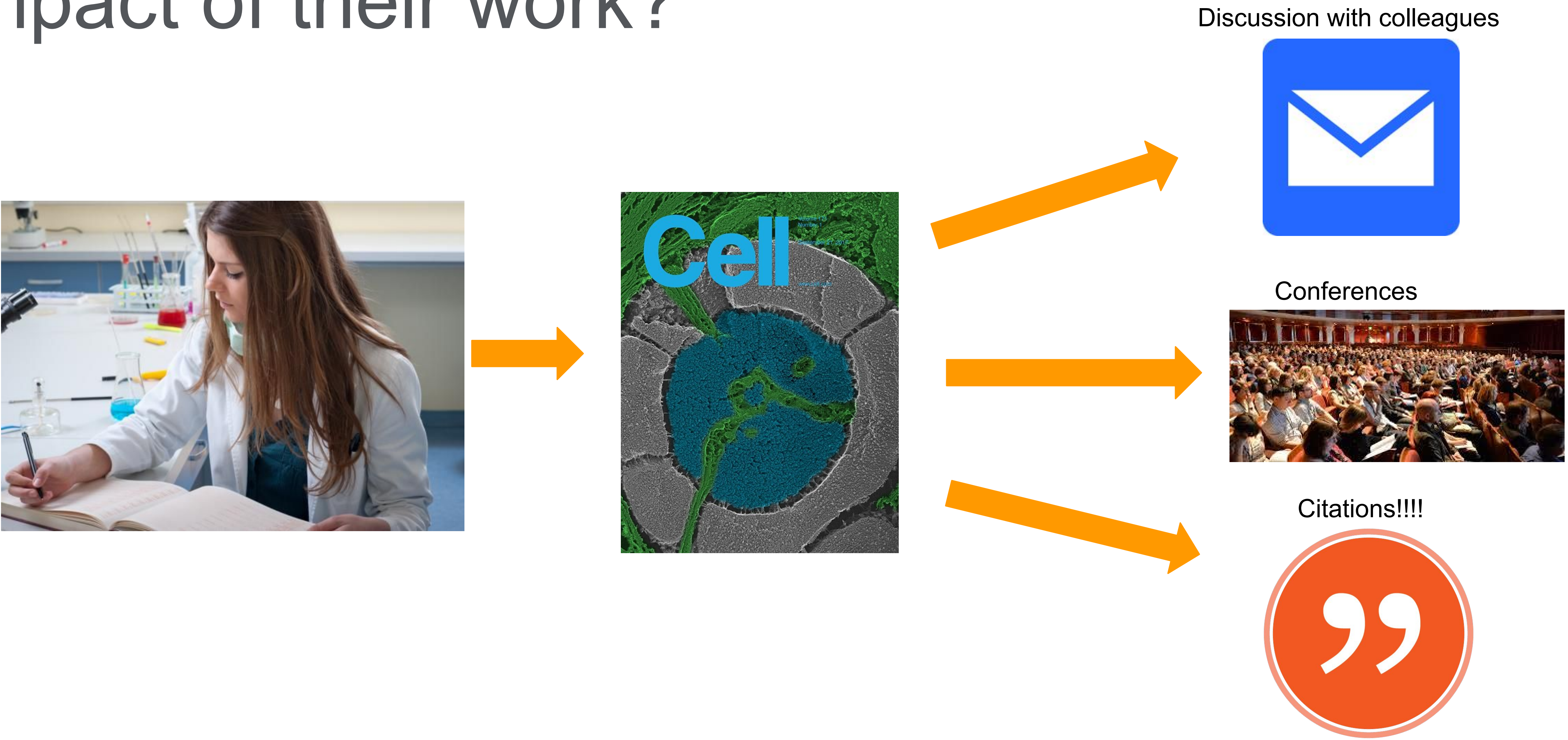
Tracing the flow of knowledge using Pyspark

Elsevier Labs
Pygotham 2017

Jessica Cox
j.cox@elsevier.com
@jessienapcox

Corey Harper
c.harper@elsevier.com
@chrpr

How do scientists evaluate the impact of their work?



