西安电子科技大学
XIDIAN UNIVERSITY
计算机科学与技术学院
School of Computer Science and Technology

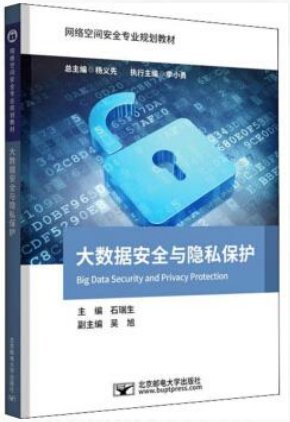
第1章 大数据安全与隐私绪论

彭延国

ygpeng@xidian.edu.cn

§1.0 课程信息 - 简介

- 课程名称：大数据安全与隐私保护
- 课程安排：40学时，2.5个学分
  - 32个理论学时(含8个在线学时)
  - 16个实验学时



- 主要参考书目：
  - 石瑞生，《大数据安全与隐私保护》，北京邮电大学出版社，2019年
  - 冯登国等，《大数据安全与隐私保护》，清华大学出版社，2018年
  - 课程讲义及Slides

§1.0 课程信息 - 辐射面

现代密码学

对称密码、非对称密码、哈希、
数字签名、身份认证等

网络安全协议

Kerberos、X.509、IPSec、
TLS/SSL、DTLS等

大数据处理技术

可搜索加密、安全多方计算、
访问控制、机器学习等

隐私保护技术

匿名化技术、查分隐私技术等

前沿计算模式

云计算、雾计算、区块链等

§1.0 课程信息 - 培养目标

安全与隐私意识

初级阶段

中级阶段

高级阶段

了解技术原理和实践方法

大数据安全与隐私的相关技术、实现方法等

掌握技术原理

洞悉技术原理、理解大数据安全和隐私的核心

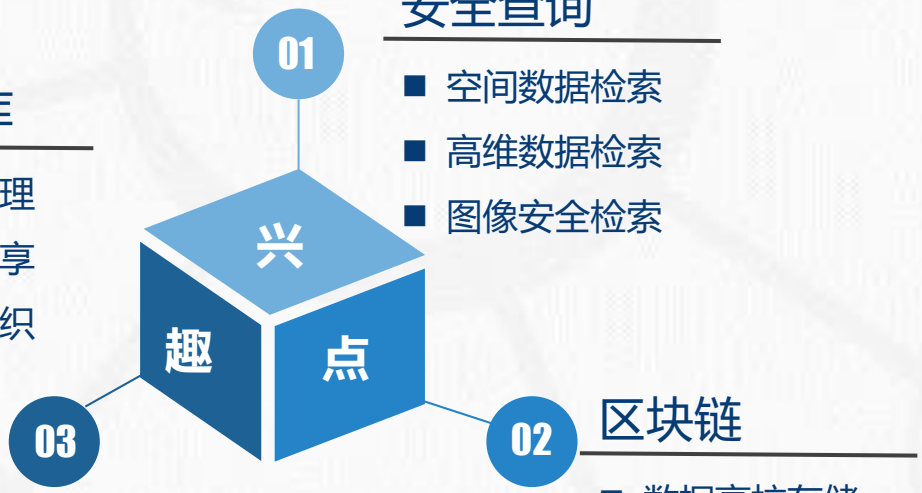
把握前沿动态

终身学习的能力

§1.0 课程信息 - 主讲人简介

个人简历

- 2009-2012 贵州大学攻读硕士学位；
- 2012-2016 西安电子科技大学攻读博士学位；
- IEEE、ACM、CCF、CACR学会会员；
- 国内外主流期刊和会议上共发表学术论文10余篇。
- http://www.ygpeng.cn/



§1.0 课程信息 - 课程考核

- 总成绩(100%)=自主实验(70%)+平时成绩(10%) +期末考核(20%) 。

- 自主实验(70分):
  - 内容：4~5个相关的大数据安全与隐私实验；
  - 验收：视情况进行线上、线下验收；
  - 评分：根据实验完成程度进行灵活打分；
  - 实验报告质量(±10分)。

- 平时成绩(10%):
  - 出勤(含理论课)：-2分/人·次，随机点名，雨课堂；

- 期末考核(20分):
  - 在线考核(通过雨课堂进行)。

大数据安全与隐私 - 2021春-CS205307-01

邀请码:8T359Z

§1.0 课程信息 - 教学团队

- 助教(3名):
  - 研究生: 马文杰、李鑫。
  - 大四: 王利。

- 助教任务:
  - 实验线上/线下验收;
  - 课程群内答疑;
  - 课程考核组织。

§1.0 内容提纲

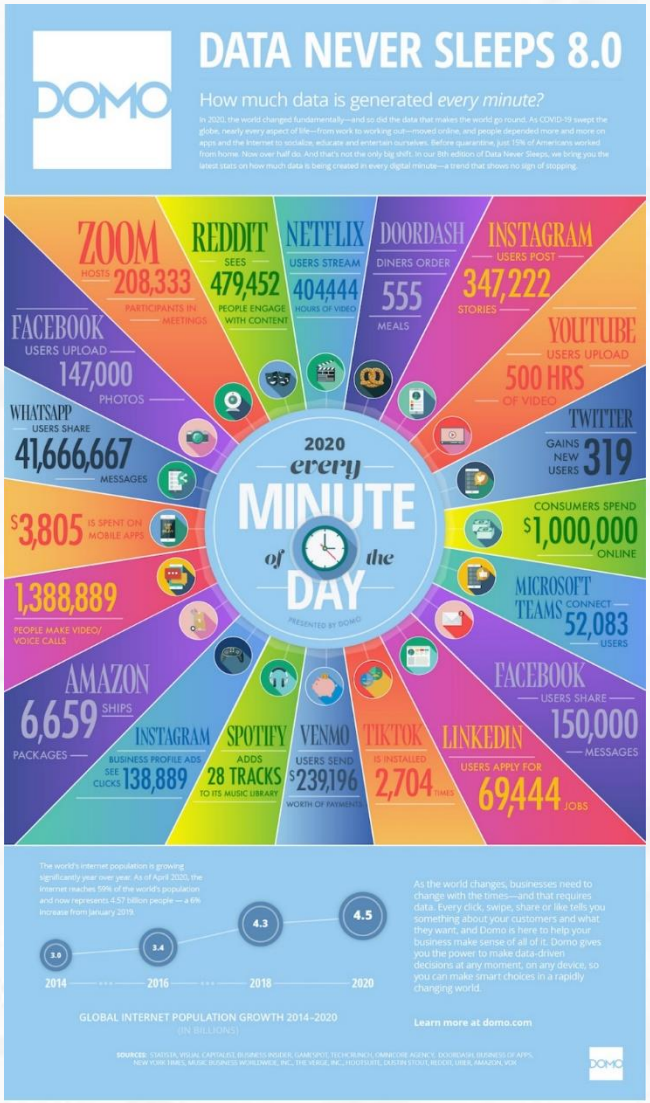
5V特征

§1.1 大数据的概念及内涵

§1.1 大数据的概念及内涵 -大数据时代与安全

- 数据在改变世界
  - 20世纪60年代IBM 360系列计算机的推出，人类开启了计算机的商业化进程。从此，信息技术开始逐步渗透到人类社会生活的方方面面。
  - 经过五十多年的发展，人类已经进入了“互联网+”时代，人类社会生活中的大部分活动都开始与数据的创造、采集、传输、使用发生关系，大数据时代已经伴随着互联网的浪潮悄然而至。
- 安全（隐私）相伴而来
  - 安全技术是一切新兴技术的伴生技术，那么大数据安全作为大数据技术的伴生技术，是我们在大数据时代保障安全的必不可少的技术。

