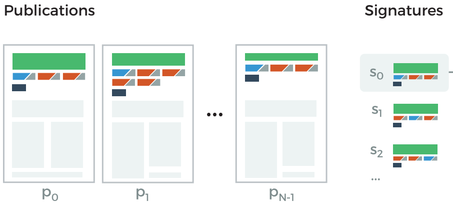
Machine Learning for Author disambiguation

Gilles Louppe

CERN

October 14, 2015

From publications to signatures



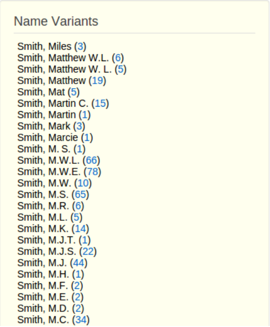
Signature for Doe, John

Title	Lorem ipsum dolor sit amet, consectetur adipiscing elit
Author	Doe, John
Affiliation	University of Foo
Co-authors	Smith, John; Chen, Wang
Year	2015

Author disambiguation

For each author, group together all his signatures, and only those.

`M.S.Smith.1`

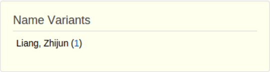


Name Variants

Smith, Miles (3)
Smith, Matthew W.L. (6)
Smith, Matthew W. L. (5)
Smith, Matthew (19)
Smith, Mat (5)
Smith, Martin C. (15)
Smith, Martin (1)
Smith, Mark (3)
Smith, Marcie (1)
Smith, M. S. (1)
Smith, M.W.L. (66)
Smith, M.W.E. (78)
Smith, M.W. (10)
Smith, M.S. (65)
Smith, M.R. (6)
Smith, M.L. (5)
Smith, M.K. (14)
Smith, M.J.T. (1)
Smith, M.J.S. (22)
Smith, M.J. (44)
Smith, M.H. (1)
Smith, M.F. (2)
Smith, M.E. (2)
Smith, M.D. (2)
Smith, M.C. (34)

No more

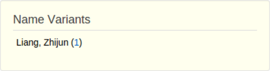
`Z.Liang.4`



Name Variants

Liang, Zhijun (1)

`Z.Liang.5`

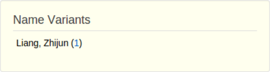


Name Variants

Liang, Zhijun (1)

...

`Z.Liang.83`

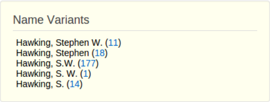


Name Variants

Liang, Zhijun (1)

No less

`S.W.Hawking.1`



Name Variants

Hawking, Stephen W. (11)
Hawking, Stephen (18)
Hawking, S.W. (177)
Hawking, S. W. (1)
Hawking, S. (14)

But all and only the correct ones

Spread of the problem

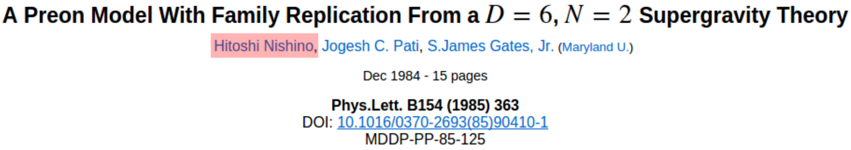
As extracted from claimed publications in INSPIRE,

- Authors have on average 2.06 name variants (synonyms)
  Eg. : Doe, John ; Doe, J.
- Unique name variants are shared on average by 1.04 authors (homonyms)

Clustering on same surnames and same given name initials, should yield very good results on average.

But, disambiguation issues are expected to amplify with the rise of Asian researchers : Caucasian names (now representative of INSPIRE authors) are almost never ambiguous, while Asian names are very often.

How would *you* fare ?

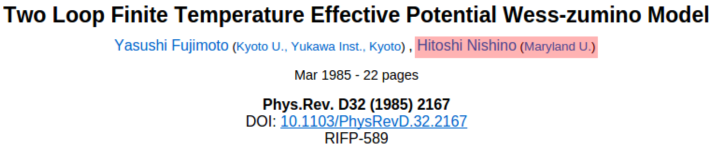


A Preon Model With Family Replication From a $D = 6, N = 2$ Supergravity Theory

Hitoshi Nishino, Jogesh C. Pati, S.James Gates, Jr. (Maryland U.)

