中国中文信息学会战略研讨会，江西婺源

互联网时代的中文信息处理

孙 乐

中国科学院软件研究所
中国中文信息学会

2012年4月13日

提纲

- 背景
- 互联网时代
- 中文信息处理再认识

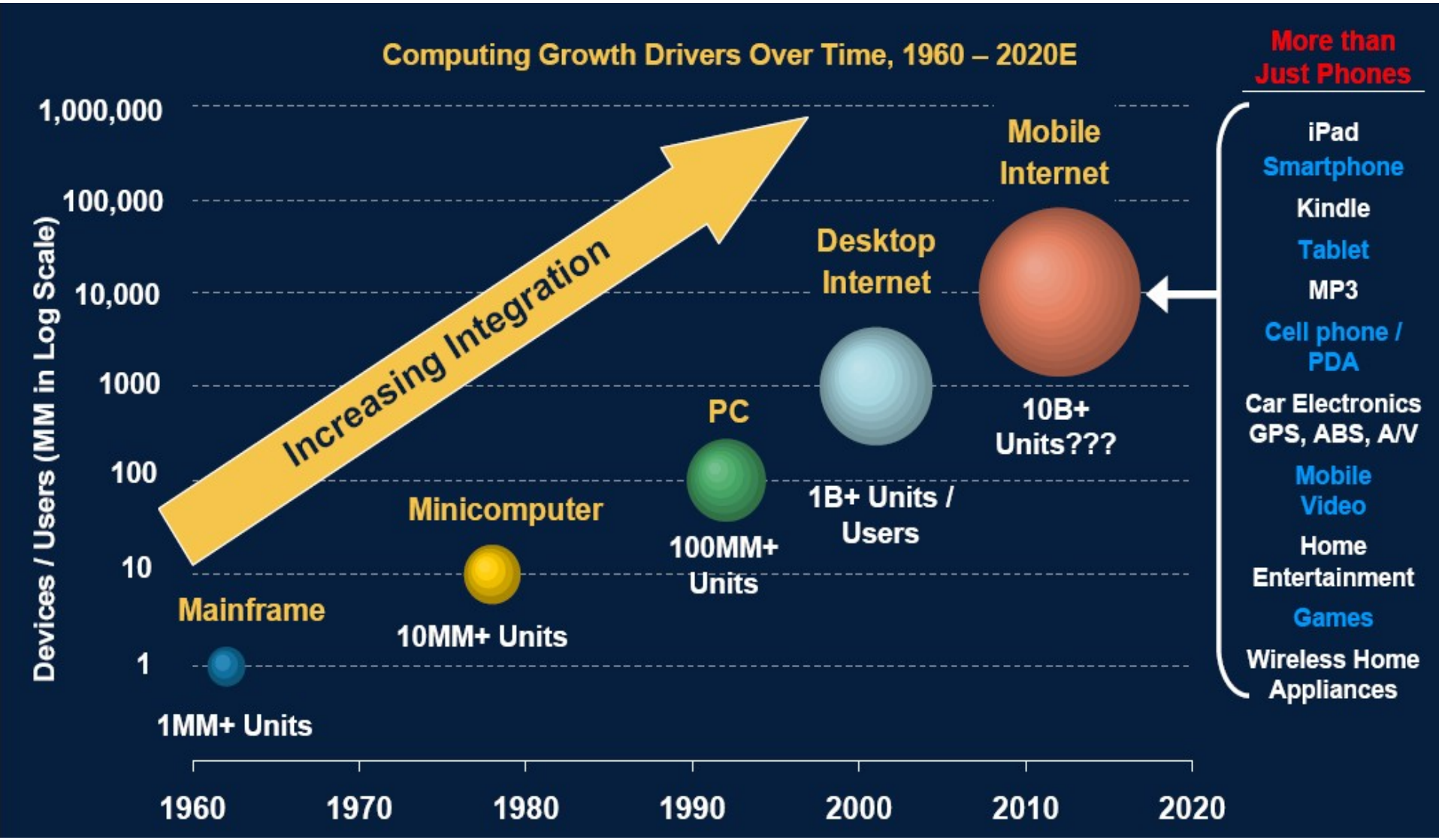
背景

- **信息**是当前知识经济社会的基础支柱，信息将成为自然资源、土地、资本和劳动力之外创造财富的重要来源