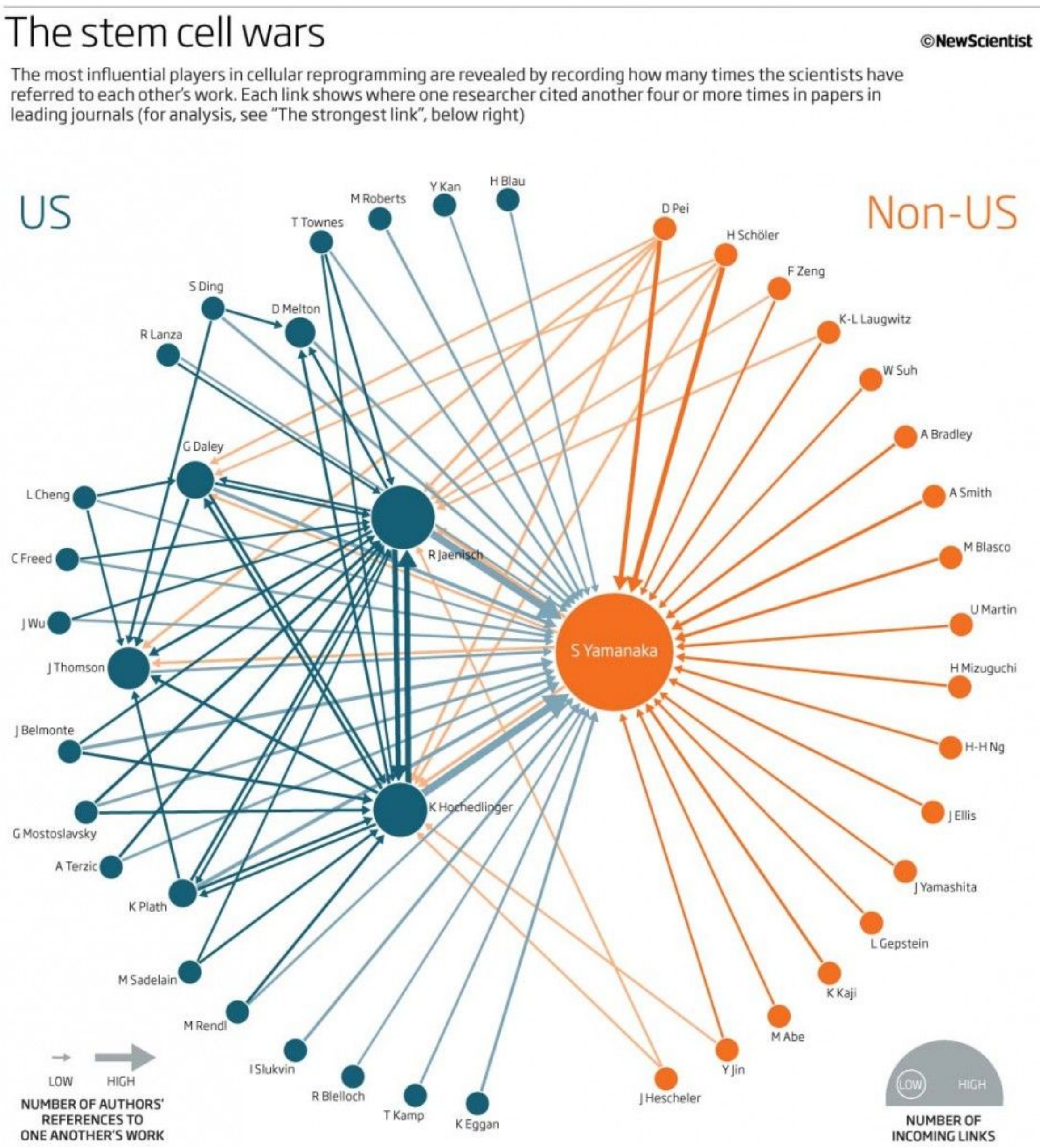
Discussion with colleagues

Conferences

Citations!!!!