Dec 1984 - 15 pages

Phys.Lett. B154 (1985) 363
DOI: 10.1016/0370-2693(85)90410-1
MDDP-PP-85-125



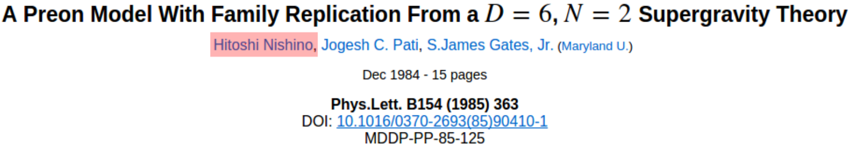
Two Loop Finite Temperature Effective Potential Wess-zumino Model

Yasushi Fujimoto (Kyoto U., Yukawa Inst., Kyoto) , Hitoshi Nishino (Maryland U.)

Mar 1985 - 22 pages

Phys.Rev. D32 (1985) 2167
DOI: 10.1103/PhysRevD.32.2167
RIFP-589

How would *you* fare ?

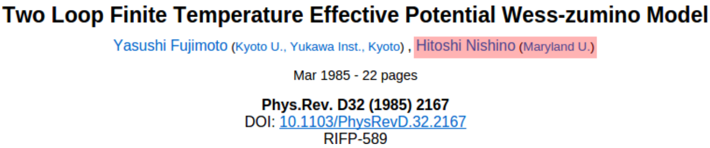


A Preon Model With Family Replication From a $D = 6, N = 2$ Supergravity Theory

Hitoshi Nishino, Jogesh C. Pati, S.James Gates, Jr. (Maryland U.)

Dec 1984 - 15 pages

Phys.Lett. B154 (1985) 363
DOI: 10.1016/0370-2693(85)90410-1
MDDP-PP-85-125



Two Loop Finite Temperature Effective Potential Wess-zumino Model

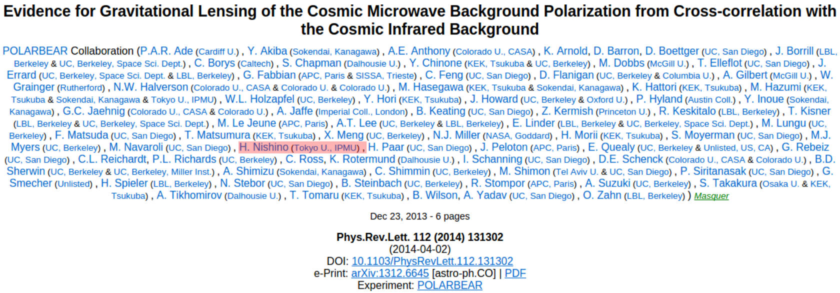
Yasushi Fujimoto (Kyoto U., Yukawa Inst., Kyoto) , Hitoshi Nishino (Maryland U.)

Mar 1985 - 22 pages

Phys.Rev. D32 (1985) 2167
DOI: 10.1103/PhysRevD.32.2167
RIFP-589

✓ Same authors

How would *you* fare ?

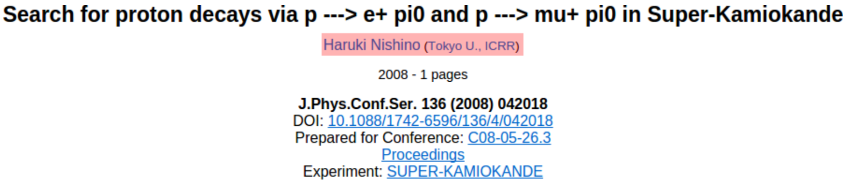


Evidence for Gravitational Lensing of the Cosmic Microwave Background Polarization from Cross-correlation with the Cosmic Infrared Background

