Data science: з печі до столу

Vsevolod Solovyov
CTO at Prophy Science
@murkt

Область:

- Существует 100+ млн научных публикаций
- Население растёт, количество учёных - тоже
- С каждым годом учёные пишут всё больше статей
- А ещё патенты, техническая документация, клинические исследования...

Что нужно людям?

- Поиск
- Рекомендации нового
- Поиск экспертов
- Охватить неизвестную тему

Кто клиенты?

- Грантовые агентства
- Научные издательства
- Фармакологические компании
- Исследователи
- Internal documentation hell

Научная публикация

- Заголовок, текст
- Авторы
- Ссылки на предыдущие работы
- Место публикации (журнал, конференция)

Простой текст неудобен

Все любят сокращать

- We study the impact of a warm dark matter (WDM) cosmology on dwarf galaxy formation...
- ...in both CDM and WDM models. WDM halos ...
- ...the most massive WDM galaxy (Halo m10k) collapses...
- ...their CDM counterparts, as can be seen by comparing the colored lines (WDM)...

Все любят сокращать

Warm dark matter	3
WDM	219

Обычный стемминг

- polar, polars, polarize, polarized, polarizations → polar
- GAN, GaN → gan
- AND → and
- anyone, anyon → anyon
- WDMS, WDM → wdm

Поищем в тексте

Alpha thalassemia	2236
Alpha thalassaemia	748
α thalassemia	1046
α thalassaemia	276
SUM(1 + 2 + 3 + 4)	4306
OR(1 + 2 + 3 + 4)	4074

Решение

- Специализированный стеммер
- Стоп-слова (anyone)
- Онтология (списки терминов)
  - Warm dark matter, *WDM*
  - Wavelength division multiplexer, *WDM*
  - Galaxy cluster, cluster of galaxies, *GC*, *cluster*

Расширение онтологии

- Keyphrase extraction
    - On the structure and oxygen transmission rate of biodegradable cellulose nanobarriers
- Synonym detection
    - octadecylphosphonic acid (OPA, ODPA)
    - convolutional LSTM (ConvLSTM, convolutional long short-term memory)
    - decoupled extended Kalman filter (DEKF, decoupled EKF)
    - human mesenchymal stem cells (human bone marrow stem cells, hMSC, HBMSC)

А автор кто?

Simon Brand
@TartanLlama

Follow

Programming difficulties:
Easy - multiplying 1,000,000 matrices in the blink of an eye
Medium - making sure that any memory allocated is freed correctly
Hard - checking if a picture is of a bird
Nightmare - handling dates
Ultra-Nightmare - handling people's names

12:52 AM - 27 Jul 2018

765 Retweets **2,276** Likes

52 765 2.3K

J Smith

- John Smith
- Jekyll E Smith
- Jekyll M Smith
- Jekyll EM Smith
- J David Smith

J Smith

- John Smith
- Jekyll E Smith
- Jekyll M Smith
- Jekyll EM Smith
- J David Smith
- Jane Smith-Krueger
- Jekyll Smith Jr.
- Jekyll Smith III

J Smith

- John Smith
- Jekyll E Smith
- Jekyll M Smith
- Jekyll EM Smith
- J David Smith
- Jane Smith-Krueger
- Jekyll Smith Jr.
- Jekyll Smith III
- JekillE Smith
- JEKYLL. E. SMITH

JL Smith, MB Salamon

Кто из:

- James L. Smith, Myron B. Salamon
- Janet L. Smith, Theodora Hatziioannou

JL Smith, MB Salamon

Кто из:

- James L. Smith, Myron B. Salamon
- Janet L. Smith, Theodora Hatziioannou
- Jerald L. Smith, Miriam Salamon

J Smith

- Больше 45 тысяч статей, подписанных каким-нибудь J Smith (в нашей базе)
- Это больше тысячи реальных людей
- Кто-то написал сотни статей, кто-то - одну-две

Ох, китайцы

Странный народ

Топ имен

Wei Wang	24000
Wei Zhang	20100
Wei Li	18400
Lei Zhang	15100
Lei Wang	14600
Jing Wang	14100
Yan Li	14000

