Big Data Technology and its Applications

Introduction

张宁 ningzhang@tsinghua.edu.cn

Outline

- What is big data
- What is machine learning
- Big data and machine learning in power system
- Course information

About Big Data

- ***Big Data*** is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.
- **Features**: “4V”
  - Volume
  - Velocity
  - Variety
  - Veracity/Value

Scale of Data
- 2.3 Trillion gigabytes of data are created every day

Analysis of Streaming Data
- Smart cars have more than 100 sensors that monitor

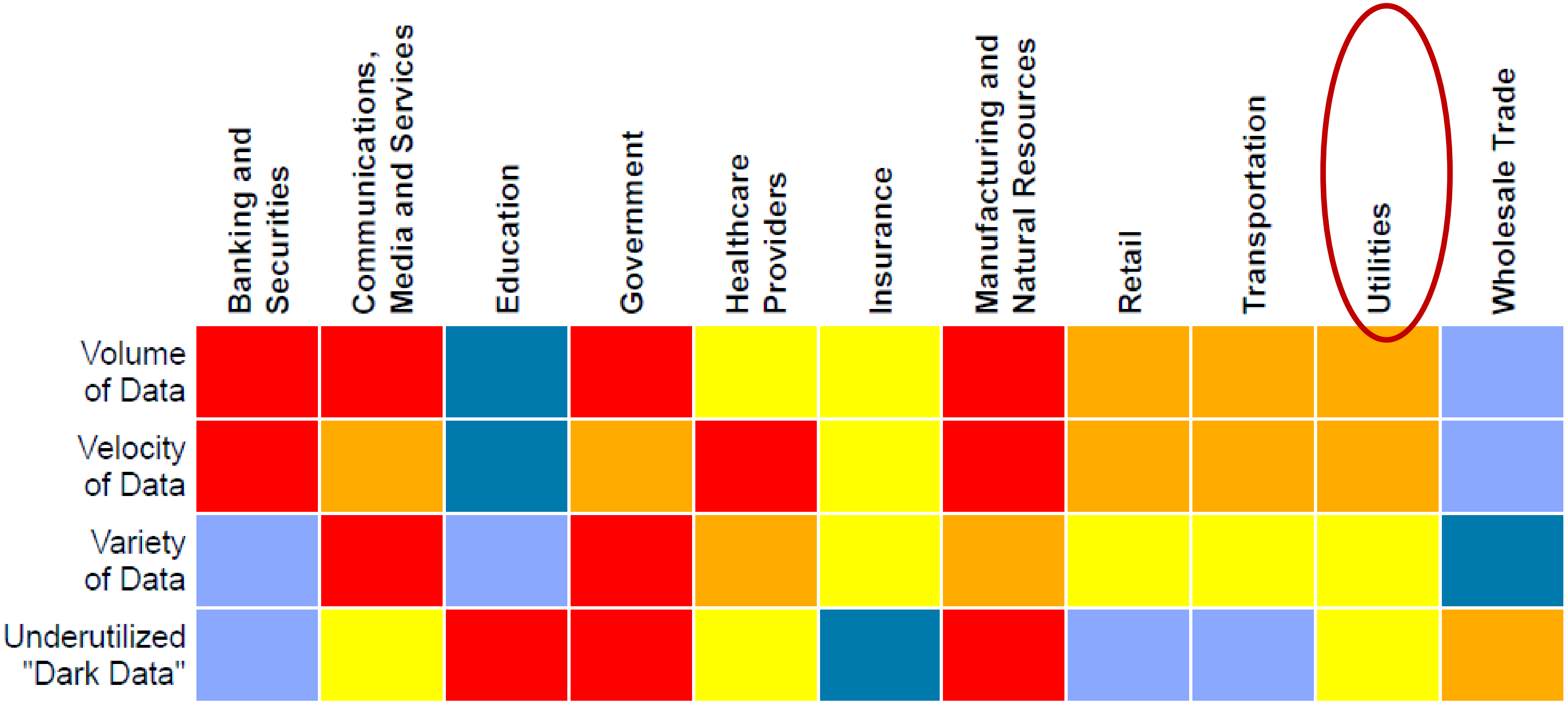
Different Forms of Data
- Text, image, video, machine data

Uncertainty of Data
- Lots of irrelevant information
- Unsure of how much of data was accurate

Which area can the big data be helpful in?

- Name three areas that you think big data help the most in your area.

Features of Big Data

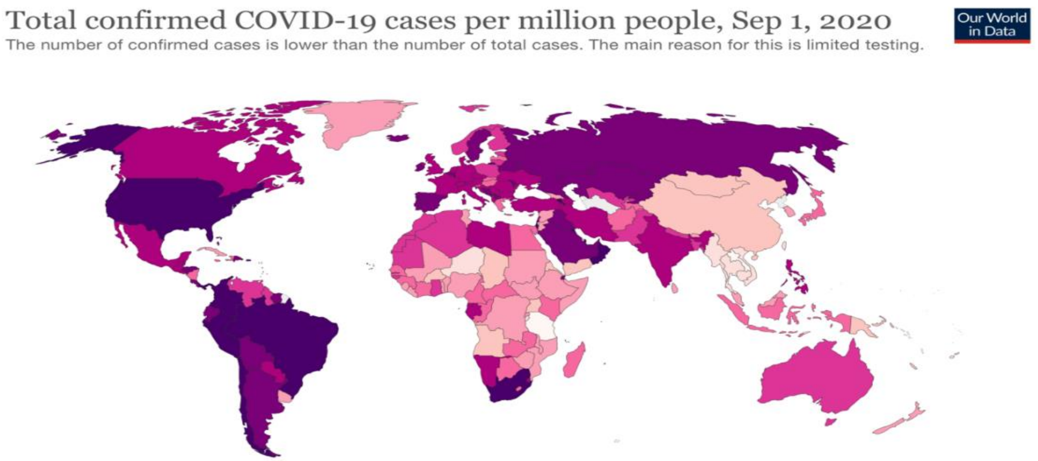


	Banking and Securities	Communications, Media and Services	Education	Government	Healthcare Providers	Insurance	Manufacturing and Natural Resources	Retail	Transportation	Utilities	Wholesale Trade
Volume of Data	🟥	🟥	🟦	🟥	🟨	🟨	🟥	🟧	🟧	🟧	🟪
Velocity of Data	🟥	🟧	🟦	🟧	🟥	🟨	🟥	🟧	🟧	🟧	🟪
Variety of Data	🟪	🟥	🟪	🟥	🟧	🟨	🟧	🟨	🟨	🟨	🟦
Underutilized "Dark Data"	🟪	🟨	🟥	🟥	🟨	🟦	🟥	🟪	🟪	🟨	🟧

Potential big data opportunity on each dimension is:

🟥 Very hot (compared with other industries)
🟧 Hot
🟨 Moderate
🟪 Low
🟦 Very low (compared with other industries)

Big Data is Changing the World


Total confirmed COVID-19 cases per million people, Sep 1, 2020

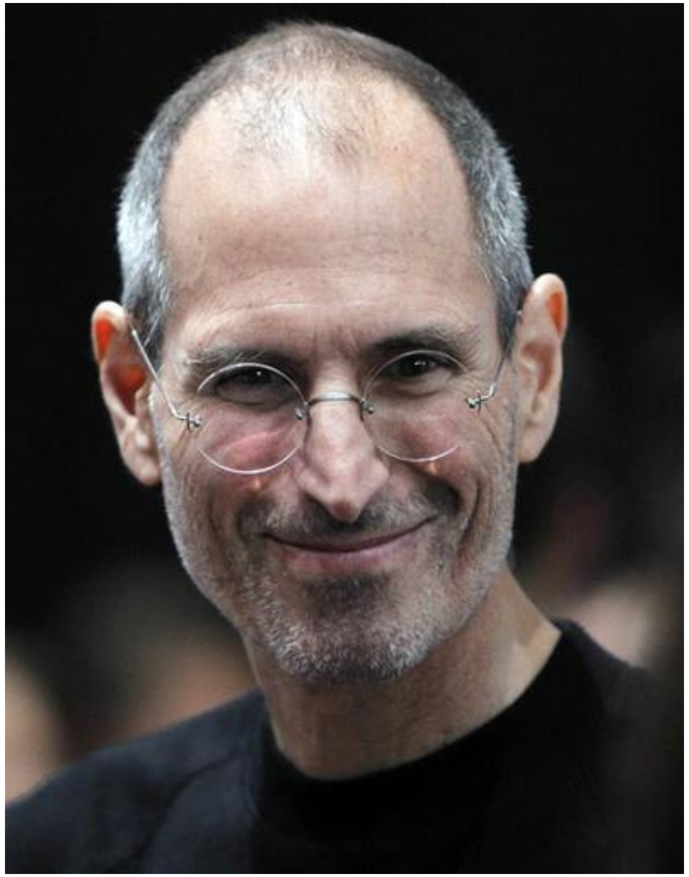
The number of confirmed cases is lower than the number of total cases. The main reason for this is limited testing.

Forecast COVID-19 with big data

Source: European CDC – Situation Update Worldwide – Last updated 1 September, 10:04 (London time)
OurWorldInData.org/coronavirus • CC BY

Recommendation algorithms of APPs

Ezioni's *Farecast* has 10 trillion pieces of data to predict airfare prices


Jobs' cancer fight (2003-2011), sequencing of his own DNA and tumor DNA

Application: Facebook Data Helps The Presidential Election

Easter special

Chickens, eggs... and chocolate

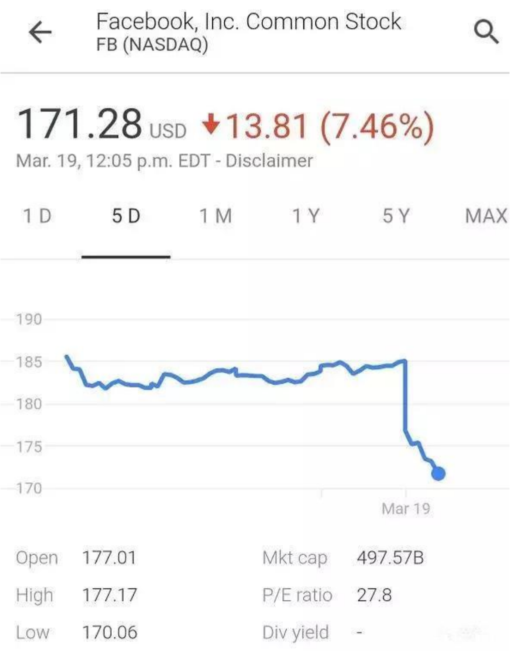
Ireland crush England for the grand slam

Stewart Lee: Don't blame the Russians

The Observer

Revealed: 50m Facebook files taken in record data breach

The Cambridge Analytica Files



Facebook, Inc. Common Stock
FB (NASDAQ)

