城市计算：概念，方法和应用

城市一体化进程的快速提高在大大提高人们生活现代化水平的同时也产生了一些较大的问题，例如交通拥挤，能源消耗与污染问题。城市计算致力于使用从城市中集成的数据来处理这些问题（如交通流，人员移动和地理位置数据）。城市计算把城市遥感，数据管理，数据分析和服务提供联系起来，然后将其转换为一个周期性稳步提高人们生活水平，城市操作系统和环境的过程。城市计算是一个将计算机科学和传统城市相关领域（如交通，土木工程，环境，经济，生态和社会学等城市范围内的学科）融合的交叉学科领域。本文首先介绍了城市计算的概念，从计算机科学的角度讨论它的框架和关键挑战。第二，我们将城市计算的应用划分为7类，包括城市规划，交通，环境，能源，社会，经济和公共安全，同时给出各类有代表性的情况。第三，我们将城市计算的代表性技术划分为4个方面，分别是城市遥感，城市数据管理，机构数据的知识融合以及城市数据可视化。最后，我们给出了城市计算的发展前景，提出了几个在这个领域内还未进行过的研究课题。

分类和主题描述：H.2.8[数据集管理]：数据集应用——数据挖掘，空间数据集和GIS；J.2[工程物理学]：地球和大气科学，数理统计；J.4[社会行为学]：经济，社会学；G.1.6[优化]；G.1.2[近似]；E.1[数据结构]；E.2[数据存储表现]

一般术语：算法，度量，实验

关键词组：城市计算，城市信息学，大数据，人类移动，城市动力学，城市遥感，知识融合，异构数据计算，轨迹

# 1 简介

城市一体化进程的快速提高促使产生了一批大城市，这在提高人们生活现代化水平的同时也产生了一些较大的问题，例如空气污染，不断增长的能源消耗以及交通拥堵。考虑到城市的复杂性和动态性，这些问题的解决显得几乎不可能。近来，遥感技术和大规模计算基础设施的建设促使产生了多样的城市空间内的大数据（例如人类移动，空气质量，交通模式和地理数据等）。如果能正确利用这些蕴含这丰富的城市信息的大数据，那么将可以有效的帮助解决上面提出的问题。例如，我们可以通过分析城市范围内的人类移动数据来检测城市道路网络中的潜在问题。这些发现可以更好的帮助城市未来的规划。另一个例子就是通过学习空气质量和起来数据源（如交通流，特征点）之间的相关关系，我们可以发现城市空气污染源。

受建立更智能城市的机遇的启发，我们提出一种版本的城市计算，旨在充分释放从城市空间中收集的大容量异构数据的隐藏信息，然后将这些信息应用到解决当今城市面临的主要问题上。简而言之，我们可以使用大数据来解决大城市中的大问题，如图1所示。

