补充

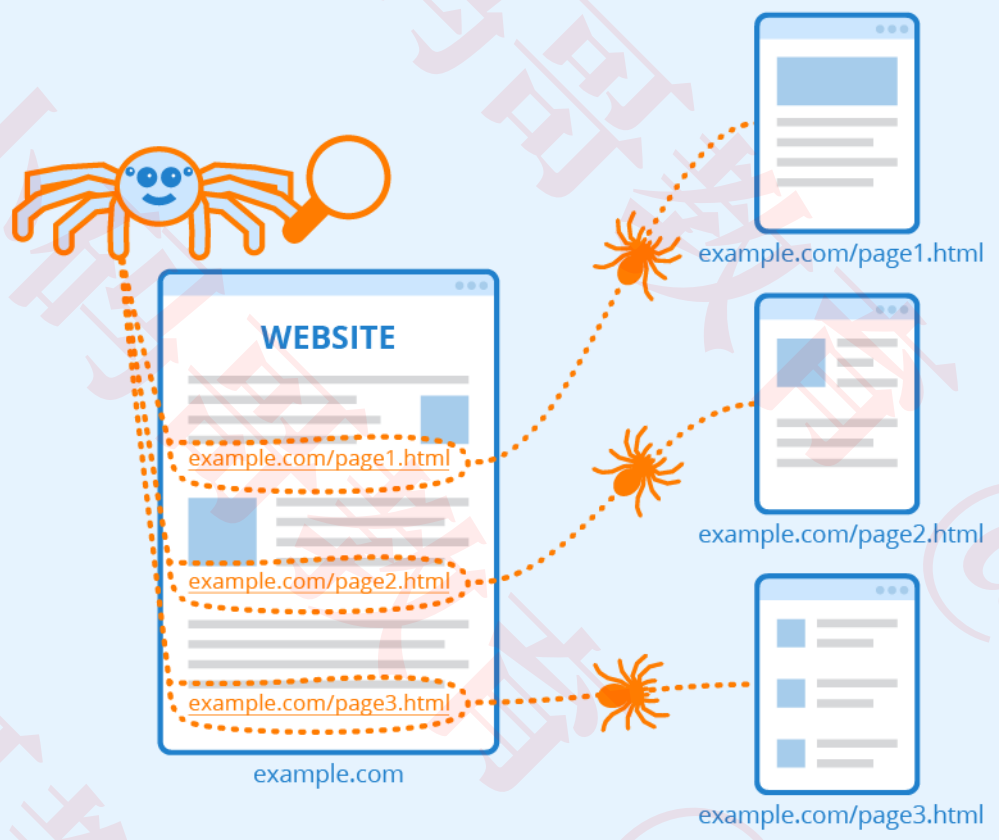
@M了个J
李明杰

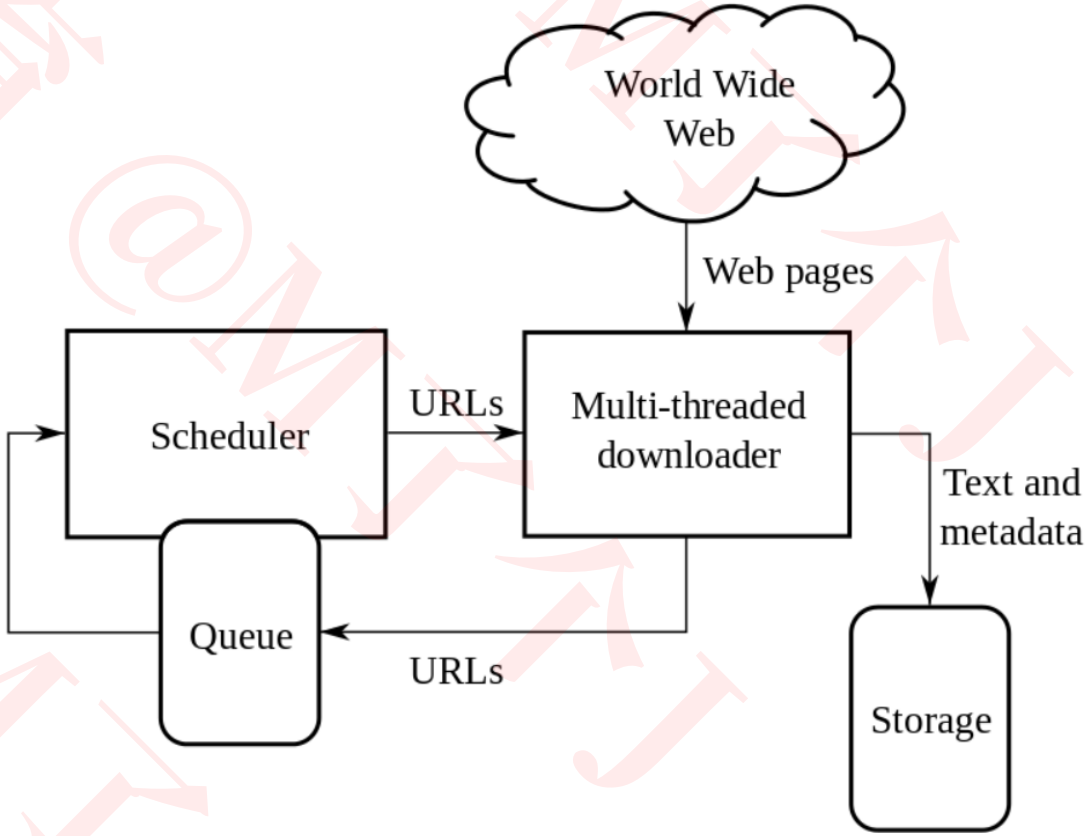
https://github.com/CoderMJLee

https://space.bilibili.com/325538782

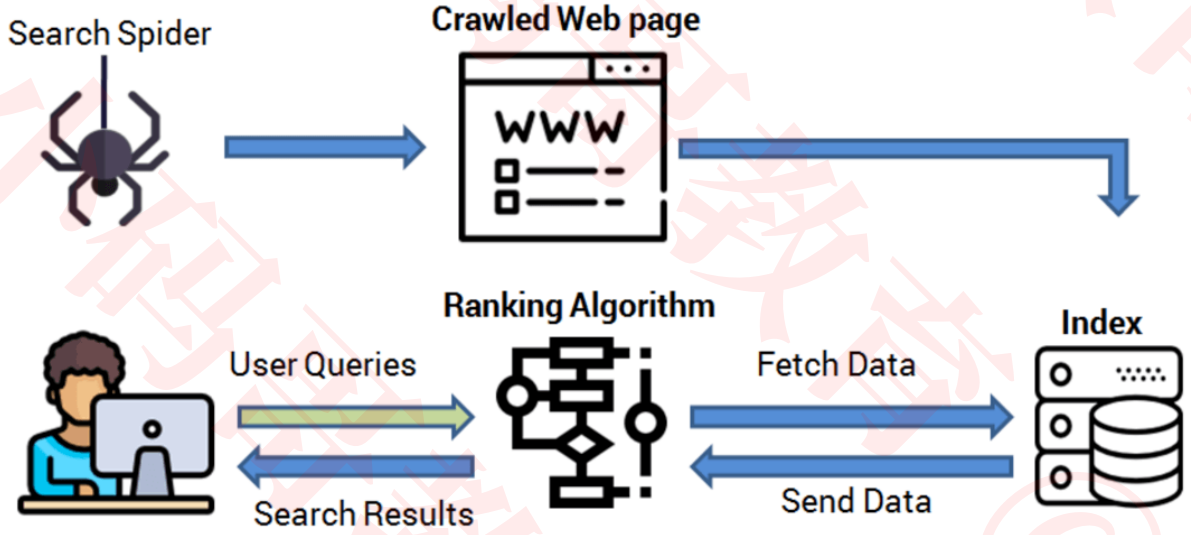
网络爬虫

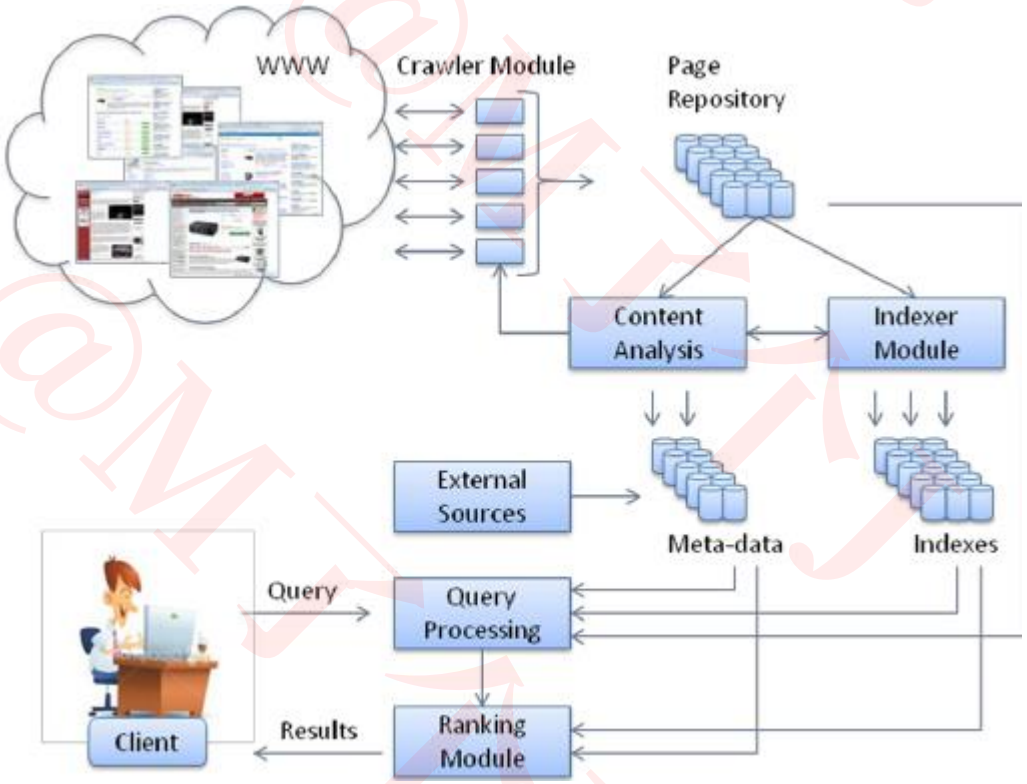
- 网络爬虫（Web Crawler），也叫做网络蜘蛛（Web Spider）
- 模拟人类使用浏览器操作页面的行为，对页面进行相关的操作
- 常用爬虫工具：Python的Scrapy框架





网络爬虫 – 搜索引擎


Search Spider
Crawled Web page
WWW
Ranking Algorithm
Index
User Queries
Fetch Data
Search Results
Send Data


WWW
Crawler Module