171.28 USD ↓13.81 (7.46%)

Mar. 19, 12:05 p.m. EDT - Disclaimer

Open	177.01	Mkt cap	497.57B
High	177.17	P/E ratio	27.8
Low	170.06	Div yield	-

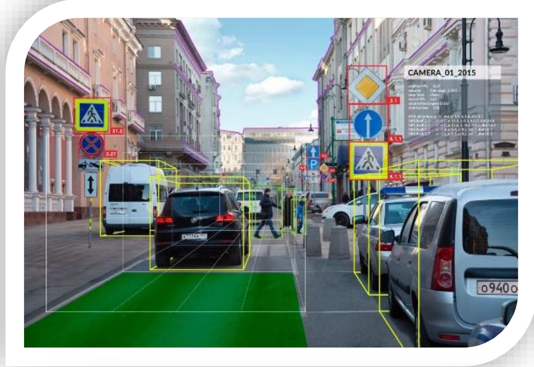
- Facebook shares 7.46% on March 19, evaporating $36.4 billion in market value

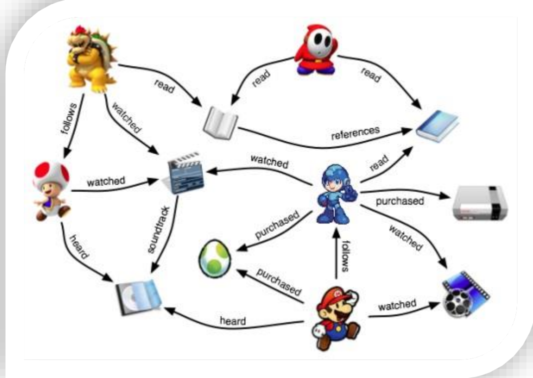
- Data analysis company Cambridge Analytica obtained massive user data from more than 50 million North American users on Facebook. he divided the user data of all American adults into 32 different personalities, and then contacted the advertising company to post 175,000 different ads promoting Trump's views on Facebook.

- Knicks replied this way: "Just give me 68 likes on Facebook, and I can infer *the person's skin color, sexual orientation, political leaning, intellectual level, religious preference, whether he drinks alcohol, drugs, and whether his parents are divorced, etc. information*"

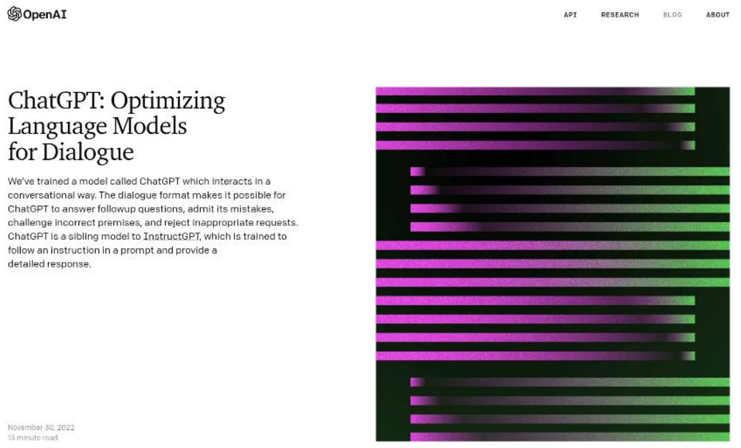
Frontiers of Big Data

- Computer Science
  - Natural language processing (NLP)
    - How to program computers to process and analyze large amounts of natural language data.
  - Computer vision (CV)
    - How computers can gain high-level understanding from digital images or videos.
  - Recommendation system
    - Predict the *rating* or *preference* a user would give to an item.
  - …





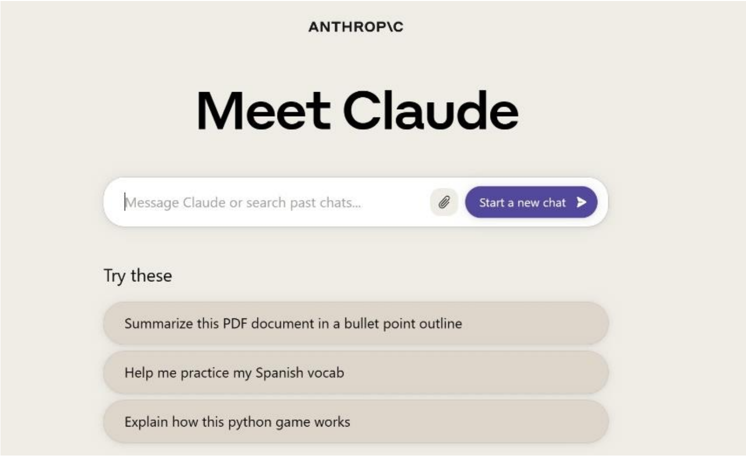
NLP models



ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

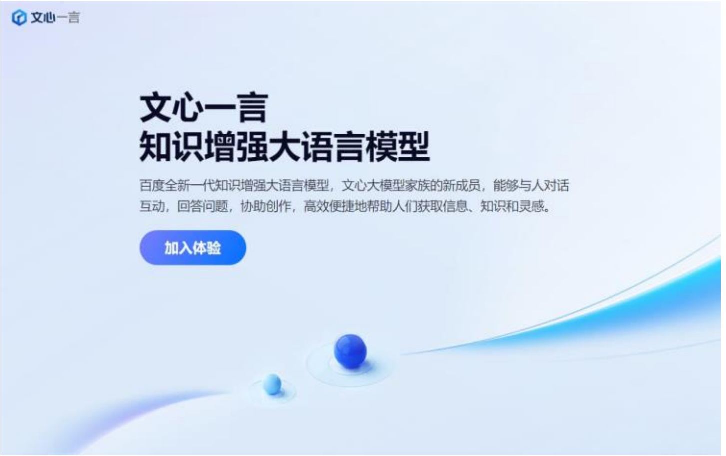
November 30, 2022
13 minute read



Meet Claude

Try these

- Summarize this PDF document in a bullet point outline
- Help me practice my Spanish vocab
- Explain how this python game works



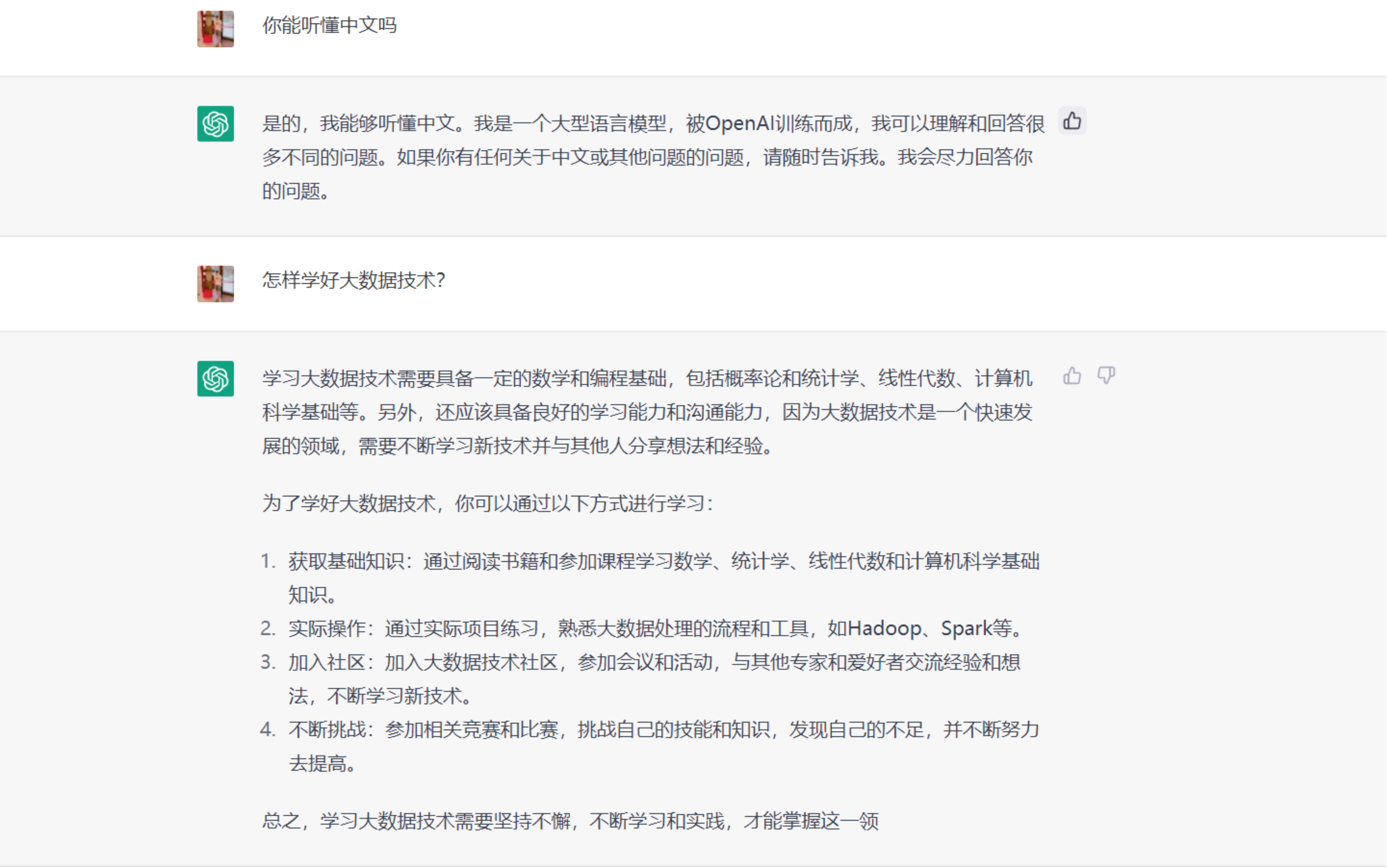
文心一言
知识增强大语言模型

百度全新一代知识增强大语言模型，文心大模型家族的新成员，能够与人对话互动，回答问题，协助创作，高效便捷地帮助人们获取信息、知识和灵感。

OpenAI

ANTHROP\C

Bai du 百度



你能听懂中文吗

是的，我能够听懂中文。我是一个大型语言模型，被OpenAI训练而成，我可以理解和回答很多不同的问题。如果你有任何关于中文或其他问题的问题，请随时告诉我。我会尽力回答你的问题。

怎样学好大数据技术?

