



数据科学与绿色文明

爱护地球 保护我们的家园



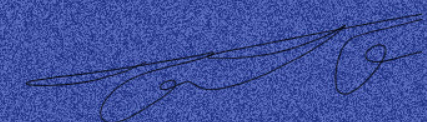
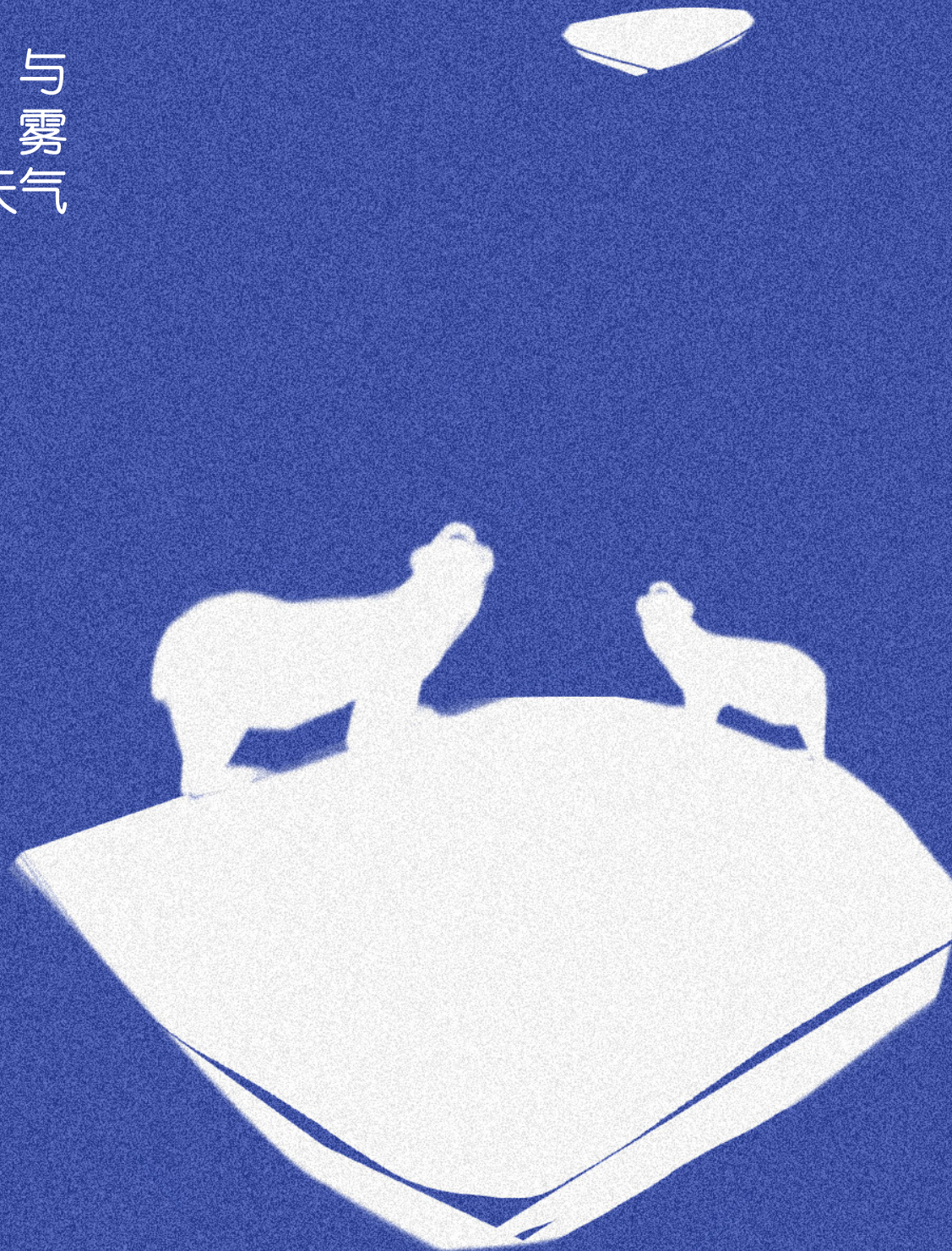
中国大学生社会实践知行促进计划
China Education Support Project



