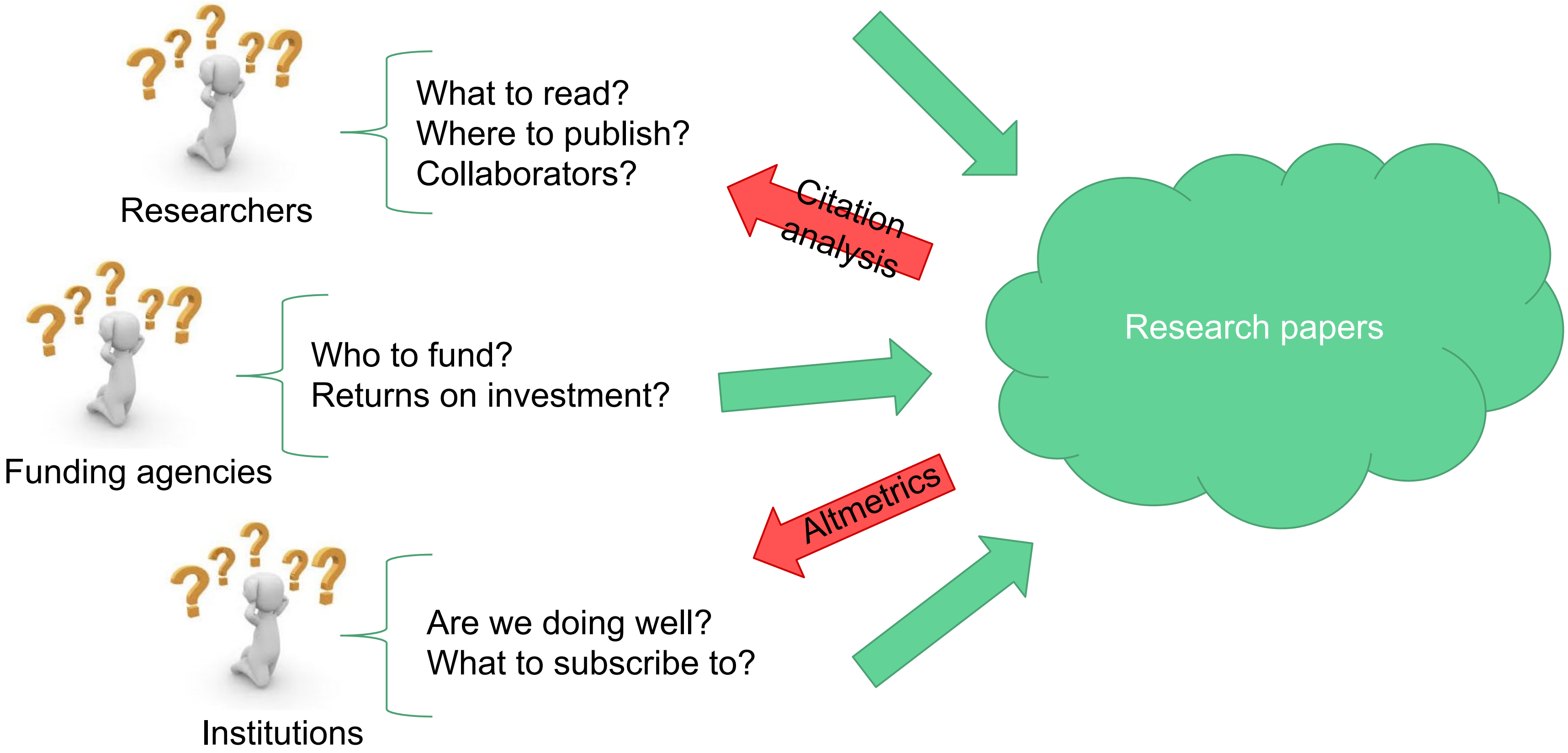
Do Citations and Readership Predict Excellent Publications?

Dasha Herrmannova, The Open University, UK
Robert Patton, Oak Ridge National Laboratory, USA
Petr Knoth, The Open University, UK
Chris Stahl, Oak Ridge National Laboratory, USA

Research question

Are current research evaluation metrics sufficient for identifying highly influential papers?

Why care about metrics?



Researchers

What to read?
Where to publish?
Collaborators?

Funding agencies

Who to fund?
Returns on investment?

Institutions

Are we doing well?
What to subscribe to?

Citation analysis

Altmetrics

Research papers

Finding what works

- ML approach
  - Evaluate all methods in terms of precision-recall/accuracy/...
  - Requirement: **ground truth**
- Research evaluation
  - No ground truth
  - Authority often established axiomatically
  - JIF, h-index, etc.
- Can we build a ground truth dataset?

Our understanding of "impact"

Low impact

VS

High impact

Our understanding of "impact"

Low impact

Survey papers:

"A general view, examination or description of someone or something"

VS

High impact

Seminal works:

"Strongly influencing later developments"

Creating a dataset

- Online questionnaire
  - Discipline?
  - Reference to a survey paper
  - Reference to a seminal paper
- Collected 314 papers
  - Labels (seminal, survey)
  - Title, authors, year of publication, abstract, DOI, ...
- Available online
  - http://trueid.semantometrics.org
- Analysis
  - Seminal papers on average 10 years older
  - Seminal papers cited on average 5 times more

Do citations/readership predict excellent papers?

- Classify papers using citations and readership as features
- Model
  - Select a threshold *t*
  - If *cit(d)* ≥ *t* → label as seminal
  - Else → label as survey
  - Use threshold with best accuracy on the training set
- Leave-one-out cross-validation
- 3 experiments
  - Aggregate
  - Per discipline
  - Per year

Results

Model	Data	Accuracy	Upper bound
Baseline	Citations	-	52.87%
	Readership	-	52.87%
Aggregate	Citations	63.06%	63.38%
	Readership	42.68%	52.87%
Discipline based	Citations	45.28%	68.11%
	Readership	42.13%	62.60%
Year based	Citations	55.23%	68.62%
	Readership	51.05%	65.27%

Conclusion

- Both citations and readership provide an improvement over the baseline
- Neither of the two metrics is optimal

What next?

- Ideal dataset
  - Multi-disciplinary
  - Time span
  - Publication types
  - Peer review judgement
- Better metrics
  - Citation context
  - Analyzing content

Thank you!

Questions?

http://trueid.semantometrics.org