学习大数据技术需要具备一定的数学和编程基础，包括概率论和统计学、线性代数、计算机科学基础等。另外，还应该具备良好的学习能力和沟通能力，因为大数据技术是一个快速发展的领域，需要不断学习新技术并与其他人分享想法和经验。

为了学好大数据技术，你可以通过以下方式进行学习：

1. 获取基础知识：通过阅读书籍和参加课程学习数学、统计学、线性代数和计算机科学基础知识。
2. 实际操作：通过实际项目练习，熟悉大数据处理的流程和工具，如Hadoop、Spark等。
3. 加入社区：加入大数据技术社区，参加会议和活动，与其他专家和爱好者交流经验和想法，不断学习新技术。
4. 不断挑战：参加相关竞赛和比赛，挑战自己的技能和知识，发现自己的不足，并不断努力去提高。

总之，学习大数据技术需要坚持不懈，不断学习和实践，才能掌握这一领

Outline

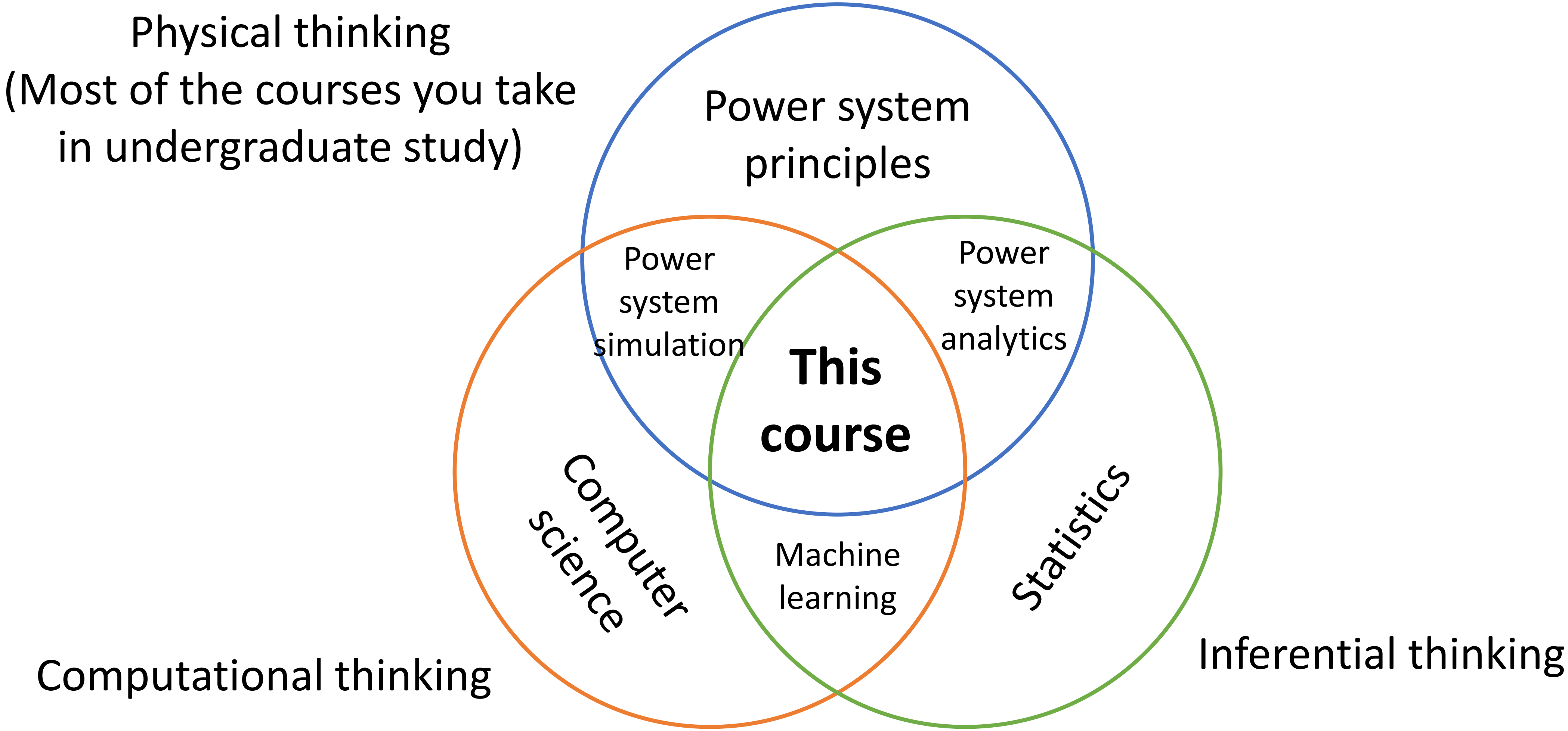
- What is big data
- What is machine learning
- Big data and machine learning in power system
- Course information

What is machine learning

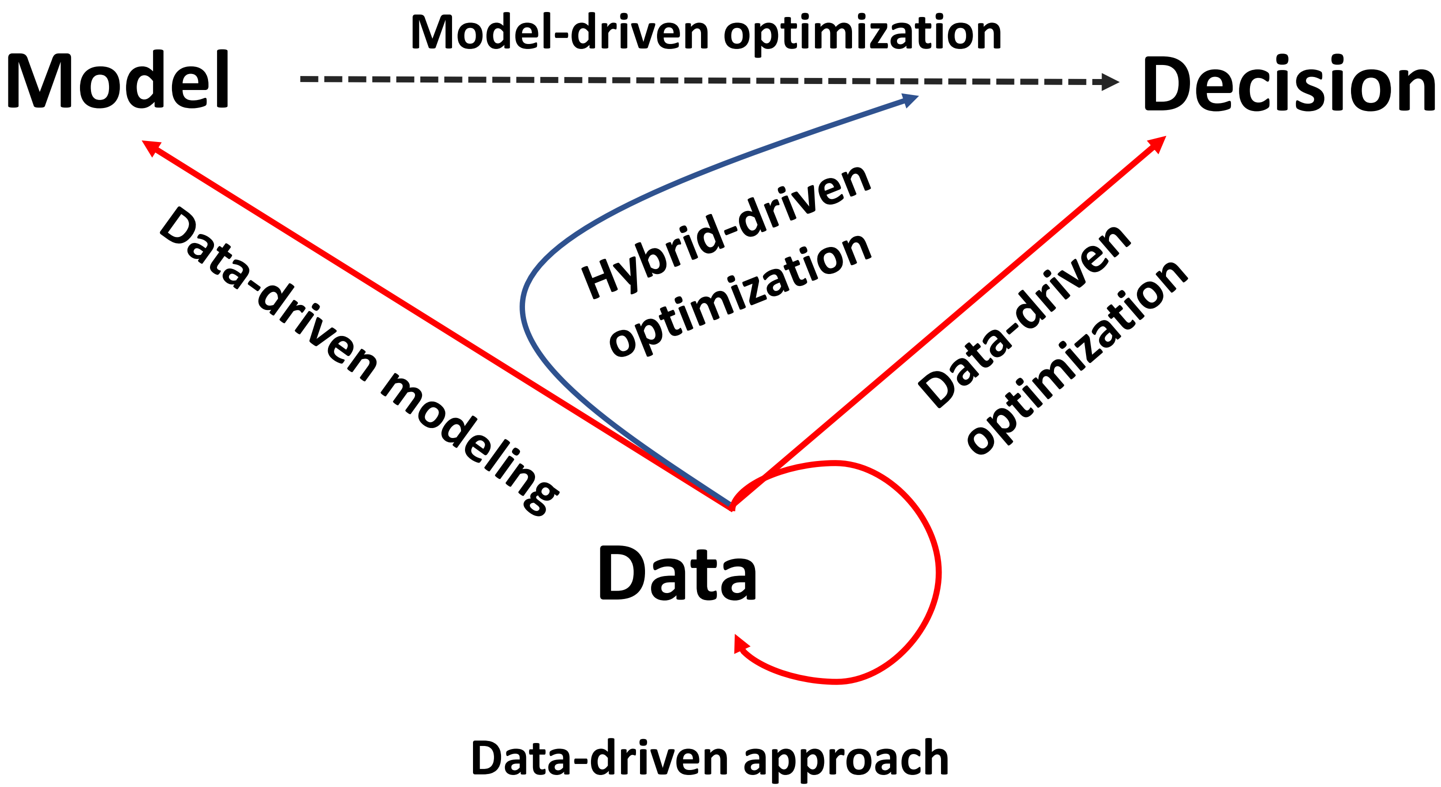
- Machine learning gives computers the ability to learn without being explicitly programmed. Machine learning explores the study and construction of algorithms which can learn and make predictions on data.

- https://simple.wikipedia.org/wiki/Machine_learning

What is machine learning



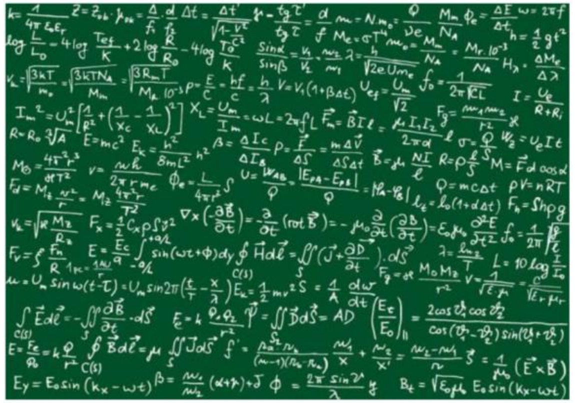
Model / Physical driven VS Data driven



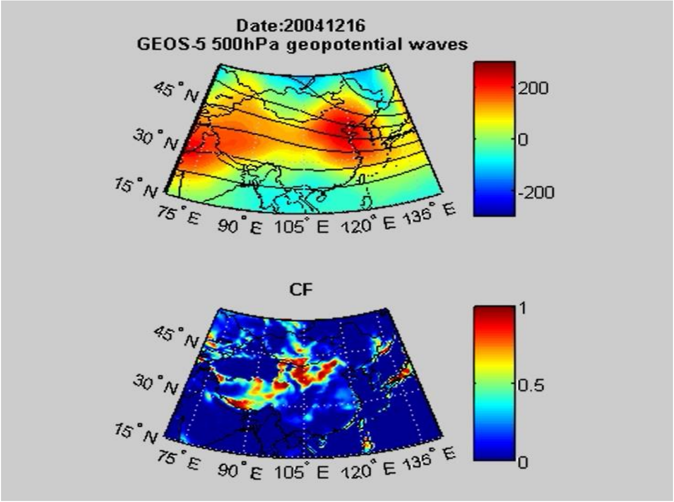
The Fourth Science Paradigm?

