# Introduction

关于智慧城市的讨论与实践在近些年来越来越多。虽然关于究竟什么是“智慧城市”的定义有诸多版本，但是它们都有一个共同点，就是试图用海量的城市数据结合算法去解决传统方法不能解决的问题。城市是一个各要素高度耦合的复杂系统，因此想要让城市变得智慧起来是一个很有挑战的系统性工程。可以预见，需要理清城市这个系统在运行时各个主体是什么，它们如何组成城市的各个子系统，这些子系统怎样相互作用并最终形成一个完整的系统。如同解剖一样知道城市如何运作，得到一个超高精度的数字孪生只是一个开始。更重要的是要能够知道当对城市进行一些改动时，整个城市的各方各面会产生什么样的连锁反应。对状态信息的描述是预测效果的前提，而预测的准确率也是衡量描述方法优劣的一个重要参考。在准确描述和预测的基础上，根据需求改变系统的状态，并能够提前预知相应改变可能带来的后果，才能说是形成了一个智慧的城市管理闭环。

智慧城市所涉及的问题复杂且庞大，所以这篇论文并没有不自量力到要去一举解决这个问题。相反，在遵循描述-预测研究范式的前提下，这篇论文选择了一种局限性视角下的交通状态预测问题作为切入点。这样做的原因有三点。其一，对于像城市这么复杂的研究对象，即便是选择一个较为局限前提展开研究也能在实际当中发挥作用，不是非要等到大一统的模型建立不可。其二，城市的各个方面虽然都跟城市相关，但是其实差异性很大，逐个各自采用不同的方法才是正常的。其三，在有限的时间、资源（特别是数据）和能力下选择一个有希望解决的子集是一个更为务实的选择。

交通是城市能够正常运转最不可或缺的功能之一。交通和物流的高效是城市高效的必要条件，如果交通的运行与管理并不智能，那么城市的智能也难以实现。不同于公共交通系统的中心化特性，私人交通具有去中心化特性。简单来说，公共交通系统的调度一般都由一个调度中心进行控制，停运或者进行改动相对容易是可能的，只不过需要考虑这样做是否恰当。而目前为止，对于私人交通的调度还是一个没有实现的事情。这个局面虽然主要的阻力来自于立法方面（譬如公权力部门是否有权力去收集私人汽车的数据），但是技术上确实并没有很成熟的解决方案。当前的交通管制，更多是以道路区段为单位的临时性管制，以暂时改变环境的特性为手段。虽然在法理层面这种基于道路封闭的交通管制是合理的，但是其在提高城市交通效率这一点上的效果确实不大。一种可能的解决方案是采用基于车辆的动态交通管制，即依据交通状态对整个交通系统内的车辆进行定制化管制，从而提高整体效率（譬如，降低整体通行时间、提高平均速度等）。

若要实现这个目标，那么就需要对城市交通系统内的各种车辆的运行规律有足够的了解。前边提到，虽然对于公共交通的了解和掌控水平较高，但是对于分布式的私人交通的了解由于种种原因并不足够。这时候对于出租车的研究就体现出价值了。出租车在交通系统中的行为模式与私人交通非常接近，但是其又有一定的中心化特征（即，由出租车公司统一运营管理，数据收集和利用更为便利），这就使它成为了绝佳的研究对象。

研究这个问题有两个直接的作用和一个潜在益处。第一，这篇论文的对于出租车交通状况的研究结果可以直接为相应的交通部门提供信息。第二，这是对城市交通状态预测方法的探索与创新，在适当的变换下可以应用于其它场景。潜在益处在于给出了一个将深度学习与传统工程领域的问题结合的研究方法案例，这对于类似的研究可以提供参考。

综上所述，这篇论文研究的是预测街区尺度下的出租车交通状态预测问题。预期达到的目标是在街区尺度提供一个出租车出行状态较为准确的预测模型，从而为交通调度提供有价值的参考。

# Problem Definition

## Concepts

如前述，这篇论文研究一种限定前提条件下的交通状态预测问题。这一节的主要工作就是严格地定义本文研究问题相关的概念，formalize将要研究的问题。研究的内容是预测街区尺度下的出租车交通状态，那么首先要定义以下3个核心概念。