“全球变暖”这一现象指的是全球平均气温逐渐升高。

全球变暖的现象首当其冲的是南北极冰盖和北极圈永久冻土融化，北极熊找不到冰块停歇而溺亡。

由于陆地温室气体排放造成大陆气温升高，与海洋温差变小，近而造成了空气流动减慢，雾霾无法短时间被吹散，造成很多城市雾霾天气增多，影响人类健康。

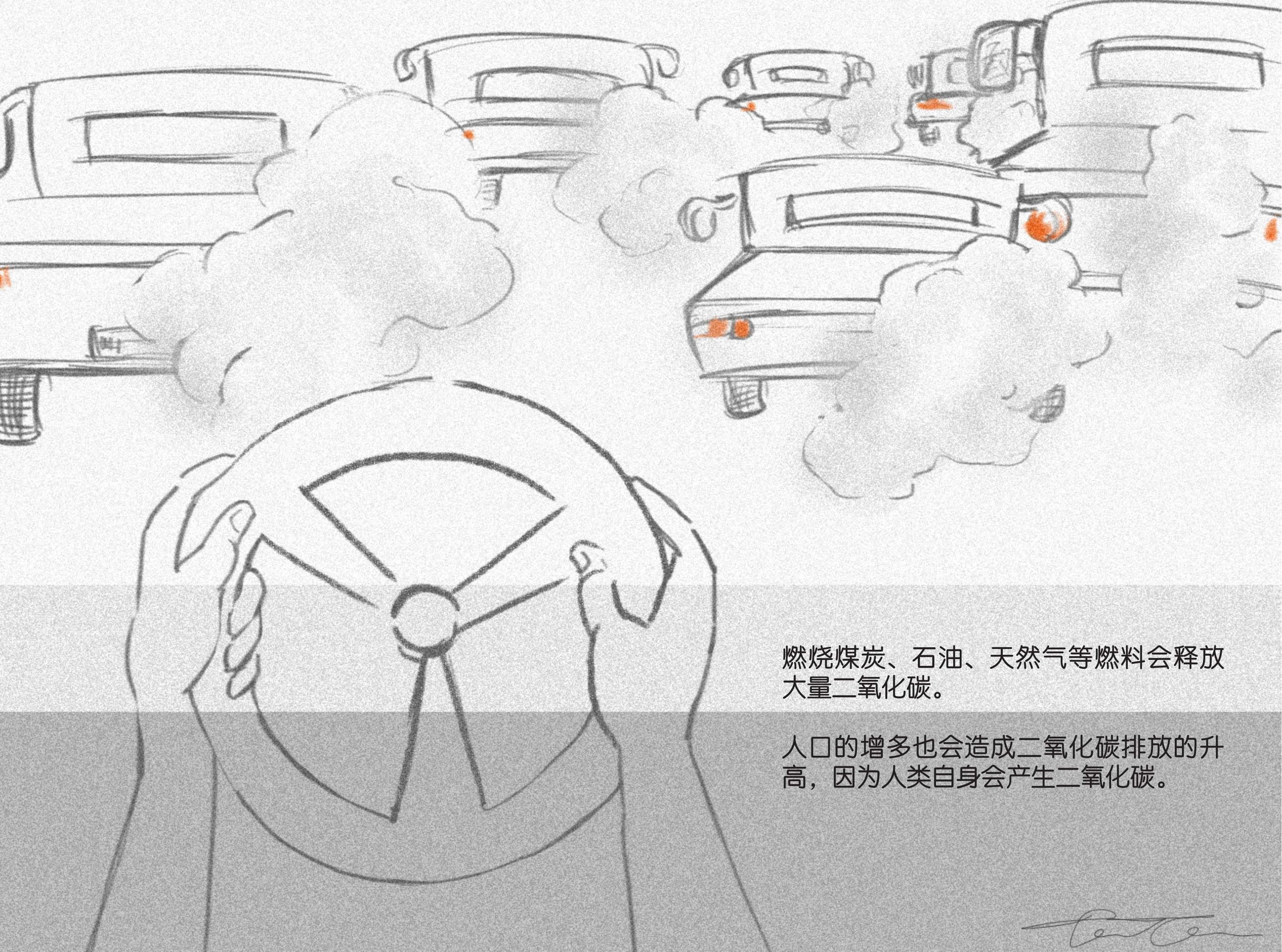




