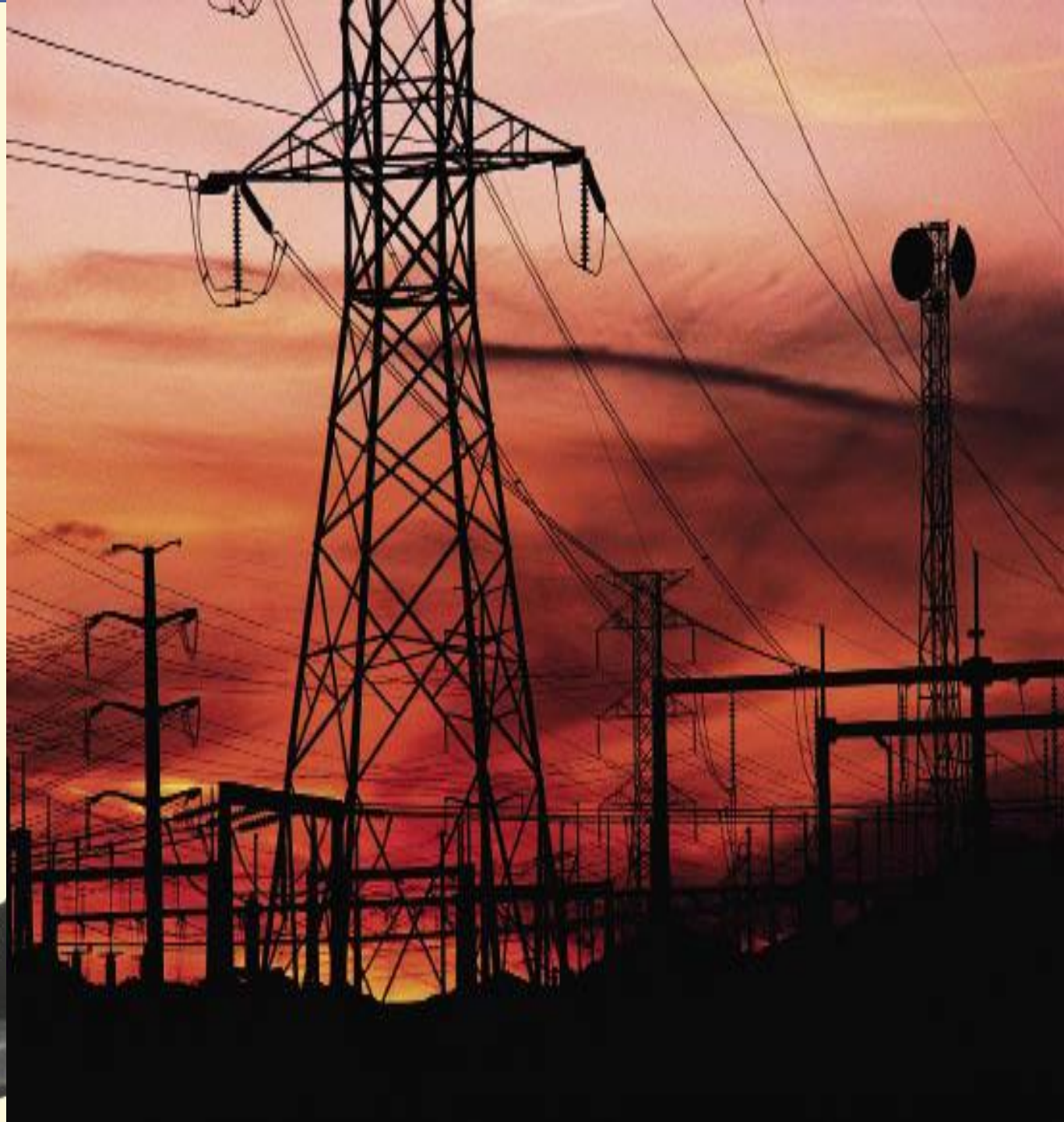
信息科学与技术学院

走近信息科学

京师信科



专题讲座2

从互联网到大数据

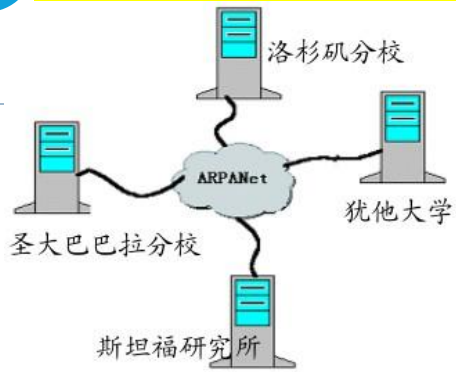
——追寻信息文明轨迹，感受信息技术价值

郭俊奇
北京师范大学　信息科学与技术学院

2015-09-25

1 原始信息交互

2 互联网前身：阿帕网



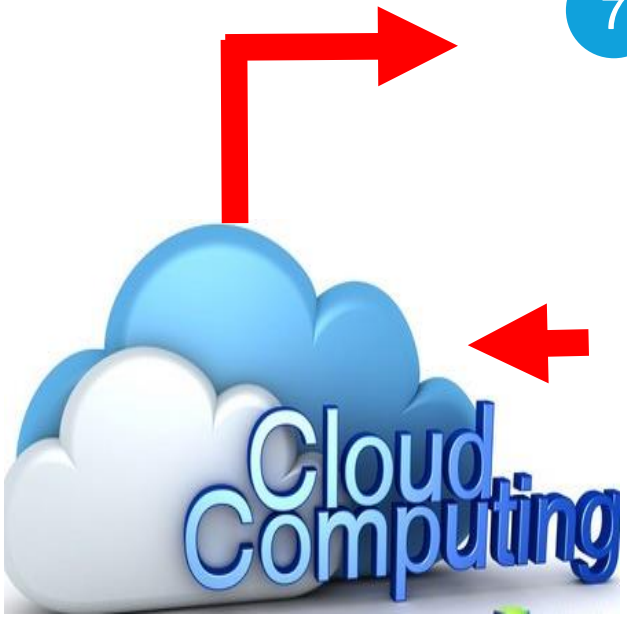
3 现代互联网

7 互联网+

未来形态：互联网+

4 物联网

5 大数据

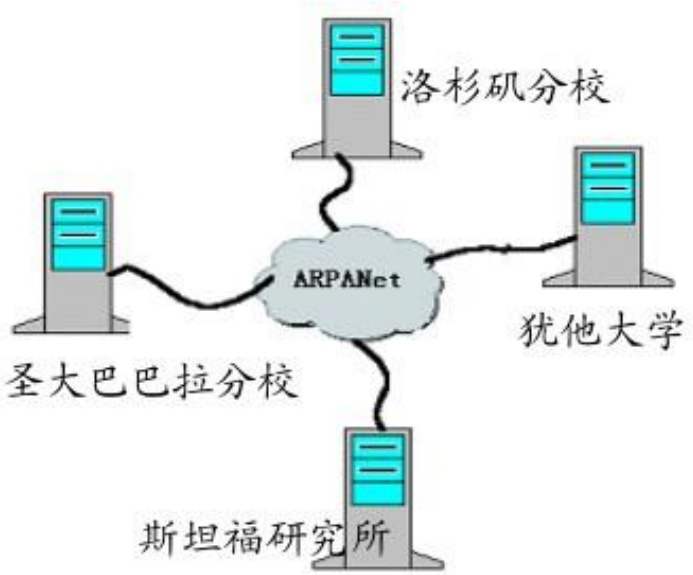


6 云计算

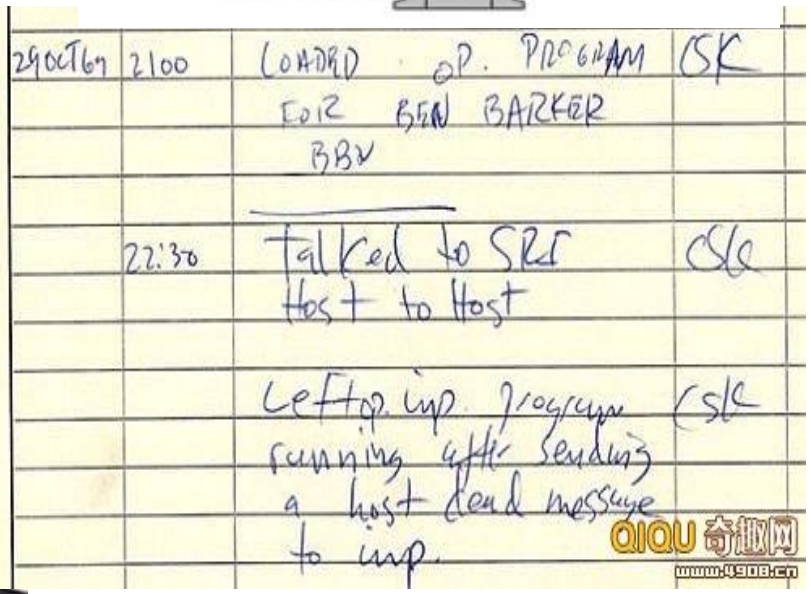
1 原始信息交互

信息传输方式	声波在空气中传播
信息处理方式	大脑控制听说反应
信息交互规则	相同语言

2 互联网前身：阿帕网（1969年）



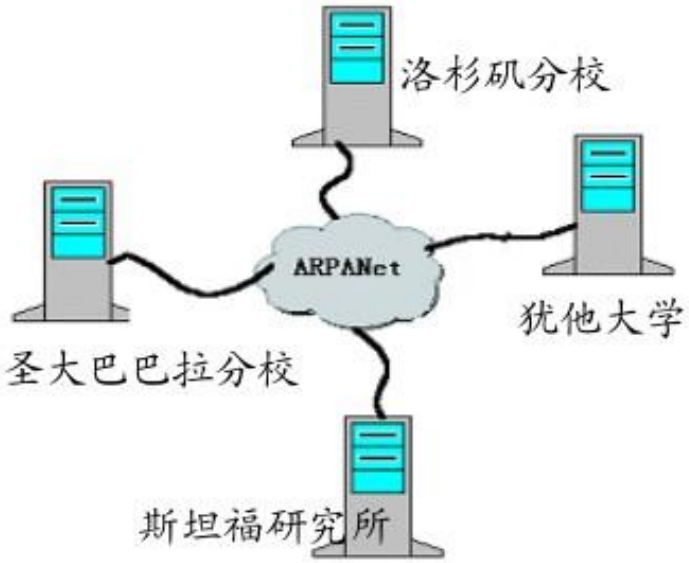
1969年，ARPANET创建者



互联网的第一个日志文件

2

互联网前身：阿帕网（1969年）



信息传输方式	有线：同轴电
信息处理方式	模拟调制解调
信息交互规则	网络控制程序（N

电磁理论课程

高频电路课程

编程类课程

3 现代互联网（1991年~至今）

➢现代互联网：

- 即Internet，是指诸多固定计算机和各类智能移动终端以有线或无线的方式，
- 依托TCP/IP协议互连形成逻辑上单一且巨大的全球化网络，
- 现代互联网可实现信息共享，它是信息社会的基础。

