# Cellular Metabolism
## Published Research Analysis

- A data science project for REDOX
  Danhui Zhang
  April 25, 2019

## Overview

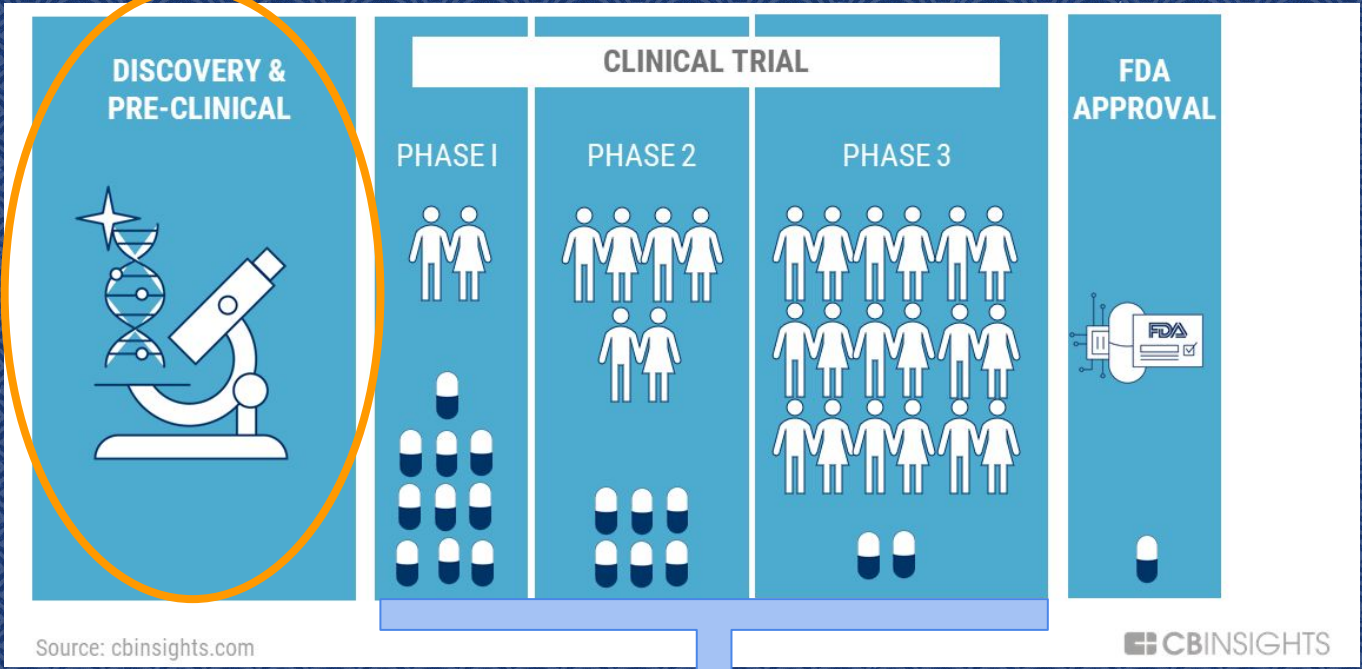
- Intro to REDOX
- Motivation for Dataset Creation
- Dataset Composition
- Data Collection Process
- Relevance of Data Analysis to REDOX Sprint
- Exploratory Data Analysis
- Unsupervised Learning
  - Merge Datasets
  - Data Preprocessing
  - Next Steps

## Overview

- **Intro to REDOX**
- Motivation for Dataset Creation
- Dataset Composition
- Data Collection Process
- Relevance of Data Analysis to REDOX Sprint
- Exploratory Data Analysis
- Unsupervised Learning
  - Merge Datasets
  - Data Preprocessing
  - Next Steps

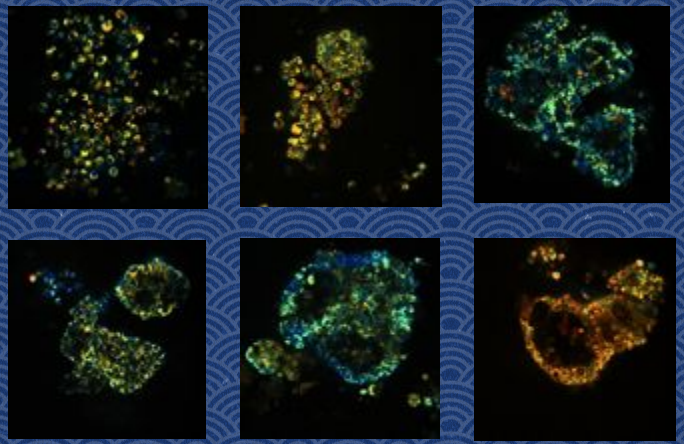



Provide a **vital decision-making tool** for **preclinical-stage drug researchers** to **verify** their **hypothesized drug mechanism of action**, so that **drugs that lack efficacy can be eliminated** from the development pipeline early on in the less costly stages, **before incurring massive costs in the clinical stages.**