**尺度**

所谓尺度，即一种衡量状态的观测精度的概念。一般可以不严谨地用“微观”、“中观”和“宏观”尺度这样暧昧的叙述对尺度进行大致归类。在本研究中，特别指对空间的划分。譬如，街区尺度就是指以街区为空间基本单元，作为trip的起始点或终结点的归类依据。虽然存在trip A 和trip B的起始经纬度坐标a和b不同，但是当且仅当a和b同时落在街区单元X当中时，则在这一尺度下可以认为trip A和trip B的起始点相同。

**状态**

所谓状态，指某种在时间维度呈现变化的事件被观测时所呈现的特征。在本研究中，状态即是在一个时间区间内，用一种数据结构来反映出租车的trip所呈现出的街区间OD总体特征，即对时间区间内交通需求的描述。在本文中状态即是需求状态的意思。

**预测**

所谓预测，指一种对未知的未来状态进行推断的行为。在本研究中进行的预测主要是基于过去的状态来推断未来的交通状态。

基于上述三个核心概念，对研究问题进行formalization。

## Assumptions

这篇论文研究的问题可以认为是时间序列预测问题，预测的对象是交通状态。交通的本质是时空间的物质移动，交通的state是对这一过程的描述。不直接预测具体的交通流量，而是预测交通的状态。

为了实现这个目标，我们需要使用一种恰当的representation来描述交通的状态，使得如果存在任何feature那么都会被包含在内。如前述，交通状态的定义实际上是把一个连续的车流时空间位移状态离散化的过程。这样做的好处在于可以通过一些先验知识把问题进行简化，同时将交通预测问题formalize为时间序列预测问题，而后者有许多成功的方法。为了formalize问题，需要提出三个假设。在假设的基础上，将交通预测问题转化为对邻接矩阵的数值作为时间的函数的拟合问题。在后边的内容中，将对假设进行实验与讨论。

**1. 城市交通过程不是完全随机的，而是一些可直接观测和难以直接观测的因素共同影响下的产物**。

假如城市的交通状况是完全随机的，意味着预测是不可能的。这种完全随机的事件自然界当中存在，譬如无法预测exact moment when a radioactive atom will decay。然而对于城市交通这种复杂系统而言，更可能的情况是它并非全然随机，而仅仅是因为我们掌握的信息不够所以看起来是随机的/无法预测的。

To be specific, 可以直接观测的因素譬如天气灾害，事故，道路临时性封闭等直接作用于物质空间，使得交通系统的environment发生了暂时性的器质性改变，从使得交通行为发生了改变。而不可直接观测的因素则例如特殊事件（体育赛事、演唱会和节假日等）造成的临时交通需求激变，社会经济学因素造成的改变（房租价格上涨导致工作者迁出城市中心区，间接增加了通勤需求）等。可直接观测因素和部分routine的不可直接观测因素都是可以被掌握的，即便获取这些信息并将它们与交通的状态建立一个可靠的映射是困难的。即便如此，一些明显的模式，譬如早晚的交通高峰拥堵，都是明显可以被感知和预测的。

**2. 城市交通在不同时间点会emerge特定的特征，只通过OD记录alone就可以获得足够有用的信息**。

多数可观测因素和不可观测因素在相对长的时间区间内（譬如，1年）不会发生剧烈变化，因此可以看作是一个常量或者周期性明确的周期函数。譬如，在一段不短的时间内，城市的人口增长，经济发展的速度不足以显著地影响交通情况，而道路网络的修改也是小概率事件。在这种情况下，主要变量是交通需求。

日常生活中，我们大多数时候仅仅依靠时间或者一小段时间的车流移动趋势就可以相对准确地预测到几分钟后是否会发生拥堵。这种感性判断能够发挥作用的主要原因在于我们掌握一个重要信息，就是人是要上下班通勤的，而大多数公司的上下班时间重合，这就导致了交通需求会呈现出明显的潮汐特征。虽然每天并非能精确到秒级预测高峰拥堵发生的事件，但是由于存在这样一个大背景，所以总能预期高峰发生的大致时段。结合周围车流的状况作为信号，那么预测到这种几乎每日必定发生的状态就不难了。这个情境中，交通需求被掌握了，那么在环境可以看作一个常量的时候，预测就成为了可能。