- Experimental Science

- Theoretical Science

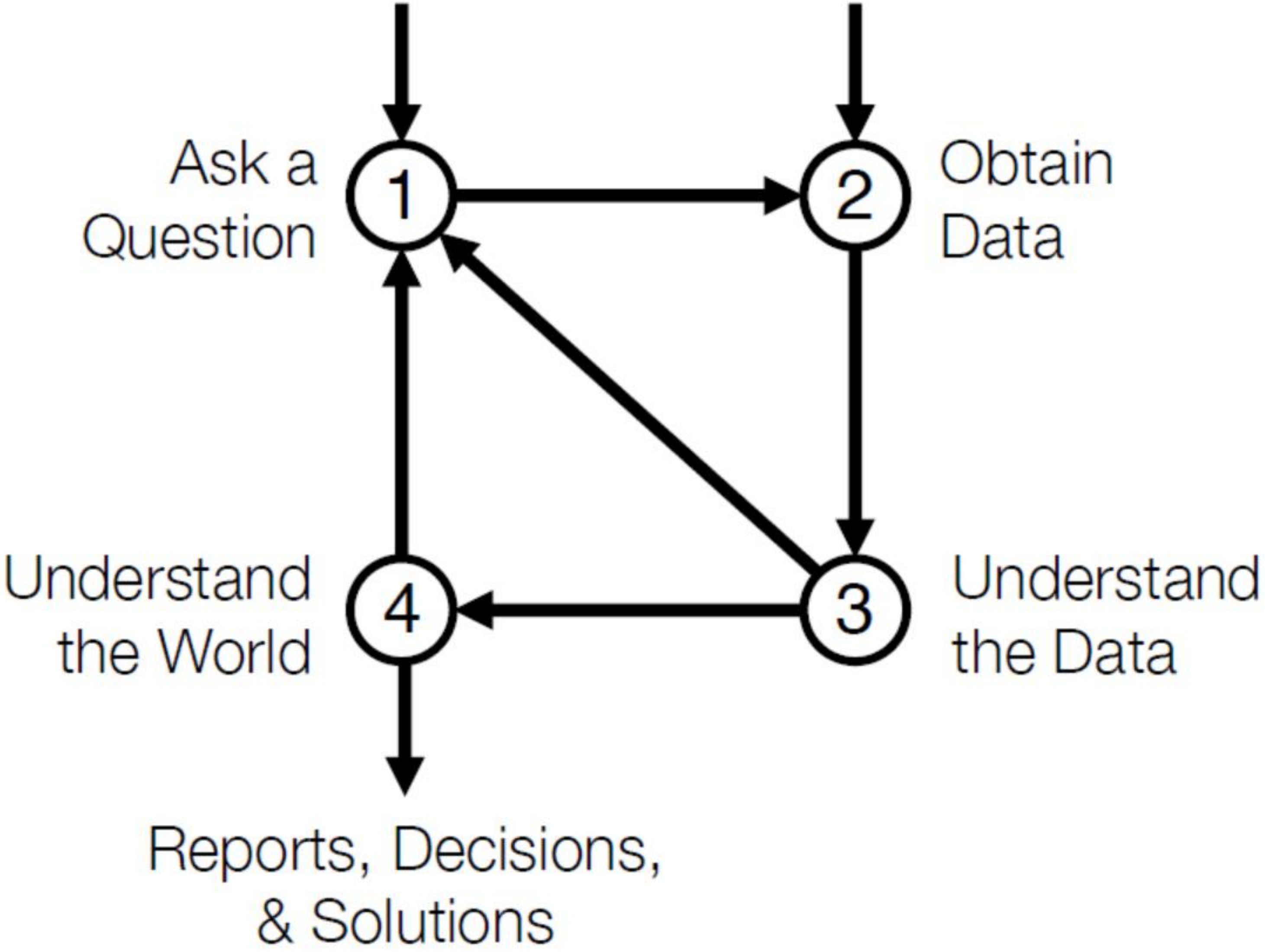


- Computational Science



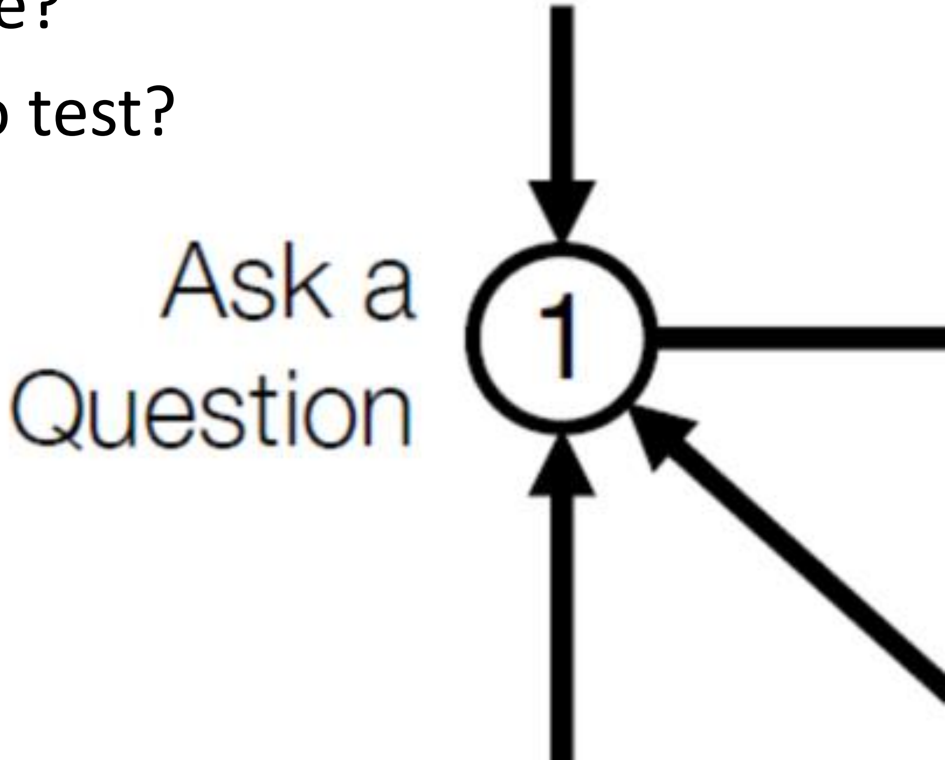
- Data Science

How to use machine learning to solve a problem



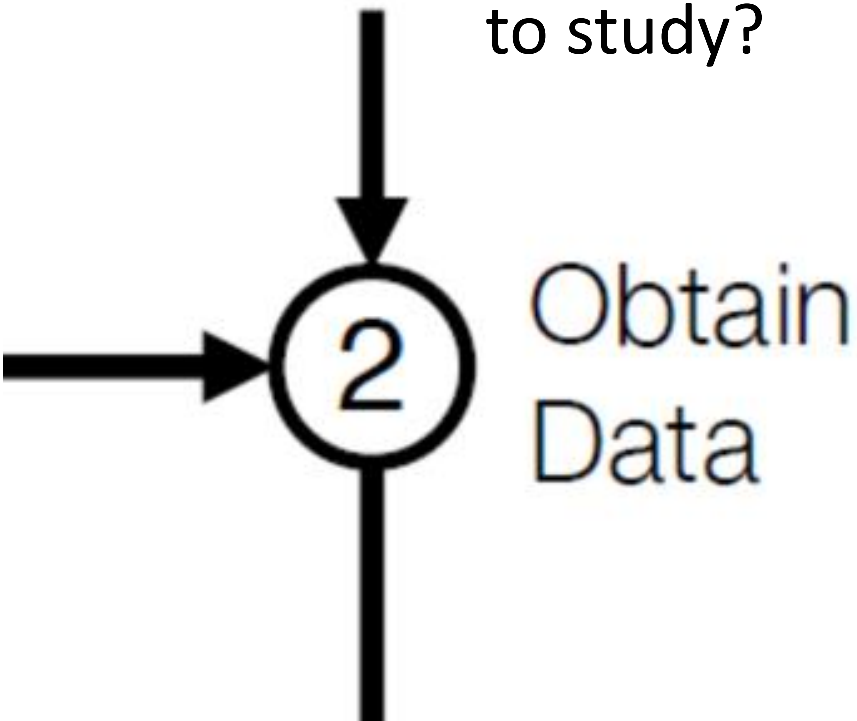
Problem Formulation

- What do we want to know?
- What problems are we trying to solve?
- What are the hypotheses we want to test?
- What are our metrics of success?

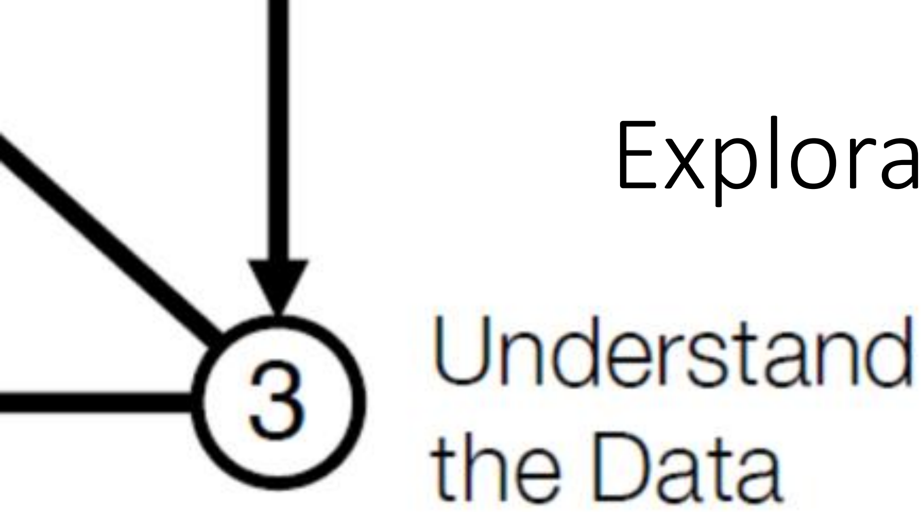


Data Acquisition and Cleaning

- What data do we have and what data do we need?
- How will we sample more data?
- Are the data enough for the study?
- Is our data representative of the population we want to study?



Exploratory Data Analysis & Visualization



- How is our data organized and what does it contain?
- How do we transform the data to enable effective analysis?
- What do the data tell us (what information is contained in the data)?

Predictions and Inference

Understand the World
4
Reports, Decisions, & Solutions

- What does the data say about the world?
- Does it answer our questions or accurately solve the problem?
- How robust are our conclusions and can we trust the predictions?

Outline

- What is big data
- What is machine learning
- Big data and machine learning in power system
- Course information

Power Systems Big Data

No.	System/ Data	Data Source	Data Type	Frequency	Data Structure
1	Economic Information	Statistic Bureau	GDP、CPI、PMI（Purchasing Managers Index）、Sales Value、Prosperity Index	Per Month	Non structural
2	Energy Consumption Data	Energy Efficiency Platform	Electrical Load、Output、Power Quality、Temperature	15Min	Non structural /Structural
3	Meteorological Data	Meteorological Bureau	Temperature、Humidity、Rainfall	Per Day	Structural
4	EV Charging Data	Charging-Pile RTU	Current、Voltage、Charging Rate、State of Charge	15Min	Structural
5	Customer Service Voice Data	Customer Service System	Customer Voice Data	Real Time	Non structural

“4V” Characteristics:

High “Variety”

High generation “Velocity”

Huge “Volume”

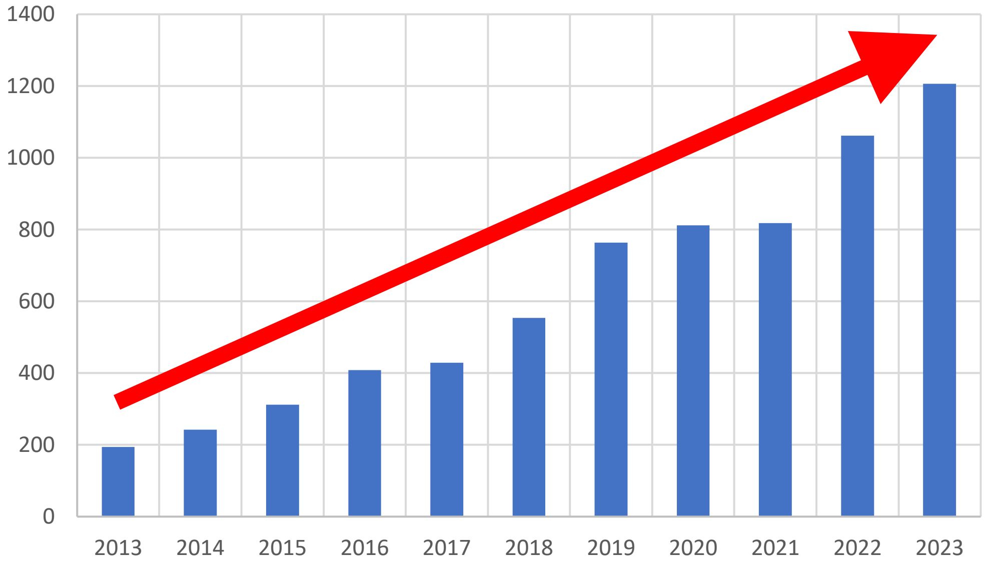
Relatively low “Value”

10 million Smart Meters, 15min ⇨ 60GB per day, 21TB per year.

