# Stylometry with R

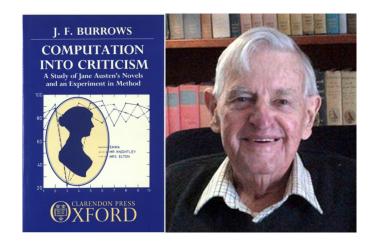
Part 1. Distances and uncertainty

Joanna Byszuk and Artjoms Šeļa



## 1. Quick recap of Burrows' Delta

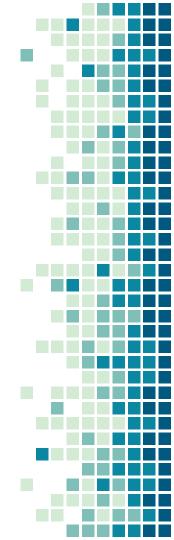
"Wealth of variables, many of which may be weak discriminators, almost always offer more tenable results than a smaller number of strong ones. [...] At all events, a distinctive 'stylistic signature' is usually made up of many tiny strokes."

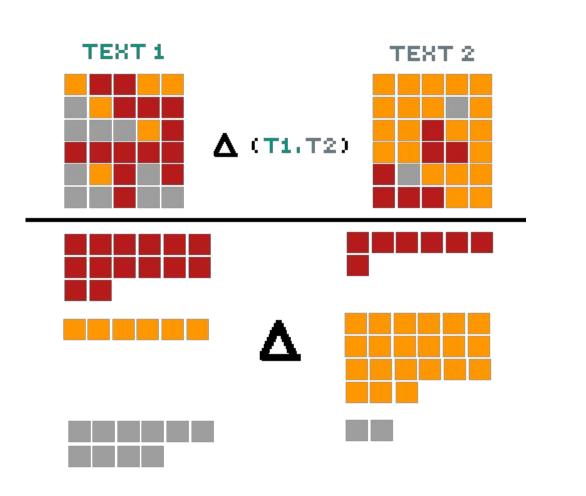


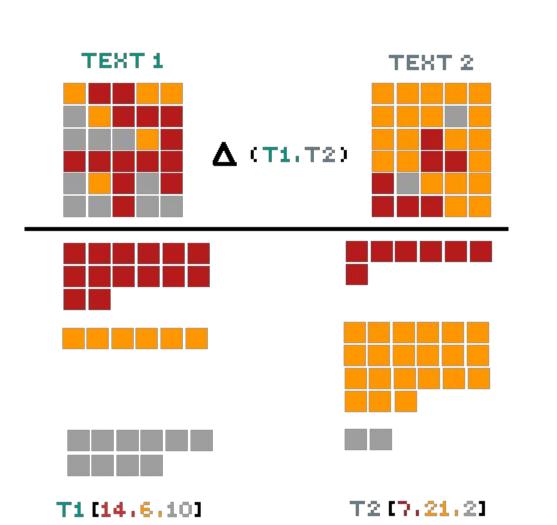
$$\Delta = \sum_{i=1}^{n} \frac{|z(x_i) - z(y_i)|}{n}$$

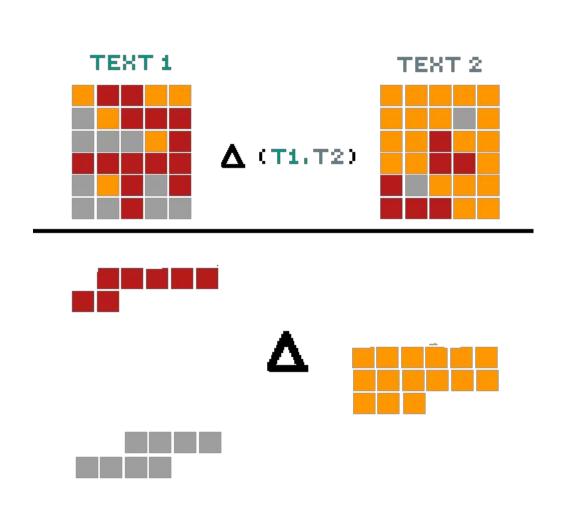
John Burrows (1928-2019)