POLARBEAR Collaboration (P.A.R. Ade (Cardiff U.) , Y. Akiba (Sokendai, Kanagawa) , A.E. Anthony (Colorado U., CASA) , K. Arnold, D. Barron, D. Boettger (UC, San Diego) , J. Borrill (LBL, Berkeley & UC, Berkeley, Space Sci. Dept.) , C. Borys (Caltech) , S. Chapman (Dalhousie U.) , Y. Chinone (KEK, Tsukuba & UC, Berkeley) , M. Dobbs (McGill U.) , T. Elleflot (UC, San Diego) , J. Errard (UC, Berkeley, Space Sci. Dept. & LBL, Berkeley) , G. Fabbian (APC, Paris & SISSA, Trieste) , C. Feng (UC, San Diego) , D. Flanigan (UC, Berkeley & Columbia U.) , A. Gilbert (McGill U.) , W. Grainger (Rutherford) , N.W. Halverson (Colorado U., CASA & Colorado U. & Colorado U.) , M. Hasegawa (KEK, Tsukuba & Sokendai, Kanagawa) , K. Hattori (KEK, Tsukuba) , M. Hazumi (KEK, Tsukuba & Sokendai, Kanagawa & Tokyo U., IPMU) , W.L. Holzapfel (UC, Berkeley) , Y. Hori (KEK, Tsukuba) , J. Howard (UC, Berkeley & Oxford U.) , P. Hyland (Austin Coll.) , Y. Inoue (Sokendai, Kanagawa) , G.C. Jaehnig (Colorado U., CASA & Colorado U.) , A. Jaffe (Imperial Coll., London) , B. Keating (UC, San Diego) , Z. Kermish (Princeton U.) , R. Keskitalo (LBL, Berkeley) , T. Kisner (LBL, Berkeley & UC, Berkeley, Space Sci. Dept.) , M. Le Jeune (APC, Paris) , A.T. Lee (UC, Berkeley & LBL, Berkeley) , E. Linder (LBL, Berkeley & UC, Berkeley, Space Sci. Dept.) , M. Lungu (UC, Berkeley) , F. Matsuda (UC, San Diego) , T. Matsumura (KEK, Tsukuba) , X. Meng (UC, Berkeley) , N.J. Miller (NASA, Goddard) , H. Morii (KEK, Tsukuba) , S. Moyerman (UC, San Diego) , M.J. Myers (UC, Berkeley) , M. Navaroli (UC, San Diego) , H. Nishino (Tokyo U., IPMU) , H. Paar (UC, San Diego) , J. Peloton (APC, Paris) , E. Quealy (UC, Berkeley & Unlisted, US, CA) , G. Rebeiz (UC, San Diego) , C.L. Reichardt, P.L. Richards (UC, Berkeley) , C. Ross, K. Rotermund (Dalhousie U.) , I. Schanning (UC, San Diego) , D.E. Schenck (Colorado U., CASA & Colorado U.) , B.D. Sherwin (UC, Berkeley & UC, Berkeley, Miller Inst.) , A. Shimizu (Sokendai, Kanagawa) , C. Shimmin (UC, Berkeley) , M. Shimon (Tel Aviv U. & UC, San Diego) , P. Siritanasak (UC, San Diego) , G. Smecher (Unlisted) , H. Spieler (LBL, Berkeley) , N. Stebor (UC, San Diego) , B. Steinbach (UC, Berkeley) , R. Stompor (APC, Paris) , A. Suzuki (UC, Berkeley) , S. Takakura (Osaka U. & KEK, Tsukuba) , A. Tikhomirov (Dalhousie U.) , T. Tomaru (KEK, Tsukuba) , B. Wilson, A. Yadav (UC, San Diego) , O. Zahn (LBL, Berkeley)) Masquer

Dec 23, 2013 - 6 pages

Phys.Rev.Lett. 112 (2014) 131302
(2014-04-02)
DOI: 10.1103/PhysRevLett.112.131302
e-Print: arXiv:1312.6645 [astro-ph.CO] | PDF
Experiment: POLARBEAR



