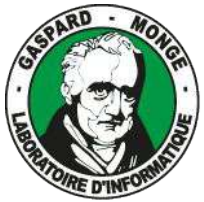
EADH 2021, the Second International Conference of the European Association for Digital Humanities

https://tinyurl.com/TreeOrderEadh2021

Evaluating Hierarchical Clustering Methods for Corpora with Chronological Order

Philippe Gambette, Olga Seminck, Dominique Legallois, Thierry Poibeau



Université Gustave Eiffel

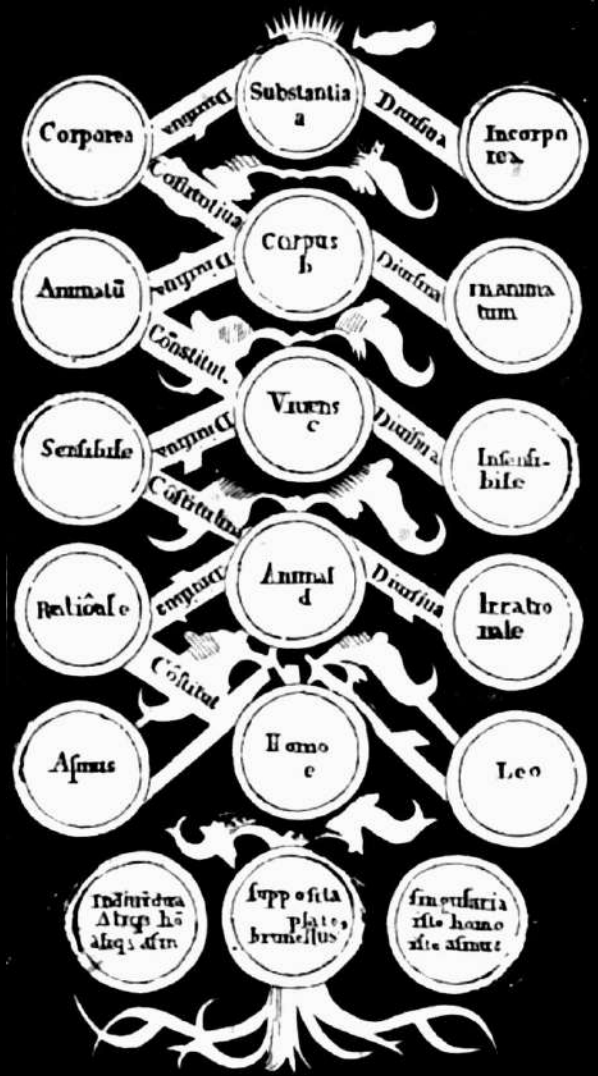
Sorbonne Nouvelle
université des cultures

ENS | PSL

Lattice

Outline of the talk

- When the result of a clustering displays a chronological signal: some context
- 2 criteria to evaluate the consistency between a tree and a chronological order
- Finding the optimal order for each criterion
- Evaluating the significance of the obtained results
- Perspectives and conclusion



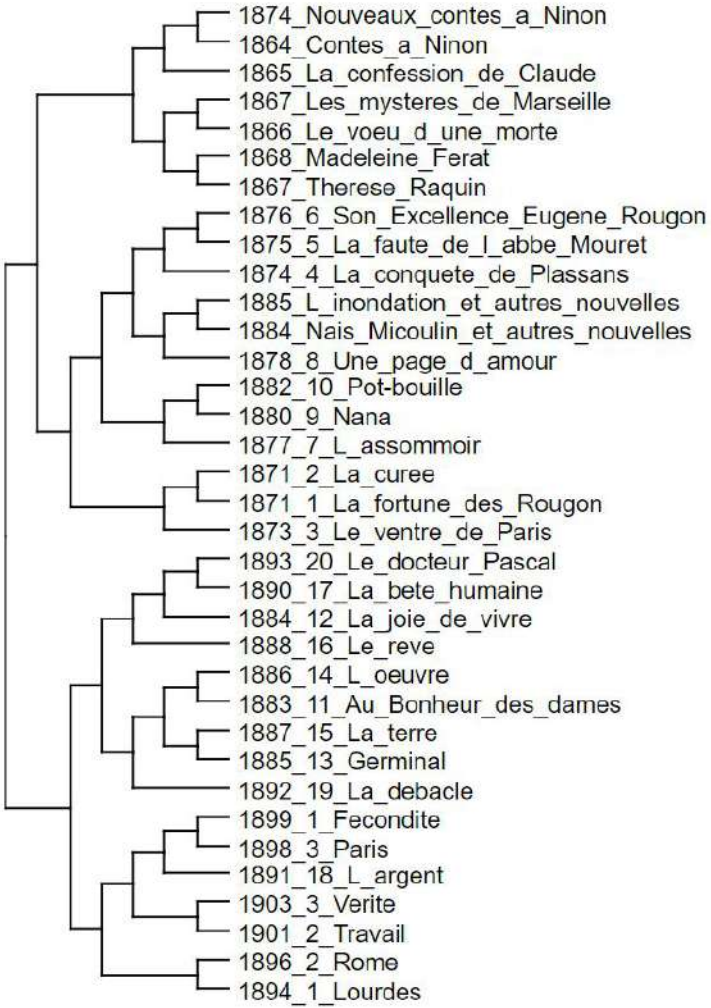
When the result of a clustering displays a chronological signal: some context

Source: Boethius, J.-P. Migne (ed.), *Patrologiae cursus completus. T. 64, Manlii Severini Boetii opera omnia*…, 1847, Staats- und Stadtbibliothek Augsburg, Google Books 9Vpm6G4A8aAC, p. 41.

Context of this study

Studying the evolution of the **idiolect of French XIXth century authors** (CIDRE corpus):

- a natural first step: hierarchical clustering
- does the clustering group together novels published in consecutive years?



Novels by Zola classified using motifs (Legallois, Charnois & Larjavaara, 2018) with the R package `stylo`

Context of this study

Studying the evolution of the idiolect of French XIXth century authors:

- a natural first step: hierarchical clustering
- does the clustering group together novels published in consecutive years?

A question relevant for other studies in digital humanities:

- historical linguistics (evolution of languages)
- political discourse analysis
- literature analysis
- etc.