温室就像是用玻璃来存住热量，给植物保温。

温室效应是指大气中的吸热气体像温室的玻璃一样，防止热量散失，使地球变暖。

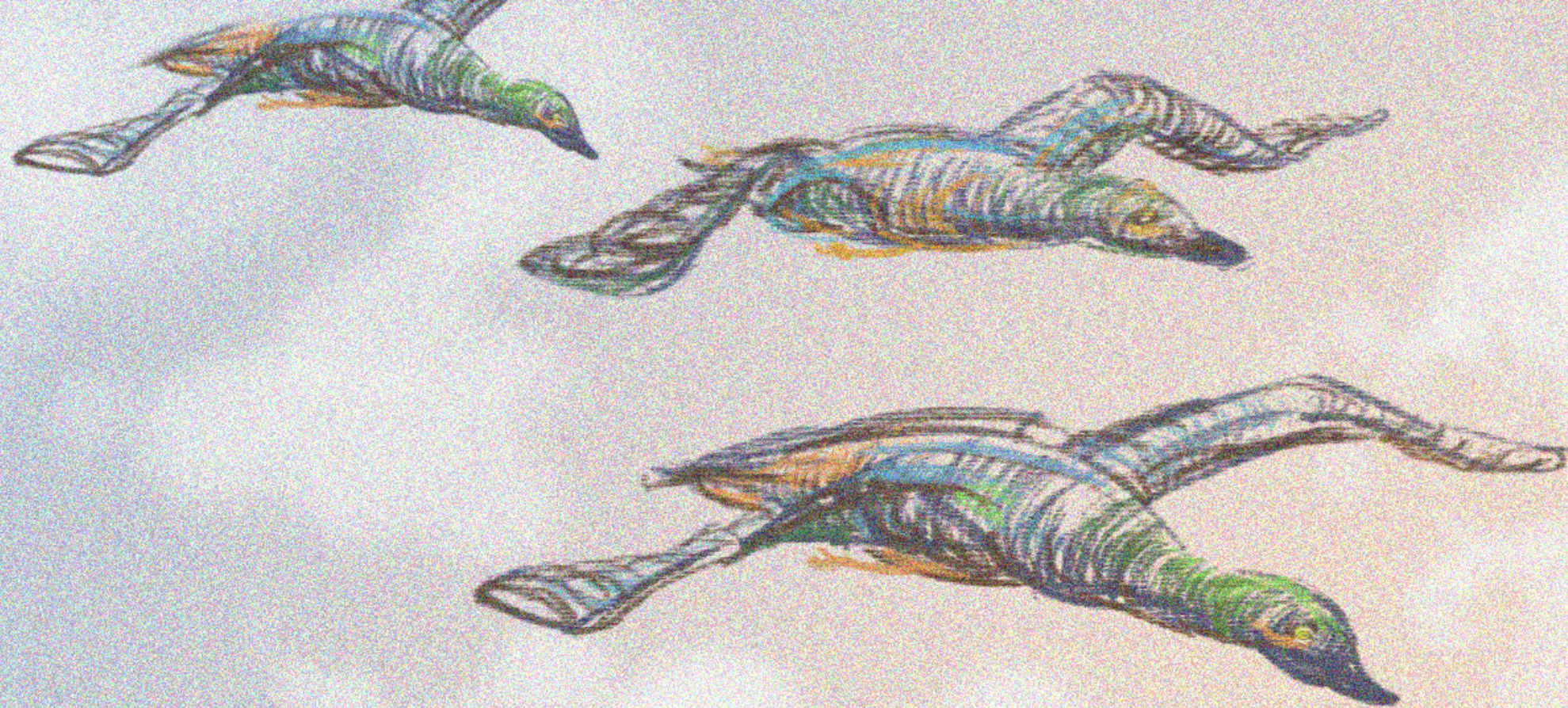
温室效应产生的主要原因是大气中的二氧化碳含量逐年增加。



燃烧煤炭、石油、天然气等燃料会释放大量二氧化碳。

人口的增多也会造成二氧化碳排放的升高，因为人类自身会产生二氧化碳。

Cartoon