Еще исключительные случаи

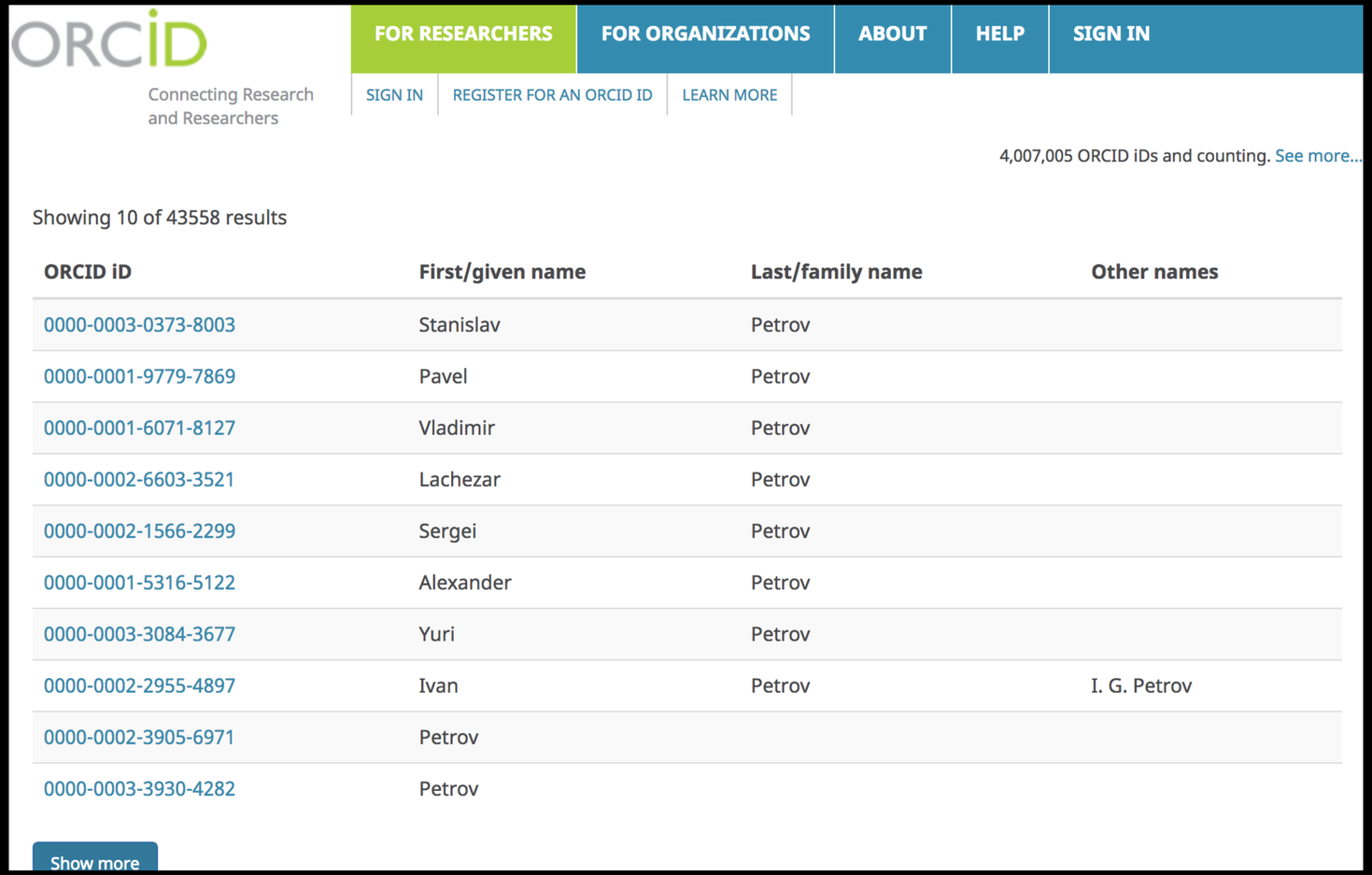
- Тысячи авторов в одной статье
- Два автора с одинаковым именем в статье
- Группы авторов
  - ATLAS collaboration
  - Investigators of the European Huntington's Disease Network

Идентификаторы спешат на помощь

- Специальные идентификаторы для научных авторов
  - ORCID – 0000-0001-8073-3068
  - ScopusID – 13204492100
  - ResearcherID – E-3698-2015
- Email

Специальные идентификаторы

- Редко указывают в статьях
- Не человекочитаемы
- Их путают в статьях
- В базах идентификаторов куча бреда



V Petrov'ы

stanislaV, paVel, Vladimir
Lachezar? Sergei?

