# Feature Spec

v-yiminfan@microsoft.com

1. 通过数据源推荐视频片段
   1. 动机：一般来说，用户提供的数据是表格化的。用户往往希望通过使用不同数据图表从不同角度对数据进行分析和可视化。在选择使用的数据片段时，在DataClip的界面中，用户需要指明自己所要可视化的数据列和使用的图表。我们希望达到的效果是，用户选定自己想可视化的数据列，然后算法为用户自动推荐几个适合用来可视化有关内容的图表类型。
   2. 输入：用户指定的数据列
   3. 输出：几个（3-5个）推荐的图表类型
   4. 预期效果：用户选取的数据中有涉及到年份、时间的，一般认为选折线图、条形图的可能性较大（强调沿时间轴的对比）；用户选取的数据涉及到一组实体所占的比例，一般认为采取饼图的比例较大；用户选取的数据是涉及到不同实体数量的值（不考虑比较关系）可以采用条形图；用户选取的数据涉及到国家的，可以推荐地图相关的图表视频。涉及到人之间的对比的，可以采取和人有关的对比图……
   5. 技术实现：在没有数据的情况下首先考虑采用规则驱动的方法。首先，需要对给定数据的类型进行识别，识别数据是何种类型（时间数据、离散类别数据、浮点数据、比例数据），有没有涉及到特殊实体（国家、地区、性别）。随后，设计一组规则，将选取的数据类型映射到图表类型即可。在经过一段时间的使用后有一定的数据积累后，采用机器学习的方法解决，可以训练一个模型，以选取的数据类型构成的序列作为输入并进行embedding，以选取的图表类型作为输出，实现自动化的推荐。
2. 数据驱动的自动化统计分析
   1. 动机：一般来说，用户提供的数据都是原始数据，而许多图表可能需要经过处理的数据（简单举例：数据的最大值、最小值、均值、方差、众数、比例，按月的原始数据按年呈现等等），我们可以在软件中集成简单数据处理的功能，比如计算简单的均值、方差、比例，节省用户手动处理操作的时间，可以将简单的数据处理功能直接在用户输入数据后进行展示
   2. 输入：用户输入的原始数据
   3. 输出：数据的部分统计量和数据的简单可视化
   4. 预期效果：用户想制作的视频中涉及到将原始以月为单位的数据转换为年呈现，可以采取简单的设定进行实现；用户的原始数据中，有男性和女性（或中国和美国）的对比，可以直接生成简单的统计量（习惯比例，人数比例等等）。尽管是简单的数据分析，但可以帮助用户快速挖掘insights。
   5. 技术实现：不需要使用机器学习的方法。只需要预设统计计算的规则即可。
3. 通过自然语言交互简化视频制作步骤
   1. 动机：在DataClip中，用户完成交互的方式是通过拖拽图表片段、指定数据的方式进行视频制作，而我们希望采用更先进的交互方式，也即采用自然语言处理的方法进行交互。给用户更好的使用体验。
   2. 输入：用户的自然语言输入。
   3. 输出：初步的视频片段，可以由用户进一步校正。
   4. 预期效果：用户给出的原始数据是中国和美国自1990年以来每季度的GDP数据和自1990年以来的每年的人口数据。用户输入语言描述：我需要以年为单位表现自1990年以来中国和美国的GDP和人均GDP变化情况。理想情况下得到的输出为，中国和美国GDP和人均GDP数据的折线图视频片段，以年为横坐标，以GDP为纵坐标。
   5. 技术实现：首先明确该技术的实现需要以一、二作为基础。我们的模型首先需要具有将数据源映射到图表类型和从原始数据到初步加工数据的映射。其次，需要使用自然语言处理技术。我们需要对用户的输入进行语义分割和命名实体识别，随后根据处理后得到的描述对数据进行加工处理（也即一、二的实现）