



北大汇丰

PHBS FINANCIAL FRONTIER DIALOGUE

金融前沿对话

2019 年第 28 期 总第 39 期



PHBS HFRI
北京大学汇丰金融研究院

主办单位：北京大学汇丰金融研究院

院长：海闻

执行院长：巴曙松

秘书长：本力

编辑：陈博闻（执行） 叶静 曹明明 鞠琳琳 方垚豪 朱伟豪

北京大学汇丰金融研究院简介

北京大学汇丰金融研究院（The HSBC Financial Research Institute at Peking University，缩写 HFRI）成立于 2008 年 12 月，研究院接受汇丰银行慈善基金会资助，致力于加强国内外著名高校、金融研究机构、以及知名金融学者之间的交流，构建开放的学术、政策交流平台，旨在提高中国金融理论与政策的研究水平，促进学术繁荣与发展，加强与政府决策部门的联系与合作，为政府决策提供参考意见，为国际金融体制改革和中国金融业的发展做出贡献。

北京大学汇丰金融研究院院长为北京大学校务委员会副主任、北京大学汇丰商学院院长海闻教授，执行院长为中国银行业协会首席经济学家、中国宏观经济学会副会长巴曙松教授。

从“比特”的视角理解数据和信息的含义

【对话主持】

巴曙松（北京大学汇丰金融研究院执行院长、中国银行业协会首席经济学家、中国宏观经济学会副会长）

【特邀嘉宾】

胡本立（协同数据联合创始人，国际数据管理协会中国区创始人、主席）

一、“比特”的由来

无论从哪个领域或视角，在数字时代，个人及单位都需要同“一串比特”打交道，不论这个比特是加密的或没加密的，基础层的或应用层的。数字化时代是从比特开始的，我希望通过对比特的历史和一些相关基本概念的介绍，有助于更好、更有效地讨论目前数字经济中出现的许多概念问题，包括大数据、机器学习、人工智能、认知科学到数字经济、电子/虚拟/加密/数字货币以及区块链等。本次讨论简述从比特延伸出来的两个与它紧密相关的基本概念——数据和信息，其它的相关概念都是建立在这两个基本概念之上。

比特的词源是 bit, bit 是从 bite（“一小块/咬一口”）简化而来。bit 作为信息的二进制位的缩写，最早出现在 1947 年 John W.

Tukey 写给贝尔实验室同事的备忘录中。随后香农(Claude Shannon)于 1948 年在他最著名的“通信的数学理论”中，将其定义为计算机中数据的最小单位，即要么 0 要么 1 的最小单位，后来也被称“香农比特”。

二、“信息”是什么——比特与信息论的简史

中文版的《信息简史》这本书对“信息”一词的提出、使用和意义等问题有过详细介绍。但是，在本次讨论中，我想让大家知道的是，“信息”是绝对抽象的一个词，因而有多种不同定义。

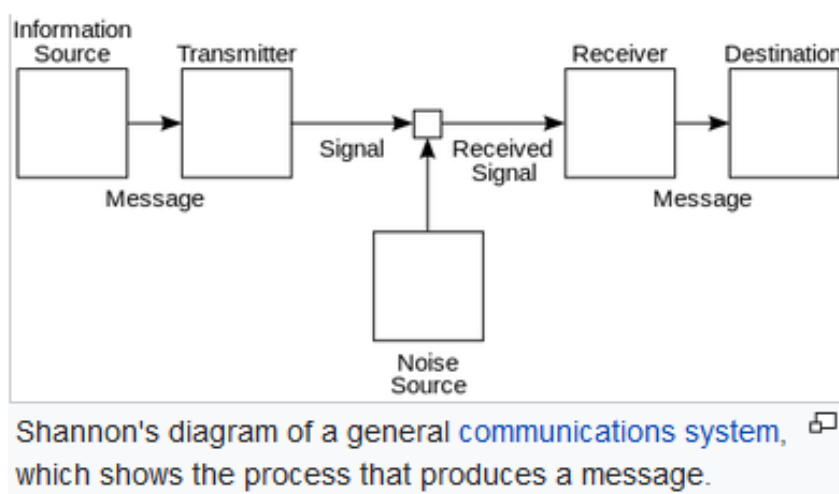


图 1

谈到“信息”在科学和工程上的定义，大家都会讨论香农的那篇文章。图 1 是香农对通信系统的一般描述，包括一串比特的产生、传递以及恢复的全过程（注意它涉及的范围和没有包括的内容）。

