Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
iCity & Big Data—Perspective

大数据研究在意大利的远景

Sonia Bergamaschi [a], Emanuele Carlini [b], Michelangelo Ceci [c], Barbara Furletti [d], Fosca Giannotti [d], Donato Malerba [c,e,*], Mario Mezzanzanica [f], Anna Monreale [d], Gabriella Pasi [g,*], Dino Pedreschi [d,h], Raffele Perego [b], Salvatore Ruggieri [h]

[a] *Department of Engineering "Enzo Ferrari," University of Modena and Reggio Emilia, Modena 41125, Italy*
[b] *High Performance Computing Laboratory, Institute of Information Science and Technologies of the Italian National Research Council (ISTI-CNR), Pisa 56124, Italy*
[c] *Department of Computer Science, University of Bari Aldo Moro, Bari 70125, Italy*
[d] *Knowledge Discovery and Data Mining Laboratory, ISTI-CNR, Pisa 56127, Italy*
[e] *Big Data Laboratory, National Interuniversity Consortium for Informatics, Rome 00185, Italy*
[f] *Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan 20126, Italy*
[g] *Department of Computer Science, Systems and Communications, University of Milano-Bicocca, Milan 20126, Italy*
[h] *Department of Computer Science, University of Pisa, Pisa 56127, Italy*

ARTICLE INFO

Article history:
Received 16 December 2015
Revised 4 June 2016
Accepted 13 June 2016
Available online 30 June 2016

关键词
大数据
智慧城市
能源
工作机会
隐私

摘要

这篇文章的目的在于综述在大数据背景下一些意大利大学正在从事的研究项目。本文不求面面俱到，目的是提供从意大利不同领域收集到的有关大数据管理方面的问题的实际解决方案。

© 2016 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. 引言

在过去的几年中，无论是在研究中心、学院还是行业，与大数据相关的计划、活动和项目激增。每天产生的与社会生活各个方面相关的数据(包括手机数据、社会数据、城市相关数据、基于网络的数据和健康相关数据)为观察和了解人们的喜好和行为，以及利用这些信息以改善人们生活的某些方面，提供了前所未有的机会。

针对这个颠覆性的变化——一个开辟了新的经济学领域的变化——欧洲委员会要求各国政府要意识到这场"大数据"革命[†]。与美国相比，欧洲的数字经济确实在接受数据革命方面进展较慢，而且也缺乏可比的工业能

* Corresponding author.
E-mail address: donato.malerb@uniba.it; pasi@disco.unimib.it

† "Towards a thriving data-driven economy," communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee, and the Committee of the Regions, Brussels, 2014 Jul 2.

2095-8099/© 2016 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
英文原文: Engineering 2016, 2(2): 163–170
引用本文: Sonia Bergamaschi, Emanuele Carlini, Michelangelo Ceci, Barbara Furletti, Fosca Giannotti, Donato Malerba, Mario Mezzanzanica, Anna Monreale, Gabriella Pasi,Dino Pedreschi, Raffele Perego, Salvatore Ruggieri. Big data research in Italy: A perspective. *Engineering*, http://dx.doi.org/10.1016/J.ENG.2016.02.011

力。为了摆脱这种落后的局面，一笔数量可观的基金已经建立起来，并将由欧盟委员会以及欧洲各国提供用以支持由大数据产生的价值相关的研究和创新活动。为了妥善地达到这个目标，必须考虑各种问题，包括功能强大而技术上满足支持密集的数据驱动计算的设施的确定(包括硬件和软件)，以及建立能从不同领域的数据中正确并高效地提取知识的多学科团队。

尽管起步较晚，欧洲的大数据市场占据了第二大市场份额，占到了全球大数据市场收入的20%[1]。德国、英国、法国和意大利是这个市场的核心国家。尤其是意大利的大数据市场在去年增长迅猛，并且预计不久会有私人和公共部门的重要投资。这篇简短的调查报道了意大利大学在应对大数据挑战时所做的一些应用和项目，重点报道了与提高城市居民生活相关的项目。作为与大数据管理相关的技术应用的有趣案例，第4部分描述了以监测能源生产和消耗为目的的系统；与此同时，第5部分全面地介绍了一个以分析从5个欧盟国家搜集到的网络空缺职位及从数据中提取职位所需要的技能为目的的模型。

值得注意的是，这篇调查并没有面面俱到，无论是与之相关的活跃的研究小组(这种小组的数量远比文中提到的要多)还是项目，在意大利都远远多于调查中所报道的这些。这篇调查的主要目的在于给读者提供一些关于目前学术界正在解决这一重大问题的思考。

最近一项关于大数据的国家倡议是以CINI大数据实验室为代表提出。CINI(www.consorzio-cini.it)是意大利国家校际信息联盟，由41所意大利公立大学组成，致力于促进和协调在计算机科学和计算机工程内多个领域的、涵盖理论和应用的科学研究活动和技术转让。这个联盟是大数据价值协会(www.bdva.eu)的创始成员。大数据价值协会是欧盟委员会在制定和实施欧洲大数据战略研究议程的行业领袖型的契约伙伴。此外，CINI大数据实验室关注数据在全国领土内的分布，致力于成为意大利大数据及数据科学领域知识和技术发展的专家中心。33所意大利大学和将近300位研究人员目前已经加入这项倡议。

第2部分简要介绍了不同的意大利大学和研究机构在解决大数据问题和致力于提高人们生活的各个方面所做的一些项目。这些项目与不同应用领域息息相关，包括了解城市动态、意大利医疗保健系统、预测光伏发电厂的能源产量和管理工作机会。最后两个部分分别论述了隐私和大数据可用性这两个重要议题。

