北京郵電大學
Beijing University of Posts and Telecommunications

大数据安全的概念

石瑞生
网络空间安全学院

提纲

- 大数据的概念
- 大数据的应用
- 大数据安全
- 隐私的概念及其发展
- 相关法规

大数据时代与大数据安全

- 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。
  - 互联网的高速发展和广泛应用使人类进入了真正的大数据时代。
- 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。

大数据概念与应用

大数据的概念

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。例如，网页、图片、视频，等等。对于非结构化、半结构化数据的处理，需要引入传统关系型数据库技术之外的新的数据处理技术。

多样性(variety)：数据类型复杂

高速性(velocity)：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。

例如，微博数据，不仅数量大，而且时效性高。
如果按照传统的搜索引擎的模式去处理，花上几天甚至几周时间去做数据采集、建立索引，这些数据由于对时效性要求高（例如，新闻事件，应急事件，等），传统的出具处理方法在这种场景下不再有效。

大规模(volume)：超出常规数据库工具的处理能力。例如， 搜索引擎。

最早的大数据服务系统就是搜索引擎系统，它需要采集互联网上所有的网页并为其建立索引、对全球几十亿用户提供实时的网页搜索服务。

面对这么大规模的数据集，传统的信息技术无能为力。谷歌公司为了应对大数据的挑战，设计了MapReduce计算模式、GFS分布式文件系统、BigTable数据管理系统，成为云计算技术的先驱。

Veracity准确性：数据质量

- 网上的大量无标注数据、人为错误数据；
- 物联网采集的正常数据中混杂着由于设备故障、环境原因、精度原因造成的错误数据

劣质信息的挑战，包括数据采集系统（例如，传感器网络）本身的质量问题，社交媒体上的谣言，电子商务网站上的水军，等等。

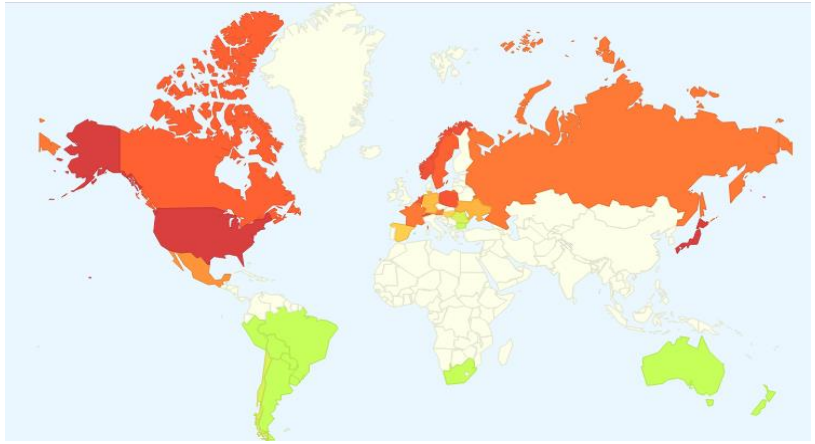
技术挑战

- Volume和Velocity，对信息处理的性能提出了挑战（更大、更快），主要靠系统构建技术来解决，云计算技术应运而生。
  - Hadoop、Storm、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，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。
  - 人机结合：众包，人脑接口，等。
- 最终的目标：提取价值（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。
  - 而且，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。

大数据的应用

- 大数据技术在很多领域得到了广泛的应用，对人类的生活、健康、经济、政治等方方面面产生了重要的影响。
- 介绍几个广为人知的大数据应用的成功案例，来帮助我们建立起对大数据的感性认识。
  - 先讲一个发生在美国的真实的故事：几年前，一个美国家庭收到了一家商场投送的关于孕妇用品的促销券，由于很明显促销券是冲着这个家庭中的那位16岁女孩来的，女孩的父亲觉得受到了侮辱，于是怒气冲冲地找到了这家商场讨说法。为了平息这位父亲的怒气，商场做出了诚恳的道歉。但数天后，这位父亲赫然发现，其16岁的女儿真的未婚先孕了。
  - 那家商场之所以能未卜先知地知道该女孩怀孕，是因为该商场通过若干种商品的消费数据建立了一个怀孕预测指数，以此来预知其顾客的怀孕情况。可以说，这只是一个典型的大数据应用案例。