§1.1 大数据的概念及内涵 - 数据永无眠



数据永无眠(2020年数据)

§1.1 大数据的概念及内涵 - 5V特征(1)

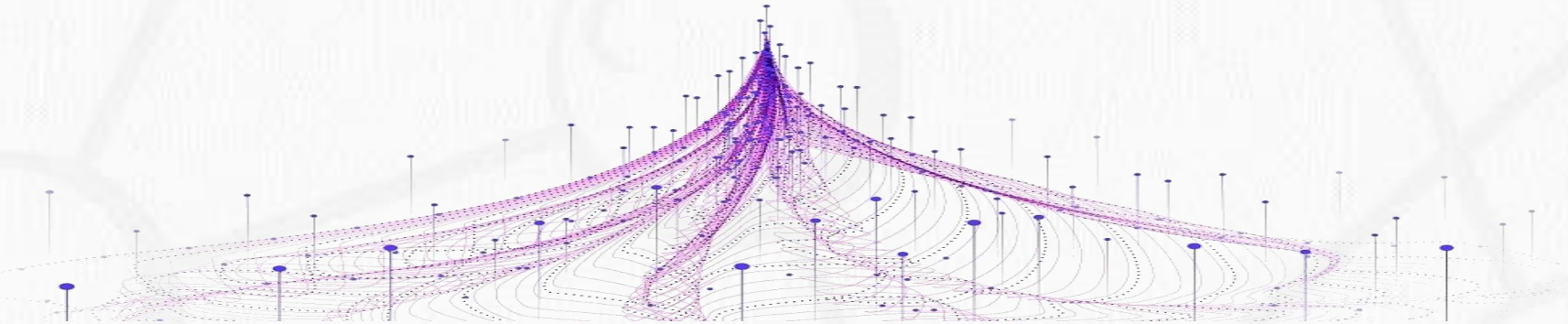
- 数量（Volume）：大数据是指大小超出常规数据库工具获取、存储、管理和分析能力的数据集。
- 速度（Velocity）：大数据不仅数据量大，而且数据产生的速度快，对数据的实时处理能力提出了非常高的要求。
- 种类（Variety）：大数据不仅有传统数据库管理的结构化数据，还有各种非结构化、半结构化数据。
- 价值（Value）：需要有效的技术才能从价值密度低的大量数据中，以可以接受的成本，创造出价值。
- 质量（Veracity）：数据的质量参差不齐。

§1.1 大数据的概念及内涵 - 5V特征(2)

- Volume和Velocity，对信息处理的性能在容量和速度方面提出了挑战，这些问题主要靠系统构建技术来解决，云计算技术、雾计算应运而生。
  - Hadoop、Storm和Spark这些开源系统在各种大数据服务系统中都得到了广泛的应用。
- Variety和Veracity对计算机系统的数据理解能力提出了挑战，它们要求计算机系统不仅能够理解规范的结构化的数据，还要能够理解不规范的半结构化和非结构化数据，甚至需要像人类一样能够识别出错误数据、恶意数据、不准确的数据。
  - 应对这个挑战，需要人工智能技术取得突破性的进展。
- 最终的目标：提取数据的价值（Value）。
  - 从数据中发现价值，在应用中创造价值。大数据，数量虽然大，然而价值密度不见得高。类似于长尾效应，需要有效的技术才能够从价值密度低的大量数据中，以可以接受的成本，创造出价值。
  - 判断一个数据集是否有价值也是很困难的事情：今天也许认为没有价值的数据，将来也许会找到很大的价值。

§1.1 大数据的概念及内涵 - 数据质量

- 数据质量（Veracity）：数据的质量参差不齐。
  - 互联网上的大量无标注数据。
  - 人为的错误数据（网络水军、恶意攻击等）。
  - 物联网设备采集的正常数据中混杂着由于设备故障、环境原因、精度原因导致的错误数据。



西安電子科技大學
XIDIAN UNIVERSITY

大数据意味着大价值

§1.2 大数据的典型应用

§1.2 大数据的典型应用 - 从一段对话开始(1)

某比萨店的电话铃响了，客服人员拿起电话

- 客服：XXX比萨店。您好，请问有什么需要我为您服务?
- 顾客：你好，我想要一份……
- 客服：先生，烦请先把您的会员卡号告诉我。
- 顾客：18646146…
- 客服：陈先生，您好！您是住在XXX路XXX小区一号12楼1205室，您家电话是2646…，您公司电话是8820…，您的手机是1391234…。请问您想用哪一个电话付费?
- 顾客：你为什么知道我所有的电话号码?
- 客服：陈先生，因为我们联机到客户关系管理(CRM) 系统。

§1.2 大数据的典型应用 - 从一段对话开始 (2)

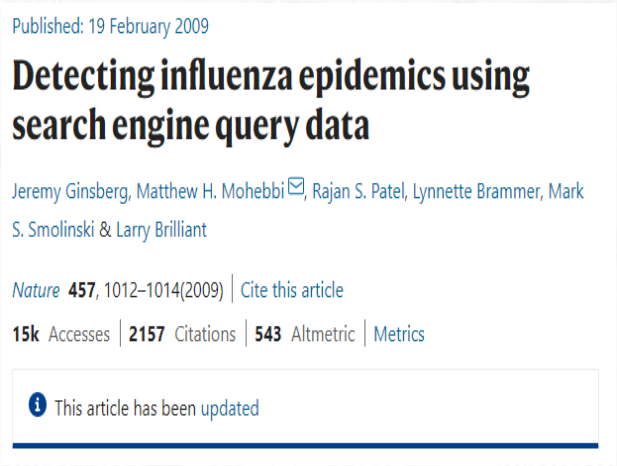
- 顾客：我想要一个海鲜比萨…
- 客服：陈先生，海鲜比萨不适您
- 顾客：为什么?
- 客服：根据您的医疗记录，你的血压和胆固醇都偏高
- 顾客：那你们有什么可以推荐的?
- 客服：您可以试试我们的低脂健康比萨
- 顾客：你怎么知道我会喜软吃这种的?
- 客服：您上星期一在中央图书馆借了一本《低脂健康食谱》
- 顾客：好。那我要家庭特大号比萨要付多少钱?
- 客服：99元，这个足够您一家六口吃了，但您母亲应该少吃，她上个月刚刚做了心脏搭桥手术，还处在恢复期

§1.2 大数据的典型应用 - 从一段对话开始(3)

- 顾客：那可以刷卡吗?
- 客服：陈先生，对不起。请您付现款，因为您的信用卡已经刷爆了，您现在还欠银行4807元，而且还不包括房贷利息。
- 顾客：那我先去附近的提款机提款
- 客服：陈先生，根据您的记录，您已经超过今日提款限额
- 顾客：算了，你们直接把比萨送我家吧，家里有现金。你们多久会送到?
- 客服：大约30分钟，如果您不想等，可以自己骑车来。
- 顾客：为什么?
- 客服：根据我们 CRM 全球定位系统的车辆行驶自动跟踪系统记录。您登记有一辆车牌号为58-748的摩托车，而目前您正在XXX路东段XXX商场右侧骑着这辆摩托车