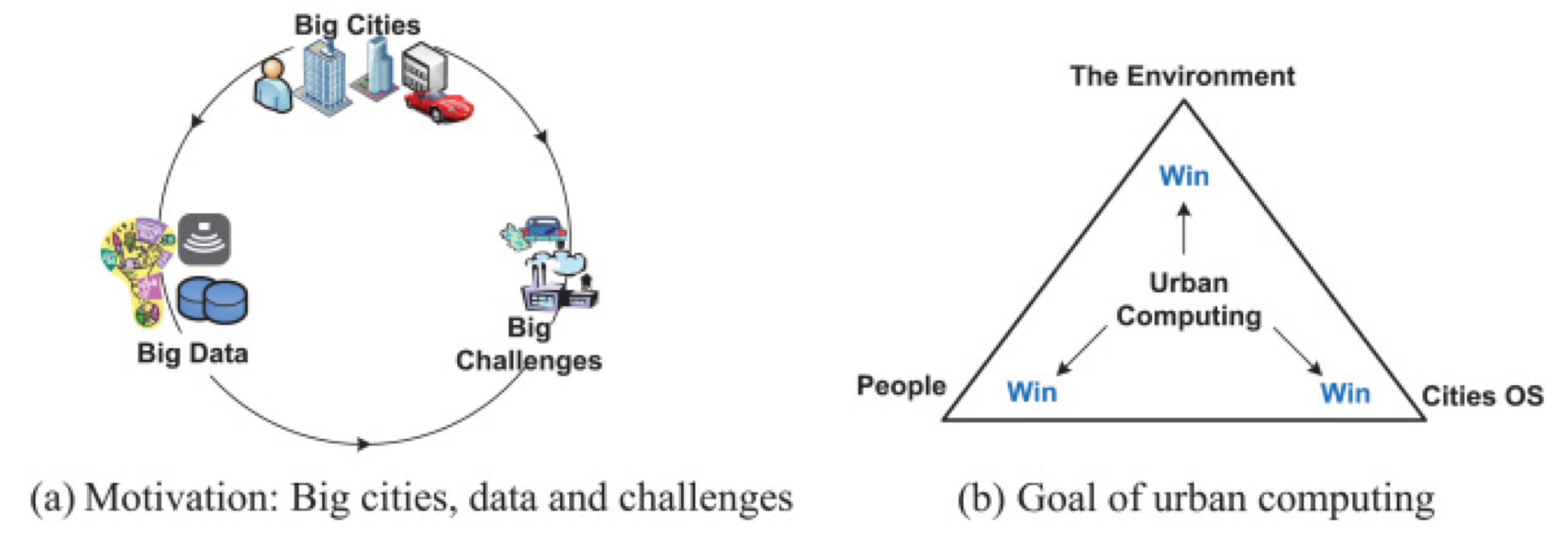


图 1 城市计算驱动和目标

虽然城市计算的概念在2007年就已经提出，但到目前为止，其还是一个比较模糊的概念，有许多问题悬而未决。例如，城市计算的核心研究问题是什么？研究主题面临的挑战是什么？城市计算的主要方法有什么？这个领域的代表应用有什么以及城市计算系统是如何运作的？

为了解决这些问题，我们在这片文章中系统的给出了城市计算的介绍，包括总体框架，主要研究问题，方法和应用。本文将有助于领域内更好的理解和探索城市计算这个初期领域，然后能给出有质量的研究结果和系统帮助建设更加绿色和智慧的城市。除此之外，城市计算是一个将计算机科学和传统城市相关领域融合的交叉学科领域，其中传统城市相关领域有交通，土木工程，环境，经济，生态和社会学等。本文主要从计算机科学方面讨论上述问题。

后文主要按照如下方式组织：在第二部分，我们介绍城市计算的概念，给出总体框架以及框架中每一步的主要挑战，同时也讨论了城市计算领域常用的数据集。在第三部分，我们将城市计算的应用分为7类，然后给出了每一类的代表情况。在第四部分，我们介绍了在具体的城市计算的情况中常用的四种方法。在第五部分，我们对文章进行总结，同时提出了几个未来研究方向。

# 2 城市计算框架

## 2.1 定义

城市计算是一个获取，集成，分析从城市中各种数据源中得到的异构大数据来解决城市现有问题（如空气污染，持续增长的能源消耗和交通拥堵）的过程，数据源包括传感器，设备，汽车，建筑和人类。城市计算将普通的城市中无处不在的传感器技术，先进的数据管理和分析模型，新颖的可视化方法联系起来，以此来创造三赢的提高城市环境，人类生活质量和城市操作系统的解决办法，如图1(b)所示。城市计算也帮助我们认识城市现象的本质，甚至预测城市的未来。城市计算也是一个融合计算机科学和传统领域（如城市交通，土木工程，经济，生态和社会学）的交叉学科领域。

## 2.2 总体框架

图2表示城市计算的总体框架，其由四层构成，分别是城市遥感，城市数据管理，数据分析和服务提供。以城市异常检测为例，如下为简要的框架运作介绍。

在城市遥感这一步，我们利用GPS传感器或者人类的移动手机信号来不断的获得人类移动信息（如城市道路网络中的路径选择行为）。我们也不断的收集人们发送到因特网上的社交信息。在数据管理这一步，人类移动和社交数据通过一些索引结构很好的组织起来，此索引结构能同时合并时空信息和内容来支持高效数据分析。在数据分析这一步，一旦发现异常，我们可以识别到人们移动显著有别于其原有模式的位置。同时我们可以通过从位置和时间相关的社交数据中挖掘具有代表性的条目来描述异常。在服务提供这一步，发生异常的道路位置及其描述就会被发送给附近的司机，这样他们就可以选择避开此异常路段。除此之外，异常信息也会被发送给交通部门来分散交通，诊断异常。此系统不断重复上面的过程来保证城市异常被及时有效检测到，以此来提高人们的驾驶体验同时减少交通拥堵。