Page Repository
Content Analysis
Indexer Module
External Sources
Meta-data
Indexes
Query
Query Processing
Client
Results
Ranking Module

网络爬虫 – 简易实例

- 可以使用Java的一个小框架Jsoup爬一些简单的数据
- jar包
  - ✓ https://jsoup.org/packages/jsoup-1.13.1.jar
  - ✓ https://mirror.bit.edu.cn/apache//commons/io/binaries/commons-io-2.8.0-bin.zip
- 爬取目标：https://ext.se.360.cn/webstore/category

```
String url = "https://ext.se.360.cn/webstore/category";
Elements apps = Jsoup.connect(url).get().select(".appwrap");
for (Element app : apps) {
    String img = app.selectFirst("img").attr("src");
    String name = app.selectFirst("h3").text();
    String intro = app.selectFirst(".intro").text();
    System.out.println(name + "_" + intro + "_" + img);

    String filepath = "F:/imgs/" + name + ".jpg";
    FileUtils.copyURLToFile(new URL(img), new File(filepath));
}
```

网络爬虫 - robots.txt

- robots.txt是存放于网站根目录下的文本文件，比如https://www.baidu.com/robots.txt
  - 用来告诉爬虫：哪些内容是不应被爬取的，哪些是可以被爬取的
  - 因为一些系统中的URL是大小写敏感的，所以robots.txt的文件名应统一为小写

