INSTITUT FÜR DEUTSCHE SPRACHE

Marc Kupietz & Cyril Belica, Institut für Deutsche Sprache

BIG LANGUAGE DATA FOR ACADEMIC AND COMMERCIAL USE

Innovation Days 2013, Berlin, 2013-12-10

Mitglied der Leibniz-Gemeinschaft

BIG LANGUAGE DATA: GROWING USABILITY GAP

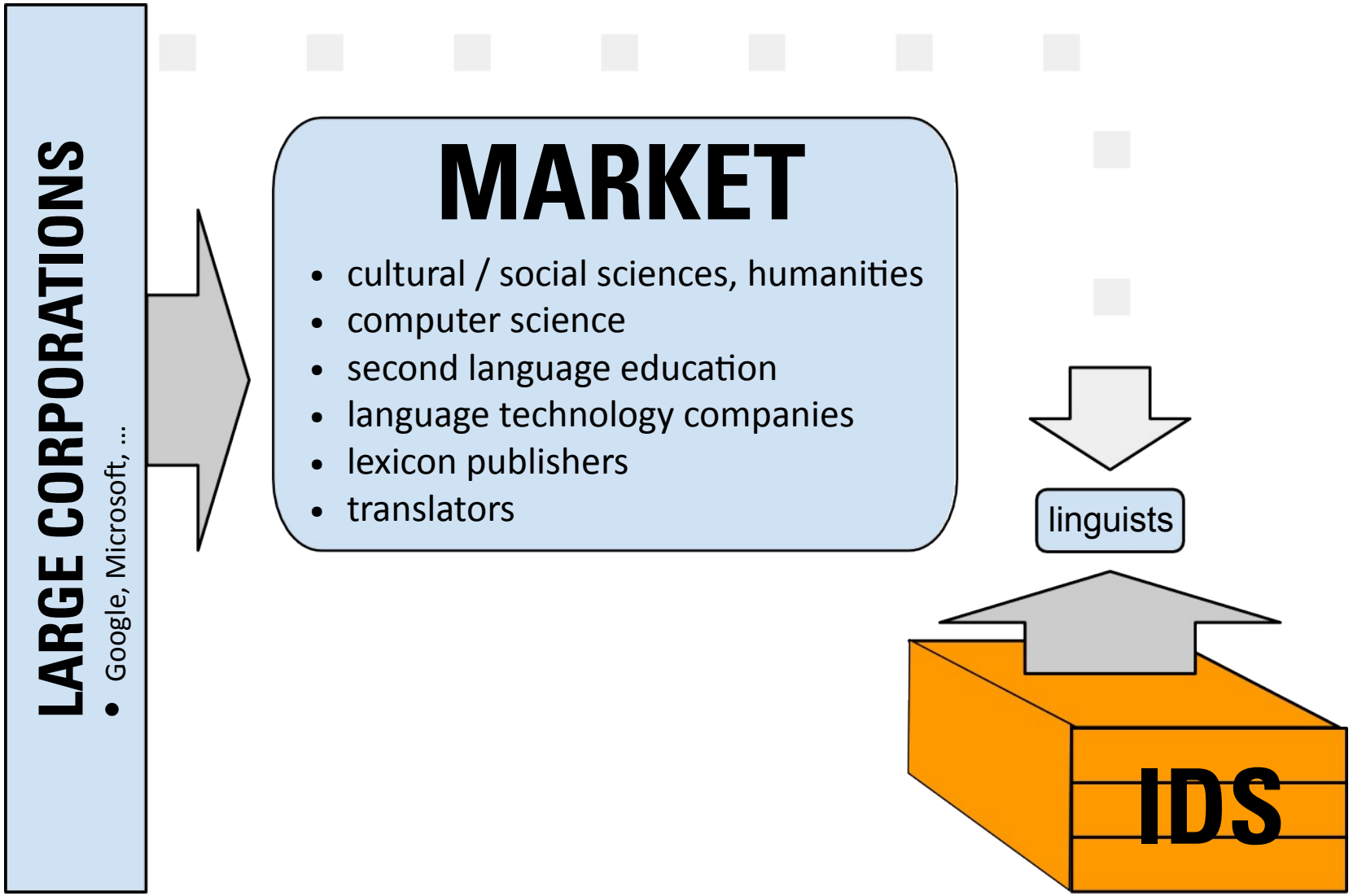
- general trend to “empiricalization” based on very large *corpora* in all areas concerned with language
- largest “corpus” almost exclusively held by Google
- largest non-web-based German corpus held by IDS