与其它基于单一数据-任务框架的系统（如网络搜索引擎）相比，城市计算拥有多数据-多任务的框架。城市计算的任务包括提高城市规划水平，降低交通拥堵，减少能源消耗以及降低空气污染。另外，在一个任务中，我们通常需要利用多样的数据源。例如，上面提到的异常检测就用到了人类移动数据，道路网络和社交数据。不同的任务可以通过结合从框架中的不同层使用不同数据获取方式，管理方式和分析技术的得到的不同数据源完成。

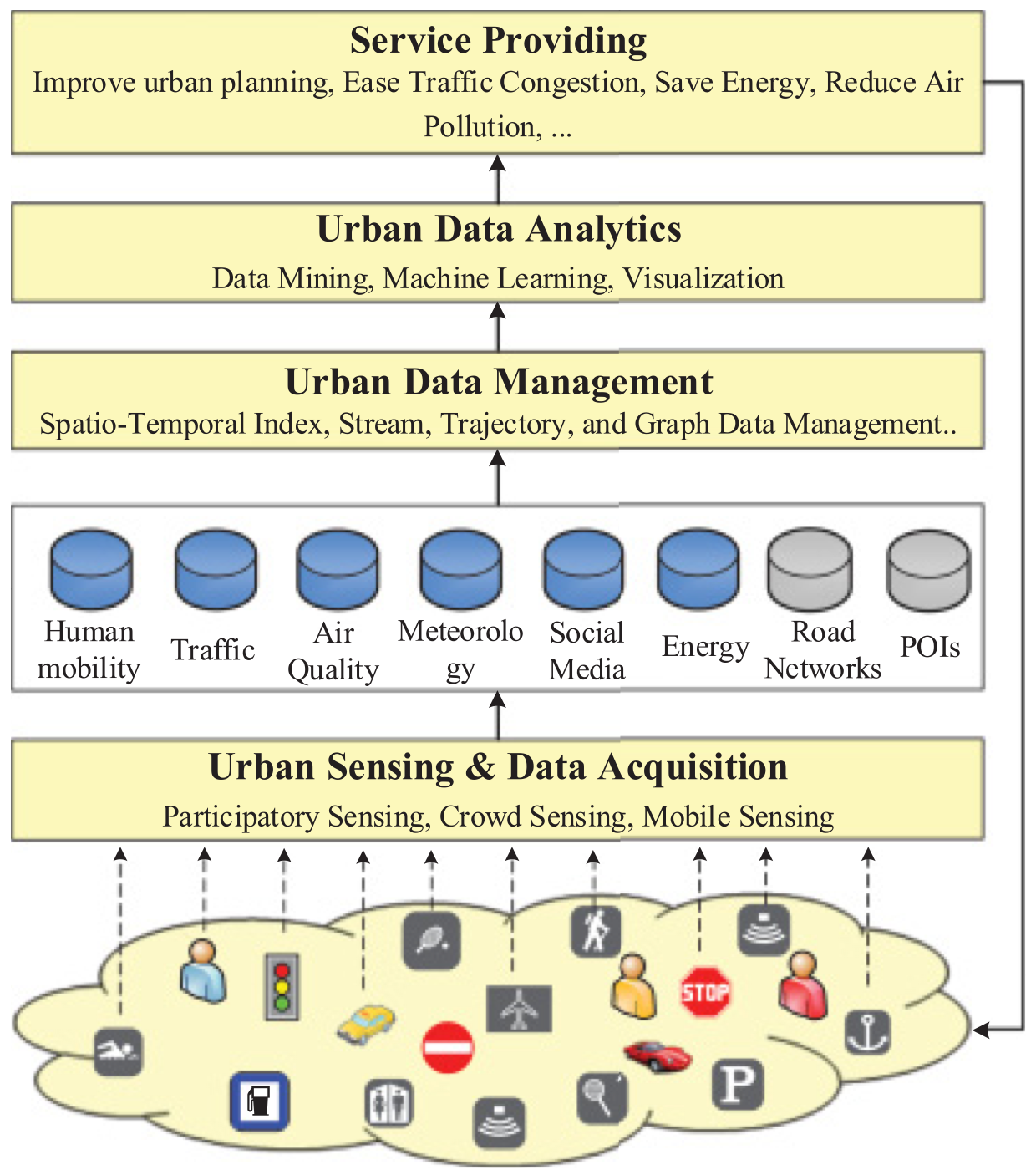


图 2 城市计算总体框架

## 2.3 主要挑战

城市计算的目标和框架导致如下三种主要挑战：

### 1 城市遥感和数据获取

首先的困难是可以*无处不在*的搜集*城市范围*内的数据的数据获取技术。考虑到斜体条目的叙述，这是一个重要的问题。监测一段道路上的交通流很容易，但是持续不断的获取城市范围内的交通状况十分困难，因为并不是每一段道路上均有传感器。建立新的遥感基础设施可以达到这一目的但是反过来会加重城市负担。探索如何巧妙的利用现有的城市基础措施达到目的显得尤为必要。将人类看做传感器是一个很好的可以帮助解决这一问题的概念。例如，当用户在社交网站上发布消息时，他们也同时帮助我们了解到其周围发生的事件。当许多人在一条道路网络中时，他们的GPS轨迹也反映了交通模式和异常。然而，任何事物都有两面性，尽管人类作为传感器有其灵活性和智能性，同时也带来了三个困难（我们将在4.1中进行讨论）：

* 能源消耗与政策：这是一个重要的参与式遥感应用问题，用户主动的贡献他们的数据（通常使用智能手机）来节省智能手机能源同时在遥感过程中保护自身隐私。这是在能源，隐私和共享数据使用下的一个折衷。