>> $10,000 per patient

# Our Service



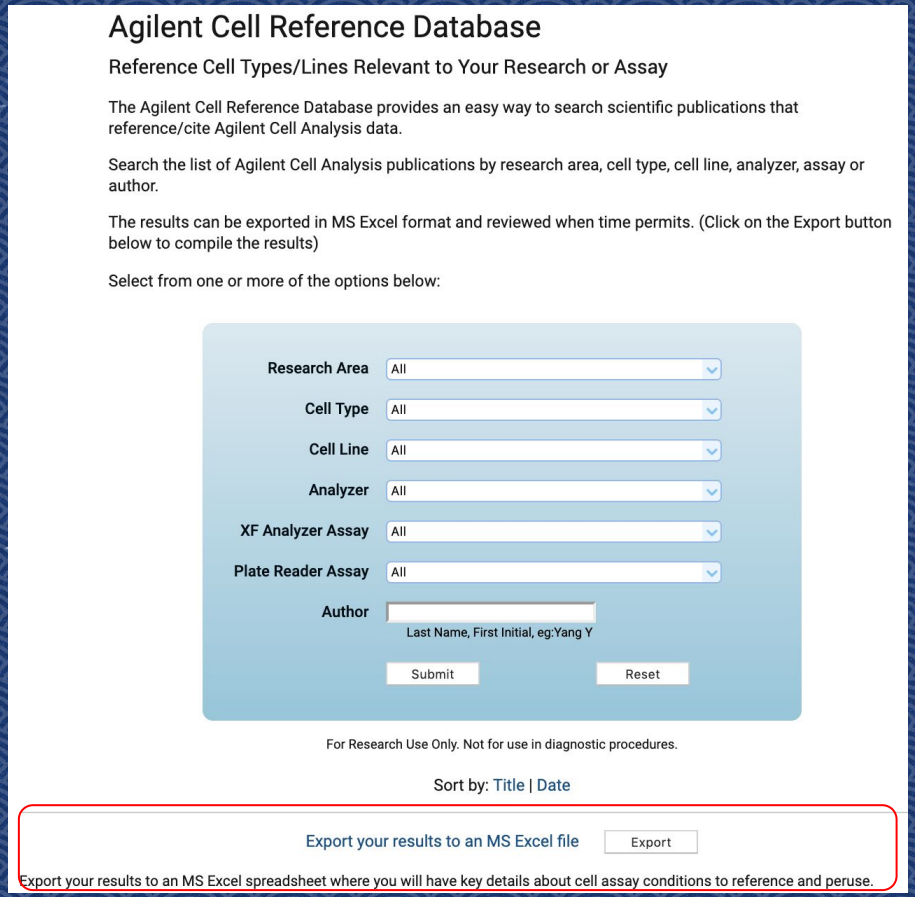


Contrary to prevailing practices of looking at bulk metrics, our service can provide information on **metabolic changes on a per-cell basis**

**Non-destructive** ⇒ can monitor **dynamic** changes over time

## Overview

- Intro to REDOX
- **Motivation for Dataset Creation**
- Dataset Composition
- Data Collection Process
- Relevance of Data Analysis to REDOX Sprint
- Exploratory Data Analysis
- Unsupervised Learning
  - Merge Datasets
  - Data Preprocessing
  - Next Steps




**Agilent** (our competitor) provides **researchers** (our target customers) an easy way to look up **scientific publications** that use **Agilent Cellular Metabolism assays** (an alternative to our service)

Inbound marketing technique to provide incentive to purchase Agilent's labware

## Overview

- Intro to REDOX
- Motivation for Dataset Creation
- **Dataset Composition**
- **Data Collection Process**
- Relevance of Data Analysis to REDOX Sprint
- Exploratory Data Analysis
- Unsupervised Learning
  - Merge Datasets
  - Data Preprocessing
  - Next Steps

| | Title | Authors | Journal | Publication date | Research area | Cell Line | Cell Type | Species | Analyzer | XF Assay | Plate Reader Assay | Seeding density | Plate coating |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | _-1-Antitrypsin (AAT)-modified donor cells sup... | Marcondes AM, Karoopongse E, Lesnikova M, Marg... | Blood | Oct 1 2014 12:00 AM | Immunology Research | T-cells | T-cells | Mouse | 24 | Cell Mitochondrial Stress Test | NaN | 8.0x10^5 cells/well | Not Specified |
| 1 | _-1-Antitrypsin (AAT)-modified donor cells sup... | Marcondes AM, Karoopongse E, Lesnikova M, Marg... | Blood | Oct 1 2014 12:00 AM | Immunology Research | Natural Killer (NK) cells | Immune Cells | Mouse | 24 | Cell Mitochondrial Stress Test | NaN | 8.0x10^5 cells/well | Not Specified |
| 2 | _-enolase regulates the malignant phenotype of... | J. Dai, Q. Zhou, J. Chen, M. L. Rexius-Hall, J... | Nat Commun | Sep 21 2018 12:00 AM | Cell Physiology Research | Pulmonary Artery Smooth Muscle Cells (PASMC) | Pulmonary Artery Cells | Human | 24 | Cell Mitochondrial Stress Test | NaN | 3.0 x10^4 cells/well | Not Specified |
| 3 | _-enolase regulates the malignant phenotype of... | J. Dai, Q. Zhou, J. Chen, M. L. Rexius-Hall, J... | Nat Commun | Sep 21 2018 12:00 AM | Cell Physiology Research | Pulmonary Artery Smooth Muscle Cells (PASMC) | Pulmonary Artery Cells | Human | 24 | Glycolysis Stress Test | NaN | 3.0 x10^4 cells/well | Not Specified |
| 4 | _-ketoglutarate orchestrates macrophage activa... | P. S. Liu, H. Wang, X. Li, T. Chao, T. Teav, S... | Nat Immunol | Sep 1 2017 12:00 AM | Immunology Research | Bone Marrow-Derived Macrophages (BMDM) | Immune Cells | Mouse | 96 | Cell Mitochondrial Stress Test | NaN | 1.0 x10^5 cells/well | Not Specified |



- 9634 instances
- **Each instance**: a single assay type done on **one cell line** and **one cell type** of **one species** (animal or human), considered **relevant in one research area**.

- Timeframe: Aug 2009 - Present
- Agilent, the data collector, has the incentive to be exhaustive in data collection
- Initial work done: modify encoding through Google Drive

## Overview

- Intro to REDOX
- Motivation for Dataset Creation
- Dataset Composition
- Data Collection Process
- **Relevance of Data Analysis to REDOX Sprint**
- Exploratory Data Analysis
- Unsupervised Learning
  - Merge Datasets
  - Data Preprocessing
  - Next Steps