- 人类创造和传承文明的最主要载体是**语言和文字**

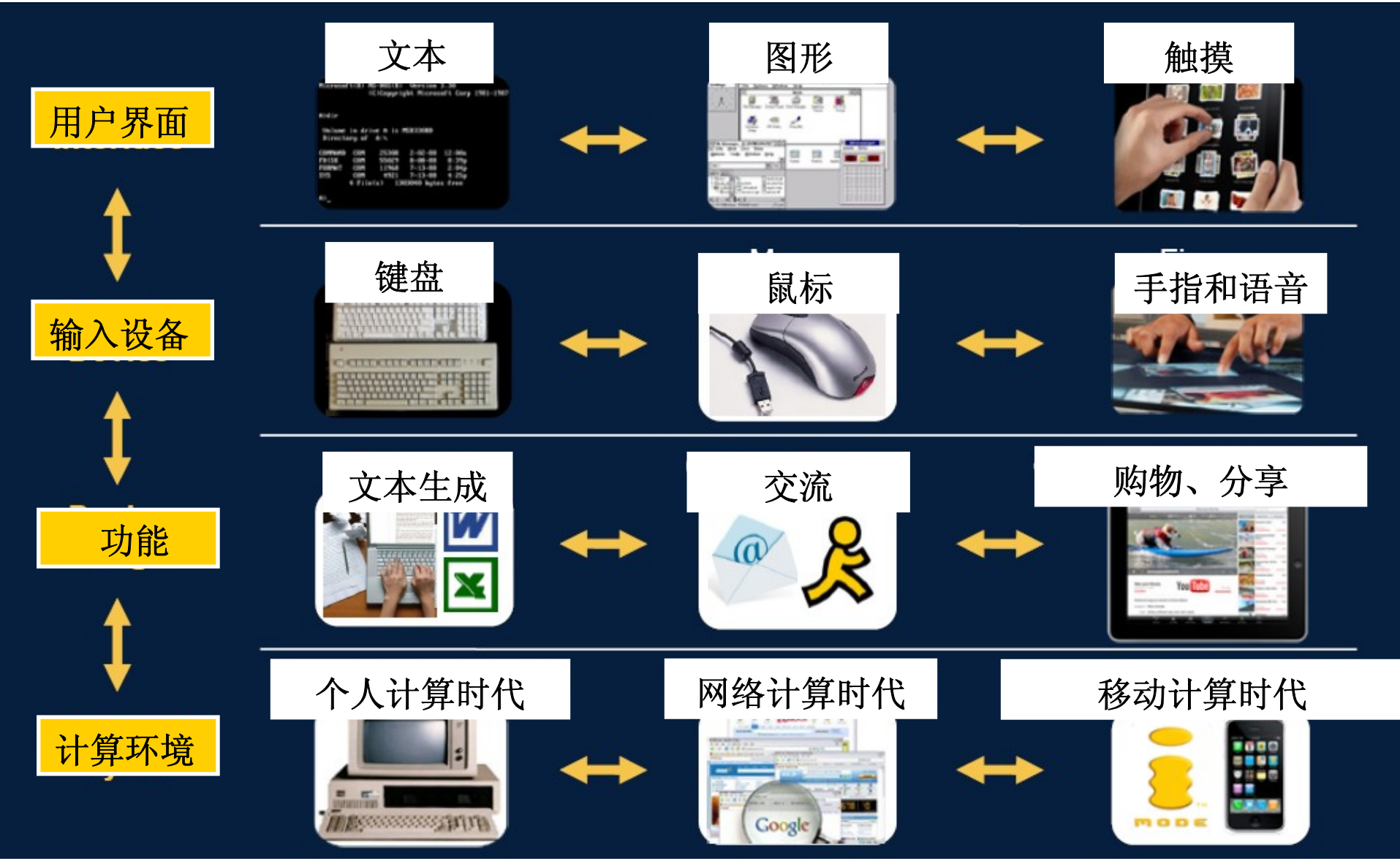
- **中文信息处理技术**是以计算机为工具对中文（包括各种少数民族语言）信息（语音和文本）进行输入、存储、传输、输出、识别、转换、压缩、检索、分析、理解和生成等各种类型处理和加工的技术

