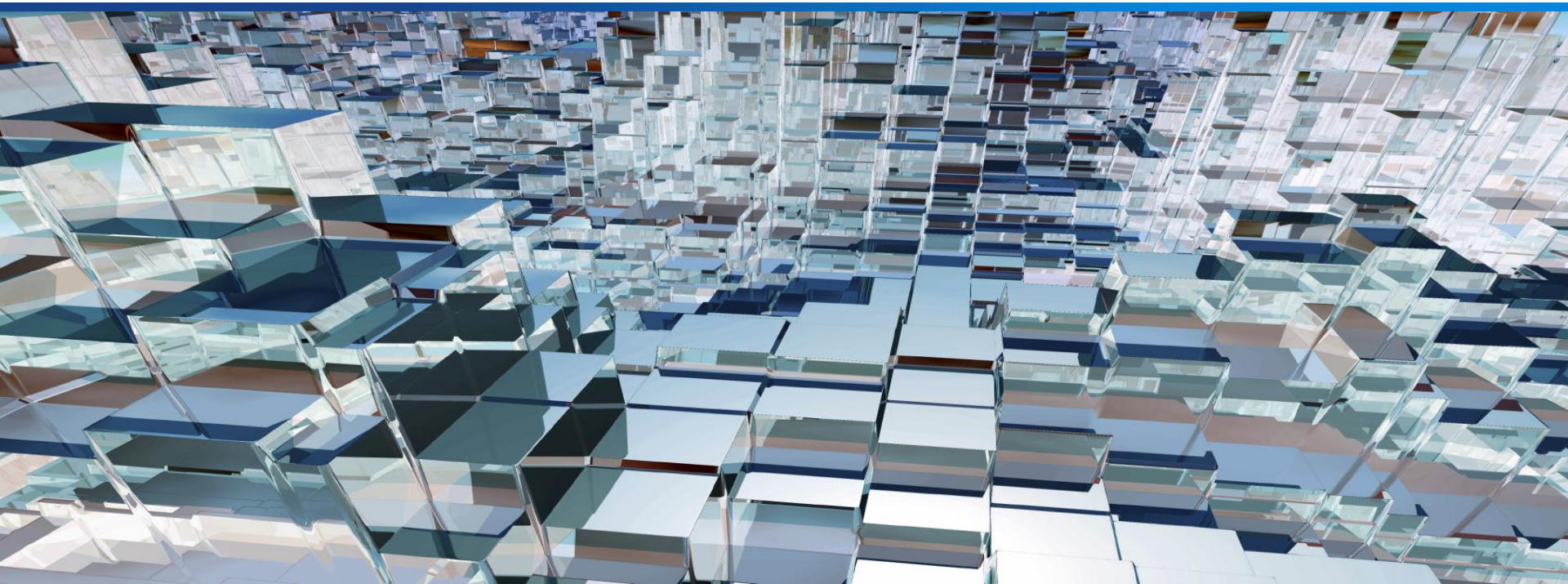
Science-Metrix
2002 – 2017

# Influence of OA, Gender, Co-authorship on Citation

## Science & Technology Indicators 2017 (Paris)

September 6–8, 2017



# Obligatory warm-up slide

- Order of operations for today:
  - Context
  - Methodology
  - Analysis
  - Some notes on ongoing follow-up work
  - Discussion
  - Conclusions, wild speculation

- Your divided attention:

**@DBStruck**  **@ScienceMetrix**

# Study context

- Citations used to evaluate impact of research (impact w/in academic community)
- Citation rates normalized by year, subfield, doc type to provide a level playing field
- Other parameters affect citation rates as well:
  - International collaboration
  - OA citation advantage (maybe?)
  - Gender
- These parameters inter-related as well:
  - International collabs more likely to be available in OA.
  - Women participate less often in international collab than men do.
- This project aims to disentangle this knot.