3 现代互联网（1991年~至今）

信息传输方式	有线、无
信息处理方式	数字信 理
信息交互规则	TCP/IP协

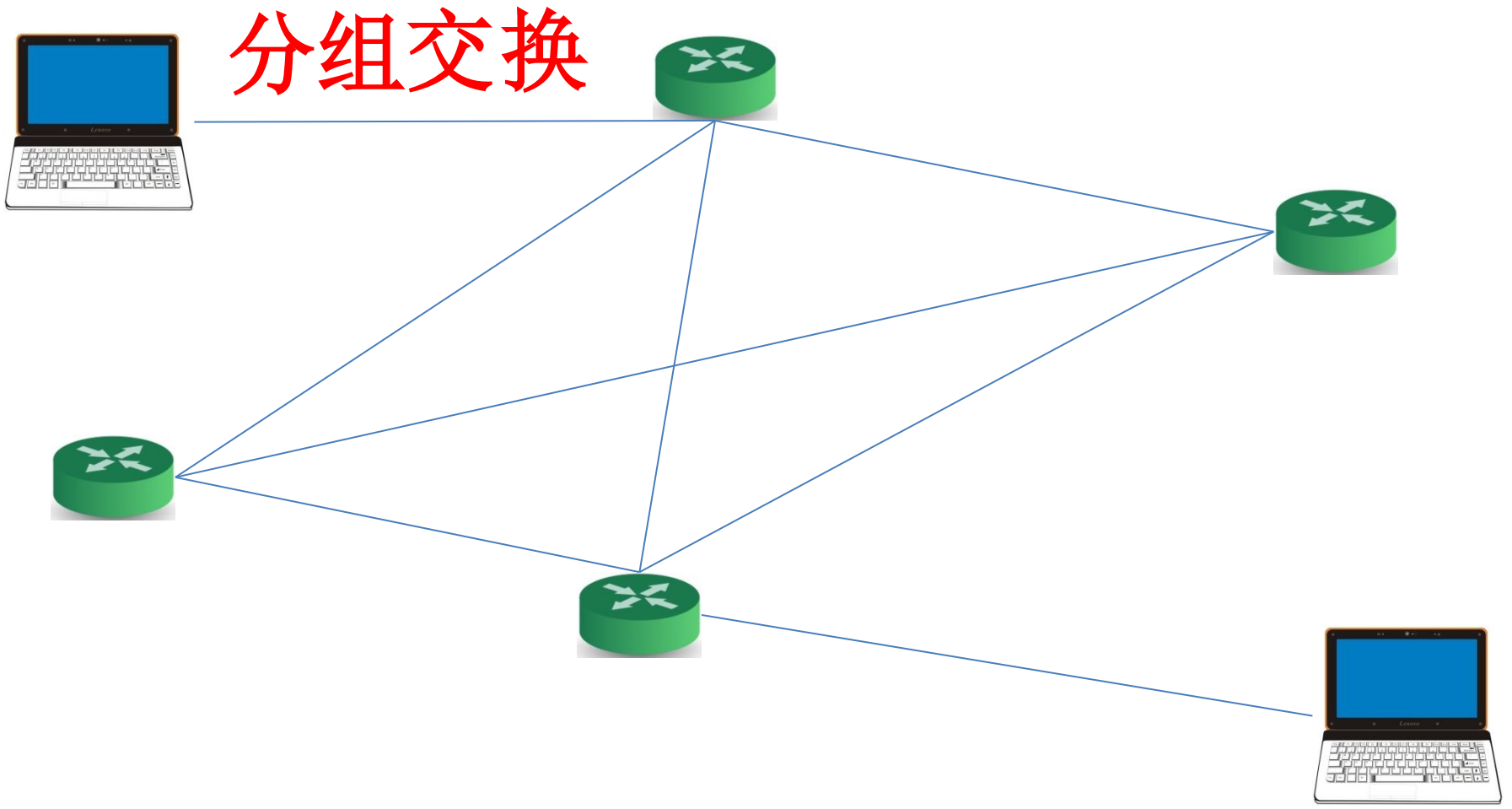
移动通信课程

数字信号处理课程

计算机网络课程

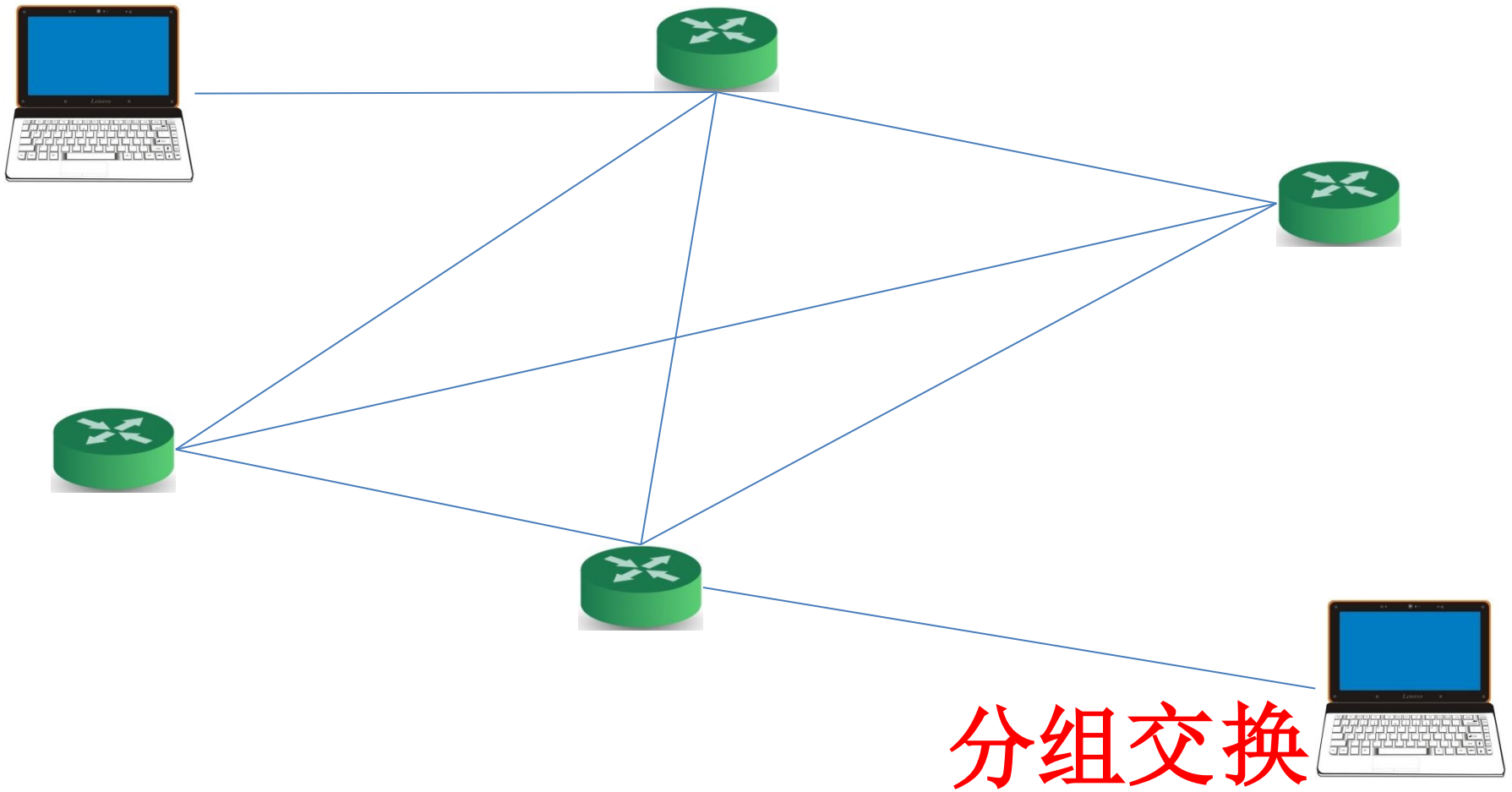
➢TCP/IP：传输控制协议/网际协议

(Transmission Control Protocol / Internet Protocol，TCP/IP)



➢TCP/IP：传输控制协议/网际协议

(Transmission Control Protocol / Internet Protocol，TCP/IP)



4 物联网（2009年~至今）

1995 Bill Gates《未来之路》首次提及物联网

1999 “物联网”定义---EPC global: 把全世界所有物品通过信息传感设备与互联网连接起来，实现智能化管理

2008 IBM、奥巴马“智慧地球”

2009 温家宝“感知中国”