* 松散控制和不均匀分布的传感器：我们可以将传统传感器安装在任何地方，同时设置其以固定的频率发送传感数据。然而，我们不能控制人类发送数据的时间和频率。在一些区域内某时段内甚至没有人出现（如没有遥感数据），不可避免的造成数据丢失和稀疏问题。另一方面，一些区域的用户集成内容可能出现冗余，带来不必要的存储负担。除此之外，我们获取的数据总是来自一部分用户，因为不是所有人都共享数据。样本数据的分布可能与整体数据集的分布存在偏差，其取决于人类的移动。
* 非结构化，隐含，噪声数据：传统传感器集成的数据通常是结构化的，明确的，干净的以及易懂的。然而，由用户贡献的数据通常格式不一例如有文字信息和图片或不能像使用传统传感器数据那样带给我们一个明确的结论。

同时，以人类作为传感器的信息噪声显著。

以论文[[1]](#footnote-1)为例来说明这两个挑战。在例子中，张等旨在利用配有GPS的出租车司机作为传感器来检测一个加油站的排队时间以及推测该地在加油的汽车数量。这样做的目标是估计一个加油站的汽油消耗量，最终估计出在一个时间段内城市的汽油消耗量。在这项应用中，我们获得的是出租车司机的GPS轨迹，其并没有明确的高速我们结果。除此之外，我们不能保证在每个加油站的任何时间都有出粗车司机，由此带来数据丢失的问题。。同时加油站中出租车的存在可能与其他类型的车不同；例如观察加油站中的更多出租车并不代表更多其他的汽车。还有就是出租车司机可能将车停在加油站附近的位置休息或等红灯。这些来自GPS轨迹的数据都是噪声。简而言之，我们通常需要从人类传感器集成的局部，有偏差，噪声和隐式的数据中学习我们真正需要的内容。