# Study methodology

- Data sources:
  - Scopus
  - 1science
  - NamSor API
- Scoping out a sample:
  - NamSor very reliable for US context
  - 1science database also covers US very well (~95% recall)
  - Citation windows + OA backfilling effect exclude most recent years
  - Not all subfields have similar gender composition, so gender dynamics might vary between them

# Study methodology

- Sample selected:
  - US publications (including their co-authors abroad)
  - Only pubs where all authors tagged by NamSor w/100% confidence
  - Publication year 2010 (follow-up work covers 2010–2012)
  - Subfields: Developmental Bio, Cardio. System & Hematology (follow-up work covers all subfields, except under Arts & Hum.)
- Filtering and bias:
  - Each of the two subfields represented ~2% of initial article pop. (US).
  - After filtering for NamSor, 1science coverage: 1.9% of population.
  - Dev. Bio sample: 3 000 papers, 32% of initial pub count
  - Cardio. System & Hema. sample: 3 500 papers, 36% of initial count

# Study methodology

- Coding variables, for each paper:
  - Citations: log (cite + 1), *pace* Thelwall & Wilson 2014
  - OA status: binary (no gold/green)
  - Gender:
    - Any women involved in team: binary
    - Share of women in team: scalar [0–1]
    - Lead author is female: binary
  - Collaboration
    - Number of authors: scalar [1–∞[ (actual max ≈ 20)
    - Number of institutions: scalar [1–∞[ (actual max ≈ 20)
    - Number of countries: scalar [1–∞[ (actual max ≈ 10)
    - International collaboration: binary

# Analysis—Developmental biology

- Inter-correlations between variables ($R^2$):

| DEV_BIO | OA | Female$ | Prop_F | Lead_F$ | n_authors | n_addresse | n_countrie | Internat$ | nb_cit | log(cite) |
|---|---|---|---|---|---|---|---|---|---|---|
| OA | 100% | 2% | 0% | 0% | 4% | 3% | 1% | 1% | 0% | 10% |
| Female$ | 2% | 100% | 67% | 25% | 16% | 5% | 2% | 2% | 0% | 2% |
| Prop_F | 0% | 67% | 100% | 47% | 2% | 0% | 0% | 0% | 0% | 0% |
| Lead_F$ | 0% | 25% | 47% | 100% | 0% | 0% | 0% | 0% | 0% | 0% |
| n_authors | 4% | 16% | 2% | 0% | 100% | 43% | 23% | 17% | 0% | 7% |
| n_addresses | 3% | 5% | 0% | 0% | 43% | 100% | 32% | 25% | 1% | 6% |
| n_countries | 1% | 2% | 0% | 0% | 23% | 32% | 100% | 80% | 0% | 3% |
| Internat$ | 1% | 2% | 0% | 0% | 17% | 25% | 80% | 100% | 0% | 3% |
| nb_cit | 0% | 0% | 0% | 0% | 0% | 1% | 0% | 0% | 100% | 30% |
| log(cite) | 10% | 2% | 0% | 0% | 7% | 6% | 3% | 3% | 30% | 100% |