OD记录就是一种对交通需求最直接的描述。

OD矩阵本质上是交通需求的一个后验统计。当OD矩阵所覆盖的时间足够短时，可以近似认为时对交通系统状态的实时描述。

**3. 状态之间会发生转移，下一个状态受到之前N（N>=1）个状态的加权影响**。特别地，当N=1时，交通需求状态的转变过程满足马尔可夫性。

由于城市系统的复杂性，因此这个假设很难得到数学意义上的严格证明。但在“微观”层面这个假设具有一定的合理性。

## Formulation

基于前述三个假设，本文的主要研究问题就是对于时间序列的预测问题，其中预测的每一个元素是一个描述交通状态的数据结构。这一数据结构反映出租车在城市不同区域内的OD联系特征，具体的形式包括但不限于邻接矩阵。

数据结构的具体的定义将在第四章中进行，预测模型的构建和实验将在第五章中进行。

## Summary

因此，这篇论文的主要工作就是把交通预测问题形式化为对一个数据结构随着时间变化的推断问题后，再利用相应的模型和算法解决之。

# Related Work

前人的工作总是有的。这部分可以从RNN survey和RP当中节选一些内容。

## 特征学习（auto encoder）

## 时间序列预测

## 循环网络

## 交通预测中的时间序列预测应用

## Autoencoder和RNN的结合在序列预测中的应用

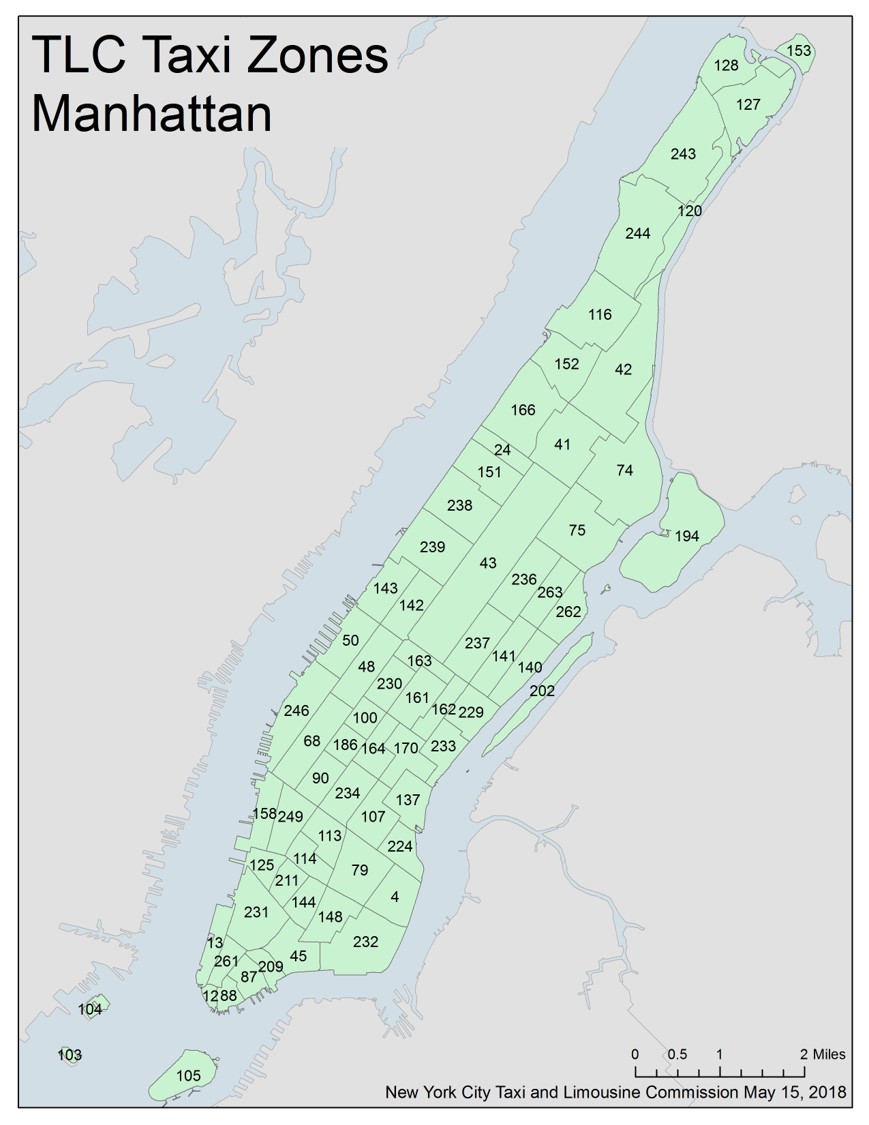
（把收集到的论文简述一下就好，没有必要堆积工作量。）

# State Representation

本文要预测交通状态，那么如何描述交通状态是很关键的。如果representation比较好，那么就更有可能捕捉到重要的信息。相反，也可能因为不合理的representation导致关键信息被丢失或者掩盖，从而使得预测不可能。因此，这一部分的主要工作就是对原始数据进行清洗和预处理，并采用几种不同的数据结构来represent交通状态，并从中选择效果最好的一种用于后续预测模型的构建。

## Data Preparation

原始数据是出租车的行驶记录，记录了包括上下车的地点编号和时间这两组核心信息，同时还包括搭乘人数，车费，支付方式等其它信息。数据的来源是纽约市出租车信息网站，数据的种类主要有三种，分别为yellow cab, green cab and for-hire-vihcle三种。考虑到数据的完整性和研究范围因素，选取了yellow cab这种活动范围限定在Manhattan地区的出租车作为研究对象。研究选择了2017年和2018年全年的yellow cab共计超过2亿条行驶记录。研究的区域一共分为69个block，编号地图如下：