大数据的应用



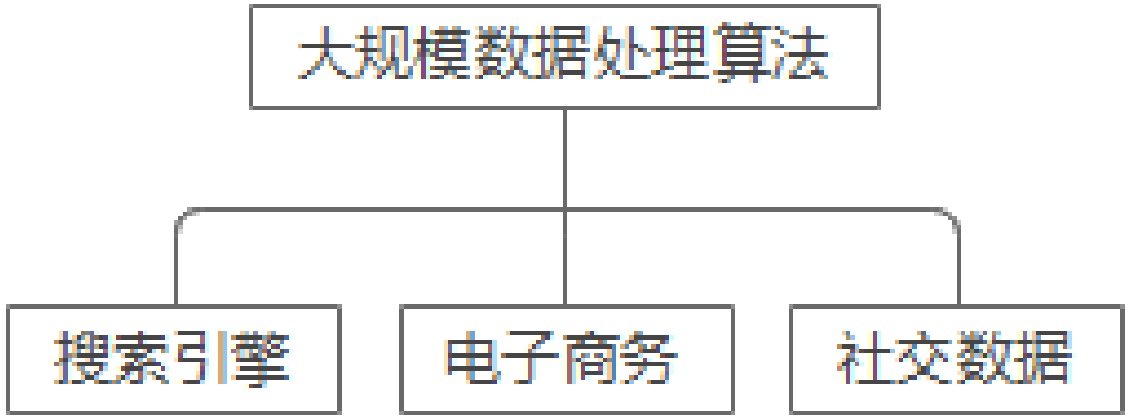
Google利用网络大数据预测流感

华尔街利用微博数据预测股票

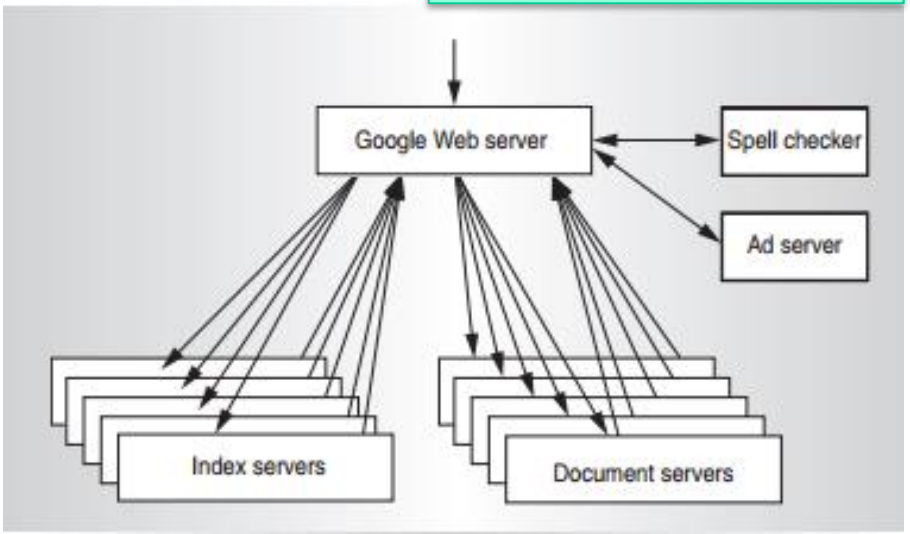
利用大数据预测美国大选

- 健康：谷歌流感趋势（Google Flu Trends，GFT）未卜先知的故事，常被看做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