2. 通过手机数据了解人和城市动态

城市一直都是一个由人、物、环境和活动组成的复杂系统，他(它)们的快速演变也不可避免地增加了复杂性。这个事实也促使科学家摈弃了传统的模型驱动分析范式，而更热衷于数据驱动的方法，开启了大数据分析时代。人们每天通过与设备、社交媒体和其他技术系统交互产生的数字信号，为从多种角度研究和理解城市动态和社会行为提供了前所未有的机会。理解这些动态意味着可以预见现象所造成的影响，并帮助政策和规划者回应公民的需求。

事实上，手机数据可用于研究和度量城市以及市民的位置，它让我们在城市尺度上分辨出人们的位置[2–4]，重现他们的流动性[5–8]和社会性[9]，并研究城市中这些事件的影响[4,10]。

2.1. 手机和出发地–目的地 (OD) 矩阵估计

对人们在老地方之间的位置和流动进行评估可用于重建一个出发地–目的地(OD)矩阵[5,6]，这有助于推断出交通需求以及理解人们对基础设施的要求。在文献[6]中，作者对个人通话轨迹进行了长期的分析，在两个最重要的地点之间重建具有规律性的活动(即高频率活动)。这样的地点是通过分析一个人在哪些地点打了最多的电话所确定的，通常与家和工作联系在一起。在确定了这些地点之间的规律性活动之后，OD矩阵总结了空间区域之间的预期交通流。

2.2. 手机用于新人口统计学和城市用户估计

估计和监测社会现象的可能性增加了人们使用大数据支持官方统计的兴趣[5]。由于管理类数据无法高频次地搜集，并且往往不包含准确的移动信息，因此，通话数据正在被越来越多地用于整合传统数据源，如用于建设城市的长期观测数据库[3]和确定城市用户的实际类型。在文献[2]和[4]中提出了社会性测量量表——一个旨在将手机用户分为不同行为类别的分析框架。分析过程从构建时空分布开始，整合了人在所有兴趣地的定位。然后通过运用一种数据挖掘的方法，了解不同人的类别，这样属于居民、动态居民、通勤者和游客的注释简介就产生了。在文献[5]中，从社会性测量量表的结果出发，创建了一个在直辖市尺度下的OD矩阵，目的是用于观测城市内部的人的流动。这种统计数据可以与国家统计局(意大利)的统计数据相媲美，它为整合现有

人口和由手机数据实时估算出的流动统计数据提供了一种安全途径。

2.3. 手机、流动多样性和经济发展

当需要通过调查社会现状以改善生活条件时，研究就会变得越来越具有挑战性。在文献[8]中，作者从全国手机数据中提取了一种分析每个人的流动多样性和流动数量的方法，并研究了外部社会经济指标之间的相关性。多样性被定义为一个用户轨迹的熵，流动的量是通过测量一个人移动的特征距离所得到的。实验表明，流动性是和幸福指标(如教育水平、失业率、收入和免职)相关的，证明了流动行为可以很好地预测城市的社会经济发展。在另外一项社会层面的探索中，文献[9]通过比较流动性和从电话中提取的社会网络得到了一个很有趣的结论。运动中的相似性和社会网络中的临近性呈现出很强的关联，从中可以得出那些在网络中不连接的人但在拓扑上紧密并有相似移动模式的人，很有可能在未来建立起社会联系。

2.4. 手机和大事件预测

监测并记录人们对于大事件的反应的可能性引起了公共管理学的极大兴趣[10]。类似的研究可以应用在对城市事件的影响的分析上[4]，以设计出针对安全和流动性的充足预案。文献[10]提出了相关性模式分析——一个提取了由城市中各种事件引起的不同区域之间的内在关系的过程。通过在同一尺度上分析通话的密度，人的数量就可以被估计，并且利用时间和空间约束的序列模式就可以分析出人的数量的显著协变量。文献[4]提出了一种在城市尺度通过由社会性测量量表得出的人口分析结果衡量事件(如节日、音乐和艺术表演以及季节性活动)影响力的方法。通过统计学方法和多分类分析，对一个兴趣区域内和在一个特定时间范围内城市用户构成的变化进行了分析。多分类分析使我们发现当将分析从一个小区域(如城市历史中心)移动到大区域(如城市郊区)时人口的构成是如何变化的。实验证明，社会性测量量表可以在城市尺度确定人的构成，并且整套方法对于测量大型活动在中小城市中的影响力是有效的。

2.5. 人口流动性的手机模型

由于这些数据具有无处不在的特性并不断扩散，人口动力学出现了新的特征。在将近10年被视为随机游走或者Lévy flight(列维飞行)的人口流动，如今已经显现出了一种不排除模式上异质性的高度的时空规律。通过研究手机轨迹的回旋半径，发现人们只在少数地点度过大量的时间。这个结果允许科学家更深入地研究流动性，并发现个人的移动距离特征的多变性也意味着其未来位置的可预测程度高。这个明显的矛盾可以通过进一步分析系统性的运动来解释，研究发现两种新的旅行者类型：回归者和探索者。回归者的系统性流动是通过他们的回旋半径估计的，特征是反复在几个首选地点之间移动。而探索者则倾向于在更多的不同地点间移动，并且他们的系统性流动对于整体流动的贡献很小[7]。

3. 意大利医疗保健大数据案例

基于由大量医疗数据提取出的医学知识证据，标准的医疗保健体系正在逐步地建立。在所有发达国家，医疗保健供应商收集并管理了大量复杂、种类不同的数据。大量的可用数据实际上保证了医疗保健体系不断完善的可能，涵盖了个性化医疗、疾病预防和有效的医疗保健组织[11]。然而，由于大量患者从不同供应商获得医疗保健，造成了电子健康数据在众多组织中碎片化地传播。因此，整合和协调这些数据，变得越来越重要。

