

“可执行文件”数据挖掘初探

利用数据挖掘技术探求计算机软件领域所面临问题的解决途径

吴章金 <wuzhangjin@gmail.com>



目录

1

软件领域所面临的问题

2

寻求解决问题的突破口

3

引入数据挖掘技术的可行性

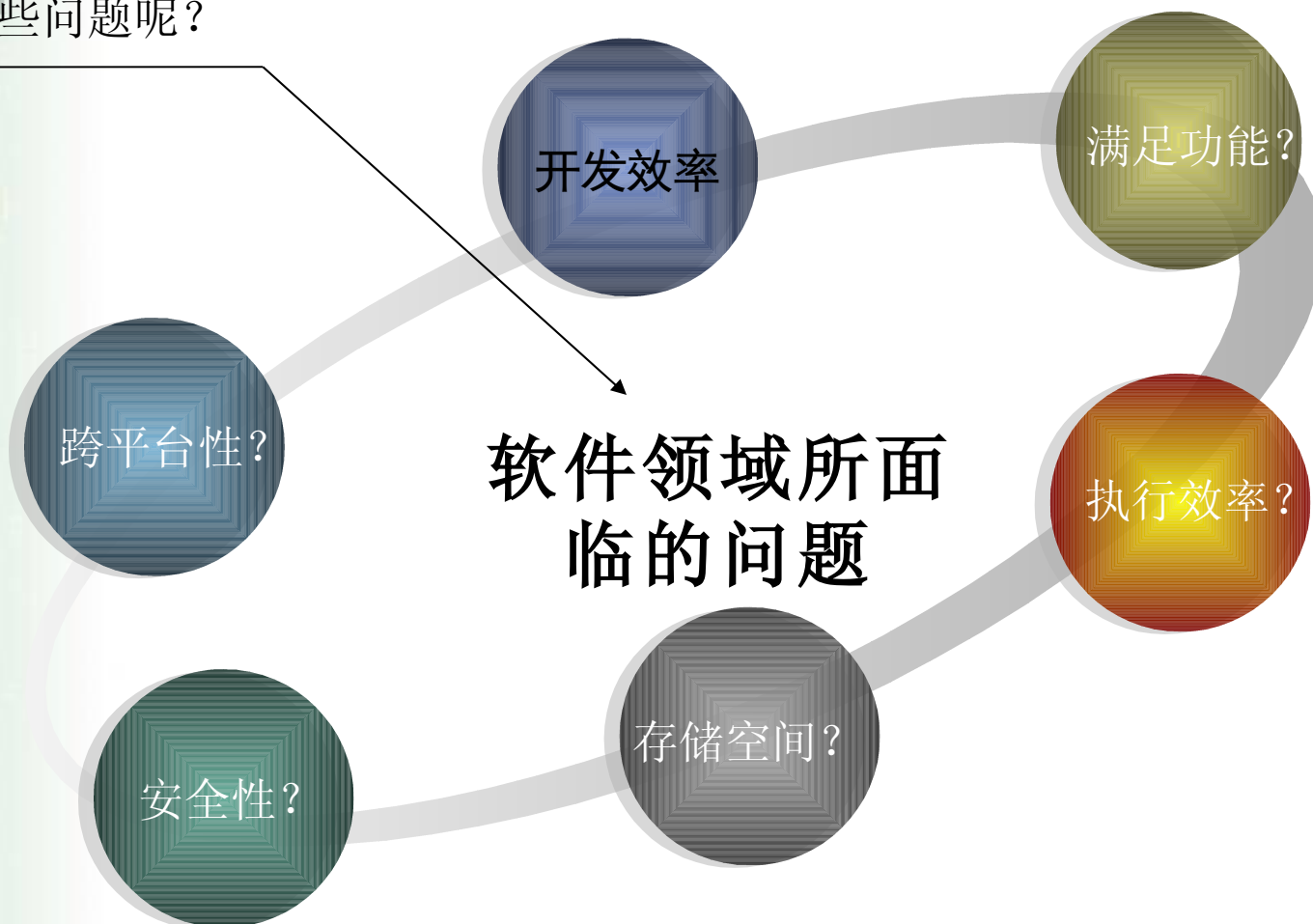
4

实例介绍



软件领域所面临的问题

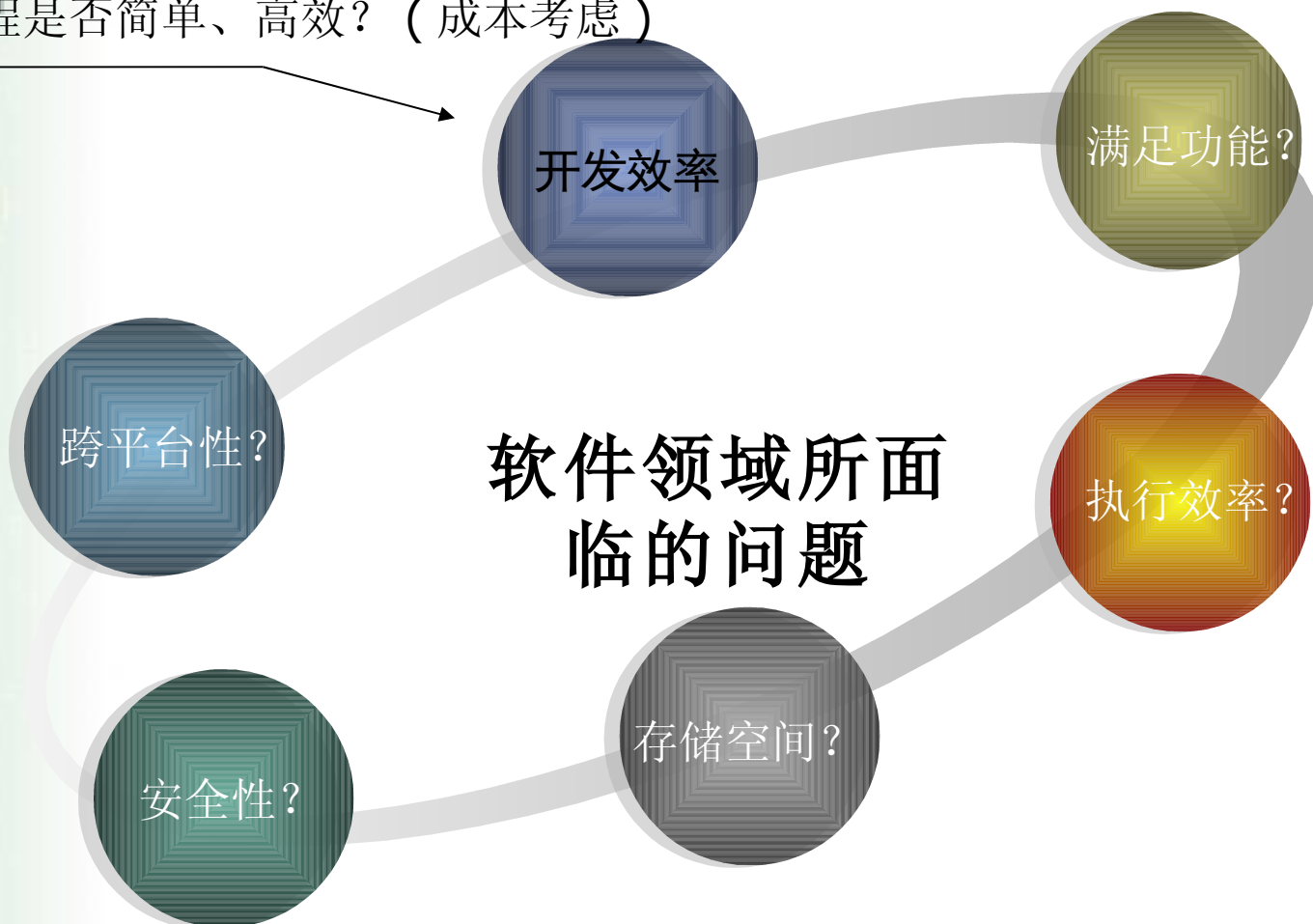
面临哪些问题呢？





软件领域所面临的问题

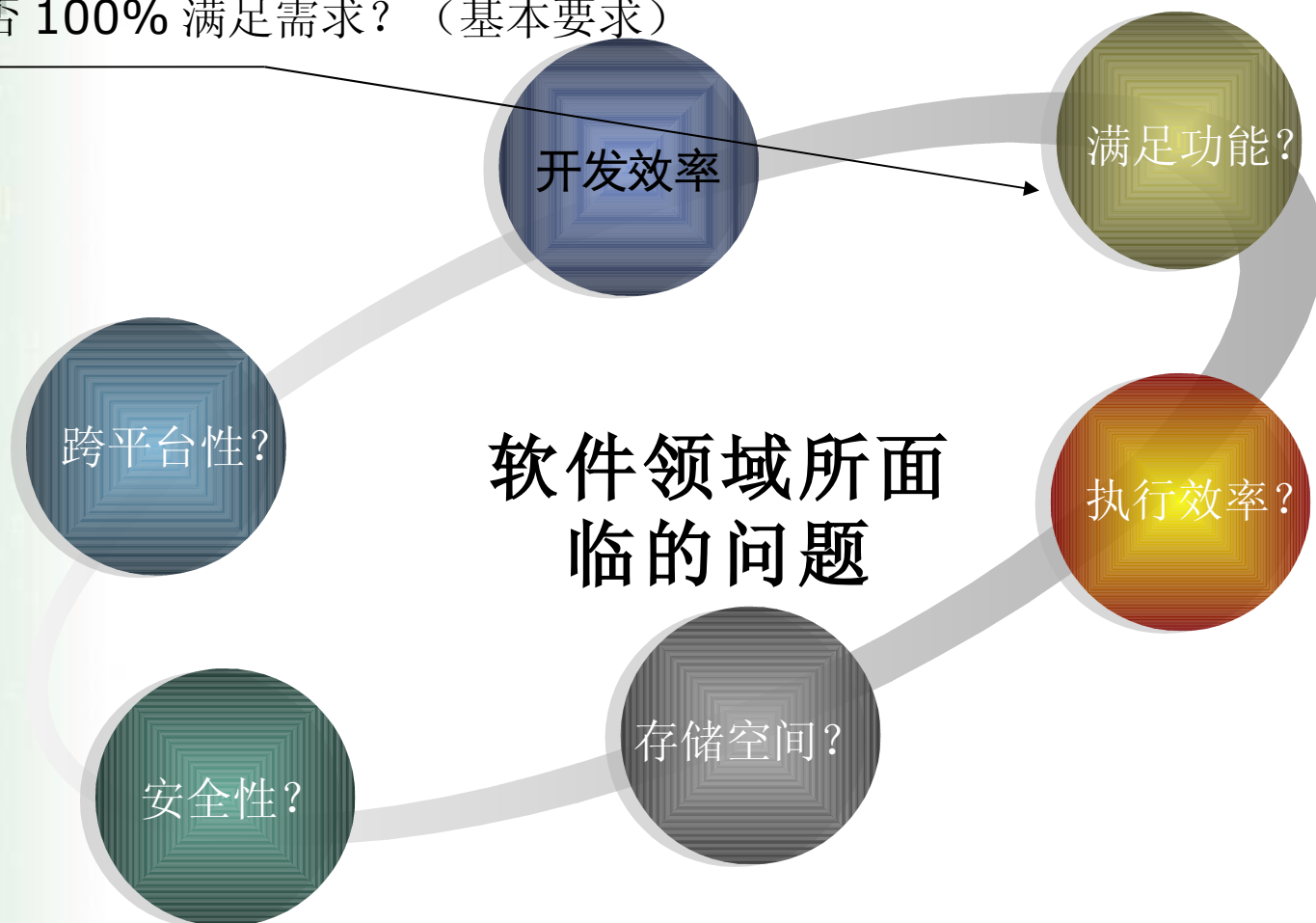
开发过程是否简单、高效？（成本考虑）