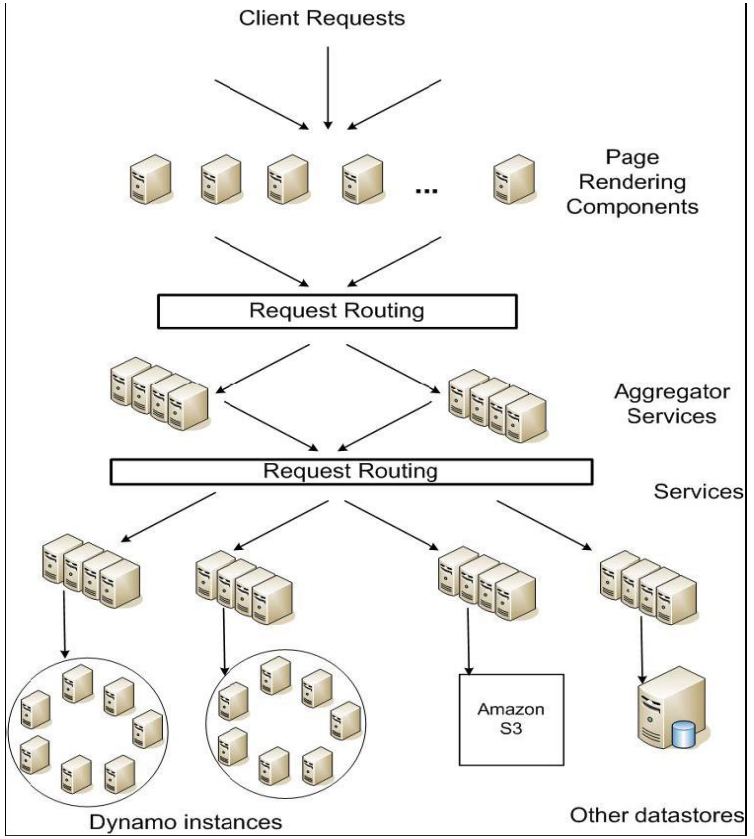
商业上的成功 – FLAG，BAT



谷歌的搜索引擎架构



亚马逊的云服务架构



大数据安全与隐私

理解大数据安全

- 怎么理解大数据安全呢?
  - 大数据安全是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。

隐私的概念与起源

- 隐私权的概念在人类发展中的首次出现
  - 1890年，隐私权概念的提出
- 隐私权第一次写入法律
  - 1902年，纽约，“面粉店与少妇”事件
- 20世纪60-70年代，关于隐私权的法律在欧美国家得到了充分发展
  - 1974年，水门事件导致尼克松总统辞职
  - 1974年，《隐私法》


01.隐私的起源-189


02.隐私权第一次入法-


04.隐私权法律的


03.水门事件-1974

技术进步与隐私保护

隐私权的保护也在发生深刻的变化，其重心不断转移，经历了从住宅到人、再到信息的转变。

- 1876年，贝尔发明了电话；
- 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。
- 1946年，电子计算机被发明；
- 1989年，万维网的出现，开启了一个新的时代

→ 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

→ 开始发生侵犯肖像权的案件。

→ 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年，《中央数据银行》计划浮出水面

→ 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(Total Information Awareness)。
2013年7月，斯诺登事件，美国的“棱镜计划”

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
    - 国家级敌手（State-level adversary）
    - 2013年7月，斯诺登事件，美国的“棱镜计划”
      - 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，像武器一样容易购买
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

数据从封闭环境走向开放环境

- 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……

大数据带给我们的困扰

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。
  - 无论是主流服务提供商，还是用户（无论是普通用户，还是安全专家）

如何保护我们的数据与隐私？

- 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

网络隐私 -- 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，包括facebook上的社交信息，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；

网络隐私 -- 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
  - Facebook，泄漏用户隐私数据事件

欧盟《通用数据保护条例》（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。

GDPR要点概述

- 规定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
- 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。

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。

“被遗忘权”（right to forgotten）

- 2014年5月13日欧盟法院就“被遗忘权”（right to forgotten）一案作出裁定，判决谷歌应根据用户请求删除不充足的，无关紧要的，不相关的数据，以保证数据不出现在搜索结果中。
- 在大数据时代，加强对用户个人权利的尊重才是时势所趋的潮流。

新技术与新挑战：
例如，区块链的不可删除特性与被遗忘权的冲突

CCS 2019 - Five Years of the Right to be Forgotten

我国的《个人信息安全规范》

- 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》（以下简称“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**。