当时，半导体的发明让物理层次上实现数字化成为可能。香农主要的兴趣是通信，如何保证这一串比特从一个地方能不失真地传到另

一地方，这一过程涉及到编码、统计、加密或解密一系列过程。在上述过程中，需要给这个不完全是物理的、又与物理有关的、并且有不少数学工具（如布尔代数、概率等）可用来描述的东西取名（与牛顿在他的力学中为“力”取名的情况类似，牛顿最后用了“force”这个词，而这个词当时是用来指军队，今天也还是）。经过与周围的各学科的同事讨论，香农最后决定把“这个东西”称为“信息”，同时信息的最小单位是“比特”。

然而，当时使他和现在许多人困惑的是，信息论中的“信息”与通常大家在用的“信息”不是一回事，有时甚至正好相反。而且他进一步发现他的信息论中的数学公式与物理中描述无序或混乱概念的“熵”的公式一样，这一现象在数学上称为“同构”。于是，信息又与物理联系上了，而比特又被称为“香农熵”。

香农是个很低调、不希望夸大并且很严谨的科学家，他清醒地不把这一串比特的含义纳入他的信息论里进行讨论和处理。人类一直有兴趣去发现规律，所谓规律即是它的可预测性。一般认为信息越多越可预测，但按香农的信息论来讲，越可预测的信息它含有的信息量应越少，因为你已不需要更多的信息来预测了。这有点有悖常理，但科学或工程上所指的信息（也就是香农发明的信息），它与我们一般认为在计算机和手机储存上确定的1和0不是一个概念或事物。

另外，信息世界的比特与物理世界的原子也常被一起讨论，其中一个原因可能是希望理解两个世界究竟有无联系；如果有，两者是如

何联系的。另一个原因可能是比特和原子是这两个世界中的最小度量和元素，如果我们能理解两个世界中最小、最基本的元素之间的关系，最终就可以理解和建立这两个世界的联系。MIT 的尼葛洛庞帝早在 1995 年在他的“Being Digital” (Bits and Atoms) 一书中，从管理的角度，讨论和列举了许多由互联网带来的“信息化”、“数字化”的例子。我觉得目前的数字化、比特币、数字货币和数字化转型只是那些概念的继续或延伸。

总之，现在大量关于“数字化”的概念和讨论，与比特和信息论中的数字化完全是不同层次的概念和问题。信息论中的数字化是将模拟信号数字化（即一个比特），从而在噪音很大或者无限的环境下，保证发出方和接收方的比特是一样的，避免了模拟信号传输的失真问题。所以当时的数字化是革命性的，提高了信息传递的质量（需要注意的是，香农只是把中间传递的过程数字化了，他有意回避了最后的接受者，但从整个数据的或者信息的生命周期来讲，它只不过是中间的一部分，更难以处理的是后面涉及到人的部分。这个问题人工智能希望能够解决，可能也取得了一定的成绩，但是距离概念化的过程还是挺大）。但是与当时一样，我们始终面临的是如何理解和处理这一串比特究竟是什么的问题。全面理解它们的区别和联系有助于我们在概念层面理解目前的数字经济和虚拟经济如何为实体经济服务，有助于更好地讨论数字货币和数字化改革等概念。