软件领域所面临的问题

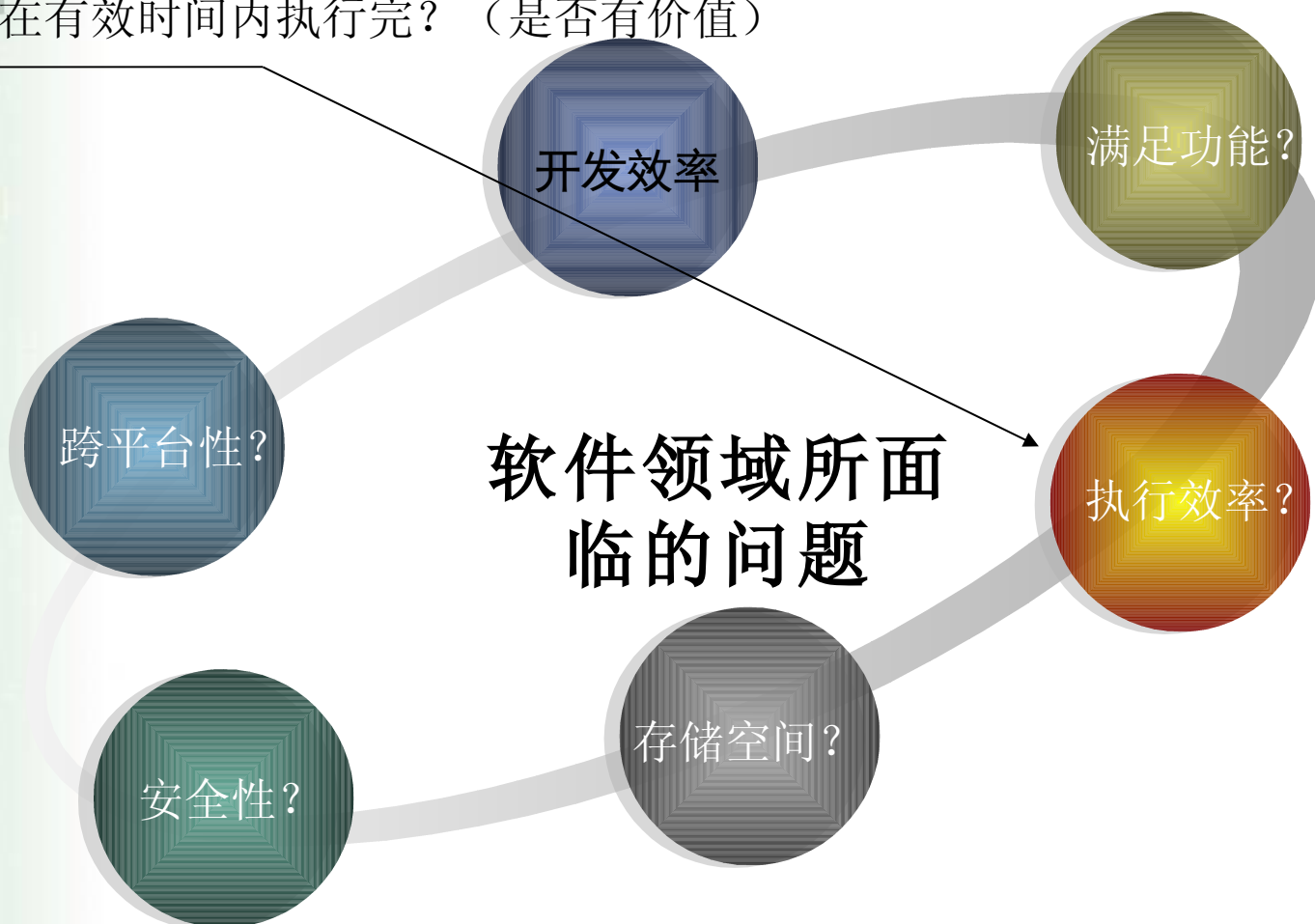
程序是否 100% 满足需求？（基本要求）





软件领域所面临的问题

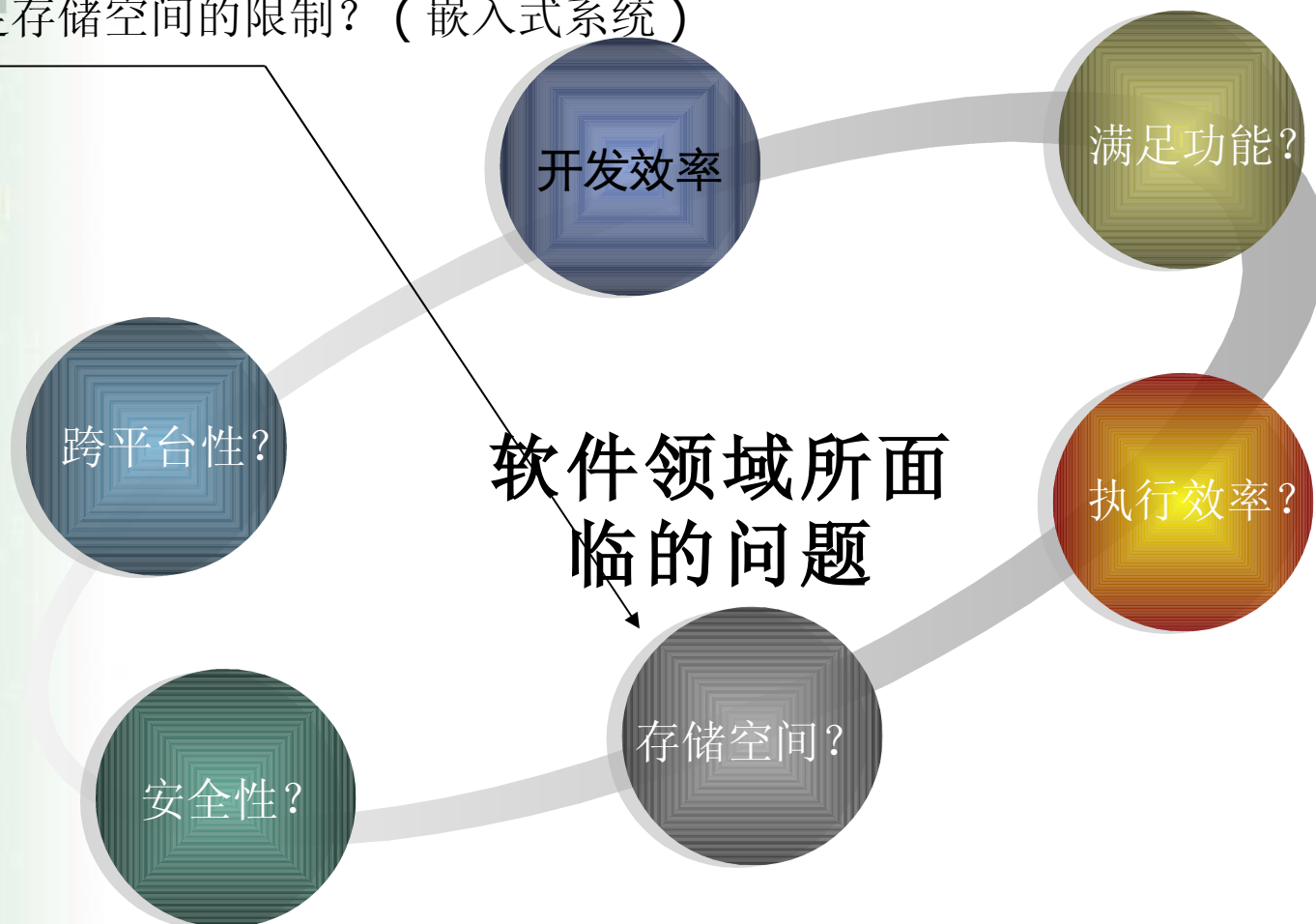
代码是否在有效时间内执行完？（是否有价值）





软件领域所面临的问题

是否满足存储空间的限制？（嵌入式系统）





软件领域所面临的问题

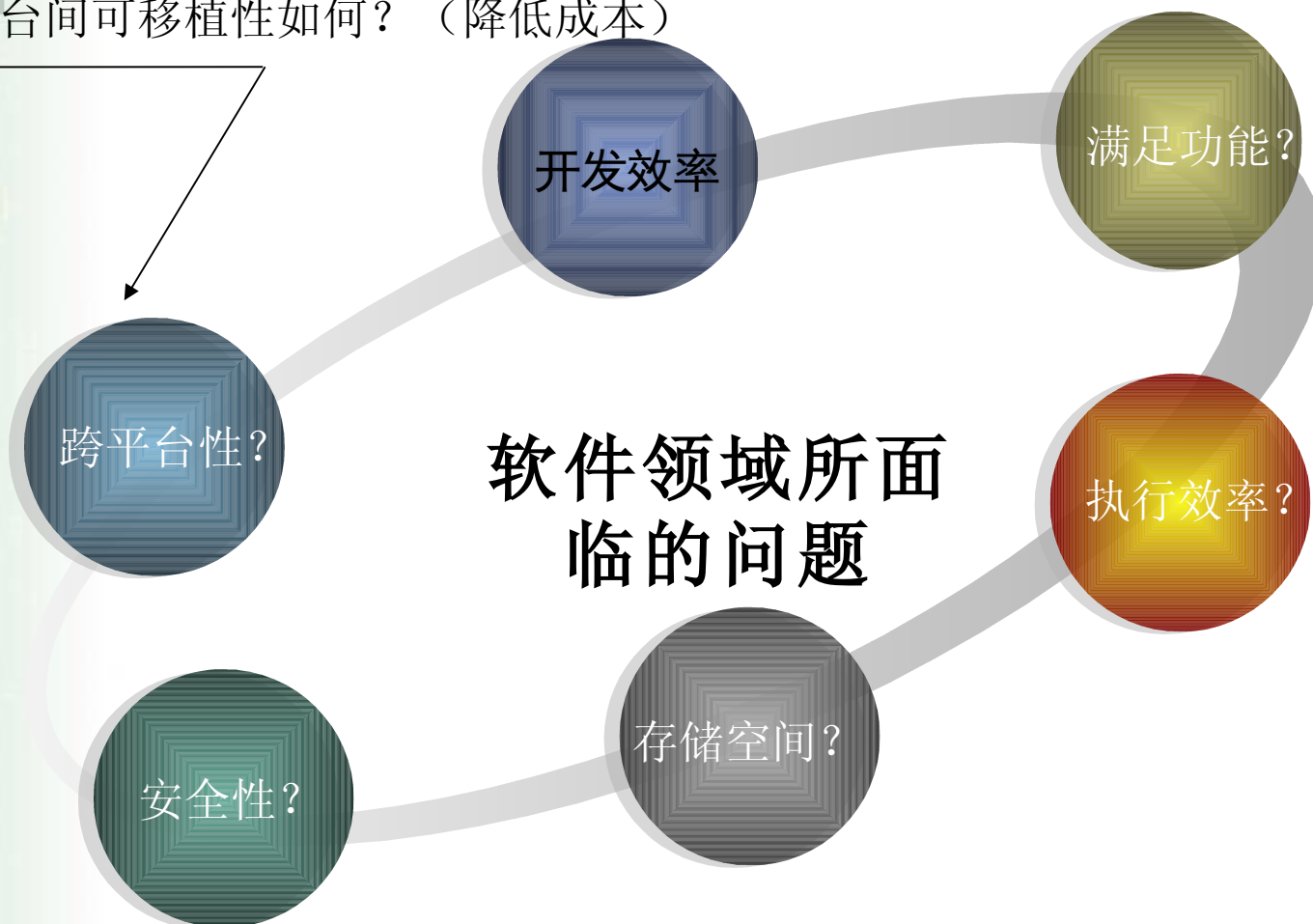
容易注入病毒吗？（空洞：全零区域）





软件领域所面临的问题

在不同平台间可移植性如何？（降低成本）





目录

1