小结

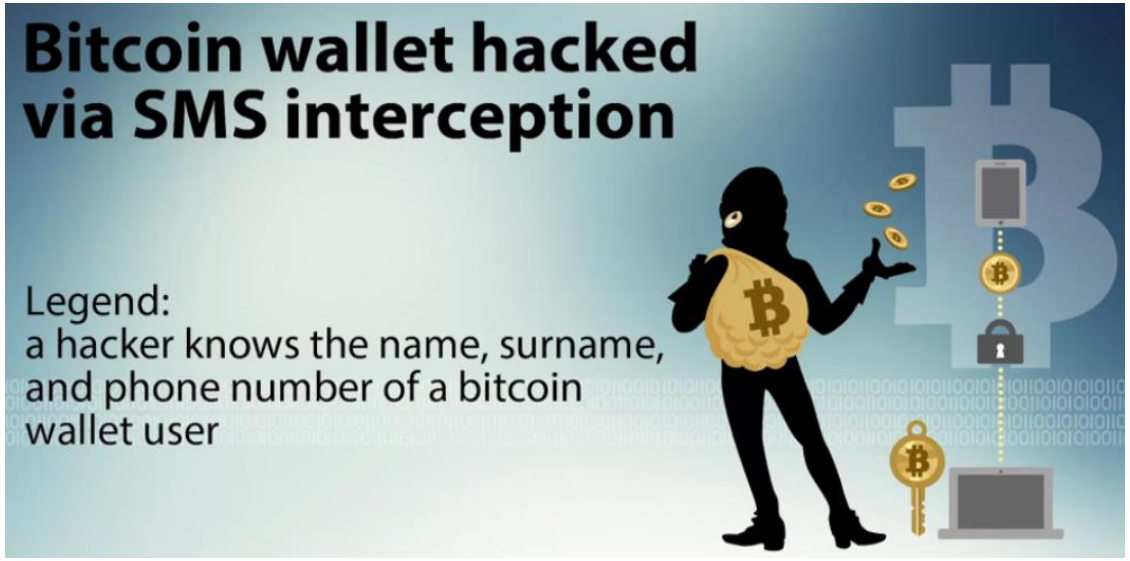
- 大数据的概念
  - 5V
- 大数据的应用：生活，健康，经济，政治，等
- 大数据安全
  - 两个方面：1）如何实现大数据系统的安全？2）基于大数据技术的攻防
- 隐私的概念及其发展
  - 大数据时代的新挑战：遍布互联网的个人数据，使得个人数据安全与隐私保护变得更为复杂
- 相关法规：数据的隐私保护已经成为企业不容忽视的问题。
  - 欧盟《通用数据保护条例》（General Data Protection Regulation，GDPR）
  - 我国的《信息安全技术 个人信息安全规范》

思考

- 个人数据安全与个人隐私的关系
  - 个人隐私的泄漏，可能会危害到个人数据安全。

- Q1）有人说，我的重要数据都需要口令才能访问。
  - 个人隐私信息的泄漏，使得定向攻击更容易
    - 例如，研究表明，口令猜测的定向攻击的成功率比人们预想的高得多；成功的关键是获得受害者的个人隐私信息，例如，生日，年龄，工作信息，家庭情况，等等。