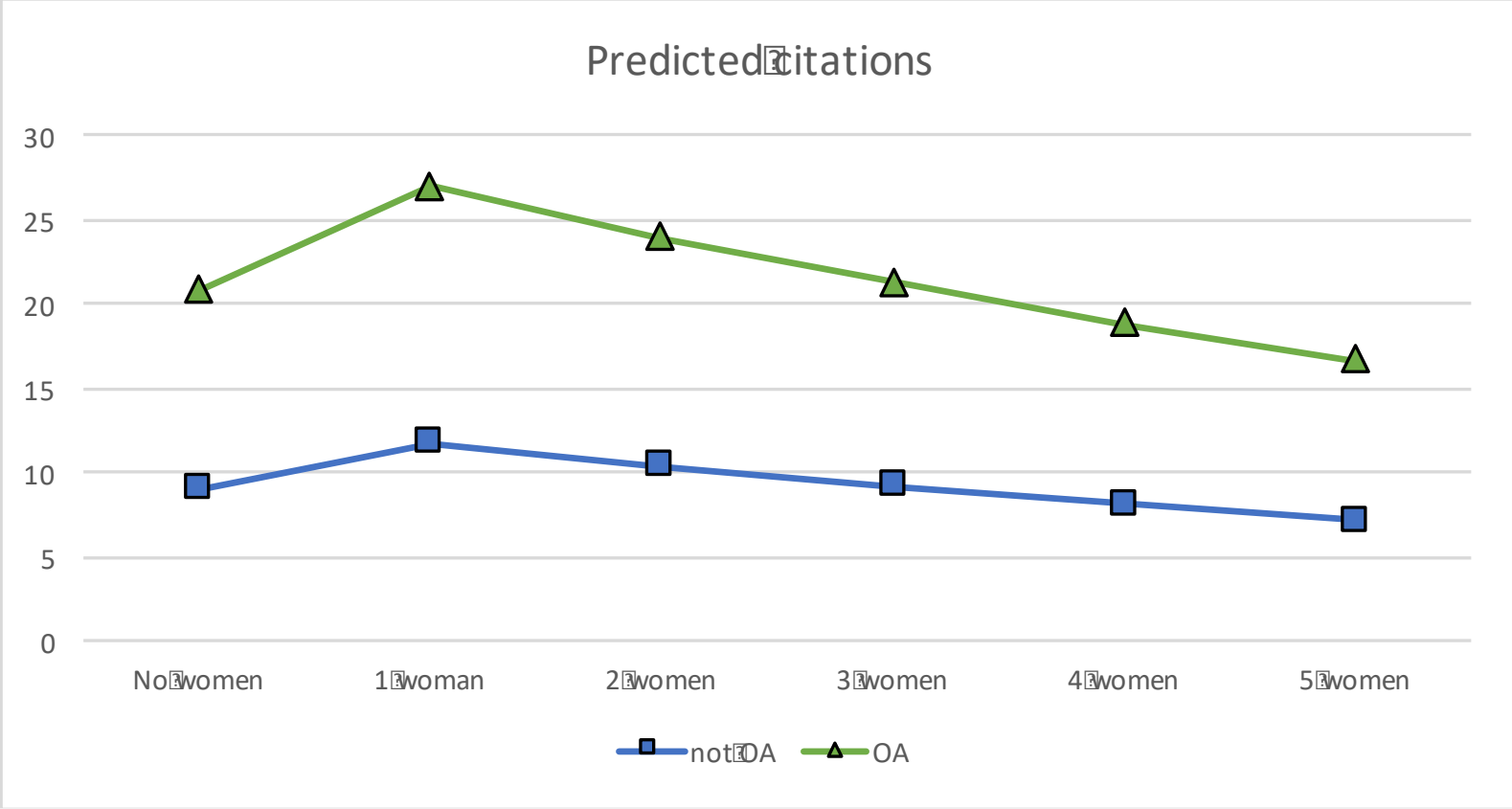
# Analysis—Developmental biology

- Coefficient tables:

| Variable | Type | Coefficient | Stat. significance |
|---|---|---|---|
| Open Access (OA) status | Binary | 0.341 | p<0.001 |
| Women involved in authorship | Binary | 0.158 | p<0.001 |
| Woman as corresponding author | Binary | **not significant** | |
| Proportion of women in research team | Scalar | -0.249 | p<0.001 |
| Number of authors | Scalar | 0.034 | p<0.001 |
| Number of institutions | Scalar | 0.040 | p<0.001 |
| Number of countries | Scalar | **not significant** | |
| International co-authorship | Binary | **not significant** | |
| Model constant | *n/a* | 0.749 | p<0.001 |
| Model overall | *n/a* | *n/a* | p<0.001 |

# Analysis—Developmental biology

- Predicted citation scores (5 authors, 2 institutions):



# Analysis—Cardiovascular systems & Hematology

- Inter-correlations between variables ($R^2$):