Context of this study

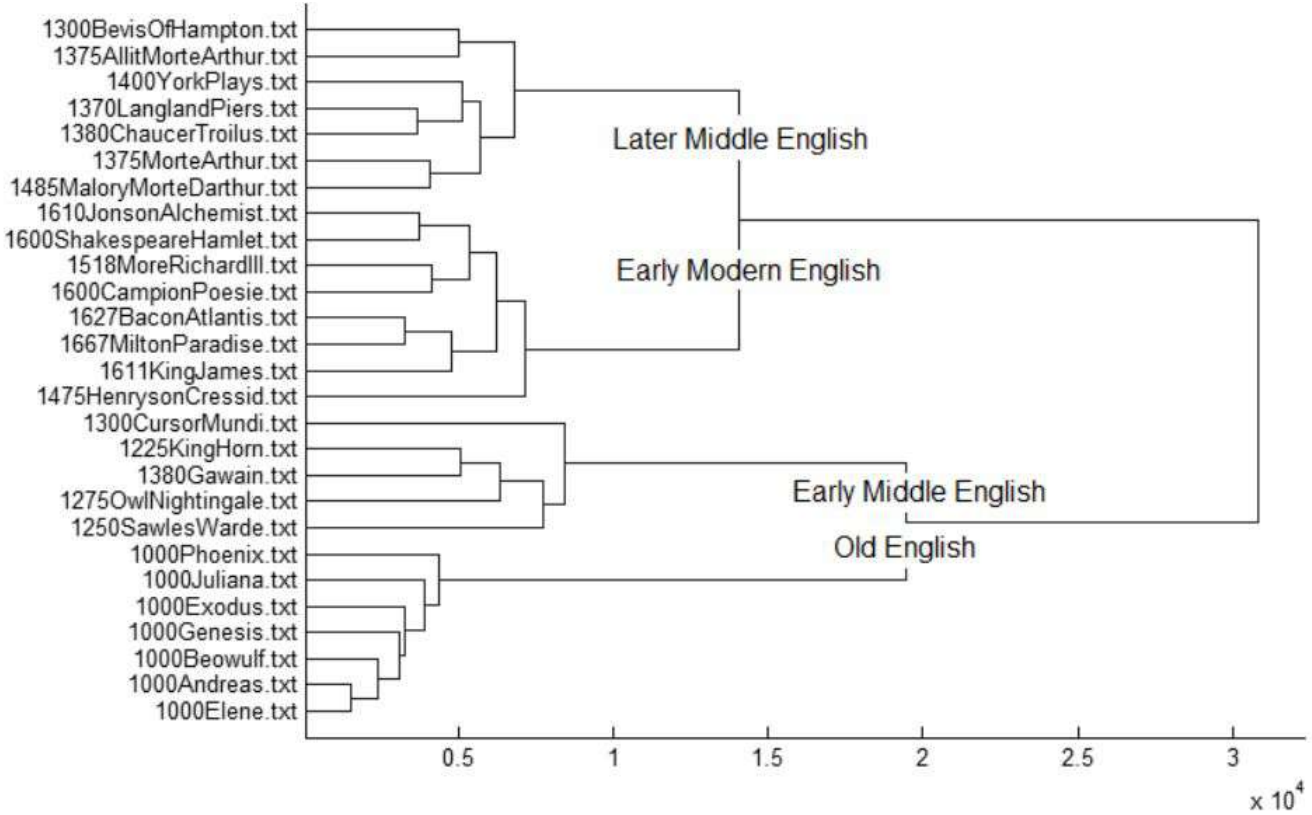


Figure 15. Hierarchical analysis of M' based on proximity matrix D in Table 6

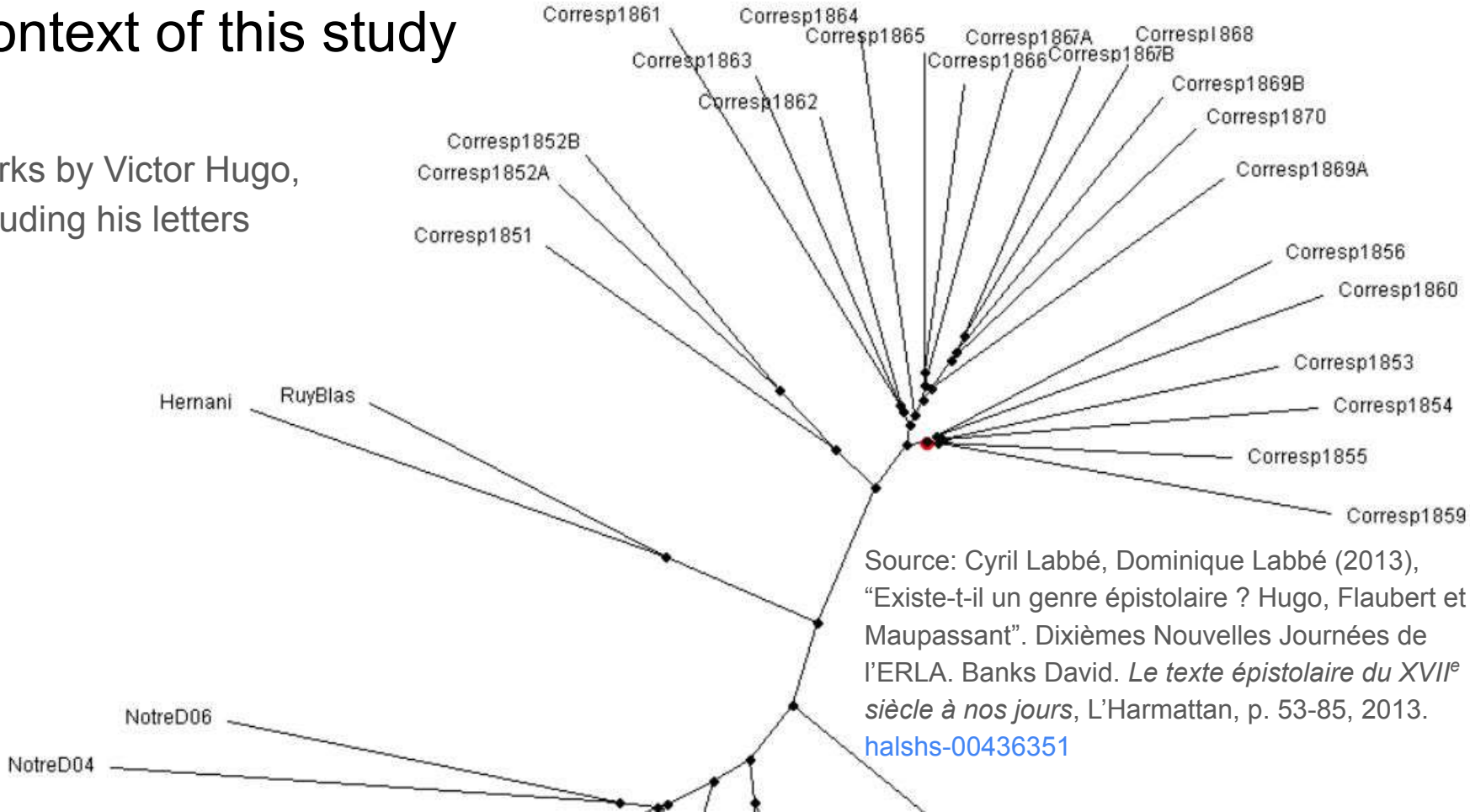
Old English texts

Source: Hermann Moisl (2020) “How to visualize high-dimensional data: a roadmap”. *Journal of Data Mining & Digital Humanities*, Special issue on Visualisations in Historical Linguistics - doi.org/10.46298/jdmdh.5594

Context of this study

Works by Victor Hugo, including his letters

Figure 1. Classification arborée sur les œuvres d'Hugo



Source: Cyril Labbé, Dominique Labbé (2013), "Existe-t-il un genre épistolaire ? Hugo, Flaubert et Maupassant". Dixièmes Nouvelles Journées de l'ERLA. Banks David. *Le texte épistolaire du XVII^e^ siècle à nos jours*, L'Harmattan, p. 53-85, 2013. halshs-00436351