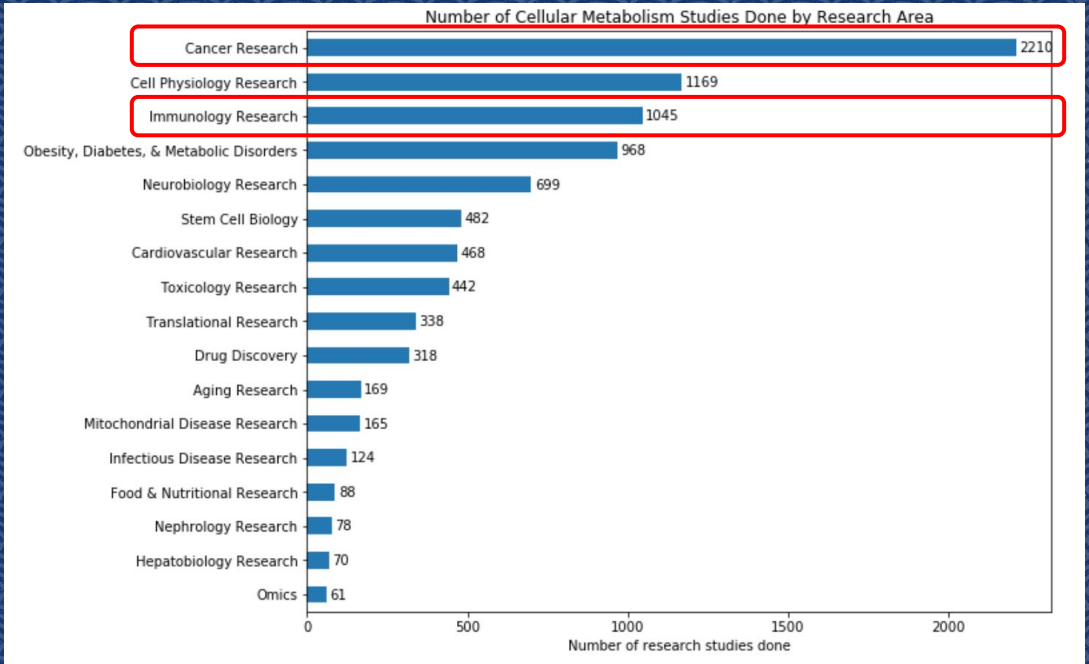
## Goals of this project for REDOX

1. **Market Research**:
   a. **Which Research Areas** Are Scientists Currently Looking for Cellular Metabolism Insights?
   b. What Is the Recent **Trend** for This Interest in Cellular Metabolism Studies?
   c. What Is Our **Initial Market Size**?
2. **Competitor Analysis**:
   a. What Are the Most Popular Products with Our Competitor? (Metabolic Assay Types)
3. **Develop Technical Value Prop:**
   a. What Are the Most Investigated Research Areas?
   b. What Are the Most Investigated Cell Types?
   c. What Are the Initial Disease Models To Focus On?
4. **Develop Implementation Plan:**
   a. How Much Time Do We Need To Service Each Customer?

## Overview

- Intro to REDOX
- Motivation for Dataset Creation
- Dataset Composition
- Data Collection Process
- Relevance of Data Analysis to REDOX Sprint
- **Exploratory Data Analysis**
- Unsupervised Learning
  - Merge Datasets
  - Data Preprocessing
  - Next Steps

# Q1-a Which Research Areas Are Scientists Currently Looking for Cellular Metabolism Insights?

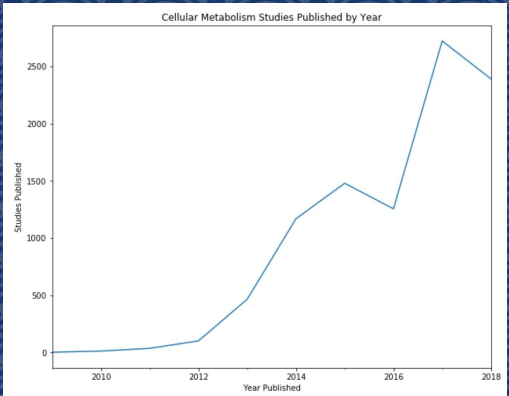



Work done:

- Data Preprocessing
  - Clean up labels
  - Remove irrelevant labels (e.g. “Review Paper”)
- Data Visualization
  - Plot bar chart
  - Add bar labels
  - Keep categories with values > threshold

# Q1-b What Is the Recent Trend for This Interest in Cellular Metabolism Studies?

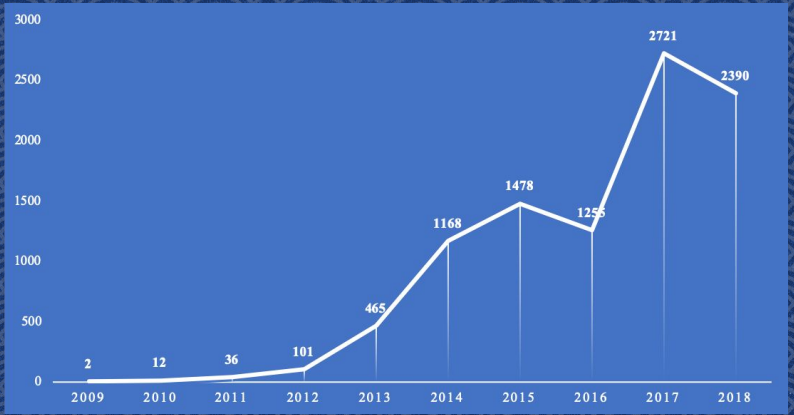
Plotted in Jupyter:




Work done:

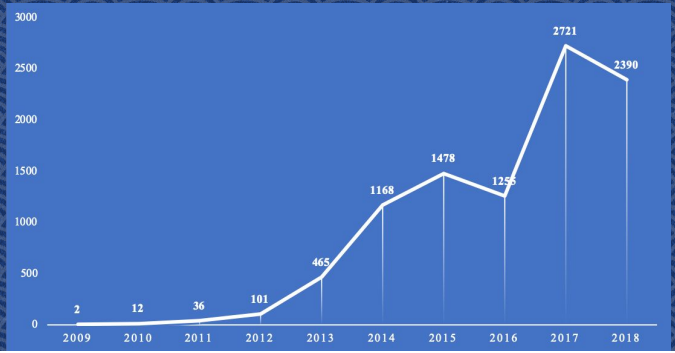
- Data Preprocessing
  - Convert object to string, then split to extract year
  - Sort values & aggregate based on year
- Data Visualization
  - Plot line graph

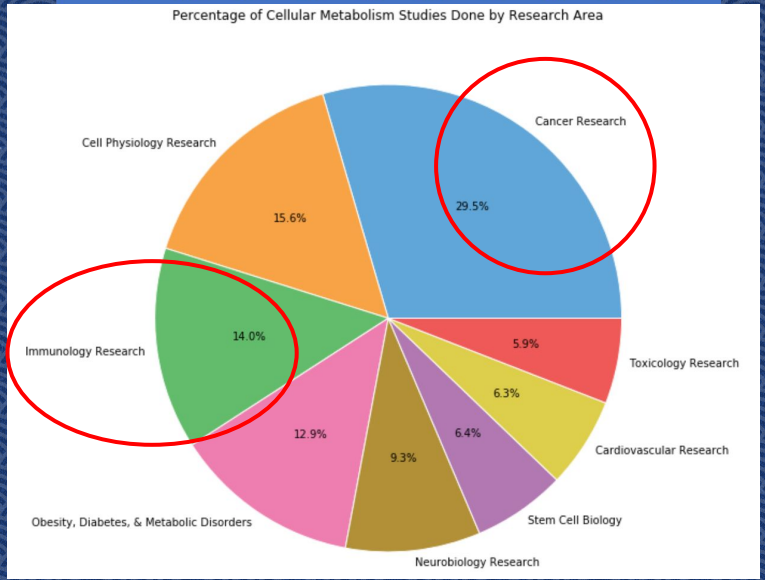
Plotted in Excel using data analyzed in Jupyter:




| Publication date |
| --- |
| Oct 1 2014 12:00 AM |
| Oct 1 2014 12:00 AM |
| Sep 21 2018 12:00 AM |
| Sep 21 2018 12:00 AM |