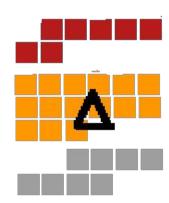




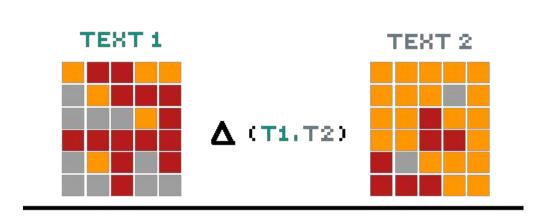


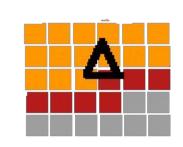






 $\triangle$  (T1.T2)=[6.15.10]



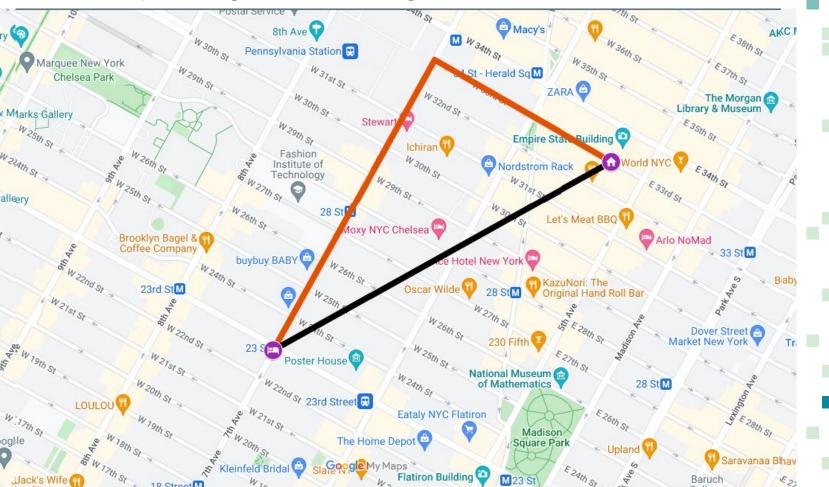


Manhattan, or city-block distance!
But also reinvented by Burrows
(with important adjustment)