| CARDIO | OA | Female$ | Prop_F | Lead_F$ | n_authors | n_addresse | n_countrie | Internat$ | nb_cit | log(cite) |
|---|---|---|---|---|---|---|---|---|---|---|
| OA | 100% | 2% | 0% | 0% | 3% | 2% | 1% | 1% | 5% | 11% |
| Female$ | 2% | 100% | 62% | 20% | 18% | 8% | 2% | 3% | 1% | 3% |
| Prop_F | 0% | 62% | 100% | 40% | 1% | 0% | 0% | 0% | 0% | 1% |
| Lead_F$ | 0% | 20% | 40% | 100% | 0% | 0% | 0% | 0% | 0% | 0% |
| n_authors | 3% | 18% | 1% | 0% | 100% | 47% | 18% | 13% | 5% | 10% |
| n_addresses | 2% | 8% | 0% | 0% | 47% | 100% | 28% | 18% | 5% | 7% |
| n_countries | 1% | 2% | 0% | 0% | 18% | 28% | 100% | 71% | 3% | 4% |
| Internat$ | 1% | 3% | 0% | 0% | 13% | 18% | 71% | 100% | 2% | 3% |
| nb_cit | 5% | 1% | 0% | 0% | 5% | 5% | 3% | 2% | 100% | 53% |
| log(cite) | 11% | 3% | 1% | 0% | 10% | 7% | 4% | 3% | 53% | 100% |

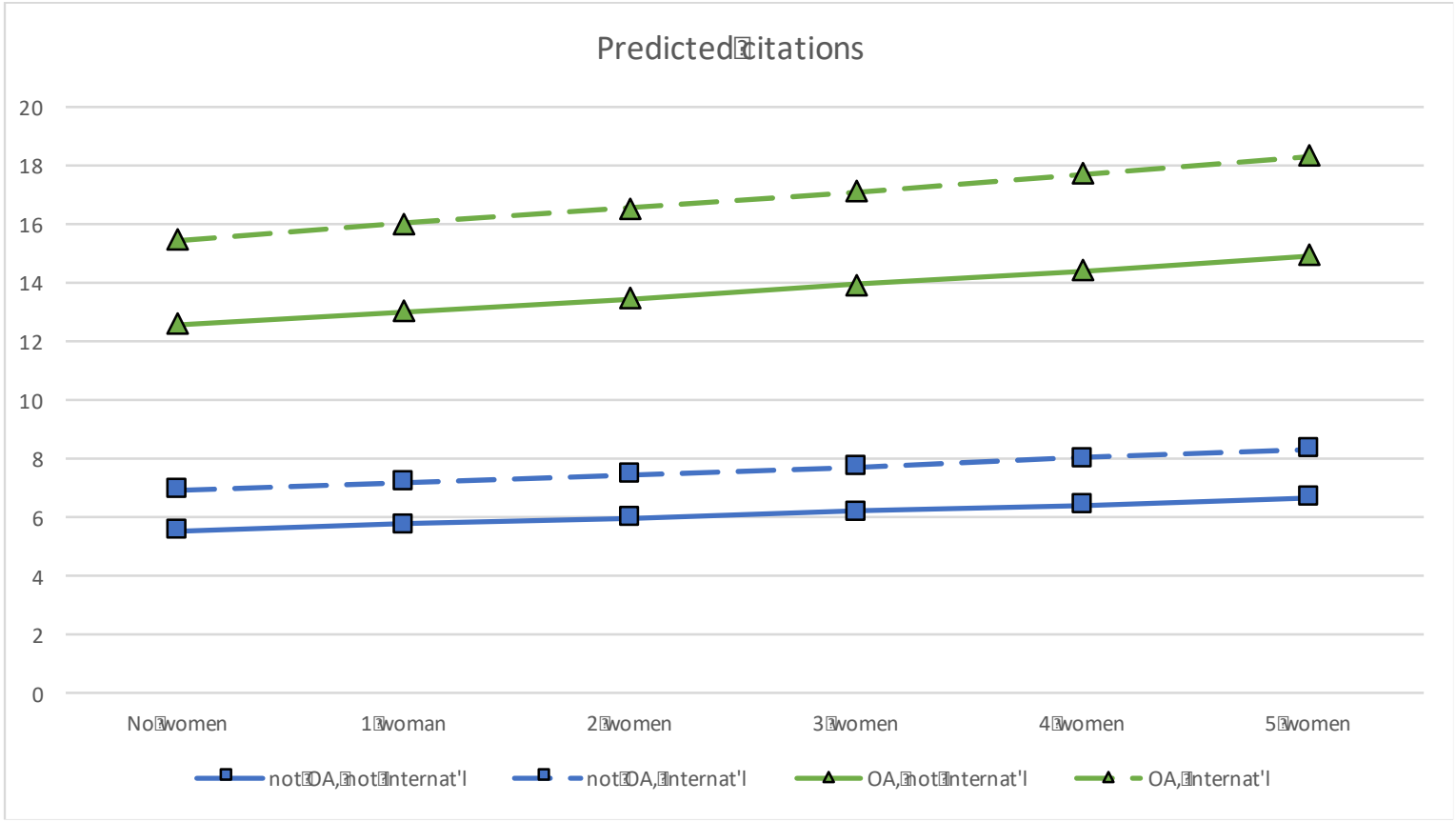
# Analysis—Cardiovascular systems & Hematology