Context of this study

New Year's addresses by French presidents from 1959 to 2001

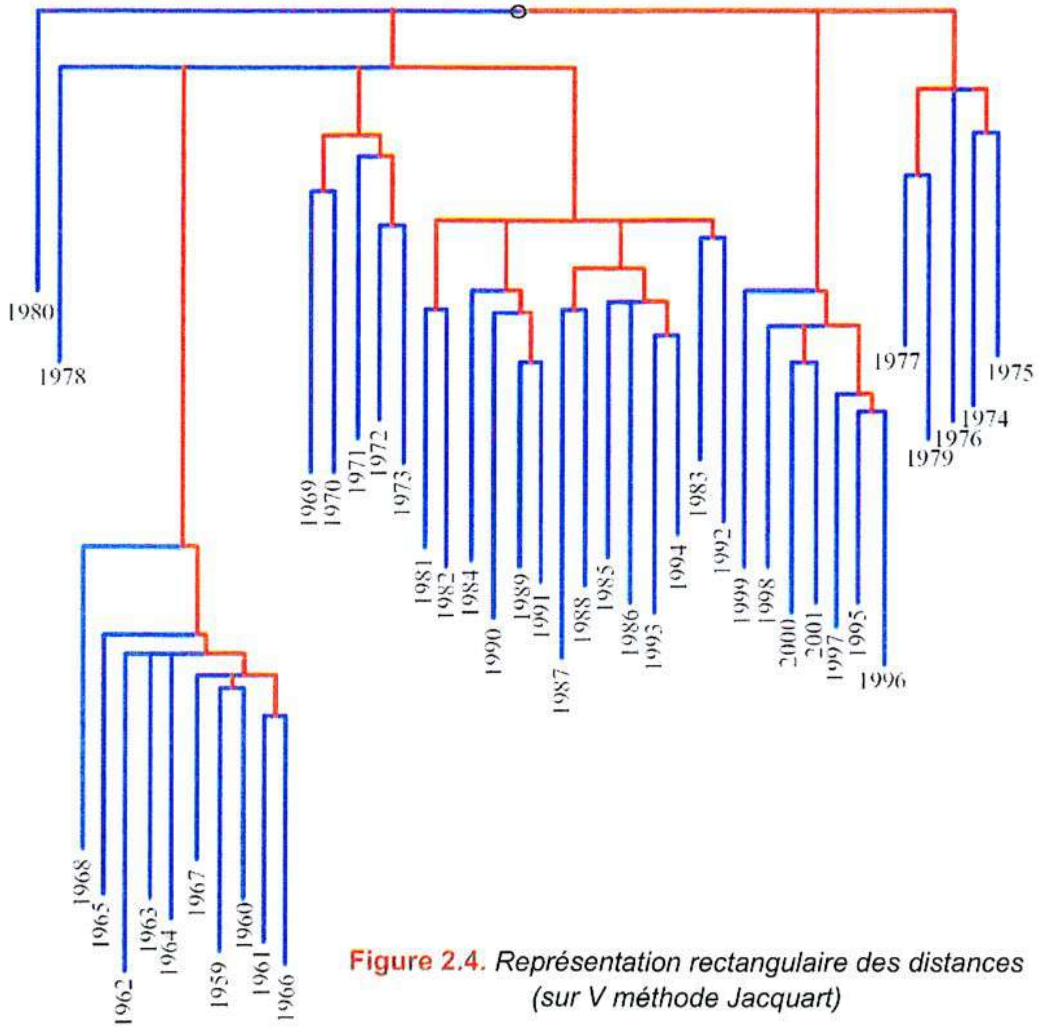


Figure 2.4. *Représentation rectangulaire des distances (sur V méthode Jacquart)*

Source: Jean-Marc Leblanc (2016), *Analyses lexicométriques des vœux présidentiels*. ISTE editions, p.64

Context of this study

New Year's addresses by French presidents from 1959 to 2001

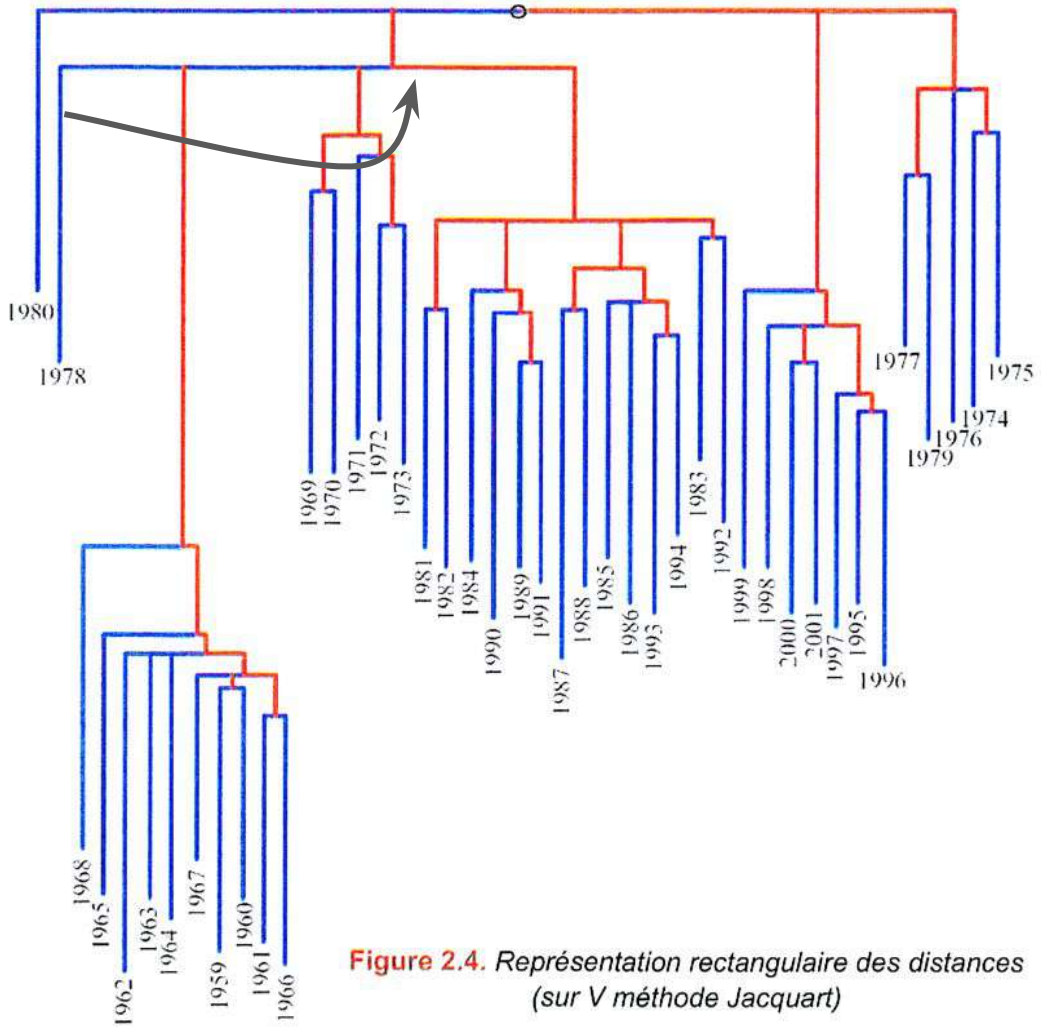


Figure 2.4. *Représentation rectangulaire des distances (sur V méthode Jacquart)*

Source: Jean-Marc Leblanc (2016), *Analyses lexicométriques des vœux présidentiels*. ISTE editions, p.64

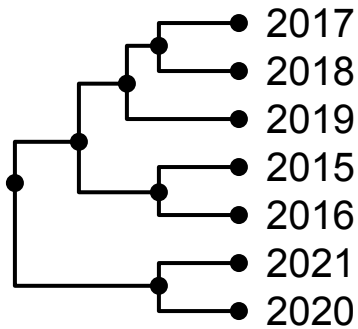
Our first sub-problem

Reordering the (dated) leaves of a tree/dendrogram

in order to

best fit with the chronology

First example:



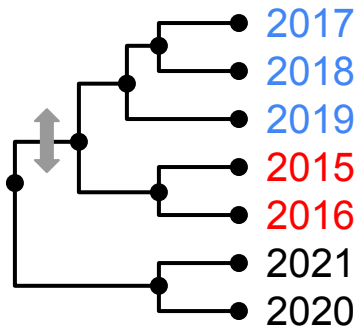
Our first sub-problem

Reordering the (dated) leaves of a tree/dendrogram

in order to

best fit with the chronology

First example:



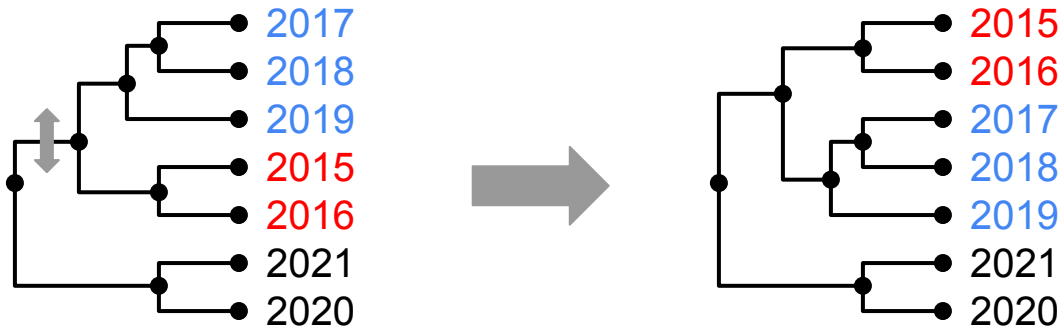
Our first sub-problem

Reordering the (dated) leaves of a tree/dendrogram

in order to

best fit with the chronology

First example:



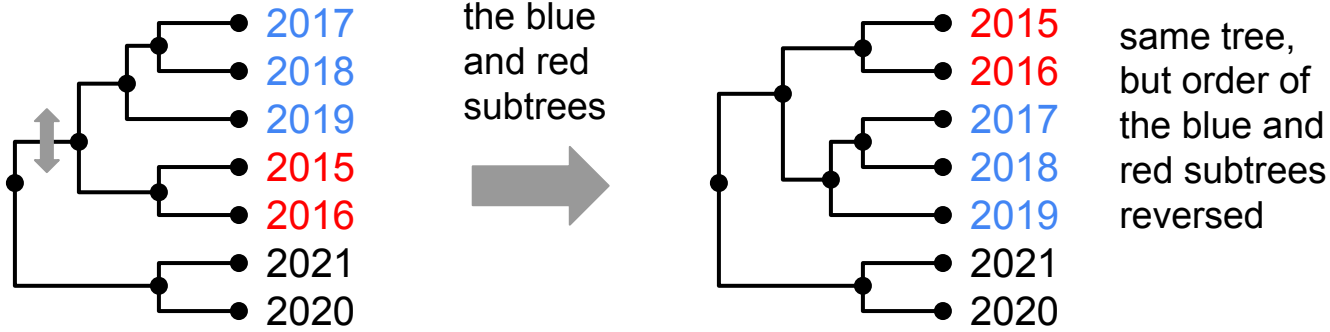
Our first sub-problem

Reordering the (dated) leaves of a tree/dendrogram

in order to

best fit with the chronology

First example:



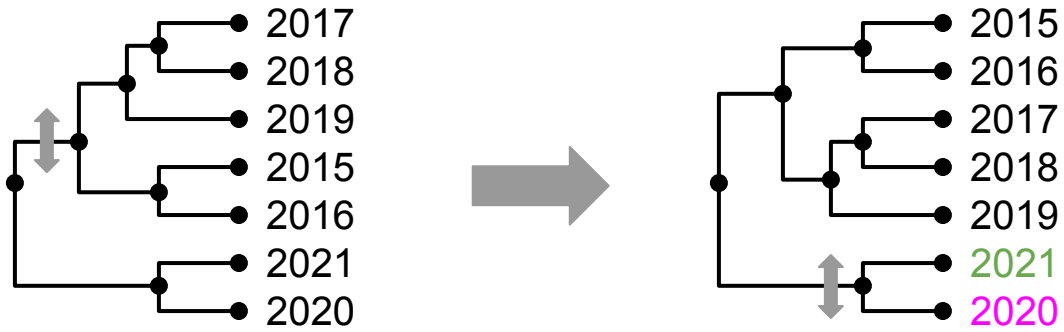
Our first sub-problem

Reordering the (dated) leaves of a tree/dendrogram

in order to

best fit with the chronology

First example:



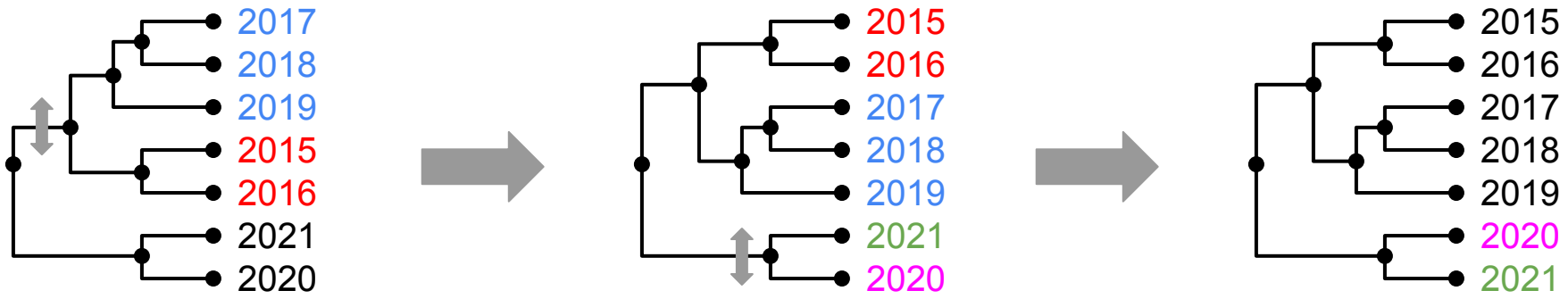
Our first sub-problem

Reordering the (dated) leaves of a tree/dendrogram

in order to

best fit with the chronology

First example:



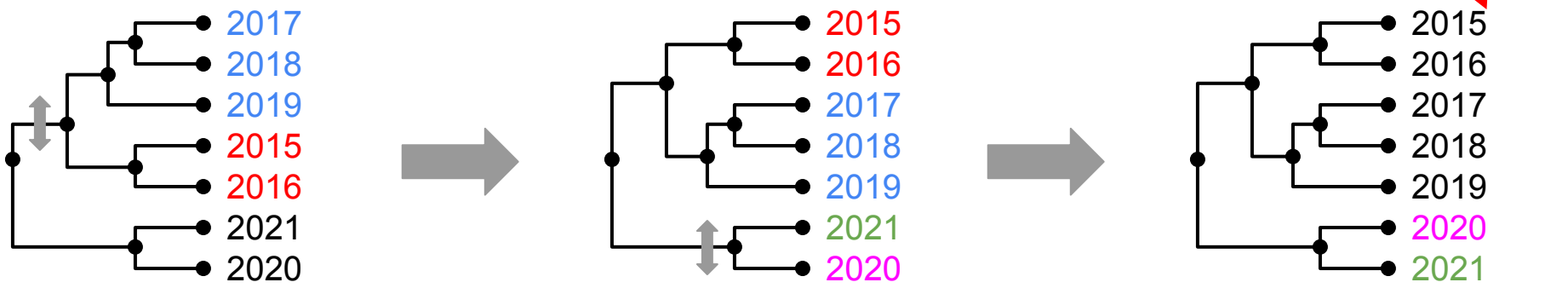
Our first sub-problem

Reordering the (dated) leaves of a tree/dendrogram

in order to

best fit with the chronology

First example:



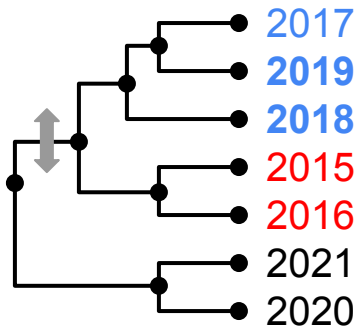
Our first sub-problem

Reordering the (dated) leaves of a tree/dendrogram

in order to

best fit with the chronology

Second example:



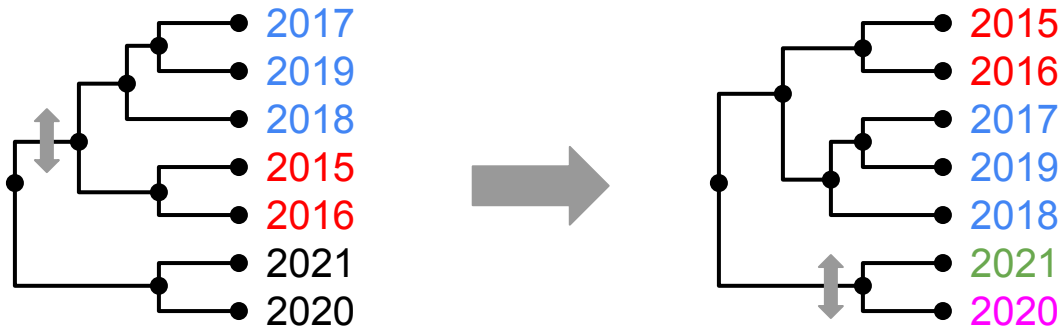
Our first sub-problem

Reordering the (dated) leaves of a tree/dendrogram

in order to

best fit with the chronology

Second example:



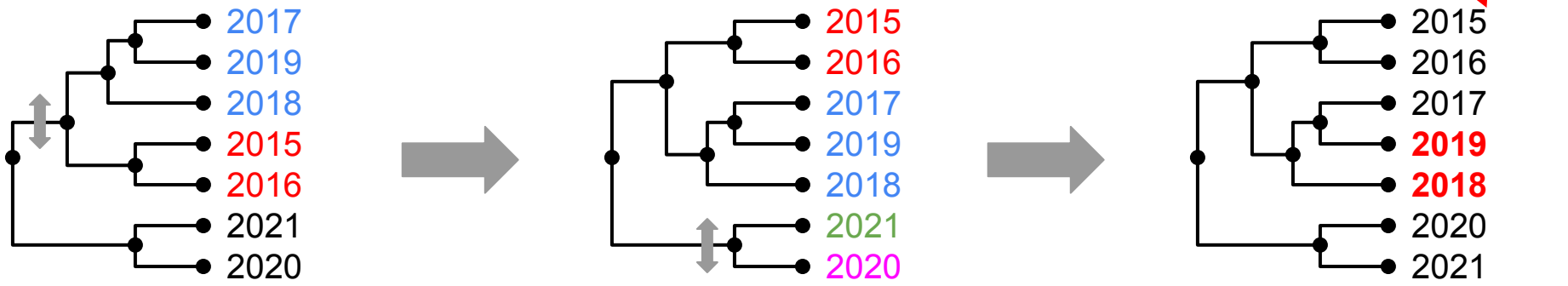
Our first sub-problem

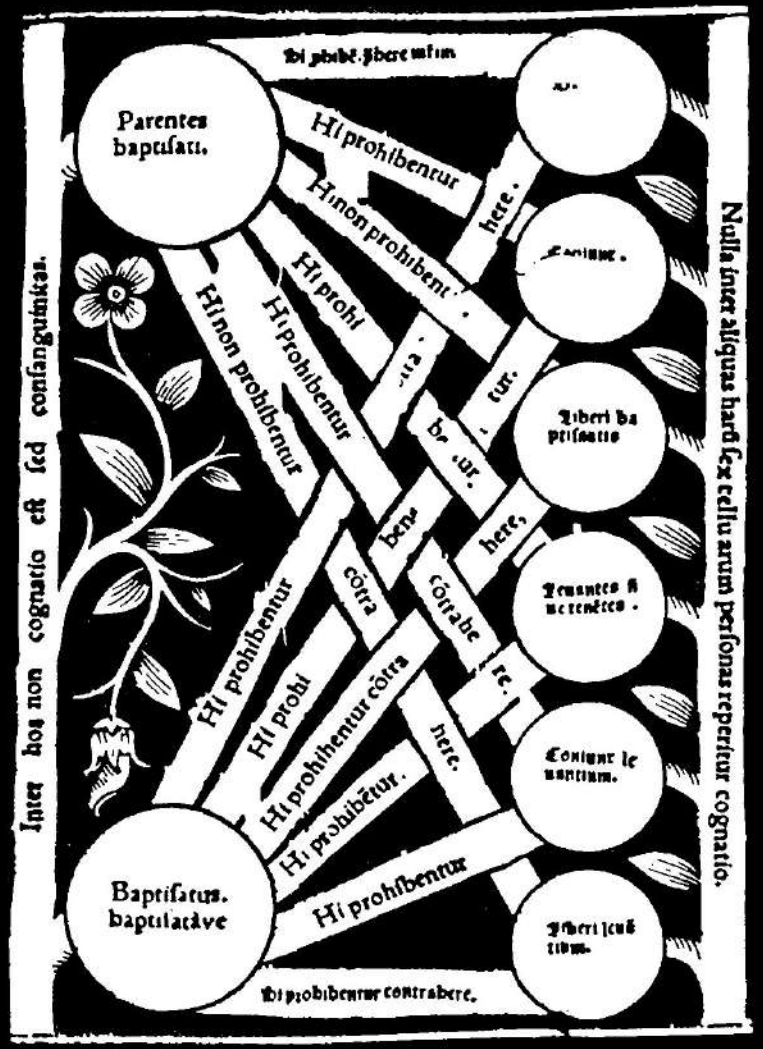
Reordering the (dated) leaves of a tree/dendrogram

in order to

best fit with the chronology

Second example:



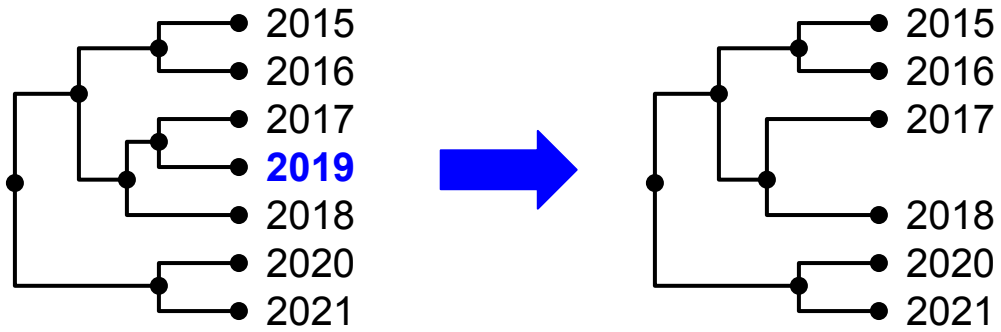


2 criteria to evaluate the consistency between a tree and a chronological order

Source: Arbre généalogique (Tholosae, 1542), Bibliothèque municipale de Toulouse, Gallica btv1b10585389q