# Q1-c What Is Our Initial Market Size?

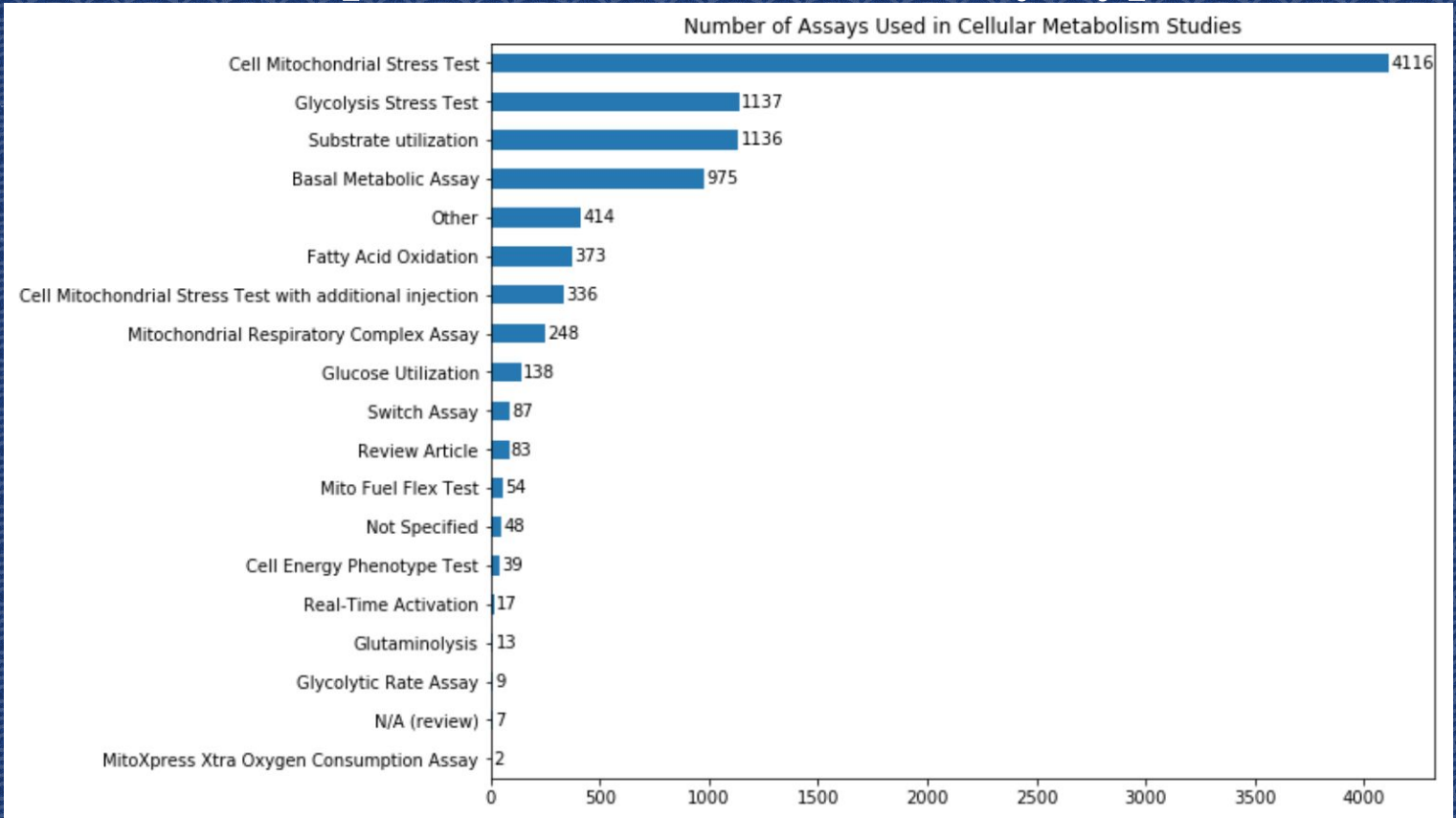


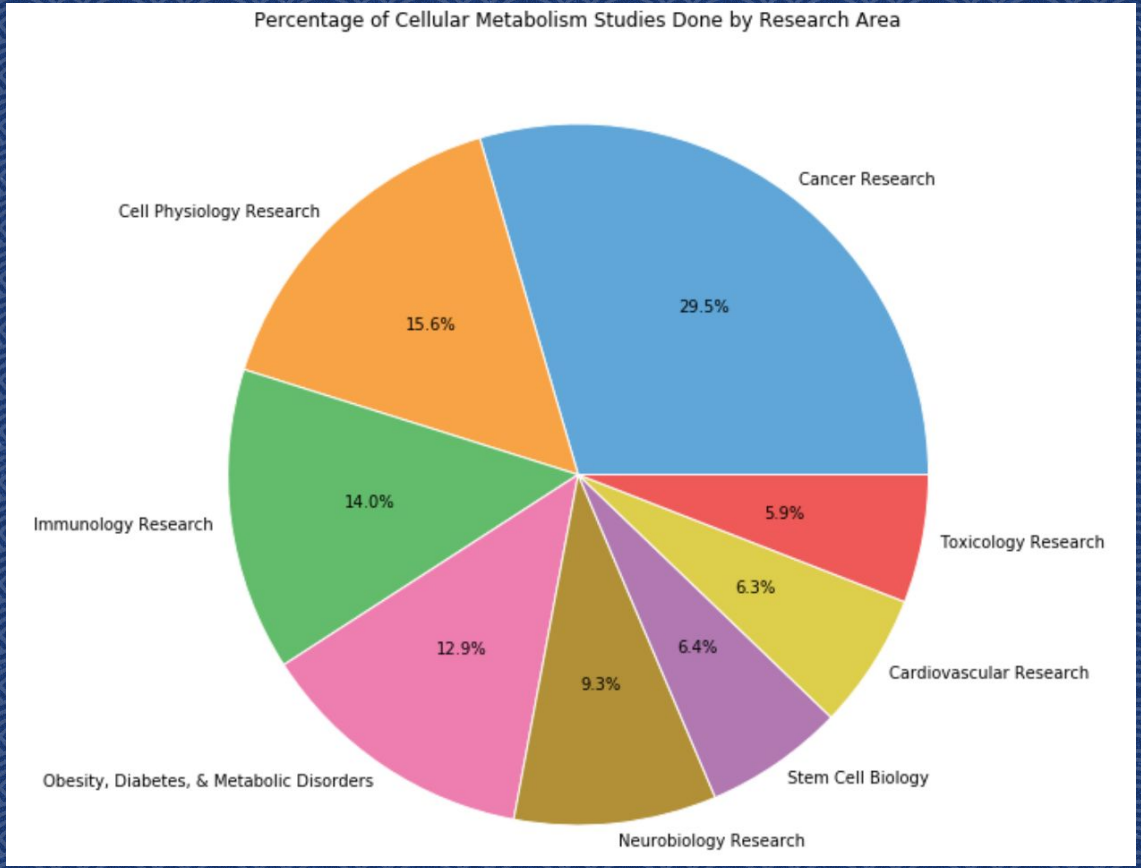

Work done:

- Data Visualization
  - Plot pie chart
  - Add percentage, specify decimal place
- Calculate initial market size from
  - Assays/year (average of last two years)
  - Percentage of cancer & immunology research

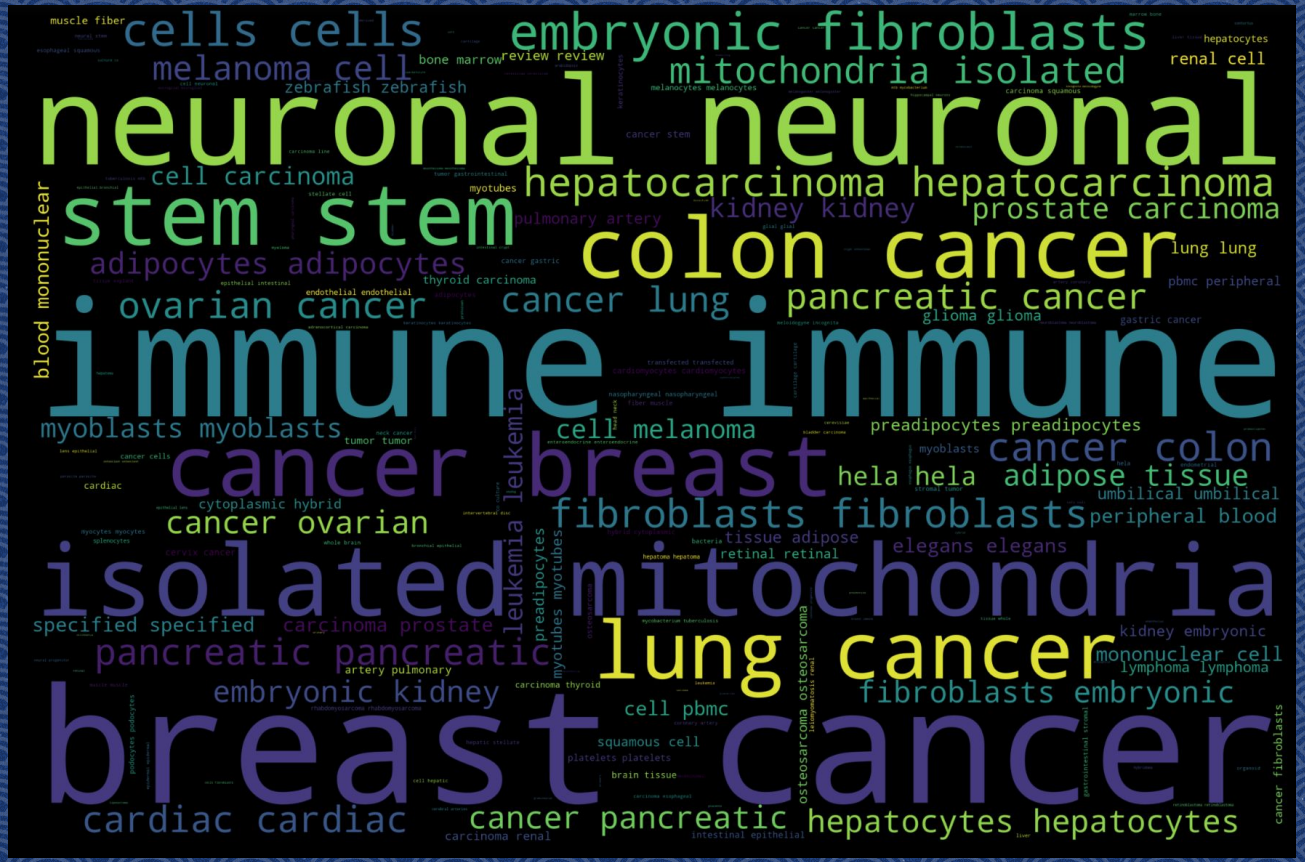
# Q2-a What Are the Most Popular Products with Our Competitor? (Metabolic Assay Types)



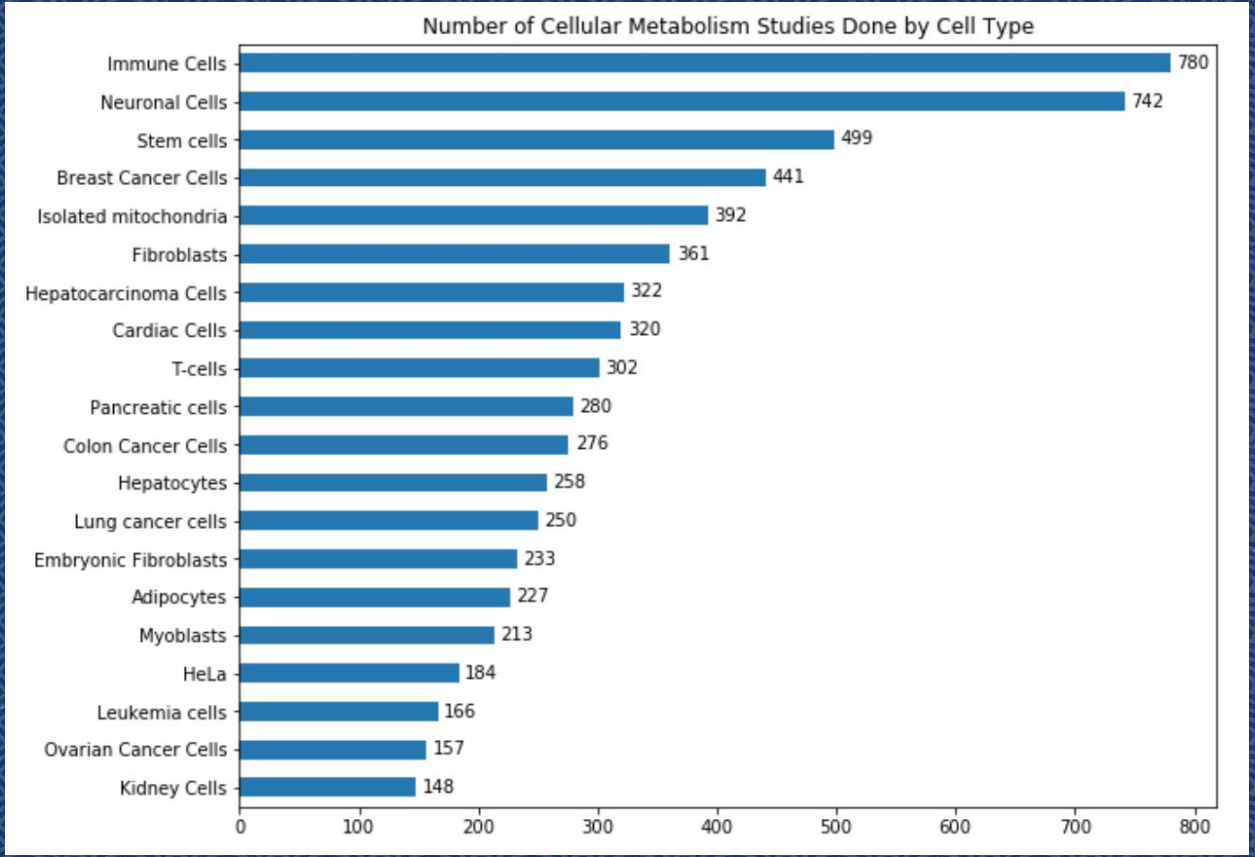
# Q3-a What Are the Most Investigated Research Areas?