软件领域所面临的问题

2

寻求解决问题的突破口

3

引入数据挖掘技术的可行性

4

实例介绍

寻求解决问题的突破口

可执行文件

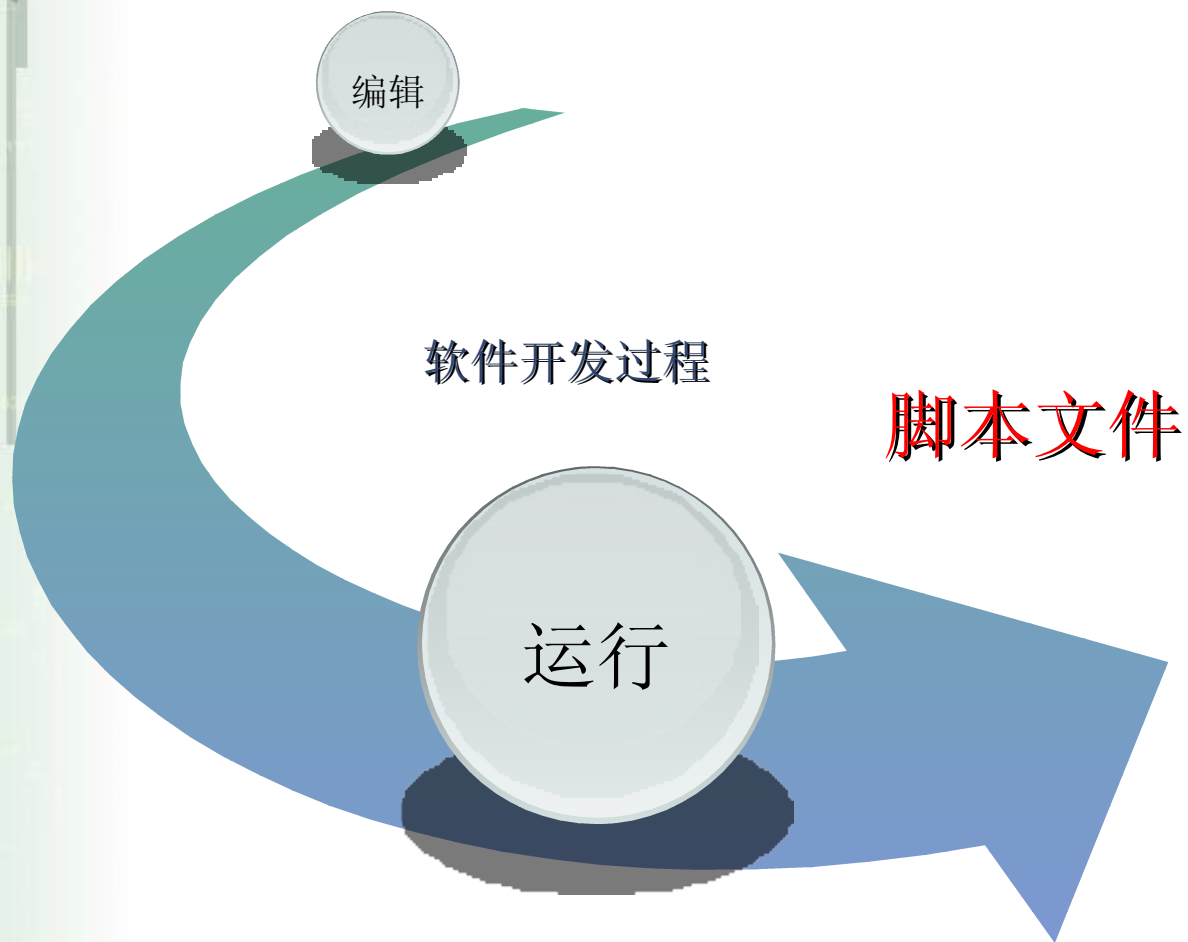
in computer science, is a file **whose contents are meant to be treated as a program by a computer.**
(wikipedia.org)

a file that **may be executed by typing its name as a command.**
(engineering.purdue.edu)

