攻击大数据

从D.O.S.三平面看

潘柱廷　中国计算机学会　常务理事、副秘书长
CCF大数据专家委　委员
启明星辰　首席战略官

2012年12月1日

微解读

【5】大数据安全和隐私问题

- 安全和隐私，永远的问题
- 随着数据的增多，组织面临的重大风险跨越了一个复杂的威胁面，要遵守更多合规规定，传统的数据保护方法常常无法满足

- 挑战
  - 大数据规模的密码学
  - 分布式编程框架中的安全计算
  - 非关系型数据存储
  - 安全的数据存储和事务日志
  - 终端输入的确认/过滤
  - 实时安全/合规监测
  - 可扩展的、可组合的、脱敏(无隐私)的数据挖掘和分析
  - 强制的访问控制和安全通信
  - 粒度访问控制
  - 数据来源和数据通道

微解读

【6】大数据安全（29票）

- 大数据的安全令人担忧
- 大数据的保护越来越重要---大数据的不断增加，对数据存储的物理安全性要求会越来越高，从而对数据的多副本与容灾机制提出更高的要求。
- 网络和数字化生活使得犯罪分子更容易获得关于人的信息，也有了更多不易被追踪和防范的犯罪手段，可能会出现更高明的骗局。大数据已经把你出卖。

微解读

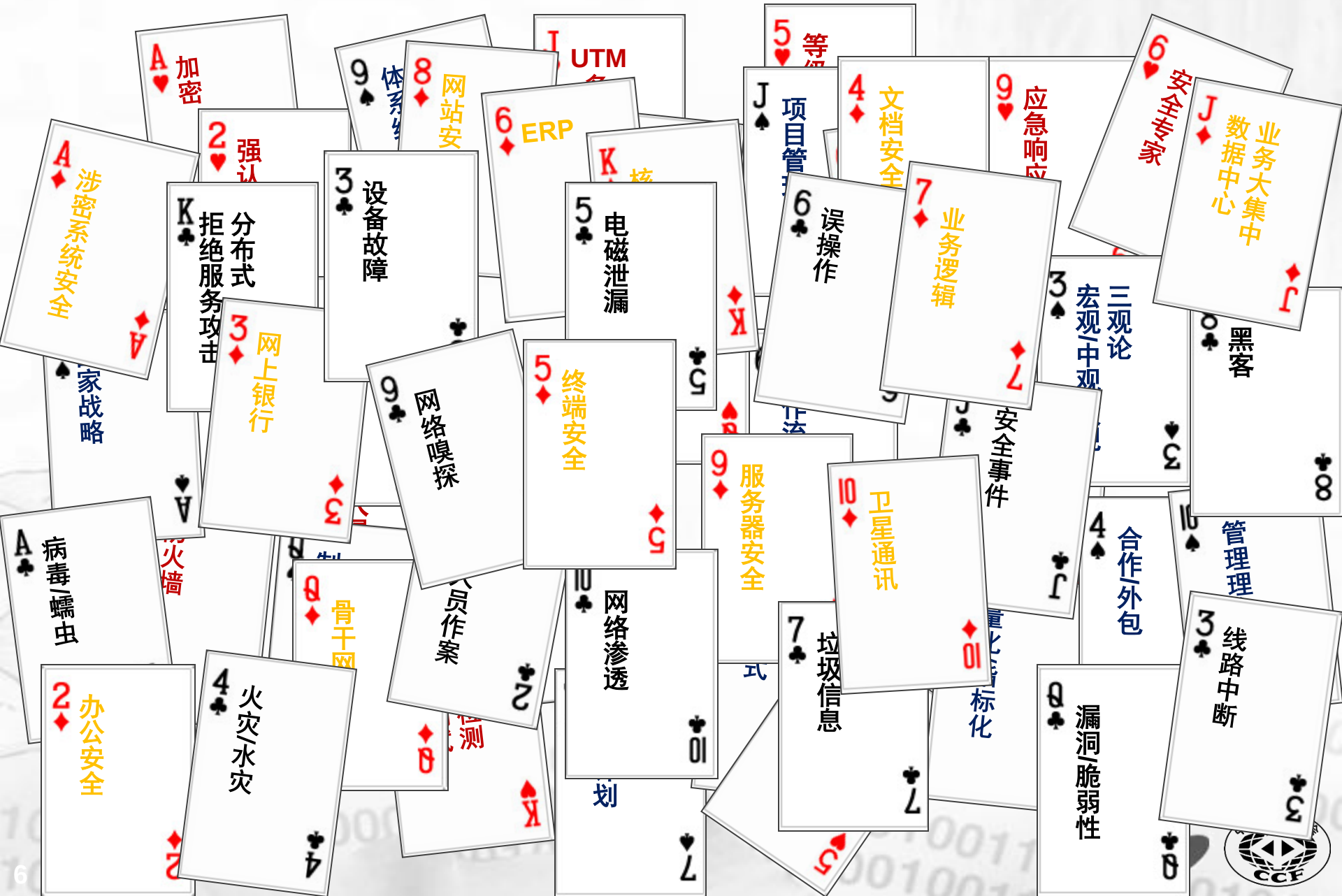
【2】大数据隐私问题（44/70票）

- 大数据对于隐私是一个重大挑战
- 2013年隐私相关的标准和条例颁布
- 现有的隐私保护法规和技术手段难于适应大数据环境
- 有偿隐私服务可能出现
- 个人隐私越来越难以保护
- “面罩”流行