怎样才能降低二氧化碳排放，从而减缓全球变暖呢？

我们可不可以分析出导致二氧化碳排放量升高的主要因素，然后采取行动呢？

看看数据科学如何帮助我们！

数据科学是从数据中提取有用知识的一系列技能和技术。



1. 数据收集与整理

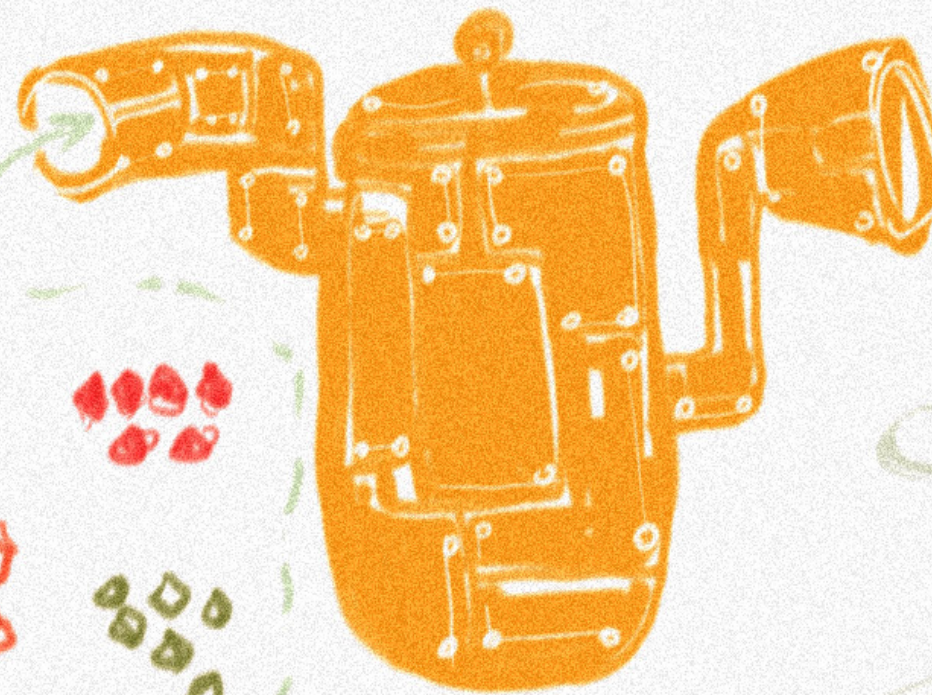


2. 数据分析



3. 数据清洗与预处理

SUPER!