§1.2 大数据的典型应用 - 典型应用

- 健康：谷歌流感趋势（Google Flu Trends, GFT）未卜先知的故事，常被看作大数据分析优势的明证。
- 经济：华尔街利用微博数据预测股票。
- 政治：利用大数据预测美国大选。


Published: 19 February 2009

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski & Larry Brilliant

Nature **457**, 1012–1014(2009) | Cite this article

15k Accesses | **2157** Citations | **543** Altmetric | Metrics

This article has been updated

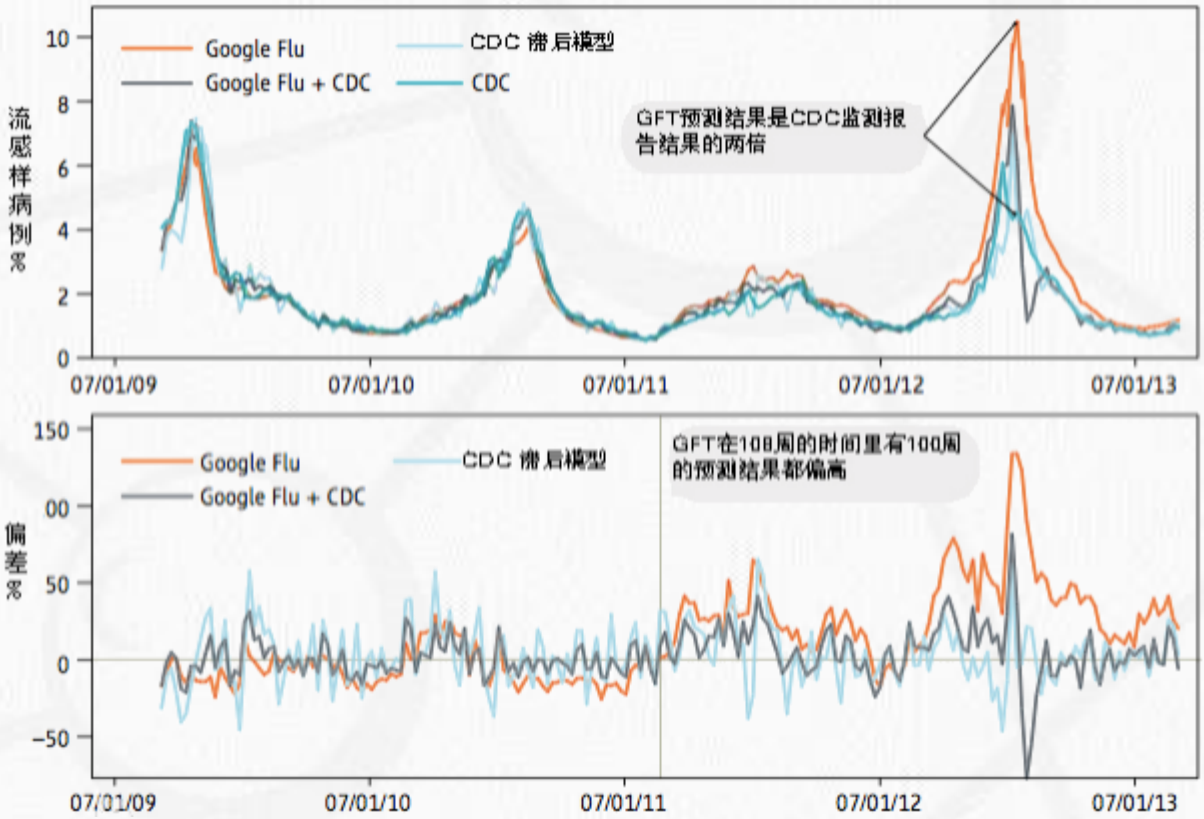
Google利用网络大数据预测流感

华尔街利用微博数据预测股票

利用大数据预测美国大选

§1.2 大数据的典型应用 - 谷歌流感趋势

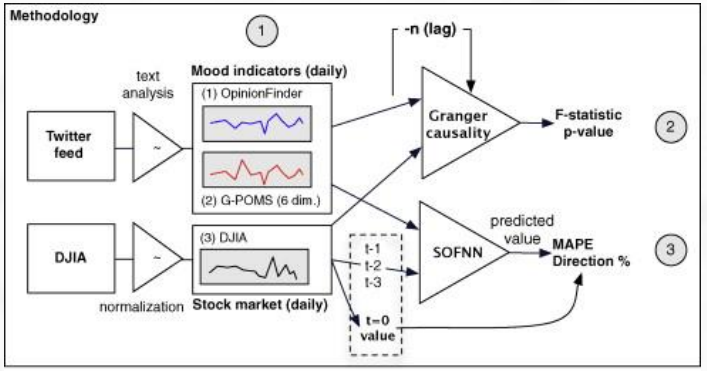
- Google流感趋势（Google Flu Trends，GFT）是Google于2008年推出的一款预测流感的产品。Google认为，某些搜索字词有助于了解流感疫情。Google流感趋势会根据汇总的Google搜索数据，近乎实时地对全球当前的流感疫情进行估测。
- 技术应用：
  - Logistic回归
- 热门原因：
  - 大就是一切！
  - 大数据解决大问题！
  - 不需要复杂算法！
- 结论

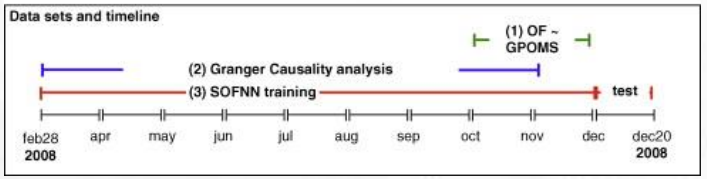


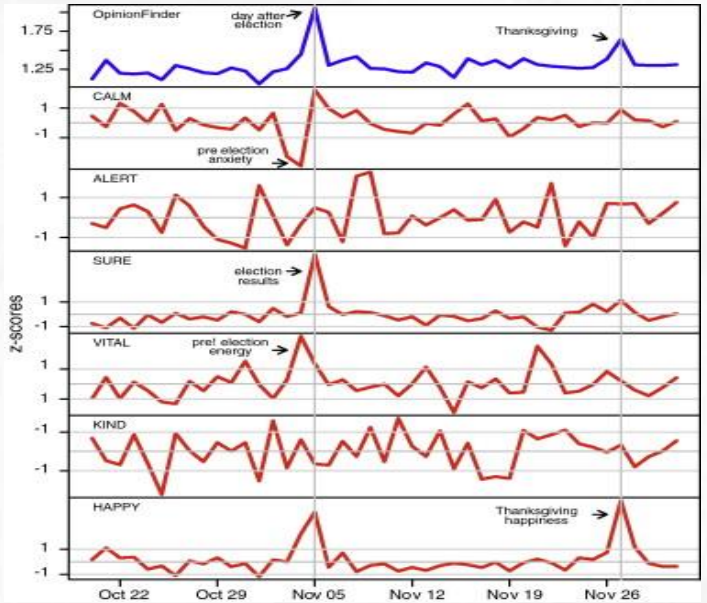
[1] J Ginsberg, M H Mohebbi, R S Patel, et al. Detecting influenza epidemics using search engine query data[J]. Nature, 2009.