4

物联网
（2009年~至今）

信息传输方式	蓝牙、W…
信息处理方式	射频识…
信息交互规则	Zigbee协…

无线通信课程

嵌入式系统课程

组网技术课程

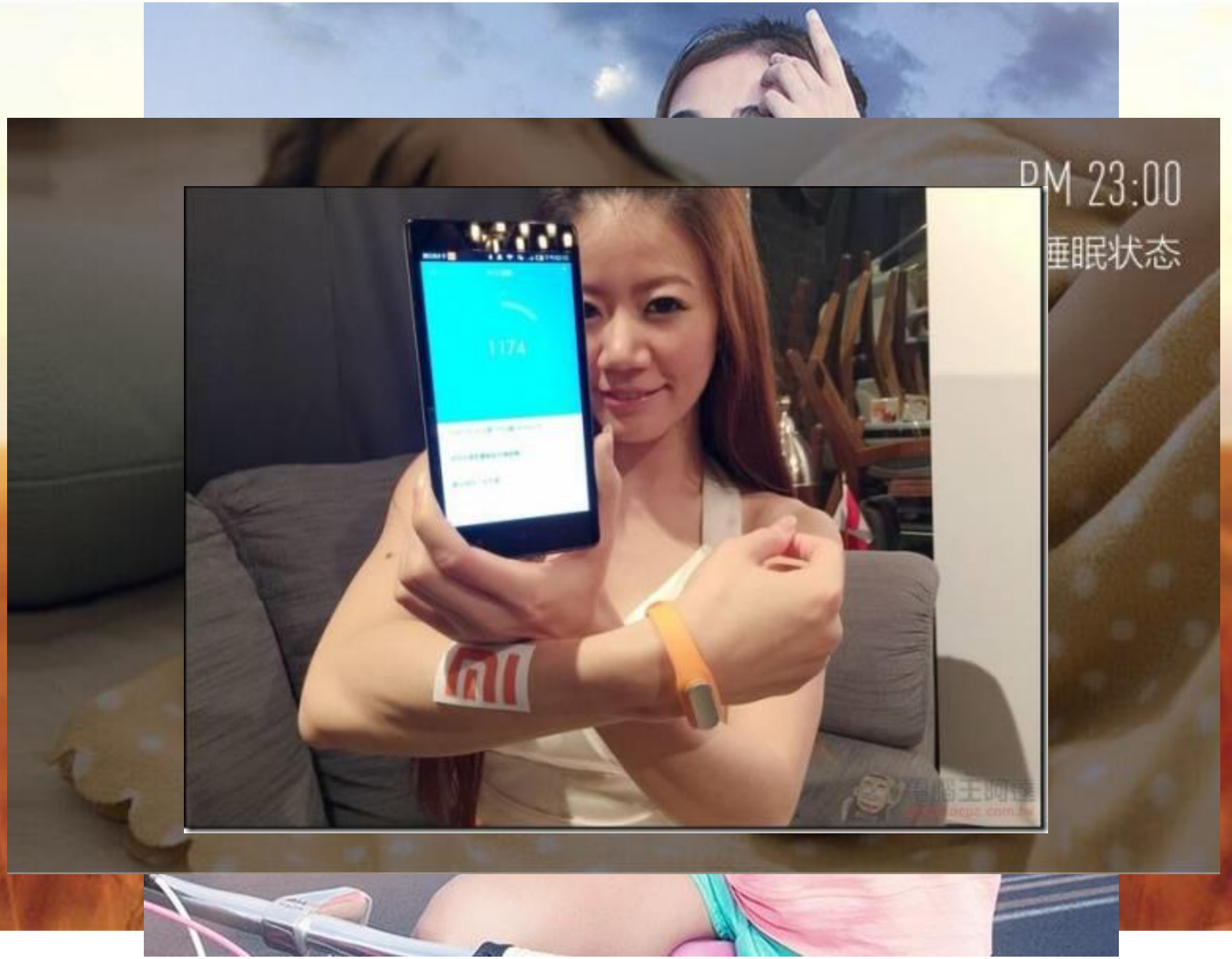
重在“传感”（sensor）

感

感受到物理世界：热/光/气/力/湿敏传感器

传

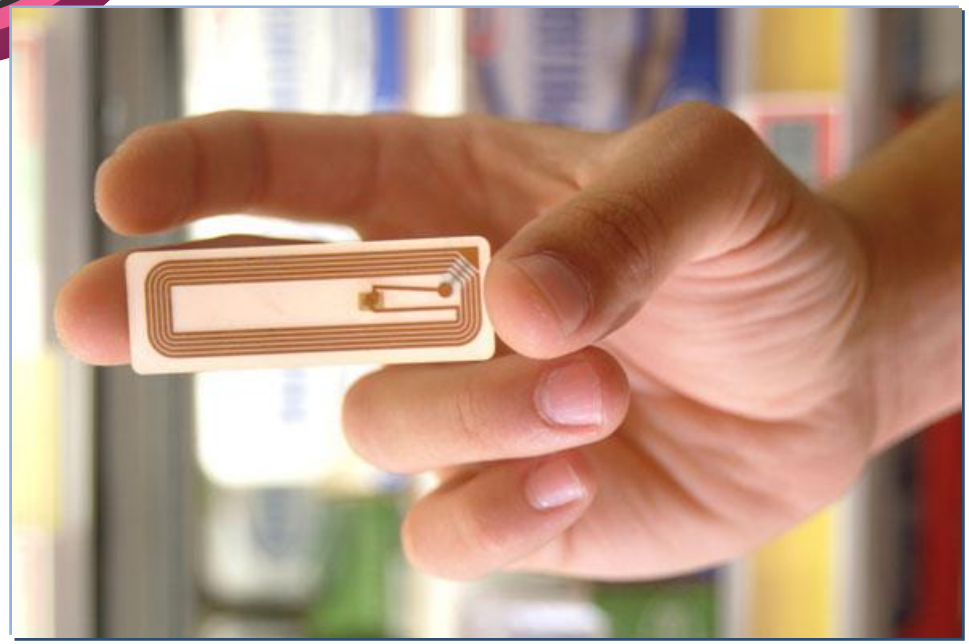
以数据读写形式传达信息：射频识别（RFID）


PM 23:00
睡眠状态

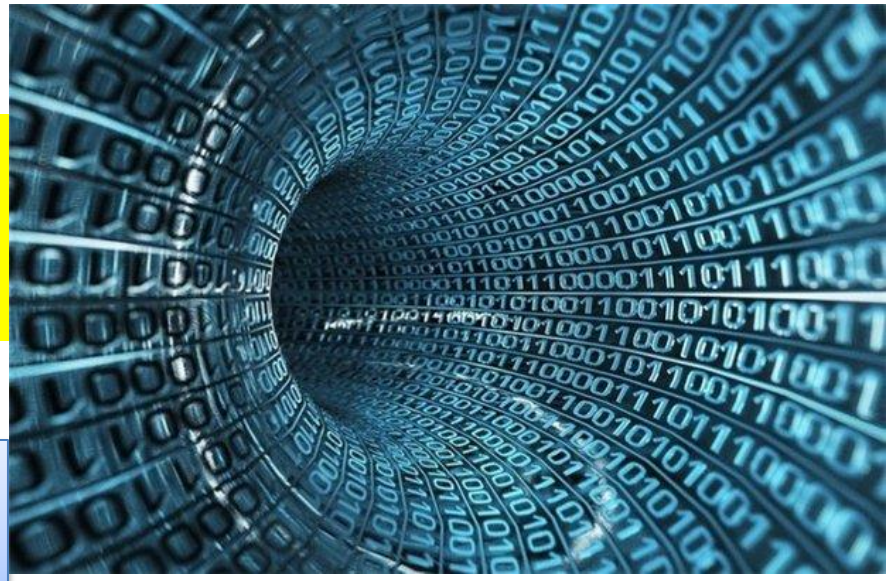
重在“传感”（sensor）

传

以数据读写形式传达信息：射频识别（RFID）



5 大数据（约2010年~至今）



大量Volume	TB量级以上
多样Variety	非结构化数据
价值Value	蕴藏有用信息
速度Velocity	实时分析处理

4V特征

1B = 8 bit

1KB = 1024 Bytes ≈ byte

1MB = 1024 KB ≈ byte

1GB = 1024 MB ≈ byte

1TB = 1024 GB ≈ byte

1PB = 1

1EB = 1 ≈ byte = 1 000 000 000 000 000 000 byte

1ZB = 1 ≈ byte = 1 000 000 000 000 000 000 000 byte

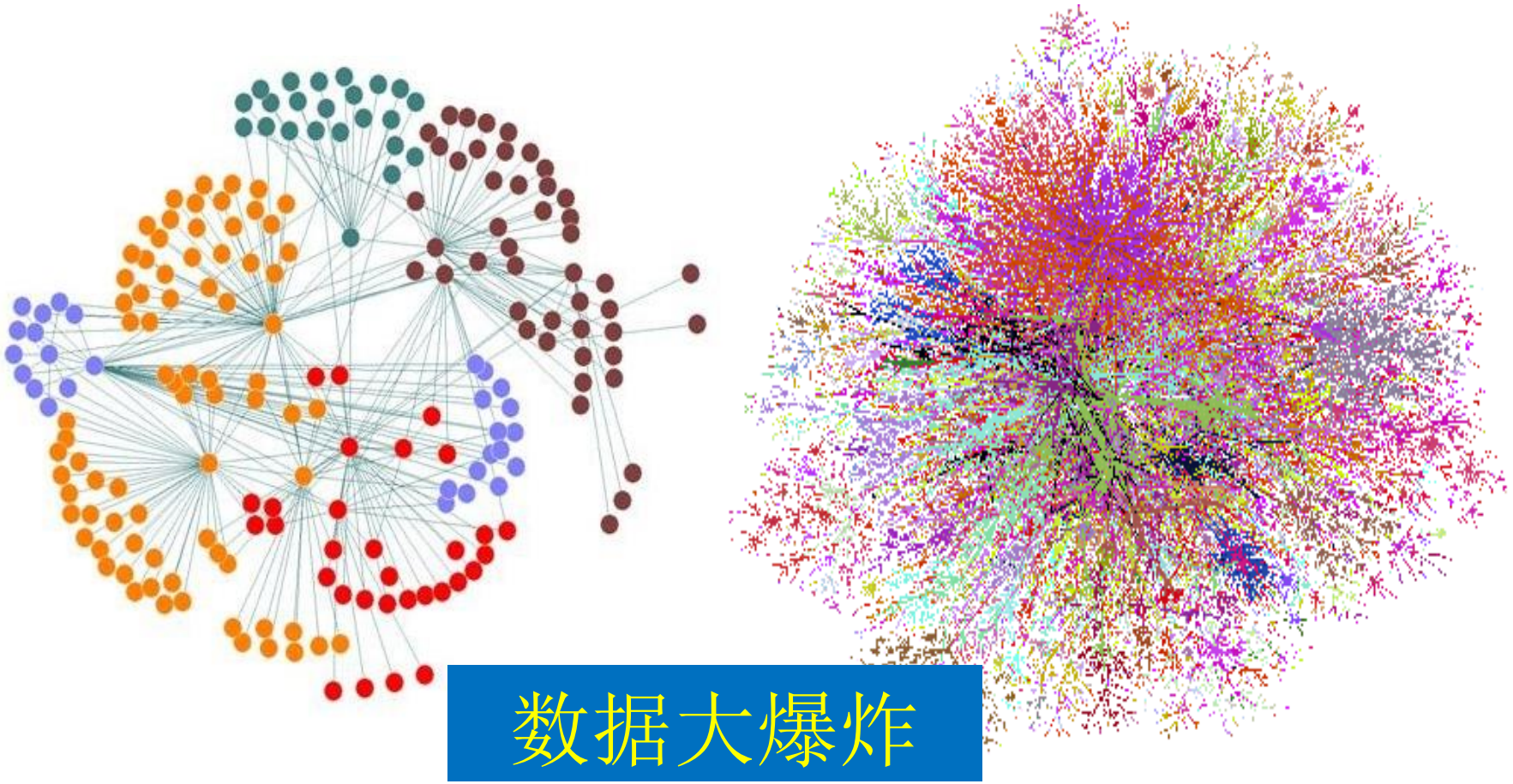
1YB = 1 ≈ byte = 1 000 000 000 000 000 000 000 000 byte