- A total of **3.583 million** data sheets and **10.874 PB** data are collected in the data center of State Grid headquarters + provincial companies!

Power Systems Big Data

- Analysis of electricity consumption big data shows a continuous growth trend



WoS索引论文数量（2023-12-31）

Web of Science核心合集
TS=(("smart meter" OR "consumption" OR "demand" OR "load") AND "data" AND ("household" OR "resident" OR "residential" OR "building" OR "industrial" OR "individual" OR "customer" OR "consumer") AND ("energy theft" OR "demand response" OR "clustering" OR "forecasting" OR "profiling" OR "classification" OR "abnormal" OR "anomaly") AND ("smart grid" OR "power system"))

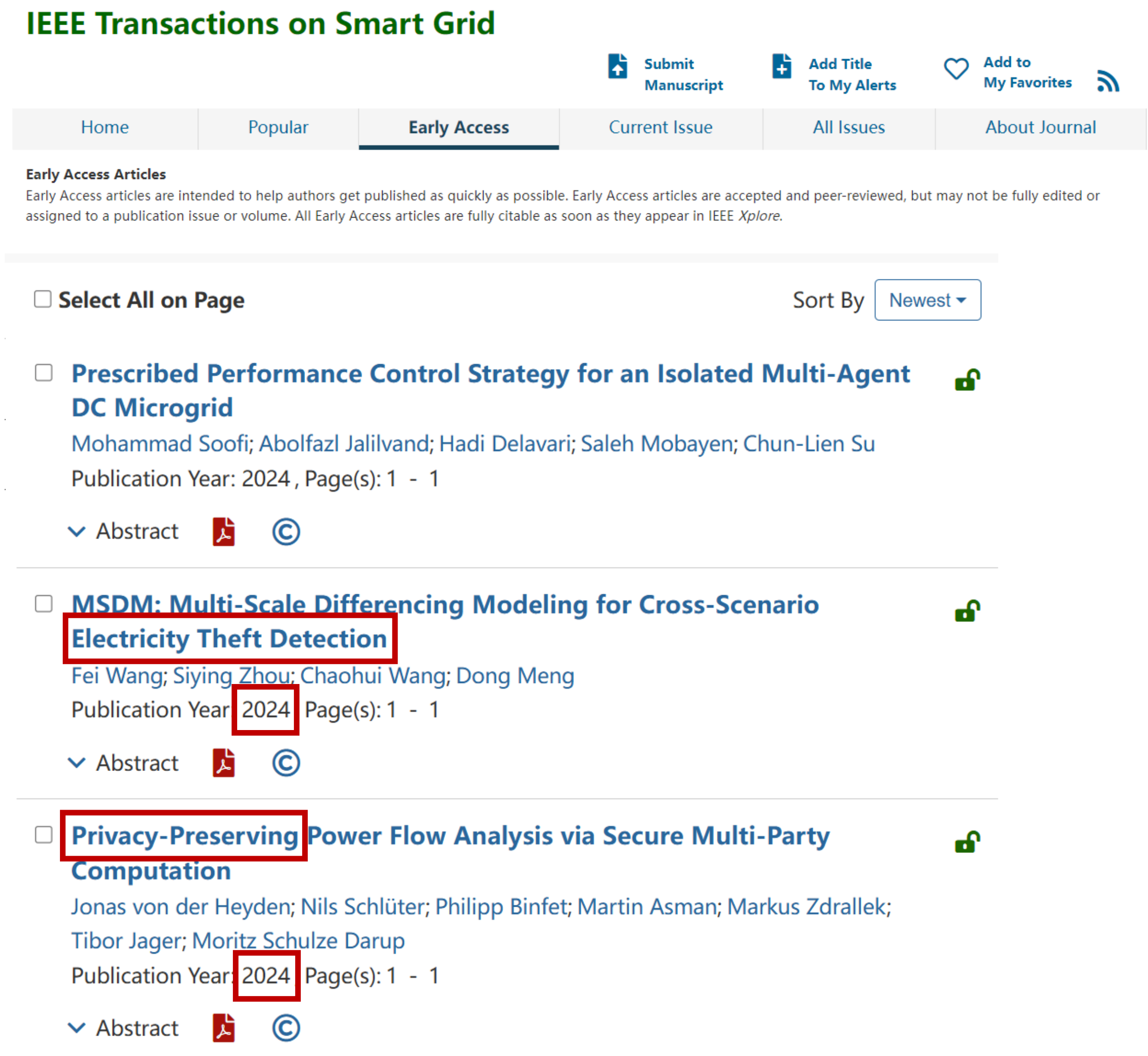
- **What is Data Analytics：**

- Data Analytics is commonly dissected into three stages: **descriptive analytics** (what do the data look like), **predictive analytics** (what is going to happen with the data), and **prescriptive analytics** (what decisions can be made from the data).

- **Descriptive：** What is the existing data？
- **Predictive：** What will happen in the future？
- **Prescriptive：** What kind of decision can be made？

Bertolucci J. Big data analytics: Descriptive vs. predictive vs. prescriptive[J]. Information Week, 2013.

Power Systems Big Data



IEEE Transactions on Smart Grid

Early Access Articles

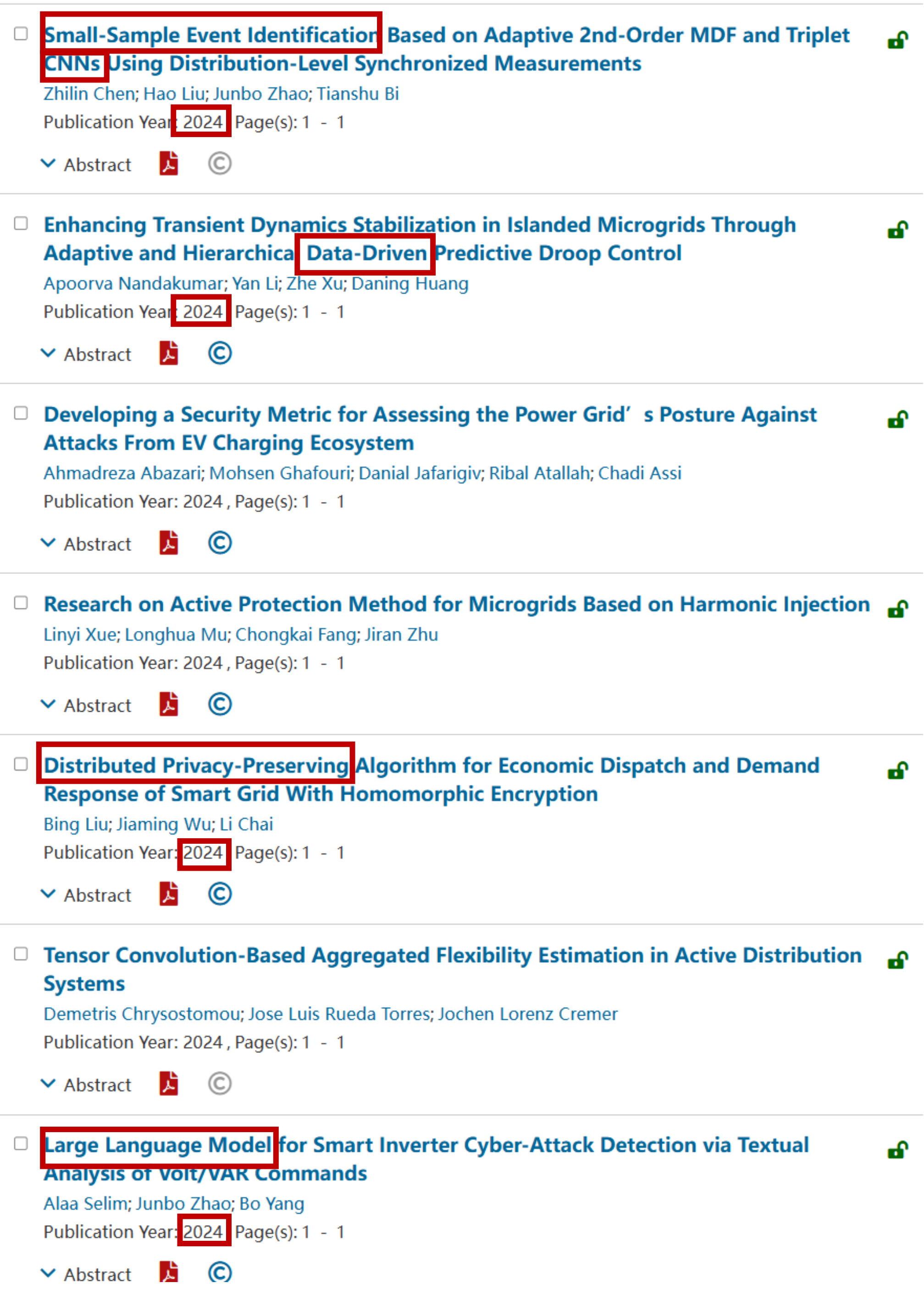
Early Access articles are intended to help authors get published as quickly as possible. Early Access articles are accepted and peer-reviewed, but may not be fully edited or assigned to a publication issue or volume. All Early Access articles are fully citable as soon as they appear in IEEE *Xplore*.