Citation networks

Example network is constructed based on how many times authors in the field cite one another.



The stem cell wars

©NewScientist

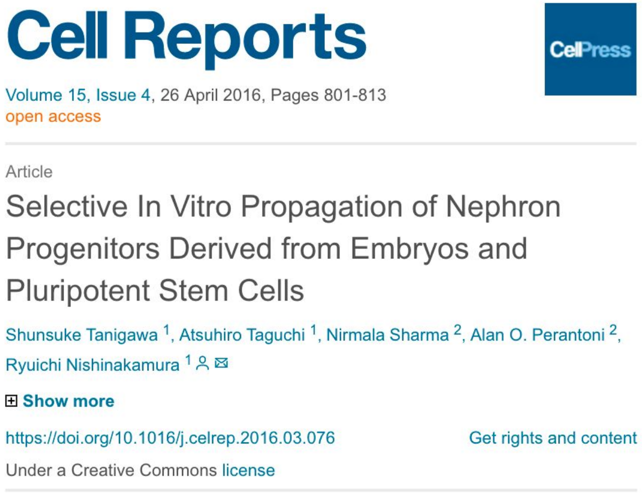
The most influential players in cellular reprogramming are revealed by recording how many times the scientists have referred to each other's work. Each link shows where one researcher cited another four or more times in papers in leading journals (for analysis, see "The strongest link", below right)

New Scientist. Inside the stem cell wars. 2010.

Differences in citation language

Materials and Methods

Human nephron progenitors were induced from iPSCs (201B7) (Takahashi and Yamanaka, 2006), based on the protocol that we previously established (Taguchi et al., 2014).



Cell Reports

Volume 15, Issue 4, 26 April 2016, Pages 801-813

open access

Article

Selective In Vitro Propagation of Nephron Progenitors Derived from Embryos and Pluripotent Stem Cells

Shunsuke Tanigawa [1], Atsuhiro Taguchi [1], Nirmala Sharma [2], Alan O. Perantoni [2], Ryuichi Nishinakamura [1]

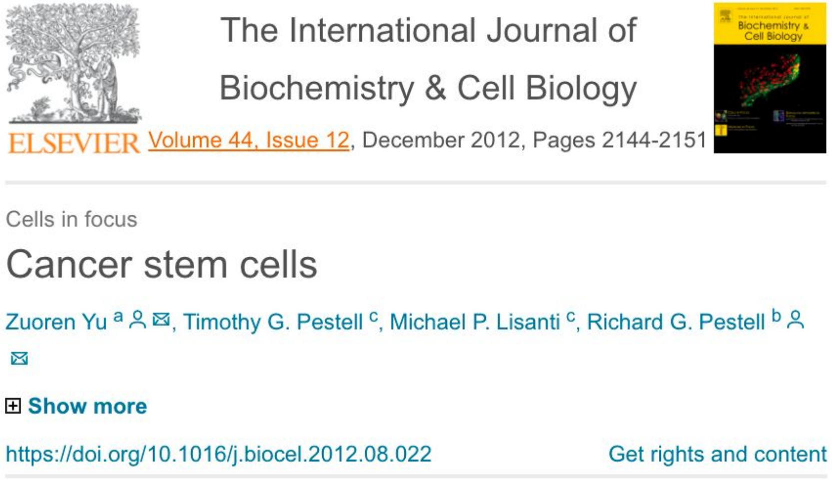
⊞ Show more

https://doi.org/10.1016/j.celrep.2016.03.076 Get rights and content

Under a Creative Commons license

Introduction

Researchers have successfully reprogrammed somatic cells into stem-like cells – known as induced pluripotent stem cells (iPSCs) – which share many of the characteristics of ESCs (Takahashi and Yamanaka, 2006).