在此背景下，由于其普遍的覆盖范围和区域管理，意大利以税收为基础的公共医疗保健系统面临着独特的挑战。意大利是世界上人口老龄化最严重的国家之一，为了防止并发症和残疾(以确保国家经济的可持续性)，对于患者而言，有效地管理慢性疾病[12]是最重要的。

意大利的医疗保健系统的组织是分层并分散的。国家层面负责确定医疗保健系统的总目标和基本原则。在另一方面，区域政府(共21个)负责通过基层医疗卫生单位(LHU，平均每个区域10所)提供医疗保健服务。由于这些医疗卫生单位分散独立，因此医疗保健数据管理系统不可互操作。在这种背景下，国家区域医疗服务机构(AGENAS——协调区域医疗保健系统的国家机构)，与托斯卡纳的区域健康机构(ARS)和意大利国家研究委员会合作，共同开发大数据分析平台，致力于提供对区域单位管理的电子健康记录做统一分析的工具。

THEMATRIX

THEMATRIX平台支持全生命周期的所有大数据分析，包括从分布式数据采集、存储到设计和分析的并行部署及结果的展示。平台允许通过支持公共数据的提取和重新映射，实现区域信息系统极端多样化的隐藏。这

一过程通过记录每位市民和公共医疗保障系统的所有交互来实现。

数据模型和数据存储技术所在区域层面的多样性和异质性集中构成了数据收集工作面临的挑战。尽管数据模型在国家层面已经很普遍，但很少或者根本没有应用于地区层面。此外，LHU在选择数据管理技术方面有绝对的自由度，导致了数据存储服务和访问接口(包括开源安装到全面企业数据库)方案的不断繁殖。正在开发的THEMATRIX平台，从LHU收集到了长期的数据，并将这些数据以共同的格式管理以供综合研究。数据收集的一个重要方面就是数据的匿名性。事实上，当电子健康数据被管理时，隐私就是最引人关注的问题之一[11]。我们的数据收集机制根据由国家隐私局(National Authority of Privacy)设立的指导方针将病患记录做了匿名处理。在隐藏患者身份的同时，数据的模糊性允许执行非常有用的跨区域分析。这个过程在地区层面实施并在国家层面推行，个人身份信息的隐藏并没有削弱国家层面进行分析获取的价值。

数据分析界面让流行病学家能灵活地访问数据，还为他们提供了可以用来定义提取(基于一定规则的算法)到的信息的图形界面。数据转换和分析开发了一种灵活的特定领域语言，为研究区域内部和全国范围关于人口健康和疾病状态的模式、原因和影响提供了可能。可编程计算引擎将运算工作组织为一个有向非循环图(DAG)，其中每一个节点代表一个要应用在患者记录流(stream of patient records)上的任务。

先前对选定飞行员进行的研究都集中在识别和预测少数几种慢性病，如糖尿病或心血管疾病。由于大量数据可用，用来计算这些情况的效率和有效性的关键绩效指标允许区域公共医疗保健系统可以在比较客观的基础上进行比较，预测算法的质量也得以加强。到目前为止，该平台已经在意大利全国试点LHU内和两个区域机构内部署并测试。可用的数据包括4年内700万居民的行政记录。为了完善基于医疗保健数据的预测模型，60万患者已经与特定患者的健康状况做了匿名匹配，并由在意大利负责调解患者和公共医疗保健系统关系的主要保健医生进行评估。明年，计划在至少10个区域(覆盖一半以上意大利人口)部署THEMATRIX分析解决方案的项目。为了支持这个国家级的大型数据分析挑战，DAG计算的并行化将会增强。其要求是提供一套针对LHU硬件异质性特点的灵活而有效的开发，硬件类型涵盖了从低规格商品机器到大企业集群支持的Apache Spark和Hadoop。

4. 能源大数据

减少污染排放这一迫切需求使得可再生能源成为一个战略领域[13]，尤其是对欧盟而言。这导致了可再生能源的涌现以及具有重要意义的分布式发电的产生。这个新能源市场面临的主要挑战有网格集成、负载均衡和能源交易。首先，将这样的分布式可再生能源整合到电网中，同时要避免依赖降低和配电损失，是一项艰巨的任务。事实上，可再生能源，如光伏数组，在它们的能源输出中是可变的和间断的，因为产生的能量也可能取决于一些不可控因素，如天气情况[14]。其次，能源市场上的主角——在供应链中扮演供求双方的分销商和一些小公司，在为他们的顾客计划能源供给时，也不得不面对需求和供给的不确定性。再次，由单一来源(尤其是从可再生能源)生产的能源有助于确定每天或者每小时市场的最终结算价格[15]，这使得能源市场非常有竞争性，对于局外人而言犹如迷宫。

为了应对这样的挑战，在地区以及全球层面监测能源的生产和消费，储存历史数据，并设计新的可靠的预测工具具有极其重要的地位。虚拟电源运行中心(Vi-POC)项目致力于原型的设计和实施，以便达成该目标[16,17]。由于数据量庞大且具有异质性，为了能高效地访问这些不能通过传统数据管理手段获得的数据，利用合适的大数据分析技术是非常有必要的。然后，由于新的(低成本)技术的可用性，小的供应商也能够收集到关于他们自己业务的数据。事实上，从小的发电厂获得的数据是相当异质的，这些源源不断的数据持续快速地增长。这些数据按照持续的(快速的)频率到达并且数量不断增长。此外，为了考虑不可控制的因素，如天气条件，有必要储存来自于气象服务组织的天气观测信息和预测信息(如温度、湿度、风速等)。