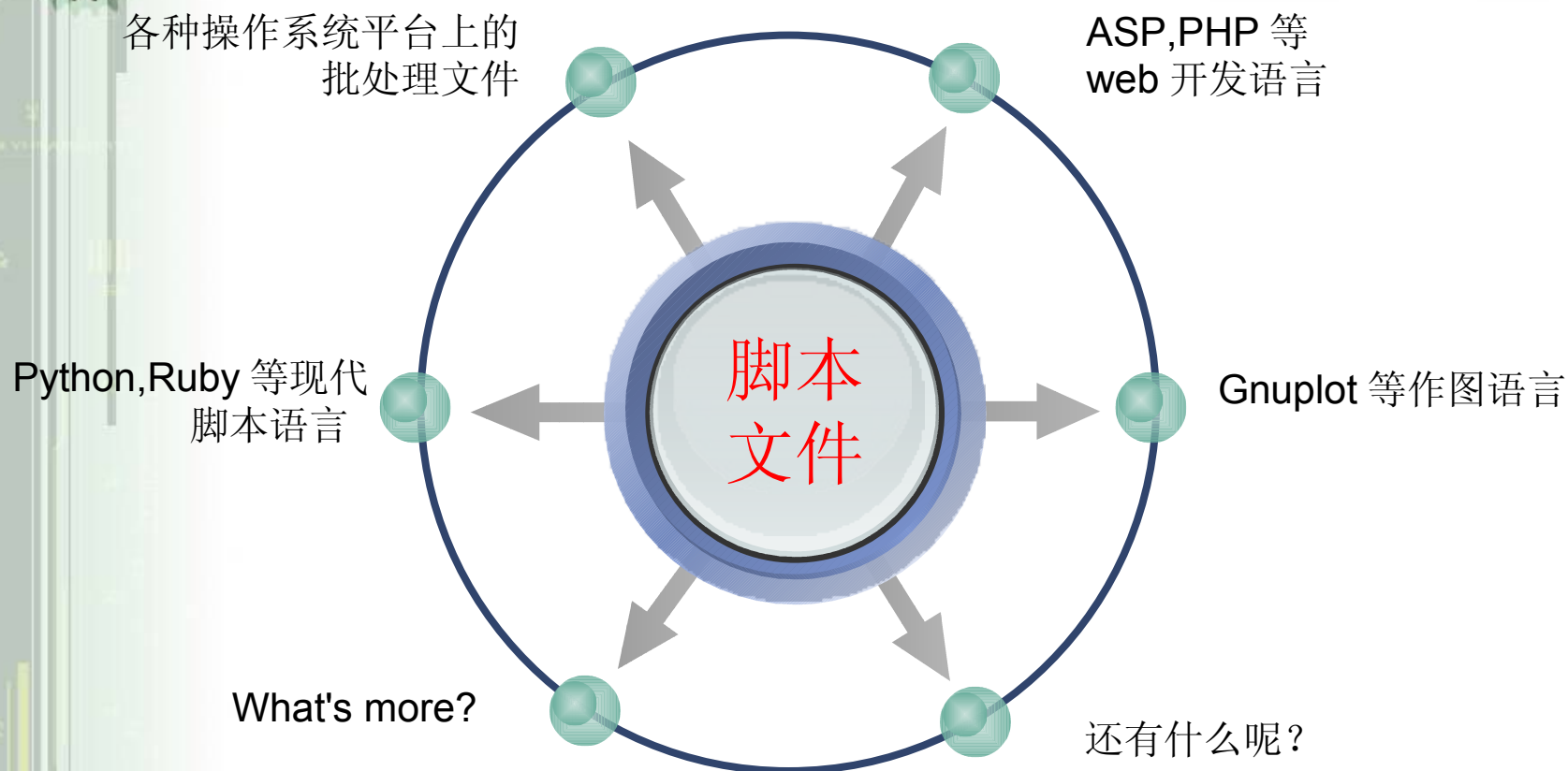


寻求解决问题的突破口

DSL³LAB



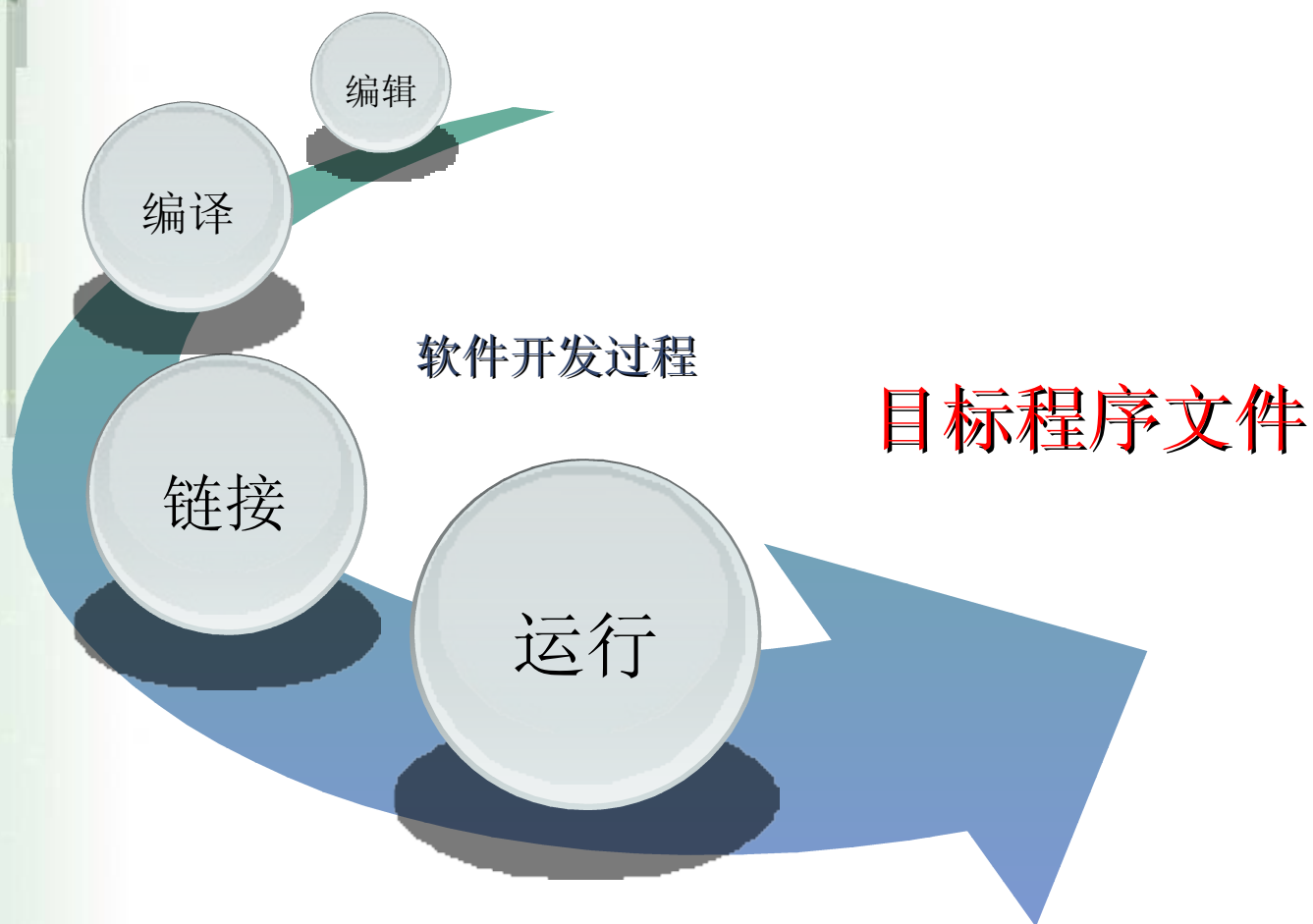
寻求解决问题的突破口



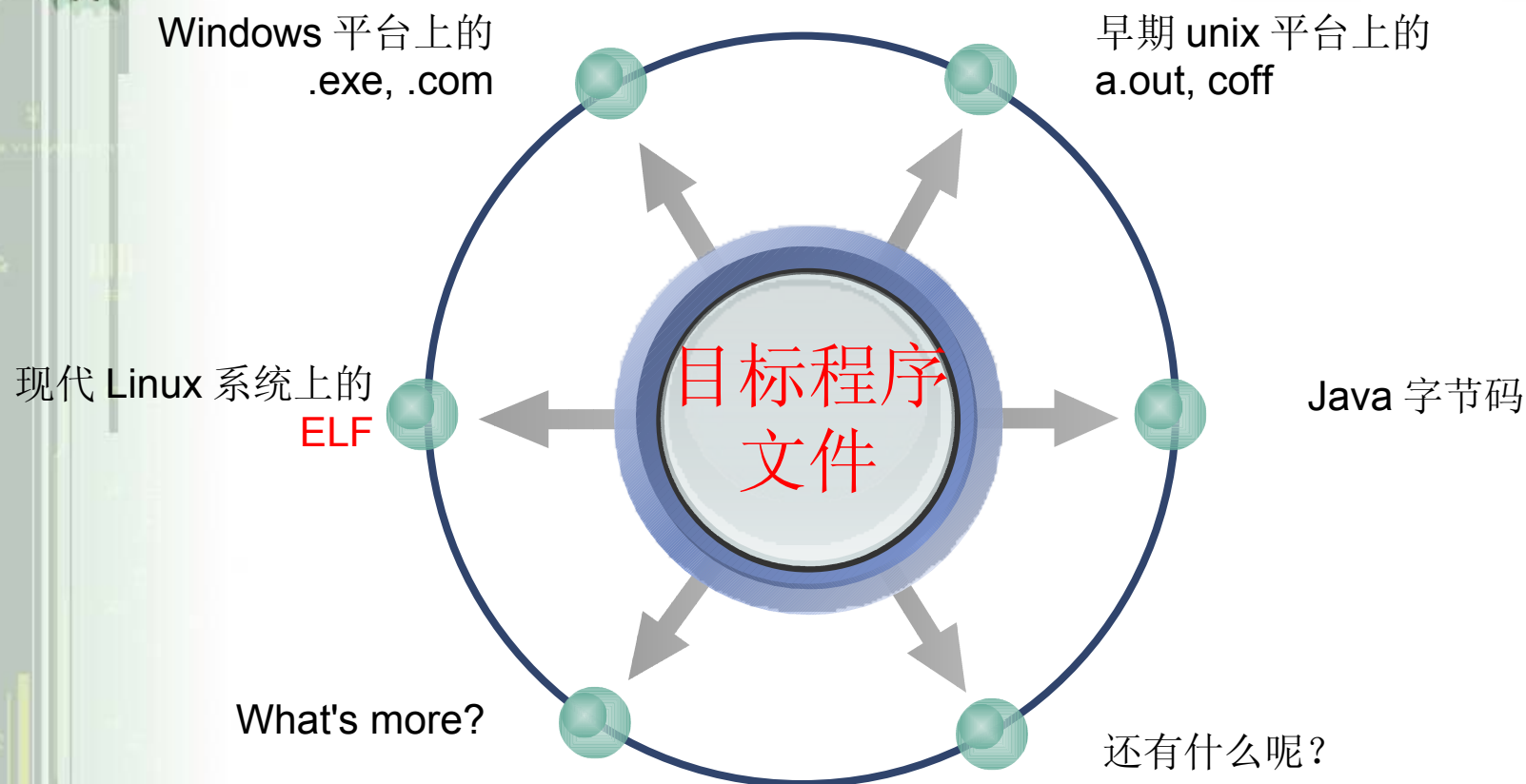


寻求解决问题的突破口

DSL³LAB




寻求解决问题的突破口





寻求解决问题的突破口

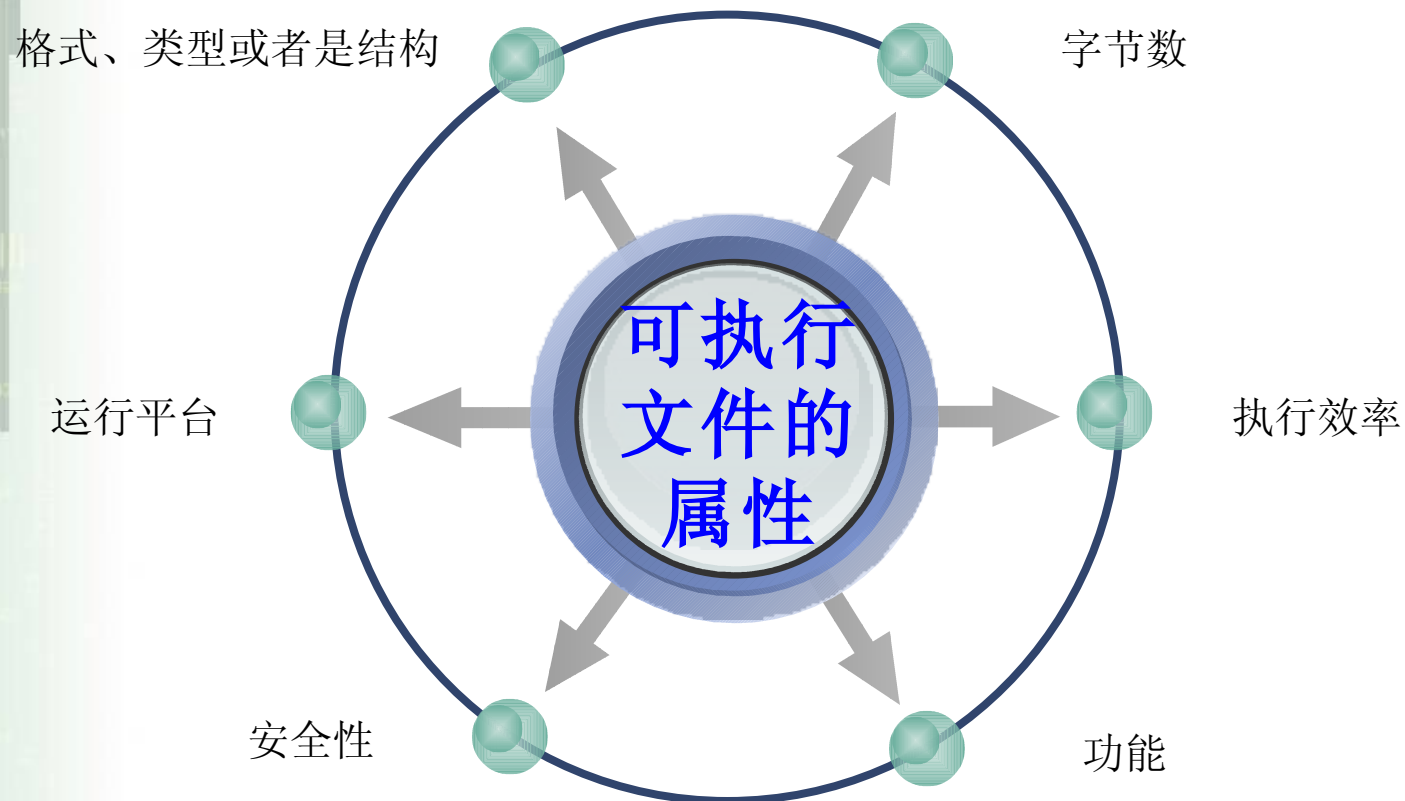
DSLΔB



可执行
文件

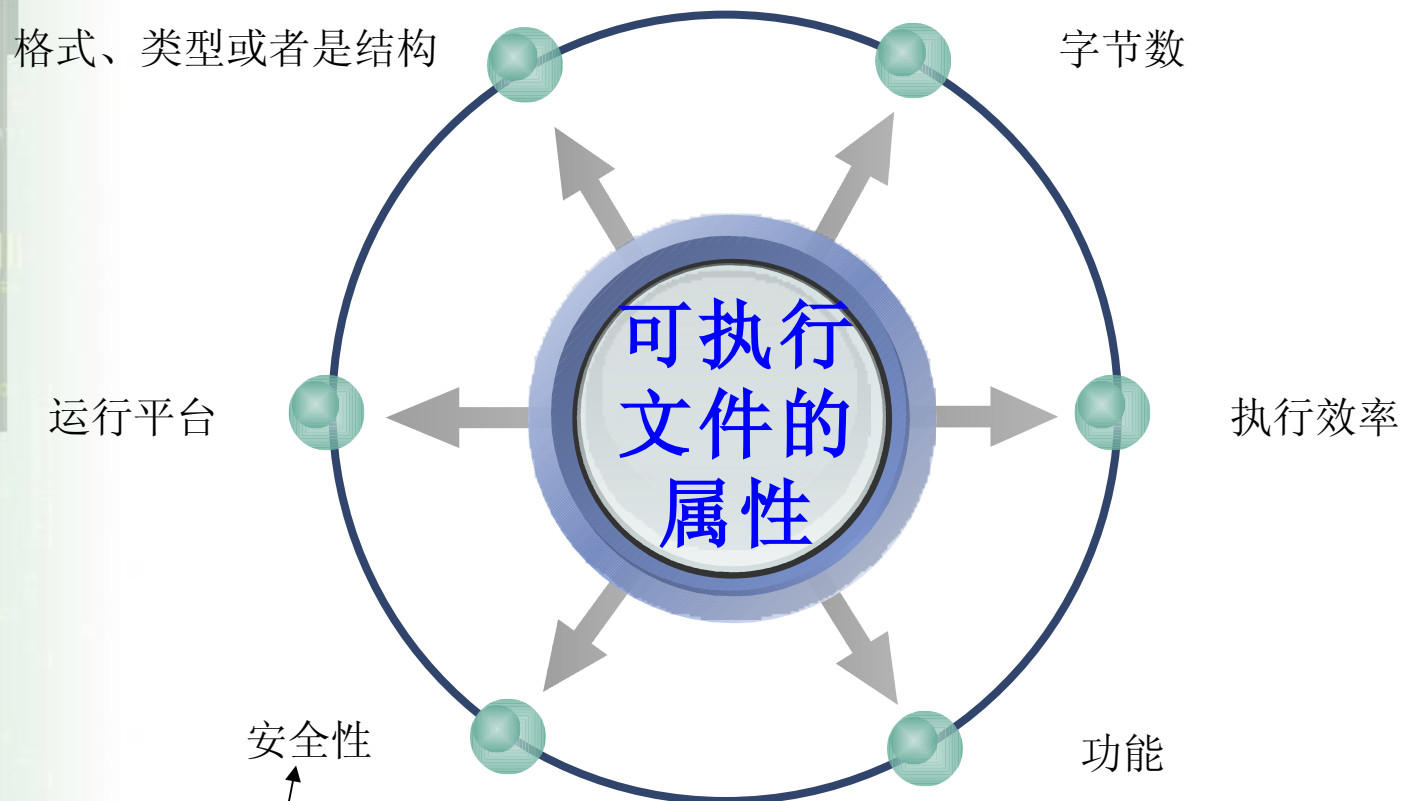
—— 软件的**最终载体和形式**

寻求解决问题的突破口





寻求解决问题的突破口



软件安全

安全性

寻求解决问题的突破口

格式、类型或者是结构

字节数

运行平台

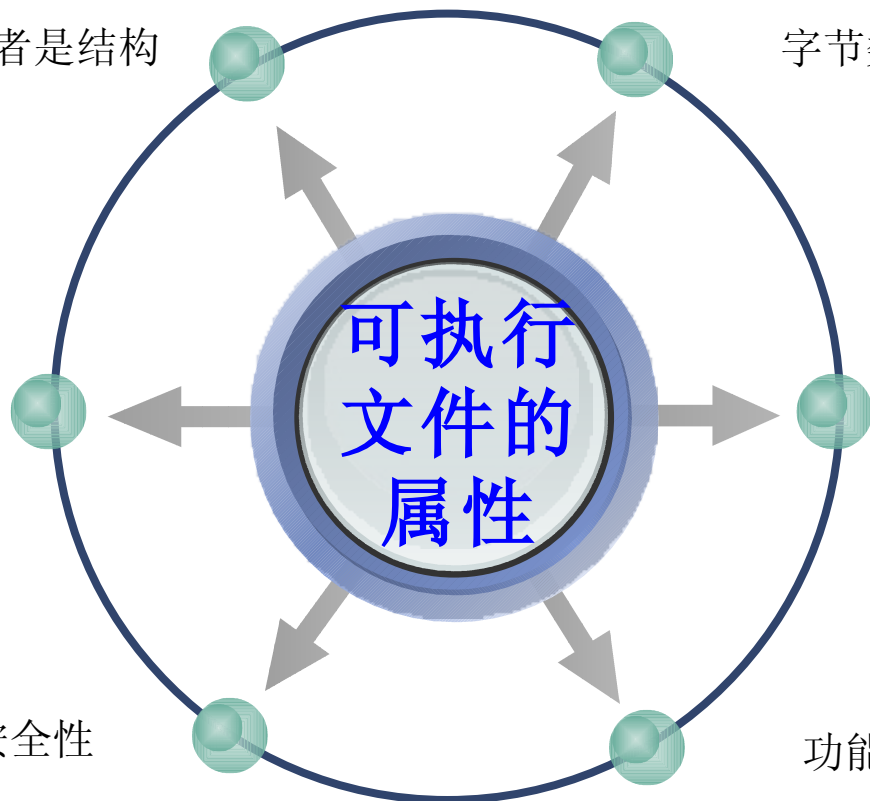
执行效率

安全性

功能

