Part1:

（漏洞检测（业务）+ 大数据基础架构（承托））

源代码安全检测 是安全中的重要 研究领域，通过 静态代码扫描 对软件脆弱性进行智能分析，企业能大幅度提升其代码质量，减少被攻击风险。（注：蓝盾在此前并没有开展相关研究工作）

本研究项目的目标是，构筑1个基于深度学习和大数据的漏洞检测引擎。这在工业界和学术界均为第一（已知文献只有VulDeePecker）。新方法和旧方法的不同之处在于：

* 旧方法 基于 模式匹配（规则），需要领域专家介入撰写大量规则，而新方法却不用；
* 新方法能有效检测未知漏洞而旧方法只对已知漏洞有效

通过此引擎，我们能快速对源代码进行漏洞检测，并输出具体漏洞编号和行号。More，此引擎能快速部署到企业（蓝盾）云平台中，与客户开发和测试环境无缝整合，以此加速整个SDLC生命周期，同时符合从DevOp到DevSecOp的安全诉求。

（缺少数据支撑，换句话说，还没有来得及做实验（也可以说成是下一步目标），已经做的：checkmarx，flawfinder，vuddy的试用以及读书笔记）

承托 蓝盾漏洞检测业务 的是其 底层高效的大数据体系架构。直到目前为止，我们集群系统的计算力如下：？CPU，？内存，？硬盘。对于数据存储和容错性需求，我们选择Hadoop的HDFS；对于大数据计算和分析需求，我们选择Spark；**我们把此集群的计算力公开，任何蓝盾的同学或团队感兴趣，均可以使用（请大声说）。**

值得一提的是，我们在此基础上构筑了大量脚本，以增加大数据体系的自动化程度，如

* 爬虫（Crawler），周期性从各数据源中收集数据（如病毒样本，病毒分析报告以及程序源代码等）；
* AI模型推送 （Model Serving & Pushing），定时无缝更新客户环境中的AI模型以达到最优效能；
* 大规模数据分析（Data Analysis at scale），对千万级别数据样本作交互性分析（如使用Spark/Python Notebook等）；
* 大规模机器学习流水线（Machine Learning Pipeline），模块化data collection，data preprocessing，feature extraction，model training及prediction。使其拥有流水线的借偶性优点
* 提供数百种AI算法模型供调用（包括传统AI和深度学习）

（说得可能比较虚）

Part2:

数据图表

* 总数据量，每日新增量，每日下载量等（举出数据下载的例子）
* backblaze截图，hdfs截图(容错性，数据交换，视频共享，分析平台中重要的输入元)，spark截图(那一台是master，有多少个worker，所以提升内存以及网络的速度（在每一次进行算法迭代的过程当中，我们都有meta data需要传递）)，spark notebook截图（交互性数据分析），并提供内网IP地址
* （陈述的目标：是告诉他们，大数据是一套工具）
* 数据驱动