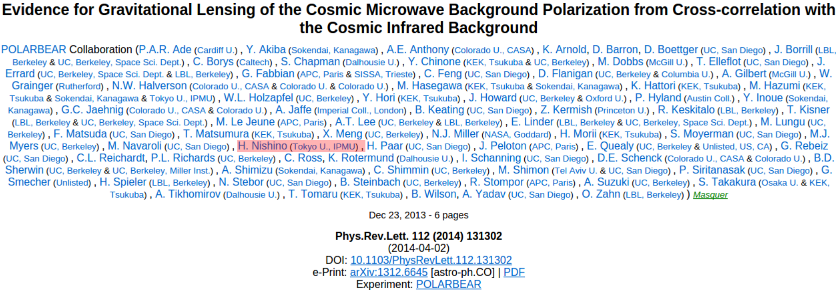
Search for proton decays via p ---> e+ pi0 and p ---> mu+ pi0 in Super-Kamiokande

Haruki Nishino (Tokyo U., ICRR)

2008 - 1 pages

J.Phys.Conf.Ser. 136 (2008) 042018
DOI: 10.1088/1742-6596/136/4/042018
Prepared for Conference: C08-05-26.3
Proceedings
Experiment: SUPER-KAMIOKANDE

How would *you* fare ?

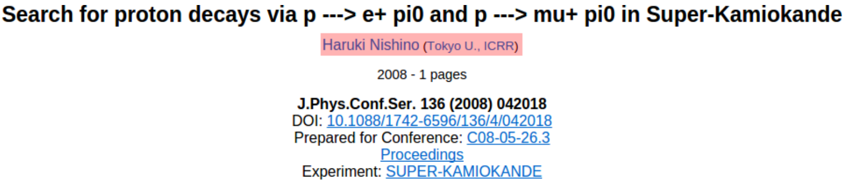


Evidence for Gravitational Lensing of the Cosmic Microwave Background Polarization from Cross-correlation with the Cosmic Infrared Background

POLARBEAR Collaboration (P.A.R. Ade (Cardiff U.) , Y. Akiba (Sokendai, Kanagawa) , A.E. Anthony (Colorado U., CASA) , K. Arnold, D. Barron, D. Boettger (UC, San Diego) , J. Borrill (LBL, Berkeley & UC, Berkeley, Space Sci. Dept.) , C. Borys (Caltech) , S. Chapman (Dalhousie U.) , Y. Chinone (KEK, Tsukuba & UC, Berkeley) , M. Dobbs (McGill U.) , T. Elleflot (UC, San Diego) , J. Errard (UC, Berkeley, Space Sci. Dept. & LBL, Berkeley) , G. Fabbian (APC, Paris & SISSA, Trieste) , C. Feng (UC, San Diego) , D. Flanigan (UC, Berkeley & Columbia U.) , A. Gilbert (McGill U.) , W. Grainger (Rutherford) , N.W. Halverson (Colorado U., CASA & Colorado U. & Colorado U.) , M. Hasegawa (KEK, Tsukuba & Sokendai, Kanagawa) , K. Hattori (KEK, Tsukuba) , M. Hazumi (KEK, Tsukuba & Sokendai, Kanagawa & Tokyo U., IPMU) , W.L. Holzapfel (UC, Berkeley) , Y. Hori (KEK, Tsukuba) , J. Howard (UC, Berkeley & Oxford U.) , P. Hyland (Austin Coll.) , Y. Inoue (Sokendai, Kanagawa) , G.C. Jaehnig (Colorado U., CASA & Colorado U.) , A. Jaffe (Imperial Coll., London) , B. Keating (UC, San Diego) , Z. Kermish (Princeton U.) , R. Keskitalo (LBL, Berkeley) , T. Kisner (LBL, Berkeley & UC, Berkeley, Space Sci. Dept.) , M. Le Jeune (APC, Paris) , A.T. Lee (UC, Berkeley & LBL, Berkeley) , E. Linder (LBL, Berkeley & UC, Berkeley, Space Sci. Dept.) , M. Lungu (UC, Berkeley) , F. Matsuda (UC, San Diego) , T. Matsumura (KEK, Tsukuba) , X. Meng (UC, Berkeley) , N.J. Miller (NASA, Goddard) , H. Morii (KEK, Tsukuba) , S. Moyerman (UC, San Diego) , M.J. Myers (UC, Berkeley) , M. Navaroli (UC, San Diego) , H. Nishino (Tokyo U., IPMU) , H. Paar (UC, San Diego) , J. Peloton (APC, Paris) , E. Quealy (UC, Berkeley & Unlisted, US, CA) , G. Rebeiz (UC, San Diego) , C.L. Reichardt, P.L. Richards (UC, Berkeley) , C. Ross, K. Rotermund (Dalhousie U.) , I. Schanning (UC, San Diego) , D.E. Schenck (Colorado U., CASA & Colorado U.) , B.D. Sherwin (UC, Berkeley & UC, Berkeley, Miller Inst.) , A. Shimizu (Sokendai, Kanagawa) , C. Shimmin (UC, Berkeley) , M. Shimon (Tel Aviv U. & UC, San Diego) , P. Siritanasak (UC, San Diego) , G. Smecher (Unlisted) , H. Spieler (LBL, Berkeley) , N. Stebor (UC, San Diego) , B. Steinbach (UC, Berkeley) , R. Stompor (APC, Paris) , A. Suzuki (UC, Berkeley) , S. Takakura (Osaka U. & KEK, Tsukuba) , A. Tikhomirov (Dalhousie U.) , T. Tomaru (KEK, Tsukuba) , B. Wilson, A. Yadav (UC, San Diego) , O. Zahn (LBL, Berkeley)) Masquer