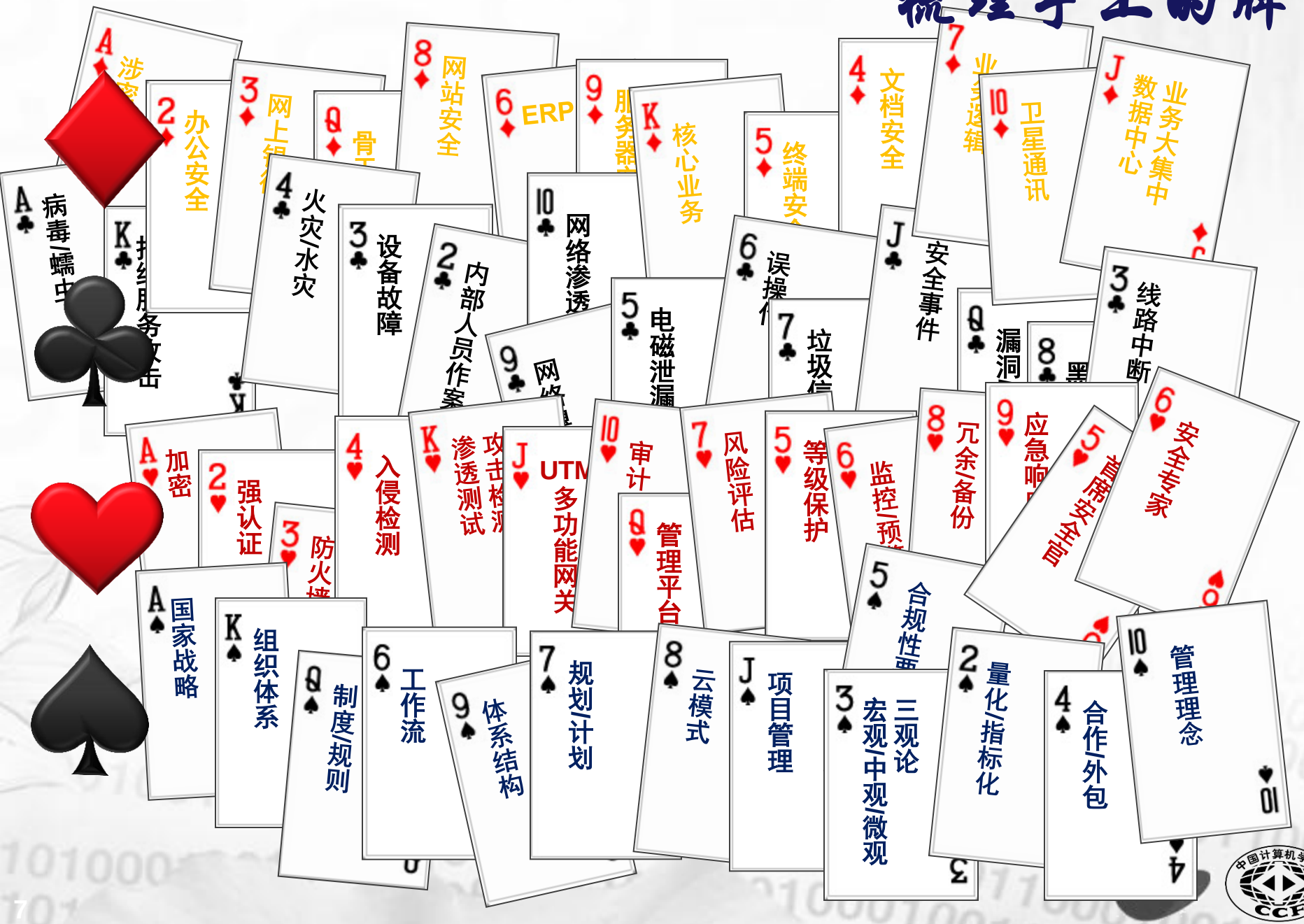
- 用大数据解决安全问题
- 大数据自身的安全问题

大数据安全

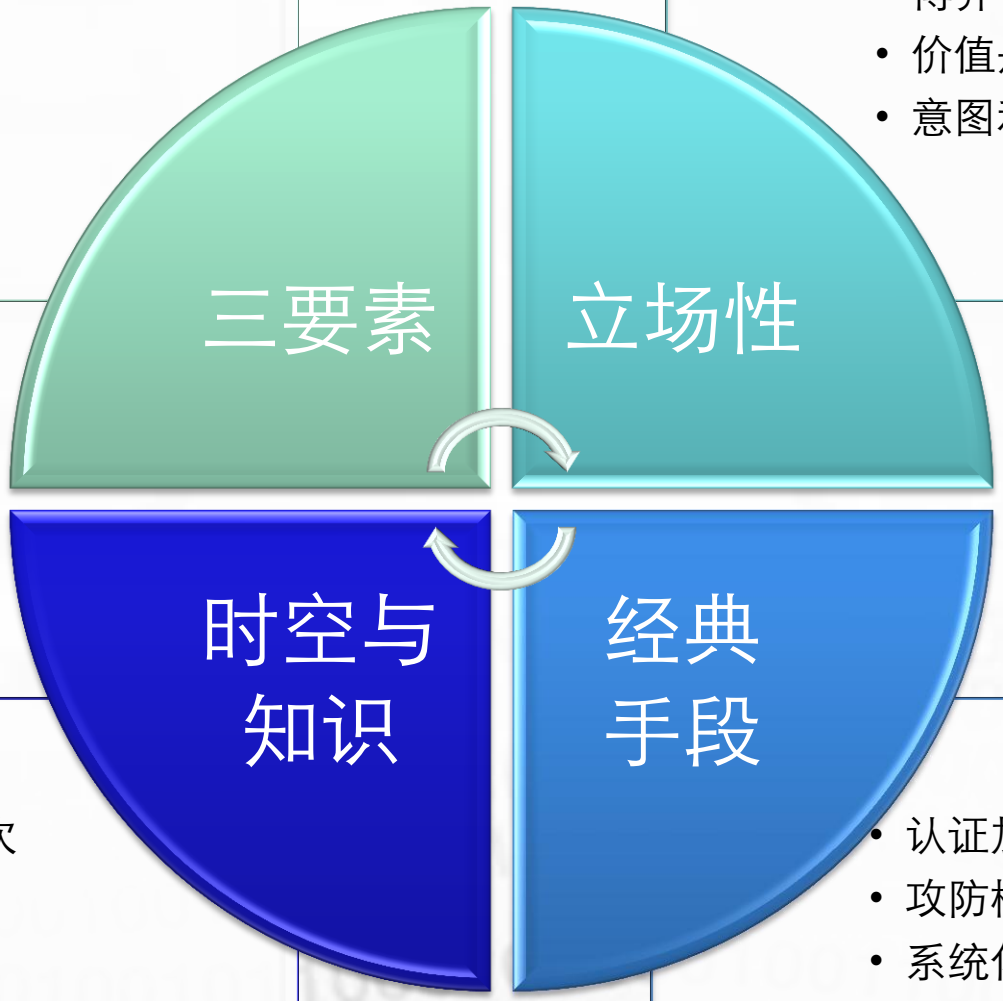
谈安全涉及到的方方面面



梳理手上的牌


网站安全
ERP
核心业务
终端安全
文档安全
卫星通讯
业务大集中
数据中心
办公安全
网络渗透
安全事件
线路中断
火灾/水灾
设备故障
内部人员作案
误操作
垃圾信
漏洞
病毒/蠕虫
电磁泄漏
加密
强认证
入侵检测
渗透测试
UTM
多功能网关
审计
管理平台
风险评估
等级保护
监控/预警
冗余/备份
应急响应
安全专家
国家战略
组织体系
制度/规则
工作流
体系结构
规划/计划
云模式
项目管理
合规性
三观论
宏观/中观/微观
量化/指标化
合作/外包
管理理念

安全思维



- 资产
- 威胁
- 安全措施

- 博弈、对抗、合作
- 价值是有归属的
- 意图和随机扰动

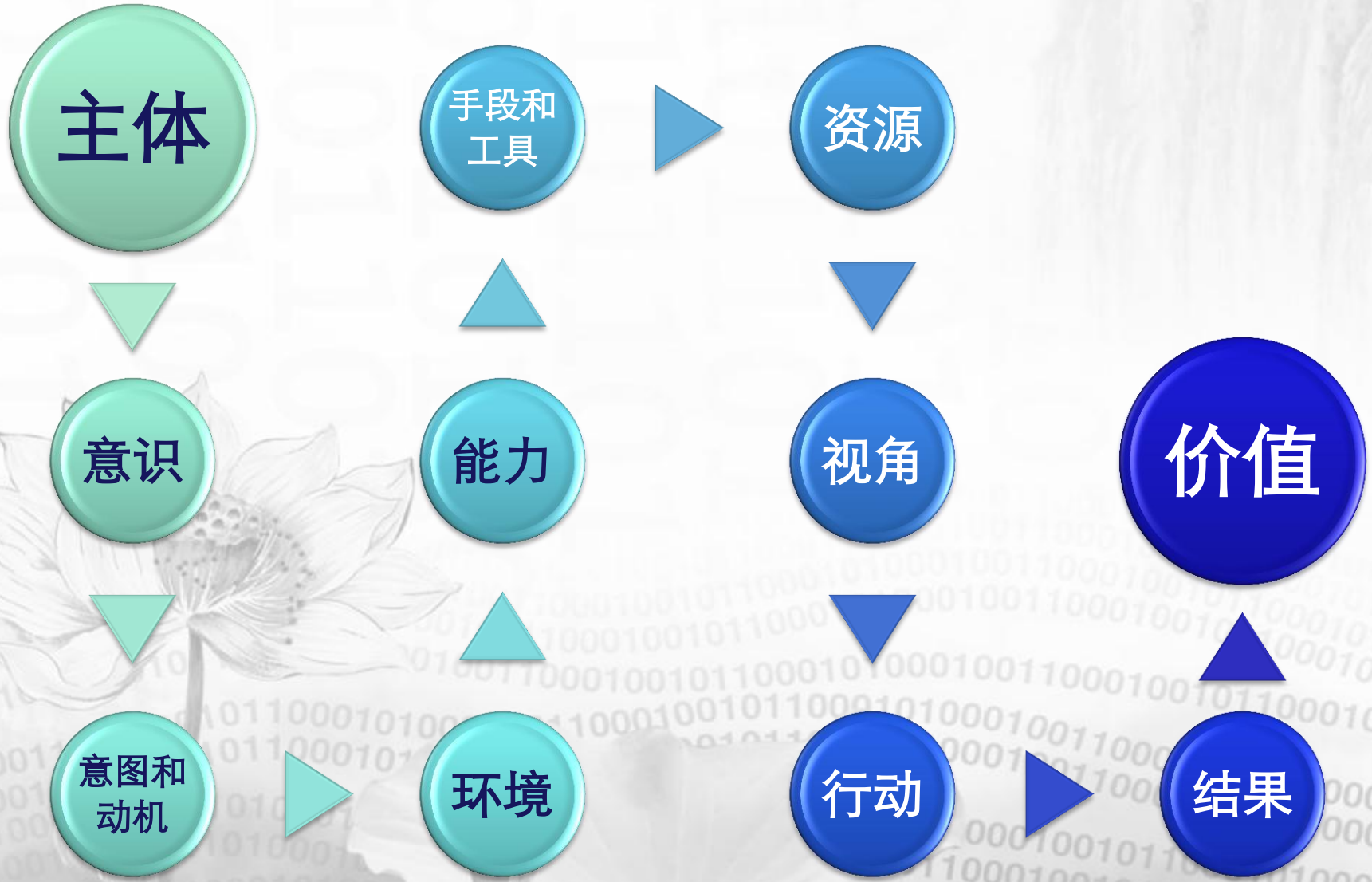
- 分布和层次
- 生命周期
- 流和用例
- 知识维簇

- 认证加密类
- 攻防检测类
- 系统化风险管理类

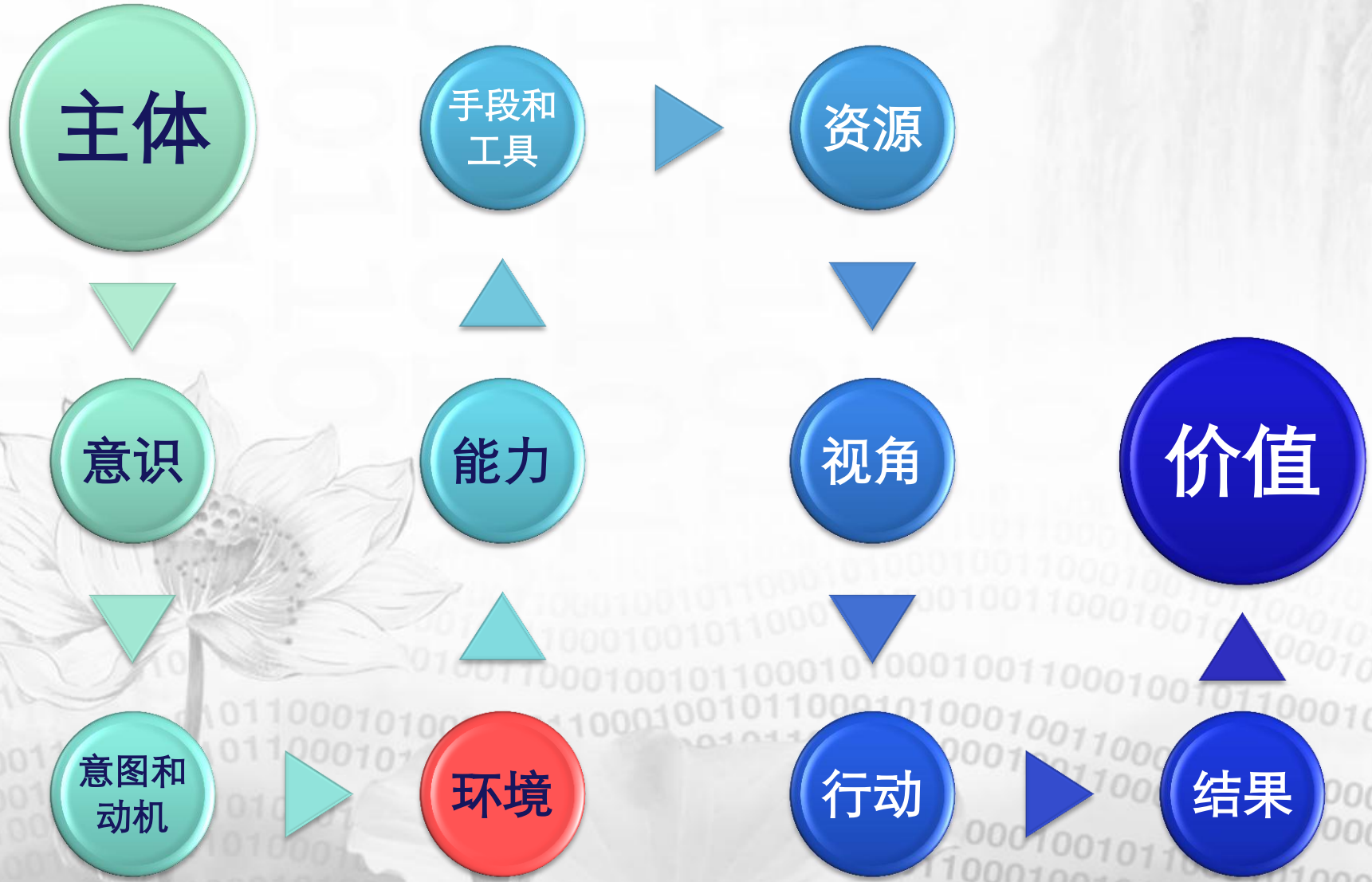
威胁场景 Threat Case

- **背景Background：前提、假设、条件等**
- **来源Agent：包括攻击者、误用者、故障源、自然（灾害）等**
- **对象Object：攻击目标和破坏对象，也就是要被保护的对象**
- **环境Environment：攻防所处的主要计算环境、网络环境、物理环境等**
- **内因——脆弱性Vulnerability：自身保护不当的地方**
- **模式和方法Mode**
- **过程Process**
  - **途径Route**：指威胁必须通过才能实现的一些部分。比如，要通过网络、要在物理上接近设备、要欺骗人等等。
  - **时序Sequence**：威胁要实现所必经的步骤和顺序。与威胁的途径是一个从空间上，一个从时间上表达。也可以将这两个因素结合起来表达威胁的过程。
- **结果——事件Event/Incident：威胁具体实现之后所造成的结果**
  - 威胁的可能性：威胁产生结果变成事件的概率。
  - 威胁的影响范围：威胁产生结果后的影响大小。以及影响进一步扩散的特性。