三、数据是什么——比特与数据的理解和讨论

数据应包括两个方面，一是数据是如何产生的和被使用（处理）的全过程，二是数据过程是一个从不终止的循环动态过程，也是人和所谓的机器的学习过程。

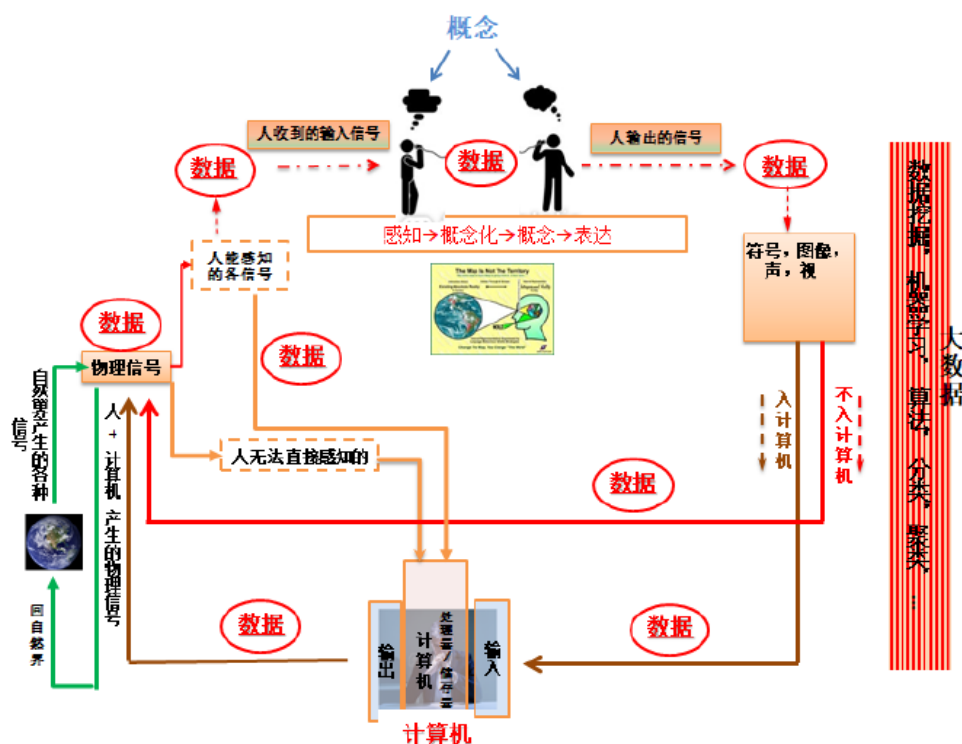


图 2

图 2 描述了数据的概念，其中出现数据的地方都已经标示在图中。图中最关键的部分就是两个人拿了两个电话在打电话，也就是说数字化绕不过中间“人”的存在。虽然图中有很多环状结构，但是中间总是有一个过程是圆的，包括机器的设计、计算机的设计、人工智能的算法都避免不了人的因素。所以很多人做了很多尝试，希望机器能够脱离人，但是我的基本观点还是自动化，也就是机器的背后还是需要依靠人来设计。

下面我从数据的角度讲“三个世界”。如图 3 所示，有一个男孩

和一个女孩，左边有两条狗，一条是狼狗，一条是对人友好的狗。在整个过程中，这两个小孩学会怎么用词汇表达这两条不同的狗。实际上如果狗有被接触或者有感觉看见它时，人对狗的概念已经形成了，但狗应该叫“狗”还是叫“犬”这个概念还没有形成。人在感知到“狗”之后，信号传递到人脑之中，脑根据狗的基本特征达成共识，从而从感官中的“狗”形成了狗的概念。最后，右上角有一个老人开始教他，男孩和女孩就明白了这个东西就是狗。