☐ **Select All on Page** Sort By Newest

☐ **Prescribed Performance Control Strategy for an Isolated Multi-Agent DC Microgrid**
Mohammad Soofi; Abolfazl Jalilvand; Hadi Delavari; Saleh Mobayen; Chun-Lien Su
Publication Year: 2024 , Page(s): 1 - 1
Abstract

☐ **MSDM: Multi-Scale Differencing Modeling for Cross-Scenario Electricity Theft Detection**
Fei Wang; Siying Zhou; Chaohui Wang; Dong Meng
Publication Year: 2024 , Page(s): 1 - 1
Abstract

☐ **Privacy-Preserving Power Flow Analysis via Secure Multi-Party Computation**
Jonas von der Heyden; Nils Schlüter; Philipp Binfet; Martin Asman; Markus Zdrallek; Tibor Jager; Moritz Schulze Darup
Publication Year: 2024 , Page(s): 1 - 1
Abstract



☐ **Small-Sample Event Identification Based on Adaptive 2nd-Order MDF and Triplet CNNs Using Distribution-Level Synchronized Measurements**
Zhilin Chen; Hao Liu; Junbo Zhao; Tianshu Bi
Publication Year: 2024 , Page(s): 1 - 1
Abstract

☐ **Enhancing Transient Dynamics Stabilization in Islanded Microgrids Through Adaptive and Hierarchical Data-Driven Predictive Droop Control**
Apoorva Nandakumar; Yan Li; Zhe Xu; Daning Huang
Publication Year: 2024 , Page(s): 1 - 1
Abstract

☐ **Developing a Security Metric for Assessing the Power Grid’ s Posture Against Attacks From EV Charging Ecosystem**
Ahmadreza Abazari; Mohsen Ghafouri; Danial Jafarigiv; Ribal Atallah; Chadi Assi
Publication Year: 2024 , Page(s): 1 - 1
Abstract

☐ **Research on Active Protection Method for Microgrids Based on Harmonic Injection**
Linyi Xue; Longhua Mu; Chongkai Fang; Jiran Zhu
Publication Year: 2024 , Page(s): 1 - 1
Abstract

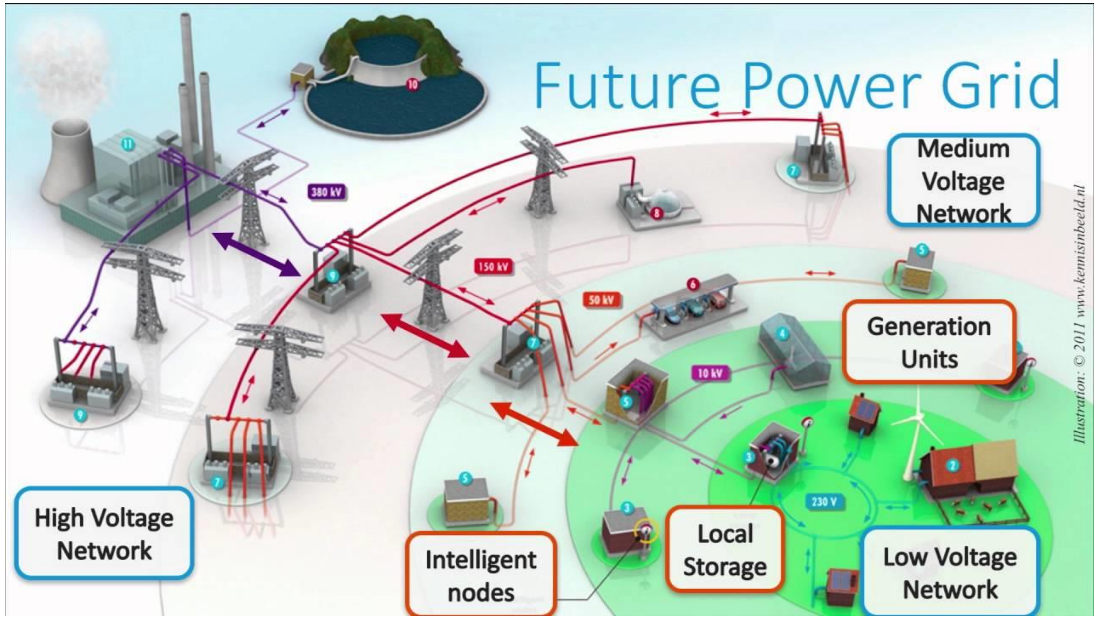
☐ **Distributed Privacy-Preserving Algorithm for Economic Dispatch and Demand Response of Smart Grid With Homomorphic Encryption**
Bing Liu; Jiaming Wu; Li Chai
Publication Year: 2024 , Page(s): 1 - 1
Abstract

☐ **Tensor Convolution-Based Aggregated Flexibility Estimation in Active Distribution Systems**
Demetris Chrysostomou; Jose Luis Rueda Torres; Jochen Lorenz Cremer
Publication Year: 2024 , Page(s): 1 - 1
Abstract

☐ **Large Language Model for Smart Inverter Cyber-Attack Detection via Textual Analysis of Volt/VAR Commands**
Alaa Selim; Junbo Zhao; Bo Yang
Publication Year: 2024 , Page(s): 1 - 1
Abstract

Digital: The driving force of the energy transition

- China's big energy companies, such as **State Grid Corporation of China, China Southern Power Grid Corporation and State Oil and Gas Pipeline Network Corporation**, have developed digital transformation strategies to treat **data** as an important factor of production.



国家电网
STATE GRID

中国南方电网
CHINA SOUTHERN POWER GRID

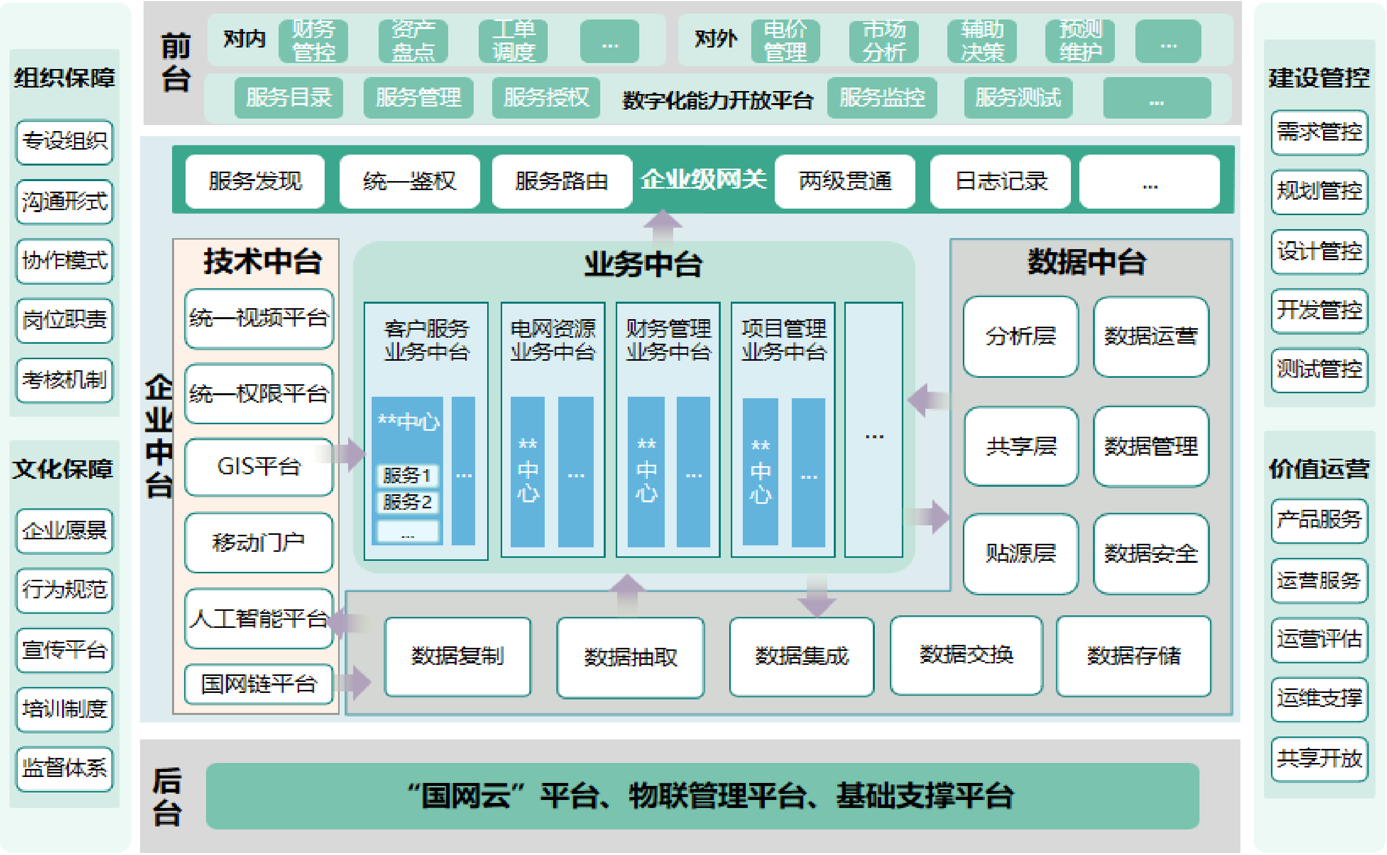
- ➢ Big Data Center of the State Grid ,
  Power Internet of Things
- ➢ Digital Grid Research Institute, Transparent Grid

Big Data Center of the State Grid

国家电网公司 STATE GRID CORPORATION OF CHINA
国家电网
STATE GRID

Started construction in March 2019 and was officially listed on May 21. It is the **professional support organization, data service platform, digital innovation platform, and data sharing platform** for the company's digital transformation.

Carry out the **whole chain management, operation and service of data business**, break through the data barriers of various specialties, and realize data convergence, fusion and sharing.



组织保障
专设组织
沟通形式
协作模式
岗位职责
考核机制

文化保障
企业愿景
行为规范
宣传平台
培训制度
监督体系

前台
对内 财务管控 资产盘点 工单调度 …
对外 电价管理 市场分析 辅助决策 预测维护 …
服务目录 服务管理 服务授权 **数字化能力开放平台** 服务监控 服务测试 …

企业中台
服务发现 统一鉴权 服务路由 **企业级网关** 两级贯通 日志记录 …

技术中台
统一视频平台
统一权限平台
GIS平台
移动门户
人工智能平台
国网链平台

业务中台
客户服务业务中台：**中心（服务1、服务2、…）…
电网资源业务中台：**中心 …
财务管理业务中台：**中心 …
项目管理业务中台：**中心 …
…

数据中台
分析层 数据运营
共享层 数据管理
贴源层 数据安全

数据复制 数据抽取 数据集成 数据交换 数据存储

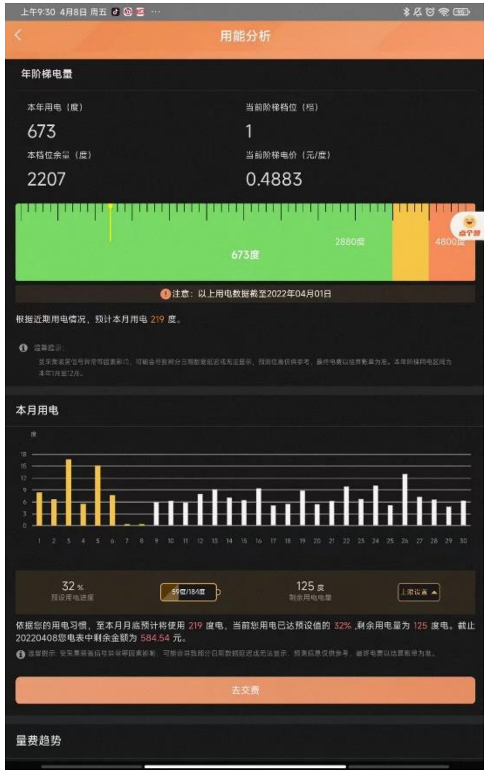
后台
"国网云"平台、物联管理平台、基础支撑平台

建设管控
需求管控
规划管控
设计管控
开发管控
测试管控

价值运营
产品服务
运营服务
运营评估
运维支撑
共享开放

能源数字化转型白皮书
(2021)