威胁场景的要素



大数据影响威胁场景



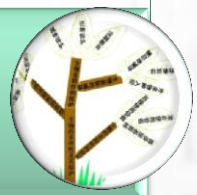
微解读

【3】大数据特性与数据态

53

- 多来源多模态数据：图像、视频、音频、数据流、文本、网页…
- 关联关系异质、结构模式复杂
- 互为因果，动态变化

关系维簇



44

- 三元空间大数据的产生、状态感知与采集
- 柔性粒度数据传输、移动、存储与计算
- 数据空间范围和数据密度的非均衡态

空间维簇

63

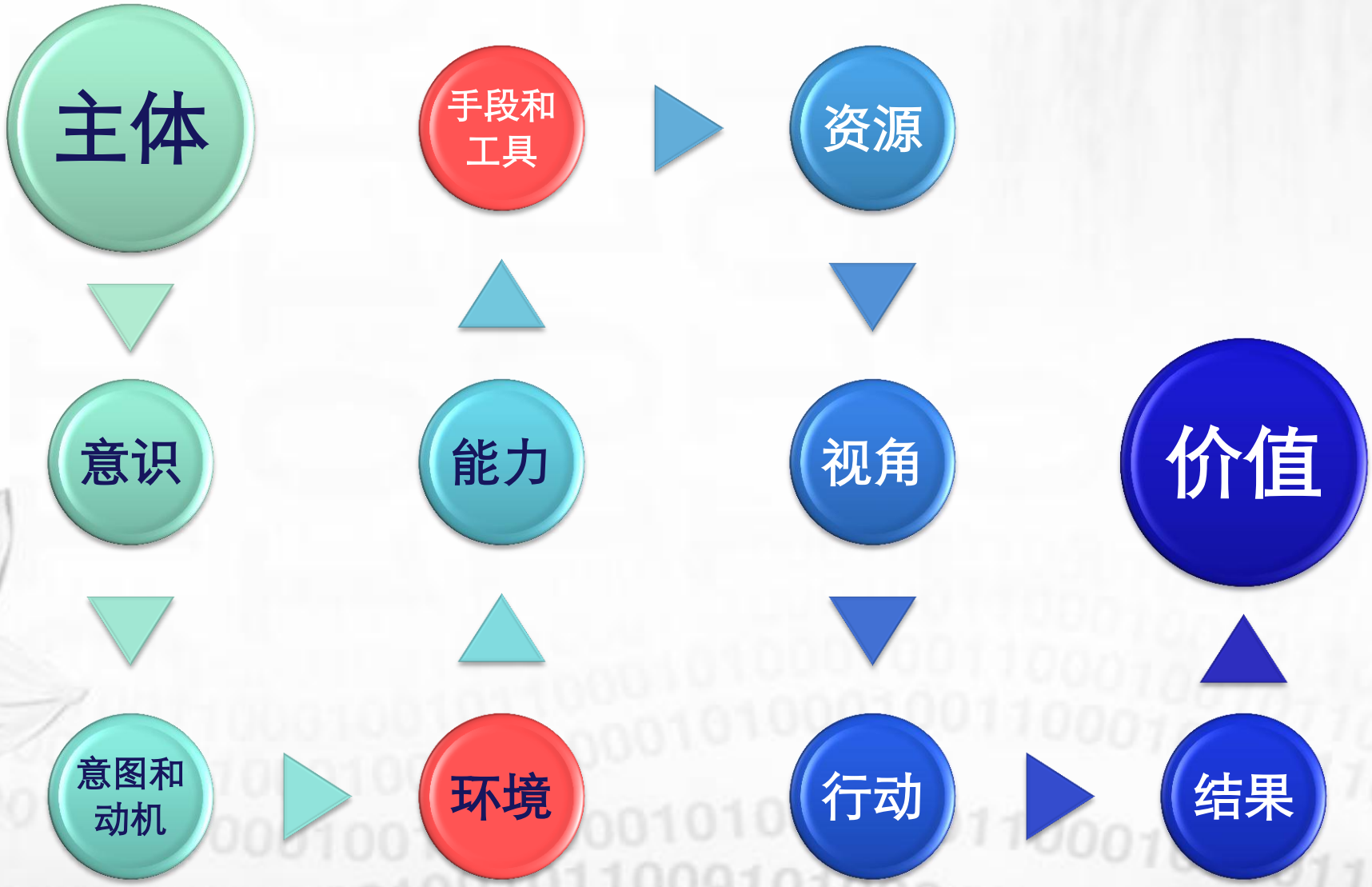
- 数据的生命周期
- 数据的时间维状态与特征
- 流化分析、增量学习、在线推荐
- 离线与在线时效性要求

时间维簇

大数据核心问题

——CCF大数据专家委

大数据思维改变了手段和工具



【2】数据计算的基本模式与范式

- 数据密集型计算的基本范式?
- 数据计算的效率评估与数据计算复杂性理论?
- 从中心化的/top-down模式转为去中心化的/自组织的计算模式?
- 基于数据的智能：会有越来越多靠“数据的体量+简单的逻辑”的方法去解决复杂问题

第四范式的科学方式

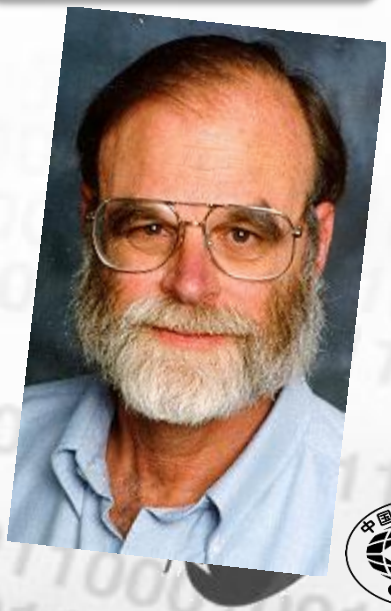
几千年来 科学实验

数百年来 模型归纳

数十年来 模拟仿真

今天 数据密集型

PARADIGM



在第一范式到第四范式

几十年来	数百年来	数十年来	今天
科学实验	模型归纳	模拟仿真	数据密集型

- 渗透测试
- 事件分析
- 漏洞挖掘
- 地址随机
- …

- 检测引擎
- 病毒特征
- 漏洞特征
- 攻击特征
- 关联规则
- …

安全内在的知识高依赖

颈

腰

足

高端

- APT检测
- 宏观态势感知
- 多点/长时/多类型

单一特征

- 特征库检测
- 单点/短时/单项

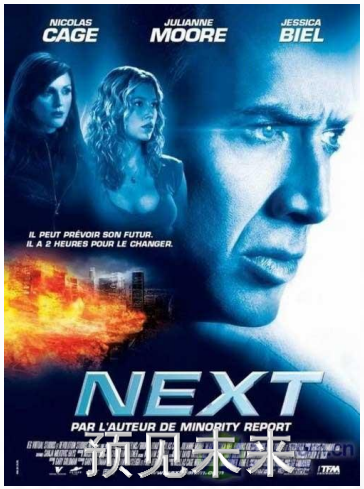
特征使用的计算复杂度
<< 特征获取的计算复杂度
特征可积累、可共享、可升级
特征库是一种知识库，需要知件理论
特征可生产