- 它并不是一个规范，而只是约定俗成的，所以并不能保证网站的隐私
  - 只能防君子，不能防小人
  - 无法阻止不讲“武德”的年轻爬虫爬取隐私信息

网络爬虫 - robots.txt

允许所有的爬虫：

```
User-agent: *
Disallow:
```

另一写法

```
User-agent: *
Allow:/
```

仅允许特定的爬虫：（name_spider是爬虫名字）

```
User-agent: name_spider
Allow:
```

拦截所有的爬虫：

```
User-agent: *
Disallow: /
```

仅禁止爬虫访问特定目录：

```
User-agent: name_spider
Disallow: /private/
```

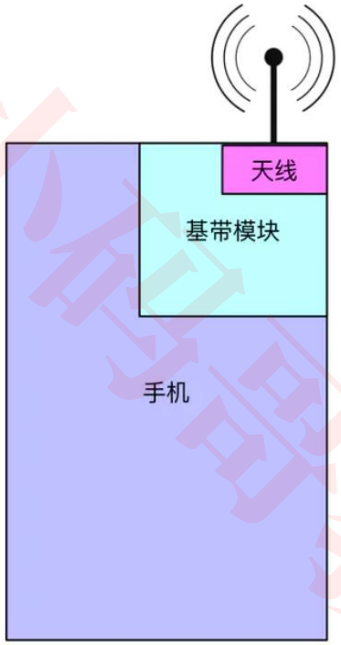
禁止所有爬虫访问特定目录：

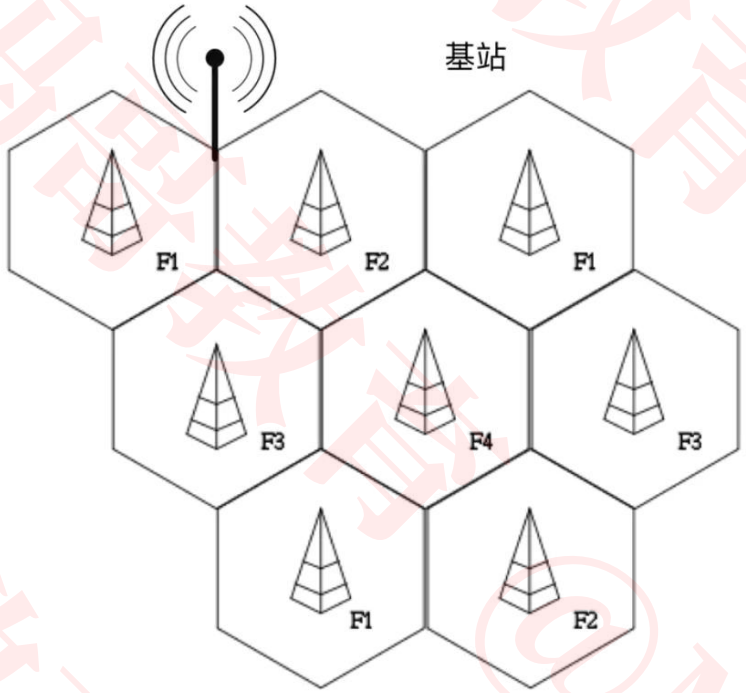
```
User-agent: *
Disallow: /cgi-bin/
Disallow: /images/
Disallow: /tmp/
Disallow: /private/
```

禁止所有爬虫访问特定文件类型

```
User-agent: *
Disallow: /*.php$
Disallow: /*.js$
Disallow: /*.inc$
Disallow: /*.css$
```