### 2 异构数据计算

* 从异构数据中学习交互增强知识：解决城市挑战包括范围很广的因素（如探索环境污染包括交通流，气象学和土地使用的同步学习）。然而，现有的数据挖掘和机器学习技术通常掌控一种类型的数据。例如，计算机视觉是处理图像，自然语言处理基于文字。根据文献[[2]](#footnote-2)和文献[[3]](#footnote-3)，同等的对待从不同数据源中得到的特征（如简单的将这些特征置于一个特征向量中，然后将其放入分类器模型中）并不能得到最好的结果。除此之外，在应用中使用多元数据源会产生高维空间，从而加重数据稀疏化问题。如果不正确的处理，过多的数据源甚至会降低模型的表现。因此需要一种先进的数据分析模型来从不同的异构数据源（包括传感器，人类，汽车和建筑数据）中学习交互增强知识。
* 高效的学习能力：许多城市计算情况需要及时的反馈，如检测交通异常和监测空气质量等。除了仅仅通过增加机器数量来提高计算速度之外，我们需要将数据管理和挖掘以及机器学习算法综合运用到计算框架中来提供高效的知识发现能力。除此之外，传统的数据管理方法技术通常是为单一模型数据源设计的。一种先进的能组织多元数据（如流，空间和文本数据）的管理方法目前还没有。因此，多元异构数据计算是数据和算法的融合。在4.3中做详细讨论。
* 可视化：大量数据带来了大量需要被很好的表示的信息。将原始结果可视化可以激发出解决问题的新观点，同时将计算结果可视化可以直观的表现其蕴含的知识以此来帮助做决策。数据可视化也可指彰显不同因素之间的相关关系或因果关系。城市计算具体情况中的多元数据的可视化会导致高维的视图，例如空间，时间和社会。如何以不同的视图关联各种数据，检测模式和趋势是很具挑战性的。除此之外，当面对多种类型和大容量数据时，如何得到能够帮助人们提出新假设的交互式可视化结果更加困难。这要求将实时的数据挖掘技术综合运用到城市计算目前还未考虑的可视化框架中。

### 3 现实与虚拟世界复合系统

与数据的集成和消耗都在数字虚拟世界的搜索引擎或电子游戏不同，城市计算从两个世界同时集成数据（如将社交与交通结合起来）。可选的，数据（如汽车的GPS轨迹）在现实世界中集成然后再传送到数字虚拟世界，如云系统。当数据与其他数据源在云端被处理后，从中获得的信息就会通过移动客户端反馈给现实世界中的用户（如驾驶方向建议，出租车拼车以及空气质量监督）。这种双世界系统的设计比传统的只在单个世界里运行的系统更加具有挑战性，因为这样的系统需要同时和以不同的格式和频率接收和发送数据的许多设备和用户交流。

## 2.4 城市数据

在这一部分，我们将简要的介绍在城市计算中常用的数据源，同时简要的说明在使用这些数据源时常见的问题。

### 2.4.1 地理数据

在城市计算的实例中，道路网络数据是最常用到的一种地理数据，例如，交通监控和预测，城市规划，路线，以及能源消耗分析等，其通常被表示为由一系列边（代表道路段）和一组点（代表道路交叉口）构成的图。每一个点拥有独立的地理坐标；每一条边由两个点和他们之间的地理空间点序列描述。其他的一些属性，如长度，限制，道路类型，道路数量也被关联到边中。

POI，例如一个餐厅或商店，通常用名字，地址，类型和一个空间坐标描述。一个城市中有大量的POIs，它们的信息可能随时间而改变（如一个餐厅更改了它的名字）。因此，搜集POI数据不简单。通常来说，有两种产生POI数据的方法。一种是从已有的黄页数据中获取，使用一种地理编码方法自动的将文字地址转换为地理位置坐标。另一种方法是在现实世界中手动搜集POI信息，例如，携带一个GPS记录仪取记录一个POI的地理坐标。一些地图数据提供者主要使用后一种方法，如Navinfo何AutoNavi。最近，一些基于位置的社交网络服务如Foursquare允许终端用户在系统中创建新的POI，如果此POI尚未被包含在内。为了包含足够大的POIs，大量使用的在线地图服务，如Bing和Google地图，通常结合上述的两种方法来搜集POI数据。因此也产生了许多问题。例如，我们如何确认哪一种POI信息是正确的呢？有时一个错误的POI地理坐标会将人们带向一个错误的地方。或者说我们如何将使用不同方法搜集到的POI数据结合起来呢？

国家使用数据描述一个地区的功能，例如住宅区，郊区和森林，实际上由城市规划者规划，概括的用卫星图像测量。例如美国地理调查局将每30m30m的方形土地划分到21类的地被植物中，如草地，水和商业。在许多发展中国家，城市中随时间推移增加了许多新的基础设计，同时一些老旧建筑被拆除，城市的现状也许同其原始的规划有所不同。由于卫星图像不能区分有划分的条形地区，例如教育，商业和住宅地区，获取大城市的现有土地使用数据并不容易。