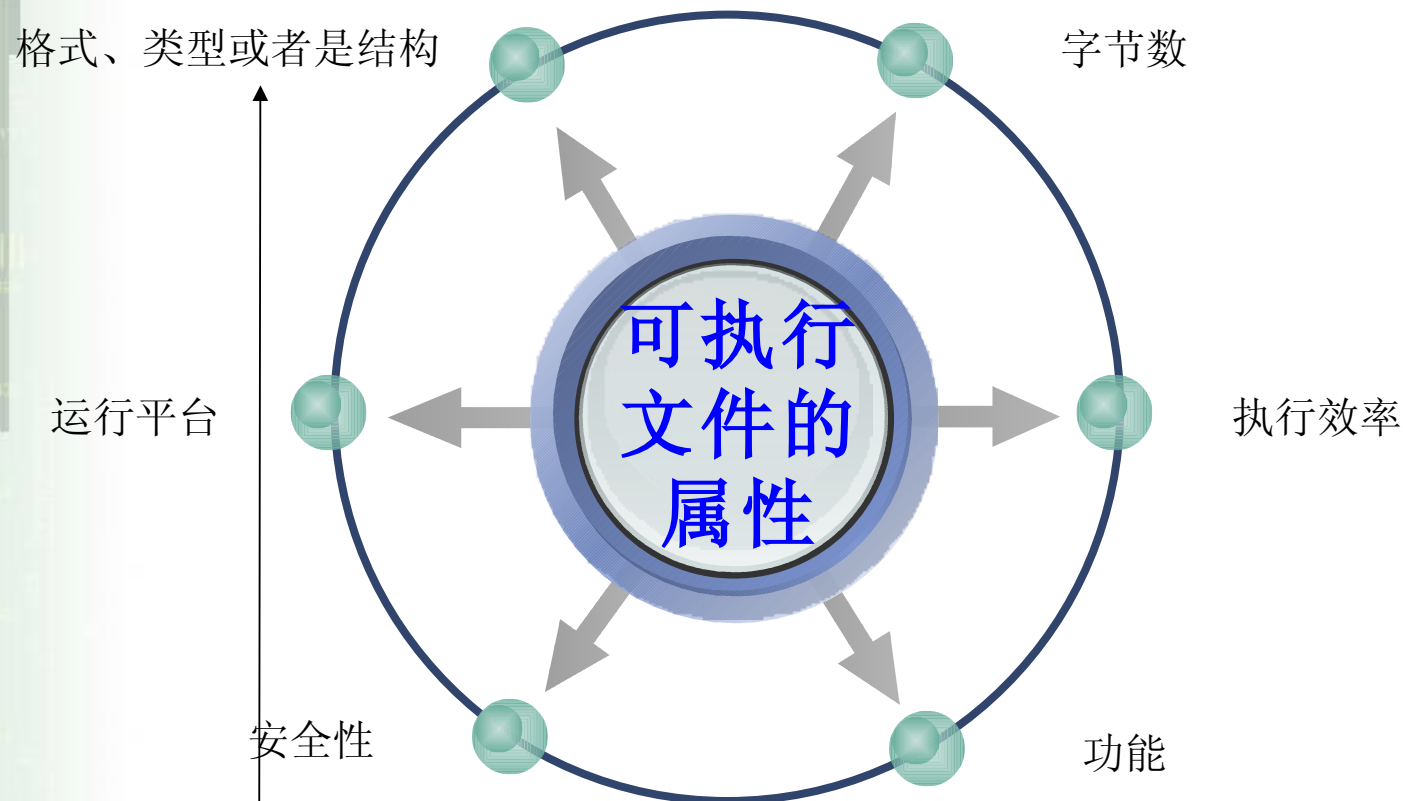
可执行
文件的
属性

软件移植性



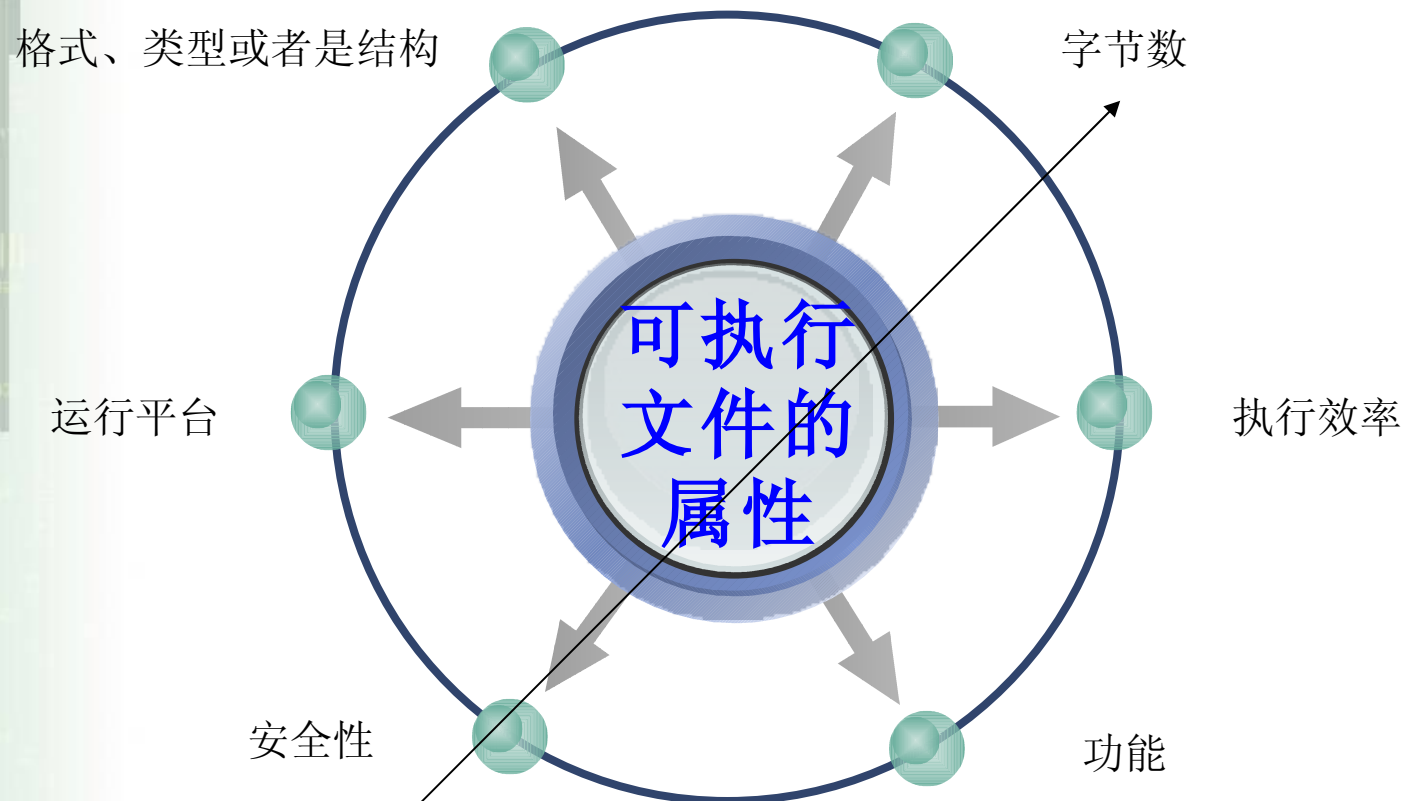


寻求解决问题的突破口



软件开发效率

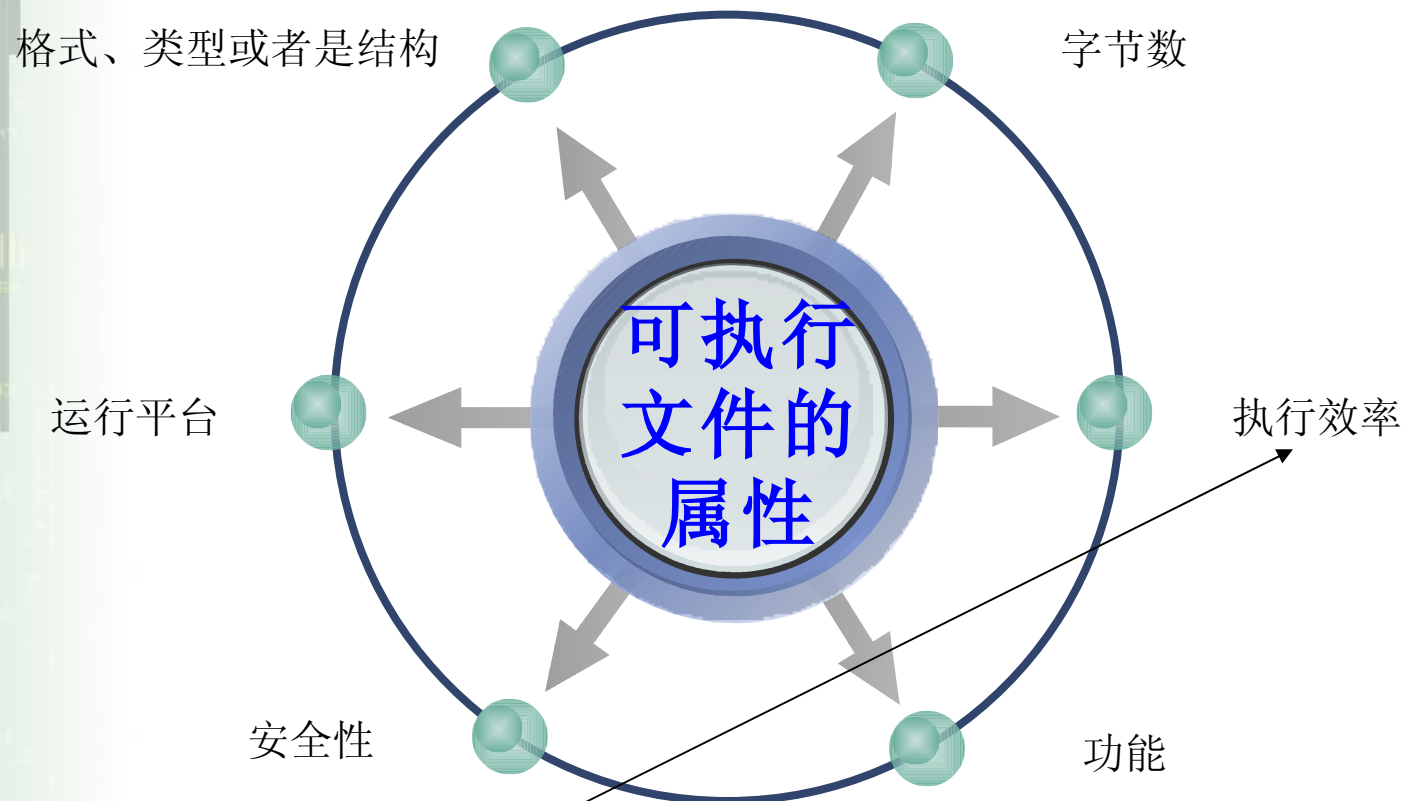
寻求解决问题的突破口



软件应用范围



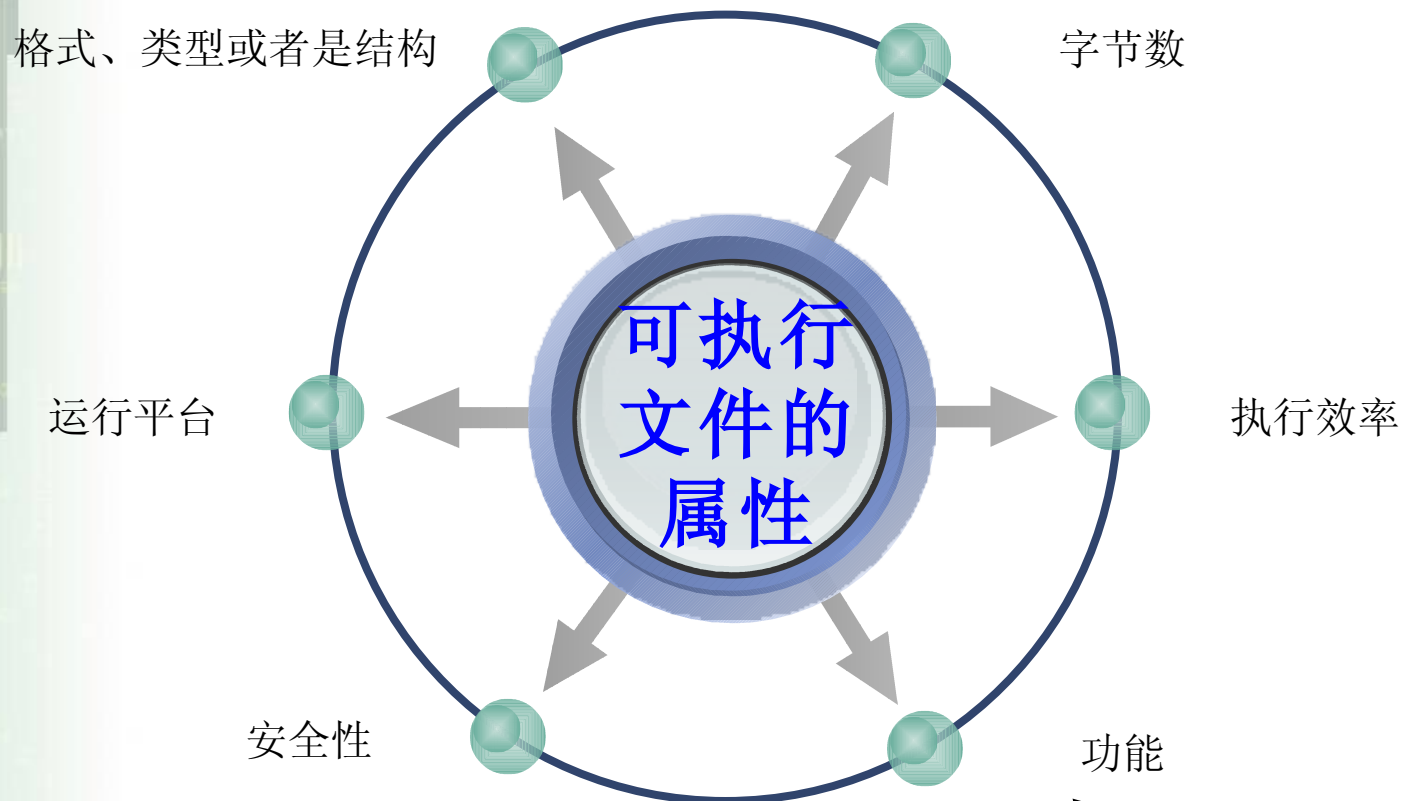
寻求解决问题的突破口



软件有效性



寻求解决问题的突破口



软件是否满足需求



寻求解决问题的突破口

DSLΔB

可执行
文件

—— 和软件领域所面临问题密切相关



寻求解决问题的突破口



—— 因此，可以考虑把“**可执行文件**”
作为**解决计算机软件领域所面临问题**
的**突破口**



目录

1

软件领域所面临的问题

2

寻求解决问题的突破口

3

引入数据挖掘技术的可行性

4

实例介绍



引入数据挖掘技术的可行性

1

有大量的相关数据和技术