§1.2 大数据的典型应用 - 利用微博数据预测股票

- 华尔街利用微博数据预测股票
  - 2008年2月28日至2008年12月19日近1000万条推特文本。
  - 推特反映出的情绪能在一定程度上预测3~4天后的股市变化。
  - 模型预测的准确性73.3%~86.7%。
- 2013年，美国证监会允许上市公司在社交网络披露公司信息。
- 之后，包括汤森路透、彭博社等提供社交网络数据分析服务。



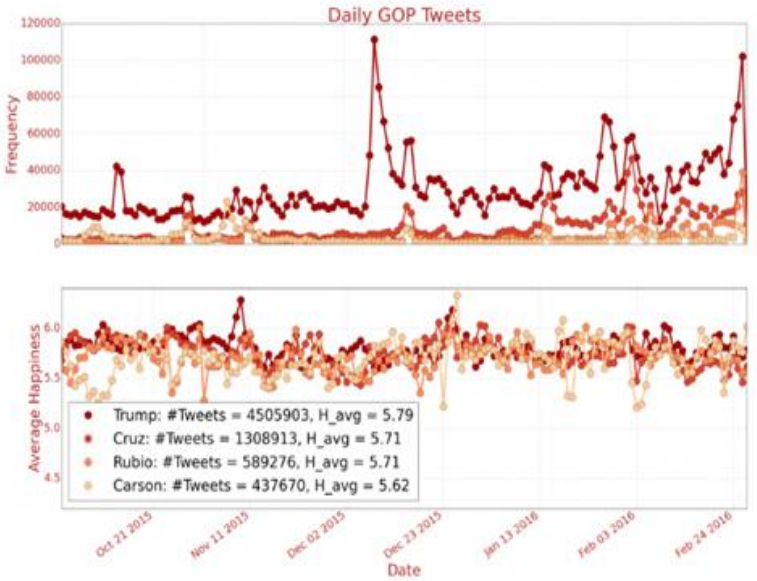




[1] J Bollena, H Mao, X Zeng. Twitter mood predicts the stock market[J]. Journal of Computational Science. 2011.

§1.2 大数据的典型应用 - 美国大选

- 2012年美国大选(奥巴马VS.罗姆尼):
  - 奥巴马团队拥有 “核代码”：数据挖掘
  - 80% 的美国选民认为奥巴马比罗姆尼让他们感觉更加重视自己
  - 奥巴马团队筹得的第一个 1 亿美金中， 98% 来自于小于 250 美金的小额捐款
- 2016年美国大选(特朗普VS.希拉里):
  - 特朗普在推文中的提及率占有明显的领先优势
  - 特朗普背后的大数据团队着重于希拉里过去的演讲，通过关键词和数据分析来洞悉希拉里演讲的漏洞和缺点，从而为特朗普提供有力的攻击武器。
  - 希拉里有一支堪比硅谷公司的大数据团队——50名专业的程序员和开发者。



§1.2 大数据的典型应用 - 商业上的成功

- 大数据的商业价值主要体现在对“人”的画像，通过人的数据对于人的需求或者潜在需求做出判断，从而及时精准地为人提供产品/服务，获得商业利益。

大数据全生命周期、安全、隐私

§1.3 理解大数据安全

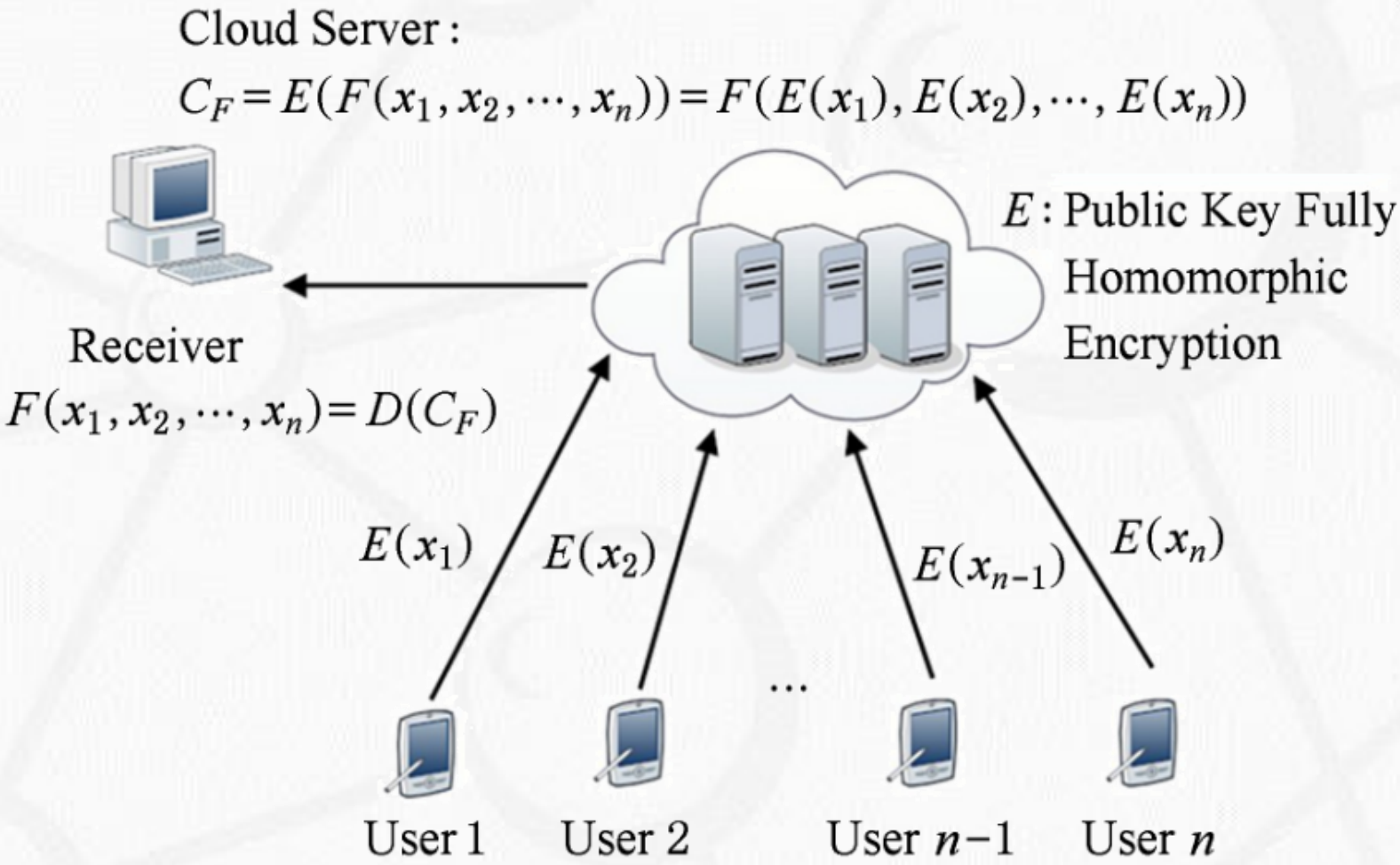
§1.3 理解大数据安全 - 如何理解大数据安全?