从这个角度看，开发Vi-POC项目为(可再生)能源供应商提供了一个数据收集、储存、分析、查询和检索框架，而这些数据来自于广泛分布的多样化的发电厂(如光伏、风、地热、斯特林发动机和自来水)。此外，Vi-POC项目的一大特点便是整合发电厂数据和天气服务数据的产能实时预测系统。

Vi-POC项目设计了一个用于存储天气信息数据和工厂传感器数据的HBase存储系统。通过客户端运行数据挖掘算法，这些数据可以用来预测工厂未来24/48小时内的产能。每个发电厂都会定时发送传感器搜集的数据。时间的间隔是基于发电厂的类型和容积所设定的。

由于是在给定的时间内通过多个传感器搜集，从发电厂搜集到的数据往往包含了不同的测量方式。事实上，发电厂之间传感器的数量和型号可能存在差异。另外，预测的数据包含基于给定时间和地点的多种天气预测参数。

对于(可再生)能源的预测，在文献中，已经提出了几种数据挖掘的方法。研究人员通常在两类方法之间进行区分：物理的和统计的。前者依靠基于物理考量(障碍物和山志学)[18]和测量数据(模型输出统计方法或者MOS)[19]的天气预测数值的优化；而后者则基于模型建立历史数值和预测变量之间的关系。

尽管已经存在应用于可再生能源预测中学习自适应模型的数据挖掘算法[15,20]，但是在时空信息、学习环境和使用算法的考量上，仍未达成一致。Vi-POC框架中应用的预测模型包括以下分析。

(1) 时空自相关性[21]：地球物理现象具有这样的特性能获得更准确的预测。空间自相关性是通过对两个空间统计数据进行分析，即空间联系的局部指标(LISA)和主轴邻距法(PCNM)，而时间自相关性是通过分析不同形式的空间统计。

(2) 学习环境：这个可通过每小时使用一个简单的输出预测或者使用一个结构化的输出预测模型(即一个24个元素的向量对应第二天的24个小时)。

(3) 学习算法：在学习自适应模型方面，我们将常被用作预测光伏发电的人工神经网络，与回归树和*k*近邻算法(或应用于Apache Spark框架中的最短*k*-NN[22])相比。两组数据得出的结果显示考虑时空自相关性是有益的。

然而，最重要的方面是学习环境：结构性的输出预测设置在很大幅度上优于非结构性的输出预测设置。最终，结果显示回归树能提供比人工神经网络和*k*-NN预测模型更好的模型。

5. 工作供给和大数据

通过专门的网络劳动力市场端口和服务招聘的空缺职位在过去几年内急速增长，使得招聘(也被称为“e招聘”)和劳动力市场分析(也被称为劳动力市场智能)有了新方法。通俗地说，一个网络空缺职位可以被看作一段在不同网站资源上刊登多次的原始文本，内容详述了职务名称和一段长度不限的介绍，往往包含了一个应聘者需要的技能。正如人们所想的，大量数据的收集、净化、归类和推理，对于公共和私营的劳动力市场运营者而言都是非常值得关注的，应该考虑从不同观点(如领土面积、新兴职业和技能)来描述劳动力市场的趋势和动态。在这样的背景下，欧盟一直在努力定义一个国际技能/职业分类系统(如ESCO[†])，这将为跨国和跨语言研究劳动力市场动态的劳动力市场分析师和政策制定者提供一种通用语。

在2015年，CRISP-UNIMIB[‡]与米兰比可卡大学(UNIMIB)计算机科学系统和通信学院中的信息检索实验室(IR-Lab)协作，开始研究由Cedefop[††]赞助的欧洲项目，该项目的目的是构建一个(系统)原型，用以分析在5个欧盟国家网络上发布的空缺职位以及必要的技能。项目背后的原理就是提取网络上发布的空缺职位数据，将其转化为支持劳动力市场智慧的知识(从而价值)。为此，著名的数据库知识挖掘过程(KDD)[23]已经被用作一种方法论框架。事实上，除了这个项目与全欧洲劳动力市场监测系统之间的关系，它还体现了在大数据全景下的一些有趣的方面，因为它需要处理大数据背景下的4个“V”：数据的“量”(例如，随着时间的推移所搜集到的空缺职位的数量不断增加)，通过哪家招聘平台发布最新的和之前的空缺职位的“速度”，每个网络资源(如半结构化和非结构化数据)的不同数据特性的“多样性”，以及“准确性”，由于在多个资源中存在重复的职位空缺，或者需要被识别和处理的缺失信息。在下面的讨论中，提供了一个过程的概述，突出了每一步中所针对的“V”以及所用的技术。

在数据来源选择这个步骤(准确性)，根据领域内的专家们提供的质量标准(如更新后的职位和领土的粒度)对70个网络数据来源进行排名。在数据收集步骤(数量、速度、多样性和准确性)建立了由3个不同组件组成的一个抓取模块，这3个组件分别是：①一个检索网页的下载器；②一个识别空缺职位主要(招聘)要求的提取器，并将这些要素储存在一个数据库中；③一个周期性地计划和执行所有抓取过程的监视器。这个模块已在内部建成，通过使用Spring框架和Talend任务流程来处理网络资源的高异质性。在3个月中，已经搜集到了欧洲

[†] ESCO is the multilingual classification of European Skills, Competences, Qualifications, and Occupations built on top of the International Standard Classification System (ISCO). ESCO is part of the Europe 2020 strategy.
[‡] The Interuniversity Research Centre on Public Services–University of Milan-Bicocca.
[††] The European Center for the Development of Vocational Training.