Dec 23, 2013 - 6 pages

Phys.Rev.Lett. 112 (2014) 131302
(2014-04-02)
DOI: 10.1103/PhysRevLett.112.131302
e-Print: arXiv:1312.6645 [astro-ph.CO] | PDF
Experiment: POLARBEAR



Search for proton decays via p ---> e+ pi0 and p ---> mu+ pi0 in Super-Kamiokande

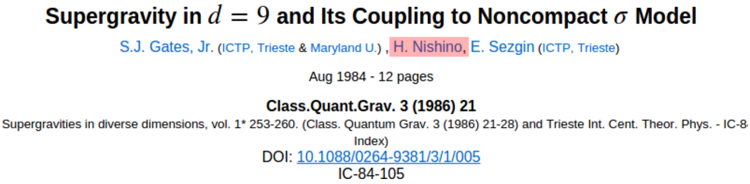
Haruki Nishino (Tokyo U., ICRR)

2008 - 1 pages

J.Phys.Conf.Ser. 136 (2008) 042018
DOI: 10.1088/1742-6596/136/4/042018
Prepared for Conference: C08-05-26.3
Proceedings
Experiment: SUPER-KAMIOKANDE

✓ Same authors

How would *you* fare ?



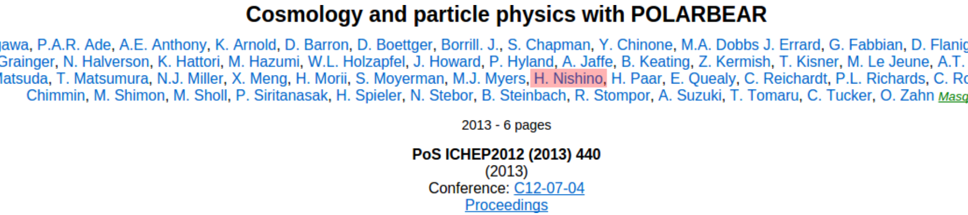
Supergravity in $d = 9$ and Its Coupling to Noncompact σ Model

S.J. Gates, Jr. (ICTP, Trieste & Maryland U.) , H. Nishino, E. Sezgin (ICTP, Trieste)

Aug 1984 - 12 pages

Class.Quant.Grav. 3 (1986) 21

Supergravities in diverse dimensions, vol. 1* 253-260. (Class. Quantum Grav. 3 (1986) 21-28) and Trieste Int. Cent. Theor. Phys. - IC-8
Index)
DOI: 10.1088/0264-9381/3/1/005
IC-84-105