国家电网
STATE GRID
国网能源研究院有限公司



年阶梯电量

本年用电（度）	当前阶梯档位（档）
673	1
本档位余量（度）	**当前阶梯电价（元/度）**
2207	0.4883

注意：以上用电数据截至2022年04月01日

根据近期用电情况，预计本月用电 219 度。

本月用电

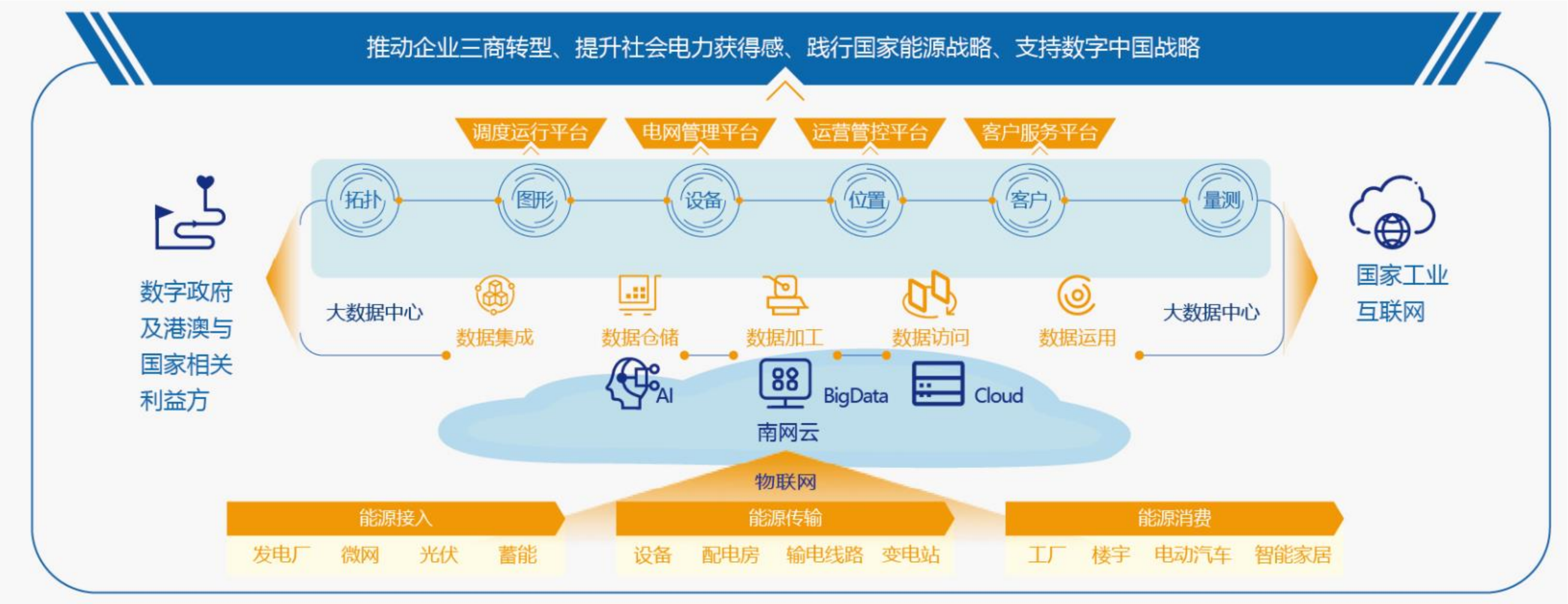
32 % 预设用电进度
59度/184度
125 度 剩余用电电量

依据您的用电习惯，至本月月底预计将使用 219 度电，当前您用电已达预设值的 32% ,剩余用电量为 125 度电。截止20220408您电表中剩余金额为 584.54 元。

量费趋势

Digital transformation strategy of China Southern Power Grid

"Digital South Grid": In 2019, China Southern Power Grid put forward the digital transformation project. Constructions of **four digital business platform, three basic digital platform** (the digital grid platform, the global Internet platform, the whole-stack cloud platform) are proposed. Achieving the connections to the national industrial Internet and digital government). **The data center of the integration of cloud and data** is going to be constructed.



Motivation

- Information traffic in power systems.
  - State Grid Electric Power Dispatching & Control Center

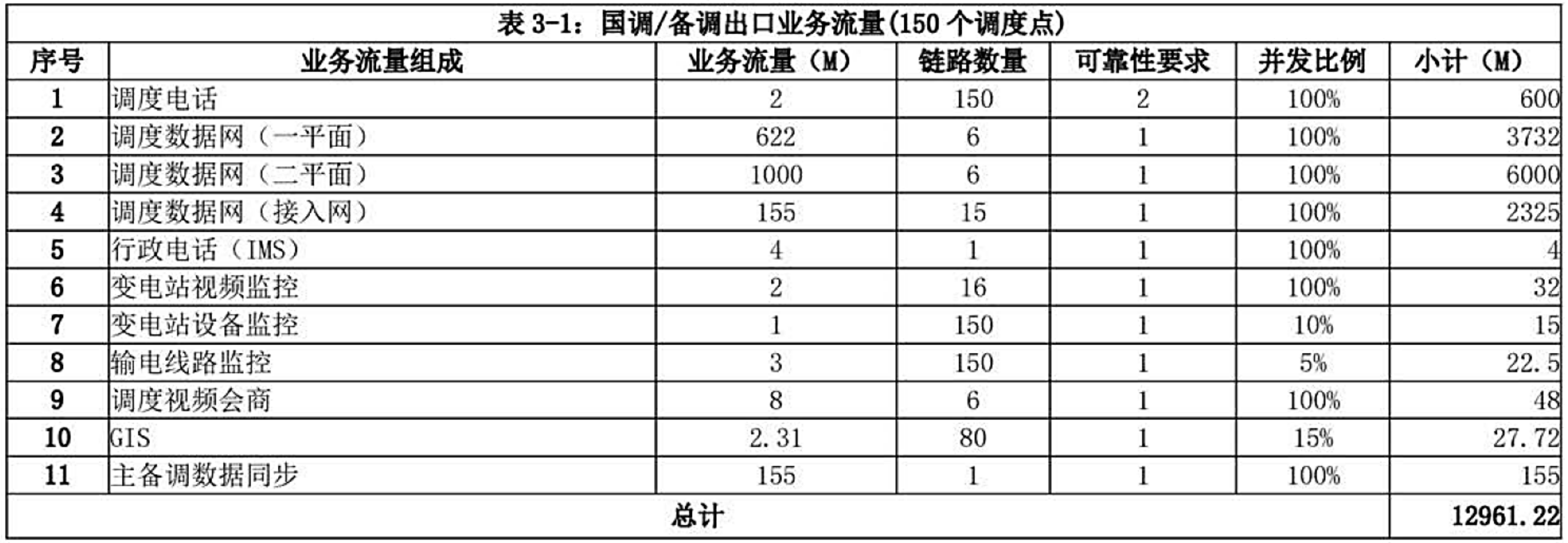


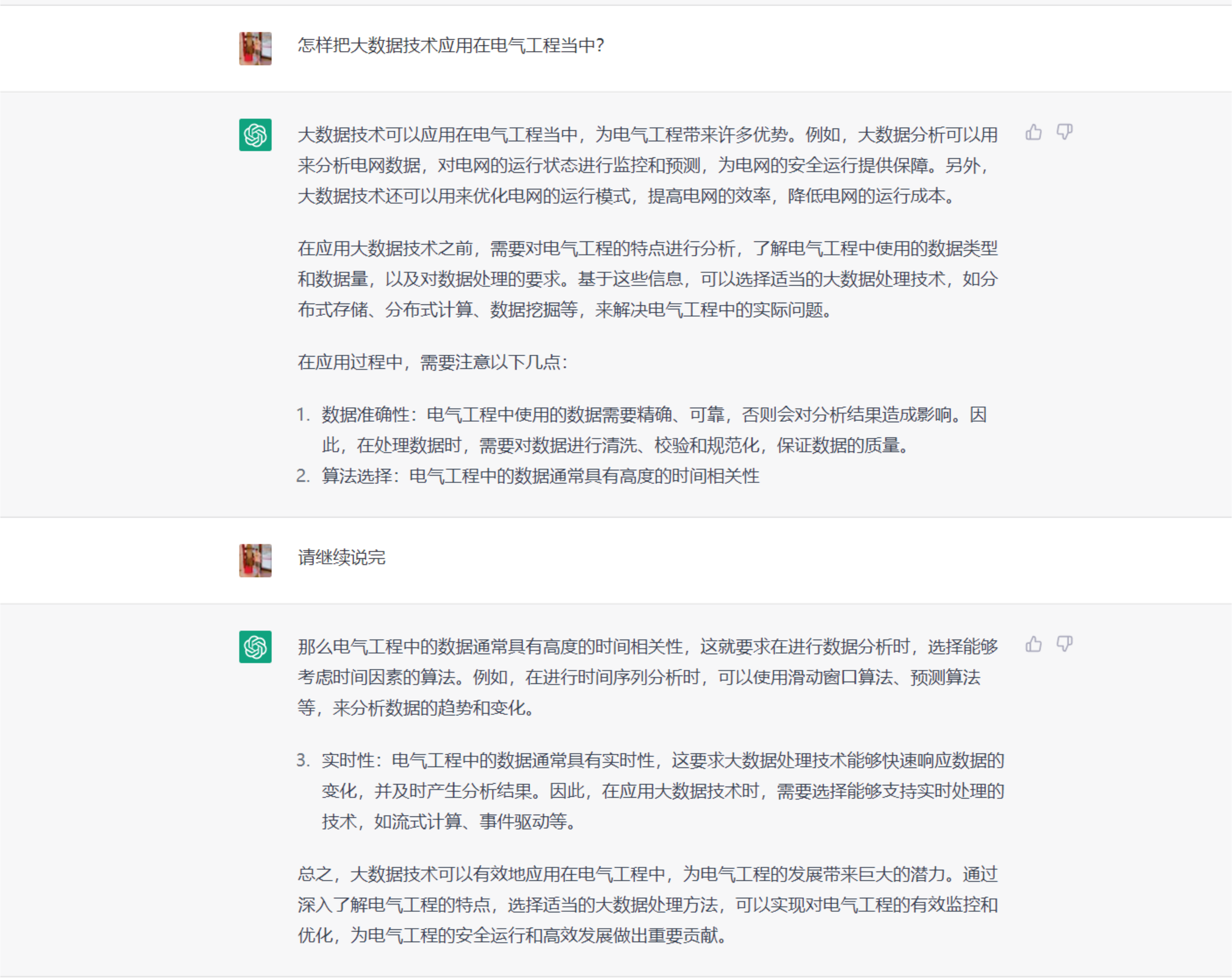
表 3-1：国调/备调出口业务流量(150 个调度点)

序号	业务流量组成	业务流量（M）	链路数量	可靠性要求	并发比例	小计（M）
1	调度电话	2	150	2	100%	600
2	调度数据网（一平面）	622	6	1	100%	3732
3	调度数据网（二平面）	1000	6	1	100%	6000
4	调度数据网（接入网）	155	15	1	100%	2325
5	行政电话（IMS）	4	1	1	100%	4
6	变电站视频监控	2	16	1	100%	32
7	变电站设备监控	1	150	1	10%	15
8	输电线路监控	3	150	1	5%	22.5
9	调度视频会商	8	6	1	100%	48
10	GIS	2.31	80	1	15%	27.72
11	主备调数据同步	155	1	1	100%	155
总计						**12961.22**

  - State Grid Corp. of China (During the 13th Five-Year Plan)
    - SGCC Headquarter：133000 Mbps
    - Regional branch: 39000 Mbps
    - Provincial company: 68000 Mbps

Digitalization of the Power System

Data Acquisition	High-efficiency and low-power intelligent sensors under complex working conditions
Data Transmission	Wide area data transmission with wired (optical fiber…) and wireless (5G…) channels
Data Storage	Centralized and distributed storage of energy data
Data Sharing	The sharing of energy data across systems while ensuring data security and personal privacy
Data Analysis	Big data and AI serving power system generation, transmission, distribution and transformation
Data Decision	Power system planning, operation, control, scheduling, cloud/edge computing, etc.