大数据范畴

3G手机/每秒

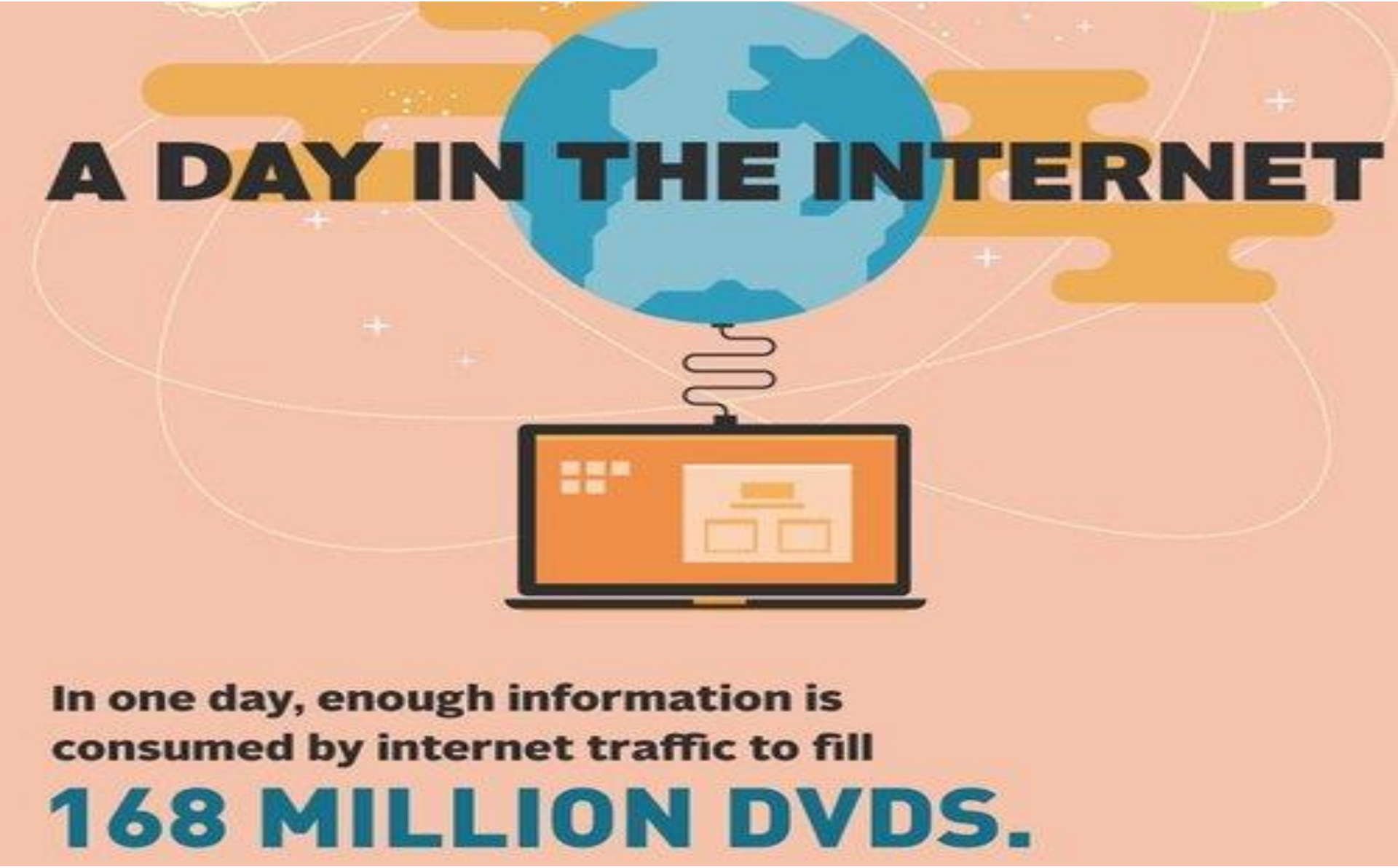
一部高清电影

6626 亿本红楼梦
或
4462个美国国会图书馆

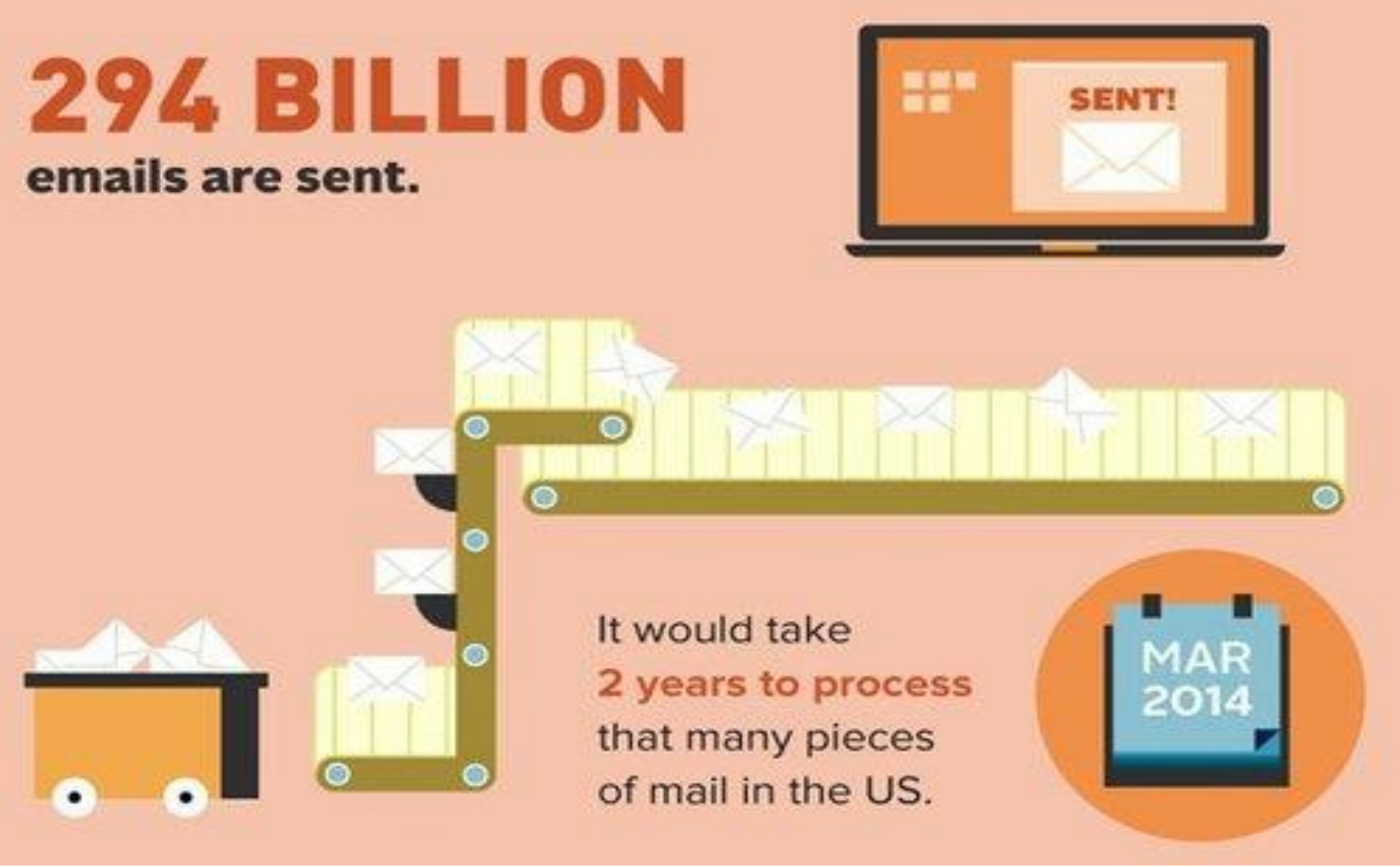


数据大爆炸

过去两年产生的数据占人类历史数据总量的90%，并且预计到2020年，人类所产生的的数据量将达到今天的44倍



互联网一天产生的内容足够刻满1.68亿张碟



一天发出2940亿封邮件，相当于美国两年纸质信件的数量

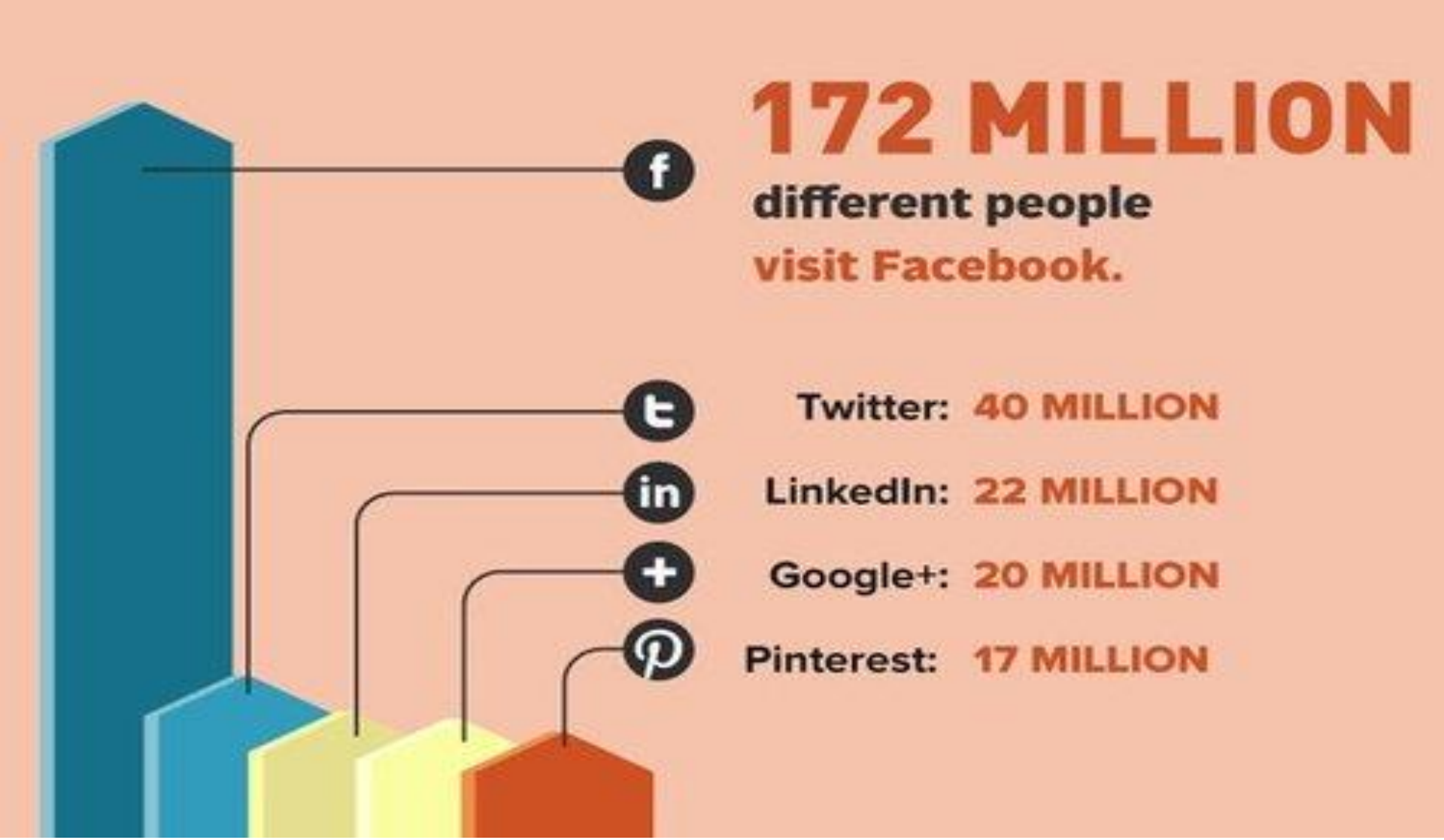
2 MILLION BLOG POSTS

are written.