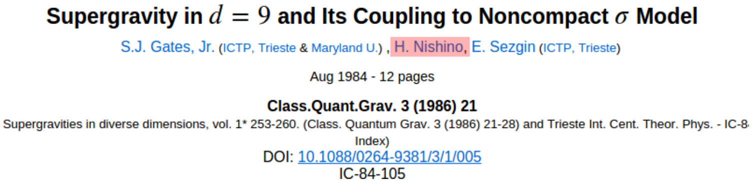
Cosmology and particle physics with POLARBEAR

awa, P.A.R. Ade, A.E. Anthony, K. Arnold, D. Barron, D. Boettger, Borrill. J., S. Chapman, Y. Chinone, M.A. Dobbs J. Errard, G. Fabbian, D. Flanig
Grainger, N. Halverson, K. Hattori, M. Hazumi, W.L. Holzapfel, J. Howard, P. Hyland, A. Jaffe, B. Keating, Z. Kermish, T. Kisner, M. Le Jeune, A.T.
latsuda, T. Matsumura, N.J. Miller, X. Meng, H. Morii, S. Moyerman, M.J. Myers, H. Nishino, H. Paar, E. Quealy, C. Reichardt, P.L. Richards, C. Ro
Chimmin, M. Shimon, M. Sholl, P. Siritanasak, H. Spieler, N. Stebor, B. Steinbach, R. Stompor, A. Suzuki, T. Tomaru, C. Tucker, O. Zahn *Masq*

2013 - 6 pages

PoS ICHEP2012 (2013) 440
(2013)
Conference: C12-07-04
Proceedings

How would *you* fare ?



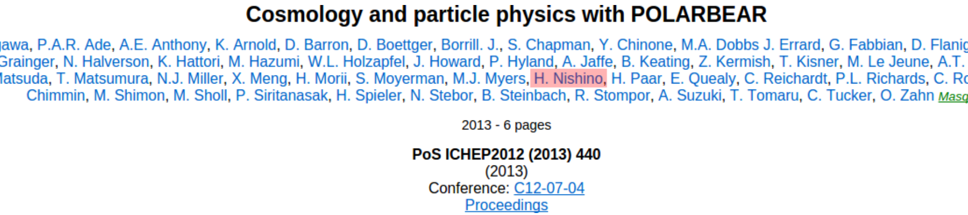
Supergravity in $d = 9$ and Its Coupling to Noncompact σ Model

S.J. Gates, Jr. (ICTP, Trieste & Maryland U.) , H. Nishino, E. Sezgin (ICTP, Trieste)

Aug 1984 - 12 pages

Class.Quant.Grav. 3 (1986) 21

Supergravities in diverse dimensions, vol. 1* 253-260. (Class. Quantum Grav. 3 (1986) 21-28) and Trieste Int. Cent. Theor. Phys. - IC-8
Index)
DOI: 10.1088/0264-9381/3/1/005
IC-84-105



Cosmology and particle physics with POLARBEAR

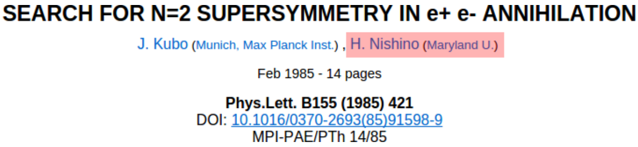
awa, P.A.R. Ade, A.E. Anthony, K. Arnold, D. Barron, D. Boettger, Borrill. J., S. Chapman, Y. Chinone, M.A. Dobbs J. Errard, G. Fabbian, D. Flanig
Grainger, N. Halverson, K. Hattori, M. Hazumi, W.L. Holzapfel, J. Howard, P. Hyland, A. Jaffe, B. Keating, Z. Kermish, T. Kisner, M. Le Jeune, A.T.
latsuda, T. Matsumura, N.J. Miller, X. Meng, H. Morii, S. Moyerman, M.J. Myers, H. Nishino, H. Paar, E. Quealy, C. Reichardt, P.L. Richards, C. Ro
Chimmin, M. Shimon, M. Sholl, P. Siritanasak, H. Spieler, N. Stebor, B. Steinbach, R. Stompor, A. Suzuki, T. Tomaru, C. Tucker, O. Zahn *Masq*

2013 - 6 pages

PoS ICHEP2012 (2013) 440
(2013)
Conference: C12-07-04
Proceedings

✗ Different authors

How would *you* fare ?