THE GAP

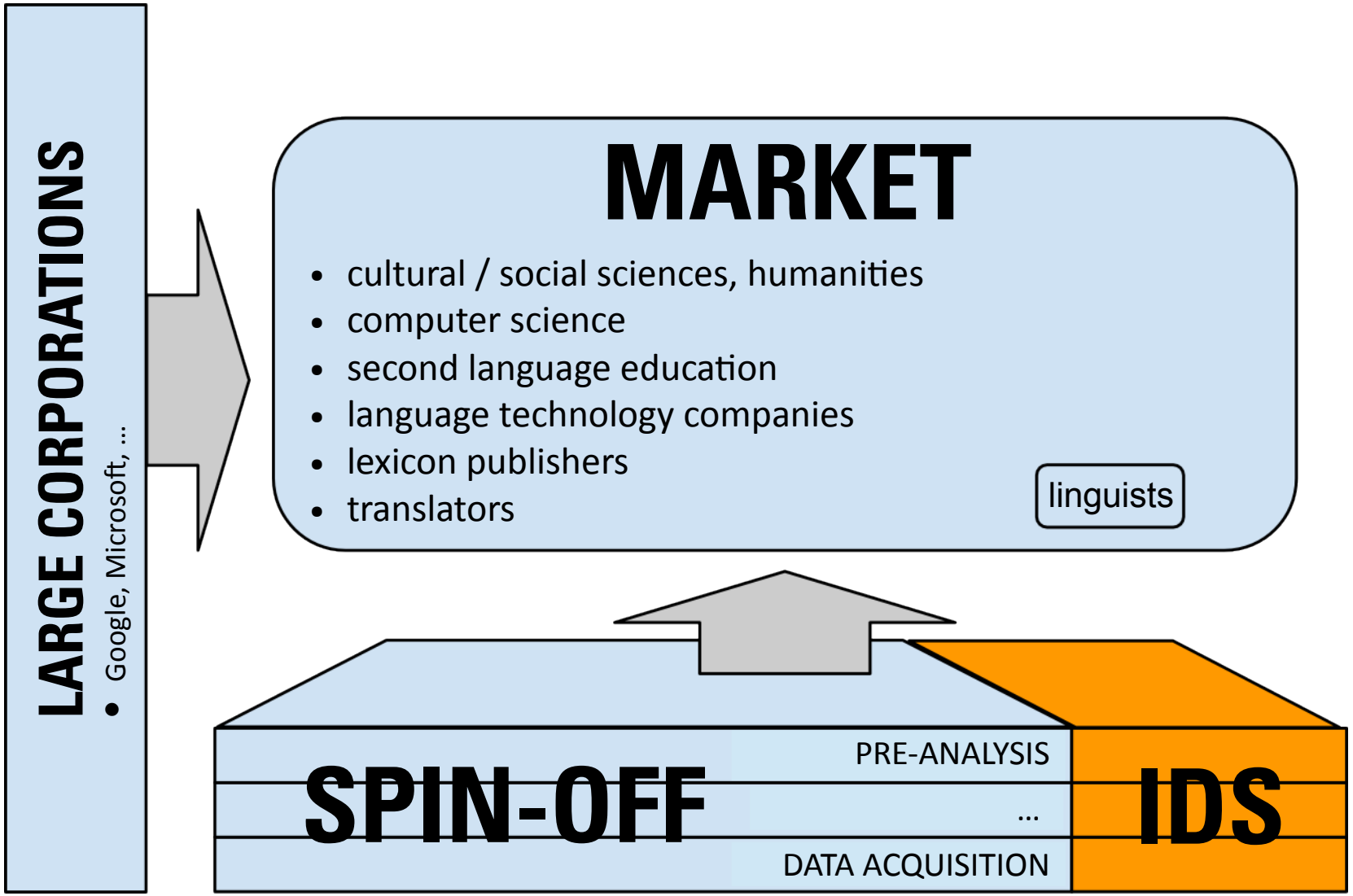
Growing gap in usability of big language data in academia and industry:

- linguistics and language technology cannot be based on google corpora only
  - not enough metadata: like archaeological findings without information about where they were found, etc.
- IDS-corpora can only be made available for very specific purposes
  - legal reasons, economical, license, technical, ... trade-offs

BIG LANGUAGE DATA SUPPLY FLOW: CURRENT STATE



BIG LANGUAGE DATA SUPPLY FLOW: PROPOSED INNOVATION



IDS-BASED SPIN-OFF IDEA
»OPEN WHAT WE HAVE TO NEW CUSTOMER GROUPS«

INTENDED PRODUCTS AND SERVICES:

- big, metadata-rich and clean corpora
- integration of existing and acquisition of new corpora
- metadata curation for the construction of controlled samples
- license acquisition and brokering
- automatic linguistic annotation
- a corpus query and analysis system
- technology for coping with IPR and limited bandwidth
- methodologies for the analysis
- interfaces to research infrastructures
- distillation of linguistic (and semantic) information

DEVELOPMENT STAGE

- for the special customer group »linguistics« (30,000 customers)
  all services have already been in production stage for several years
- spin-off is in proof of concept stage

- we are in discussion with …
  - project »Verwertung Geist« (BMBF-funded)
  - research infrastructure initiatives
  - potential partners
    - other research institutes, computer science departments, publishers
  - potential customers
  - rights holders
  - funding agencies
  - investors

AIM OF PITCHING

- looking for co-operation partners
- looking for language data rights holders (publishers, …)
- looking for seed-capital / investors
- discuss public interest aspects

INSTITUT FÜR DEUTSCHE SPRACHE

THANK YOU FOR YOUR ATTENTION

corpuslinguistics@ids-mannheim.de

Mitglied der Leibniz-Gemeinschaft