（9）罗加帝（[illegible]）：斯坦福[illegible]

（10）多诺霍（David Donoho）：[illegible]

（11）伯恩（Kirk Borne）：2014 年被评为 IBM 大数据与分析英雄

（12）梅森（Hilary Mason）：Fast Forward Labs 发起人，知名学者

（13）杨立昆（Yann Lecun）：纽约大学数据科学中心的负责人

（14）哈默巴赫（Jeff Hammerbacher）：Cloudera 项目的创始人以及首席科学家

（15）阿钦（Jeremy Achin）：Data Robot 创始人

（16）扎哈里亚（Matei Zaharia）：Spark 的主要开发者，Databricks 的创始人之一

（17）克恩（Gary King）：哈佛教授

（18）金特里（Carla Gentry）：Analytical Solution 的数据科学家

（19）朝乐门：国内第一部系统阐述数据科学专著的作者，国家精品在线开放课程“数据科学导论”的主讲人，北京市优质教材《数据科学理论与实践》作者，数据科学领域本体（Data Science Ontology）的研发者

7．相关工具

（1）Anaconda：全球最受欢迎的数据科学平台之一

（2）Jupyter Notebook：IBM 的开源、支持多种编程语言的开发工具

（3）RapidMiner Studio：数据科学的通用平台

（4）Databricks：数据科学统一分析平台

（5）IBM Watson Studio：IBM 提供的数据科学工具

（6）DataRobot：自动化实现机器学习平台

（7）Trifacta：数据加工的工具

（8）Paxata：数据准备工具

（9）Weka：用 Java 编写的数据挖掘软件

（10）Tableau 和 D3：数据可视化和工具

（11）SAS 和 SPSS：数据分析与建模

（12）谷歌的 Tensotflow 与 Faccbook 的 PyTorch：深度学习框架

（13）Open CV：计算机视觉与图像处理

习　题

一、选择题

1．大数据挑战主要体现在（　　）。

A．数据量（Volume）的几何级增长

B．数据类型（Variety）的多样化

C．数据价值（Value）的发现越来越困难

D．数据处理速度（Velocity）要求越来越高

2. DIKW 金字塔(DIKW Pyramid)模型揭示了数据、信息、知识和（　　）之间的区别与联系。

A. 资料　　B. 能源　　C. 智商　　D. 智慧

3. 以下四种描述中，正确的是（　　）。

A. 大数据和海量数据是同一个事物的不同描述

B. 数据和数值是同一个事物的不同描述

C. 数据和数字是同一个事物的不同描述

D. 以上说法均不正确

4. IBM 认为，大数据是拥有以下 4 个共同特点（又称“4V”）中任意一个的数据源：极大的数据量级、以极快的速度移动、极广泛的数据源类型，以及（　　）。

A. 极高的准确性　　B. 极高的多样性

C. 极高的长久性　　D. 极高的真实性

5.（　　）指从“数据视角”提出问题、在“数据层次”上分析问题、“以数据为中心”解决问题，以及将“数据”当作决策制定的决定因素，提高决策制定的信度与效度。

A. 模型驱动型决策支持　　B. 数据驱动型决策支持

C. 任务驱动型决策支持　　D. 算法驱动型决策支持

6. 在大数据时代，尤其在数据科学中，人们对数据的认识与研究视角是（　　）。

A. 我能为数据做什么　　B. 如何设计算法和模型

C. 数据能为我做什么　　D. 如何降低计算复杂度

7. 从知识体系看，数据科学主要以（　　）为理论基础，其主要研究内容包括数据科学基础理论、数据加工、数据计算、数据管理、数据分析和数据产品开发。

A. 统计学　　B. 机器学习

C. 数据可视化　　D.（某一）领域知识

8. 图灵奖获得者吉姆·格雷提出的科学研究的第四范式——数据密集型科学发现（Data-intensive Scientific Discovery）描述了数据科学的（　　）。

A. 三世界原则　　B. 三要素原则

C. 数据复杂性原则　　D. 从简原则

9. 以下描述中错误的是（　　）。

A. 商务智能主要关注的是对“过去时间”的“解释性研究”，主要回答的是诸如“上一个季度发生了什么？”“销量如何？”“哪里存在问题？”“在什么情况下出现的？”等问题

B. 数据科学主要关注的是对“未来时间”的“探索性研究”，主要回答的是诸如“如果……将来会怎么样？”“最佳业务方案是什么？”等

C. 商务智能的主要处理对象以非结构化数据为主

D. 数据科学建立在数据工程之上，属于“基于数据的处理与管理”，主要关注的

是如何基于数据进行辅助决策（或决策支持）、商业洞察、预测未来、发现潜在模式，以及如何将数据转换为智慧或产品

10.（　　）技术支持将源代码、注释、文字、段落、图表混排在一起，是数据科学家的常用工具之一。

A．VizQL　　B．Markdown　　C．HTML　　D．Excel

11．以下能力中，数据科学家需要具备的能力或素质是（　　）。

A．提出“好”的研究假设或问题，并完成对应的试验设计

B．喜欢团队合作与协同工作

C．掌握数据科学的理论基础——统计学、机器学习和数据可视化

D．学会数据科学的基础理论，尤其是其主要理念、原则、理论和方法

12．以下能力中，数据工程师需要具备的能力或素质是（　　）。

A．数据保障　　B．数据的 ETL 操作

C．数据的备份与恢复　　D．主数据管理及数据集成

13．以下能力中，数据分析师需要具备的能力或素质是（　　）。

A．良好的沟通能力　　B．应用统计学与应用机器学习

C．数据科学　　D．一定的编程开发能力

14．与传统科学不同的是，数据科学是由（　　）驱动，即数据是业务、决策、战略、市场甚至组织结构变化的主要驱动因素。

A．目标　　B．利益　　C．数据　　D．知识

二、调研与分析题

1．结合自己的专业领域，调研数据科学及大数据在所属领域中的应用现状。

2．调查分析近 3 年在数据科学领域出版的专著。

3．调查分析数据科学家的常用方法、技术与工具。

4．调查分析近 3 年 *The Data Science Journal* 等数据科学领域的学术期刊上发表论文的主题。

5．调查分析近 3 年 IEEE、DSAA 等数据科学领域国际会议的主题。