# Q3-b What Are the Most Investigated Cell Types?



# Q3-b What Are the Most Investigated Cell Types?

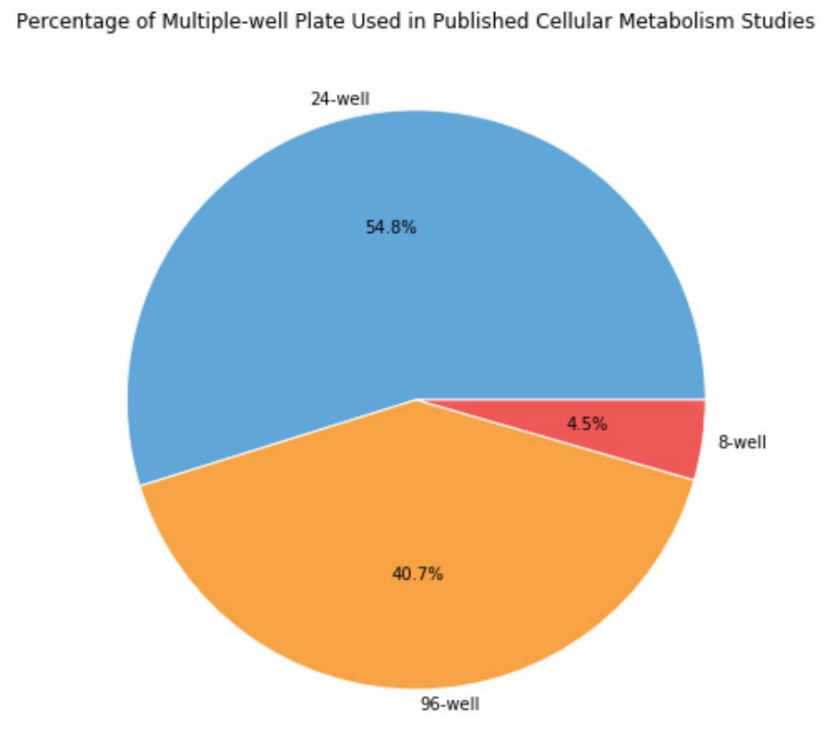


# Q3-c What Are the Initial Disease Models To Focus On?

- In addition to cancer immunotherapy, we have interest to **expand our service portfolio** to other disease models, potentially including **neurodegenerative diseases, cardiovascular diseases, diabetes, obesity,** and so on

- We may be able to figure out which research to look into first through **unsupervised learning**
  - Convert publication date to epoch
  - Merge current dataset with journal impact factor dataset
  - Create a sparse matrix
  - Use affinity propagation method to cluster data & find the instances that best represents each cluster

# Q4-a How Much Time Do We Need To Service Each Customer?




Assumptions: all our customers use multi-well plates

- We already know imaging time/well and set-up time/experiment(plate)
- With info in pie chart, we can estimate time needed to service each customer & each experiment
- Help with cost estimation & pricing strategy

## Overview

- Intro to REDOX
- Motivation for Dataset Creation
- Dataset Composition
- Data Collection Process
- Relevance of Data Analysis to REDOX Sprint
- Exploratory Data Analysis
- **Unsupervised Learning**
  - Merge Datasets
  - Data Preprocessing
  - Next Steps

# Groundwork for Unsupervised Learning
# Merging dataset

| | Rank | Sourceid | Title | Type | Issn | SJR | SJR Best Quartile | H index | Total Docs. (2017) | Total Docs. (3years) | Total Refs. | Total Cites (3years) | Citable Docs. (3years) | Cites / Doc. (2years) | Ref. / Doc. | Country | Publisher |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 18991 | Nature Reviews Genetics | journal | 14710056, 14710064 | 34,896 | Q1 | 307 | 108 | 429 | 7108 | 7296 | 167 | 38,94 | 65,81 | United Kingdom | Nature Publishing Group |
| 1 | 2 | 20315 | Nature Reviews Molecular Cell Biology | journal | 14710072, 14710080 | 32,714 | Q1 | 372 | 112 | 428 | 7278 | 8741 | 206 | 29,64 | 64,98 | United Kingdom | Nature Publishing Group |
| 2 | 3 | 18434 | Cell | journal | 00928674, 10974172 | 25,137 | Q1 | 682 | 547 | 1978 | 27123 | 43114 | 1734 | 23,61 | 49,59 | United States | Cell Press |
| 3 | 4 | 12464 | Nature Reviews Cancer | journal | 1474175X | 23,530 | Q1 | 373 | 118 | 403 | 9157 | 7687 | 198 | 37,90 | 77,60 | United Kingdom | Nature Publishing Group |
| 4 | 5 | 18990 | Nature Genetics | journal | 10614036 | 22,243 | Q1 | 511 | 291 | 811 | 13028 | 16977 | 693 | 22,68 | 44,77 | United Kingdom | Nature Publishing Group |

# Groundwork for Unsupervised Learning
# Merging dataset

| | Title | SJR | H index |
|---|---|---|---|
| 0 | Nature Reviews Genetics | 34,896 | 307 |
| 1 | Nature Reviews Molecular Cell Biology | 32,714 | 372 |
| 2 | Cell | 25,137 | 682 |
| 3 | Nature Reviews Cancer | 23,530 | 373 |
| 4 | Nature Genetics | 22,243 | 511 |

