# 数据预处理

小组选用的是Nature上的数据集“A dataset of publication records for Nobel laureates”诺贝尔奖得主发表记录数据集（1900 - 2016），数据包括1900年至2016年间545位诺贝尔奖（物理学、化学、医学和生理学）得主的出版记录（包括全部发表记录、与获得诺贝尔奖相关的发表记录）。

本次选择使用Python语言在Jupyter Notebook上进行数据预处理，主要通过pandas、numpy、copy和affiliation\_parser，并在分别导入了Chemistry publication record.csv、Medicine publication record.csv、Physics publication record.csv和Prize-winning paper record.csv的基础上进行数据规模探索、过滤缺失值、去重处理。

## Graph-1

为了便于调用，将列名之间的空格替换为“\_”，如“Laureate ID”改为“Laureate\_ID”。查看是否存在同一篇文献发表在多个平台上的情况，若有则人工进行筛选去除。

要成为诺贝尔奖得主，发表的文献数量是否需要达到特定的阈值？为了实现这个目标，需要统计各学科诺贝尔奖得主在获奖前的累计发文量。通过定义outpu\_csv函数来进行累计统计并分别将各学科结果输出到不同的csv文件中，便于后续可视化读取。

## Graph-2

前十机构筛选：根据不同领域的诺奖得主在不同的三个数据集上新增一个列subject，根据领域标上chemistry, medicine或physics,然后将三个领域的数据集集合呈一个数据集full\_data。对full\_data中的Laureate ID和Affiliation作groupby操作并利用Counter计算各机构出现的次数，最终获得拥有诺奖得主数量最多的前十个机构。这边的数据上，默认作者只要在机构出版过文献就视为有效计数，不严格以作者最后出现的地方作为计数标准。

前十机构诺奖分布数据：首先根据前十机构在full\_data中筛选出这十个机构中的作者，由于该份数据中一个作者只会在一个领域出版，不涉及多领域出版，因此可以简单靠对根据Laureate ID做去重操作后的数据作用pandas自带的函数pivot\_table计算每个Affiliation在不同subject的作者（Laureate ID）的作者数量。再对数据根据机构排名进行排列后将他们以嵌套数组的形式输出作为图的输入数据。

## Graph-3

在累计发文量的基础上，学者所在的国家受学术领域的发展、资源、文化背景等是否对发表文献量和质量产生影响？这种影响是否体现在诺贝尔得主的国家分布上？因此，利用from affiliation\_parser import match\_affil将Prize-winning paper record.csv中的发文机构数据生成所在的国家数据，并与Graph-1数据处理中各学科outpu\_csv函数的结果进行合并。同时，为了体现累计发文量的群体分布将累计发文数以“> 180”、“91 - 180”、“46 - 90”、“16 - 45”和“1 - 15”分组。

在可视化的数据准备方面，需要体现每年各国累计发文量分组三个学科的获奖人数，因此按此顺序生成嵌套列表；在学科分布上，同样输出以三个学科获奖人数及总获奖人数的嵌套数组；在体现各国诺贝尔得主学科分布上，输出每年各国三个学科获奖人数的嵌套列表。

## Graph-6

所使用的数据为Prize-winning paper, 将获奖时间（prize year）和发表时间（pub year）相减获得获奖间隔时间(prize\_pub\_year),然后观察prize\_pub\_year字段的分布情况，发现负数的异常值，不合理 ，因此对prize\_pub\_year作顺序排列，找到三个间隔时间为负数的数据，经过调查发现是数据中的发表时间有误，结合网上资料对这些异常值作修正。最后根据prize\_pub\_year作groupby操作计算文献数量

对full\_data中的Laureate ID和Affiliation作groupby操作并利用Counter计算各机构出现的次数观察获奖文献与其发表时间之间的关系。

# 数据可视化

## Graph-1

为了更好呈现诺贝尔奖得主在获奖前所出版的文献积累，同时探究各学科获奖学者累计发文量的分布模式，选择以散点图进行可视化，并以三个颜色分别表示不同学科、散点大小由小到大来表示累计发文量由少增多、该学科累计发文量最高的学者以红色标记直观地显示。此外，各学科散点图辅以平均累计发文量的虚线，能更容易了解各学科普遍的累计发文量与观察偏离值。

## Graph-2

为了观察具有实力的前十个Affiliation再不同学科领域的研究实力，选择采用弦图对数据进行可视化，对三个学科赋予不同的颜色，弦图的颜色也是从学科出发到达不同的Affiliation，因此整个图看起会比较干净利落。此外在弦图的排列上，我们让右边圆显示学科，左半圆显示Affiliation，左半圆从上往下的顺序是按照Affiliation的诺奖得主数量进行逆序排列的，拥有最多诺奖得主的Affiliation会在左半图的最上方。在内弦的排列上，将弦按从高到低的顺序在弧内进行排序，并且在弦相交时，会显示在顶部最小的弦。在外弧线的外面，把诺奖得主的数量用刻度的形式展现出来。此外还加入了交互的功能，当鼠标指向外弧线、刻度或标签名字时，相应类别下的弧线的透明度保持不变，其他未被选中的类别透明度则会被调低，突出显示所指类别在数据上的分布情况。

本图还有另外两个交互功能，在看到弦图之前，用户会先看到被弦图的外弧线所包围的对图的描述，需要点击图内的SHOW按钮后才会进行最后的可视化，把完整的弦图展示出来，点击后SHOW按钮会被隐藏，要想从完整的弦图回到介绍状态的弦图，可以通过下方的RESET按钮实现。