The International Journal of Biochemistry & Cell Biology

ELSEVIER Volume 44, Issue 12, December 2012, Pages 2144-2151

Cells in focus

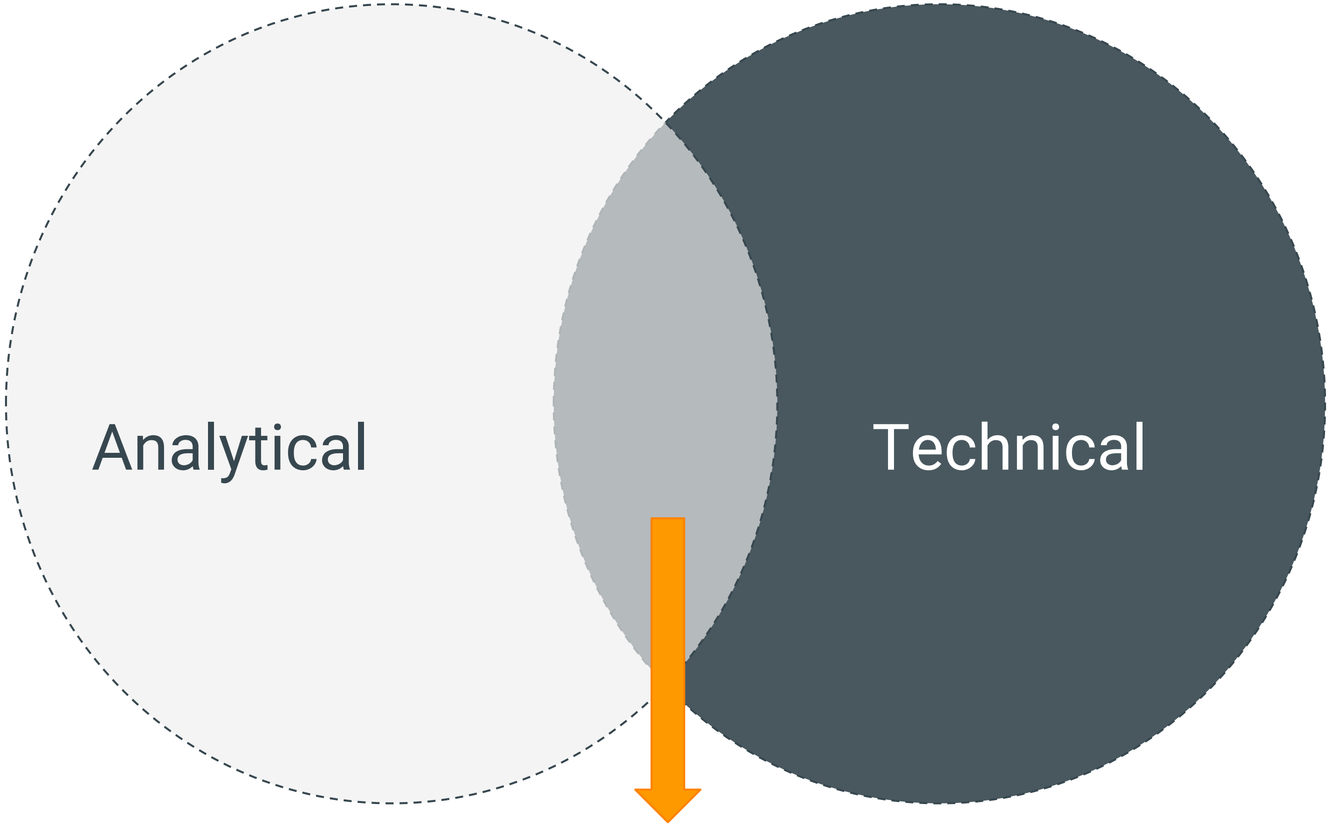
Cancer stem cells

Zuoren Yu [a], Timothy G. Pestell [c], Michael P. Lisanti [c], Richard G. Pestell [b]