2

可存储为哪些数据类型

3

可供挖掘的模式

4



有大量的相关数据

1

有大量的“可执行文件”实体以及相关的数据，比如各个操作系统平台上有为实现各种各样的功能而设计和开发的软件副本

纯数据层面

有大量的相关数据

1

有大量的“可执行文件”实体以及相关的数据，比如各个操作系统平台上有为实现各种各样的功能而设计和开发的软件副本

纯数据层面

2

积累了大量的软件相关技术，比如软件开发、软件设计、软件测试、软件维护、软件可移植性、软件安全等各方面的技术。

技术层面

有大量的相关数据

1

有大量的“可执行文件”实体以及相关的数据，比如各个操作系统平台上有为实现各种各样的功能而设计和开发的软件副本

纯数据层面

2

积累了大量的软件相关技术，比如软件开发、软件设计、软件测试、软件维护、软件可移植性、软件安全等各方面的技术。

技术层面

辅助数据挖掘过程



引入数据挖掘技术的可行性

1

有大量的相关数据和技术

2

可存储为哪些数据类型

3

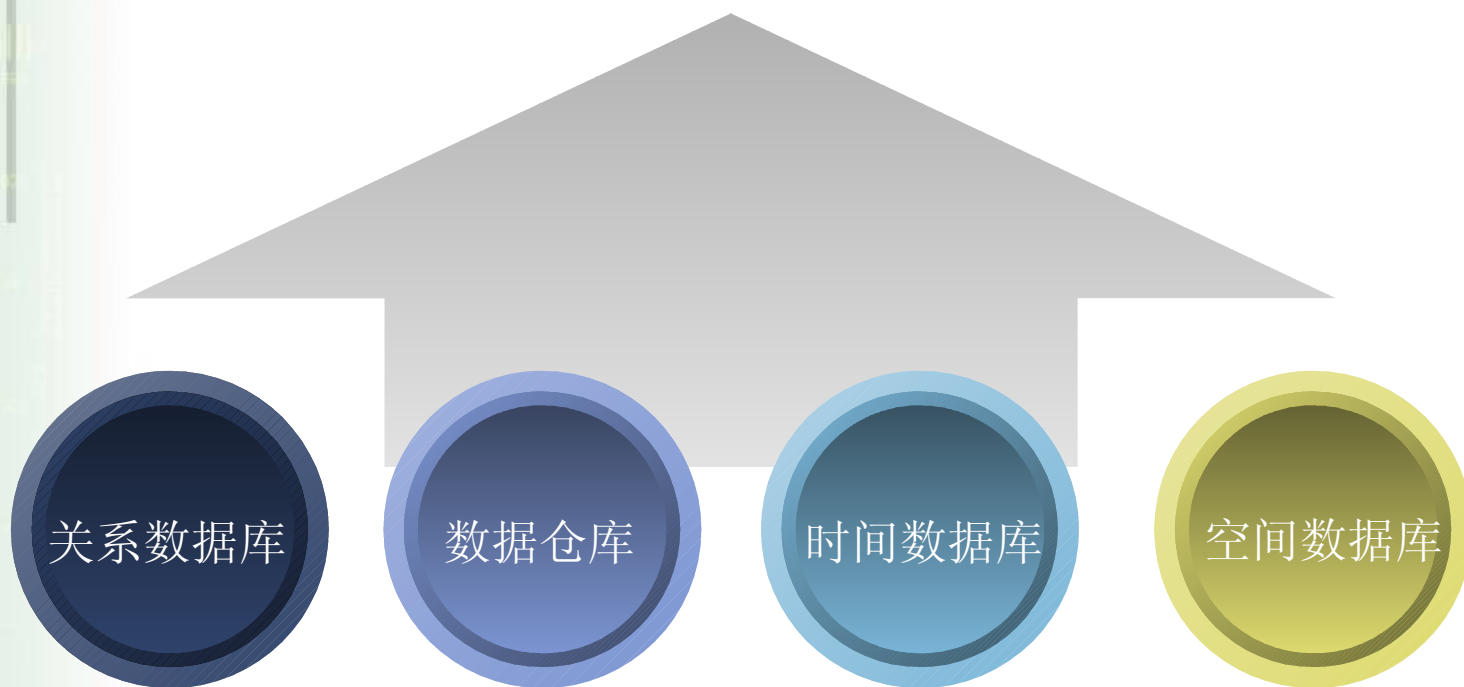