Email

- Один емейл точно принадлежит одному человеку
- У одного человека может быть несколько
  - Gmail
  - Hotmail
  - Разные университеты

Rapid Antigen Processing and Presentation of a Protective and Immunodominant HLA-B*27-restricted Hepatitis C Virus-specific CD8+ T-cell Epitope

Julia Schmidt, [1,2,3] Astrid K. N. Iversen, [4,*] Stefan Tenzer, [5] Emma Gostick, [6] David A. Price, [6] Volker Lohmann, [7] Ute Distler, [5] Paul Bowness, [8,9] Hansjörg Schild, [5] Hubert E. Blum, [1] Paul Klenerman, [9] Christoph Neumann-Haefelin,[#1] and Robert Thimme[#1,*]

звёздочка стоит

Christopher M. Walker, Editor

[1] Department of Medicine II, University Hospital Freiburg, Freiburg, Germany,
[2] Faculty of Biology, University of Freiburg, Freiburg, Germany,
[3] Centre of Chronic Immunodeficiency, University of Freiburg, Freiburg, Germany,
[4] Nuffield Department of Clinical Neurosciences, Division of Clinical Neurology, Weatherall Institute of Molecular Medicine, Oxford University, Oxford, United Kingdom,
[5] Institute of Immunology, University Medical Center of the Johannes Gutenberg University of Mainz, Mainz, Germany,
[6] Institute of Infection and Immunity, Cardiff University School of Medicine, Cardiff, United Kingdom,
[7] Department of Infectious Diseases, University of Heidelberg, Heidelberg, Germany,
[8] Medical Research Council Human Immunology Unit, Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, Oxford, United Kingdom,
[9] Nuffield Department of Clinical Medicine, Oxford, United Kingdom,
Nationwide Children's Hospital, United States of America,
[#] Contributed equally.

AKNI RT

* E-mail: astrid.iversen@ndcn.ox.ac.uk (AKNI); robert.thimme@uniklinik-freiburg.de (RT)

The authors have declared that no competing interests exist.

Email принадлежит одному человеку?

- isrn.molecular.biology@hindawi.com
- microbiologia.clinica@unt.edu.ar
- gyn-sekretariat@pius-hospital.de
- secretariat@grangettes.ch

А что делают другие?

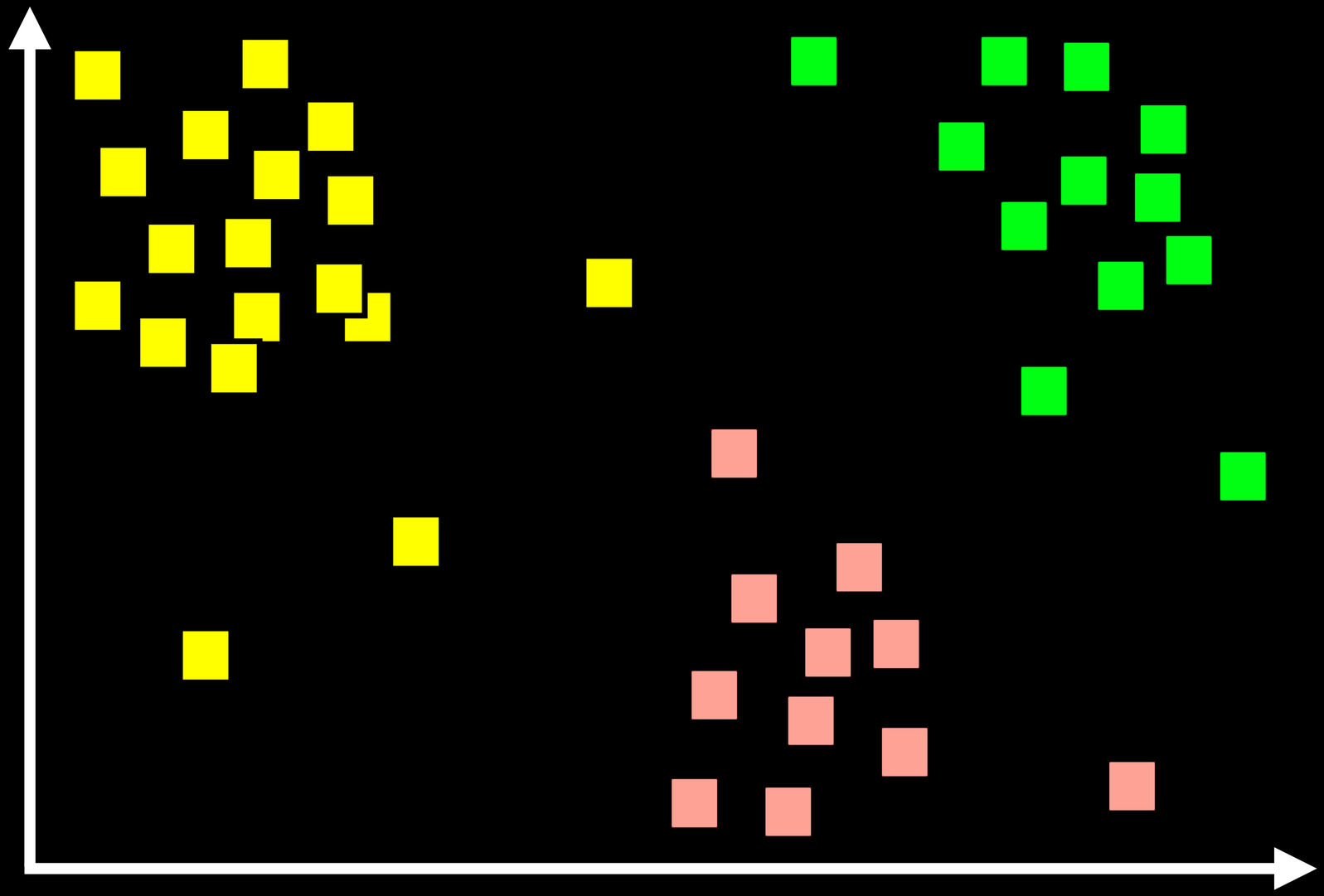
- Cited "V S Saxena" but the profile reads "Vikram S Saxena". This made it a tedious process to manually resolve out the conflicts.
- For **each pair** of publication records, we compute all basic features.
- Почти всегда разбирают заново