⊞ Show more

https://doi.org/10.1016/j.biocel.2012.08.022 Get rights and content

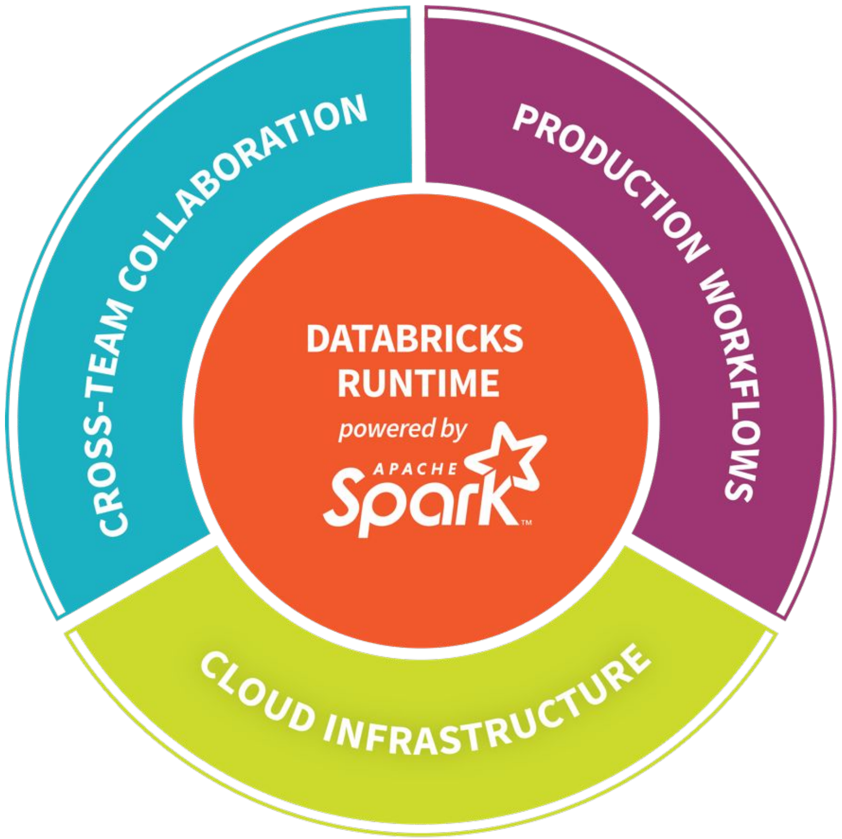
Goal and Motivation



NLP of citation data in a spark environment

Databricks

- Built on top of Apache Spark
- Allows for cross-team collaboration
- Cloud infrastructure



databricks™

COMMUNITY EDITION

Learn Apache Spark

- Mini 6GB cluster
- Interactive notebooks and dashboards
- Public environment to share your work

Spark architecture

- Driver / Executor Env.
  - Driver distributes work to executors
  - Executors load data
  - Extends Map/Reduce

- Delayed execution
  - Transformations & Actions
  - Optimized execution plan
  - Concept of *pipelines*