Enough posts to fill

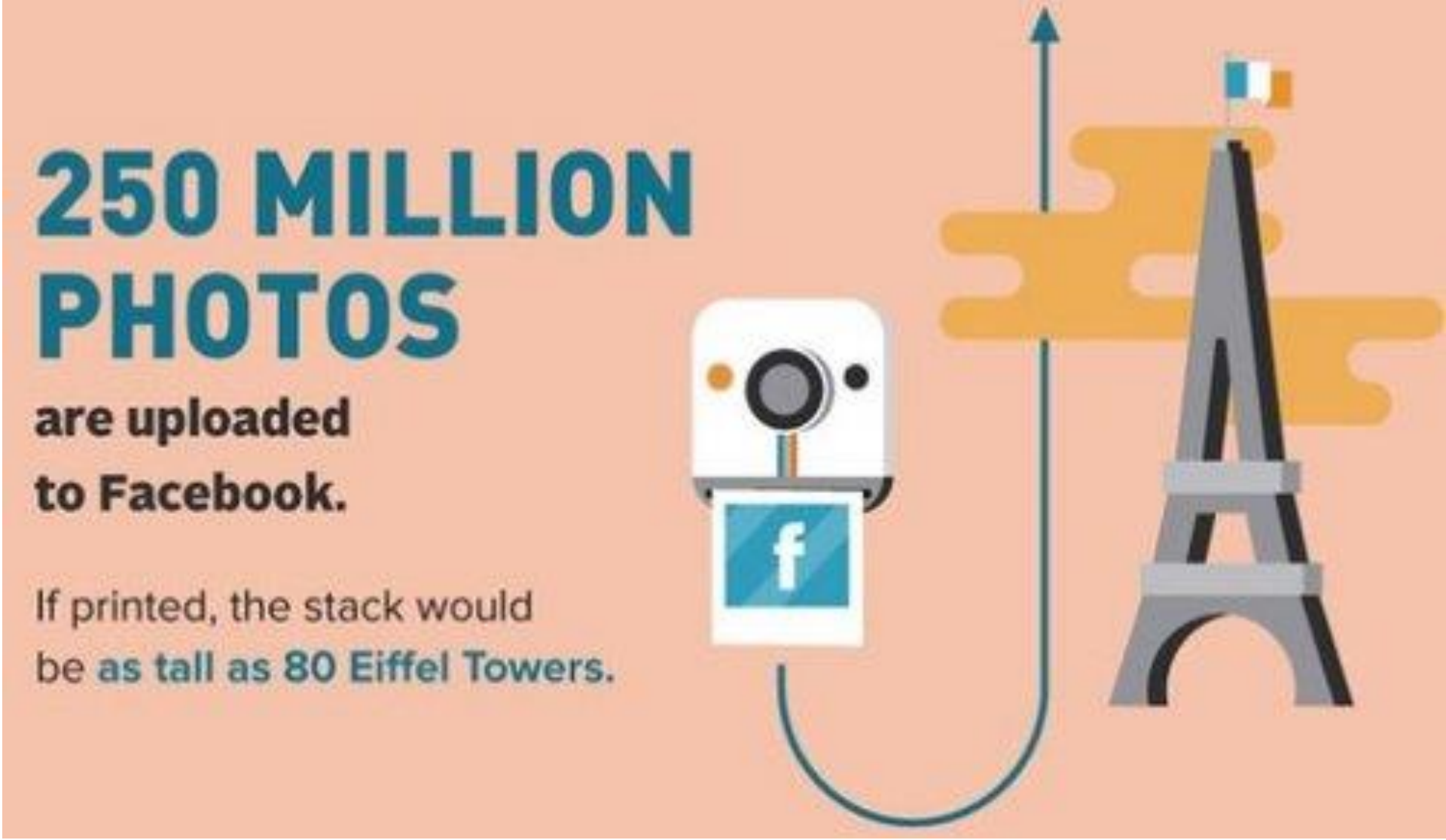
Time Magazine for 770 years.

一天的社区论坛上发出200万个帖子，相当于《时代》杂志770年的文字量



每天有1.72亿人登陆Facebook，4000万人登陆Twitter

人们每天在Facebook上耗费的时间总计47亿分钟



人们每天在Facebook上传2.5亿张图片，如果都打印出来相当于80座埃菲尔铁塔的高度

每天在YouTube上传86.4万小时视频，如果不间断全部播放完需要98年

用户每天在Pandora上听1870万小时音乐，如果Pandora从公元1年（西汉末年）开始播放，现在还在放

顾客姓名	订单日期	订单号
柳小山		54019
柳小山		20513
柳小山		36262
柳小山		36262
柳小山		36262
柳小山		39682
柳小山		4132
巴朗	8/16/2007	26949
巴朗		02
巴朗		94
巴朗		50
巴朗		63
巴朗		62
巴朗	10/30/2009	27559
巴朗	12/25/2009	20737
巴朗	12/25/2009	20737
巴朗	12/29/2009	46662
巴朗	12/29/2009	46662

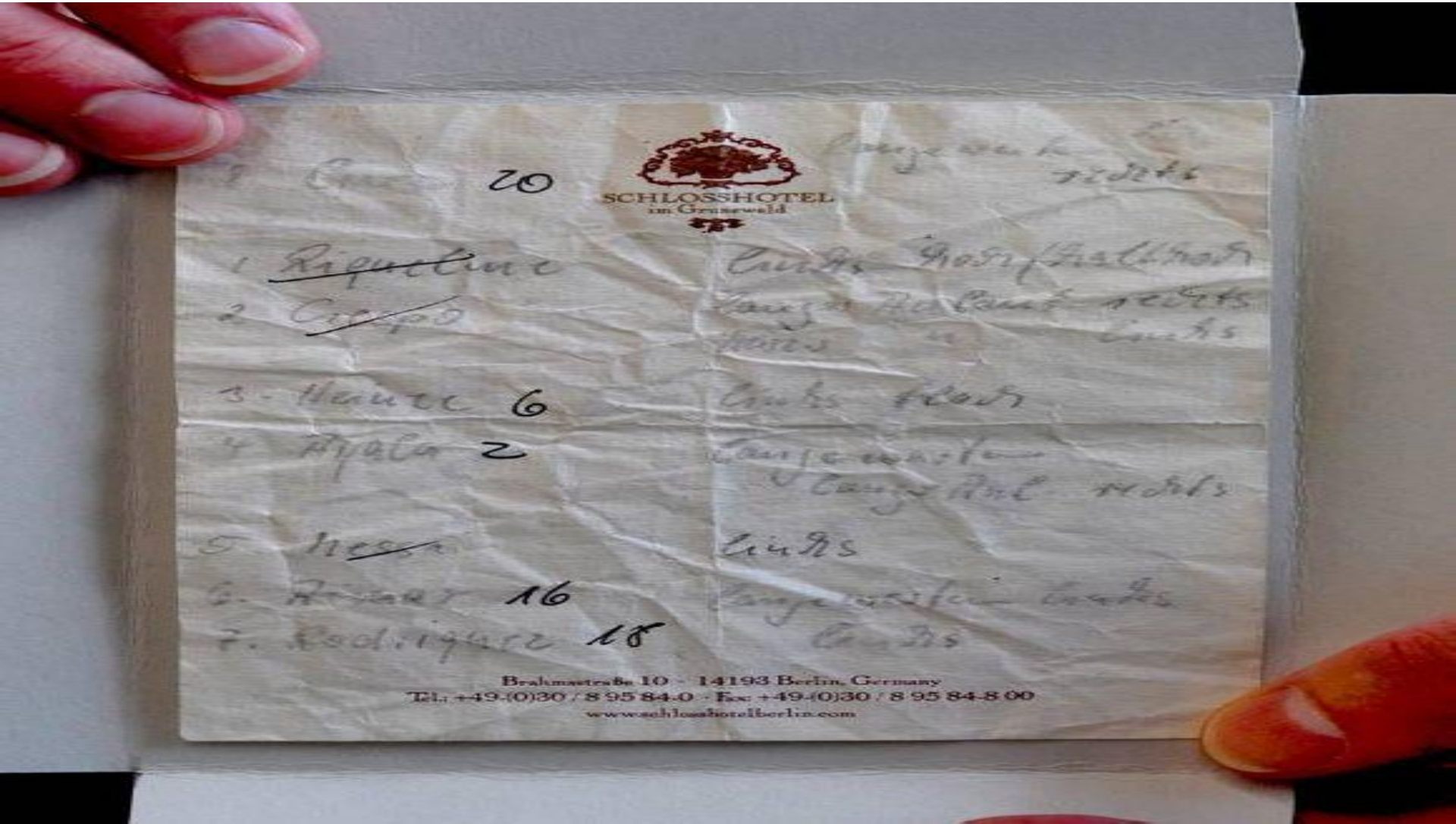
单一规整
结构化

多样繁杂
非结构化

大数据

大数据有啥用？

大数据有啥用？


SCHLOSSHOTEL
im Grunewald
Brahmsstraße 10 · 14193 Berlin, Germany
Tel.: +49-(0)30 / 8 95 84-0 · Fax: +49-(0)30 / 8 95 84-8 00
www.schlosshotelberlin.com