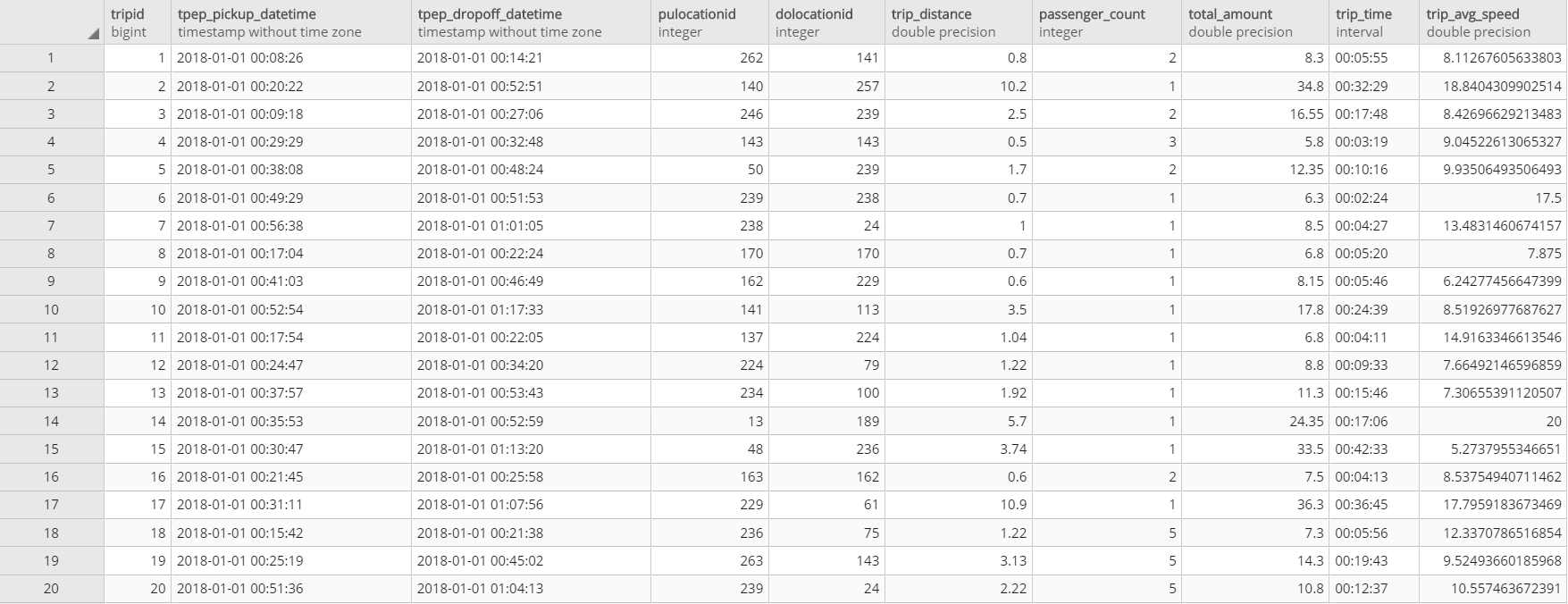


值得注意的是，地图中的编号并非从1到69，而是更大范围内的69个值。这是因为曼哈顿外部地区的和曼哈顿属于同一编号体系下。数据中的每一条记录的起点和终止点都在这69个区域当中取值。选择曼哈顿区域除了数据获取的易得性之外，还有一个原因是街道非常规则，基本上是正交街道为主，可以避免一些特殊地理因素带来的噪音。

前文提到，为了对交通状态进行预测，关键在于选择一种representation来描述状态。同时，我们做出了交通状态并非无需随机而是存在一些structure的。在本章将会提出两种representation，同时对其是否存在structure进行验证（证明）。在选择一种表示方法之前，需要先对原始数据进行explore和预处理。

**数据库建立**

由于原始数据是以.csv文件的格式提供的，且数据分散（按照月份），而这种情况下效率不高，且进行查询时不甚便利，因此在开始一切工作之前，第一步的工作将所有原始数据导入数据库，形成一个big table。



**清洗**

因为本文所用的数据是由Taxi and Limousine Commission (TLC)整理和提供，因此数据质量相对较高。唯一的问题是由于数据在收集是分布式的，由每个taxi上的记录仪获取。这意味着收集端出错的可能性是存在的。

通过初步清理，主要filter掉了以下几种错误数据：

1. 里程为0的trip
2. Duration为0的trip
3. 出发时间比到达时间晚的trip（这显然是不可能的）
4. Trip的duration长得不可思议（譬如长达几个月甚至更久）
5. 平均速度超出了合理的范围