5个国家的400万个空缺职位。数据的清洗和归类任务(数量、多样性、准确性)负责识别重复的空缺职位信息，并根据ESCO职业分类(大约436种职业项目)对其进行分类。注意到信息清洗是个曲折的过程，因为它可能会影响到随后步骤的可信度(见文献[24-26])。为此，使用了机器学习算法，因为它优于在一个领域相关基准[27]中的其他方法，并能在项目设置中达到高级别的分类精度(如从德国的79%最高可达捷克的98%)。

分类模块是使用SCiPy框架的自定义代码构建的。技能提取任务(数量、多样性和准确性)负责使用语言模型从空缺职位的描述中提取技能。根据ESCO职业分类标准中的数据分类就会被雇主所要求的技能信息所丰富，这样就能详细地描述一个通过网络发布工作机会。

最后，使用著名的D3.js可视化库对几个可视化模型进行了识别。这个过程中的一个终端产品(只关注意大利劳动力市场数据)的例子就是WollyBI[†]。

总而言之，这个项目揭示了应用智能化技术和数据工程来应对在一个真实和特定领域背景下的大数据问题。研究结果为今后的工作铺设了这几条道路：首先，根据雇主的技能要求自动对相似的职业进行归类；其次，基于图模型展示所搜集到的知识，这对于一个包含了所有空缺职位(数以百万计的节点)的一个大型而且高度动态的知识库而言是一个自然且便捷的选择。在项目部署之后，一个关于欧洲一些主要国家的网络劳动力市场数据就会被搜集。这个具有高价值的知识库将有利于劳动力市场智能领域的研究活动。

6. 大数据分析中的隐私和道德

源于人类活动细枝末节的大数据，作为我们每天使用的通信技术(ICT)系统的副产品，记录着社会生活的多重维度：自动付款系统记录了我们的消费轨迹；搜索引擎记录了我们在网络上的查询日志；无线网络和手机设备则记录了我们的移动轨迹。这些描述人类活动的大数据就在一个虚拟化的“知识社会”的中心，其中对社会现象的认识就是通过社会挖掘技术不断地从多社会维度的大数据中提取知识。因此，人类的数字轨迹的分析为理解复杂面创造了新的机会，例如，流动行为、经济和金融危机、流行病的蔓延和意见的扩散。然而，在数据处理和分析中的伦理问题的高风险，以及建议和预测所带来的伦理后果，要远远高于从这些数据中发现有趣模式的机会。几个重要的伦理风险包括：①隐私侵犯，发生在无约束地侵入研究对象的个人资料时；②歧视，当被发现的信息不公平地用于制定针对某一类人(可能这些人不知道)的歧视性的决定时。

然而，大数据和道德并非天敌。在文献中，一些研究已经表明许多基于大数据分析的实践和应用，可以被设计成一种与道德要求共存的高品质结果的形式。其秘诀是设计开发执行伦理价值要求的大数据分析技术，为公平提供保障。

在大数据分析隐私权保护的背景下，Monreale等[28]推荐最早由Ann Cavoukian在20世纪90年代提出的隐私设计范式的实例化——一种大数据分析服务设计。这个方法论在以下领域被应用以保障隐私。

6.1. 数据发布中的隐私

Monreale等[29]设计了一个移动数据发布的隐私保护方法，它使聚类分析用于理解人们在明确的城市区域内的流动性行为。发布的轨迹是通过一个适当的过程实现匿名，实现了原始轨迹的一个广义版本。通过应用这个框架所获得的结果显示轨迹是如何通过匿名化，达到高级别的保护防止再度被识别，同时保存了挖掘轨迹集群的可能性。这个方法使得更新、更强大的信息流动分析服务或定位服务成为可能。

6.2. 数据挖掘外包中的隐私

Giannotti等[30]设计一个在模式挖掘任务外包中的隐私保护方法。尤其是，结果显示了一个公司是如何将交易数据外包给一个第三方，并以保护隐私的方式获得数据挖掘的服务。在此设置中，不仅仅是基础数据，挖掘结果(战略信息)也不会共享，并且必须保留隐私。在参考文献[27]中所提出的隐私解决方案包括了应用一种加密体系通过以下步骤改造原来的数据库：①通过一个1-1替代功能取代原有数据库中的每一条目；②通过对数据库进行假的转换的方法，即每个条目变得至少和其他所有条目(k–1)是没有区别的。基于这个简单的思想，这个框架保证了不仅仅是个体条目，而且任何一组条目，都能在最坏的情况下与至少其他k组不被区分，而且事实上平均来看不止k组。这种保护意味着攻击者有更小的概率在交易数据或者挖掘结果中猜到包含在其中

[†] http://www.wollybi.com/en/

的真实内容。与之相反，数据所有者能够利用有限的计算资源有效地解密由第三方机构返还的正确的挖掘结果。

6.3. 分布式分析系统中的隐私

Monreale等[31]提出了一种分布式移动数据分析中的隐私保护方法，主要针对当一个不被信任的中央站收集了一些基于每个节点观察到的移动数据流所计算得出的汇总统计的情况。这个中央站储存收集到的统计信息，并基于从数据采集器中收集到的信息计算所有领土内的交通概况。提出的框架通过应用一个知名的隐私模型——“差异隐私”，从而保证能在个人层面保护隐私。尤其是，隐私技术能在节点的移动数据发送到不被信任的中央站前就对其进行扰乱。

6.4. 从数据中发现的歧视及其预防

在分析歧视数据的背景下，主要分为两个研究方向(见参考文献[32]中的一项调查)。从数据中发现的歧视存在于真实发现的歧视现象以及隐藏在大量历史决策记录中的做法。最初被提出的是一个关于社会组织使用分类规则进行挖掘和过滤的直接或间接歧视的过程。这个过程是由以法律为基础的歧视评估作为指导，可能包括置信度的统计检验[32]。个体歧视被一个k-NN方法模型所取代，并应用于一个研究项目资金的真实案例研究中[33]。