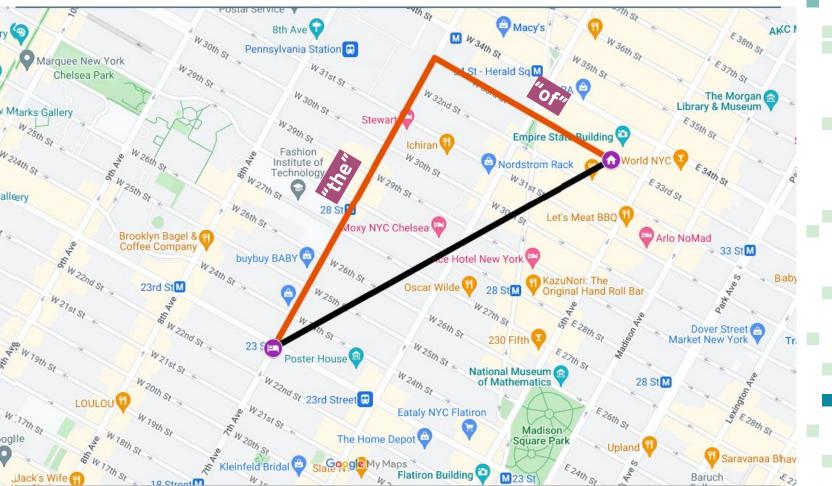
$$\triangle$$
 (T1.T2)= 7+15+8= 30

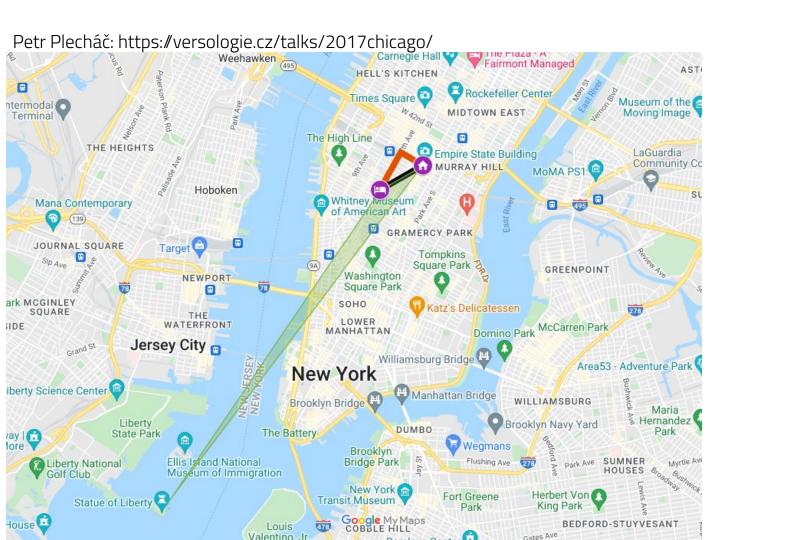
#### Petr Plecháč: https://versologie.cz/talks/2017chicago/ (a) Macy's 8th Ave AKC I ry 😙 M W 34th St E 38th St W30th St Pennsylvania Station E37th St Marquee New York W 29th St 34 St - Herald Sq M Chelsea Park W33rdSt ZARA 😩 W30th St The Morgan Library & Museum Marks-Gallery Stewart 😜 E 35th St W 29th St Empire State Building W 26th St Fashion E W S2th St W 30th St Nordstrom Rack Institute of Technology 6 W 27th St allery 28 St M Moxy NYC Chelsea Brooklyn Bagel & Coffee Company Arlo NoMad 33 St M buybuy BABY W 26th St W 24th St 5 W 22nd St Baby 23rd StM W27st St WarthSt E E 28th St # W 22nd St Dover Street A W 19th St W21st St Poster House National Museum 🔝 W 22nd Street ₩ of Mathematics 28 St M E 26th St Eataly NYC Flatiron WIZHSI & E28th St Madison W 18th St N 17th St The Home Depot Square Park Slate ogle My Maps Flatiron Building Jack's Wife Baruch E27

### Petr Plecháč: https://versologie.cz/talks/2017chicago/



#### Petr Plecháč: https://versologie.cz/talks/2017chicago/





Petr Plecháč: https://versologie.cz/talks/2017chicago/ Weehawken Fairmont Managed and 🔺 Museum of the ntermodal 6 Moving Image Terminal 5 Manhattan THE HE LaGuardia Distance Community Co ИA PS1 SL Mana Contempora 4 (139) JOURNAL SQUARE Euclidean REENPOINT 3 Distance ŭ ark MCGINLEY 278 SQUARE IDE Carren Park B Grand St Area53 - Adventure Park Cosine Distance iberty Science Center **MSBURG** Maria Hernandez Navy Yard Park ay | Evert et al. 2017 Liberty National Golf Club pert Von Statue of Google My Maps COBBLE HILL House 🖽 BEDFORD-STUYVESANT Louis

Cates Ave

Valentino Jr.

Petr Plecháč: https://versologie.cz/talks/2017chicago/ Weehawken relative frequencies of the ntermodal 👩 Image Terminal C. Brontë: Jane Eyre C. Brontë: Shirley THE ardia munity Co SL Mana Contemi (139) JOURNAL SQU SIp Ave ark MCGINLEY SQUARE standardized z-scores IDE Grand St C. Brontë: Jane Eyre C. Brontë: Shirley re Park iberty Science Cente nandez ay | Myrtle Ave Liberty Natio Statue House 🛑 COBBLE HILL THE CO. Valentino Jr. Cates Ave

### Note on distances for French

#### Burrow's Delta

(with Euclidean normalization)

#### Why Molière most likely did write his plays

FLORIAN CAFIERO 10 AND JEAN-BAPTISTE CAMPS DAUTHORS Info & Affiliations

SCIENCE ADVANCES - 27 Nov 2019 - Vol 5, Issue 11 - DOI: 10.1126/sciadv.aax5489

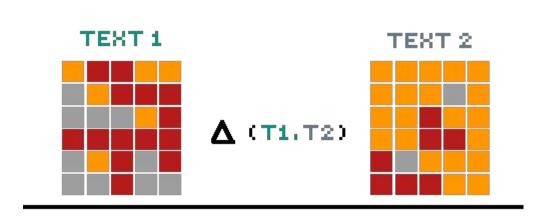
Cosine Delta (Wurzburg)

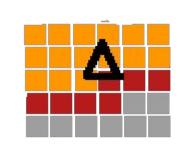
#### JOURNAL ARTICLE

### Understanding and explaining Delta measures for authorship attribution •

Stefan Evert, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, Thorsten Vitt

*Digital Scholarship in the Humanities*, Volume 32, Issue suppl\_2, December 2017, Pages ii4-ii16, https://doi.org/10.1093/llc/fqx023