6. 关键字段（起始地点，终止地点，起始时间，终止时间）缺失的数据

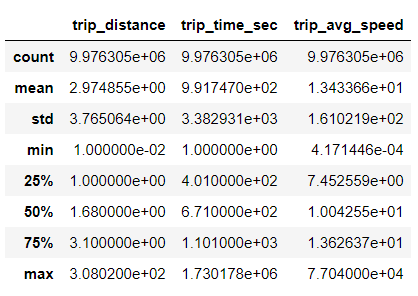
其中，第五点的标准非常宽松，只要车辆的平均速度小于F1方程式塞车的最高可能时速就不会被过滤掉。这意味着速度快得飞起的那些trip只能是因为记录仪没有被正确地使用。这些数据基本上都是由于在收集过程中发生的错误造成的无效数据，因此排除这部分数据可以避免噪声对模型拟合造成负面影响。

在经过清洗后，可用的数据量恰好稍多于2亿条。

**数据初探**

在对数据预处理过后，为了得到一个恰当的数据representation，最好先对数据有一个大概的认知。考虑到data peeking的负面因素，只以2017年其中一个月的数据作为抽样进行exploration。

可以发现，平均trip time大约在990秒左右，相当于16.5分钟。同时有95%的trip时间小于等于35分钟，99%的trip时间小于等于60分钟。中位数约为10分钟，超过一半以上的trip时长再10分钟到20分钟之间。trip的时长分布的分布特征如下：



(trip time distribution)

除了时间之外，还有许多值得explore的信息，譬如最热门的地点，交通高峰时段等等。然而，对于时间序列预测来说，最重要meta information还是如何在时间维度对数据进行处理，从而使得类似热门地点、高峰情况这些目标信息能够得到更好的展现，从而有利于预测任务。本章的主题是就是尝试构建几种representation，因此对于具体信息的统计分析和本章的主题并不十分相关，因此省略。对于具体交通状态的特征的讨论，将在第五章对预测模型的评估中展开。