歧视预防包括了从可能导致预测模型作出(可能是自发的)歧视决定的训练数据和学习算法中消除偏见。参考文献[34]研究了防止歧视的数据净化，首先将隐私的t-closeness模型变弱到一个非歧视模型，然后通过使用最先进的数据净化方法处理t-closeness。一个能同时处理隐私和歧视净化的方法见参考文献[35]。关于学习算法，有人提出了一种改良的基于规则分类器的投票机制以减少可能的歧视性规则的权重[32]。

7. 强化大数据的可用性

7.1. 大数据的实体解析

网络已经成为结构化和半机构化数据的宝贵来源。大量的高质量关系数据可以从HTML表格中提取[36]，并且随着网络数据的出现，大量作为链接数据的公开半结构化数据呈指数性增长[37]。这些数据以数量大、品种多、变化快为特点，但与此同时，它们的准确性和质量也常常是个问题[38,39]。基于以上这些原因，这样的数据常常被认为是“大数据”。数据真正的潜能往往体现在整合不同来源数据的时候，最近在网络挖掘中提取实体、关系和本体以建立大型通用知识库就可以作为论证，如Freebase和Yago[40]。对于企业、政府机构和在大型科研项目的研究人员，如果能与他们已经拥有的受限于传统数据集成过程的数据相结合，这些数据甚至可以更有价值。

能够识别指向同一实体的记录是使这些数据有意义的基础步骤。一般来说，为了能够实现实体解析(ER)，传统的技术要求在数据源之间有一个模式对齐。不幸的是，大数据典型的特点就是高异质性、高噪声和非常大的数据量，造成传统的模式对齐技术不再适用。例如，谷歌基地包含超过1万个用10万个独特的图形描述的实体类型；在这样的情况下，执行和维护一个模式对齐是行不通的[41]。

最近，已经提出了两种技术以解决这些问题：①放弃挖掘模式信息并完全依赖于冗余来限制错误匹配机会的技术[42-44]；②从数据中直接提取模糊模式信息，不执行传统的模式对齐的技术，这对于ER是有用的[45]。后者的结果是最有保障的，但被研究得最少。事实上，遵循他们建议的方向，是有可能让基于模式的ER技术支持大数据的，既保证了高查全率和精度，又不需要执行不堪忍受的传统的模式对齐步骤。

7.2. 大数据的探索

在大数据时代，新的用户界面需要与我们收集到的大量数据进行交互；否则，用户将被数据淹没。在参考文献[46]中，提出了一种解决方案，可以帮助用户将他们的注意力集中在一个小组相关的数据，使用贝叶斯方法推断用户的选择。在我们的试验中，我们研究了一种利用在大数据背景中的用户输入推断相关信息的方法。

贝叶斯网络增强的面浏览[47]，通过分析用户的选择作为概率模型，为用户推断有价值的信息。面浏览是一种通过多个步骤应用动态过滤器对数据进行探索的技术：每次使用一个过滤器，结果就会显示给用户，用户还可以应用额外的过滤器或者调整现有的。在每一步骤中，显示的过滤器和过滤器中的数值可能是不同的。

所提出的方法对于在大数据环境中探索数据是有效的，即属性的个数和值都很巨大。换而言之，面浏览所提供的优势就是过滤器的动态性。此外，对于用户为了能动态地获得最有用的过滤器，有必要利用用户目前的

选择进行推断。因此，通过利用图形贝叶斯网络概率模型对用户的选择进行分析，有可能推断出对于他们最有价值的过滤器。图模型是首选，主要因为它们易于理解、验证和解释结果。在这样的背景下，贝叶斯网络中的变量就是数据集的属性。贝叶斯网络被用于推测这些属性之间的关系，用于计算一个用户的选择和其他网络中属性之间相关性的概率，然后将最相关的属性展现为过滤器。此外，为了避免显示太多的值，可以在过滤器中推断出相似和不相似的值。为了总结过程，只有5个最相似和最不相似的值会展示给用户。

8. 结论

本文展示了一些在意大利进行的有关大数据的学术研究活动，内容涵盖了旨在提高人们生活的多个方面的应用，以及两个普遍性的重要问题——隐私和大数据的可用性。文中展现了一个多产的学术研究界，已经能够面对大数据目前在数量、速率、多样性和准确性上所带来的挑战。下一阶段是与行业更紧密地合作，共同面对最实质性的挑战：从大数据中创造价值。从这个意义上讲，参与实施由大数据价值协会制定的欧洲战略研究议程，以及从CINI“大数据”实验室获得支持，将会是至关重要的。

Compliance with ethics guidelines

Sonia Bergamaschi, Emanuele Carlini, Michelangelo Ceci, Barbara Furletti, Fosca Giannotti, Donato Malerba, Mario Mezzanzanica, Anna Monreale, Gabriella Pasi, Dino Pedreschi, Raffele Perego, and Salvatore Ruggieri declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Europe Big Data market 2015–2020 [Internet]. New York: PR Newswire Association LLC.; c2016 [updated 2016 May 30, cited 2016 Jun 12]. Available from: http://www.prnewswire.com/news-releases/europe-big-data-market-2015---2020-300276656.html.

[2] Furletti B, Gabrielli L, Renso C, Rinzivillo S. Analysis of GSM calls data for understanding user mobility behavior. In: Hu X, Lin TY, Raghavan V, Wah B, Baeza-Yates R, Fox G, et al., editors Proceedings of the 2013 IEEE International Conference on Big Data; 2013 Oct 6–9; Santa Clara, CA, USA; 2013. p. 550–5.