### 2.4.2 交通数据

有许多搜集交通数据的方法，如使用循环传感器，监控摄像头以及流动车。循环传感器通常以成对的形式装配在主要道路上（如高速路）。这种传感器监测一辆汽车穿过两个连续监测器的时间间隔而不是绝对时间。一直每一对循环监测器的距离，我们可以根据时间间隔计算得到汽车的速度。统计在一段时间内通过一堆循环监测器的汽车数量，我们就得到了这条路上的交通流量。考虑到部署和维护循环监测器的经济和人力成本巨大，此类交通监控技术通常只运用到主要道路上而非级别较低的路段。因此，循环监测器的覆盖面积有限。除此之外，循环监测器无法告知我们汽车在一段道路上和道路件的行驶方式。一辆汽车在交叉口消耗的时间（如等红灯和转弯）不能在此类数据中被识别。

监控摄像头在城市区域中广泛分布，生成大容量的反映交通状况的图片和录像。这些数据向人们展示了一种虚拟的交通状况。然而，将图片和录像自动转换为具体的交通流量和速度仍然是一件比较困难的事情。将使用一个位置训练得到的机器学习模型运用到另一个位置是十分困难的，因为他们具有不同的道路结构和不同的摄像头设置，例如高度，角度和焦点。因此，使用这种方法监测城市范围内的交通状况主要依靠人力执行。

流动车信号是通过带有GPS传感器的车辆穿梭城市获得。这些车辆的轨迹会被发送到一个中央控制系统，然后将其匹配到相应的道路网络中来获取速度和道路段。基于各种原因，许多城市早已在出租车，公交车和物流车中安装了GPS传感器，因此流动车数据已经被广泛接收。相对于循环监测器和摄像头监督的方法，流动车交通监控方法有更好的灵活性以及更低的配置成本。但是，流动车的覆盖面依靠这些车辆的分布，这种分布可能随时间而变，同时，在一个时间段内的数据可能有偏差。换句话说数据稀疏的问题依旧存在，解决这个问题需要依靠更加先进的能依靠有限数据恢复城市方位内交通状况的探索技术。文献[[4]](#footnote-4)给出了针对将出租车GPS轨迹转换为社会与社区动态的调查。

### 2.4.3 手机信号

一条电话记录（CDR）包括其详细的属性，如打电话与接电话的双方，开始时间以及持续时间。根据此信息，我们可以得到个体的行为或者建立不同用户之间的关系网络。也可以推测用户间的相似性。另一种手机信号的分类更加关注于用户的位置而不是手机之间的交流。使用三角位置算法，一个手机的位置粗略的由三个或多个基站计算得到。这种数据可以代表城市范围内的人员移动，同时亦可以永磊检测城市异常或者从长期来看，可以用来学习城市功能区域和城市规划。有时两种类型的手机数据会被集成使用（如有手机之间的交流记录和每个手机的位置）。

### 2.4.4 交换数据

人们在城市中穿梭的时候会产生大量的交换数据，例如地铁或公交车刷卡数据以及停车场买票数据。在城市交通运输系统中，有大量刷卡数据可用，人们在上地铁或公交车前都会刷RFID卡。一些系统还会要求人们在下地铁或公交车时也刷卡。每一项交易记录包括时间戳，起始/到达站，站点ID以及乘车费用。这是另一种表示城市范围内人群移动的数据。

在街边停车通常会通过停车仪进行收费。停车场的收费记录一般包括付费时间以及费用。这种数据代表了一个地方的汽车量，这不仅可以用来提升城市停车区域的基础设施建设，同时也可以用来分析人们的行车模式。后一种还可以用来帮助商业广告选址。

### 2.4.5 环境监测数据

1. Zhang, D. Wilkie, Y. Zheng, and X. Xie. 2013. Sensing the pulse of urban refueling behavior. In Proceedings of the 15th International Conference on Ubiquitous Computing. ACM, 13–22. [↑](#footnote-ref-1)
2. Zheng, F. Liu, and H. P. Hsieh. 2013b. U-Air: When urban air quality inference meets big data. In Proceedings of 19th SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, 1436–1444. [↑](#footnote-ref-2)
3. Yuan, Y. Zheng, and X. Xie. 2012b. Segmentation of urban areas using road networks. MSR-TR-2012-65. [↑](#footnote-ref-3)
4. Castro, D. Zhang, C. Chen, S. Li, and G. Pan. 2013. From taxi GPS traces to social and community dynamics: A survey. ACM Computer Survey 46, 2, Article 17, 34 pages. [↑](#footnote-ref-4)