大数据有啥用？

数据挖掘、模式识别课程

- 我们身边的事：

sina 新浪网
sina.com.cn

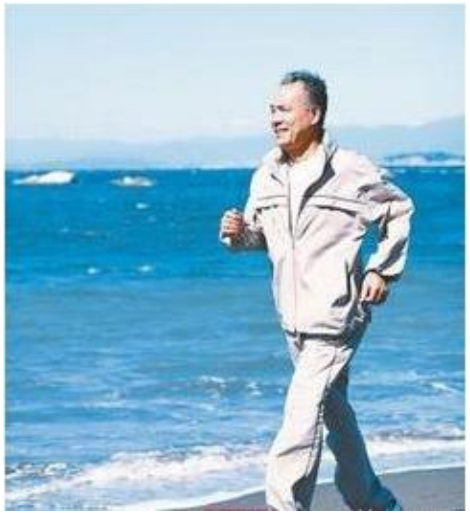
猜你喜欢

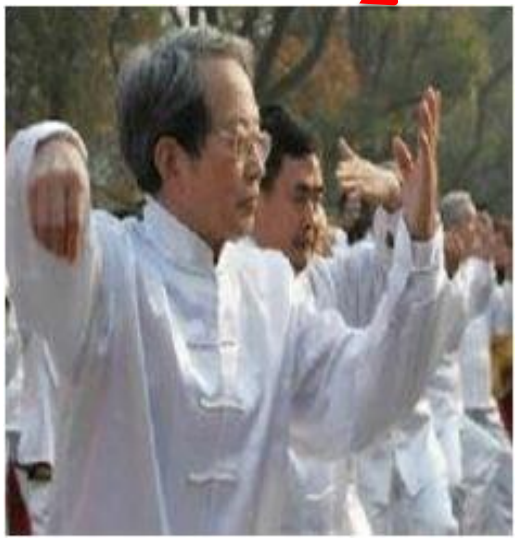
- 互联网经济，“关系”
- 北京7级阵风带来扬沙
- 赛立信通信研究:移动办
- 李光斗：“互联网”风
- 宋清辉：从中国制造走向
- 滥用互联网思维是一种病
- “互联网”概念，阿里
- “互联网”需要政府监