可供挖掘的模式

4



可存储为哪些数据类型

“可执行文件”的相关数据可存储为哪些数据类型





引入数据挖掘技术的可行性

1

有大量的相关数据和技术

2

可存储为哪些数据类型

3

可供挖掘的模式

4



特征化和区分

可供挖掘的模式

某个病毒特有的一些二进制串



特征化和区分



可供挖掘的模式

字节数					
格式					
平台					
安全					
效率					
功能					

频繁模式、关联和相关



可供挖掘的模式

	格式	平台	安全	效率	功能
字节数					
格式	X				
平台		X			
安全			X		
效率				X	
功能					X

频繁模式、关联和相关

可供挖掘的模式

DSL³LAB

2005 → 2006 → 2007 → **2008**

病毒一
技术一

病毒一
技术二

病毒三
技术四

病毒四
技术四

演变分析和预测

可供挖掘的模式

DSL³LAB

2005 → 2006 → 2007 → **2008** → **2009?**

病毒一
技术一

病毒二
技术二

病毒三
技术四

病毒四
技术四

演变分析和**预测**



目录

1

软件领域所面临的问题

2

寻求解决问题的突破口

3

引入数据挖掘技术的可行性

4

实例介绍



实例介绍

DSL³LAB

1

项目简介

2

项目架构

3

功能介绍

4

项目简介



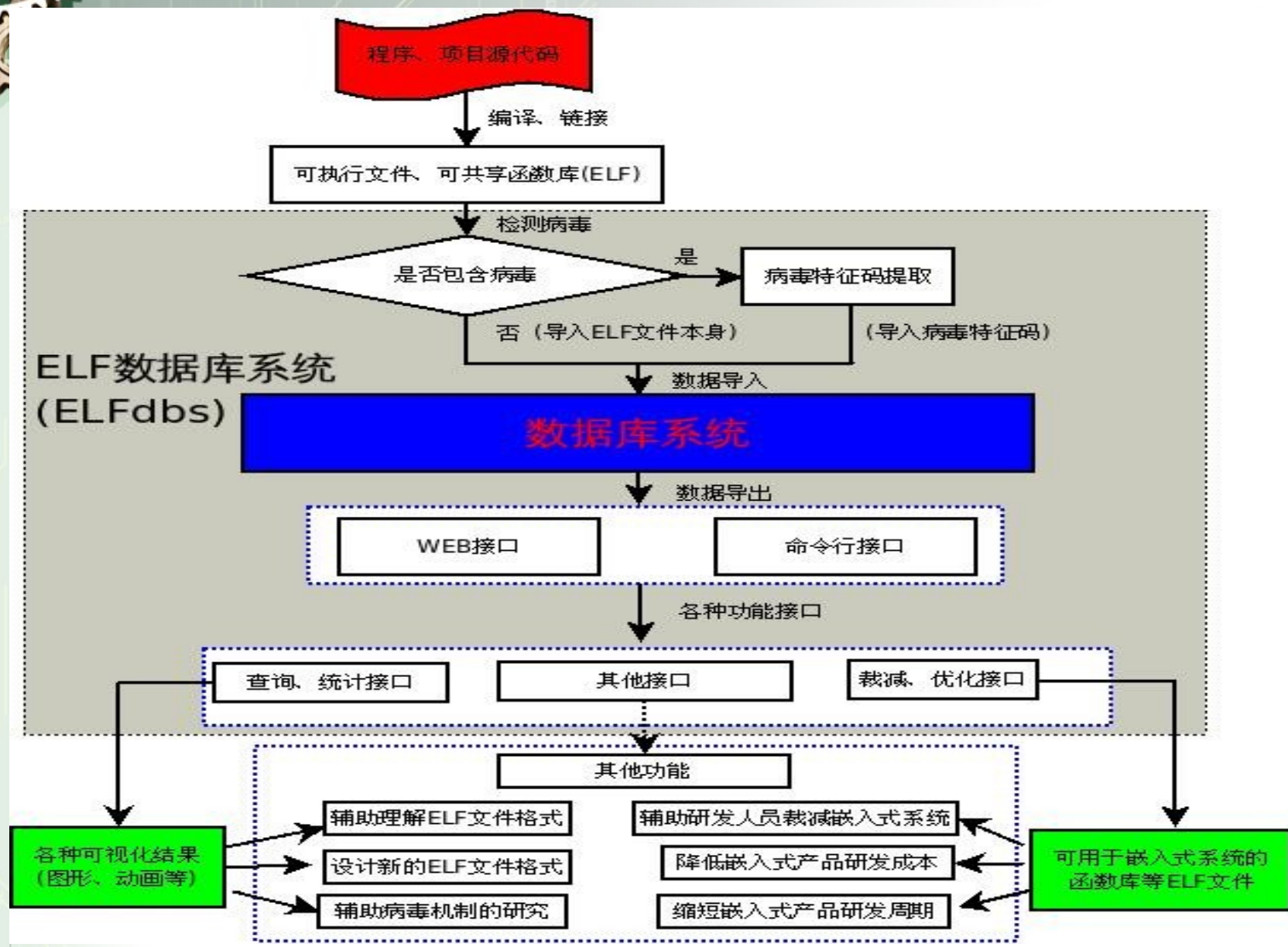
Unix 平台上的一种
可执行文件格式

关系型数据库

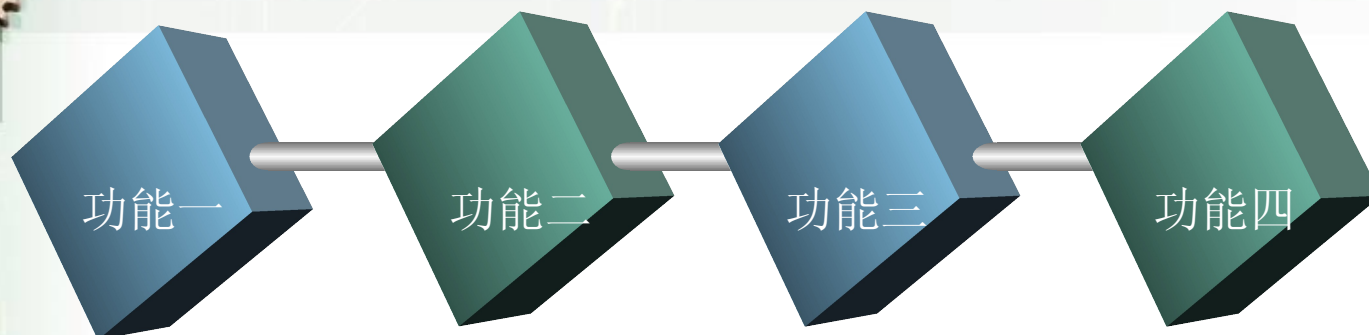
各种查询、统计、
分析、裁减接口

ELFdb : ELF Database System

项目架构



功能介绍



计算机病毒
技术研究平台

嵌入式系统
需要的函数库
的裁减平台

下一代可执行
文件格式的
研发平台

程序开发人员
的技术理论
的实践平台

Thank You !

吴章金 <wuzhangjin@gmail.com>