- 怎么理解大数据安全呢?
  - 大数据安全即针对大数据服务系统，从系统架构与认证授权、计算与存储、算法设计与数据采集等多个角度来分析其安全问题及解决方案。
  - 同时，大数据技术也可以作为解决安全问题的技术手段，加强系统安全防护能力。技术都是双刃剑，攻击者基于大数据技术，也会具备更强的攻击能力、创造出新的攻击模式。

- 大数据服务中的安全问题
  - 大数据采集、传输中的安全和隐私。
  - 大数据存储的安全和隐私。
  - 大数据处理及其安全隐私。
  - 大数据共享使用及其安全隐私。

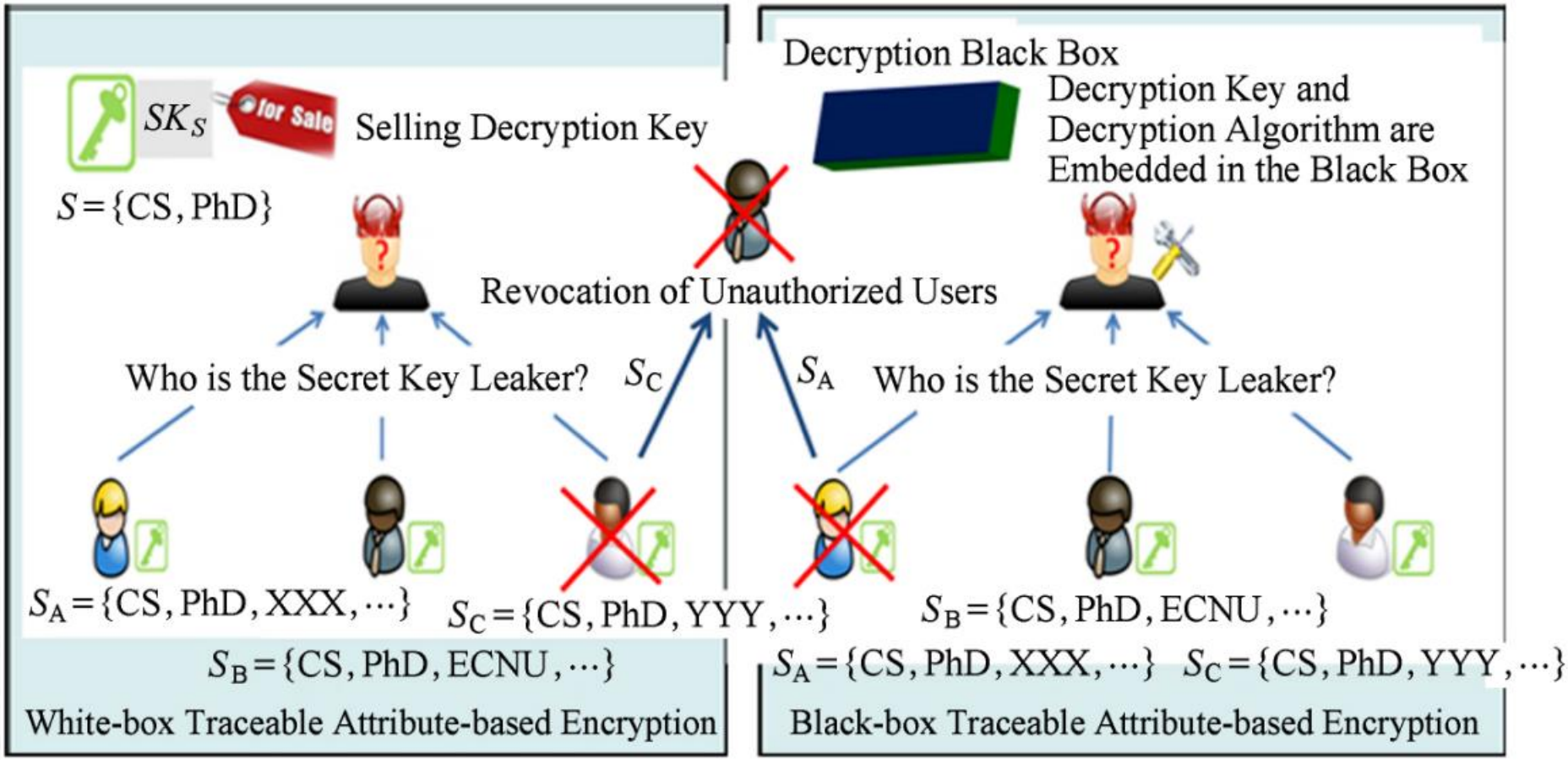
§1.3 理解大数据安全 - 采集与传输

- 典型问题：大数据服务系统架构与认证授权问题



§1.3 理解大数据安全 - 存储

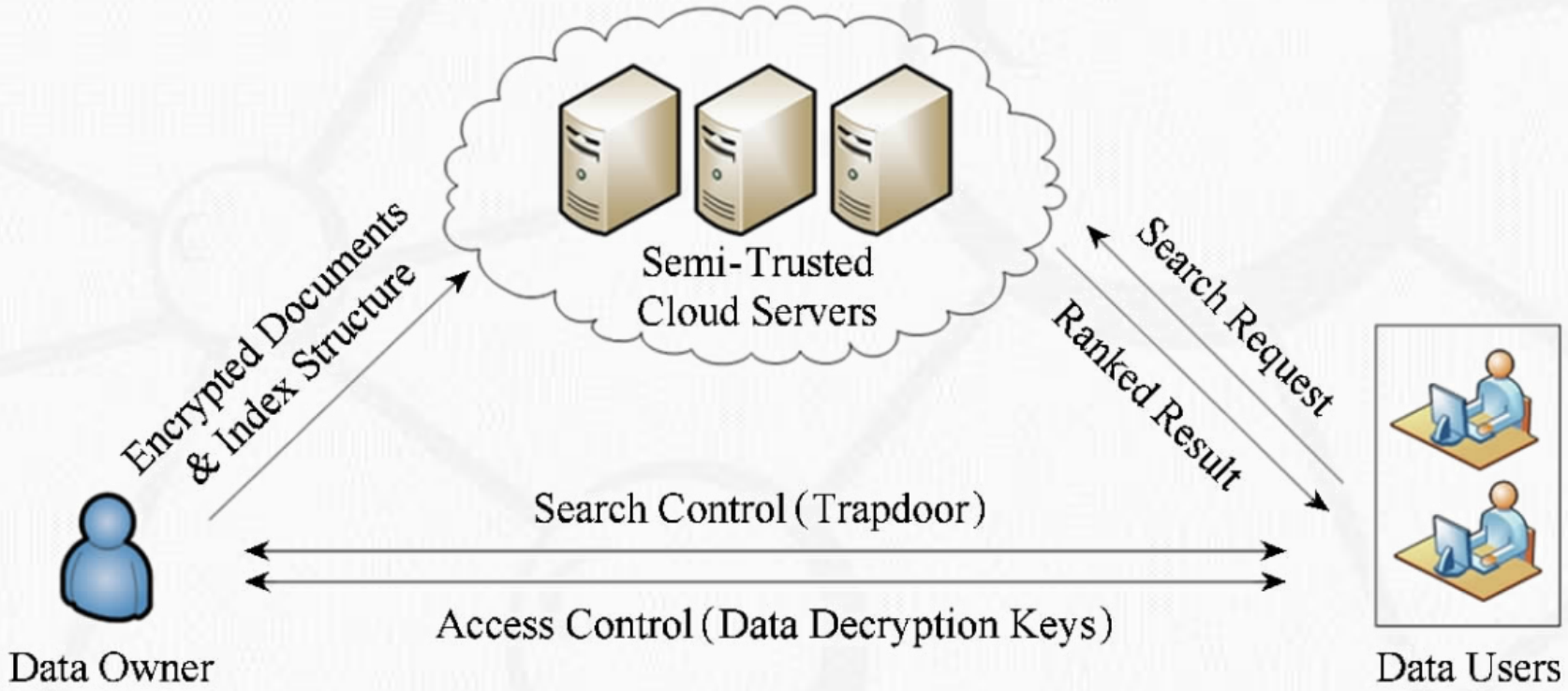
- 典型问题：大数据的访问控制



基于属性加密的访问控制

§1.3 理解大数据安全 - 处理

- 典型问题：加密数据的可用性问题(可搜索性、可计算性等)