4. 模型训练、调优与评估



5. 模型预测



数据科学 - 数据收集与整理

工人们上报所在工厂每天使用的煤、石油、天然气等的使用量。
工程师们统计工人们上报的数据，并按年汇总做成表格。



历史二氧化碳排放数据

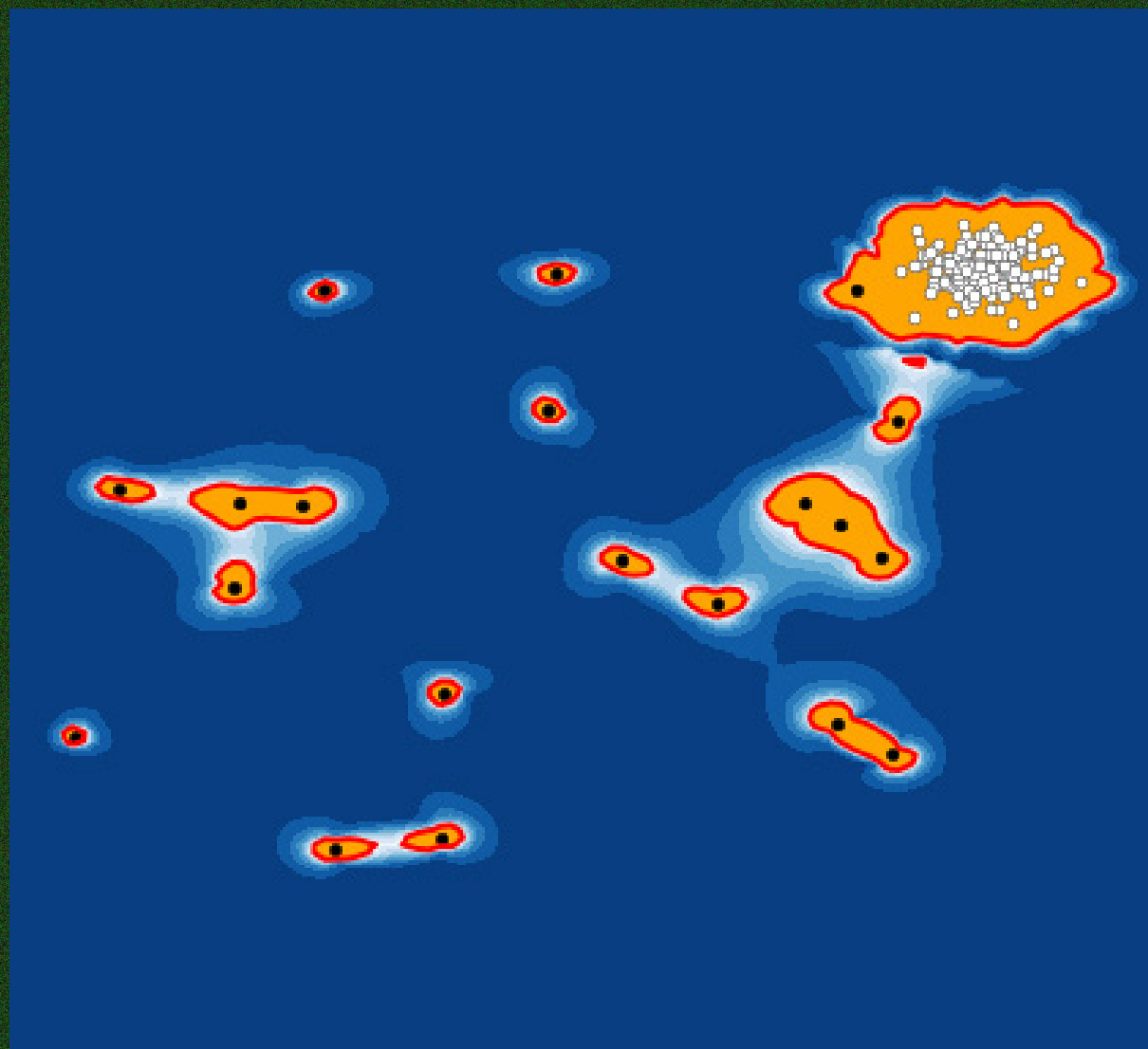
年份	二氧化碳排放 总量	固体燃料	液体燃料	气体燃料	水泥生产	天然气燃烧
1949	1419	960	362	81	16	0
1950	1630	1070	423	97	18	23
1951	1767	1129	479	115	20	-24
1952	1795	1119	504	124	22	26
1953	1841	1125	533	131	24	27
1954	1865	1116	557	138	27	27
1955	2042	1208	625	150	30	31
1956	2177	1273	679	161	32	32
1957	2270	1309	714	178	34	35
1958	2330	1336	731	192	36	35
1959	2454	1382	789	206	40	36
1960	2569	1410	849	227	43	39
1961	2580	1349	904	240	45	42
1962	2686	1351	980	263	49	44
1963	2833	1396	1052	286	51	47
1964	2995	1435	1137	316	57	51
1965	3130	1460	1219	337	59	55