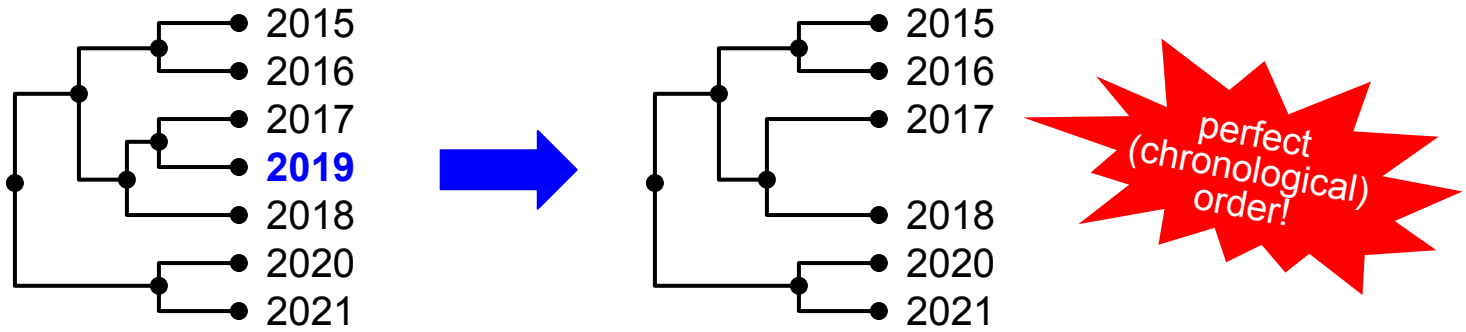
Our first criterion

Our first criterion: the number of leaves to remove



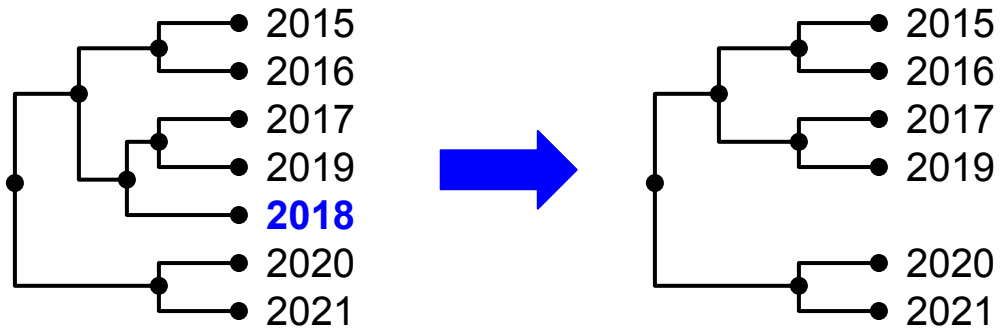
Removing leaf 2019 (or 2018 or 2017) makes the tree consistent with the chronology

Our first criterion: the number of leaves to remove



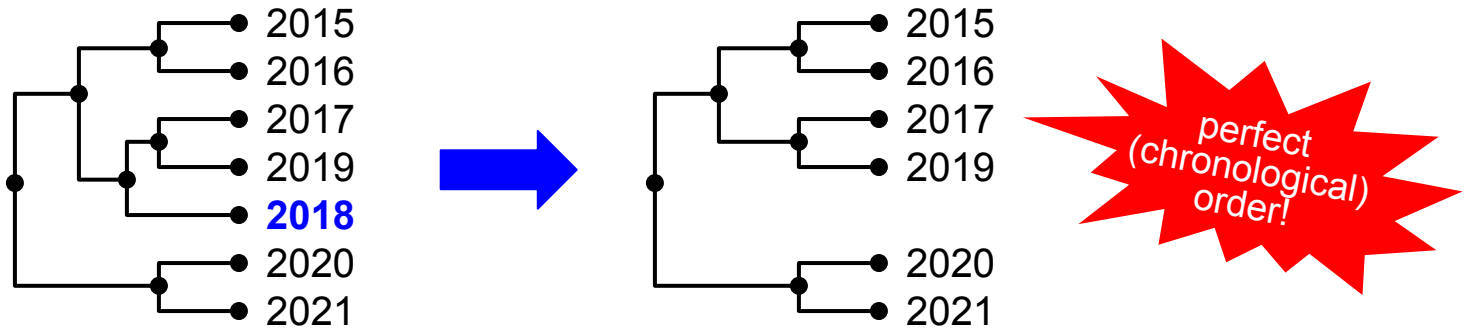
Removing leaf 2019 (or 2018 or 2017) makes the tree **consistent with the chronology**

Our first criterion: the number of leaves to remove



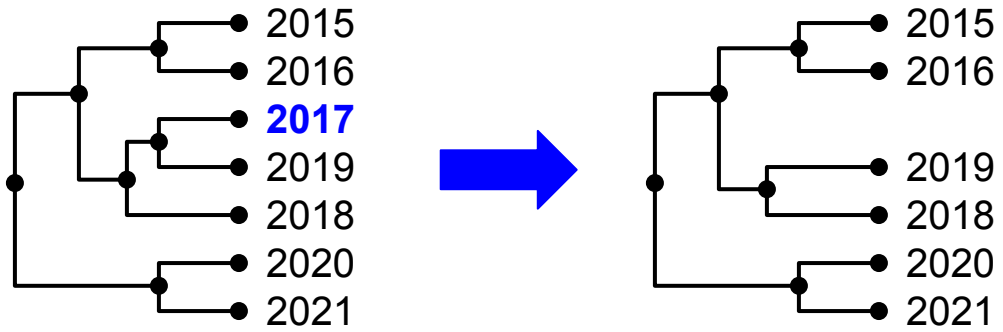
Removing leaf 2019 (or **2018** or 2017) makes the tree consistent with the chronology

Our first criterion: the number of leaves to remove



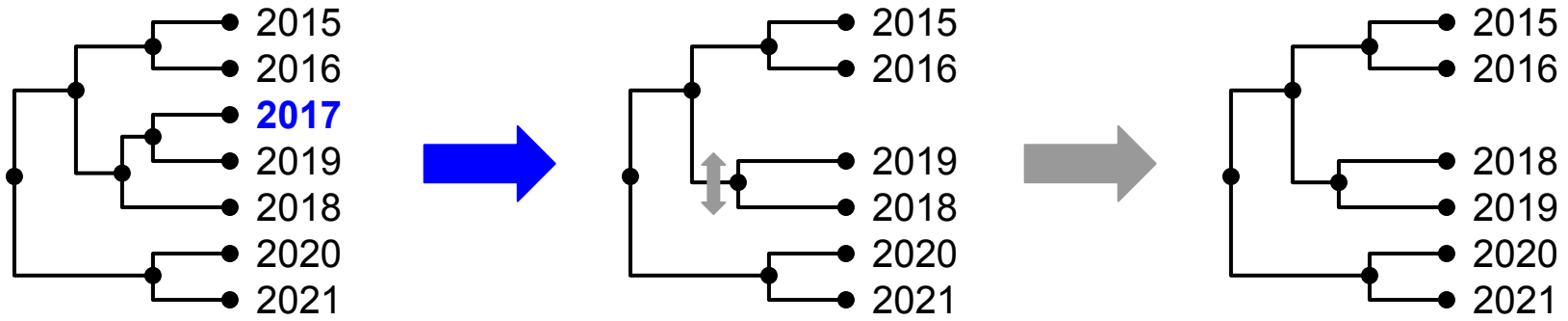
Removing leaf 2019 (or **2018** or 2017) makes the tree **consistent with the chronology**

Our first criterion: the number of leaves to remove



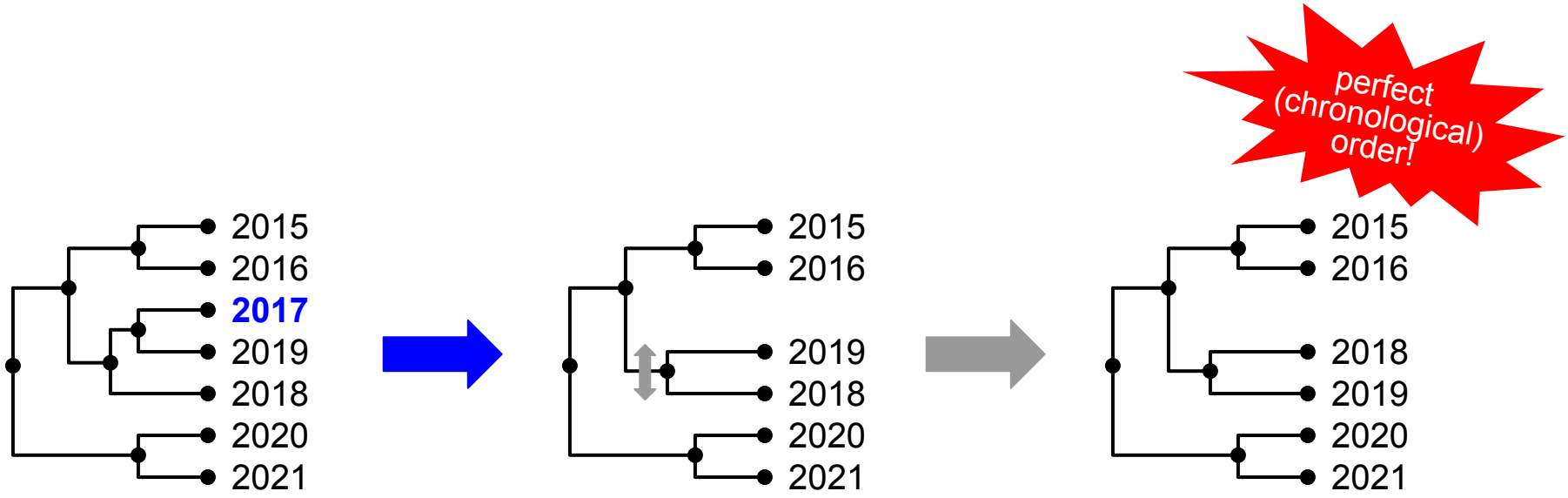
Removing leaf 2019 (or 2018 or **2017**) makes the tree consistent with the chronology