## Graph-3

为了呈现各国诺贝尔得主在国家、累计发文量以及学科的分布试图分别以不同的可视化图标来呈现，由堆积柱状图、环形图和散点图组合而成的四层嵌套图。同时为了呈现各国获奖者逐年增长的分布与趋势，选择以动态、自动播放的方式依次显示各年的可视化结果。在用户体验与人机交互方面，提供点击选择查看特定年份结果的年份按钮，也设置了继续播放和暂停播放按钮，以提供更好的交互功能。

在具体的结果可视化上，图中心的环形图主要用于显示学科获奖比例，三个颜色代表不同学科，灰色则用于提示用户在1907年至2016年数据呈现的进度。此外，以同一颜色由浅至深的五个环状来表示五个累计发文由低到高的分组，同时以直线为每个国家划分了散点表示区域，这层外面依次为国家和年份的文字显示。

## Graph-6

为了直观地显示获奖文献与其发表时间之间的关系，选择利用柱状图对数据做可视化，y轴表示文献数量，x轴表示间隔时间。在图的交互上，鼠标指向柱子会显示柱子所代表的文献数量和间隔时间的数值。另外，也做了排序功能，可以点击Sort Descending和Sort Ascending按钮来对柱子根据文献数量进行排序，同时也支持用Reset按钮恢复默认设置。

# 结果解读

## Graph-1

从三个学科的平均累计发文量来看，化学领域获奖学者们的累计发文量几乎是物理学领域以及医学和生理学领域的一倍，在散点的分布上亦体现了化学领域学者们普遍发文量较高，散点在1970年以后多分布在累计发文量较高的区间。物理学领域以及医学和生理学领域累计发文量的分布大体类似，值得注意的是物理学领域只有Omura, S学者的累计发文量远远超过该领域的其他学者，在49年内发表了949篇文献，但比起化学领域Olah,G在38年内出版了1213篇文献、Stoddart, J在46年内出版了1103篇文献，还是远远不及。医学和生理学领域的最高累计发文量为Akasaki, Isamu的39年内出版了612篇，与Amona, Hiroshi在29年内605篇以及Nakamura, Shuji在27年内出版了549篇，三位日本学者在该领域的发文量远远超出其他学者。由此，物理学领域以及医学和生理学领域的可视化结果反映出在这两个领域发文量多寡并非决定获得诺贝尔奖的主要因素，更多的是研究结果的质量。宏观而言，随着时代的进步，诺贝尔得主的累计发文量呈增长态势，从侧面反映出要获得诺贝尔奖需要用有一定的学术研究成果的积累，在化学领域这种趋势尤为明显。

## Graph-2

在这10间拥有最多诺奖得主的学校/机构中，学校共占了7间，机构占了3间，现在我们来观察它们在不同领域的获奖情况。可以看到在数据中排名前三的是哈佛大学、剑桥大学和斯坦福大学，其中哈佛大学的诺奖得主数量明显比其他学校多，跟排第二的剑桥大学比多了40%，可见其在2016年作为QS排名第一的大学是实至名归的。数据中比较有趣的现象是机构/大学在研究方向上的偏科现象。数据中Affiliation在不同学科的研究成果是不平均的，医学和生理学是三个学科中最特殊的学科，有关这个学科的研究成果更常出现在专业化程度更高的Affiliation中，又或者说相较于其他学科医学和生物学这一领域是更为专业化的，有专门的机构/大学去专研这一方向。例如Rockefeller University和National Institutes of Health的研究重心就是明显往医学和生物学倾斜的，虽然在综合实力上他们并不是最优的，但是在医学和生物学这一领域上他们是最优的，获奖作者数量远远高出其他诺奖数量更多的大学。由此可见当我们在选择大学时应结合自身兴趣和方向做选择，排名越高的大学不必然是最优解。

## Graph-3

美国是最早拥有诺贝尔获奖者的国家，从物理学开始到医学和生理学，再到化学领域，一直到1928年仍为包揽三个学科诺贝尔奖的国家，并以物理学诺贝尔奖数量最多，同时在获奖学者的累计发文量中仍主要集中在16至45篇。1929年，加拿大的一位学者在医学和生理学领域获奖，其累计发文量同样集中在16至45篇；1933年爱尔兰化学领域诺贝尔获奖者、1934年英国学者获得医学和生理学领域的诺贝尔奖同样如此，一直到1938年同样由美国学者开始累计发文量达180篇以上。

整体而言，三个学科中物理学领域的诺贝尔奖增长率在早期呈明显优势，医学和生理学领域的获奖数后来居上。同时，美国获奖学者的累计发文量明显超过其他国家的学者，其中化学领域的累计发文量由少向多递增，医学和生理学以及物理学领域的则呈由少向多递减，和Graph-1呈现的结果，即在化学领域中获得诺贝尔奖与学术积累一定程度上呈正相关关系。此外，加拿大和英国的诺贝尔奖得主相较于其他国家较多，前者在物理学领域的成果倍受认可，后者则在医学和生理学领域有出色的成果。由此可见，美国优质的学术环境与氛围、丰富的资源等为物理学、化学以及医学和生理学学者开展科研提供了基础；加拿大、英国和日本学者在这三个学科领域的获奖数也反映了该国对研究前沿、学科发展的重视。

## Graph-6

可以看到获得诺奖的文献不一定都是当年最新的文献，很多文献都等了10以上才获奖，最常见的获奖间隔时间分别是9、11、16年，大过22年才获奖的文献数量是较少的，发表后就马上获奖的文献从数据上来看也是较少的，可见文献发表后没有马上获得热烈的反响并得奖是较为正常的事情，相反的很多文献可能会因为所发表的知识过于超前或未被重视而被埋没，静待未来当时机来临时才能够体现出文献内容的价值，这种“早熟性的科学发现”或“延迟承认”的文献也被称为睡美人文献。