Какая информация полезна?

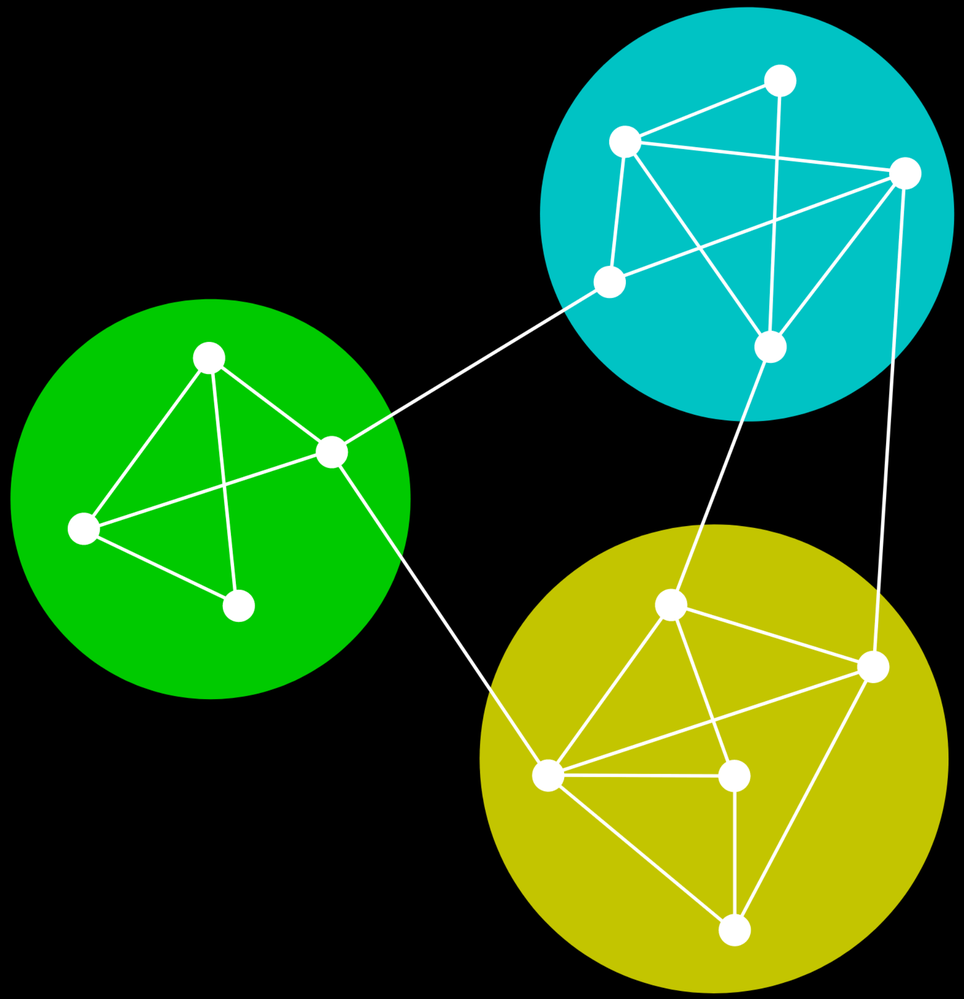
- Идентификаторы
- Ссылки на статьи
- Аффилиация (университет)
- Место публикации (журнал)
- Соавторы
- Имя
- Текст статьи

