClustering.cpp"

```
InitDataPoints(true);

InitClusters(4);
```

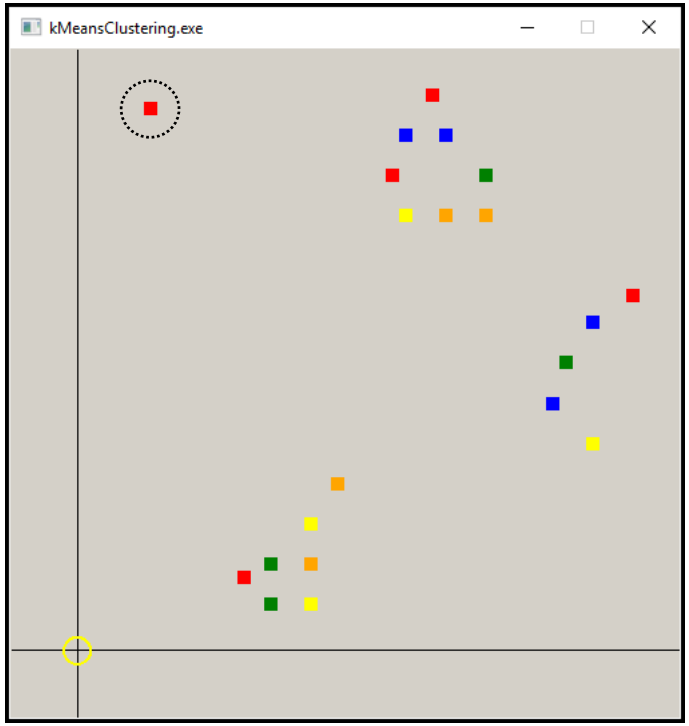


The outlier skews the **red** and **blue** clusters

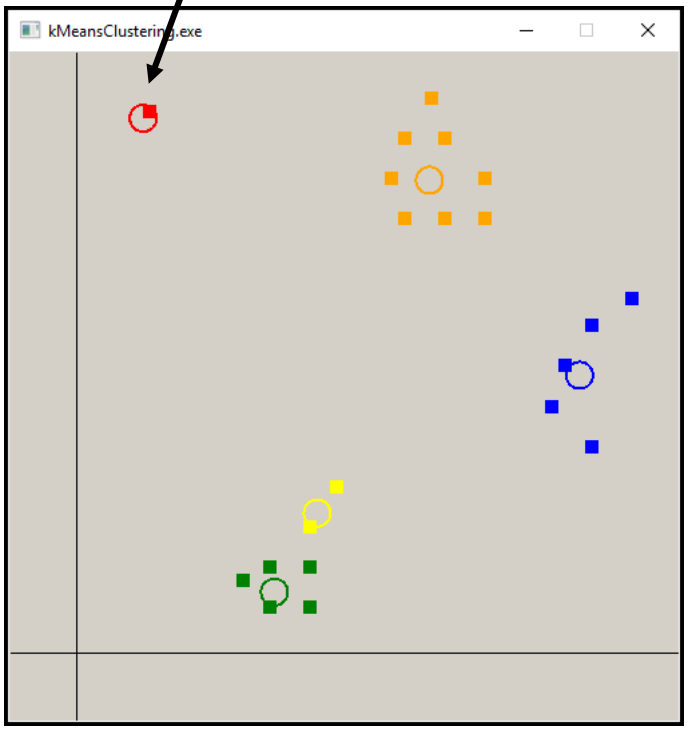


Identifying Data Outliers

```
InitDataPoints(true);

InitClusters(5);
```



The outlier is now its is own cluster



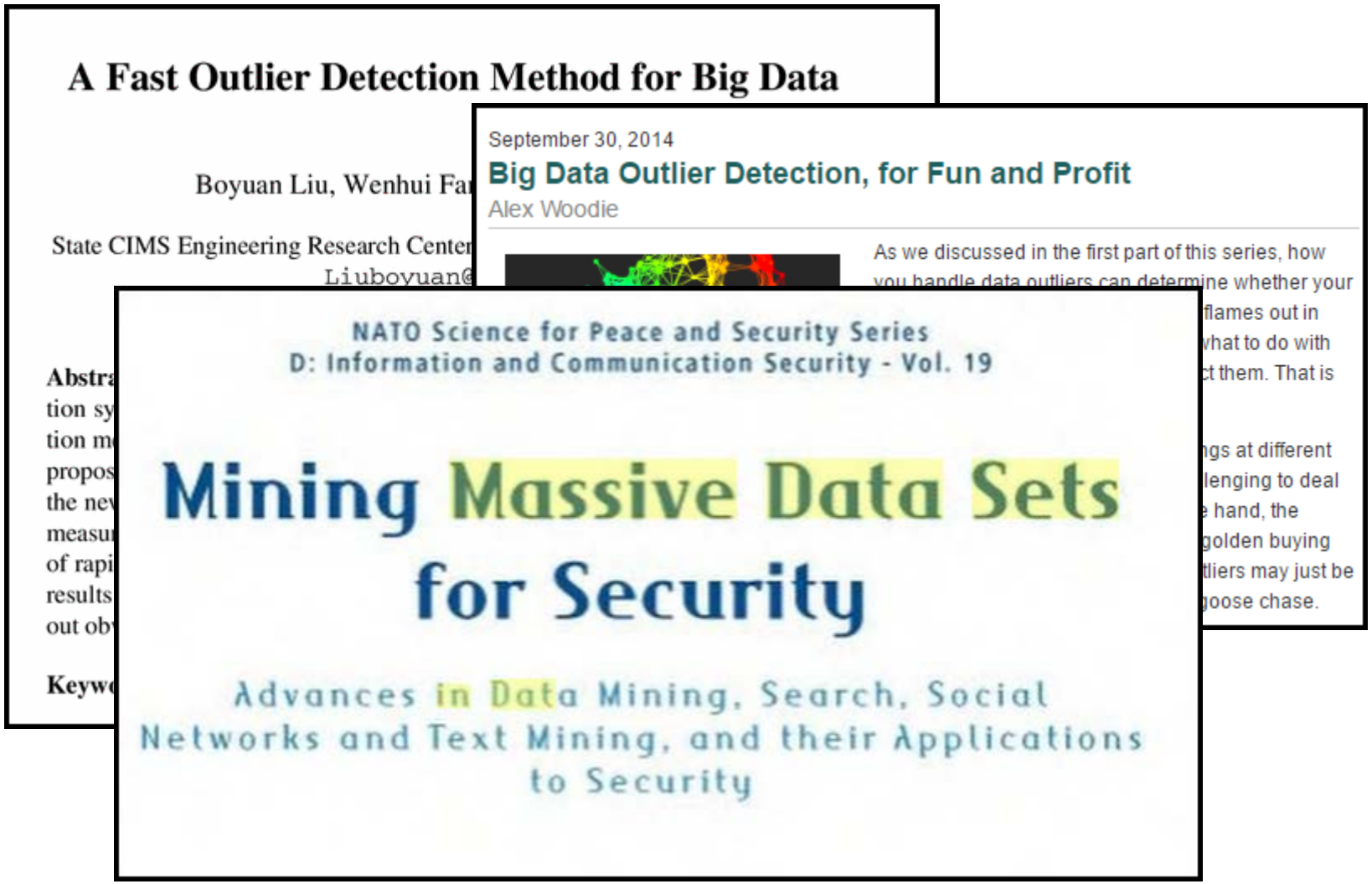
Lab 2 – Research Questions

- Update **main()** code to call `InitDataPoints(false);`
  - Then increase the number of clusters in steps from 3 to 6
  - What happens when number of clusters = 6 ?
  - What does it mean to "**over fit**" the data?
- Update **main()** code to call `InitDataPoints(true);`
  - Then increase the number of clusters in steps from 3 to 6
  - Does using more clusters help identify outliers?
  - Do all "single point" clusters contain likely data errors?
- How could variance ($\boldsymbol{\sigma^2}$) be used to gauge if a $\boldsymbol{\mu}$ point is "appropriate" for a given cluster?

Identifying Data Outliers

- Once we have data partitioned into clusters, we can eliminate outliers by **evicting** any data points whose distance from its converged cluster's $\boldsymbol{\mu}$ point is *too large*
  - The data points should be distributed around the cluster's $\boldsymbol{\mu}$ point in a normal distribution. This means that **99.97%** of data points should be $\leq \mathbf{3}\boldsymbol{\sigma}$ from the mean (geometric center) of the cluster
  - If a data point is $> \mathbf{3}\boldsymbol{\sigma}$ distance from the $\boldsymbol{\mu}$ point of its cluster, it probably is an outlier and should be evicted
  - After we evict a data point, we iterate the algorithm all over again
- If a data point winds up getting evicted from ***all*** clusters, it is probably a **data capture error** or a **statistical anomaly**

Identifying Data Outliers



Now you know…

- We can use **k-means clustering** to categorize data points by its nearness to a similar group's statistical means
  - Data sets which are homogenous (have a more uniform distribution of values) **may not converge** to a meaningful set of clusters since the data lacks sharpness
  - We can use k-means **to identify data outliers** so we can avoid allowing those anomalies from skewing the main observations – this process is called evicting a point from the cluster
- We can extend k-means clustering to analyze **multi-dimensional data sets** to look for trends
  - This is exactly how Amazon and Netflix provide you purchase recommendations based upon your shared interests & tastes compared with other buyers – a **very effective** selling technique!