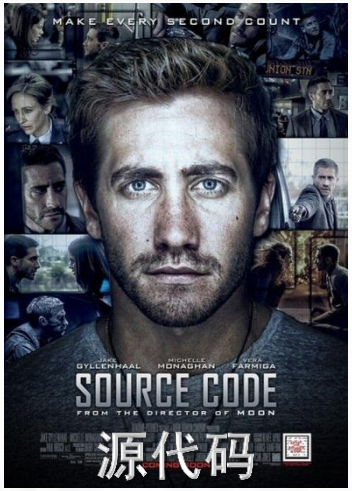
安全，从第一范式到第四范式

几十年来	数百年来	数十年来	今天
科学实验	模型归纳	模拟仿真	数据密集型

- 渗透测试
- 事件分析
- 漏洞挖掘
- 地址随机
- …

- 检测引擎
- 病毒特征
- 漏洞特征
- 攻击特征
- 关联规则
- …





- 模拟攻击
- 模拟被攻击
- …

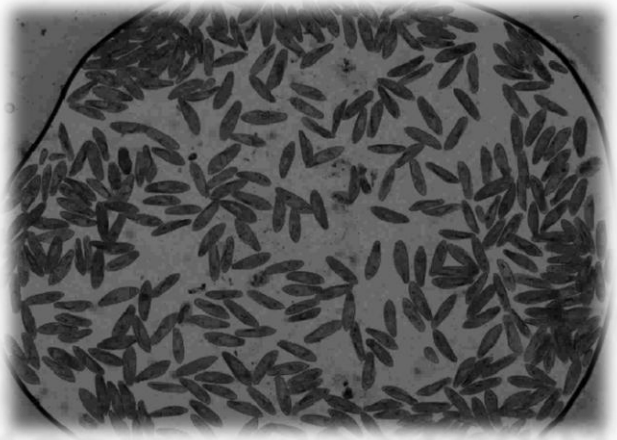
- 沙箱
- 蜜罐
- 标识检测
- …

- 数据密度
- 基于记忆
- 数据浓缩
- …

群目标研究

人群

鸟群

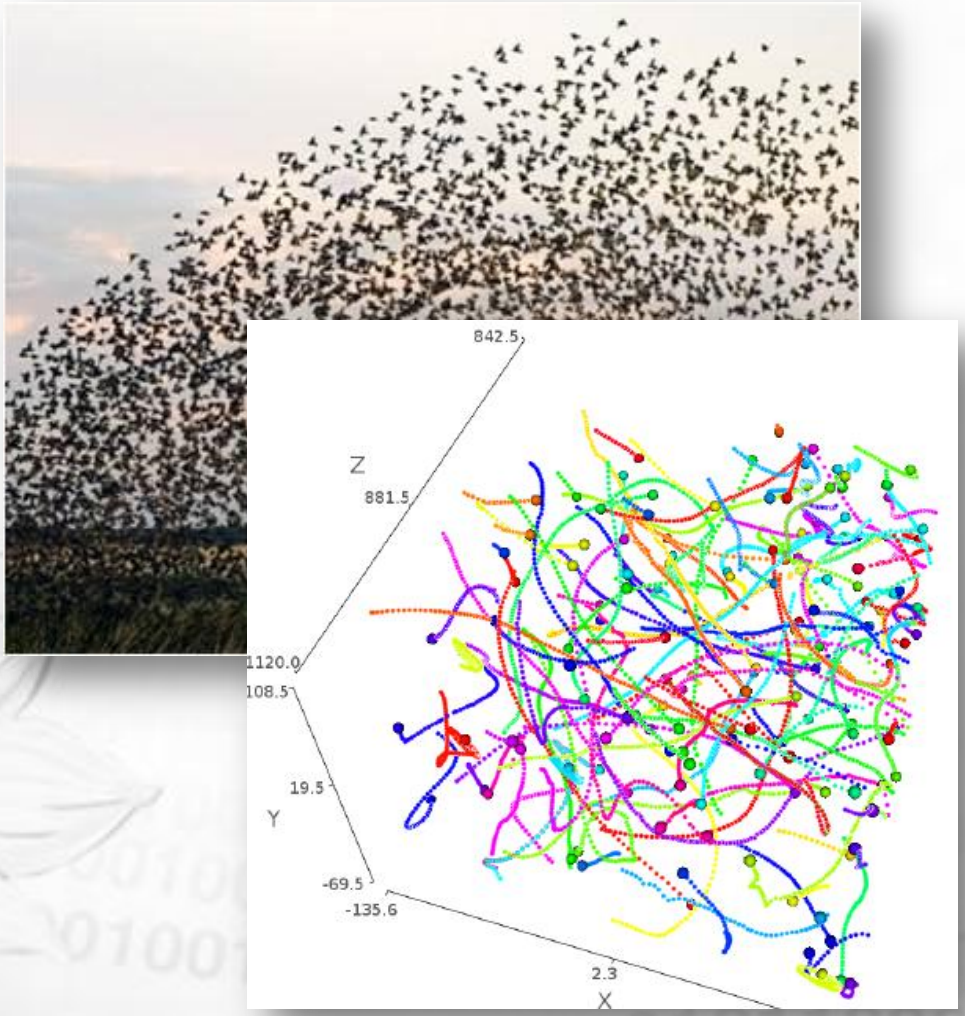


细胞群

鱼群

——摘自陈雁秋香山大数据论坛的PPT

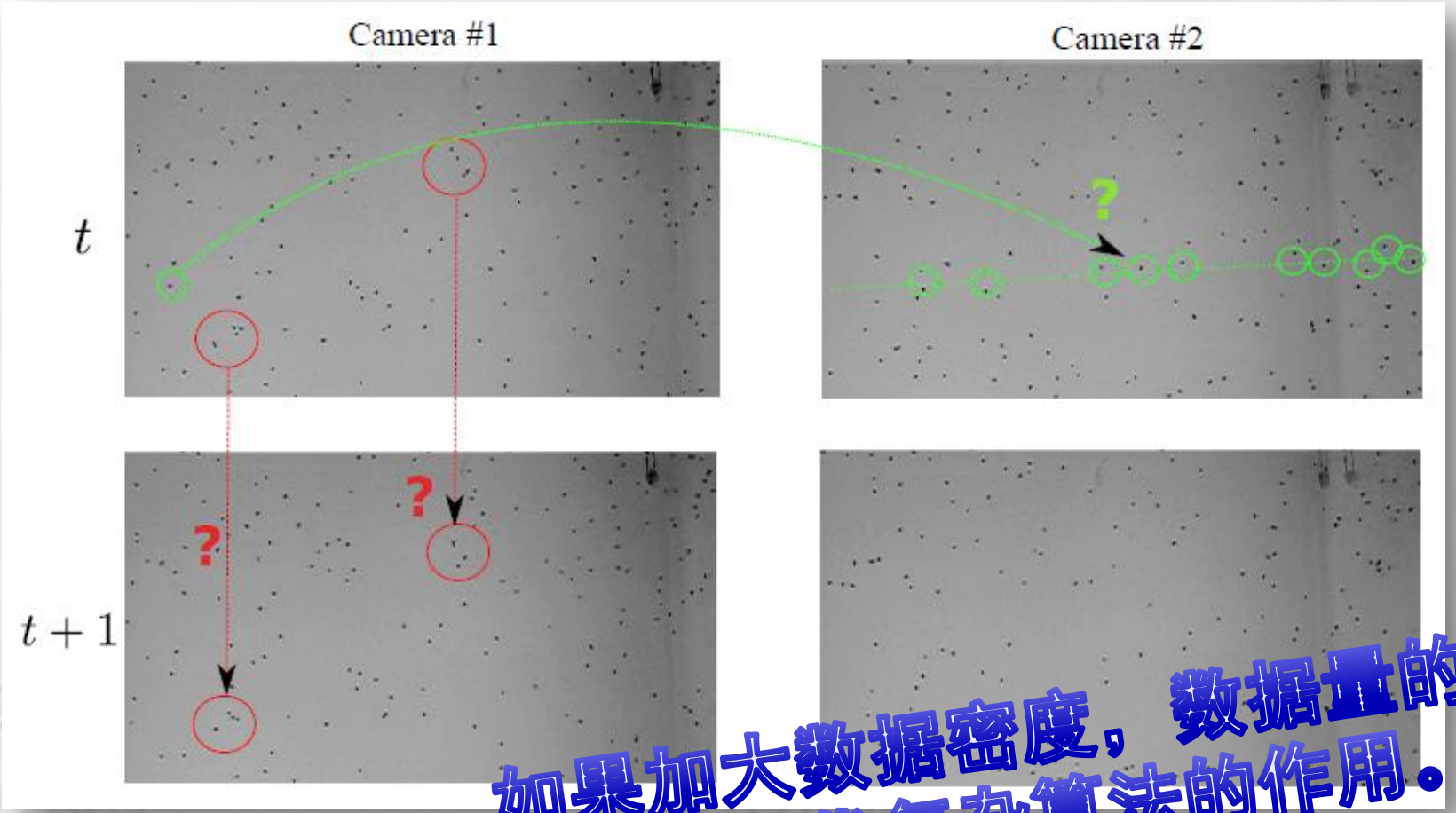
群目标三维跟踪



- 发现与归纳人群在各种场合下的运动规律，有助于场地道路的优化规划。
- 发现果蝇群、斑马鱼群、细胞群等的生物群体运动规律，有助于揭示感知认知、社会行为背后的传感与神经信息处理机理。