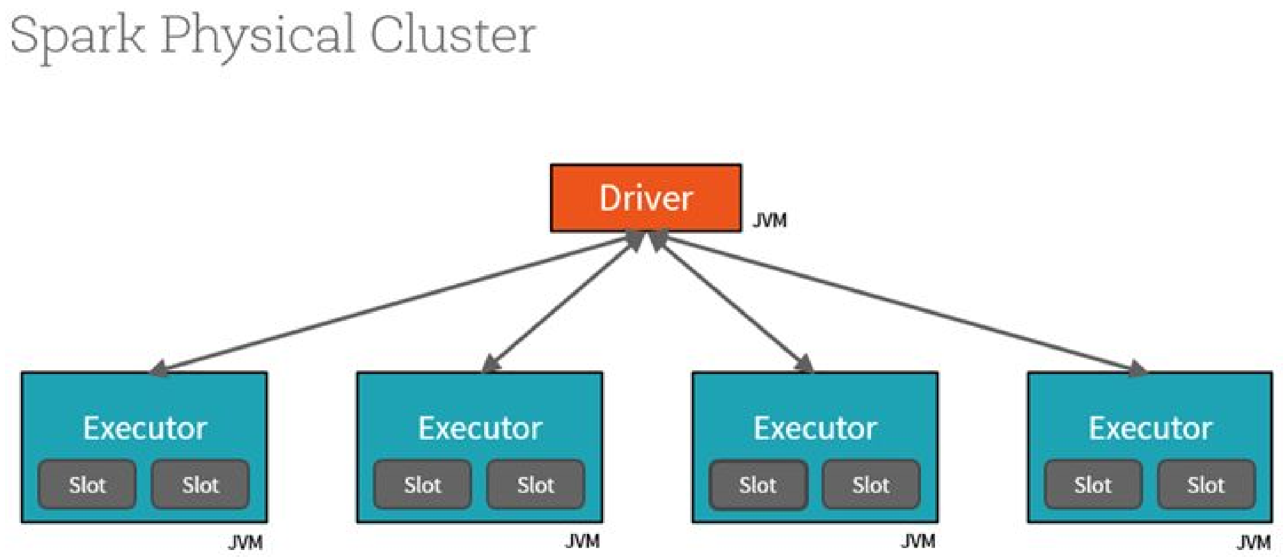
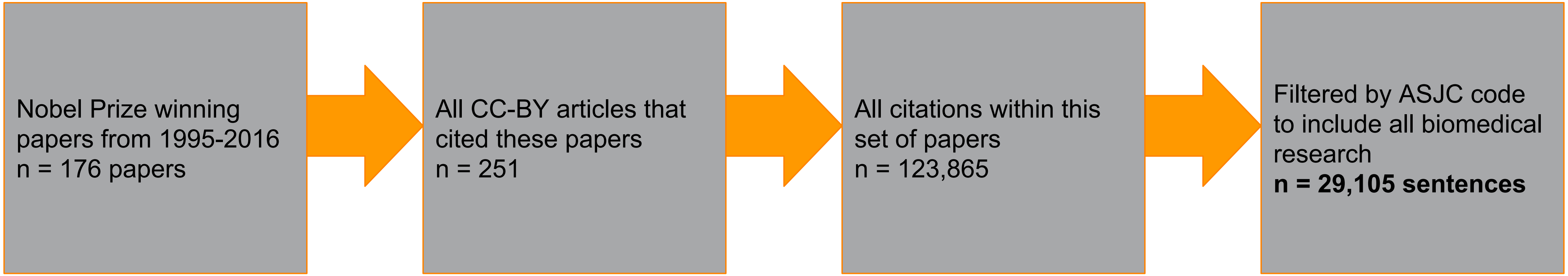


Image from :https://docs.databricks.com/spark/latest/gentle-introduction/gentle-intro.html

Sample Corpus: Nobel Prize Winners



Nobel Prize winning papers from 1995-2016
n = 176 papers

All CC-BY articles that cited these papers
n = 251

All citations within this set of papers
n = 123,865

Filtered by ASJC code to include all biomedical research
n = 29,105 sentences

Features

- ScopusIds
- ASJC (All Science Journal Classification) codes of the citing documents https://github.com/plreyes/Scopus
- Age of the citation
- Section title that citation occurred within
- Article type
  - Review
  - Original
  - Conference paper
  - Etc.
- Text
  - Sentence with the citation
  - Previous sentence
  - Next sentence

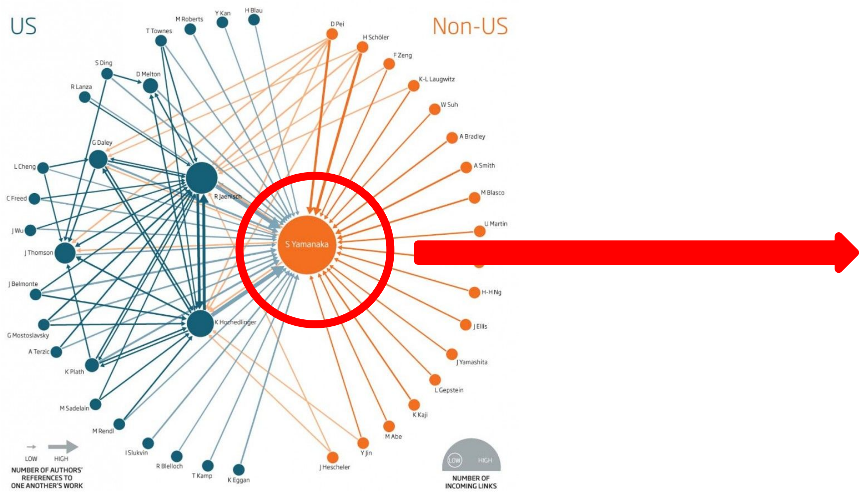
Notebooks

What about that original paper?