怎样把大数据技术应用在电气工程当中?

大数据技术可以应用在电气工程当中，为电气工程带来许多优势。例如，大数据分析可以用来分析电网数据，对电网的运行状态进行监控和预测，为电网的安全运行提供保障。另外，大数据技术还可以用来优化电网的运行模式，提高电网的效率，降低电网的运行成本。

在应用大数据技术之前，需要对电气工程的特点进行分析，了解电气工程中使用的数据类型和数据量，以及对数据处理的要求。基于这些信息，可以选择适当的大数据处理技术，如分布式存储、分布式计算、数据挖掘等，来解决电气工程中的实际问题。

在应用过程中，需要注意以下几点:

1. 数据准确性：电气工程中使用的数据需要精确、可靠，否则会对分析结果造成影响。因此，在处理数据时，需要对数据进行清洗、校验和规范化，保证数据的质量。
2. 算法选择：电气工程中的数据通常具有高度的时间相关性

请继续说完

那么电气工程中的数据通常具有高度的时间相关性，这就要求在进行数据分析时，选择能够考虑时间因素的算法。例如，在进行时间序列分析时，可以使用滑动窗口算法、预测算法等，来分析数据的趋势和变化。

3. 实时性：电气工程中的数据通常具有实时性，这要求大数据处理技术能够快速响应数据的变化，并及时产生分析结果。因此，在应用大数据技术时，需要选择能够支持实时处理的技术，如流式计算、事件驱动等。

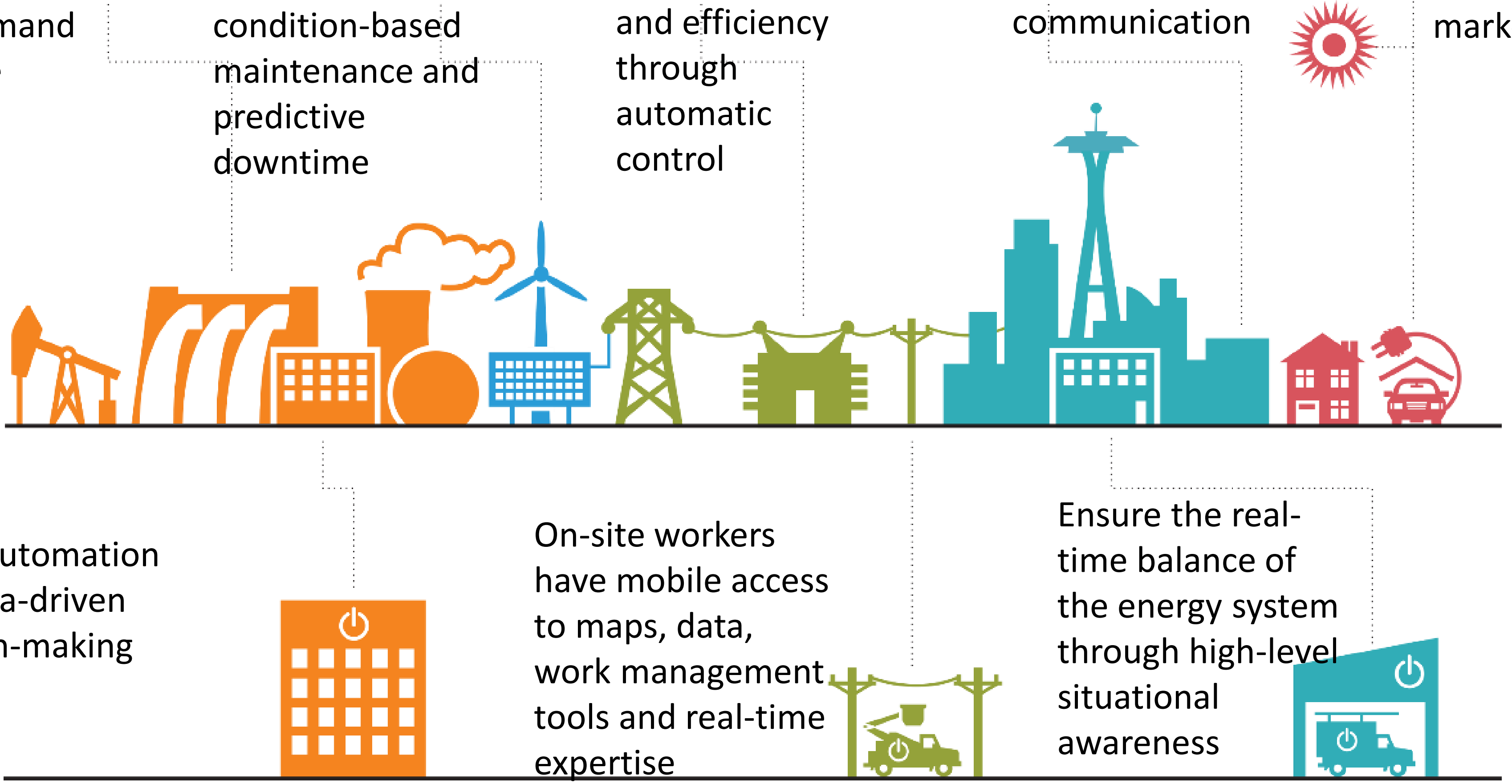
总之，大数据技术可以有效地应用在电气工程中，为电气工程的发展带来巨大的潜力。通过深入了解电气工程的特点，选择适当的大数据处理方法，可以实现对电气工程的有效监控和优化，为电气工程的安全运行和高效发展做出重要贡献。

Which area can the big data be helpful in power system?

- Name three areas that you think big data help the most.

Frontiers of Big Data – Power Systems

- The ***digitization*** of the power system will reach all part of power systems, including power generation, transmission, distribution and transformation.



Big data-driven distributed energy supply and demand balance

Data-driven strategies include preventive and condition-based maintenance and predictive downtime

Smart grid improves system resilience, safety and efficiency through automatic control

Analysis of customer behavior and personalized communication

Distributed energy and its trading market

Cloud automation and data-driven decision-making system

On-site workers have mobile access to maps, data, work management tools and real-time expertise

Ensure the real-time balance of the energy system through high-level situational awareness

Outline

- What is big data
- What is machine learning
- Big data and machine learning in power system
- Course information

Course Overview

Big Data × Power Systems

Fundamentals
Big data thinking
Know-how

Using big data science to solve power system problems

Syllabus

1. Introduction (Python)
2. Mathematics
3. Database
4. ML(1) – Regression
5. ML(2) – Artificial Neural Networks
6. ML(3) – Logistic Regression and SVM
7. ML(4) – Decision Tree
8. ML(5) – Ensemble Learning and Random Forest
9. ML(6) – Unsupervised Learning
10. Big Data in the Demand Side – Forecasting
11. Big Data of Electricity Consumption
12. Big Data Data-driven Power Flow Calculation
13. Power System Topology Identification
14. Overview of Big Data Applications –CV, NLP, and power system analytics
15. Final Assignment

Grading

- Homework (30%)
  - 5~6 assignments: Implementing a machine learning algorithm on Python (Coding)
  - Or solving a mathematic problem on the algorithm (Model constructing)

- Project (70%) = Code + Report + Presentation (All required)
  - Assigned competitive project

- Honor Code: discussion is allowed but copying or sharing the code is not allowed.

Resources

- Lessons
  - CS229: Machine Learning （Stanford）
  - CS231n: Convolutional Neural Networks for Visual Recognition （Stanford）
  - Pattern Recognition, Zhang Xuegong, Tsinghua
  - Introduction to Artificial Intelligence, CST, Tsinghua

- Books
  - An Introduction to Statistical Learning with Applications in R.
  - Introduction to Artificial Intelligence, Stuart Russell and Peter Norvig
  - 模式识别，张学工
  - 数理统计，韦来生
  - 机器学习，周志华
  - Data Mining and Analysis: Fundamental Concepts and Algorithms
  - Applied Multivariate Statistical Analysis.
  - Time Series Analysis with Applications in R
- Website
  - https://scikit-learn.org/
  - https://tensorflow.google.cn/

Instructor

- **Ning Zhang,** Department of Electrical Engineering, Tsinghua University

☐ BS (2007), Ph.D(2012) from Tsinghua University, China.

☐ Post Doc. 2012-2014, Tsinghua university, China.

☐ Research assistant, 2010.10~2011.07, University of Manchester, UK

☐ Research assistant. 2013.12~2014.03, Harvard University, US

☐ Assistant Professor. 2014.09~ 2017.01, Tsinghua University, China.

☐ Associate Professor. 2017.01~, Tsinghua University, China.

☐ **Main research interests:**
  ✓ Renewable energy analytics and grid integration
  ✓ Multiple energy systems
  ✓ Power system planning
  ✓ Power system uncertainty analysis
  ✓ Power system data driven analysis
☐ **Website**: http://www.ningzhang.net/
☐ Meeting: Monday 4:00~5:00 PM, 3-206 West main building

Teaching Assistant

Mingchen Ma

- 2019~2023, B.S. degree in Electrical engineering, Tsinghua University.
- 2023~, Ph.D. student in Electrical engineering, Tsinghua University.

Contact

- West Main Building 3-208
- 15636063070
- mmc23@mails.tsinghua.edu.cn

国庆节及中秋节上课安排

- 根据学校中秋节以及国庆节放假安排，我们的课程原第四节10月2日停上，因此，前四节课的上课时间分别是：
- 第一节：9月11日（校历第1周）星期三第1节
- 第二节：9月18日（校历第2周）星期三第1节
- 第三节：9月25日（校历第3周）星期三第1节
- 第四节：10月9日（**校历第5周**）星期三第1节
- 请大家注意上课时间。

Q&A