——摘自陈雁秋香山大数据论坛的PPT

跟踪群目标的挑战在哪里？



如果加大数据密度，数据量的增加会替代复杂算法的作用。

——图片摘自陈雁秋香山大数据论坛的PPT

安全，从第一范式到第四范式

几十年来 **科学实验**

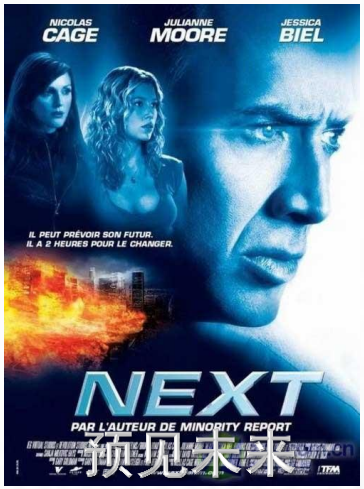
数百年来 **模型归纳**

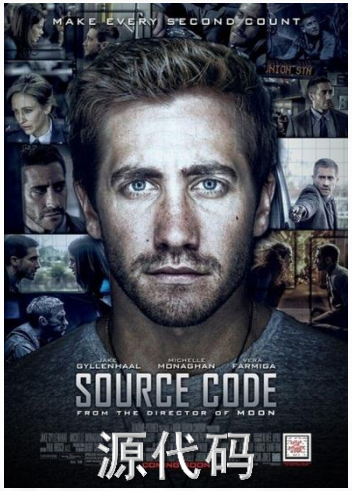
数十年来 **模拟仿真**

今天 **数据密集型**

- 渗透测试
- 事件分析
- 漏洞挖掘
- 地址随机
- …

- 检测引擎
- 病毒特征
- 漏洞特征
- 攻击特征
- 关联规则
- …





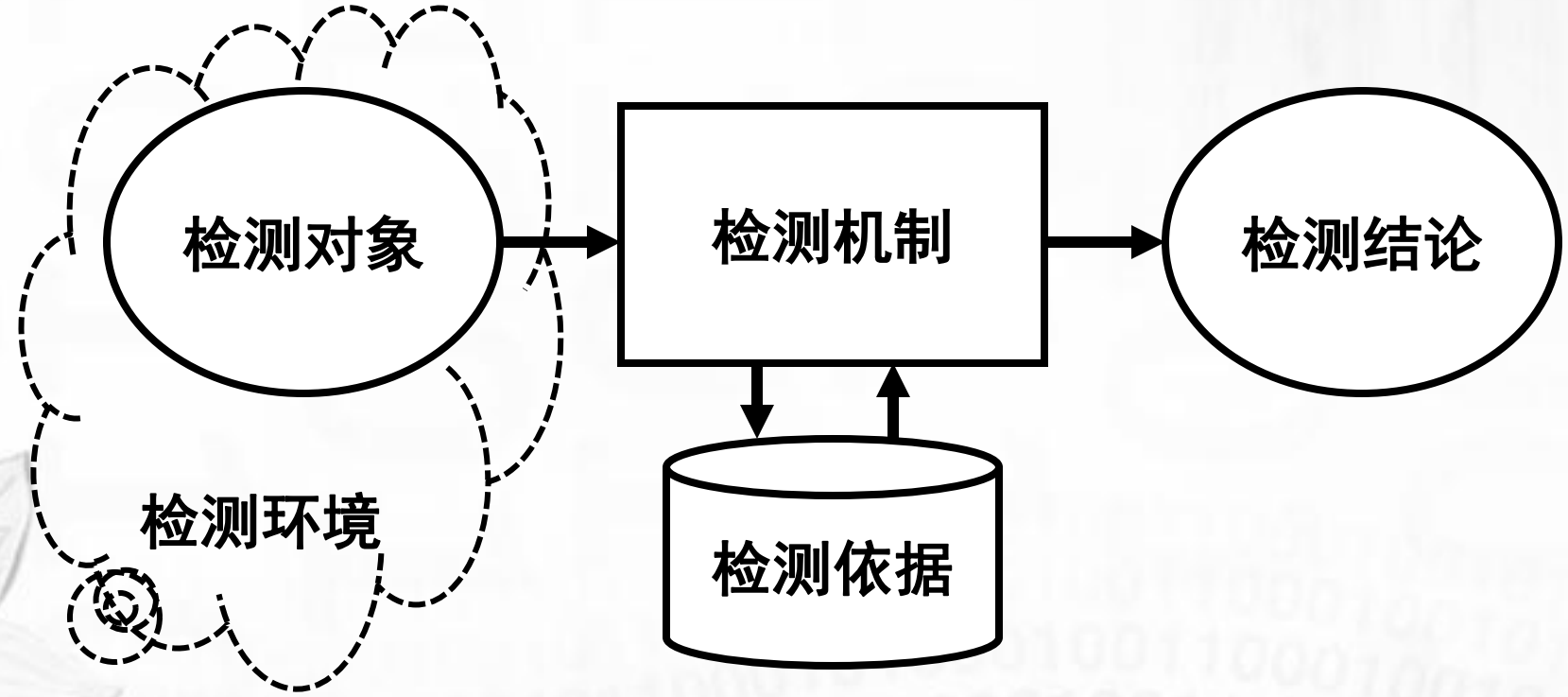
- 模拟攻击
- 模拟被攻击
- …

- 沙箱
- 蜜罐
- 标识检测
- …

- 数据密度
- 基于记忆
- 数据浓缩
- …

探寻检测的逻辑模式

- 检测的一般抽象模型



检测对象类型：A系统 B数据流 C数据体 S体系

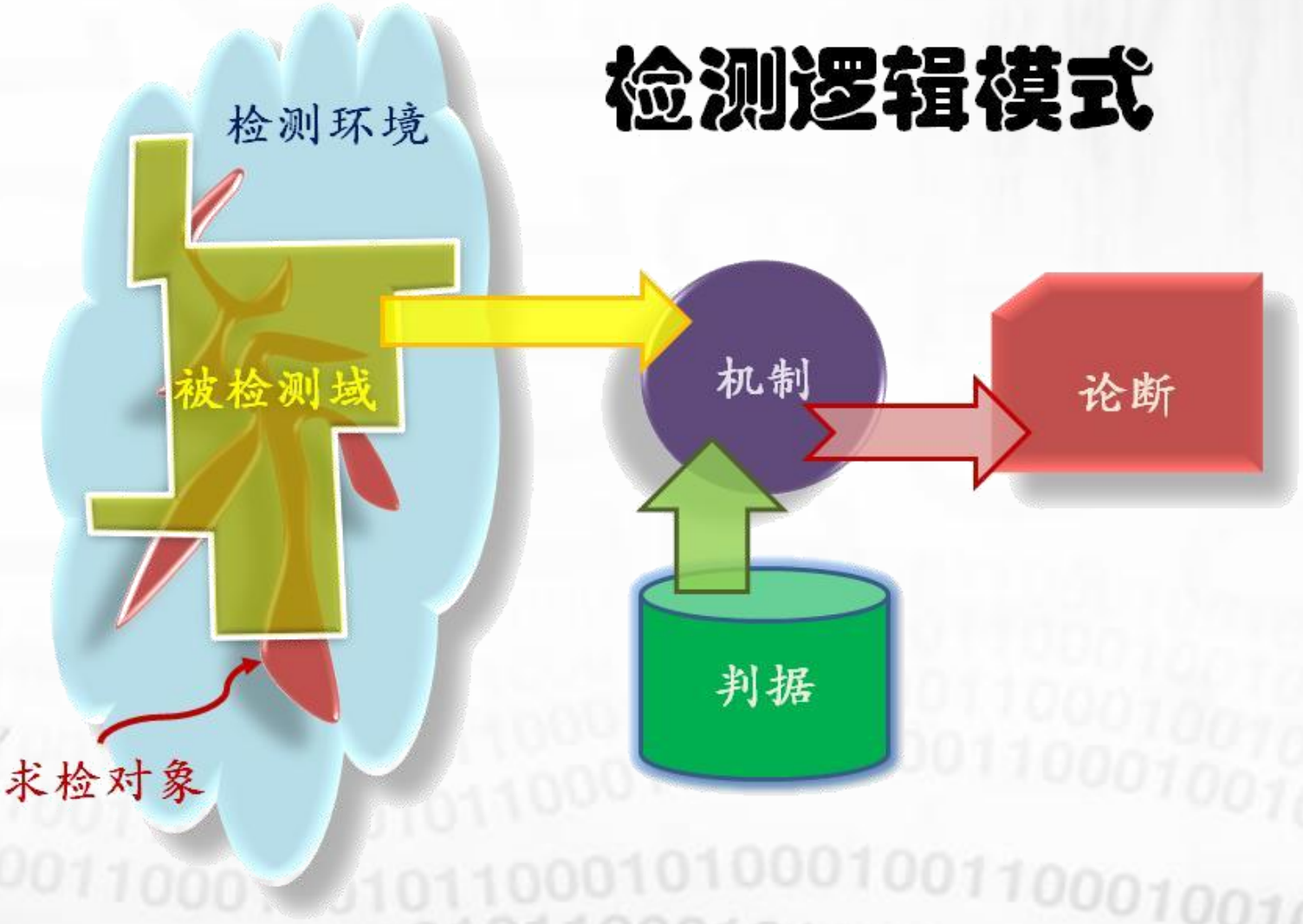
当前典型的微观检测步骤模式

设备类安全检测产品
工具类安全检测产品

SOC等安全管理平台

探寻检测的真谛

检测逻辑模式



微观检测的新步骤模式

X觉观念：比如视觉和关注，何谓视，何谓觉

认知的全程动态反馈模式

扩大　浓缩　精确　场景

对象增加
空间范围扩展
空间密度加大
时间区间扩展
时间粒度增加
速度增加
知识类型增多
…

基于记忆的检测方法群

记忆的关键
是忘记什么

【4】大数据的作用力和变换反应

将大数据变小（物理作用）

42

- 在尽量不损失价值的条件下，减小数据规模
- 不改变数据基本属性情况下的数据清洗
- 抽样、去重、过滤、筛选、压缩、索引、提取元数据等等方法，可以直接将大数据变小，这种作用类似于所谓的物理式的变小

价值提炼（化学反应）

49

- 大数据探索式考察与可视化将发挥作用，人机的交互分析可以将人的智慧作用融入
- 通过群体智慧、社会计算、认知计算对数据价值的发酵和提炼
- 从数据分析到数据制造

微解读

【4】大数据的作用力和变换反应

42

将大数据变小（物理作用）

- 在尽量不损失价值的条件下，减小数据规模
- 不改变数据基本属性情况下的数据清洗
- 抽样、去重、过滤、筛选、压缩、索引、提取元数据等等方法，可以直接将大数据变小，这种作用类似于所谓的物理[illegible]的变小

扩大　浓缩　精确　场景

49

价值提炼（化学反应）

- 大数据探索式考察与可视化将发挥作用，人机的交互分析可以将人的智慧作用融入
- 通过群体智慧、社会计算、认知计算对数据价值的发酵和提炼
- 从数据分析到数据制造

微解读

【4】大数据的作用力和变换反应

42

将大数据变小（物理作用）

- 在尽量不损失价值的条件下，减小数据规模
- 不改变数据基本属性情况下的数据清洗
- 抽样、去重、过滤、筛选、压缩、索引、提取元数据等等方法，可以直接将大数据变小，这种作用类似于所谓的物理式的变小

扩大　浓缩　精确　场景

49

价值提炼（化学反应）

- 大数据探索式考察与可视化将发挥作用，人机的交互分析可以将人的智慧作用融入
- 通过群体智慧、社会计算、认知计算对数据价值的发酵和提炼
- 从数据分析到数据制造