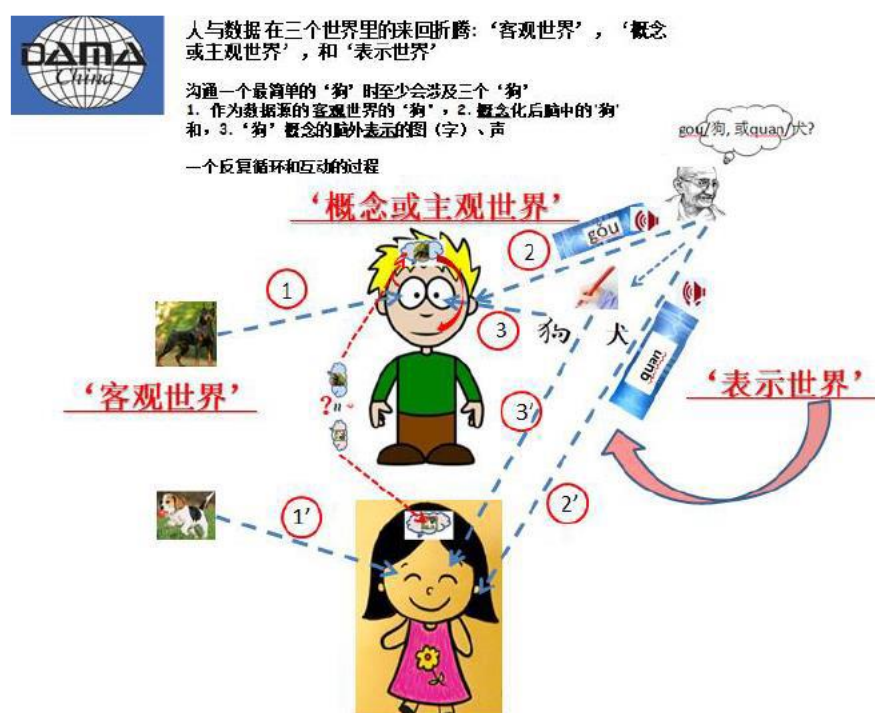


图 3

所以基于这三个事件，我们可以得到“三个世界”：左边的是客观存在的世界；中间是以概念为主的主观世界（因为严格来讲，主观事件下面是大脑的神经元，也是物质结构），概念只有在人脑中才有；右边是大量数据构成的表示世界。这三个世界应该如何打通？上图实

际上反映了这个循环和三个世界的关联，但关联的程度不一定完全一样。例如，男孩和女孩可能对狗的基本特征（如两个眼睛、四个脚）基本上不会有不同的意见，但是如果男孩是被狼狗咬过的，女孩则没有被狗咬过，那么他们对狗的感觉和理解是不一样的，当然这种不同不一定很明显地体现出来。

对于客观世界，大家都在努力发现客观世界的自然科学规律，我们现在碰到的大量问题是在表示世界里面。有人认为表示世界是虚拟的，但实际上它不是真的虚拟，它只是和左边的客观物理世界不一样。例如，声音是一种数据，但数据也是物理的，只不过不是完全像左边的客观世界、物理世界的东西。因而我们在大量讨论或辩论的就是在表示世界里面的事情，我的一个观点是包括传统数据、概念模型、逻辑模型、物理模型等，只要是在脑外的都是物理的，尤其从现在人工智能的观点来讲也更是如此。人工智能都是在处理一些信号，这些信号对人来讲都是信号。



图 4

以从事数据管理行业的人的视角来看，他们把所有的东西都认为是数据，而且所有与之相关的都是数据管理。我在这里通过一张简单的图（如图 4 所示）来展示我们的工作。左边是人工智能的机器，主要是统计分类；右边是传统的数据管理世界，包括数据库、数据仓库、数据挖掘、数据分析等，即在人工智能出现以前的整个一套 IT 系统。但是左边和右边并不能自动地连接在一起，而是需要依靠人，这个不是一个简单的过程。而且左边的分类也没有到概念层，也就是达到能够理解的、能够跟右边的数据产生、储存和使用相连接的层次。所以这里面实际上有很大的机会。

四、如何更好讨论一系列冠于或实质上的“数字”问题和挑战

第一，不同层次的问题需要在不同层次中去讨论和解决。我们现在的讨论基本上是不分层次的，从香农的信息论到下面的比特，技术上的中间层到具体业务，还有经济金融的概念，这里面有好几个层次。我觉得不同层次的问题需要在不同层次中去讨论和解决。当然跨界很重要，许多创造发明都是跨界产生的，但是跨界也有风险。

第二，不同层次的问题需要能被打通。创新如果是跨层次的，那么就需要打通这些层次。打通是数据管理的老问题了，但整合孤岛打通绝对不是一个简单的过程。一种普遍的看法就是技术能够解决这个问题，但是就算有再好的技术和算力，真的打通还是需要靠人的介入。

第三，直接与间接的问题。科技给经济和社会带来的影响可能有

两方面，一个是直接的，一个是间接的。科技一方面直接改变人的行为，另一方面通过人的行为变化产生的数据间接影响经济的模型。它们之间是有连接的，这种连接主要还在人。

第四，几种不同的概念和数据的定义以及分类问题。我觉得分类现在是个很大的问题，主要是概念的定义和分类。首先，分类是物理的还是人为的？物理的意思是这个数据的产生是由自然规律来决定的，人的因素很少或者没有。自然科学基本上希望把人类的因素都去掉，但事实上又少不了人。不管是极为抽象的广义相对论，还是其他理论，最后都是人研究出来的。上面讲的分类需要有一个验证的过程，现在许多情况很难验证，也经常变化，从而导致人也在改变这个情况。人的一些看法、理解和定义都在变化，但是物理方面的东西却比较容易达成共识，例如温度、气候。其次是不同领域和行业的分类，每个领域是从自身的某一个角度来看这个世界，如果你要他换另外一个角度来看世界，这个领域就不是原来的那个领域了，这里面有交叉问题。