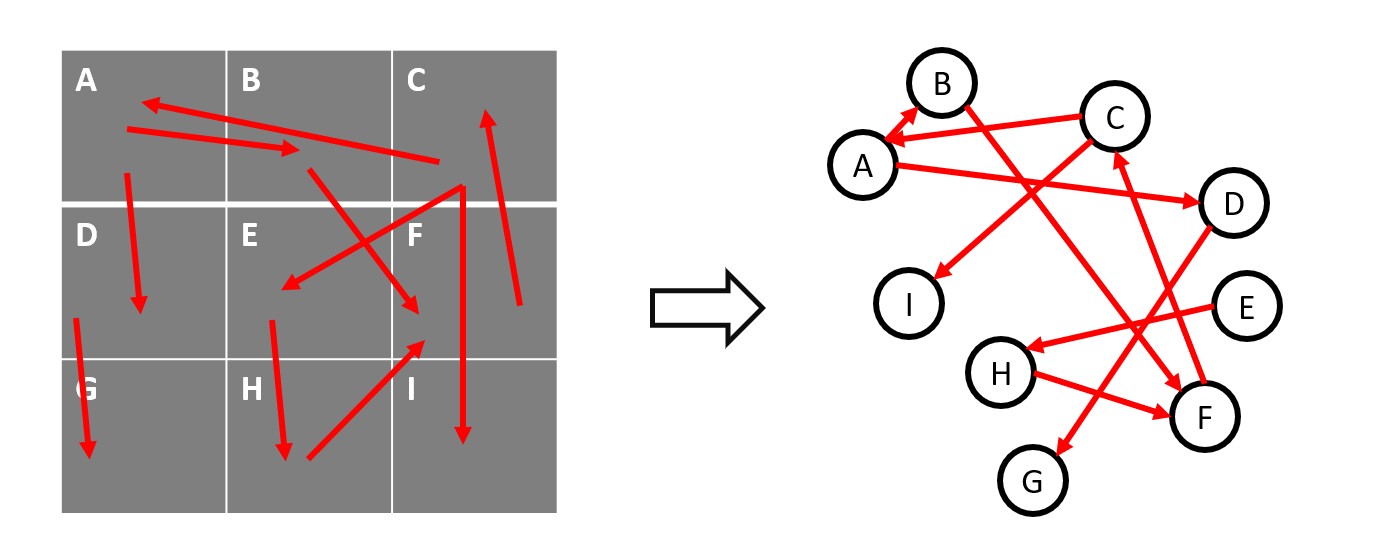
## State Data Structure

（主要提出几种数据表达方式。1. 邻接矩阵一层。2. 邻接矩阵3层。讲清楚为什么这么做。）

前面提到了OD矩阵是一种描述交通需求的有效工具。它的效率主要在于不丢失信息的前提下，用更加condense的形式描述了同样的信息量。同时，这种信息的聚合方式可以立刻呈现出交通的热门地点等信息，从孤立的记录种发掘出了相关性。