安全，从第一范式到第四范式

几十年来 **科学实验**

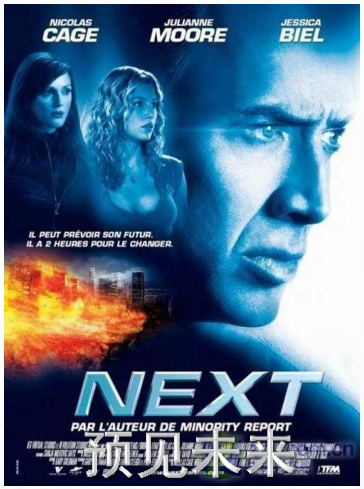
数百年来 **模型归纳**

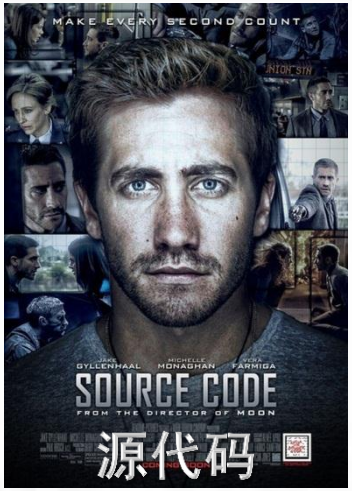
数十年来 **模拟仿真**

今天 **数据密集型**

- 渗透测试
- 事件分析
- 漏洞挖掘
- 地址随机
- …

- 检测引擎
- 病毒特征
- 漏洞特征
- 攻击特征
- 关联规则
- …





- 模拟攻击
- 模拟被攻击
- …

- 沙箱
- 蜜罐
- 标识检测
- …

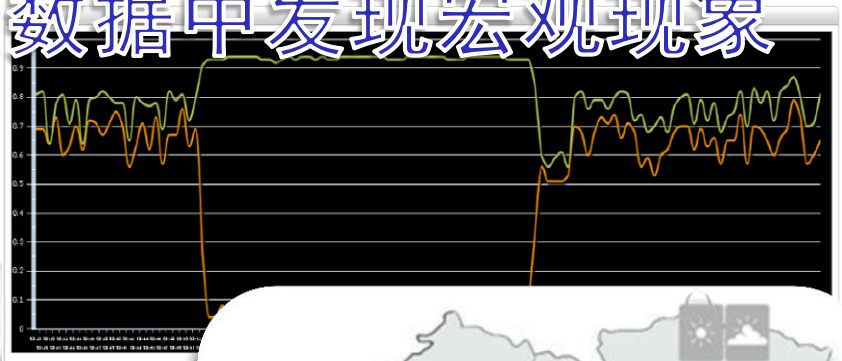
- 数据密度
- 基于记忆
- 数据浓缩
- …

高端信息安全检测都是大数据问题

全局预警——宏观态势感知

- 难点是看不全

大数据中发现宏观现象

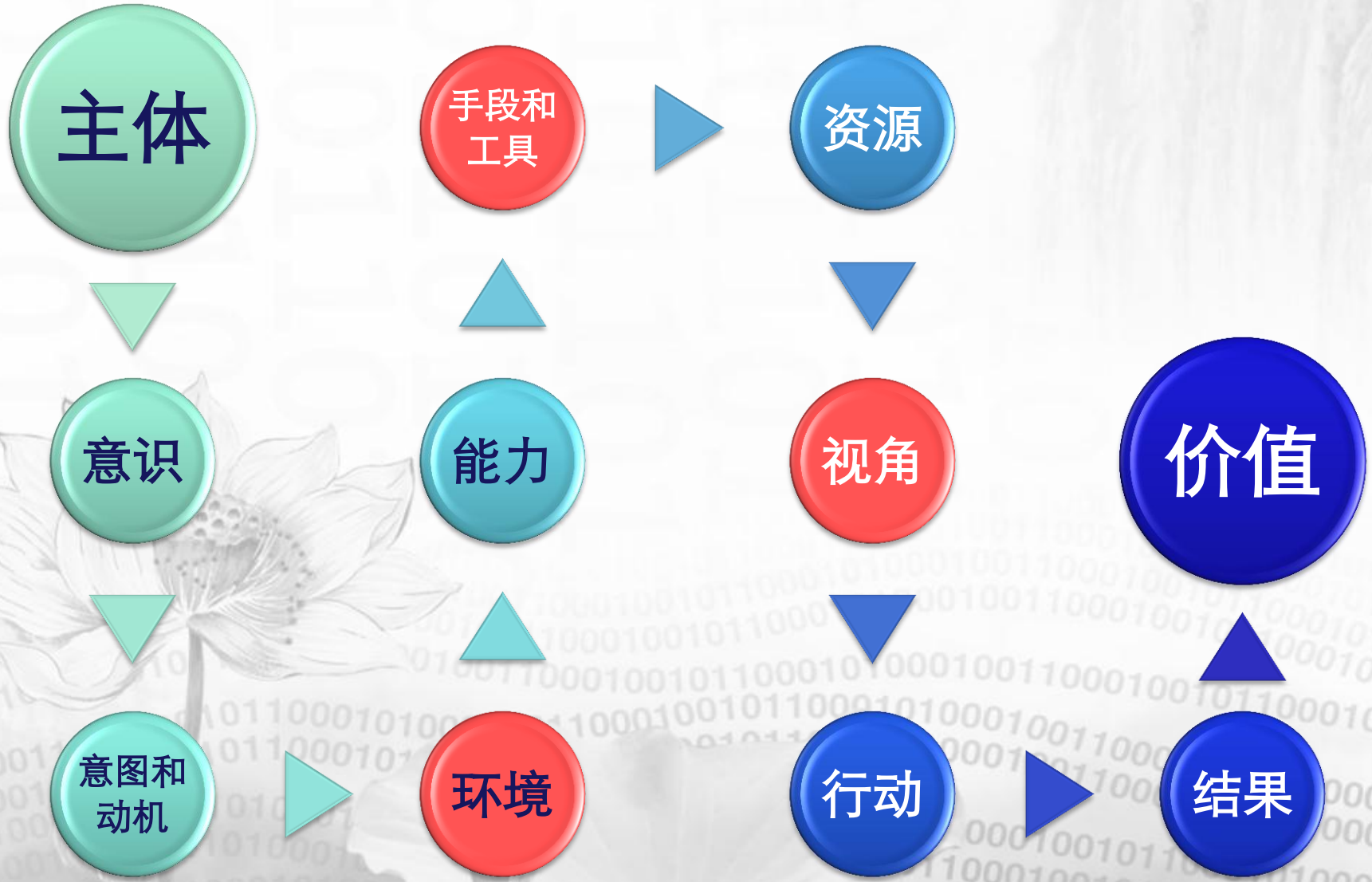


动态预防——APT防范

- 难点是看不见

大数据中发现微观事件

大数据带来了数据视角



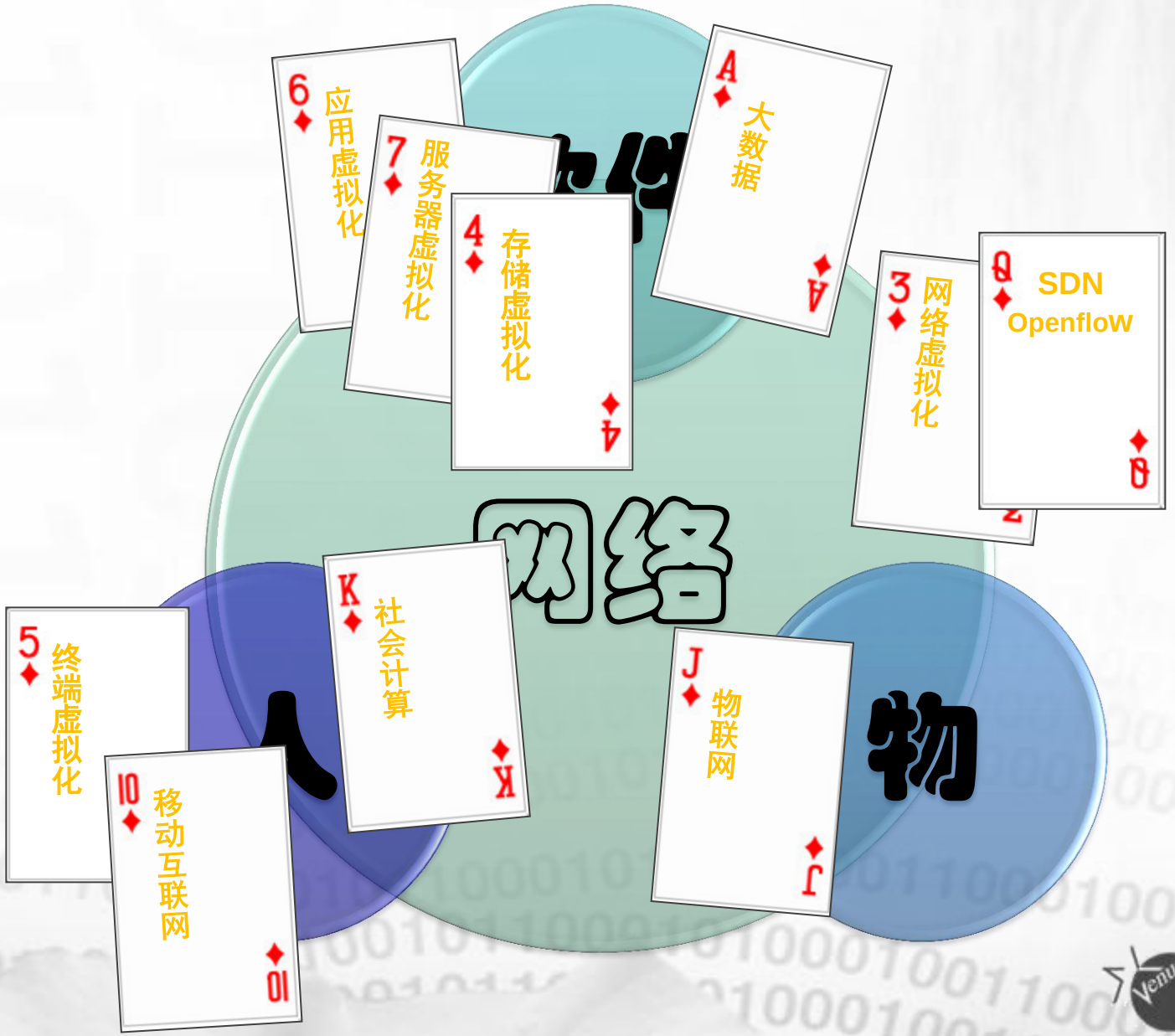
所谓新计算、新网络和新数据

- 大数据
- 社会计算