Manhattan, or city-block distance!
But also reinvented by Burrows
(with important adjustment)

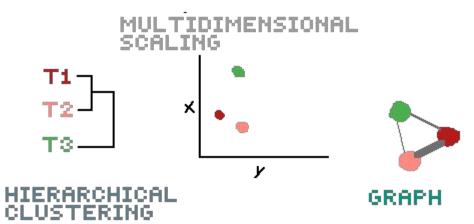
$$\triangle$$
 (T1.T2)= 7+15+8= 30

#### DISTANCE MATRIX

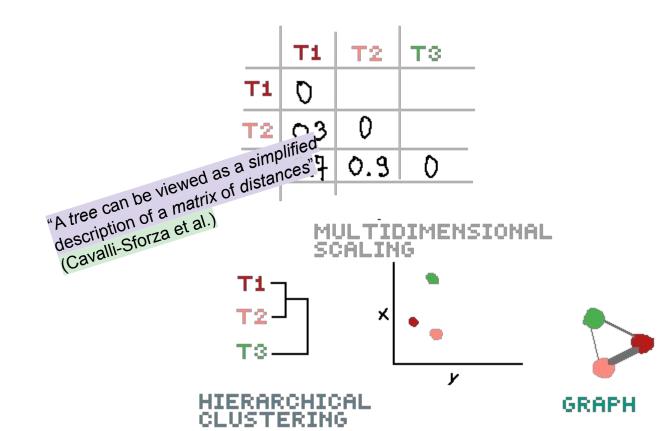
	Ti	Т2	тз
T1	O		
Т2	0.3	0	
тз	0.7	0.9	0

#### DISTANCE MATRIX

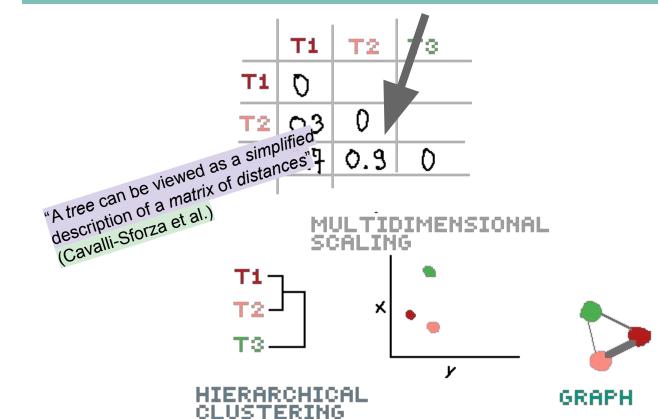
	Ti	Т2	тз
Τi	Ō		
Т2	0.3	0	
тз	0.7	0.9	0



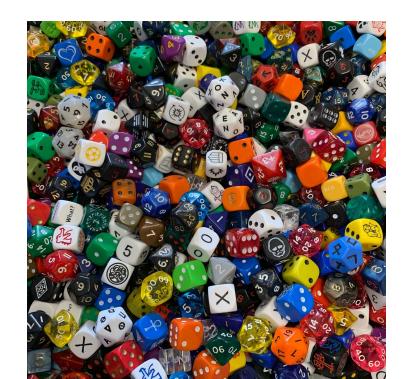
#### DISTANCE MATRIX



### OK, but how much can I trust this distance measure?



# Sampling, bootstrapping, iterations!





















Sample:

p(square) = 0.66



Sample:

p(square) = 0.66



Sample:

p(square) = 0.66

Resample 1:

Resample 2: 0.66



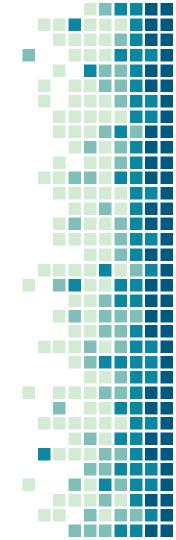
Sample:

p(square) = 0.66

Resample 1:

Resample 2: **A** 0.66

Resample 3: **A A A O**.33



Sample:

p(square) = 0.66

Resample 1:

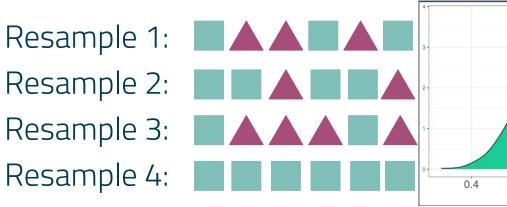
Resample 2: **A** 0.66

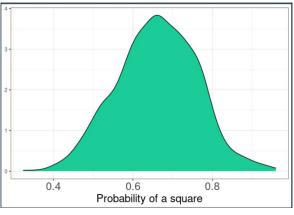
Resample 3: **A A A O**.33

Resample 4:

Sample:

p(square) = 0.66

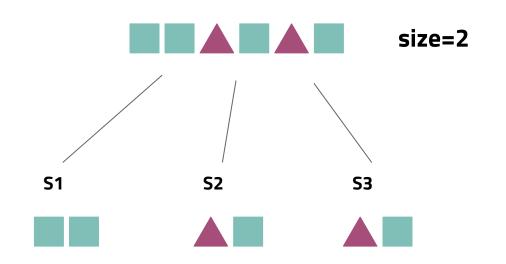




# 3. Estimating uncertainty in text similarity (within *stylo*)

- Random sampling tricks
- (Bootstrap) consensus trees (Eder 2013)
- (Bootstrap) consensus networks (Eder 2017)
- General Imposters (Kestemont et al. 2016)

# Normal vs. random sampling (in stylo)





51

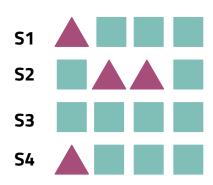


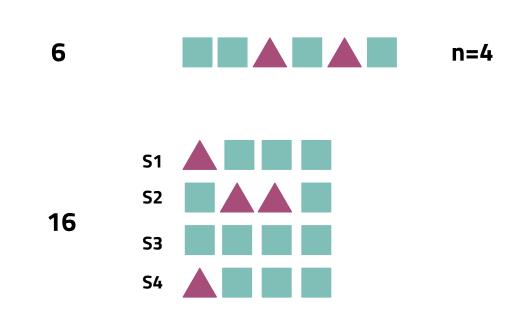


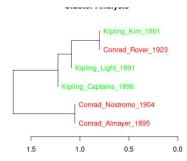




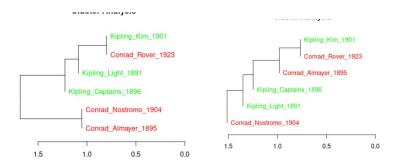






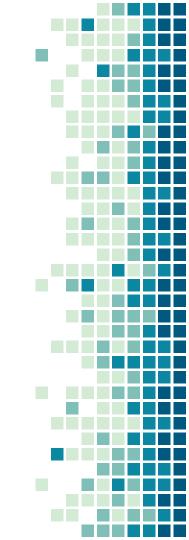


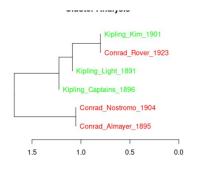
Feature set 1

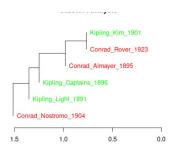


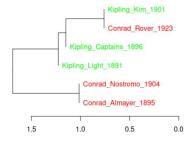
Feature set 1

Feature set 2





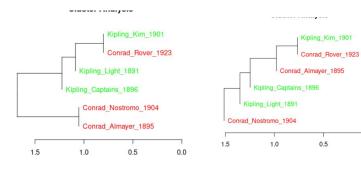


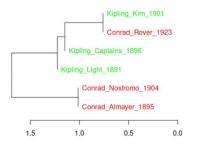


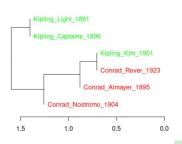
Feature set 1

Feature set 2

Feature set 3







Feature set 1

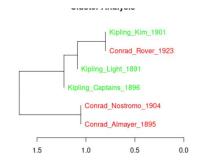
Feature set 2

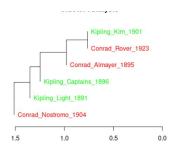
0.0

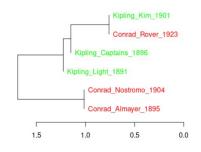
Feature set 3

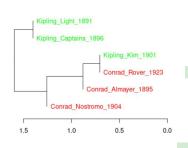
Feature set 4

### 4. Majority rule (>50% of branches)









Feature set 1

Feature set 2

Feature set 3

Feature set 4

Using stylo() off the shelf you can "bootstrap":

- MFW length
- Culling strength
- Text themselves (take samples from texts)

Using stylo() off the shelf you can "bootstrap":

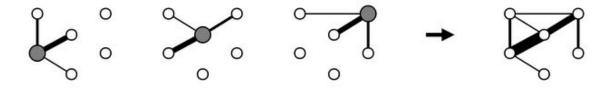
- MFW length
- Culling strength
- Text themselves (take samples from texts)

. . . .