VS

| | Title | Authors | Journal | Publication date | Research area | Cell Line | C |
|---|---|---|---|---|---|---|---|
| 0 | _-1-Antitrypsin (AAT)-modified donor cells sup... | Marcondes AM, Karoopongse E, Lesnikova M, Marg... | Blood | Oct 1 2014 12:00 AM | Immunology Research | T-cells | |
| 1 | _-1-Antitrypsin (AAT)-modified donor cells sup... | Marcondes AM, Karoopongse E, Lesnikova M, Marg... | Blood | Oct 1 2014 12:00 AM | Immunology Research | Natural Killer (NK) cells | |
| 2 | _-enolase regulates the malignant phenotype of... | J. Dai, Q. Zhou, J. Chen, M. L. Rexius-Hall, J... | Nat Commun | Sep 21 2018 12:00 AM | Cell Physiology Research | Pulmonary Artery Smooth Muscle Cells (PASMC) | Pu |
| 3 | _-enolase regulates the malignant phenotype of... | J. Dai, Q. Zhou, J. Chen, M. L. Rexius-Hall, J... | Nat Commun | Sep 21 2018 12:00 AM | Cell Physiology Research | Pulmonary Artery Smooth Muscle Cells (PASMC) | Pu |
| 4 | _-ketoglutarate orchestrates macrophage activa... | P. S. Liu, H. Wang, X. Li, T. Chao, T. Teav, S... | Nat Immunol | Sep 1 2017 12:00 AM | Immunology Research | Bone Marrow-Derived Macrophages (BMDM) | |

# Groundwork for Unsupervised Learning
# Merging dataset

| | Original | Fuzzy |
|---|---|---|
| 845 | Benef Microbes | Benef Microbes |
| 846 | Hepatol Commun | Hepatol Commun |
| 847 | Chem Rev | Chem Rev |
| 848 | Life Science Alliance | Life Science Alliance |
| 849 | Blood Cancer J | Blood Cancer J |
| 850 | BIOspektrum | BIOspektrum |
| 851 | Nature Reviews Genetics | Nature Reviews Genetics |
| 852 | Nature Reviews Molecular Cell Biology | Nature Reviews Molecular Cell Biology |
| 853 | Cell | Cell |
| 854 | Nature Reviews Cancer | Nature Reviews Cancer |

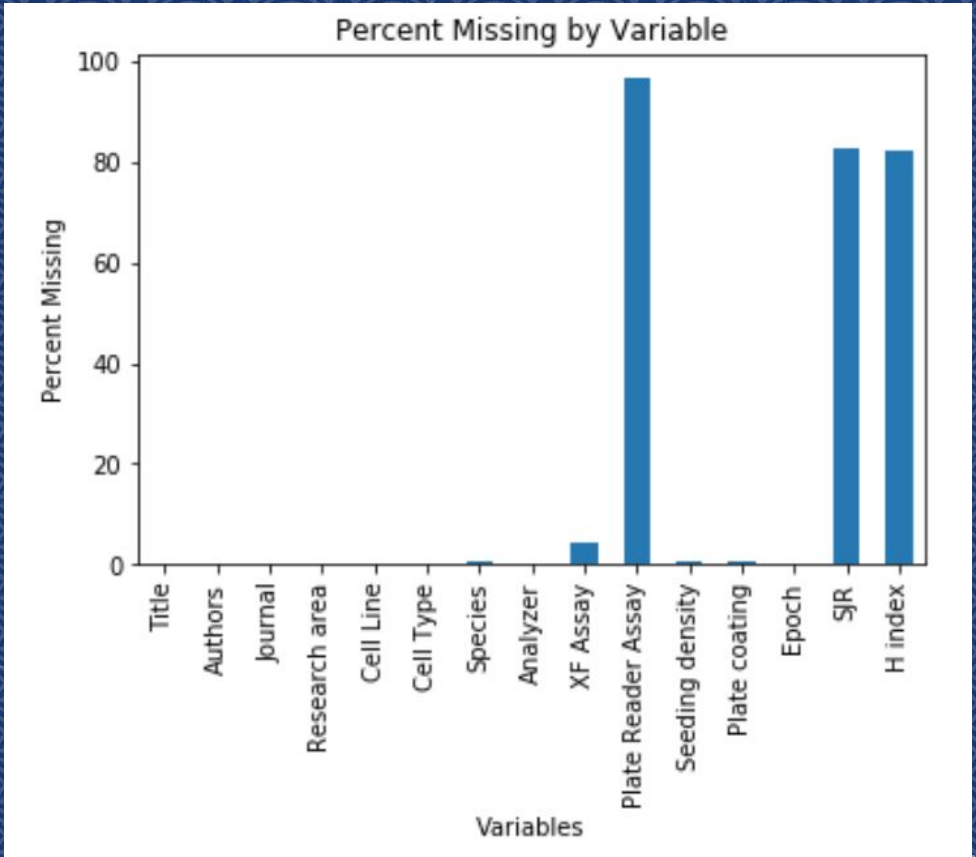
# Groundwork for Unsupervised Learning
# Merging dataset


Fuzzy matching using OpenRefine

# Groundwork for Unsupervised Learning
# Merging dataset

| Authors | Journal | Research area | Cell Line | Cell Type | Species | Analyzer | XF Assay | Plate Reader Assay | Seeding density | Plate coating | Epoch | SJR | H index |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Marcondes AM, Karoopongse E, Lesnikova M, Marg... | Blood | Immunology Research | T-cells | T-cells | Mouse | 24 | Cell Mitochondrial Stress Test | NaN | 8.0x10^5 cells/well | Not Specified | 1412121600000000000 | 6,434 | 410.0 |
| Marcondes AM, Karoopongse E, Lesnikova M, Marg... | Blood | Immunology Research | Natural Killer (NK) cells | Immune Cells | Mouse | 24 | Cell Mitochondrial Stress Test | NaN | 8.0x10^5 cells/well | Not Specified | 1412121600000000000 | 6,434 | 410.0 |
| J. Dai, Q. Zhou, J. Chen, M. L. Rexius-Hall, J... | Nat Commun | Cell Physiology Research | Pulmonary Artery Smooth Muscle Cells (PASMC) | Pulmonary Artery Cells | Human | 24 | Cell Mitochondrial Stress Test | NaN | 3.0 x10^4 cells/well | Not Specified | 1537488000000000000 | NaN | NaN |
| J. Dai, Q. Zhou, J. Chen, M. L. Rexius-Hall, J... | Nat Commun | Cell Physiology Research | Pulmonary Artery Smooth Muscle Cells (PASMC) | Pulmonary Artery Cells | Human | 24 | Glycolysis Stress Test | NaN | 3.0 x10^4 cells/well | Not Specified | 1537488000000000000 | NaN | NaN |
| P. S. Liu, H. Wang, X. Li, T. Chao, T. Teav, S... | Nat Immunol | Immunology Research | Bone Marrow-Derived Macrophages (BMDM) | Immune Cells | Mouse | 96 | Cell Mitochondrial Stress Test | NaN | 1.0 x10^5 cells/well | Not Specified | 1504224000000000000 | NaN | NaN |