Our first criterion: the number of leaves to remove



Removing leaf 2019 (or 2018 or **2017**) makes the tree consistent with the chronology

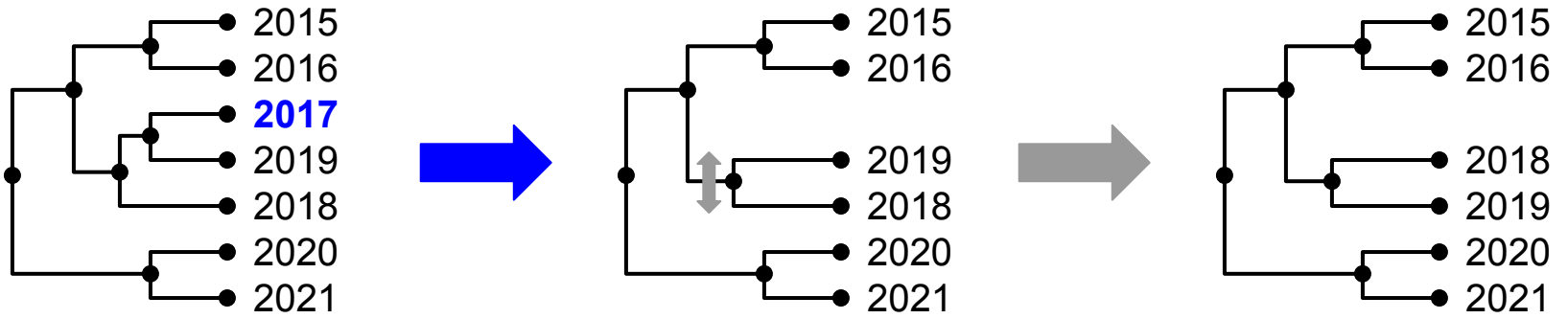
Our first criterion: the number of leaves to remove



Removing leaf 2019 (or 2018 or **2017**) makes the tree **consistent with the chronology**

Our first criterion: the number of leaves to remove

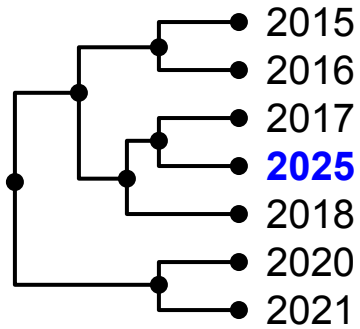
Our **first criterion**:
Reorder the leaves of the tree so that **the minimum number of leaves needs to be removed** to make the tree consistent with the chronology



Removing leaf 2019 (or 2018 or **2017**) makes the tree **consistent with the chronology**

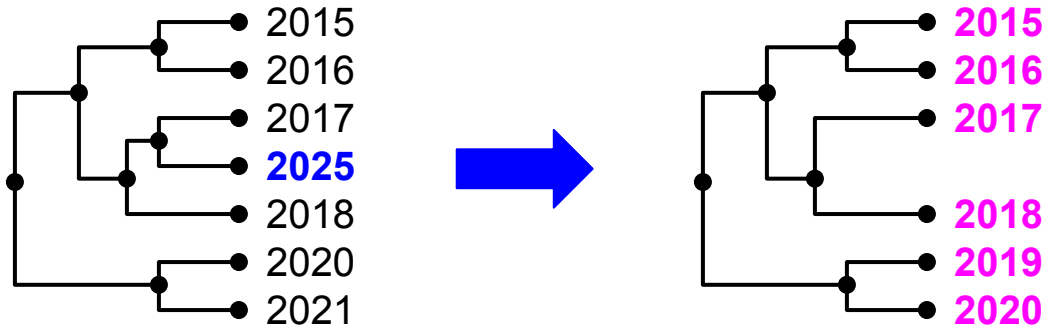
Another criterion?

Another example:



Another criterion?

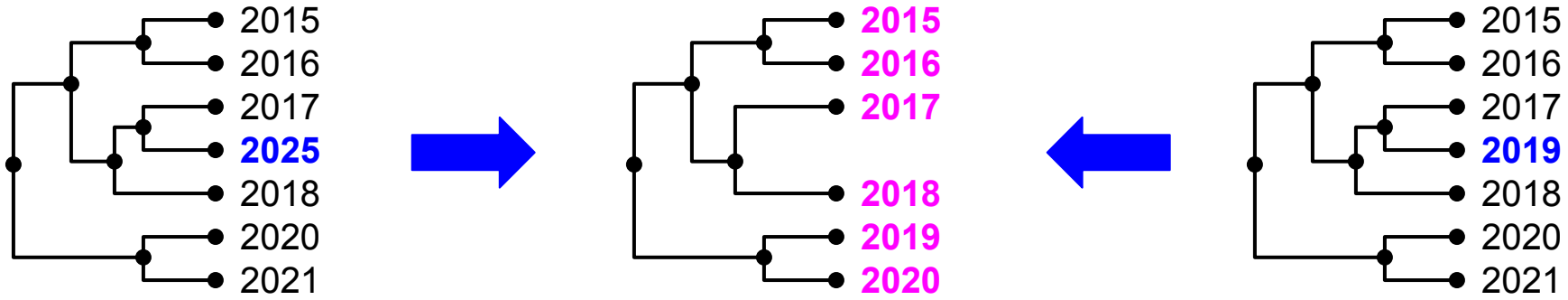
Another example:



Removing one leaf (2025) makes the tree **consistent with the chronology**

Another criterion?

Another example:

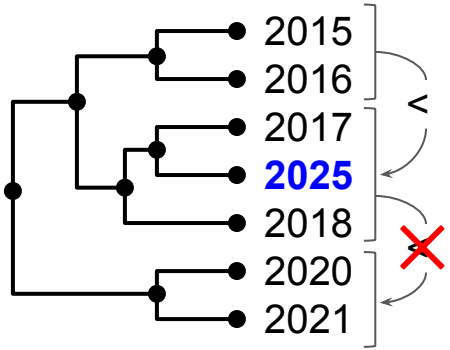


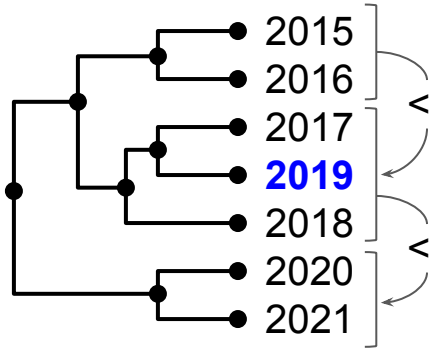
Removing one leaf (2025 or 2019) makes the tree **consistent with the chronology**

… but the 2025-tree seems “less consistent with the chronology” than the 2019-tree

Another criterion?