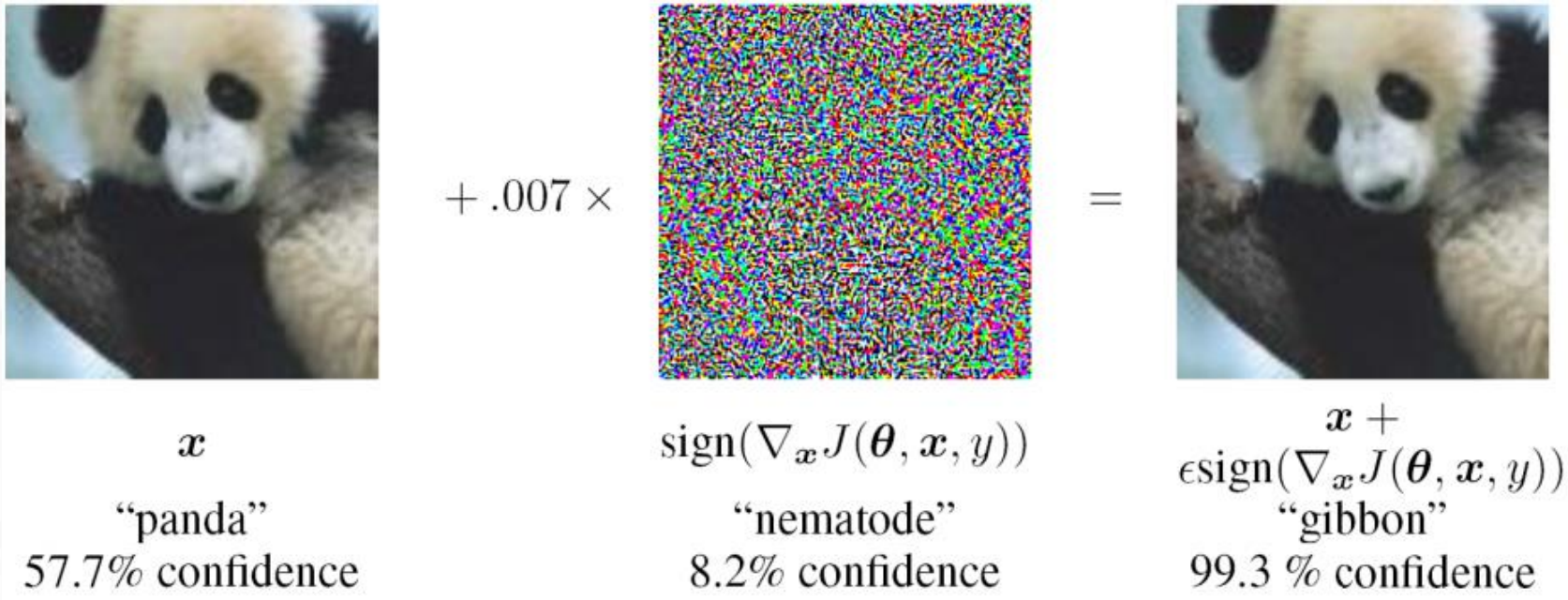
§1.3 理解大数据安全 - 共享使用（1）

- 典型问题：利用大数据服务的系统漏洞，攻击系统功能
  - “后门攻击（backdoor attack）”或“木马攻击（trojan attack）”是一种危害性很大的投毒攻击，它使攻击者能够将“后门”或“木马”植入到模型中，并在预测阶段通过简单的后门触发器完成恶意攻击行为。被植入“后门”的深度神经网络在正常样本上表现很好，但会对具有特定后门触发器的输入样本做出特定的错误预测。

停车标志及其受后门攻击的版本，后门触发器（从左到右）为黄色方块、炸弹和花朵

§1.3 理解大数据安全 - 共享使用（2）

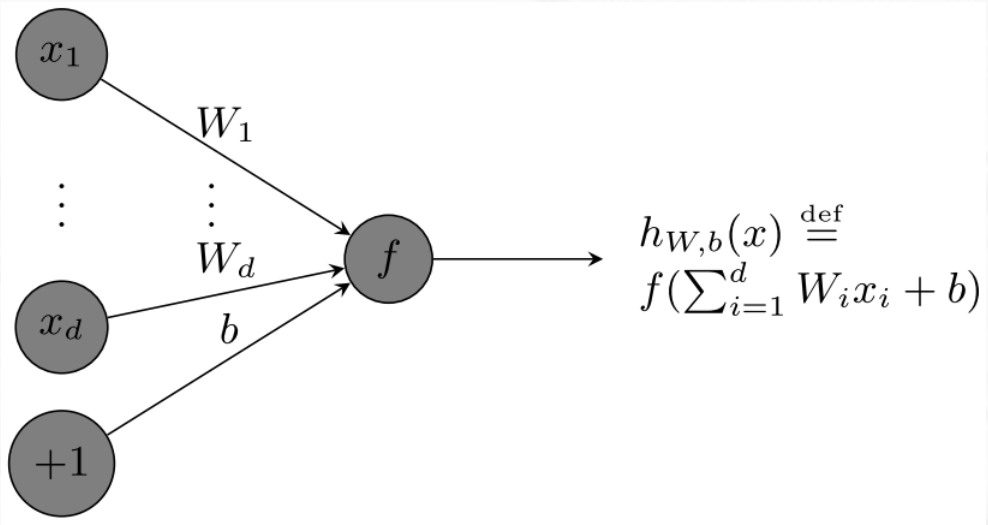
- 典型问题：利用大数据服务的算法缺陷或者训练数据管理的漏洞，注入攻击数据

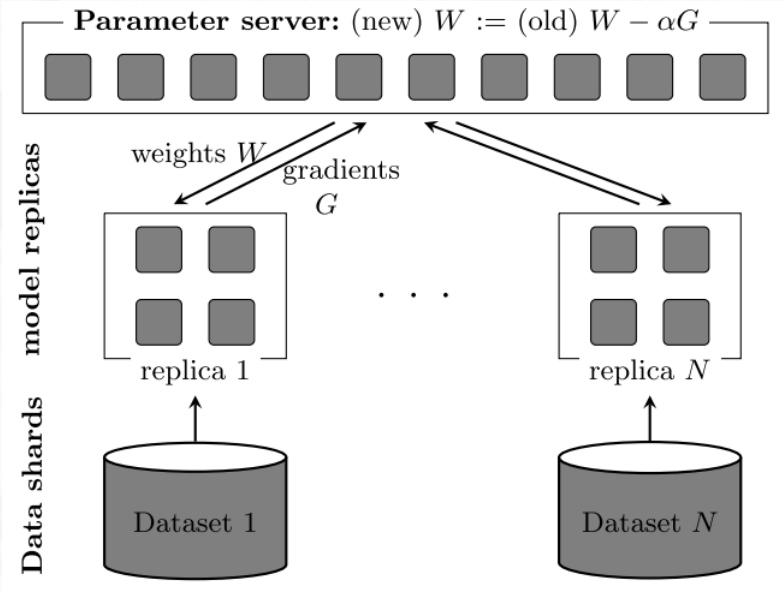


对抗样例攻击示例

§1.3 理解大数据安全 - 非法使用

- 利用大数据服务的合法功能，获取隐私信息





$$\eta_k = \frac{\delta J(W,b,x,y)}{\delta W_k} = 2(h_{W,b}(x) - y) f'(\sum_{i=1}^{d} W_i x_i + b) \cdot x_k$$

$$\eta = \frac{\delta J(W,b,x,y)}{\delta b} = 2(h_{W,b}(x) - y) f'(\sum_{i=1}^{d} W_i x_i + b) \cdot 1$$