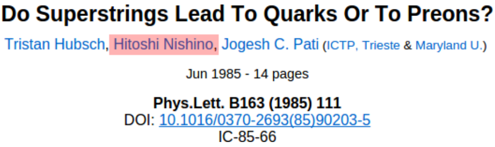


SEARCH FOR N=2 SUPERSYMMETRY IN e+ e- ANNIHILATION

J. Kubo (Munich, Max Planck Inst.) , H. Nishino (Maryland U.)

Feb 1985 - 14 pages

Phys.Lett. B155 (1985) 421
DOI: 10.1016/0370-2693(85)91598-9
MPI-PAE/PTh 14/85



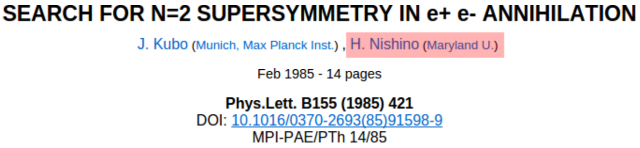
Do Superstrings Lead To Quarks Or To Preons?

Tristan Hubsch, Hitoshi Nishino, Jogesh C. Pati (ICTP, Trieste & Maryland U.)

Jun 1985 - 14 pages

Phys.Lett. B163 (1985) 111
DOI: 10.1016/0370-2693(85)90203-5
IC-85-66

How would *you* fare ?

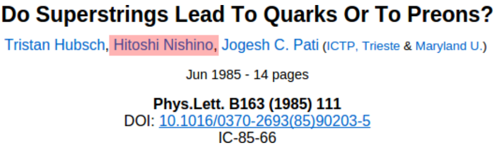


SEARCH FOR N=2 SUPERSYMMETRY IN e+ e- ANNIHILATION

J. Kubo (Munich, Max Planck Inst.) , H. Nishino (Maryland U.)

Feb 1985 - 14 pages

Phys.Lett. B155 (1985) 421
DOI: 10.1016/0370-2693(85)91598-9
MPI-PAE/PTh 14/85



Do Superstrings Lead To Quarks Or To Preons?

Tristan Hubsch, Hitoshi Nishino, Jogesh C. Pati (ICTP, Trieste & Maryland U.)

Jun 1985 - 14 pages

Phys.Lett. B163 (1985) 111
DOI: 10.1016/0370-2693(85)90203-5
IC-85-66

✓ Same authors

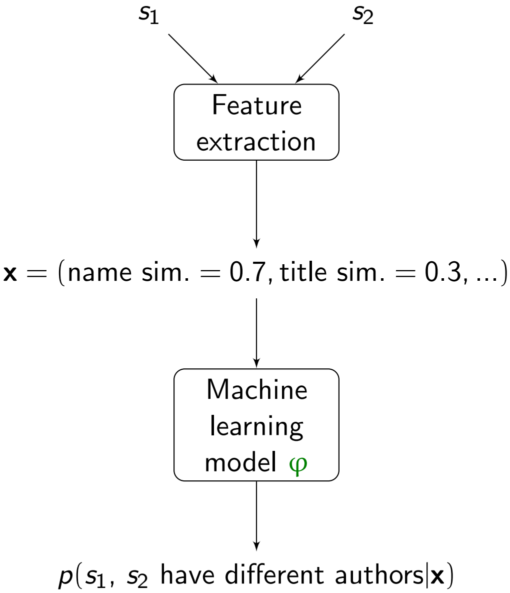
Learning from data

- Manual disambiguation is long and difficult, even for experienced curators.

