中文信息处理战略研讨

资源 ● 评测

报告人：杨尔弘

2012.4.13

主题：我国中文信息处理事业发展的战略机遇在哪里？

1. 在新的历史时期，中文信息处理领域的内涵和外延已经发生了深刻变化，请您给出一段文字扼要描述您对中文信息处理（研究或技术）在新的历史条件下的定义和理解。
2. 如果说前三十年中文信息处理领域的工作可以概括为着力于如何让中文进入计算机，那么下一个三十年中文信息处理的主要着眼点和着力点应该在哪里？
3. 中文信息处理领域的重要原始创新可能在哪里？
4. 中文信息处理相关产业发展的重要方向或方面可能是什么？**如何推动中文信息处理领域“产学研”的战略合作，其有效模式可能是什么？**
5. 如何有效推动中文信息处理进入“核高基”之类的国家重大科研计划中（在上述研讨的基础上）？

国际上开展的评测

- 组织（DARPA)、NIST、大学、公司 联合 ）
- 定义了研究任务：（NER、EDT、RDC、EDC、topic、event ，KPb，……），各种语言同时开展
- 建立了公共的数据资源（共享的，基础知识、评测数据），评价技术的参照（我们很多的研究者贡献资源）
- 研究导向

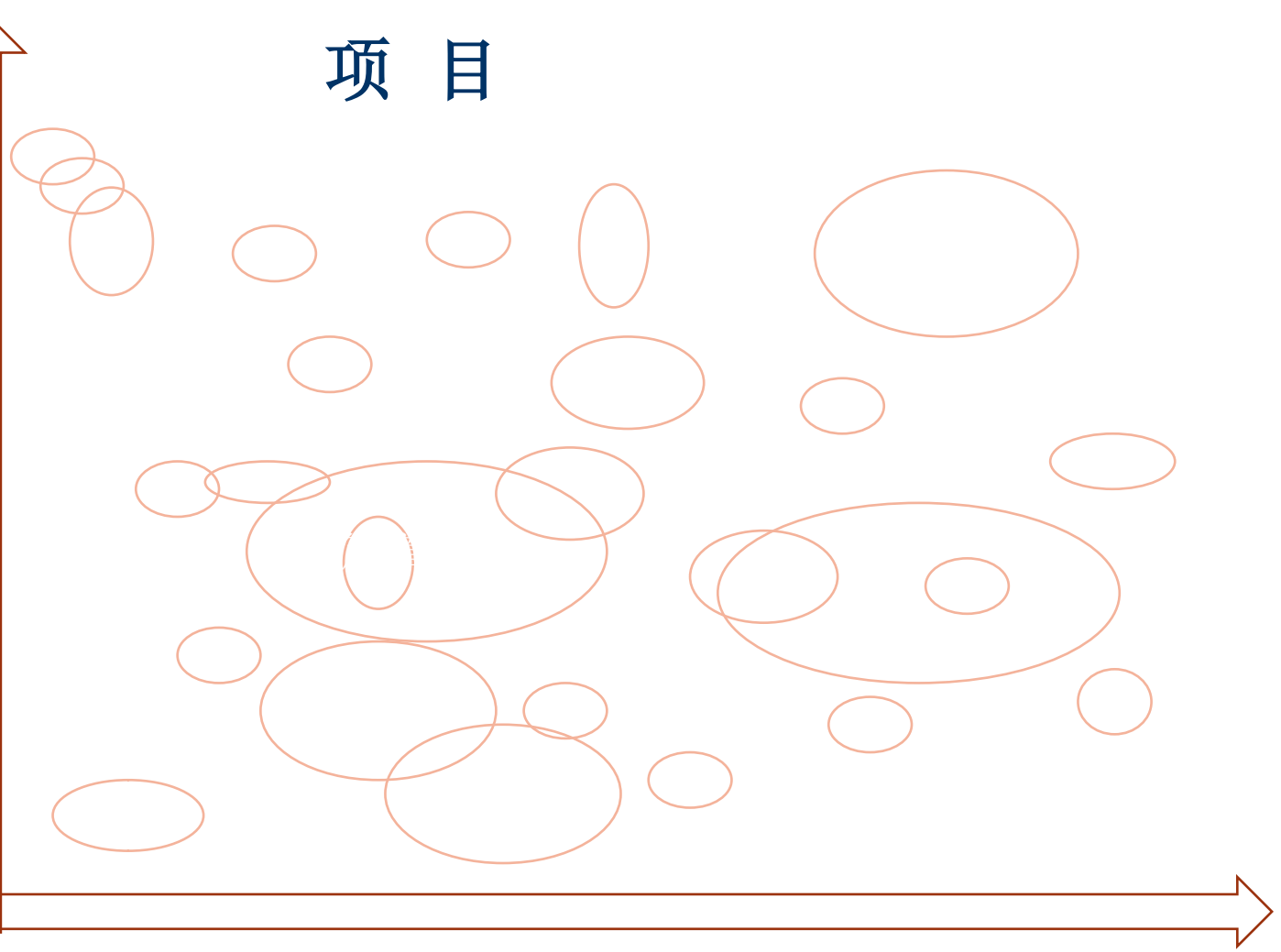
中文评测：研究驱动

- 评测
  - 863计划
    - 语音合成、汉字识别、机器翻译、自动文摘、全文检索、文本分类、分词标注、命名实体
  - Sighan、Cips
    - 分词标注、句法、人名消歧、wsi
  - Cips 机器翻译专委会
    - 举行了7届评测，每届多个项目
  - 信息检索专委会
    - 情感倾向性分析

- 任务定义
- 资源
- 导向

研究

项 目



产品

研究

项 目

- 973、863、基金……
- 校（所）企业合作项目
- ……

- **联合攻关项目？重量级话题　语义？**
  - **国家队**
  - **自由队**

- **评测机制**

产品

- 中文信息处理的专项基金
- 研究队伍
- 定义任务，定义出每**一个基本的 研究对象**（滚动的）
  - 比如：基本语义单元，推衍语义单元、意义关系 …… ，
  - ？更合适于计算的单元
- 建构资源：比如语义资源究竟应该建立什么样的核心、开放的资源？
  - 整合已有的数据资源
  - 定义众包的对象
  - 核心、外围知识资源
  - 汉语、少数民族语言
  - 形成公共的数据集
- 评测
  - 专题研讨

资源

- 狭义的资源：数据、语言、知识
- 广义：为人文学科提供语言数据、语言工具
  - 语言信息处理技术产生的数据的服务模式
    - 语言生活状况调查
  - 基于数据的人机互动的分析
    - 共时、历时数据

- 国家需求驱动，定义任务，形成研究导向，组织起来整合、建构中文（汉语、少数民族语言）的资源（语言相关）。
- 在广义的语言资源上提供研究、分析数据的服务模式。

谢谢大家！