Кластеризация

- Пространственная
  - k-means
  - DBSCAN
  - ...
- Графовая
  - Modularity
  - Clique detection
  - ...

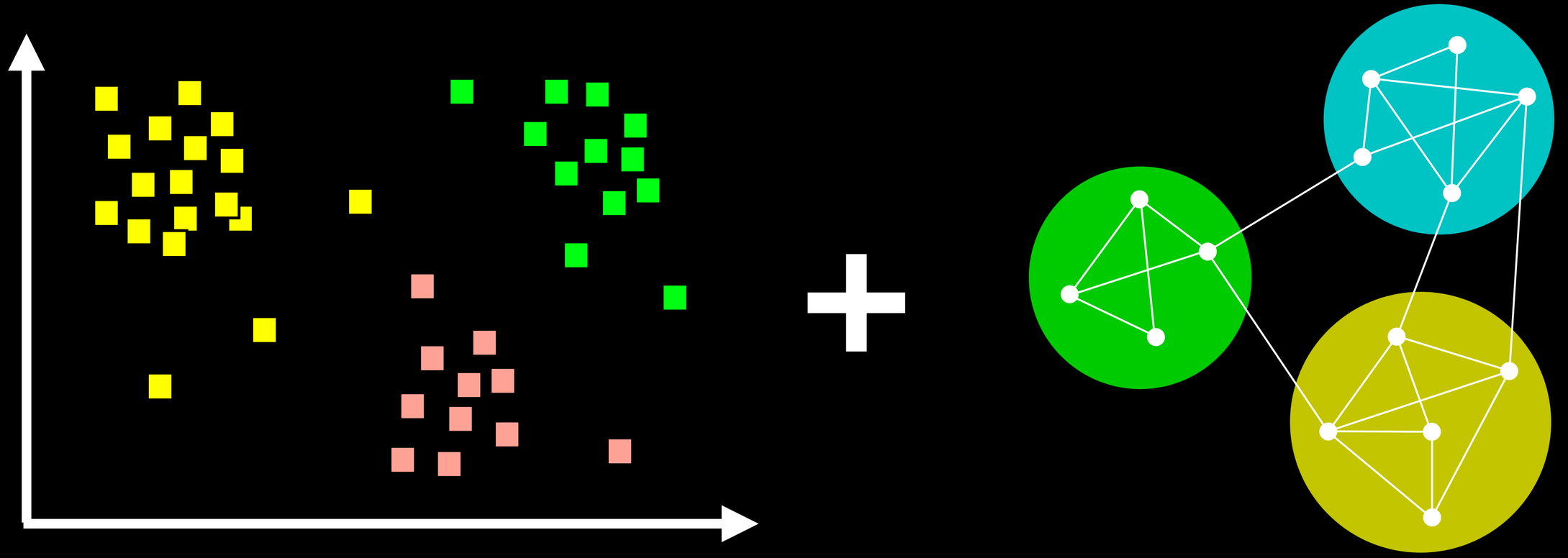


Пространственная кластеризация



Графовая кластеризация

Как совместить?

