大数据的应用举例

大数据在金融行业应用范围较广，典型的案例有花旗银行利用 IBM 沃森电脑为财富管理客户推荐产品；美国银行利用客户点击数据集为客户提供特色服务

大数据在金融行业的应用可以总结为以下方面：

（1）风险管控：依据客户消费和现金流提供信用评级或融资支持，利用客户社交行为记录实施信用卡反欺诈

（2）产品设计：利用大数据计算技术为财富客户推荐产品，利用客户行为数据设计满足客户需求的金融产品

Big data is widely used in the financial industry. Typical cases include Citibank, which uses IBM Watson computer to recommend products for wealth management clients.

Big data in the application of the financial industry can be summed up in the following aspects:

(1) risk control: according to customer consumption and cash flow to provide credit rating or financing support, using customer social behavior record the implementation of credit card fraud

(2) Product design: Use big data computing technology to recommend products for fortune customers, and use customer behavior data to design financial products to meet customer needs

由于农产品不容易保存，因此合理种植和养殖农产品对十分重要。借助于大数据提供的消费趋势报告和消费习惯报告，政府将为农牧业生产提供合理引导，建议依据需求进行生产，避免产能过剩，造成不必要的资源和社会财富浪费。大数据技术可以帮助政府实现农业的精细化管理，实现科学决策。

Because agricultural products are not easy to preserve, it is important to grow and raise them properly. With the help of the consumption trend report and consumption habit report provided by big data, the government will provide reasonable guidance for agricultural and animal husbandry production, and suggest that production should be carried out according to demand, to avoid excess production capacity and unnecessary waste of resources and social wealth. Big data technology can help the government to realize the fine management of agriculture and scientific decision-making.

交通的大数据应用主要在两个方面，一方面可以利用大数据传感器数据来了解车辆通行密度，合理进行道路规划包括单行线路规划。另一方面可以利用大活数据来实现即时信号灯调度，提高已有线路运行能力。科学的安排信号灯是一个复杂的系统工程，必须利用大数据计算平台才能计算出一个较为合理的方案。

The application of big data in traffic is mainly in two aspects. On the one hand, big data sensor data can be used to understand vehicle traffic density and rational road planning. On the other hand, large live data can be used to realize real-time semaphore dispatching and improve the running ability to existing lines. Scientific arrangement of signal lights is a complex system engineering, which requires the use of big data computing platform to work out a more reasonable scheme.

A data lifecycle

• Collect - collect the data in the same format for each state's system, allowing them to be compared and reviewed for errors

• Describe - use the same metadata for all data collected for each state's system, allowing them
to be searched using universal terminology between states and different tollways

• Preserve - store the data in the same type of database to minimise maintenance costs across
states and tollways

• Integrate - transfer the data between systems using the same format and functionality to simplify security, process and compatibility

• Analyse - use the same software, functions and visualisation where possible, so that there is a common understanding of the data and analysis between stakeholders, regardless of state or

• What is a data scientist?
A **data scientist** is a practitioner who has sufficient knowledge in the overlapping regimes of business needs, domain knowledge, analytical skills, and software and systems engineering to manage the end-to-end data processes through each stage in the data lifecycle. Moreover, **Data science** is the empirical synthesis of actionable knowledge from raw data through the complete data lifecycle process.

• What is R Markdown?
To maintain a reproducible workflow, you need to record what steps you take in a process. R Markdown is an authoring format (.Rmd files) that enables us to combine embedded R code

• How has data science changed business models?
Information brokering service: buys and sells data/information for others.
Information-based differentiation: satisfies customers by providing a differentiated service built on the data/information.
Information-based delivery network: deliver data/ information for others.
Information provider: business selling the data/ information it collects.

• How can you handle missing data?
If a small fraction of cases have several missings, drop the cases.
If a variable or two, out of many, have a lot of missings, drop the variables.
If missings are small in number, but located in many cases and variables, you need to impute these values (replace with substituted values) to do most analyses.

• What is the significance of open data?
open data provides new opportunities for business, new products and services, and can raise productivity
open data supports public understanding and citizen engagement
scientists need to better publicise their data (with help from universities, etc.)
industry sectors should work with regulators and coordinate industry collaboration
collaboration across sectors in both public and private settings

• How are temporal elements included in data?
The temporal aspect of the data can be of different types. Specific and Relative.

• what difference between Standard Value Chain and Data Lifecycle

The value chain tends to presume there is a start and an end to the data science process, as part of the business activity. The lifecycle is more about the interaction with the data (or roles/transformations of the data), regardless of where it sits in business activities. Governance should play a role in all the value chain and all the lifecycle, making sure they all are managed appropriately and efficiently. Different models represent that relationship in different ways.

- How can linked data be used?

1、把 URI 当作东西的名字使用

使用 URI 作为资源的标识，即网络上的任何事物或资源的标识 名称，如 HTML 文档、科研人员、国家等，都使用 URI 进行标识和定位，用于帮 助用户更直接的获取资源。

2、为了让人们可以查找这些名字，使用 HTTP URI。

使用 HTTP URI 来标识资源，在网络环境下，数据 资源能够通过 HTTP 协议访问获取，真正实现基于 Web 的访问和互联

3、当某个人查找某个 URI 的时候，以规范的标准(RDF, SPARQL)，提供他有用的资料。

当某个人查询一个 URI 时，使用 RDF 提供与当前资源相关的其他有用信息，为用户提供更 多有价值的关联资源。

4、在提供他的资料里，给他指到别的 URI 的连结，使他可以发现更多东西。

与更多相关资源的 HTTP URI 建立语义链接，提高用户 发现、获取和使用网络中潜在的相关信息资源的能力。

1. Use the URI as the name of something

URI is used as the identification of resources, that is, the identification name of any thing or resource on the network, such as HTML documents, researchers, countries, etc., all use URI for identification and positioning, to help users obtain resources more directly.

2. To enable people to look up these names, use HTTP URIs.

HTTP URI is used to identify resources. In the network environment, data resources can be accessed and obtained through HTTP protocol, thus truly realizing web-based access and interconnection

3. When someone looks up a URI, provide him/her with canonical standards (RDF, SPARQL).

When someone queries a URI, RDF is used to provide other useful information about the current resource, providing users with more valuable associated resources.

4. Give users links to other URIs in his profile so he can find out more.

Establish semantic links to HTTP URIs for more relevant resources, improving the user's ability to discover, retrieve, and use potentially relevant information resources on the network.