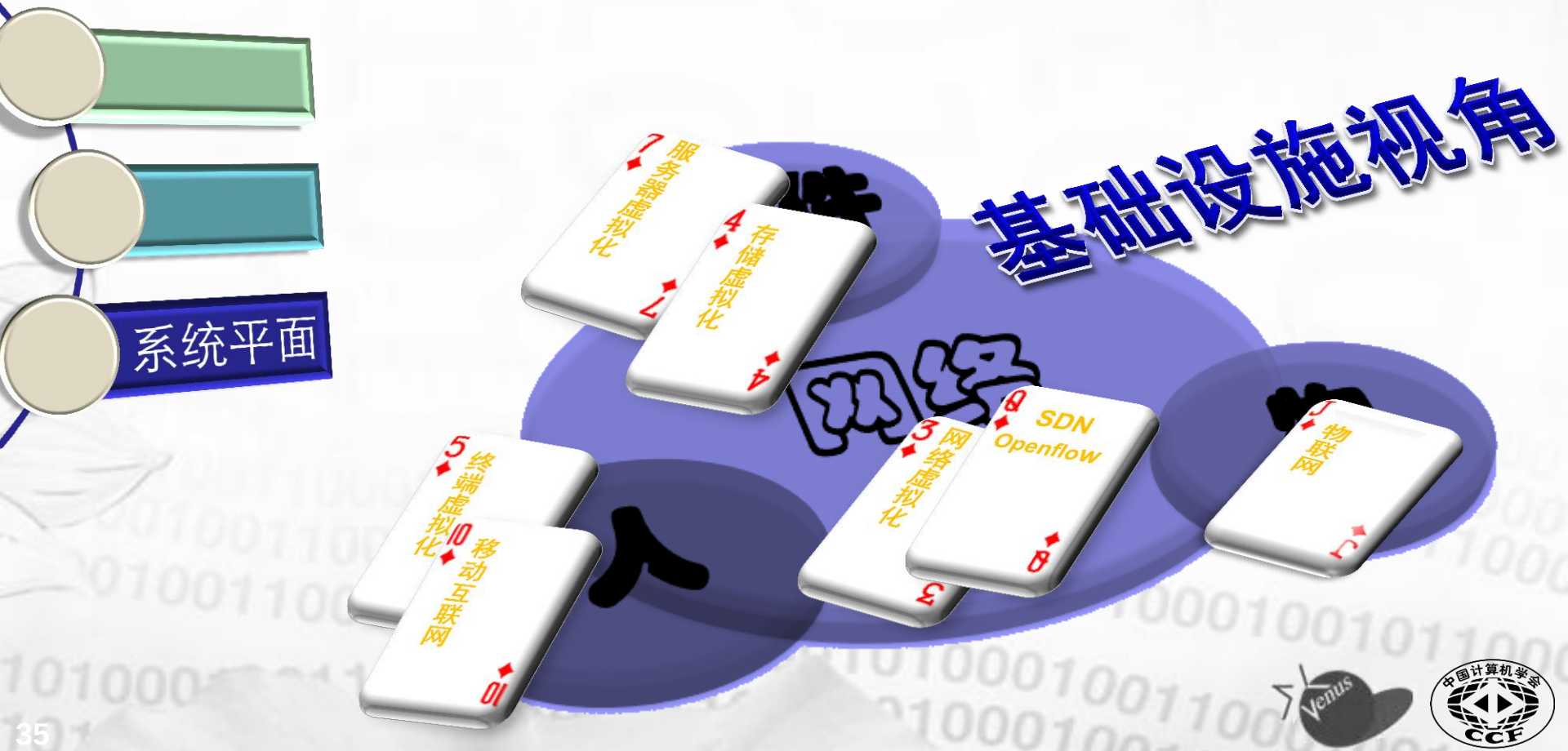
- 移动互联网
- 物联网
- SDN/Openflow

- 云计算
- 虚拟化
- 高性能

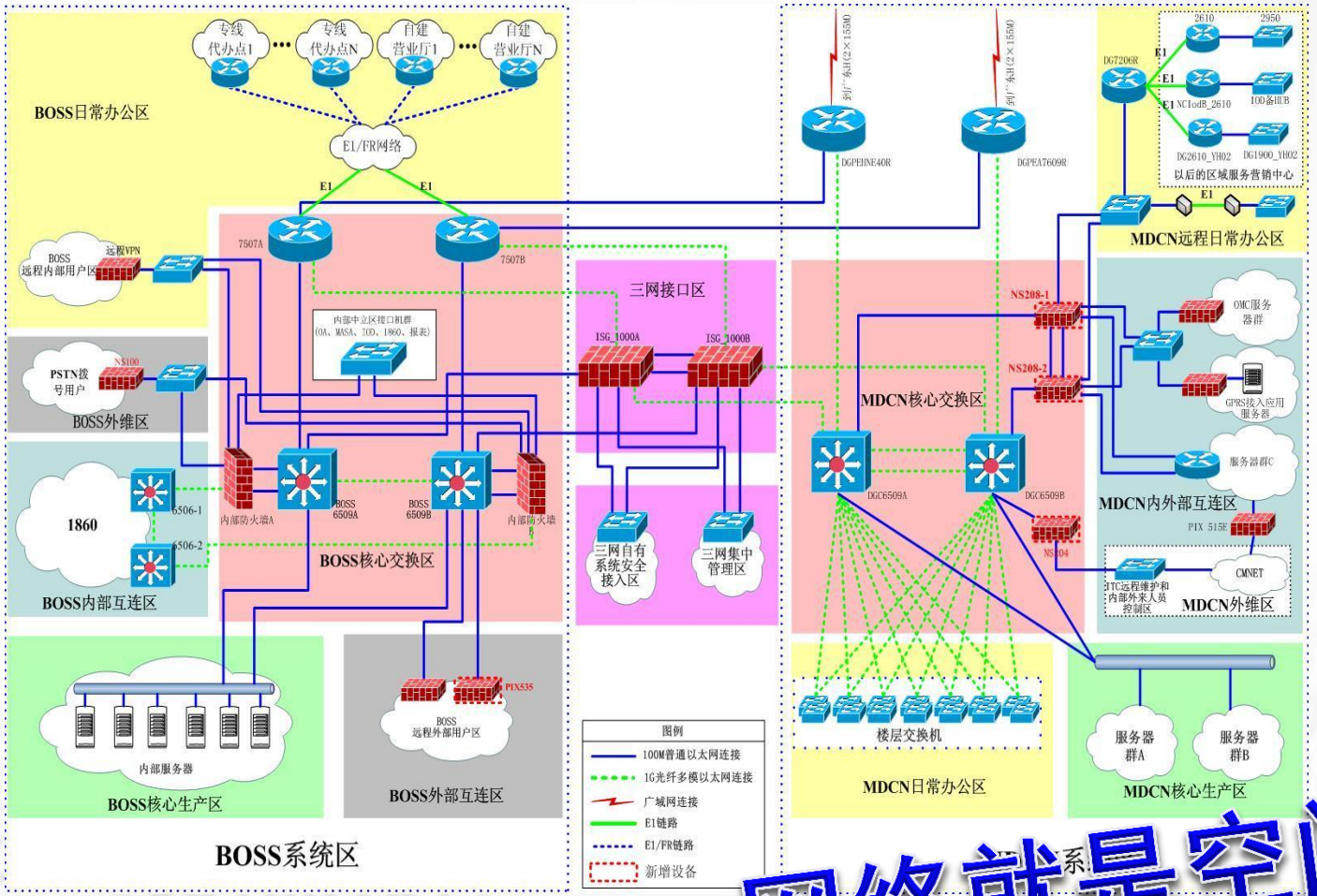
网络连接【人、物和软件】


6 应用虚拟化
7 服务器虚拟化
4 存储虚拟化
A 大数据
3 网络虚拟化
Q SDN Openflow
网络
5 终端虚拟化
10 移动互联网
K 社会计算
J 物联网
人
物

D.O.S.三个平面


系统平面
基础设施视角
网络
人
7 服务器虚拟化
4 存储虚拟化
5 终端虚拟化
10 移动互联网
3 网络虚拟化
8 SDN Openflow
J 物联网

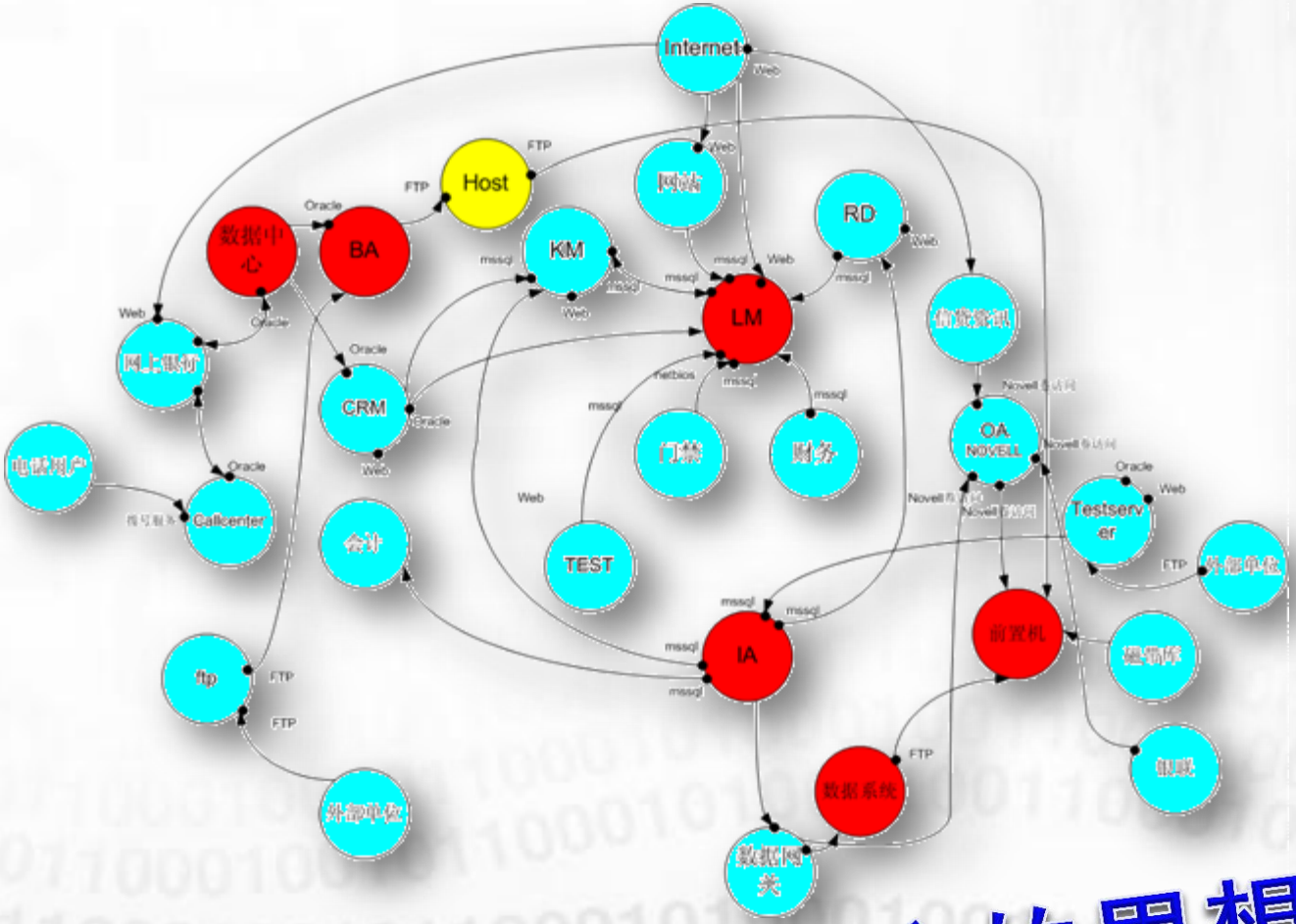
典型的系统视图



网络就是空间

D.O.S.三个平面

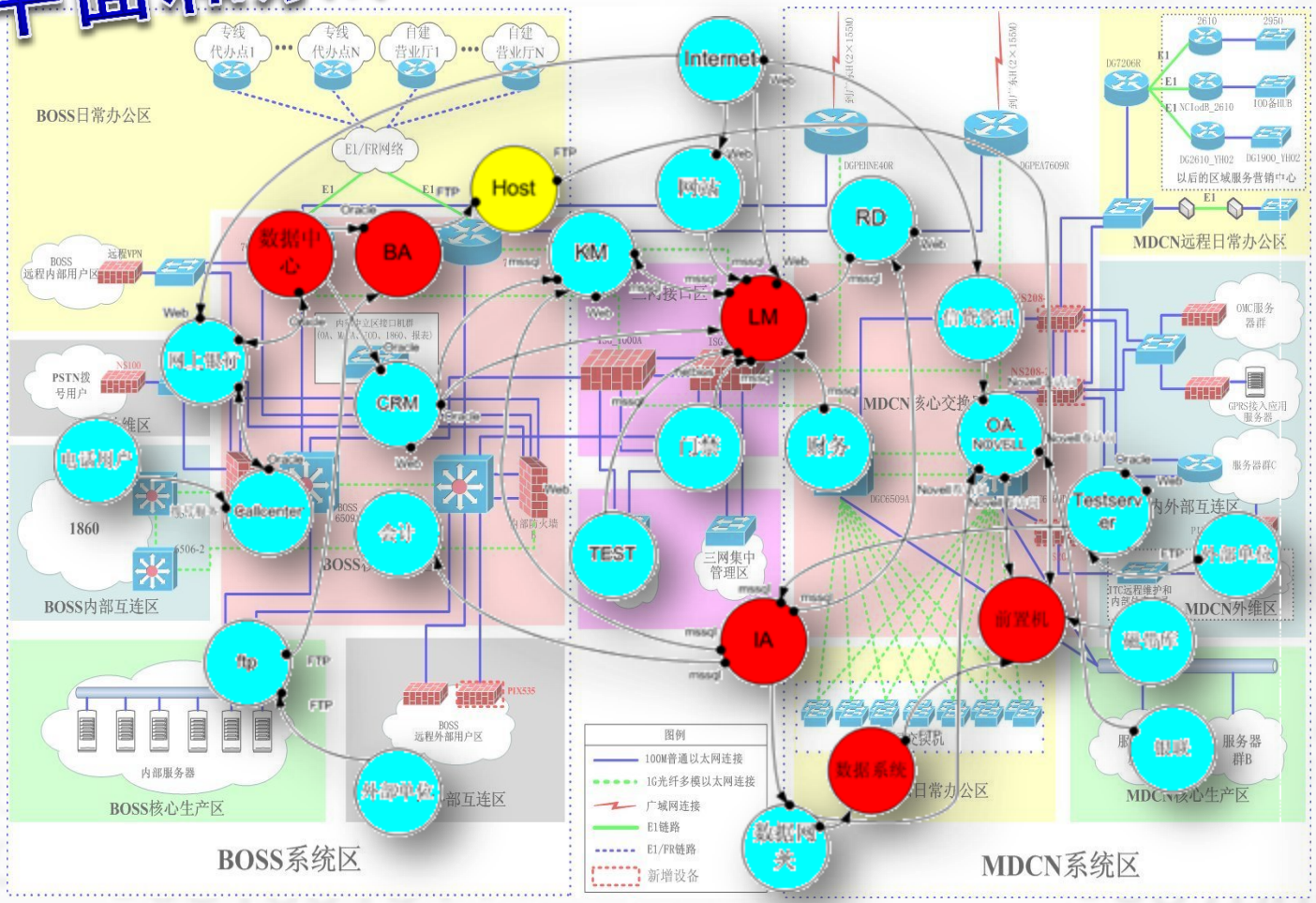
业务流


Internet
Web
Host
FTP
网站
RD
数据中心
BA
Oracle
KM
mssql
LM
网上银行
CRM
电话用户
Callcenter
财务
门禁
OA
NOVELL
Testserver
TEST
会计
ftp
IA
前置机
数据系统
外部单位
Netbios