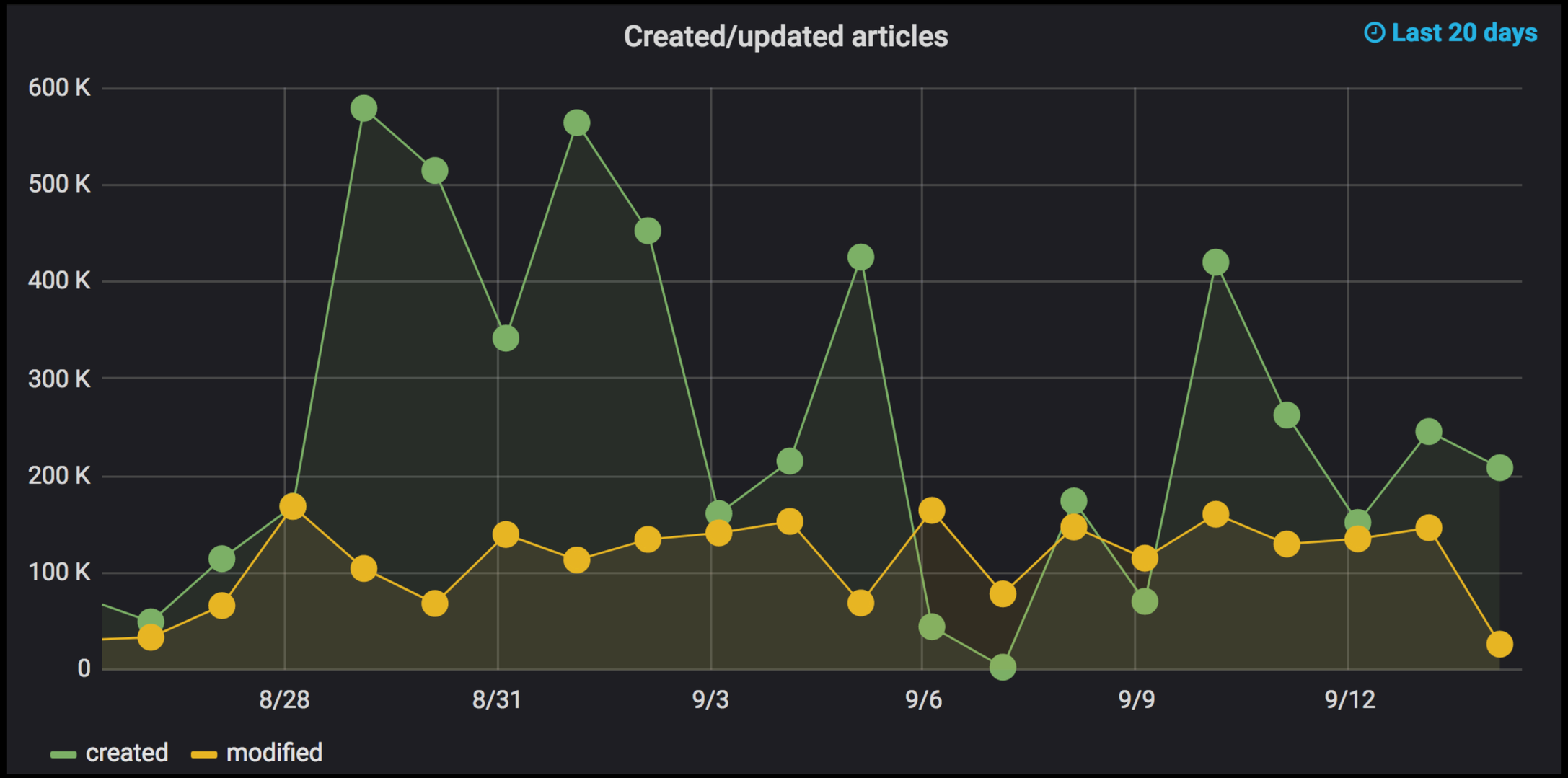


Как совместить?

- Расстояние/похожесть → ребра в графе
- Графы → координаты
  - Graph embedding
  - Metric learning

Реальность

- Не объединять Jurg и Jurgen в одного автора
- Но если e-mail или ORCID совпадает - объединять
- А если два разных ORCID - то не объединять
- Но...
- А если...
- И каждый день новые данные



Новые данные

Новые данные

- Результат должен быть стабилен
- Больше миллиона новых имен в день
- 100+ миллионов "статейных" авторов всего
- Наколенный скоринг с эвристиками

Выводы

- Знание данных бесценно
- Крайние случаи могут всё поломать
- Верить источникам нельзя
- Разнообразные хаки, чтоб работало

Фейсом об тейбл дривен девелопмент

Спасибо за внимание

Вопросы?

Vsevolod Solovyov
CTO at Prophy Science
@murkt
vsevolod.solovyov@gmail.com