[3] Furletti B, Gabrielli L, Renso C, Rinzivillo S. Pisa tourism fluxes observatory: deriving mobility indicators from GSM call habits. In: Proceedings of the 3rd International Conference on the Analysis of Mobile Phone Datasets; 2013 May 1–3; Cambridge, MA, USA; 2013.

[4] Gabrielli L, Furletti B, Trasarti R, Giannotti F, Pedreschi D. City users' classification with mobile phone data. In: Ho H, Ooi BC, Zaki MJ, Hu X, Haas L, Kumar V, et al., editors Proceedings of the 2015 IEEE International Conference on Big Data; 2015 Oct 29–Nov 1; Santa Clara, CA, USA; 2015. p. 1007–12.

[5] Furletti B, Gabrielli L, Giannotti F, Milli L, Nanni M, Pedreschi D. Use of mobile phone data to estimate mobility flows. Measuring urban population and inter-city mobility using big data in an integrated approach. In: Proceedings of the 47th SIS Scientific Meeting of the Italian Statistical Society; 2014 Jun 11–13; Cagliari, Italy; 2014.

[6] Nanni M, Trasarti R, Furletti B, Gabrielli L,Van Der Mede P, De Bruijn J, et al. Transportation planning based on GSM traces: a case study on ivory coast. In: Nin J, Villatoro D, editors Citizen in sensor networks. Cham: Springer International Publishing; 2014. p. 15–25.

[7] Pappalardo L, Simini F, Rinzivillo S, Pedreschi D, Giannotti F, Barabási AL. Returners and explorers dichotomy in human mobility. Nat Commun 2015;6:8166.

[8] Pappalardo L, Pedreschi D, Smoreda Z, Giannotti F. Using big data to study the link between human mobility and socio-economic development. In: Ho H, Ooi BC, Zaki MJ, Hu X, Haas L, Kumar V, et al., editors Proceedings of the 2015 IEEE International Conference on Big Data; 2015 Oct 29–Nov 1; Santa Clara, CA, USA; 2015. p. 871–8.

[9] Wang D, Pedreschi D, Song C, Giannotti F, Barabási AL. Human mobility, social ties, and link prediction. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2011 Aug 21–24; San Diego, CA, USA; 2011. p. 1100–8.

[10] Trasarti R, Olteanu-Raimond AM., Nanni M, Couronné T, Furletti B, Giannotti F, et al. Discovering urban and country dynamics from mobile phone data with spatial correlation patterns. Telecommu Policy 2015;39(3–4):347–62.

[11] Liu W, Park EK. Big data as an e-health service. In: Proceedings of the 2014 IEEE International Conference on Computing, Networking and Communications; 2014 Feb 3–6; Honolulu, HI, USA; 2014. p. 982–8.

[12] Gini R, Francesconi P, Mazzaglia G, Cricelli I, Pasqua A, Gallina P, et al. Chronic disease prevalence from Italian administrative databases in the VALORE project: a validation through comparison of population estimates with general practice databases and national survey. BMC Public Health 2013;13(1):15.

[13] Directive 2009/28/EC of the European Parliament and of the Council on the promotion of the use of energy from renewable sources and amending and subsequently repealing Directives 2001/77/EC and 2003/30/EC. Official Journal of the European Union L 140; 2009 Jun 5. p. 16–47.

[14] Ioakimidis CS, Oliveira LJ, Genikomsakis KN. Wind power forecasting in a residential location as part of the energy box management decision tool. IEEE Trans Ind Inform 2014;10(4):2103–11.

[15] Bessa RJ, Miranda V, Gama J. Entropy and correntropy against minimum square error in offline and online three-day ahead wind power forecasting. IEEE Trans Power Syst 2009;24(4):1657–66.

[16] Ceci M, Cassavia N, Corizzo R, Dicosta P, Malerba D, Maria G, et al. Innovative power operating center management exploiting big data techniques. In: Proceedings of the 18th International Database Engineering & Applications Symposium; 2014 Jul 7–9; Porto, Portugal. New York: ACM; 2014. p. 326–9.

[17] Ceci M, Corizzo R, Fumarola F, Ianni M, Malerba D, Maria G, et al. Big data techniques for supporting accurate predictions of energy production from renewable sources. In: Proceedings of the 19th International Database Engineering and Applications Symposium; 2015 Jul 13–15; Yokohama, Japan New York: ACM; 2015. p. 62–71.

[18] Bofinger S, Heilscher G. Solar electricity forecast—approaches and first results. In: Proceedings of the 21st European Photovoltaic Solar Energy Conference; 2006 Sep 4–8; Dresden, Germany; 2006. p. 4–8.

[19] Pelland S, Galanis G, Kallos G. Solar and photovoltaic forecasting through post-processing of the Global Environmental Multiscale numerical weather prediction model. Prog Photovoltaics 2013;21(3):284–96.

[20] Sharma N, Sharma P, Irwin DE, Shenoy PJ. Predicting solar generation from weather forecasts using machine learning. In: Proceedings of the 2011 IEEE International Conference on Smart Grid Communications; 2011 Oct 17–20; Brussels, Belgium; 2011. p. 528–33.

[21] Stojanova D, Ceci M, Appice A, Džeroski S. Network regression with predictive clustering trees. Data Min Knowl Disc 2012;25(2):378–413.

[22] Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I. Spark: cluster computing with working sets. In: Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing; 2010 Jun 22–25; Boston, MA, USA. Berkeley: USENIX Association; 2010. p. 1765–73.