# Groundwork for Unsupervised Learning
# Merging dataset



# Groundwork for Unsupervised Learning
# Data Preprocessing

```
In [351]: df['Seeding density'].iloc[410:429]

Out[351]: 410            Not Specified
          411            Not Specified
          412            Not Specified
          413            Not Specified
          414            Not Specified
          415            Not Specified
          416    2.0 x10^4 cells/well
          417    0.15x10^6 cells/well
          418     1.0x10^4 cells/well
          419     1.0x10^4 cells/well
          420            Not Specified
          421            Not Specified
          422               50 ug/well
          423               50 ug/well
          424    8.0 x10^4 cells/well
          425    8.0 x10^4 cells/well
          426            Not Specified
          427     8.0x10^2 cells/well
          428    4.0 x10^4 cells/well
          Name: Seeding density, dtype: object
```

```
In [352]: # split on the space
          # temp = df['Seeding density'][1].split(' ')[-1]
          df['Seeding density'] = df['Seeding density'].astype(str)
          df['Seeding density unit'] = df['Seeding density'].apply(lambda x: x.split(' ')[-1])
          df['Seeding density unit'].value_counts()[:10]
          # len(df['Seeding density unit'].unique())
          # take the last one
          # see how many unique
          # extract the numbers by spliting on x and ^

Out[352]: cells/well      5215
          Specified       3774
          ug/well          177
          nan               55
          worms/well        50
          islets/well       40
          protein/well      32
          cells/mL          21
          well              19
          cell/well         16
          Name: Seeding density unit, dtype: int64
```

# Groundwork for Unsupervised Learning
# Data Preprocessing

```
In [353]: fuzz.partial_ratio('cells/well', 'cells/welll')

Out[353]: 100

In [354]: fuzz.token_set_ratio('cells/well', 'cell/swell')

Out[354]: 90
```

```
In [356]: choices = df['Seeding density unit'].unique()
          possibilities = process.extract('cells/well', choices, limit=10, scorer=fuzz.token_sort_ratio)
          # Let's see everyting with a score above 88
          # Reference: http://jonathansoma.com/lede/algorithms-2017/classes/fuzziness-matplotlib/fuzzing-
          [possible for possible in possibilities if possible[1] > 88]

Out[356]: [('cells/well', 100),
           ('cell/well', 95),
           ('cells/welll', 95),
           ('cells/wll', 95),
           ('cell/swell', 90)]

In [357]: choices = df['Seeding density unit'].unique()
          possibilities = process.extract('Specified', choices, limit=10, scorer=fuzz.ratio)
          [possible for possible in possibilities if possible[1] > 88]

Out[357]: [('Specified', 100),
           ('Speccified', 95),
           ('Specifed', 94),
           ('Sepcified', 89),
           ('Speciifed', 89)]
```

Fuzzy matching using fuzzy wuzzy

# Groundwork for Unsupervised Learning
# Data Preprocessing

```
In [358]: # df.apply(lambda row: fuzz.token_sort_ratio(row['Seeding density unit'], 'cells/well'), axis=1) > 88
          df.loc[df.apply(lambda row: fuzz.token_sort_ratio(row['Seeding density unit'], 'cells/well'), axis=1) > 88,\
                 'Seeding density unit'] = 'cells/well'

          # Check that fuzzy match worked
          len(df['Seeding density unit'].unique())

Out[358]: 90

In [359]: df.loc[df.apply(lambda row: fuzz.token_sort_ratio(row['Seeding density unit'], 'Specified'), axis=1) > 88,\
                 'Seeding density unit'] = 'Specified'

          # Check that fuzzy match worked
          len(df['Seeding density unit'].unique())

Out[359]: 86

In [360]: df.loc[df.apply(lambda row: fuzz.token_sort_ratio(row['Seeding density unit'], 'ug/well'), axis=1) > 88,\
                 'Seeding density unit'] = 'ug/well'

          # Check that fuzzy match worked
          len(df['Seeding density unit'].unique())

Out[360]: 85
```

# Groundwork for Unsupervised Learning
# Data Preprocessing

In [361]:
```
cells_well = df.apply(lambda row: fuzz.token_sort_ratio(row['Seeding density unit'], 'cells/well'), axis=1) > 88
cells_well_dict = cells_well.value_counts().to_dict()
print('"cells/well" counts: ', cells_well_dict.get(True))

not_specified = df.apply(lambda row: fuzz.token_sort_ratio(row['Seeding density unit'], 'Specified'), axis=1) > 88
not_specified_dict = not_specified.value_counts().to_dict()
print('"Not Specified" counts: ', not_specified_dict.get(True))

ug_well = df.apply(lambda row: fuzz.token_sort_ratio(row['Seeding density unit'], 'ug/well'), axis=1) > 88
ug_well_dict = ug_well.value_counts().to_dict()
print('"ug/well" counts: ', ug_well_dict.get(True))

x = (cells_well_dict.get(True)+not_specified_dict.get(True)+ug_well_dict.get(True))/len(df)
print('The top three categories of units, as shown above, comprises of ',\
      "{:.1%}".format(x), ' of the data')
```

```
"cells/well" counts:  5239
"Not Specified" counts:  3780
"ug/well" counts:  178
The top three categories of units, as shown above, comprises of  95.5%  of the data
```

# Groundwork for Unsupervised Learning
# Data Preprocessing

In [363]:
```
cells_well_indices = np.where(df['Seeding density unit']=='cells/well')[0].tolist()
not_specified_indices = np.where(df['Seeding density unit']=='Specified')[0].tolist()
ug_well_indices = np.where(df['Seeding density unit']=='ug/well')[0].tolist()

l = cells_well_indices + not_specified_indices + ug_well_indices

def complement(l, universe=None):
    if universe is not None:
        universe = set(universe)
    else:
        universe = set(range(min(l), max(l)+1))
    return sorted(universe - set(l))

list_to_clean = complement(l)
len(list_to_clean)
```

Out[363]:
```
437
```

# Thank you!
# Any questions?