数据分析小例子：

- 心脏病快速诊断：临床发现心脏病与血压、胆固醇有关

病人编号	血压	胆固醇	是否有心脏病
1	73	150	无
2	85	165	无
.....	.....	.....	.....
10	110	190	有

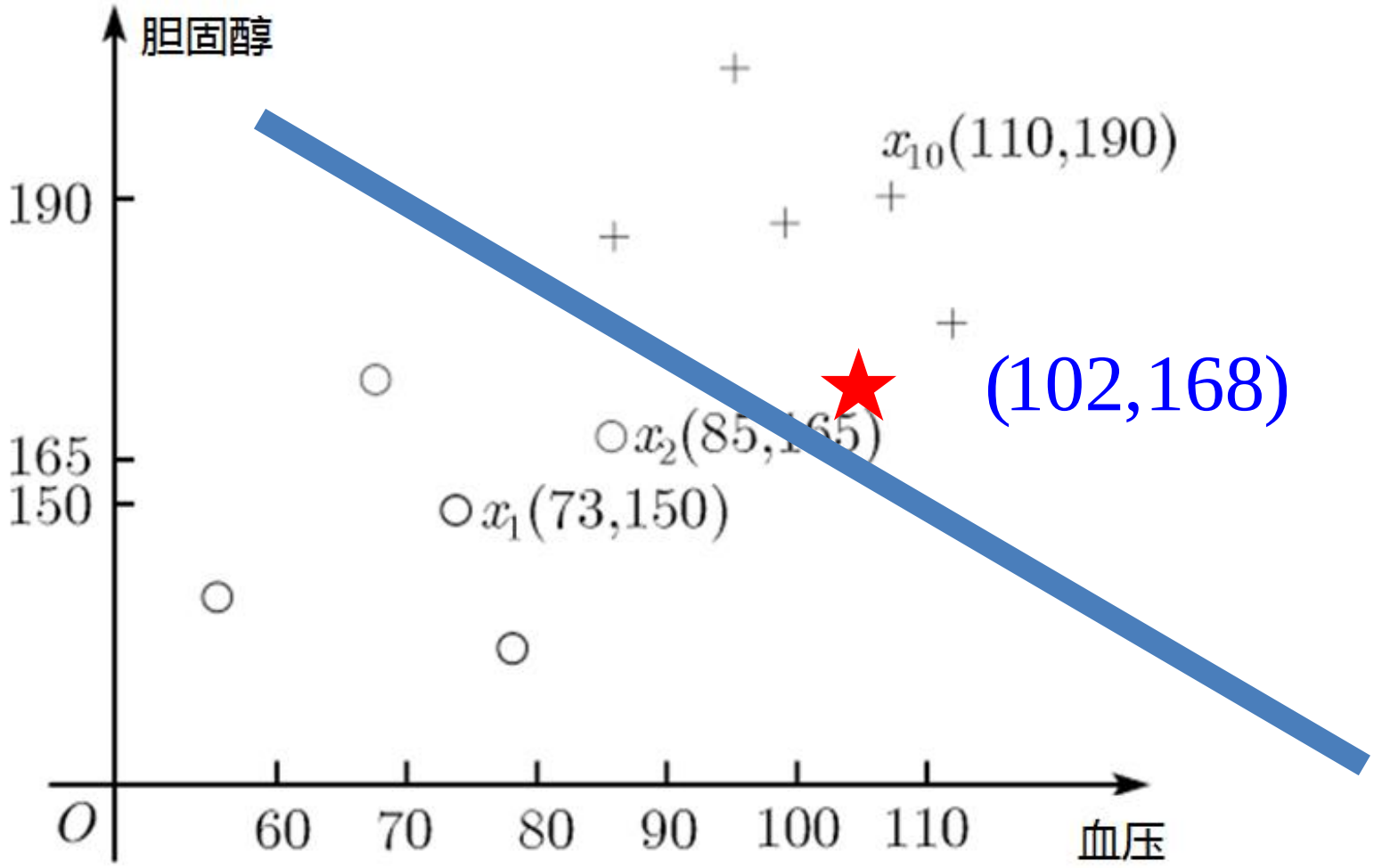


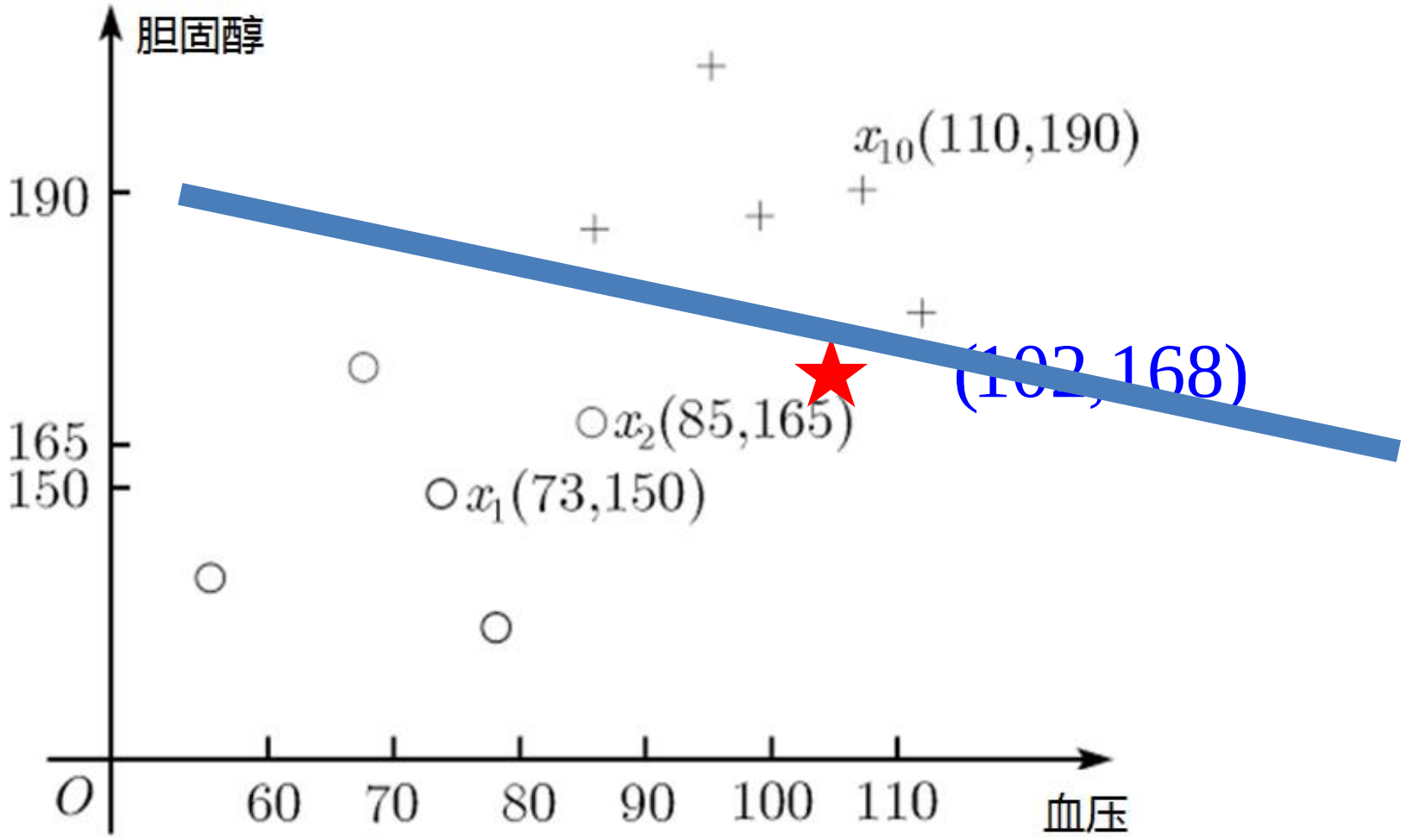


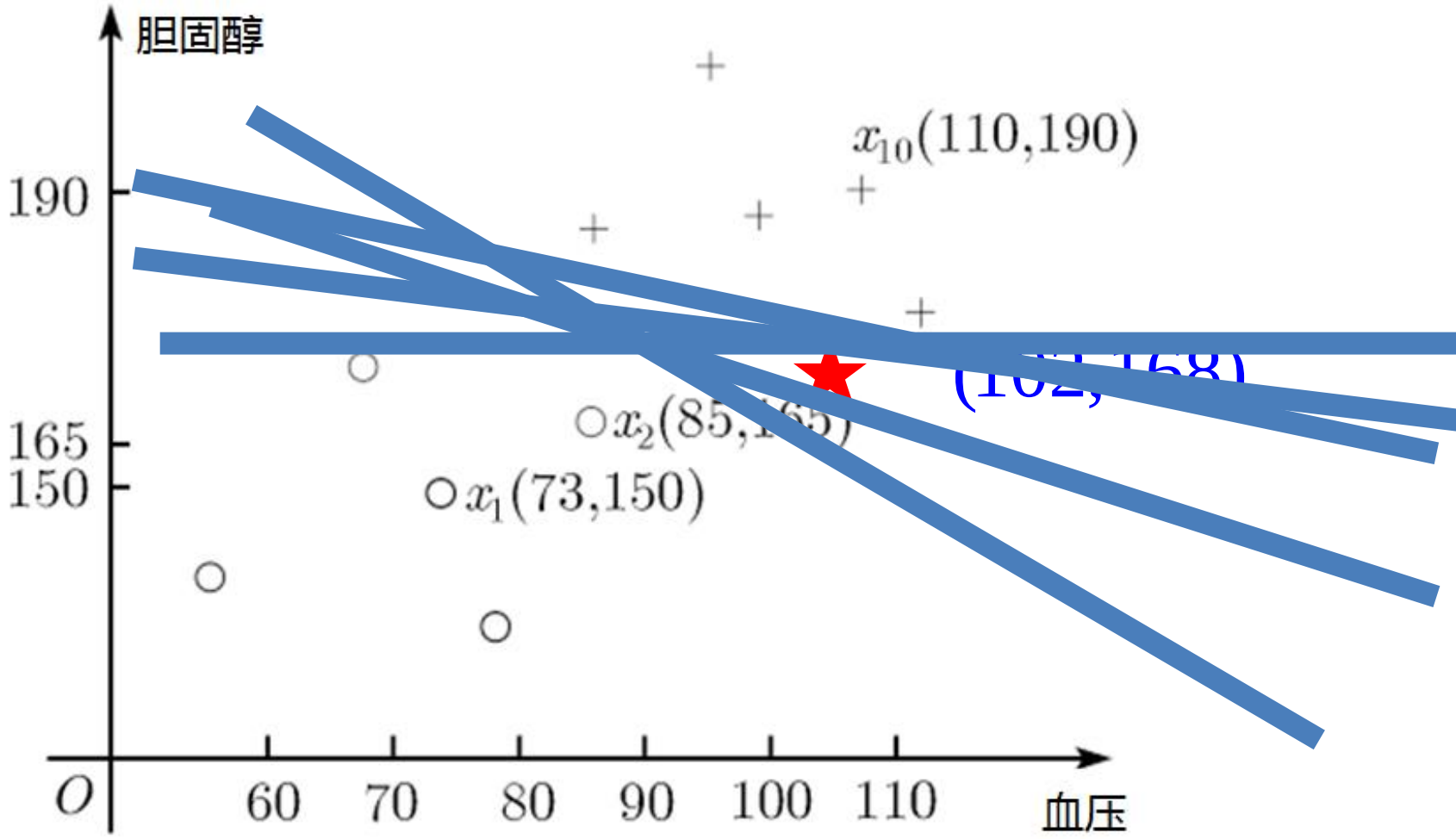
数据分析小例子：

- 心脏病快速诊断：临床发现心脏病与血压、胆固醇有关

病人编号	血压	胆固醇	是否有心脏病
1	73	150	无
2	85	165	无
.....	.....	.....	.....
10	110	190	有
11	**102**	**168**	**?**

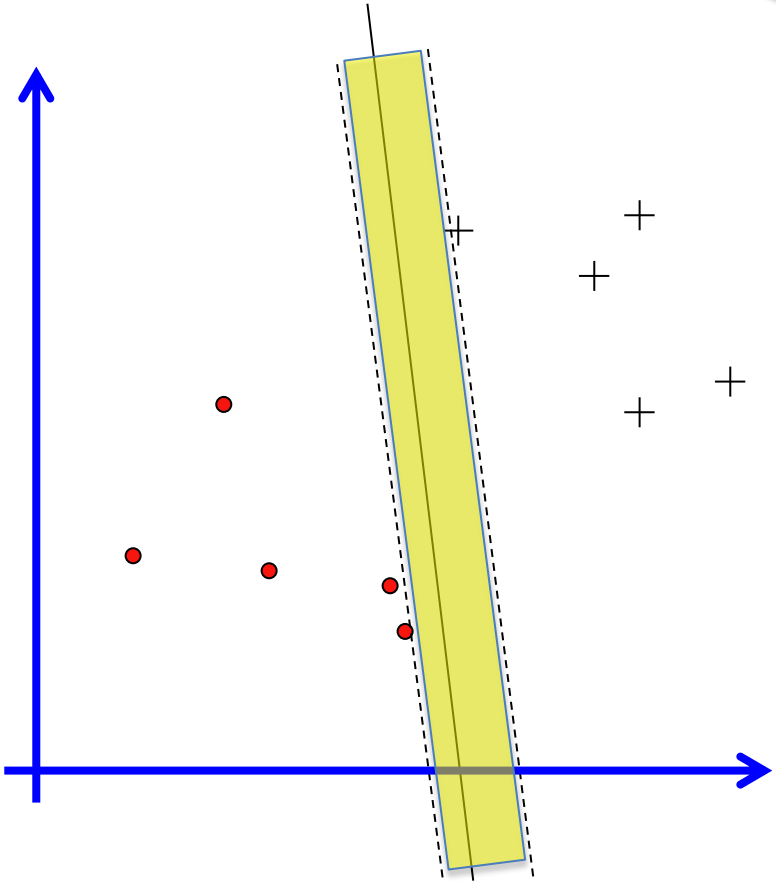

胆固醇
190
165
150
O
60
70
80
90
100
110
血压
$x_{10}(110,190)$
$x_2(85,165)$
$x_1(73,150)$
(102,168)

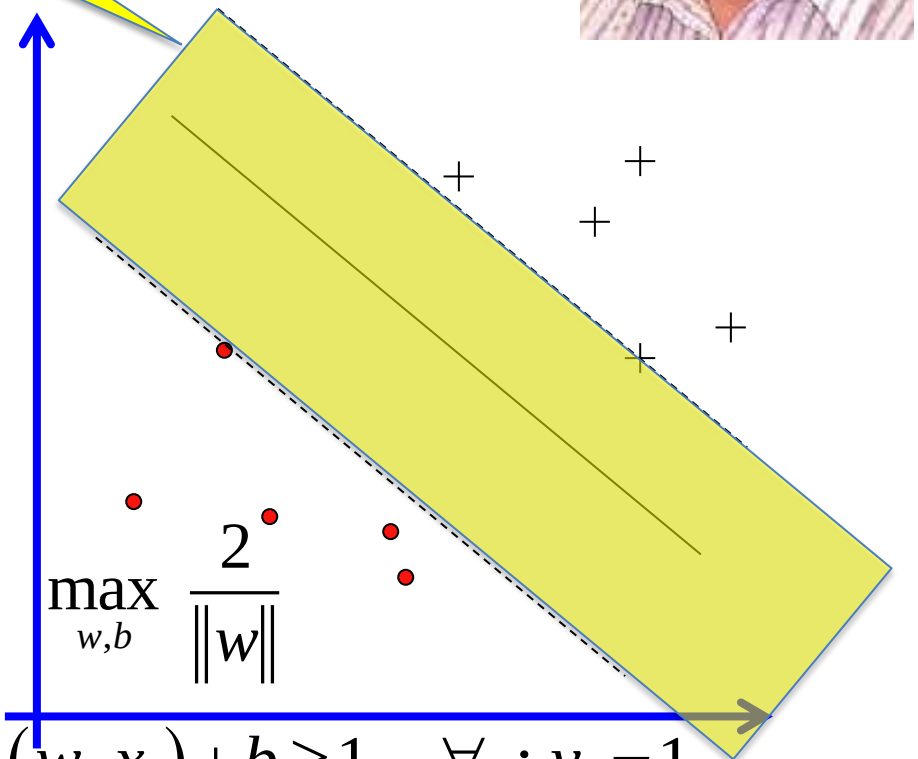

胆固醇
190
165
150
$x_{10}(110,190)$
$x_2(85,165)$
$x_1(73,150)$
(102,168)
O
60
70
80
90
100
110
血压


胆固醇
190
165
150
O
60
70
80
90
100
110
血压
$x_{10}(110,190)$
$x_2(85,165)$
$x_1(73,150)$

最大间隔原则

支持向量机





$$\max_{w,b} \frac{2}{\|w\|}$$

$$(w \cdot x_i) + b \geq 1, \quad \forall_i : y_i = 1$$

$$(w \cdot x_i) + b \leq -1, \quad \forall_i : y_i = -1$$

6

云计算
（约2008年~至今）

效率低
计算能力太弱

效率高
计算能力超强

6 云计算（约2008年~至今）

- **云计算：** 通过互联网把多个成本较低的计算实体整合成一个具有强大计算能力的完美系统。

- **简单理解：** 运营公司提供云服务器，云服务器上有强大的计算程序和广阔的存储空间，用户通过网络远程登录云服务器，并按照需要使用这些计算程序和存储空间

- 百度开放云平台


百度开放服务平台
直达号
轻应用
移动应用
开放云
开发者服务
百度媒体云
语音识别服务
重新定义人机交互
每种创新性的交互，催生一批创新性的应用
百度媒体云语音识别服务正式上线，现在就看你的了！
语音搜索词识别
通用语音识别
长文本语音输入
垂直领域识别

- 百度开放云平台

云服务(Cloud)

应用引擎(BAE)

多语言、弹性的服务端运行环境

云存储

自动扩容、独立冗余大数据对象存储

云数据库

支持MySQL、MongoDB、Redis

云推送

支持通知、消息、富媒体推送服务

媒体云

提供视频/图像/语音等相关服务

LBS云

从云到端一站式的LBS开发支持

端服务(Frontia)