在联邦学习深度神经网络中，根据梯度获取输入训练数据的信息

西安電子科技大學
XIDIAN UNIVERSITY

兼顾安全与隐私

§1.4 隐私的概念及其发展

§1.4 隐私的概念及其发展 - 隐私的概念与起源

- 大数据安全与隐私问题要求我们能够在大数据时代兼顾安全与自由、个性化服务与商业利益，在保护国家安全与个人隐私的基础上，从数据中挖掘潜在的巨大商业价值和学术价值，使得数据资源能够被充分利用，造福社会，造福人类。
  - 1890年，布兰代斯与他的同学沃伦在《哈佛法律评论》上共同发表关于隐私权的奠基之作《隐私权》。
  - 1965年，美国中央数据银行计划，把各政府部门的数据库打通整合，给全国每一个人建立数据档案，这个档案将包括每个人的教育、医疗、纳税、犯罪等等记录。
  - 1974年，尼克松的水门事件，美国《隐私法案》（Privacy Act of 1974）诞生。
  - 2002年，美国在《2002国土安全法》中重新提出中央数据银行计划，还赋予该计划一个更响亮的名字：万维信息触角计划（Total Information Awareness）。
  - 2013年，斯诺登事件导致美国的“棱镜计划”被曝光。
  - 2018年，欧洲联盟的《通用数据保护条例》于2018年5月25日开始实施。
  - 2018年，中国的《信息安全技术 个人信息安全规范》于2018年5月1日起正式实施。

§1.4 隐私的概念及其发展 - 技术进步与隐私保护

- 1876年，贝尔发明了电话。

随着电话网络的广泛应用，电话监听成为了侵犯人类隐私的一种方式。1974年的水门事件，是电话监听事件的重大事件。

- 1888年，美国柯达公司发明了世界上第一台安装胶卷的可携式方箱照相机。

开始发生侵犯肖像权的案件。

- 1946年，电子计算机被发明。

1962年，IBM开始采用集成电路技术设计计算机；
1964年，IBM 360系列计算机的推出，开启了计算机的商业化进程；
1965，美国中央数据银行计划浮出水面。

- 1989年，万维网的出现，开启了一个新的时代。

2002年，《2002国土安全法》中重新提出的中央数据银行计划，有一个更响亮的名字：万维信息触角计划。
2013年7月，斯诺登事件，美国的“棱镜计划”被曝光。

§1.4 隐私的概念及其发展 - 谁对你的隐私感兴趣?

- 随着互联网的兴起，网络隐私成为一个大家日益关注的问题
- 谁在侵犯大众的隐私?

➢ 政府：公共安全；国际政治
—2013年7月，斯诺登事件，美国的“棱镜计划”被曝光
—数据与元数据

➢ 企业：经济利益
—苹果，谷歌，BAT等互联网公司

➢ 黑客及一些犯罪组织：黑产

- 网络空间的黑色产业链
  - ➢ 系统漏洞，被明码标价
  - ➢ 入侵工具，像武器一样可以购买
  - ➢ 水军，花钱就能够发起内容攻击
  - ➢ 云计算技术，使得普通人的攻击能力日益增强

§1.4 隐私的概念及其发展 - 数据从封闭走向开放

- 思考一个问题：我们的数据在哪儿?
  - 大数据时代，一个很大的变化是我们的数据已经不仅仅保存在于我们的设备里。
  - 各种服务商：日常购物、通信服务、交通运输、水电能源、医疗教育、旅行消费……
  - 各种公共基础设施：公共监视、强制的信息登记……

- 我们所管理的数据（存储）分布在互联网上的各种账号中，通过这些账号才能够访问我们的个人数据
  - （各种网络服务的）账号安全机制能够帮助我们抵御黑客及其他攻击者的非法访问。
  - 但是，管理好这些账号（保证账号的安全）其实是一件非常困难的事情。

§1.4 隐私的概念及其发展 - 如何保护我们的隐私

- 我们所管理的数据（存储）分布在互联网上的各种账号中，通过这些账号才能够访问我们的个人数据。
  - （各种网络服务的）账号安全机制能够帮助我们抵御黑客及其他攻击者的非法访问。
- 隐私，其实不仅仅是这些我们直接管理的数据，还有很多不为我们所控制的数据。
  - 各种元数据：从“窃听”变为“监视”，例如：通话记录，而不是通话内容。
  - 各种行为数据：从网络空间到物理空间。

信息化和智能化的时代，我们被无缝、全程的信息“记录”着。
We are ‘seamlessly’ recorded by all time, in this information society.

§1.4 隐私的概念及其发展 - 网络空间

- 网络空间：个人网络行为被跟踪
  - 搜索记录，浏览记录
    - 我们什么时间，搜索过什么关键字，浏览过哪些网页
  - 电子邮件
    - 我们的电子邮件都保存在电邮供应商的日志文件中
  - 通话记录
    - 我们的通话记录都被加上时间标记备份在电话公司的大容量硬盘上；
  - 信息发布与社交网络
    - 我们所有的个人网页、空间、微博、短视频文件，还有博客的信息都被保存在多个服务器上
  - 购物记录
    - 我们何时何地买了什么东西，我们的喜好、品味以及支付能力都被信用卡提供商编目归档

§1.4 隐私的概念及其发展 - 物理空间

- 物理空间：行为被网络获取
  - 即时行踪：定位服务，WIFI服务，电信运营商的蜂窝网络
    - 我们的即时行踪完全被手机供应商和电信运营商所掌握
  - 容貌与打扮：谷歌街景
    - 我们的容貌和穿着打扮都被安装在各大商场和街角的摄像头捕捉并记录
  - 典型事件：
    - 谷歌地球，谷歌街景对个人隐私的侵犯
    - 苹果手机收集用户位置信息事件