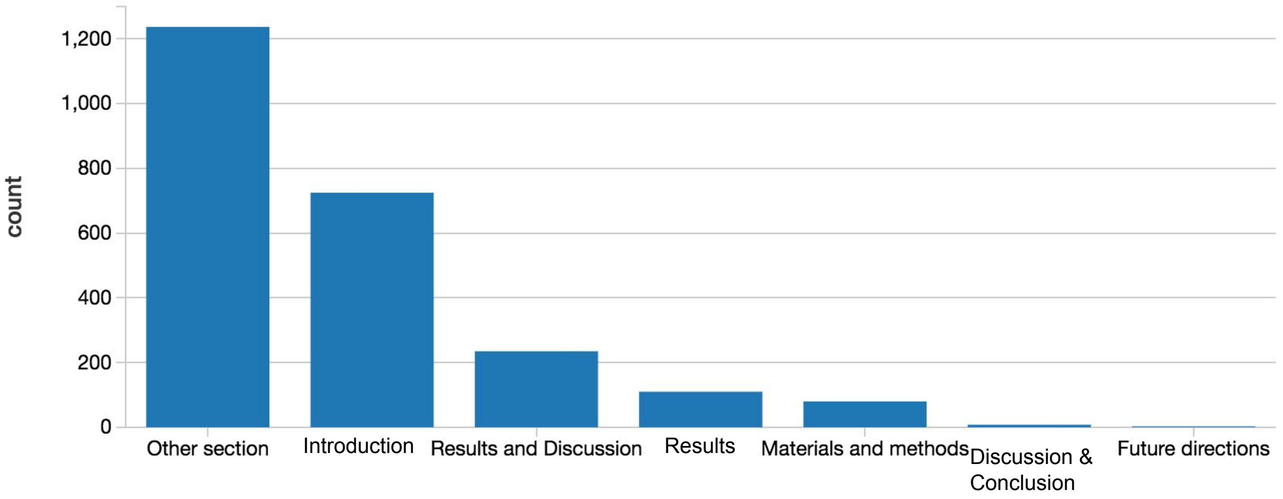
The stem cell wars

©NewScientist

The most influential players in cellular reprogramming are revealed by recording how many times the scientists have referred to each other's work. Each link shows where one researcher cited another four or more times in papers in leading journals (for analysis, see "The strongest link", below right)



We collected 2,398 sentences that cited the 2006 Yamanaka paper and performed the same analysis...



Section Title	Most frequent	Bigrams	Trigrams
Materials and Methods (n = 80)	performed	previously described	bisulfite treatment dna
Introduction (n = 725)	pluripotent	pluripotent stem	induced pluripotent stem
Results & Discussion (n = 235)	studies	stem cells	pluripotent stem cells

Conclusions

- There is more to consider than just number of times a work is cited
- NLP allows us to understand ***how*** and ***why*** work is being discussed
- Databricks and PySpark allow us to assess thousands of sentences quickly for language patterns

Future Directions

- Visualizing topic clusters to group similar uses of papers
- Using neural network techniques like word2vec and sense2vec
- Using part-of-speech parsing and tagging to look for grammatical patterns in citations

- Other applications
  - Analysis of language used in facebook posts with links
  - Categorize papers by use type

Takehomes

- Notebooks are available to run on the community edition of Databricks
- Dataset in CSV format
- Link to archived version of notebook
- Slides

http://dx.doi.org/10.17632/8kyckg3dh5.1

Mendeley Datasets

“Pygotham 2017” Jessica Cox & Corey Harper

Thank you!

Jessica Cox
j.cox@elsevier.com
@jessienapcox

Corey Harper
c.harper@elsevier.com
@chrpr