第三方帐号登录

为应用轻松搭建帐号功能

个人数据存储

支持多端同步的用户数据存储服务

社会化分享

为应用集成微博/微信/QQ等分享

移动应用统计

提供应用渠道/版本等的统计分析

用户反馈

提供统一管理反馈消息/倾听用户声音

云计算涉及的问题：

1）数据存储：一个非常庞大数据库文件，单机存不下，如何横跨多台机器存储？

我希望看到的是一个文件系统，引用的是一个文件路径，但实际的数据分成很多小块分别存放在很多不同的机器上。

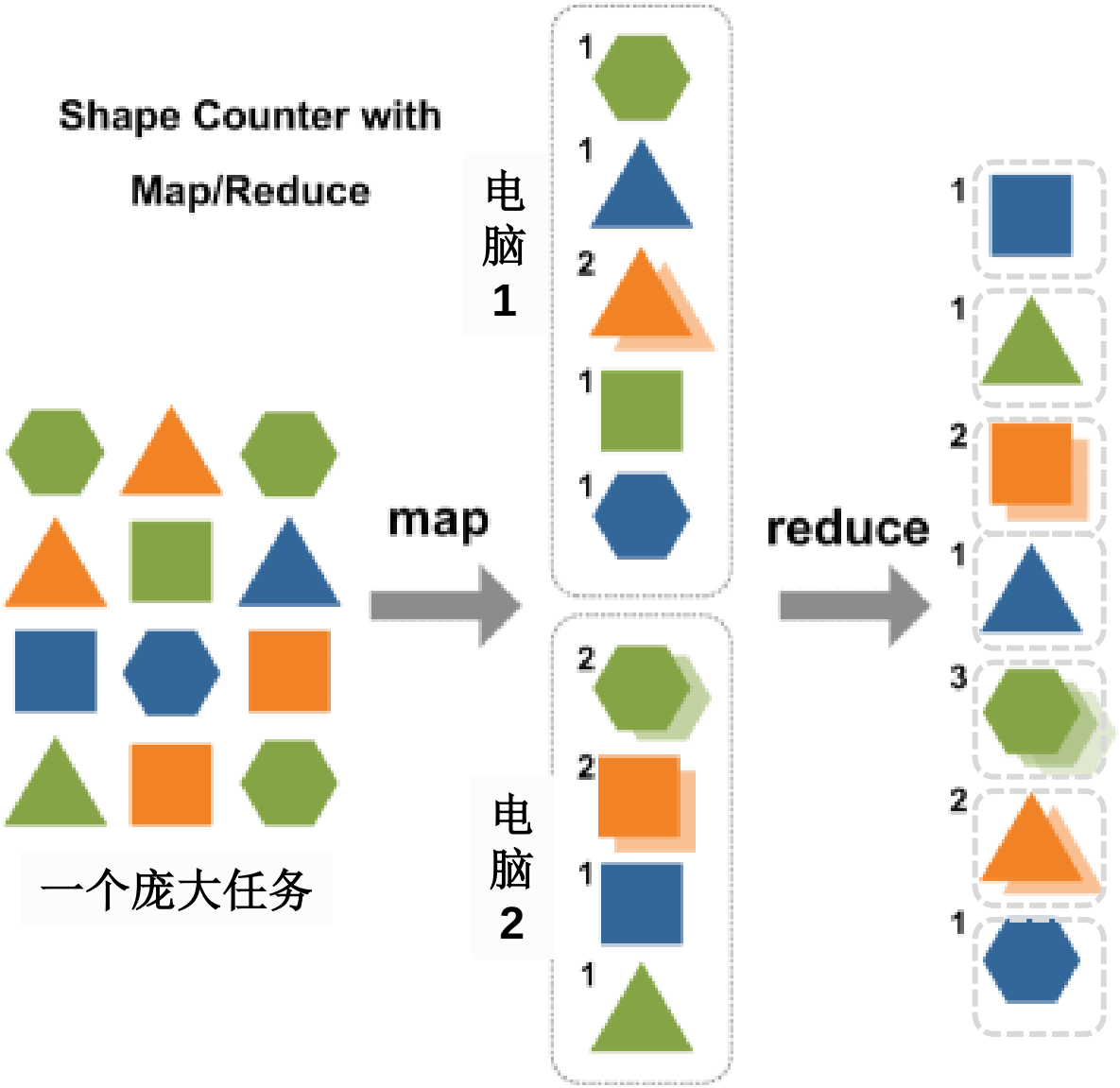
解决方案：GFS（谷歌分布式文件系统）

2）数据处理：一个非常庞大数据计算任务，单机[illegible]效率太低，如何分配给多台机器协同处理？

把一个庞大的数据分析任务划分成多个[illegible]机器同时做以节省时间，但需要考虑[illegible]启相应任务，机器之间如何相互协调等。

解决方案：MapReduce（谷歌分布式计算框架）

MapReduce 原理



7

互联网+
（2015年~至今）

- **2015年3月，全国两会，马化腾提案**
《以互联网+为驱动推进我国经济社会创新发展》

- **2015年3月，李克强总理在人大《政府工作报告》中首次提出制定“互联网+”行动计划，推动移动互联网、云计算、大数据、物联网等与现代制造业结合**

7

互联网+
（2015年~至今）

互联网+

商业价值

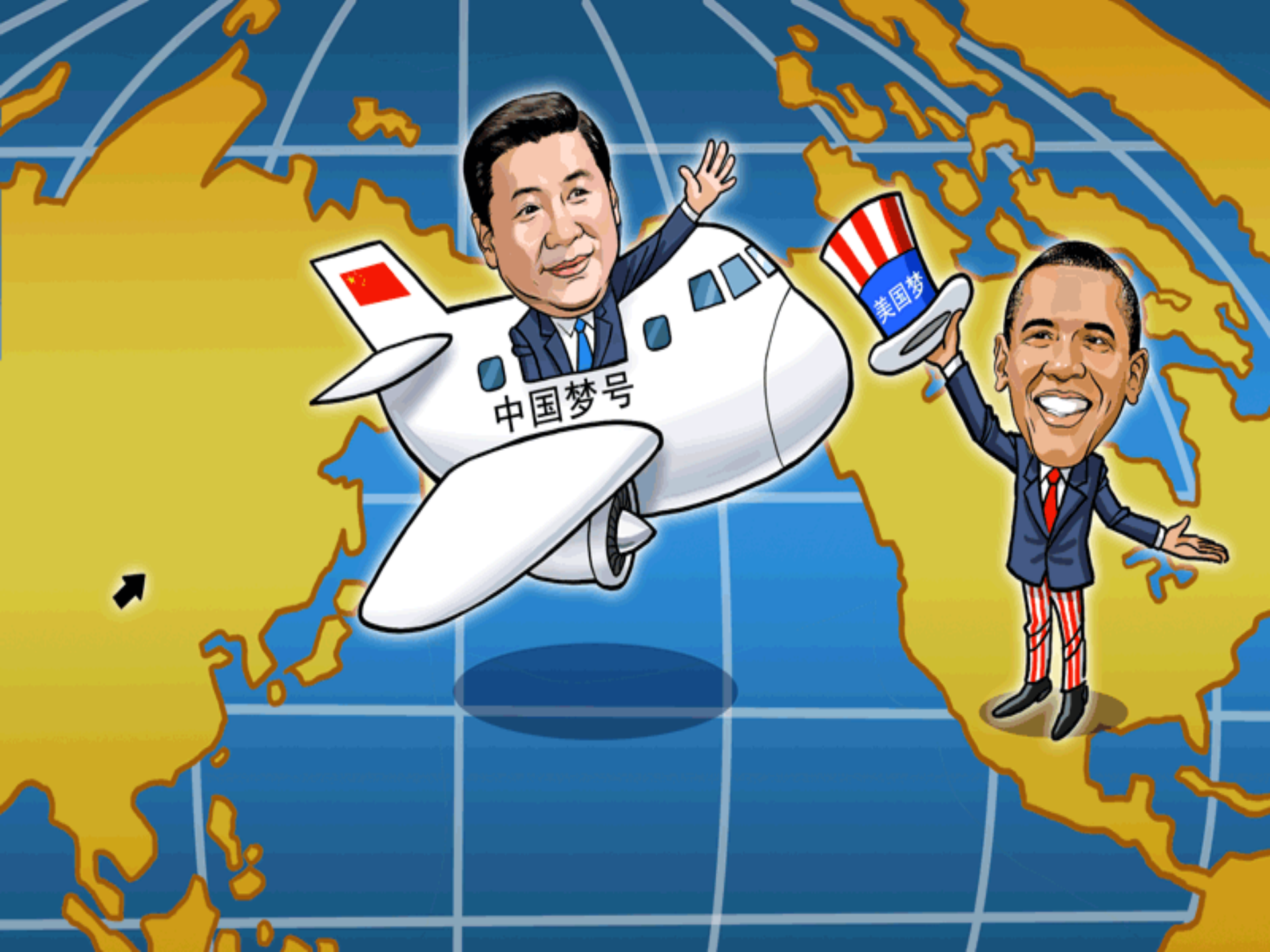
衣 淘宝，凡客

食 大众点评，美团，饿了么

住 携程，去哪儿

行 神州专车，滴滴，uber，快滴

• • •


中国梦号
美国梦

- **随习大大访美的15名企业领袖：**

阿里巴巴的马云、腾讯的马化腾、百度的李彦宏；

中远集团的马泽华、中国建筑的官庆、中国银行的田国立、工商银行的姜建清；

万向集团的鲁冠球、联想集团杨元庆、新奥能源的王玉锁、双汇集团的万隆、伊利集团的潘刚、海尔集团的梁海山；玉皇化工的王金书和天津钢管集团的李强。

- **美方参与对话的15名商界大佬所在公司：**

微软、思科、苹果、IBM、亚马逊、通用汽车、霍尼韦尔、杜邦、陶氏化学、百事、波音、星巴克


第八届中美互联网论坛
The 8th US-China Internet Industry Forum / Seattle, September 23, 2015

- 参加本届论坛的美国企业大佬包括微软公司CEO萨提亚、英特尔CEO科再奇、IBM董事长兼CEO罗睿兰、苹果CEO库克、高通CEO史蒂夫·莫伦科夫、脸谱创始人扎克伯格、领英创始人霍夫曼、亚马逊CEO贝佐斯、Airbnb联合创始人兼CEO切斯基、雅虎创始人杨致远等。

- 中方参加者包括阿里巴巴马云、腾讯马化腾、百度张亚勤、京东刘强东、新浪曹国伟、搜狐张朝阳、联想杨元庆、奇虎360周鸿祎。

高新产业都在信息技术行业！
高新产业都在信息技术行业！

谢晓波，2015毕业
腾讯

鲁珵，2015毕业
公务员（北京地税）

严红理，2015毕业
华为

- 张红阳，2014年，中国农业银行
- 许梦玲，2014年，北京银行
- 郑琛，　2014年，大唐电信
- 余晓峰，2013年，中国建设银行
- 刘晓春，2013年，百度
- 曲明超，2013年，中国人寿保险
- 张志佳，2012年，中国建设银行
- 曹明华，2012年，搜狐
- 陈勇，　2011年，百度
- 卿金坚，2011年，新加坡 国立大学
- 吴丹，　2011年，国家天文台
- 付增梅，2010年，中国科学院
- 唐异东，2009年，中国移动
- 赵大伟，2008年，阿里巴巴
- 金鑫，　2008年，美国伊利诺伊大学

谢谢各位同学！