西安电子科技大学
XIDIAN UNIVERSITY

欧盟、国内

§1.5 大数据安全与隐私的法律法规

§1.5 相关法律法规 - GDPR

- 欧盟《通用数据保护条例》（General Data Protection Regulation, GDPR）是20年来数据隐私条例的最重要变化。

- 将协调全欧洲的数据隐私法律，为所有欧盟民众保护和授权数据隐私，并将重塑整个地区的数据隐私保护形式。

- 2018年5月25日，GDPR在欧盟全面实施。

对于欧盟公民来说，GDPR增加技术公司在收集用户数据时的责任，从而保护了公民权利。

§1.5 相关法律法规 - GDPR规则

- 针对从欧盟公民处收集数据的企业：强制企业遵循Privacy by Design原则。
  - 数据转移权：该规定声明，用户可要求自己的个人数据畅通无阻地直接迁移至新的提供商，数据以机器可读的格式迁移。
    - 当用户不再使用该公司产品时，它们将会丢失大量数据。
  - 被遗忘权：每个数据主体有权要求数据控制者删除个人数据，并且不能过分延长数据留存时间。
  - 算法公平性：数据主体有权要求对算法自动决策给出解释。
    - 例如，如果贷款申请人被自动决策拒绝时，有权寻求解释。
    - 对于技术公司而言，这是对人工智能的严重限制，将大幅减缓AI技术的发展。

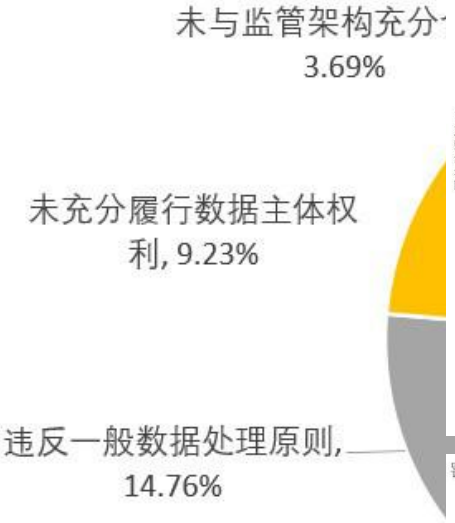
§1.5 相关法律法规 - GDPR案例（1）

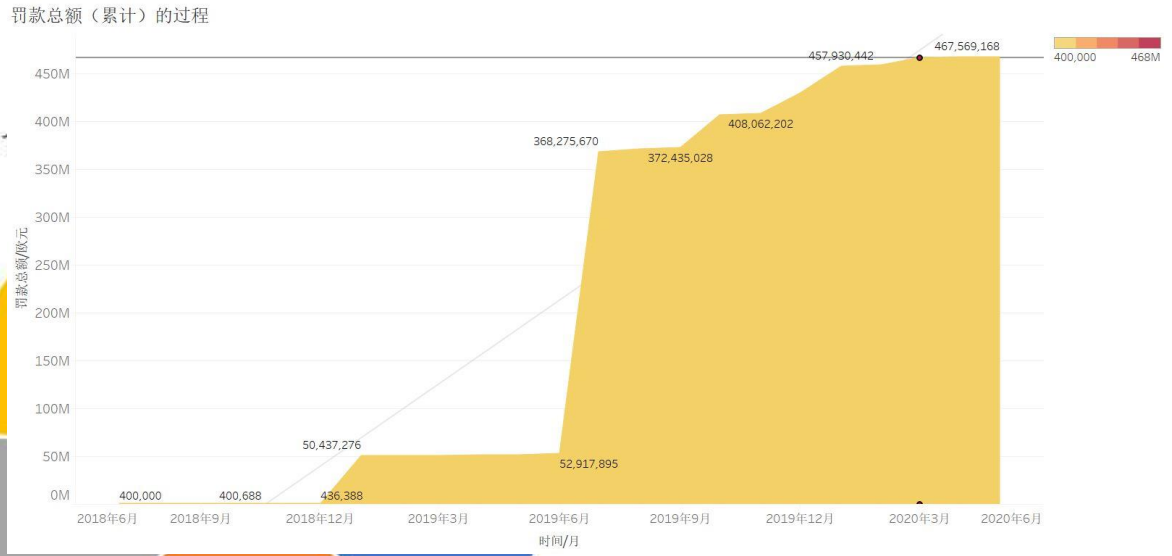
- 2016年，乐购银行承认，该银行约4万个账户被黑客攻击，其中2万个账户资金被盗。
  - 乐购银行到2016年9月底的年营业额是9.53亿英镑，但有着银行及连锁超市的乐购集团的营业额却是484亿英镑。如果此次数据泄露发生在GDPR实施之后，4%的数据隐私侵权集体诉讼罚单，将令该公司承受高达19.36亿英镑的罚款。
- 2018年5月28日的报道称，Facebook和谷歌等美国企业成为GDPR法案下的第一批被告。
  - 这些公司违反了GDPR，因为它们采取了“要么接受，要么放弃”的做法，即客户必须同意收集、共享他们的数据并将其用于定向广告，或者删除他们的账户。GDPR禁止强迫人们接受广泛的数据收集以换取使用服务的行为。

§1.5 相关法律法规 - GDPR案例（2）

- 2019年1月22日，谷歌因GDPR违规行为被罚款5700万美元。
  - 审查违规行为的委员会发现了两类违反GDPR的行为：违反了透明度和信息义务，违反了为广告个性化处理提供法律依据的义务。
- 2019年7月8日，对英国航空公司开出1.83亿英镑巨额罚单。
  - 2018年客户数据遭泄露事件。2018年9月，英国航空公司表示，该公司数据遭窃，从8月21日到9月5日，大约38万笔交易受影响。
- 仅仅一天后，2019年7月9日对万豪处以1.24亿美元的罚款。
  - 原因是万豪2018年发生客户数据泄露事件。2018年11月30日，万豪国际集团官方微博发布声明称，客房预订数据库被黑客入侵，在2018年9月10日或之前曾在该酒店预定的最多约5亿名客人的信息或被泄露。

§1.5 相关法律法规 – GDPR案例（3）





§1.5 相关法律法规 - 网络安全法

- 2017年6月1日正式实施《中华人民共和国网络安全法》。
  - 明确网络空间主权的原则
  - 明确了网络产品和服务提供者的安全义务
  - 明确了网络运营者的安全义务
  - 进一步完善了个人信息保护规则
  - 建立了关键信息基础设施的安全保护制度
  - 确立了关键信息基础设施重要数据跨境传输的规则

§1.5 相关法律法规 - 国内法规

- 2018年1月，由全国信息安全标准化技术委员会组织制订的国家标准《信息安全技术 个人信息安全规范》获批发布全文。
- 尽管这是一部推荐性的国家标准，不具有强制力，但仍引起了学界与实务界的广泛关注。
- 《规范》主要有两方面的亮点，
  - 一是在《网络安全法》和“两高司法解释”的基础上，明确了个人信息处理活动中各项术语的定义，例如“个人信息控制者”“收集”“明示同意”“用户画像”“个人信息安全影响评估”“删除”“去标识化”等。
  - 二是对个人信息收集、保存、使用、转让和披露、通用安全各个环节提出了非常明确具体的要求。

§1.5 相关法律法规 - 数据安全法

- 2020年7月3日，《中华人民共和国数据安全法（草案）》全文在中国人大网公开征求意见。
  - 确定数据作为生产要素
  - 确立了行业安全责任、监管与统筹协调主体
  - 更加重视数据安全制度的建设
  - 明确了数据分级分类保护制度
  - 要求数据活动要具备制度+技术+培训+其他安全措施
  - 首次确定数据交易中介服务机构的义务
  - 弥补电信条例上位法缺失问题

§1 小结

- 内容回顾
  - 大数据的概念及内涵：大数据5V特征
  - 大数据的典型应用：生活、健康、经济、政治等
  - 大数据隐私与安全
  - 大数据安全与隐私的法律法规

- 掌握
  - 大数据5V特征

Thanks!
Questions & Advices!