THE OHIO STATE UNIVERSITY

Linking Different Types of Data: The Case of Innovation

Bruce A. Weinberg

Ohio State, NBER, IZA

Unlinked data:
$100^1=100$

Unlinked data:

$100^1=100$

$100^1+100^1=200$

Unlinked data:

$100^1=100$

$100^1+100^1=200$

$100^1+100^1+100^1=300$

Linked data:
$100^1=100$

Linked data:

$100^{1}=100$

$100^{1+1}=10{,}000$

Linked data:

$100^{1}=100$

$100^{1+1}=10{,}000$

$100^{1+1+1}=1{,}000{,}000$

Linked data with distinct data elements leads to exponential growth

For innovation, data include:

- **Bibliometrics** – Pubs, Patents, Citations, Grants
- **Surveys** – Surveys of Earned Doctorates, Survey of Doctorate Recipients
- **Transactional** – Payments, especially to people

Transactional Data

- For innovation, people & teams are critical
  1. Innovations are not made on production lines (yet!)
  2. “Wrapped up in people” (Oppenheimer)
- UMETRICS is the back end of payments on sponsored projects on 72 campuses
- Identifies entire teams, even the “unseen” on papers and patents
- Tracks researchers into economy

The Career

Background, Training

Background: SED

Training: Transcripts, SED

Support: SED, Grants

Teams / Networks: UMETRICS, Pubs, Patents

Topics: Dissertation, Pubs, Patents, Grants

Outputs: Pubs, Patents

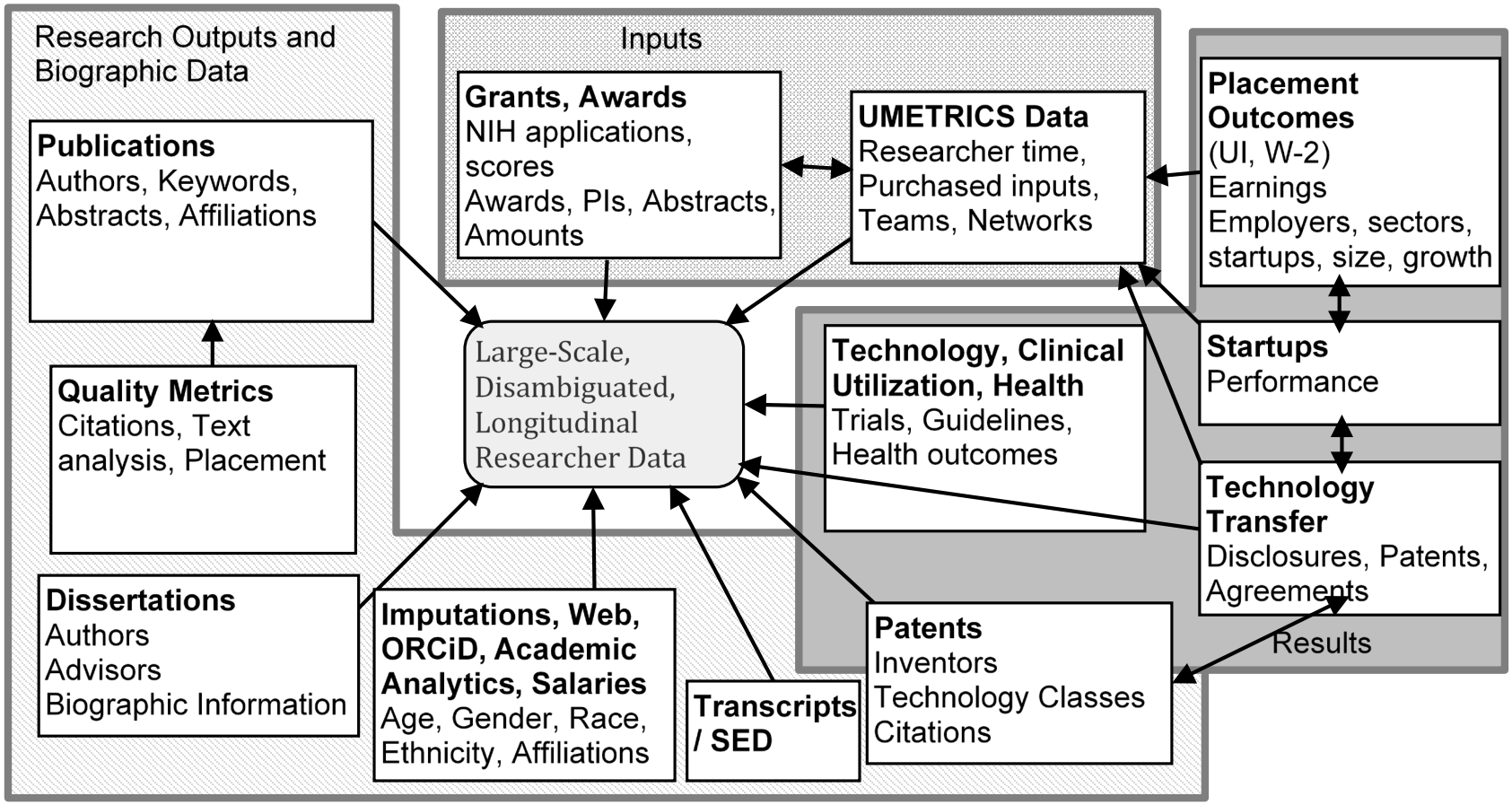
Placement, Outcomes, Value

Employment, Earnings: SDR, UI & Tax Docs

Scientific, Tech Output:
Publications, Patents, Grants,
SDR, Tech Transfer Records