[23] Fayyad U, Piatetsky-Shapiro G, Smyth P. The KDD process for extracting useful knowledge from volumes of data. Commun ACM 1996;39(11):27–34.

[24] Boselli R, Cesarini M, Mercorio F, Mezzanzanica M. Planning meets data cleansing. In: Proceedings of the 24th International Conference on Automated Planning and Scheduling; 2014 Jun 21–26; Portsmouth, NH, USA; 2014. p. 439–43.

[25] Mezzanzanica M, Boselli R, Cesarini M, Mercorio F. Data quality sensitivity analysis on aggregate indicators. In: Helfert M, Francalanci C, Felipe J, editors Proceedings of the International Conference on Data Technologies and Applications; 2012 Jul 25–27; Rome, Italy; 2012. p. 97–108.

[26] Mezzanzanica M, Boselli R, Cesarini M, Mercorio F. A model-based evaluation of data quality activities in KDD. Inform Process Manag 2015;51(2):144–66.

[27] Amato F, Boselli R, Cesarini M, Mercorio F, Mezzanzanica M, Moscato V, et al. Challenge: processing web texts for classifying job offers. In: Kankanhalli MS, Li T, Wang W, editors Proceedings of the 2015 IEEE International Conference on Semantic Computing; 2015 Feb 7–9; Anaheim, CA, USA; 2015. p. 460–3.

[28] Monreale A, Rinzivillo S, Pratesi F, Giannotti F, Pedreschi D. Privacy-by-design in big data analytics and social mining. EPJ Data Sci 2014;3(1):10.

[29] Monreale A, Andrienko G, Andrienko NV, Giannotti F, Pedreschi D, Rinzivillo S, et al. Movement data anonymity through generalization. Trans Data Privacy 2010;3(2):91–121.
[30] Giannotti F, Lakshmanan LVS, Monreale A, Pedreschi D, Wang H. Privacy-preserving mining of association rules from outsourced transaction databases. IEEE Syst J 2013;7(3):385–95.
[31] Monreale A, Wang WH, Pratesi F, Rinzivillo S, Pedreschi D, Andrienko G, et al. Privacy-preserving distributed movement data aggregation. In: Vandenbroucke D, Bucher B, Crompvoets J, editors Geographic information science at the heart of Europe. Cham: Springer International Publishing; 2013. p. 225–45.
[32] Romei A, Ruggieri S. A multidisciplinary survey on discrimination analysis. Knowl Eng Rev 2014;29(5):582–638.
[33] Romei A, Ruggieri S, Turini F. Discrimination discovery in scientific project evaluation: a case study. Expert Syst Appl 2013;40(15):6064–79.
[34] Ruggieri S. Using *t*-closeness anonymity to control for non-discrimination. Trans Data Privacy 2014;7(2):99–129.
[35] Hajian S, Domingo-Ferrer J, Monreale A, Pedreschi D, Giannotti F. Discrimination- and privacy-aware patterns. Data Min Knowl Disc 2015;29(6):1733–82.
[36] Cafarella MJ, Halevy A, Wang ZD, Wu E, Zhang Y. WebTables: exploring the power of tables on the web. In: Proceedings of the Very Large Database Endowment; 2008 Aug 23–28; Auckland, New Zealand; 2008. p. 538–49.
[37] Bizer C, Heath T, Berners-Lee T. Linked data: the story so far. In: Sheth A, editor Semantic services, interoperability and web applications: emerging concepts. Hershey: IGI Global; 2011. p. 205–27.
[38] Batini C, Rula A, Scannapieco M, Viscusi G. From data quality to big data quality. J Database Manage 2015;26(1):60–82.
[39] Firmani D, Mecella M, Scannapieco M, Batini C. On the meaningfulness of "Big Data Quality". Data Sci Eng 2016;1(1):6–20.
[40] Dong XL, Srivastava D. Big data integration. In: Proceedings of the Very Large Databases Endowment; 2013 Aug 26–30; Trento, Italy; 2013. p. 1188–9.
[41] Madhavan J, Jeffery SR, Cohen S, Dong XL, Ko D, Yu C, et al. Web-scale data integration: you can afford to Pay As You Go. In: Proceedings of the 3rd Biennial Conference on Innovative Data Systems Research; 2007 Jan 7–10; Asilomar, CA, USA; 2007. p. 342–50.
[42] Papadakis G, Ioannou E, Palpanas T, Niederée C, Nejdl W. A blocking framework for entity resolution in highly heterogeneous information spaces. IEEE Trans Knowl Data En 2013;25(12):2665–82.
[43] Papadakis G, Koutrika G, Palpanas T, Nejdl W. Meta-blocking: taking entity resolution to the next level. IEEE Trans Knowl Data En 2014;26(8):1946–60.
[44] Papadakis G, Papastefanatos G, Koutrika G. Supervised meta-blocking. In: Proceedings of the Very Large Databases Endowment; 2014 Sep1–5; Hangzhou, China.; 2014. p. 1929–40.
[45] Bergamaschi S, Ferrari D, Guerra F, Simonini G. Discovering the topics of a data source: a statistical approach. In: Proceedings of the Workshop on Surfacing the Deep and the Social Web Co-located with the 13th International Semantic Web Conference; 2014 Oct 19; Trentino, Italy; 2014.
[46] Bergamaschi S, Simonini G, Zhu S. Enhancing big data exploration with faceted browsing. In: Proceedings of the 10th Scientific Meeting of Classification and Data Analysis Group; 2015 Oct 8-10; Cagliari, Italy; 2015.
[47] Fagan JC. Usability studies of faceted browsing: a literature review. Inform Technol Libr 2010;29(2):58–66.