数据科学 - 数据分析

数据分析，是从现有收集到的数据中，获取有意义的信息和结论。

在二氧化碳的数据中，我们可以发现，天然气燃烧值有一个负值，在数据科学中被认为是异常点，需要进行人工修正。
异常值是指样本中的个别值，其数值明显不同于其他观测值，也叫离群点。

液体燃料、气体燃料、水泥生产三个数据指标对二氧化碳排放量的影响雷同，这三者之间的相关性高达 90%，所以在后续的数据科学步骤中我们以液体燃料作为代表。相关性指的是两个变量之间的关联程度，就好像天气越热我们吃的冰淇淋越多。



数据科学 – 数据清洗与预处理

在实例中，天然气燃烧数据“-24”这个值，与其他数字有明显区别，会影响到我们结果的产生，在数据科学中，我们称之为脏数据，需要进行清洗。

数据预处理包括一系列的不同的方法或操作，数据科学家使用这些方法来将原始数据处理成更方便他们进行分析的形式。



历史二氧化碳排放数据预处理结果

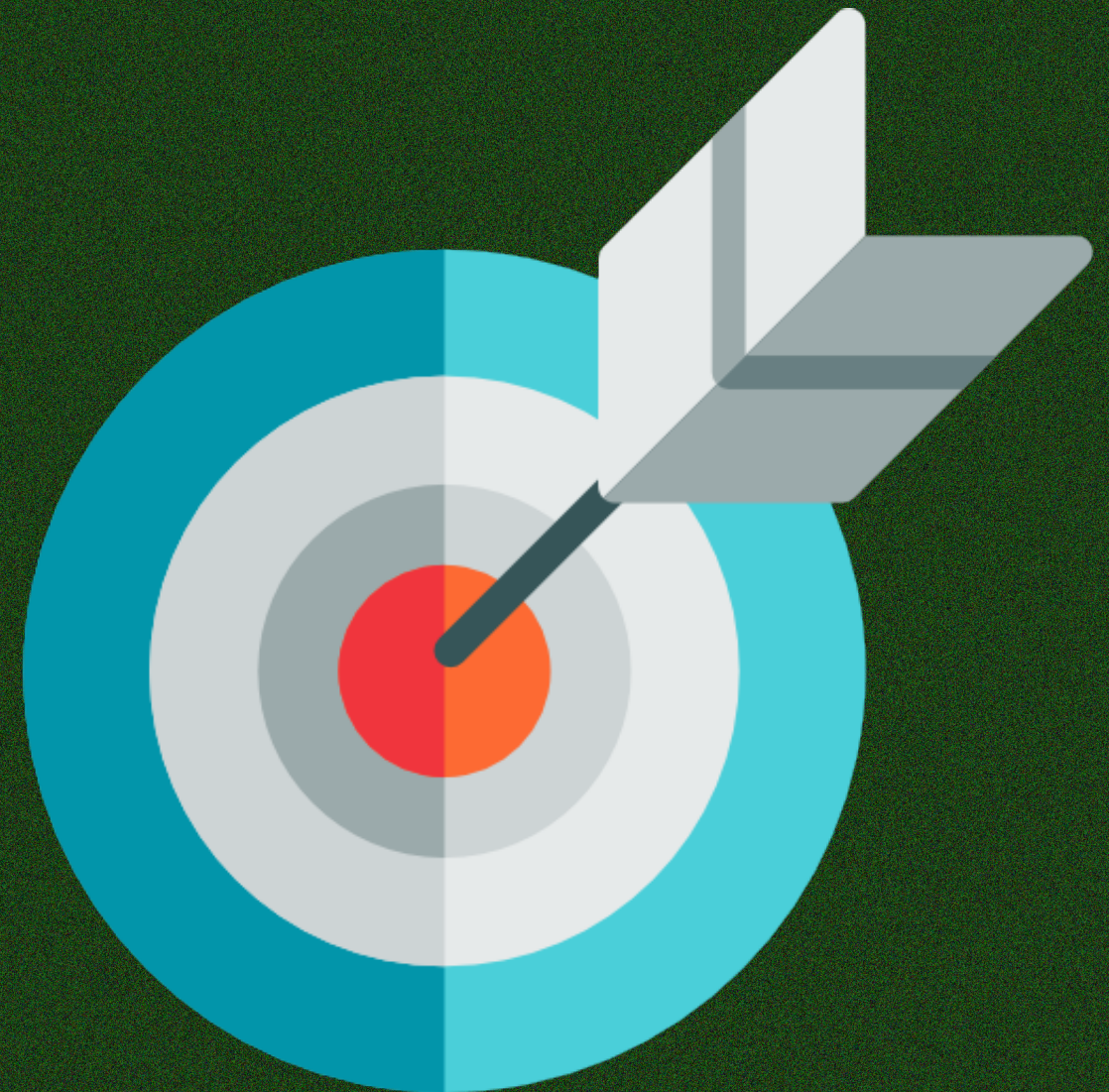
年份	二氧化碳排放总量	前一年二氧化碳排放总量	二氧化碳排放总量升高或降低 (0: 降低; 1: 升高)
1979	5357		
1980	5301	5357	0
1981	5138	5301	0
1982	5094	5138	0
1983	5075	5094	0
1984	5258	5075	1
1985	5417	5258	1
1986	5583	5417	1
1987	5725	5583	1
1988	5936	5725	1
1989	6066	5936	1
1990	6074	6066	1