- Couldn't we automatically find a set of rules to disambiguate two signatures ?

$$\varphi(s_1, s_2) = \begin{cases} 0 & \text{if } s_1 \text{ and } s_2 \text{ belong to the same author,} \\ 1 & \text{otherwise.} \end{cases}$$

- This is a machine learning task called supervised learning.



s_1

s_2

Feature extraction

$\mathbf{x} = (\text{name sim.} = 0.7, \text{title sim.} = 0.3, \ldots)$

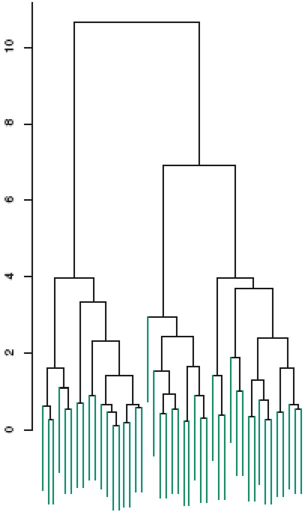
Machine learning model φ

$p(s_1,\ s_2 \text{ have different authors}|\mathbf{x})$

Feature extraction

Feature	**Combination operator**
Full name	Cosine similarity of $(2, 4)$-TF-IDF
Given names	Cosine similarity of $(2, 4)$-TF-IDF
First given name	Jaro-Winkler distance
Second given name	Jaro-Winkler distance
Given name initial	Equality
Affiliation	Cosine similarity of $(2, 4)$-TF-IDF
Co-authors	Cosine similarity of TF-IDF
Title	Cosine similarity of $(2, 4)$-TF-IDF
Journal	Cosine similarity of $(2, 4)$-TF-IDF
Abstract	Cosine similarity of TF-IDF
Keywords	Cosine similarity of TF-IDF
Collaborations	Cosine similarity of TF-IDF
References	Cosine similarity of TF-IDF
Subject	Cosine similarity of TF-IDF
Year difference	Absolute difference
White	Product of estimated probabilities
Black	Product of estimated probabilities
American Indian or Alaska Native	Product of estimated probabilities
Chinese	Product of estimated probabilities
Japanese	Product of estimated probabilities
Other Asian or Pacific Islander	Product of estimated probabilities
Others	Product of estimated probabilities

Disambiguation as a clustering problem

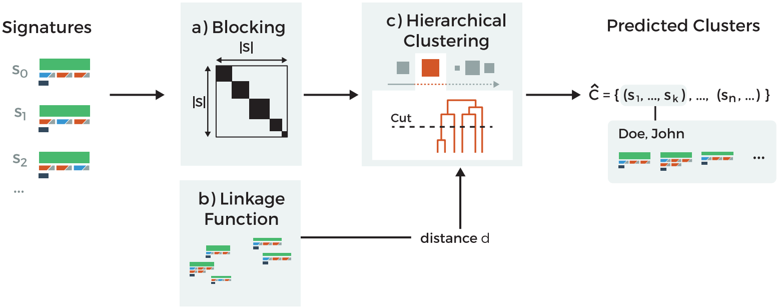


- Author disambiguation = clustering signatures that belong to the same author.

- Using our model φ, the probability that two signatures belong to different authors can be used as a (pseudo) distance metric, and e.g., plugged into a hierarchical clustering clustering.

- The complexity of hierarchical clustering is $O(N^2)$. For $N = 10^7$ signatures, this is impractical. *Solution :* pre-cluster signatures into blocks of smaller size, then cluster each of these blocks.

Workflow



Signatures

s_0

s_1

s_2

...

a) Blocking

|S|

|S|

b) Linkage Function

c) Hierarchical Clustering

Cut

distance d

Predicted Clusters

$\hat{C} = \{ (s_1, \ldots, s_k), \ldots, (s_n, \ldots) \}$

Doe, John

...

Results

	F measure
Baseline[1]	0.9409
Our model	**0.9862**

[1] Group by same surnames and same given name initials.

References

- Implementation available at https://github.com/inveniosoftware/beard.
- *Ethnicity sensitive author disambiguation using semi-supervised learning.* Gilles Louppe, Hussein Al-Natsheh, Mateusz Susik, Eamonn Maguire. http://arxiv.org/abs/1508.07744.