Teams / Networks: UMETRICS, Pubs, Patents

Emerging Data Architecture



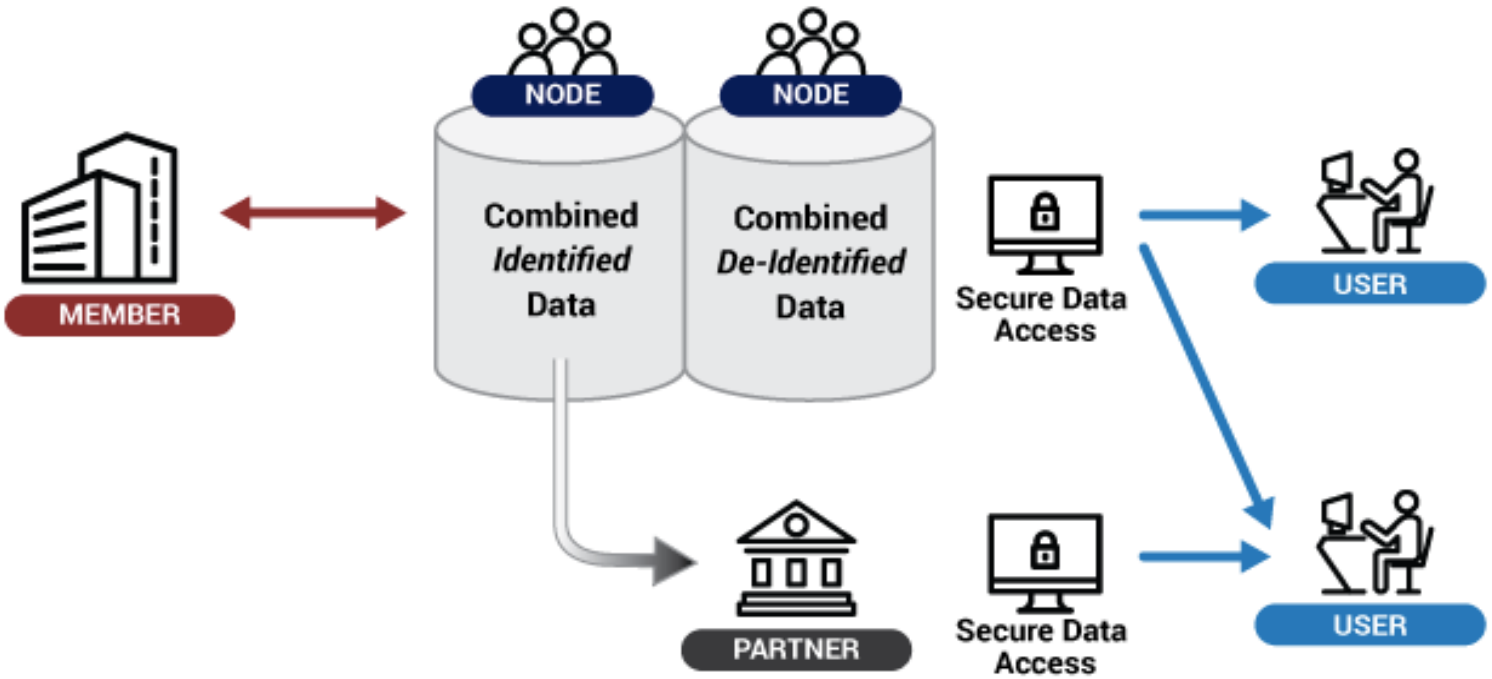
Issues

1. Data can be massive
   - 130K people named “Wang, Y” in PubMed!
2. Data can be poorly and/or inconsistently formatted
3. Ground truth can be hard to establish
4. Imputations
5. Different privacy / confidentiality / licensing issues

MEMBERS: Universities contribute data, support infrastructure and receive campus-specific and aggregate reports

NODES: Approved nodes materially improve data, develop products, and expand user communities

USERS: Approved users securely access de-identified aggregate datasets



PARTNERS: Approved partners receive data from IRIS which they improve and make accessible throught their own secure systems

What types of studies can be done?

1. Underrepresentation / intersectionality
   - Requires near population data to measure small, underrepresented groups
2. Match individuals to the industries and employers that need their knowledge
   - Help determine whether we are producing the right number / mix of researchers
   - And where they are valued

Link Data…

Link Data… Especially Different Types of Data!