无线网络

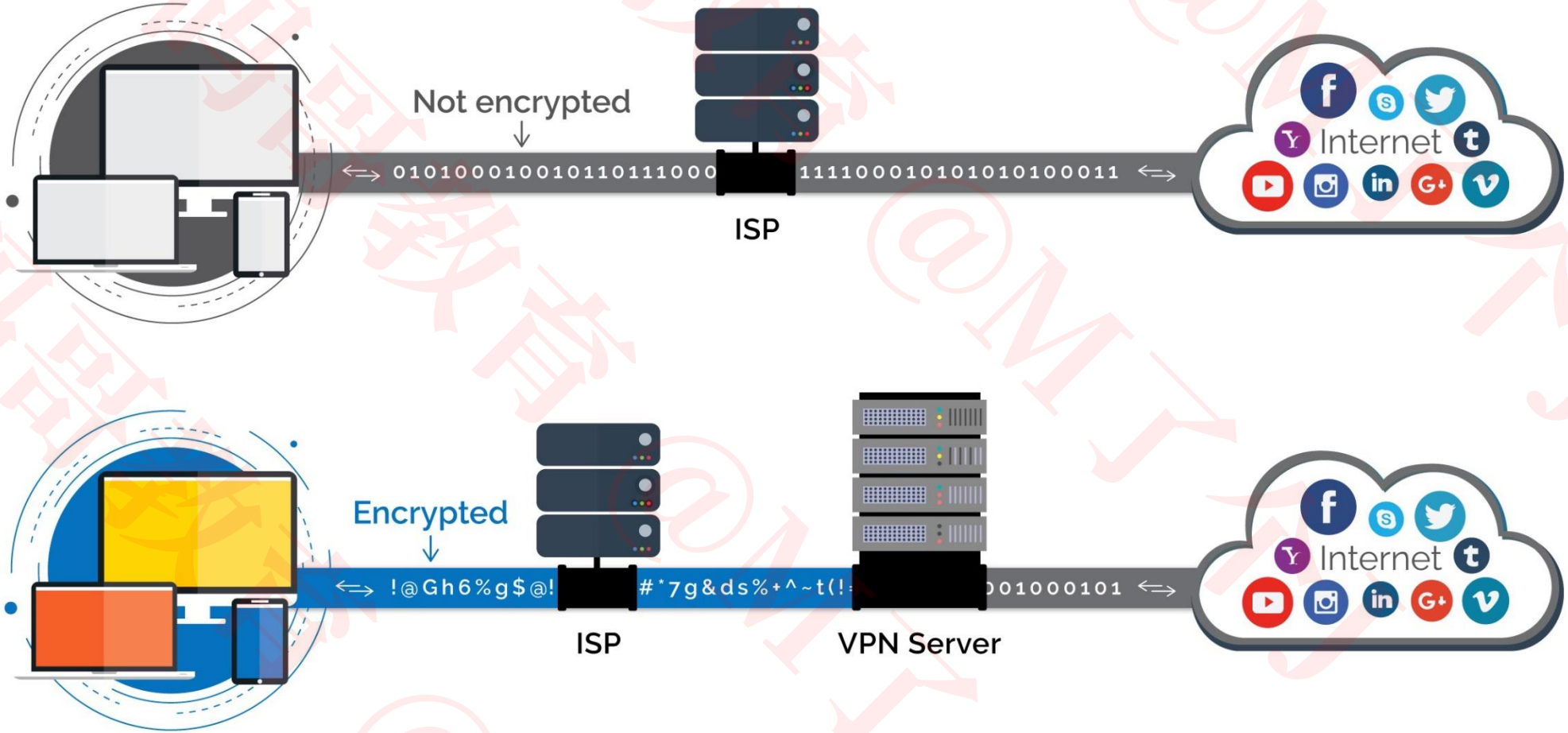




■ 无线AP（Access Point）：无线接入点

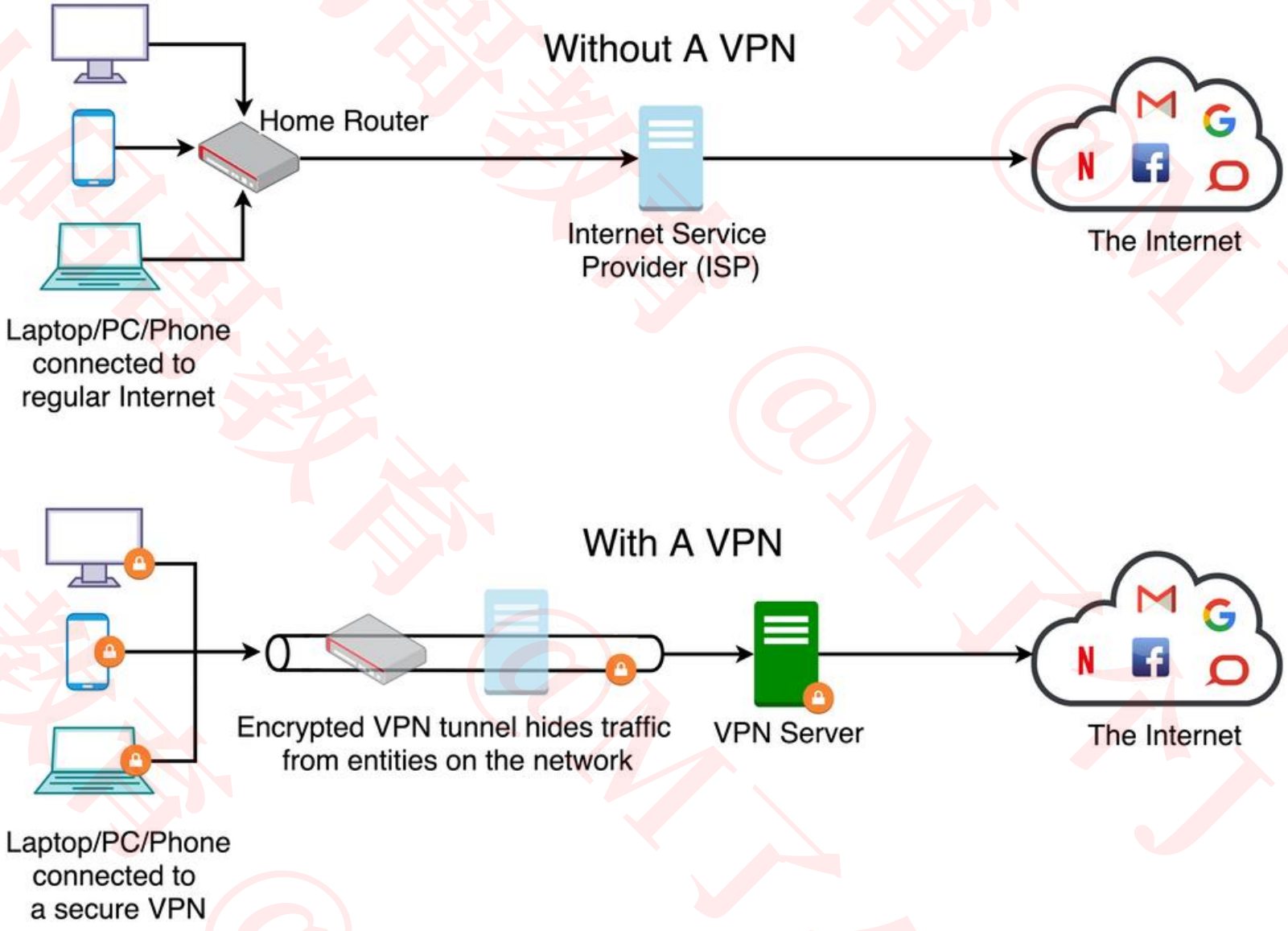
VPN

- VPN（**V**irtual **P**rivate **N**etwork），译为：虚拟私人网络
- 它可以在公共网络上建立专用网络，进行加密通讯



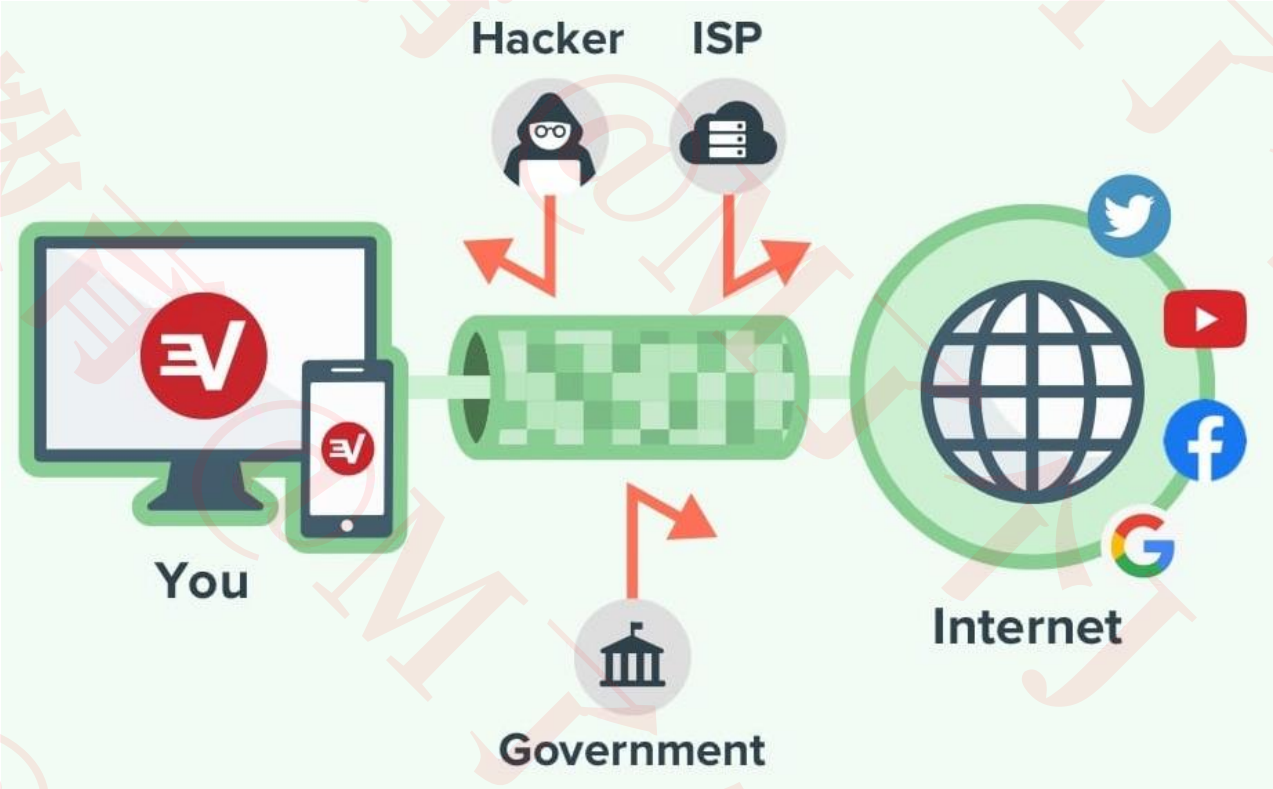
小码哥教育 SEEMYGO

VPN



VPN – 作用

- 提高上网的安全性
- 保护公司内部资料
- 隐藏上网者的身份
- 突破网站的地域限制
  - 有些网站针对不同地区的用户展示不同的内容
- 突破网络封锁
  - 因为有GFW的限制，有些网站在国内上不了
  - Great Firewall of China
  - 中国长城防火墙



VPN – 与代理的区别

- 软件
  - VPN一般需要安装VPN客户端软件
  - 代理不需要安装额外的软件

- 安全性
  - VPN默认会对数据进行加密
  - 代理默认不会对数据进行加密（数据最终是否加密取决于使用的协议本身）

- 费用
  - 一般情况下，VPN比代理贵

VPN – 实现原理

- VPN的实现原理是：使用了隧道协议（Tunneling Protocol）

- 常见的VPN隧道协议有
  - PPTP（**P**oint to **P**oint **T**unneling **P**rotocol）：点对点隧道协议
  - L2TP（**L**ayer **Two** **T**unneling **P**rotocol）：第二层隧道协议
  - IPsec（**I**nternet **P**rotocol **Sec**urity）：互联网安全协议
  - SSL VPN（如OpenVPN）
  - ……

tcpdump

- tcpdump是Linux平台的抓包分析工具，Windows版本是WinDump

- 使用手册
  - https://www.tcpdump.org/manpages/tcpdump.1.html

- 不错的教程
  - https://danielmiessler.com/study/tcpdump/