五、结束语

最后，我想借用著名的计算机科学家和统计学家迈克尔·乔丹在《哈佛数据科学评论》上发表的文章“人工智能，革命远未发生”中的两段话：

“这门新的工程学科将建立在上个世纪赋予了实质的思想之上，如信息、算法、数据、不确定性、计算、推理和优化。此外，由于新学科的大部分重点将放在来自人类和关于人类的数据上，因此，它的

发展将需要社会科学和人文学科的观点。”科技当然重要，但是目前的许多问题的瓶颈是社会和人文的，不是或不都是科技能解决的。

“无论我们是否很快就能理解‘智能’，我们都面临着一项重大挑战，那就是如何将计算机和人类结合起来，从而增强人类的生活。虽然各种构建块已经就位，但是将它们组合在一起的原则还没有到位，……我们早期的许多社会规模的推理和决策系统已经暴露出严重问题的概念缺陷。”目前需要严谨的科学意识和态度，努力和尽力解决那些“已经暴露出严重问题的概念缺陷”。

问答环节：

Q：“数字货币”和“货币数字化”这两个概念您怎么看？

A：我觉得主要还是看货币怎么定义，如果经济学家不能弄清楚货币的定义，那么数字这个东西就可以挂上去了。我理解的货币数字化就是电子支付、电子货币。整个关于货币的讨论从电子的，到后来虚拟的，再到加密的，现在又回到数字，最近有一个报告则又回到了电子。所以一些基本的概念如果不正确或者不清楚，我们可以整天就在这些名字里面兜圈子。现在比较热门大的数字化改革基本上就是业务模式的改变，这种改变是从人的行为数据里产生的，而不是直接从技术上产生的。我觉得经济学家需要重新看这个世界，需要重新突出原来的一些模型，但是又不能完全重新推翻。管理层也要重新理解数据究竟是什么，我很担心他们一知半解，只有他们对数据有更深的理

解才能立好法律规范。关于数据怎么确权、什么算隐私、什么算自己的数据等等都还没有搞清楚。现在整个数据的环境都变了，法律法规肯定是滞后的，但滞后的法律法规又要管理现在的环境，所以这里面又有很多地方值得进一步讨论的问题。我觉得国内监管方面的问题是监管的界限不清晰。虽然银监会和保监会合并了，但是机构上的合并跟数据上的融合之间还是有一定距离。

本文根据北京大学汇丰金融研究院执行院长巴曙松教授发起并主持的“全球市场与中国连线”第三百零三期（2019年9月16日）内容整理而成，特邀嘉宾为协同数据联合创始人，国际数据管理协会中国区创始人、主席胡本立先生。

胡本立先生，协同数据联合创始人，国际数据管理协会中国区创始人、主席。曾任世界银行首席信息技术官，负责世行信息技术总体架构的设计、开发、和运行，支持世行的战略和使命；继后任香港证券及期货事务监察委员会首席信息官、顾问；中国证监会战略及规划委员会委员；中国社保基金理事会高级技术顾问；国际数据管理协会顾问，副主席，负责国际数据标准；惠普中国咨询总顾问等。近年来专注数据管理，尤其是理解人、数据和计算机之间的复杂认知过程，和需要的数据思维、技术和工具。

【免责声明】

“全球市场与中国连线”为中国与全球市场间内部专业高端金融交流平台。本期报告由巴曙松教授和熊婉芳博士共同整理，未经嘉宾本人审阅，文中观点仅代表嘉宾个人观点，不代表任何机构的意见，也不构成投资建议。

本文版权为“全球市场与中国连线”会议秘书处所有，未经事先书面许可，任何机构和个人不得以任何形式翻版、复印、发表或引用本文的任何部分。



PHBS HFRI
北京大学汇丰金融研究院