流的思想

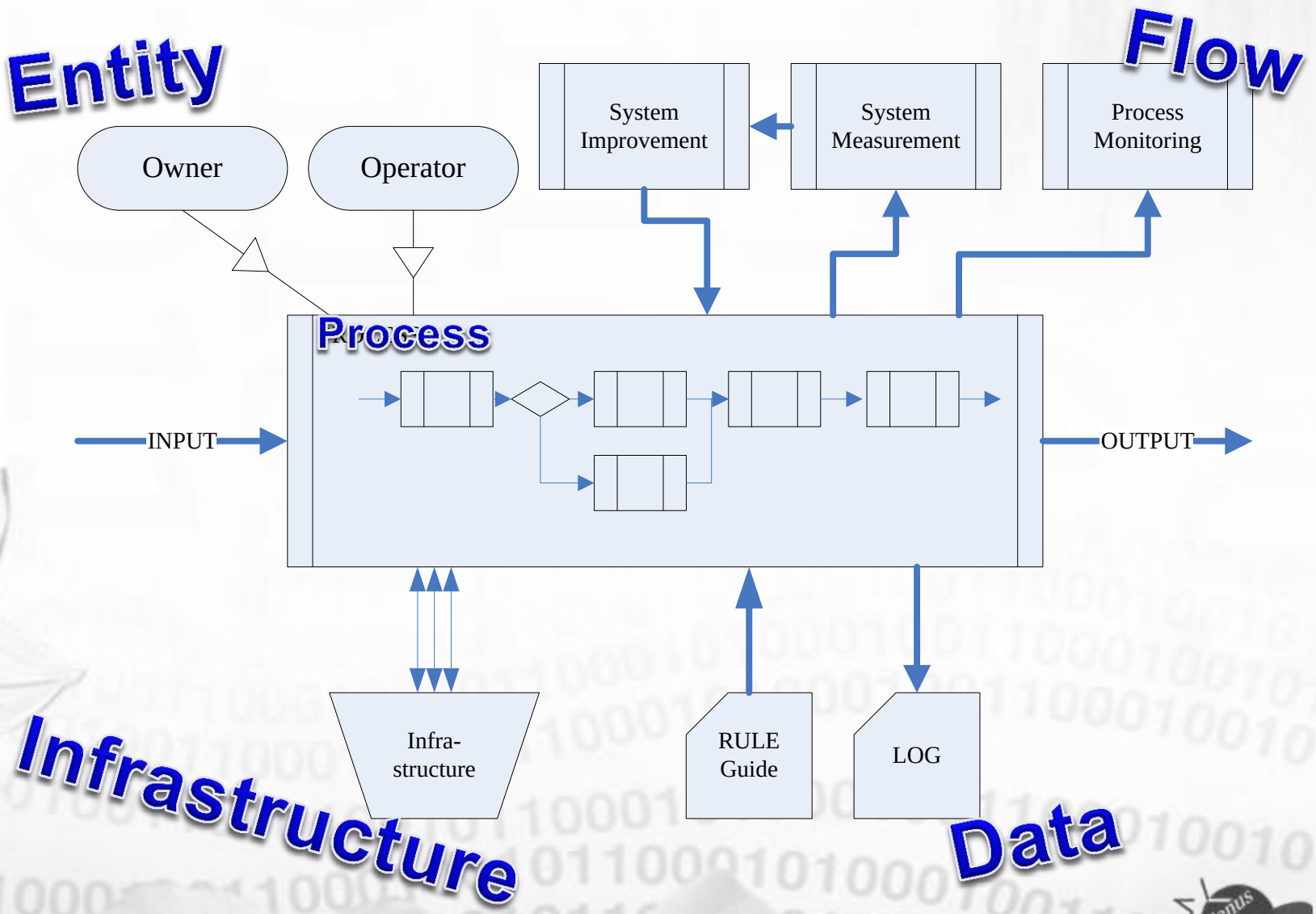
业务流@网络结构

服务平面和系统平面的交叠



比如：系统平面中的路径，在服务平面中只是一个服务交锋面

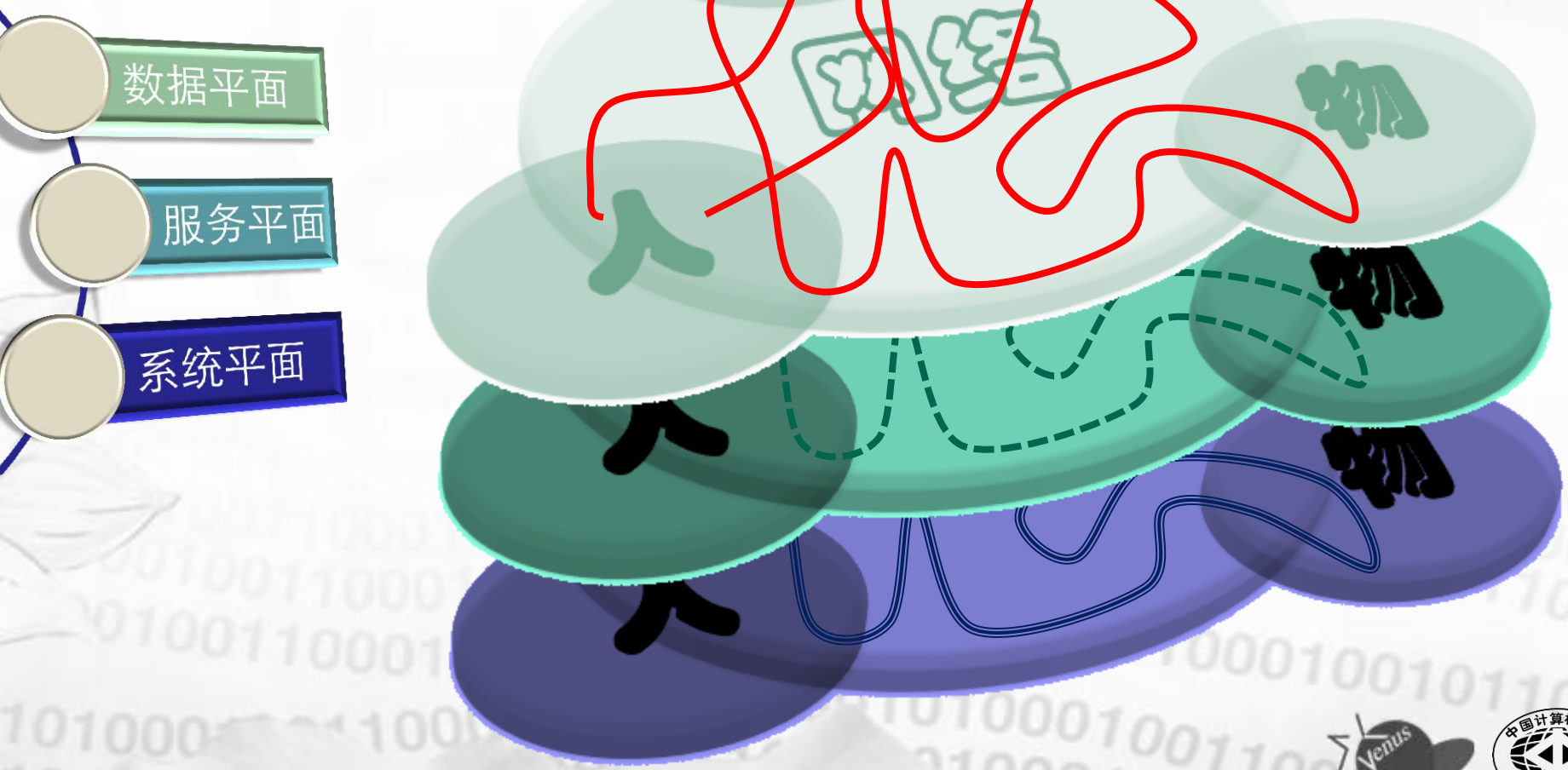
EDIF，过程认识


Entity
Flow
Owner
Operator
System Improvement
System Measurement
Process Monitoring
Process
INPUT
OUTPUT
Infra-structure
RULE Guide
LOG
Infrastructure
Data

D.O.S.三个平面

D.O.S.三个平面

数据视角、数据科学观念，比所谓大数据更有深意。



微解读

【1】数据科学与大数据的学科边界

数据的科学问题

55

- 数据界与物理界、人类社会之间的关联与差异?
- 是否存在独立于领域的数据科学?
- 数据科学的分类谱系
- 大数据的复杂性主要来自个体之间的联系
- 学习理论和认知理论等应当是数据科学的重要组成部分

大数据的基本问题

(定义、结构等)

43

- 需要对“大数据”给出科学定义，清晰说明其内涵与外延
- 大数据区别于其他数据的关键特性是什么?
  - 3V
  - 高价值总量、低价值密度
- 大数据意味着全数据?
- 需要为动态、高维、复杂大数据建立形式化、结构化描述，并在此基础上发展大数据处理技术

大数据核心问题

——CCF大数据专家委

数据视角

数据观察

- **数据的生命周期**
  - 数化-处置-价值化-逆数化
- **面向数据的XX**
  - 就像面向对象带来的变化
- **数据结构性**
  - 所谓非结构也是某种结构
  - 显性结构和隐形结构
  - 区别于程序的结构
- **数据质量问题**
- **…**

数据奇思

- **数据如何变成活体**
  - Agent化
  - 数据如何像生物一样寄生在系统之中
- **数据的谱系(分类)**
- **数据是实体**
- **数据不是实体**

大数据带来攻击的变化

有些攻击变难了

- 大数据常常意味着数据及其承载系统的分布式和鲁棒性
- 单个数据和系统的价值相对降低
- 空间和时间的大跨度，价值的稀疏，使得寻找价值攻击点更不容易

有些攻击变容易了

- 微观攻击被掩盖在大XX下面
- 完全的去中心化很难，只要存在中心就可能成为被攻击的穴道
  - 枢纽中心、管理中心
- 对于低密度价值的提炼过程也是吸引攻击的招摇过程

数据视角的独特攻击思维

系统视角和服务视角

- 原先传统的系统攻击依然有效。
  - 数据总要承载在某些系统上
- 从大数据的工作流程中，可以寻找到破坏的契机
  - 比如：DNS服务
  - 比如：人的选择迷惑

数据视角

- 数据污染
- 病毒式传播
- 奇点破坏
- 高维空间下的群聚和离群点
- 所谓信息主权，更容易体现在系统和数据

围绕价值攻击

围绕数据的攻防

- 隐私和安全
- 数据迷惑和数据隐藏
- 数据脱敏

围绕人的攻防

- 人的ID化
- 人的数量变成一个不太大的大数
- 围绕人的数据标识分类
- 人性的弱点

大数据、大价值


几十年来
科学实验
数百年来
模型归纳
数十年来
模拟仿真
今天
数据密集型
主体
手段和工具
资源
资源平面
服务平面
系统平面
软件
资源和价值视角
人
物
意识
能力
视角
价值
意图和动机
环境
行动
结果
资产和业务
Asset
保障措施
Safeguard
威胁
Threat

思索中…

微博 @潘柱廷.

微刊《信息安全美学》

微刊《大数据安全》