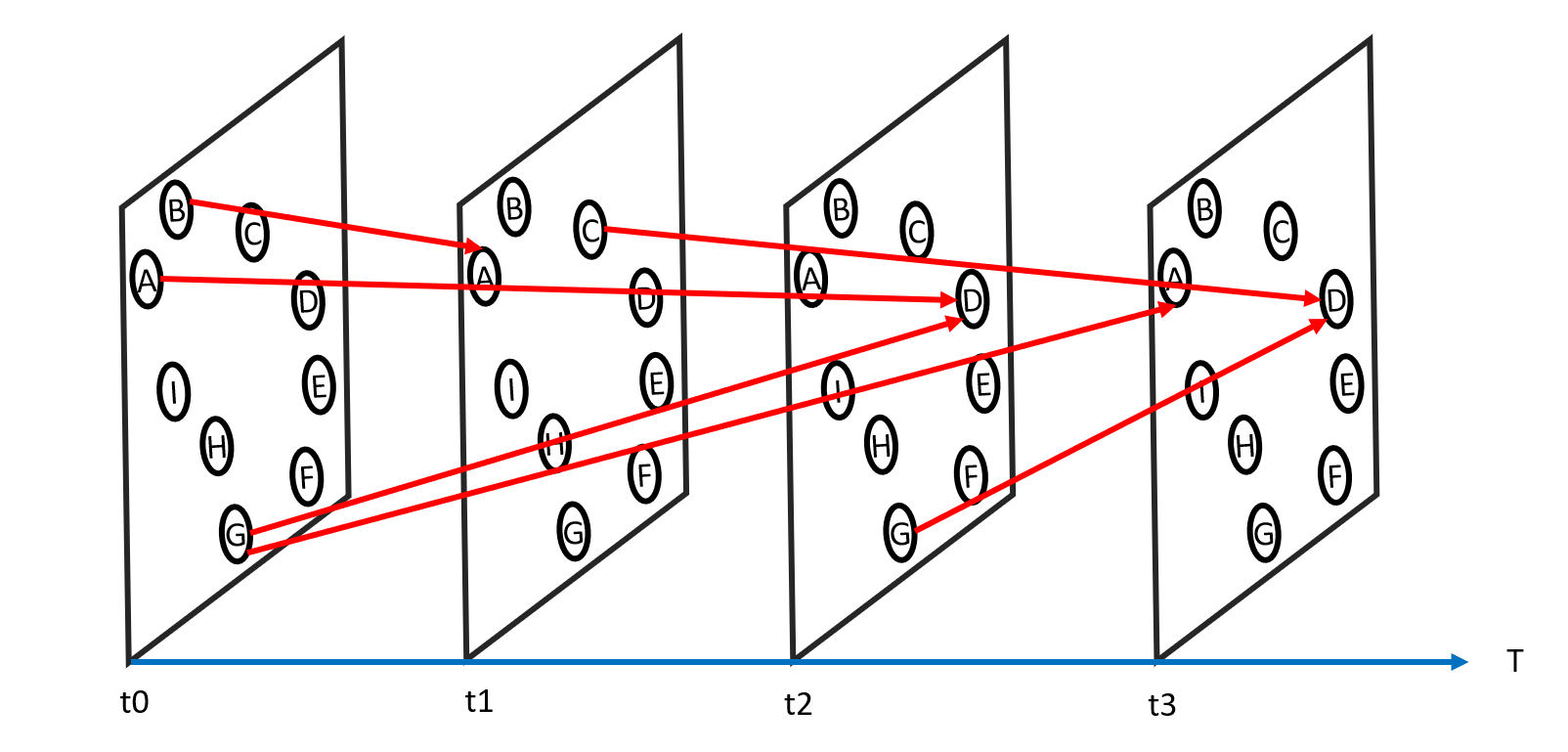
（示意图，tabular数据是一维的，而od矩阵则是二维的）

OD矩阵就是adjacency matrix，矩阵的行和列是trip出发地和目的地的编号，矩阵的每一个entry表示的是从i到j的trip数量。这个矩阵反映的是一个有向图，如下图所示：



倘若只是提供了一定数量的记录然后生成OD矩阵的话，那么这是一个非常简单的任务。但是当把OD矩阵作为交通状态的representation并且希望基于这种representation来进行预测时，就要面对如何做时间切片的问题。所谓time slice，类似于分布式算法中的snapshot，就是在单位时间内记录此时整个研究区域的od状态。由于我们的任务是预测下一个timestep的od矩阵，这就意味着必须在时间维度对trips进行切分，获得一个od矩阵在时间维度上的变化序列。这个序列中每一个元素（即od矩阵）的时间跨度应该是等长的，又称为granularity of time。把时间跨度设定为等长，是一种选择。理论上当然可以把时间步骤设定为不等长的，但是这种设定比较违反直觉。本文不打算在这个问题上进行过多的深入探讨。将整个trip数据集转化为od矩阵的序列（即，离散化）的主要trick在于如何处理一段trip的首尾在不同的time slice这种情况。

由于trip的发生具有高并发的特点，于是在保证时间切片的时间跨度reasonable和meaningful的前提下，想要找到一种干干净净地包含所有trip首尾的切片几乎是不可能完成的任务。



（示意图：不能干干净净切割）

在一个time step里边，大致有三种trip:

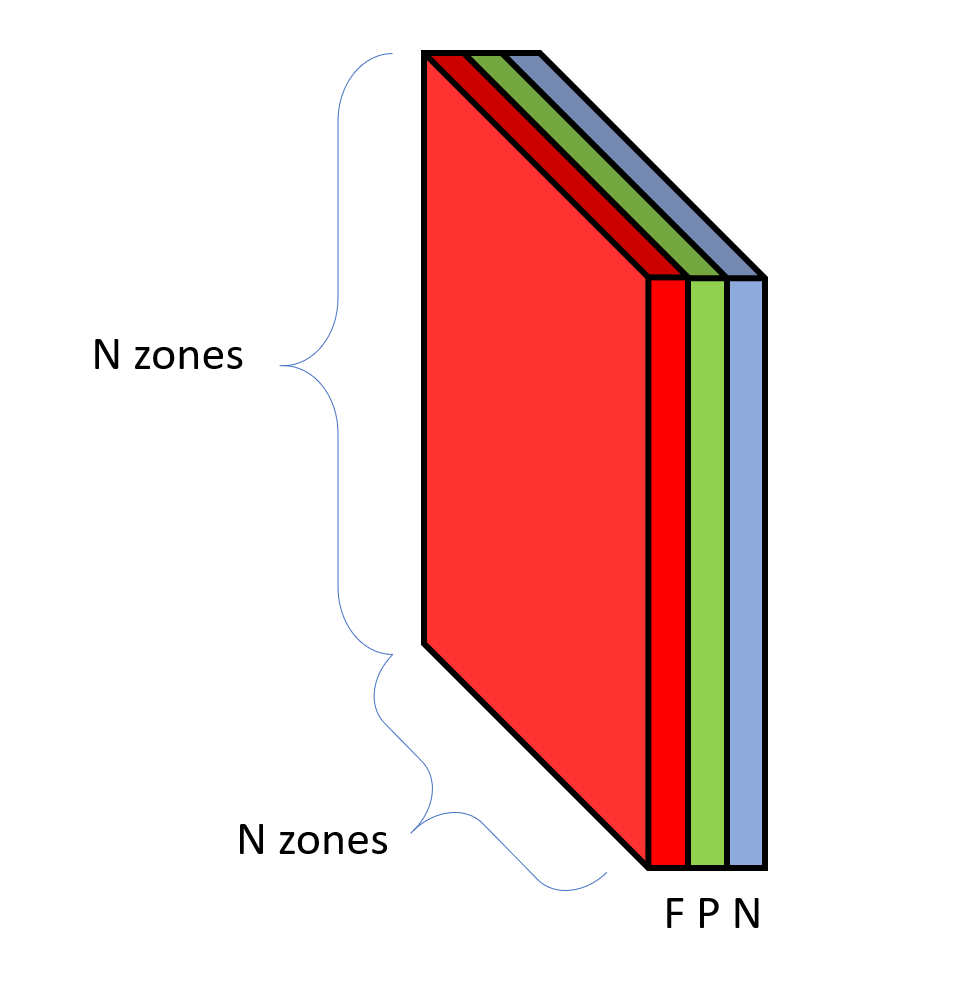
（配示意图）

1. 头在尾不在
2. 尾在头不在
3. 头尾都在

其中第一种情况表明【有一段trip在这个time step发生了】，这意味着这个event会对未来的状态产生影响。第二种情况表示有一段trip在这个time step结束了。第三种情况表是在当前时间区间内的od情况。这三种trip在一个time step中数量的多少具有各自的意义，在一个序列中来看，相互之间的关系如下：

（图解：其中，第一种trip在当前time step的意义是【尚未结束的】trip，而第二种和第三种则是【已经结束的】trip）

根据三种trip的特性，可以把每个time step的od信息用一个(3, Z, Z)tensor表示，其中Z为街区的数量，如下图所示：



（representation1示意图）

相比只有一层的od矩阵，三层的representation保留了更多的原始信息，看起来是更好的。然而果真如此吗，我们无法仅仅就保留了更多信息这个角度认为它是一种更好的representation，因为也存在一种可能，即这个多出来的信息其实对于预测并没有什么帮助，反而可能是一种噪声。先验经验在这里虽然符合直觉，但是在实际是否更好则需要通过实际的数据做实验来判定。因此在后续的内容中将会采用两种representation，作为对照分析。

## Experiment

这部分主要的工作是比较两者的性能。性能的定义是究竟能反映多少信息。

用熵衡量。

## Summary

这部分主要的工作就是清洗了一下数据，然后基于对数据的初步分析和先验知识，形成了两种描述交通状态的representation形式。这两种representation表现出来的性能差不多，那么就看后续在模型中有什么表现吧。

# Exploration of Models

## Architecture

## Loss Definition

## Experiments

## Performance

## Summary

# Conclusion

研究结论