But the possibilities are limitless

#### 6. Consensus networks

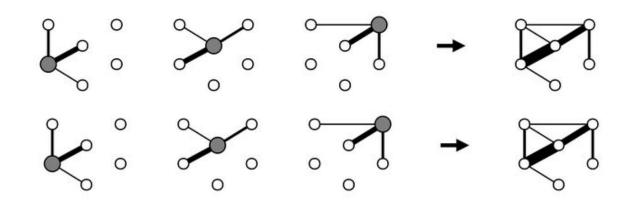
1. Look at the neighbours!



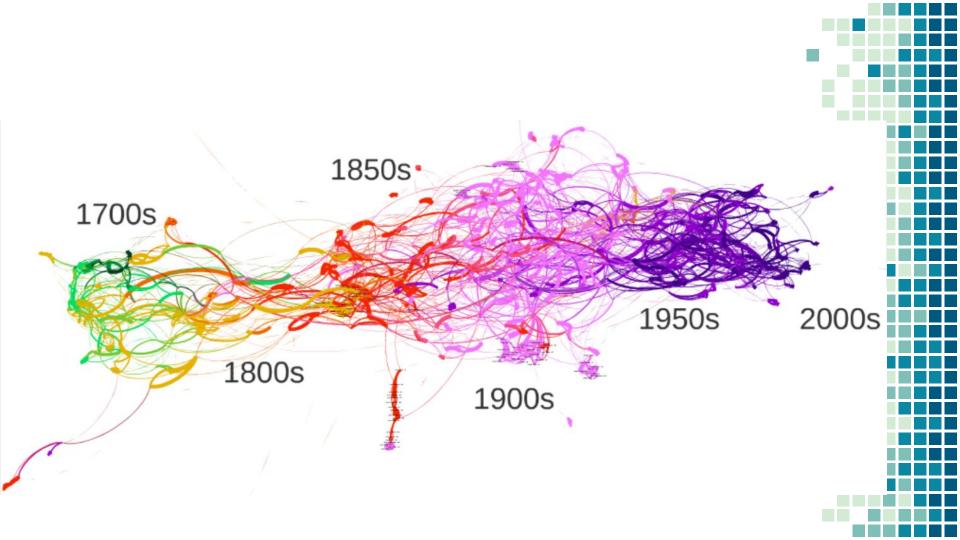


#### 6. Consensus networks

- 1. Look at the neighbours!
- 2. Then look at the neighbours many times!



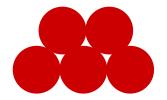




- Try using stylo.network() (alpha version!)
- Or brave the depths of Gephi
- Or work with networks from R!
  - Best tutorial I know:
  - https://kateto.net/network-visualization

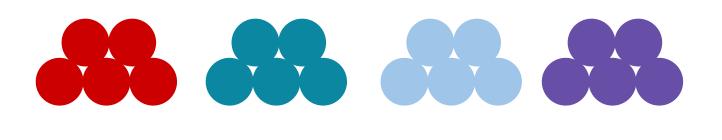


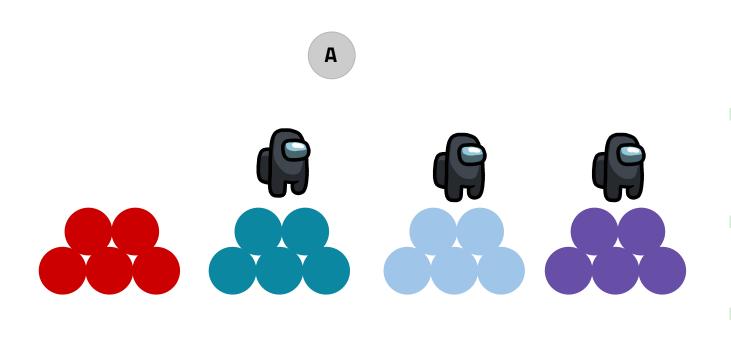
Α





Α





Random samples Random features Random imposters

Random samples
Random features
Random imposters



Random samples
Random features
Random imposters