数据科学 - 模型训练、调优与评估

模型是使用数学方法做的一个黑盒子，就好像一台超级会做饭的机器。

我们提供给模型大量的数据，并通过这台机器的调节按钮对模型进行调优，使模型有更好的表现。这就是模型训练和调优。

我们会提供给训练出的模型新的数据，并通过对比模型输出的结果和真实数据的结果，来检查和评估模型的表现。



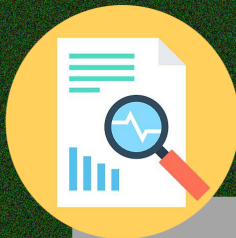
数据科学 - 模型预测

通过前面学习到的数据收集和数据分析步骤，我们可以获得有效的 2020 年固体燃料、液体燃料、气体燃料、水泥生产、天然气燃烧排放指标。

数据科学模型可以预测 2020 年碳排放情况么？
答案是肯定的！

经过模型训练和调优，模型预测结果告诉我们，如果我们可以减少固体燃料、液体燃料的使用或减少天然气燃烧，二氧化碳的排放量会有所降低。

化石资源，是一种碳氢化合物或其衍生物，由古代生物的化石沉积而来，是一次性能源。
固体燃料，如煤炭，木材。
液体燃料，如汽油，煤油。



二氧化碳排放模型预测结果

	固体燃料	液体燃料	气体燃料	水泥生产	天然气燃烧	模型预测二氧化碳排放
正常使用化石资源	4991	3087	1809	564	66	升高 
减少使用化石资源	2991	2087	1809	564	60	降低 



在日常生活中，
如何让二氧化碳排放
降低呢？

节约不可再生资源的
使用！

增加清洁能源的使用，
比如太阳能、风能等！

多种树！

我们还能做些什么
呢？