CCS 2016 - Targeted Online Password Guessing: An Underestimated Threat



泄漏账号的姓名和电话号码，会有什么后果？

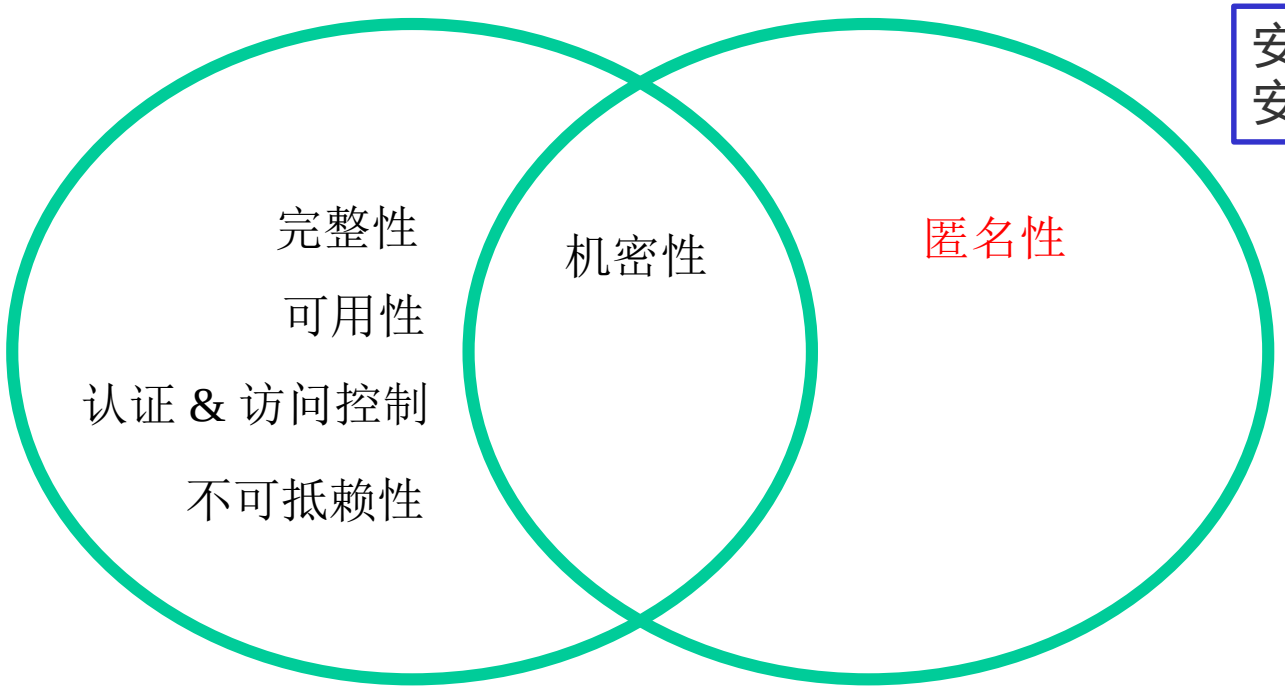
信息安全是聚焦于信息资产的安全工具和安全行为，隐私保护则是关注对个人信息的使用和保护。

Bitcoin wallet hacked via SMS interception

Legend:
a hacker knows the name, surname, and phone number of a bitcoin wallet user

思考

- 安全 vs. 隐私
- Q2） 属性分析
  - {CIA（机密性，完整性，可用性），认证，访问控制，不可抵赖性} ∩ {机密性，匿名性} = {机密性}



安全和隐私的关系：
安全能够保证信息的机密性，隐私则常常需要这种机密性

公共安全 vs. 个人隐私

- 如何在进行安全保障的同时保留民主的重要支柱，例如言论自由、集会及结社自由、还有最关键的隐私权。
- **隐私保护机制被犯罪分子用来隐藏网络行迹**

还有很多开放的问题，需要大家去探索。

阅读材料 -

1) CCS 2019 - Five Years of the Right to be Forgotten
2) NDSS 2019 - Measuring the GDPR’s Impact on Web Privacy
3) USENIX 2019 - Understanding Passwords of Chinese Web Users
4) CCS 2016 - Targeted Online Password Guessing-An Underestimated Threat