互联网时代



Morgan Stanley, Internet Trends, June 7, 2010

30年来的演化



Morgan Stanley, Internet Trends, June 7, 2010, 略有修改

30年来的演化

- 用户界面：文本 → 图形 → 触摸

  趋势：**自然**

- 功能：信息生成 → 信息交流 → 信息分享

  趋势：**社会化**

提纲

- 背景
- 互联网时代
- 中文信息处理再认识

重要成果

- **汉字激光照排系统** 王选 / 北京大学
- 国家最高科学技术奖 2001年
- 联合国教科文组织科学奖 1995年
- **国家科技进步一等奖 5 项**
- 华光Ⅲ型中文电子出版系统（1987年 北京大学）
- 联想式汉字微型机系统LX-PC（1988年 中科院计算所）
- 北大方正电子出版系统（1995年 北京大学）
- 智能型英汉机器翻译系统（1995年 中科院计算所）
- 汉王形变练笔的手写识别方法与系统

  （2001年 北京汉王科技有限公司）
- **国家技术发明奖二等奖**
- 亚伟中文速录机技术

时间	“国家科技进步二等奖”获奖项目
1989年	译星机器翻译系统（中软译星智能技术公司）
1997年	TRS 中文全文检索系统 （易宝北信信息技术公司）
1999年	THOCR - 97综合集成汉字识别系统（清华大学）
1999年	藏汉双语信息处理系统（西北民族学院）
1999年	计算机全汉字信息处理系统集成（北京中易电子公司、中国标准技术开发公司、北京中易郑码新技术公司、北京大正电子有限公司）
2000年	WPS2000智能集成办公系统(金山电脑有限公司)
2001年	维汉声图文一体化信息处理综合系统(新疆大学)
2001年	藏文视窗平台、字处理软件和藏文网站(西北民族学院、西北民族语言文字信息技术研究所、甘肃同元信息系统技术有限责任公司)
2002年	KD系列汉语文语转换系统（科大讯飞）
2003年	高性能东方文字文档智能全信息数字化系统（清华大学）
2004年	藏文计算机键盘和输入编码方法研究（青海师范大学等）
2006年	汉王OCR技术及应用（汉王科技）
2011年	智能语音交互关键技术及应用开发平台（中国科学技术大学，科大讯飞）
2011年	综合型语言知识库（北京大学）
2011年	藏文软件研发与推广应用（西藏大学）

重要成果（个人计算时代）

- 汉字输入技术

（速录、手写识别、语音、汉字识别、少数民族语言编码）

- 汉字输出技术

（汉字激光照排）

- 其他

（全文检索、机器翻译、知识库等）

30年来的演化

- 用户界面：文本 → 图形 → 触摸

  趋势：**自然**

- 功能：信息生成 → 信息交流 → 信息分享

  趋势：**社会化**

中文信息处理再认识

- 从自然的角度看

输入 → 语音
手写
图形→文字

输出 → 字形(书法、文化传承)
趋势分析（从数据/文本→图形转换）
查询结果的多维度展示
熟悉的语言（翻译）

中文信息处理再认识

- 从社会化的角度看

信息分享的前提：统一语义描述

信息分享的基础:

实体与实体之间的关系: 大规模知识库

人与人之间以及真实世界的关联: 社会计算

实体与真实世界的关联: 实体链接

信息分享的方式：

信息推荐/过滤	动态数据
信息检索/问答系统	静态数据

中文信息处理再认识

- **中文信息处理技术**是以计算机为工具对中文（包括各种少数民族语言）信息（语音和文本）进行输入、存储、传输、输出、识别、转换、压缩、检索、分析、理解和生成等各种类型处理和加工的技术

- **中文信息处理技术**

  是中文语言世界与真实世界的桥梁，

  是信息时代中国文化传承与发扬的基础，

  是互联网时代保护数字化国土的重要手段，

  是知识经济时代获取财富（信息资源）的工具。


为建设创新型国家
提供基础技术支撑
网络舆情监控、
社会化网络
群体智慧…
透析中文本质，
构建计算机用
中文语义理论
国家战略
Web演变
产业推动
中文信息处理
中文
知识资源
多学科融合
行业：速记、教育、情报、航空、公共安全…
计算机科学、脑科学、认知科学、智能科学、哲学、数学
Web资源挖掘

谢谢！

恳请批评指正！

sunle@iscas.ac.cn