- Coefficient tables:

| Variable | Type | Coefficient | Stat. significance |
|---|---|---|---|
| Open Access (OA) status | Binary | 0.317 | p<0.001 |
| Women involved in authorship | Binary | ***not significant*** | |
| Woman as corresponding author | Binary | ***not significant*** | |
| Proportion of women in research team | Scalar | 0.069 | p=0.033 |
| Number of authors | Scalar | 0.039 | p<0.001 |
| Number of institutions | Scalar | 0.022 | p<0.001 |
| Number of countries | Scalar | ***not significant*** | |
| International co-authorship | Binary | 0.084 | p<0.001 |
| Model constant | *n/a* | 0.576 | p<0.001 |
| Model overall | *n/a* | *n/a* | p<0.001 |

# Analysis—Cardiovascular systems & Hematology

- Predicted citation scores (5 authors, 2 institutions):



# Ongoing follow-up

- Looking at different modeling approaches:
  - Robust modeling: better suited to input and output variables that are non-normally distributed, which we know to be the case
  - Binning: looking at 5 bins for gender composition of research teams
  - Citation data as a category variable: trying to address challenges posed by non-normal distribution
- Applied models to all subfields, for 2010–2012, still US pubs
- Prepping data to include number of pubs from researchers & institutions involved, to figure out how much of a role these play in determining citation outcomes.

# Discussion

- OA associated with higher citation scores—though reasons still not clear!
  - Selection bias?
  - Early availability bias?
  - Prestigious institutions/researchers having funds for APCs?
- International collaboration promotes higher citation scores; best parametrised as binary variable, not scalar.
- Larger number of authors & institutions promotes citation.
- Mixed-gender teams—leaning male—seem to be optimal for promoting citations.
- Gender of lead author does not seem to have an effect.

# Discussion

- OA and international collab advantages robust across areas of research, and models. OA > international collab.
- Gender dynamics quite even across domains using approach presented today; but across subfields and across models their impact is much less consistent.
- All the models have strong statistical significance ($p<0.001$), but low fit ($R^2 \approx 0.15$): good predictor of aggregate results.
- Potential concern: with a low $R^2$, models might be picking up on different underlying patterns.

# Conclusions, wild speculation

- Citation scores partially determined by OA and international collaboration, and (seemingly) gender balance of research teams.
- Influence of each is independent of the others.
- Each can be considered a strategy for increasing citation.
- If citation scores supposed to measure quality or "excellence" of research content—distinct from visibility or uptake—should we be normalising for these strategies when assessing excellence?
- Would be interesting to inspect effect of these strategies on quality and dissemination, independent of each other.
- Ultimately, is citation-based evaluation primarily about quality or uptake? This should guide testing, interpretation.

# Shameless pitch

- If you liked the presentation, consider following on Twitter.

**@DBStruck**  **@ScienceMetrix**

- We also blog about bibliometrics, data mining and science policy at ScienceMetrics.org: check it out, sign up!

# Contact information


**CONTACT**

**Brooke Struck**

brooke.struck@science-metrix.com

**Guillaume Roberge**

guillaume.roberge@science-metrix.com

**Matthew Durning**

matt.durning@science-metrix.com

**David Campbell**

david.campbell@science-metrix.com


**Thank you!**


**Science-Metrix**

Montréal – Ottawa – Washington

**WEBSITE**

**www.science-metrix.com**

**PHONE**

1.514.495.6505

1.800.994.4761