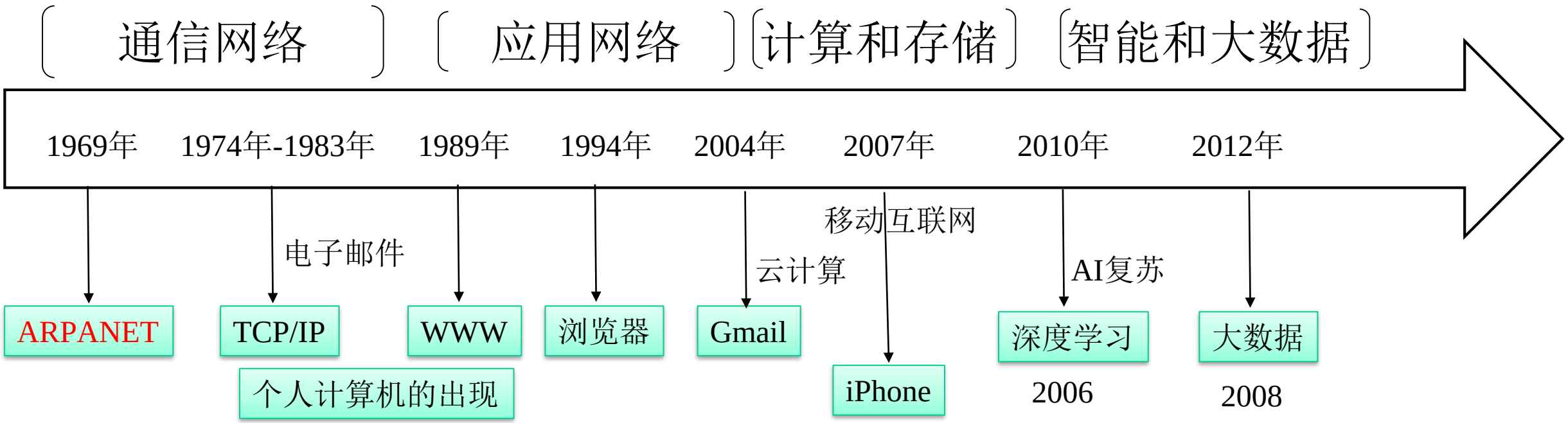
5) Acquisti, Alessandro, Leslie K. John, and George Loewenstein. "What is privacy worth?." The Journal of Legal Studies 42.2 (2013): 249-274.
6) Winegar, Angela G., and Cass R. Sunstein. "How Much Is Data Privacy Worth? A Preliminary Investigation." Journal of Consumer Policy 42, no. 3 (2019): 425-440.

北京郵電大學
Beijing University of Posts and Telecommunications

感谢聆听！

互联网的演进 – 从网络到分布式智能系统

经过WWW以来的近30年的演进，互联网已经不再只是最初（70年代-80年代）设想的一个基于包交换技术的通信网络，而是演进为一个集通信、存储、计算于一体的遍布全球的信息网络。

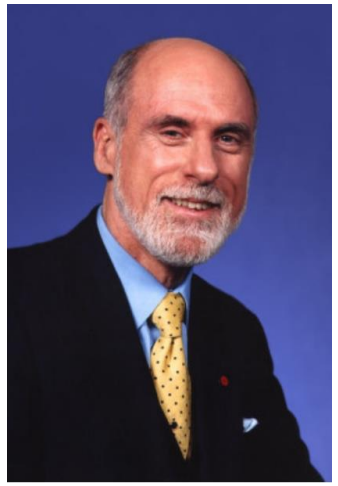


- 互联网是一台分布式大型计算机

核心协议 – TCP/IP

- 1973年，卡恩与瑟夫开发出了TCP/IP协议中最核心的两个协议：TCP协议和IP协议。
- 1974年12月，卡恩与瑟夫正式发表了TCP/IP协议
  - 为了验证TCP/IP协议的可用性， 使一个数据包由一端发出 ，在经过近10万km的旅程后到达服务端。在这次传输中，数据包没有丢失一个字节，这充分说明了TCP/IP协议的成功。
- 1983年元旦，TCP/IP协议正式替代NCP，从此以后TCP/IP成为大部分因特网共同遵守的一种网络规则。
- 1984年，TCP/IP协议得到美国国防部的肯定，成为多数计算机共同遵守的一个标准。

罗伯特·卡恩（Robert Elliot Kahn）



温特·瑟夫（Vint Cerf）

互联网的GUI - WWW

- 1989年，Tim Berners-Lee
- 1994年，浏览器
- 1997年，CDN

蒂姆 伯纳斯 李（Tim Berners-Lee）

智能与大数据

- 深度学习：2006-2010

- **大数据**：2008年，**提出概念**；2012年，奥巴马政府的《大数据研发计划》确立了大数据在世界范围的战略位置。