Another example:



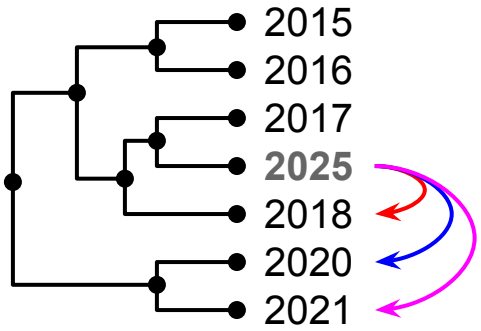


Removing one leaf (2025 or 2019) makes the tree consistent with the chronology

… but the 2025-tree seems "**less consistent with the chronology**" than the 2019-tree

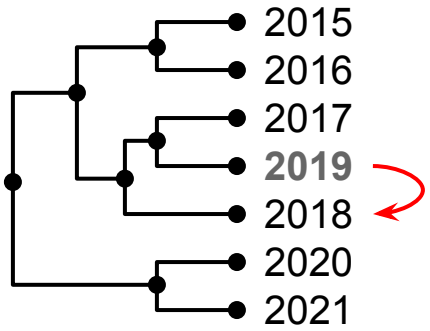
Our second criterion

Example 1:



3 conflicts: **2025>2018**, **2025>2020** and **2025>2021**

Example 2:



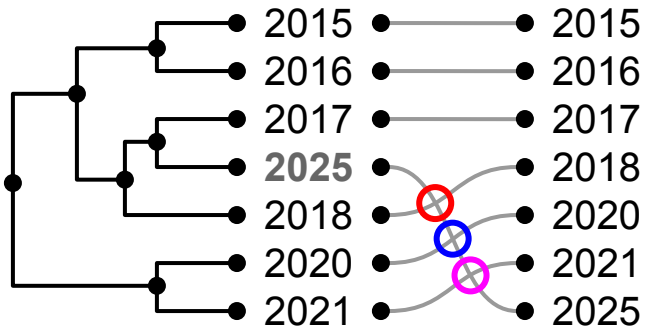
1 conflict: **2019>2018**

Our second criterion

1 conflict =
1 crossing

perfect
chronological
order

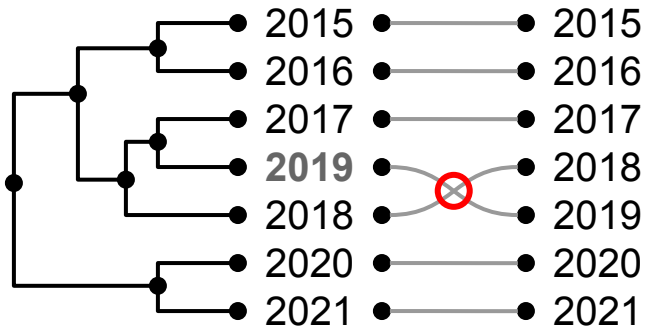
Example 1:



3 conflicts: **2025>2018**,
2025>2020 and **2025>2021**

perfect
chronological
order

Example 2:

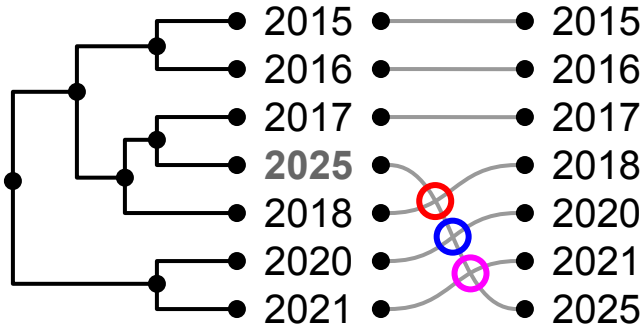


1 conflict: **2019>2018**

Our second criterion: the number of conflicts

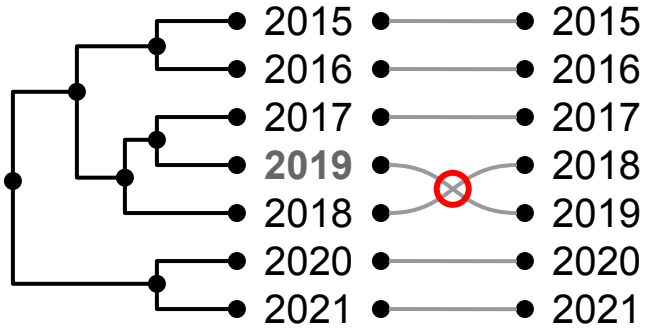
Our **second criterion**:
Reorder the leaves of the tree so that **the minimum number of conflicts** with the chronological order remain

Example 1:



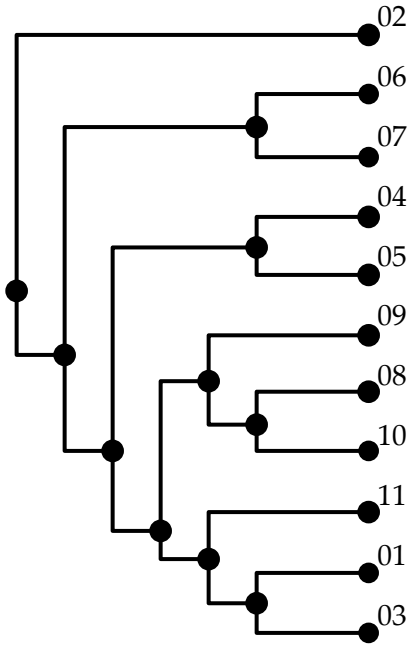
3 conflicts: **2025>2018**,
2025>2020 and **2025>2021**

Example 2:

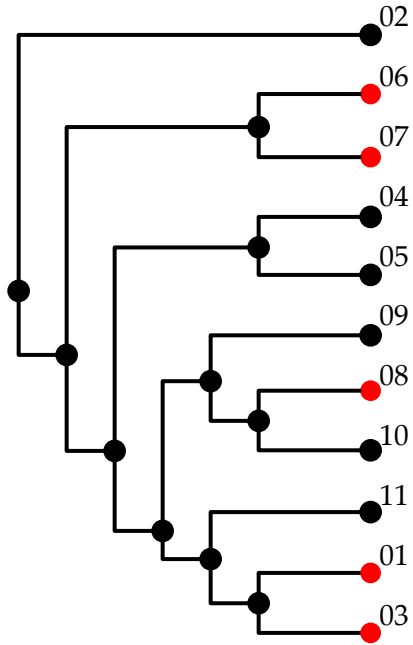


1 conflict: **2019>2018**

The two criteria are not equivalent!

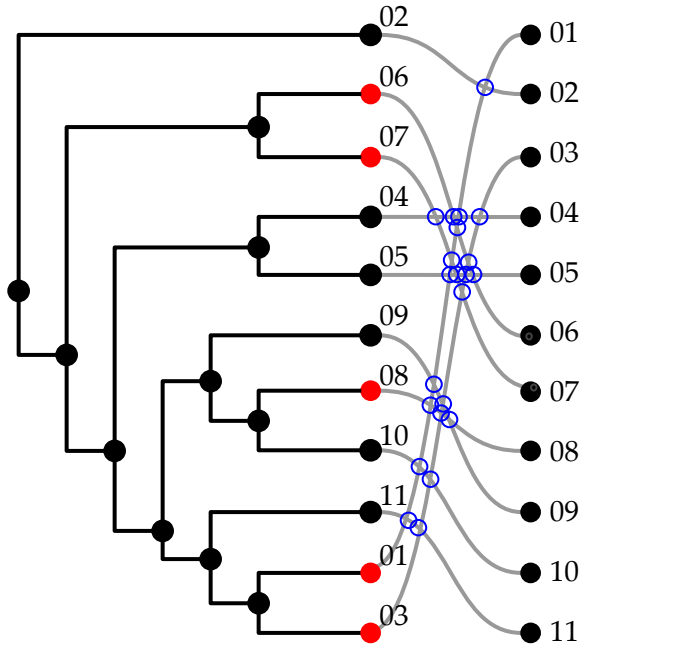


The two criteria are not equivalent!

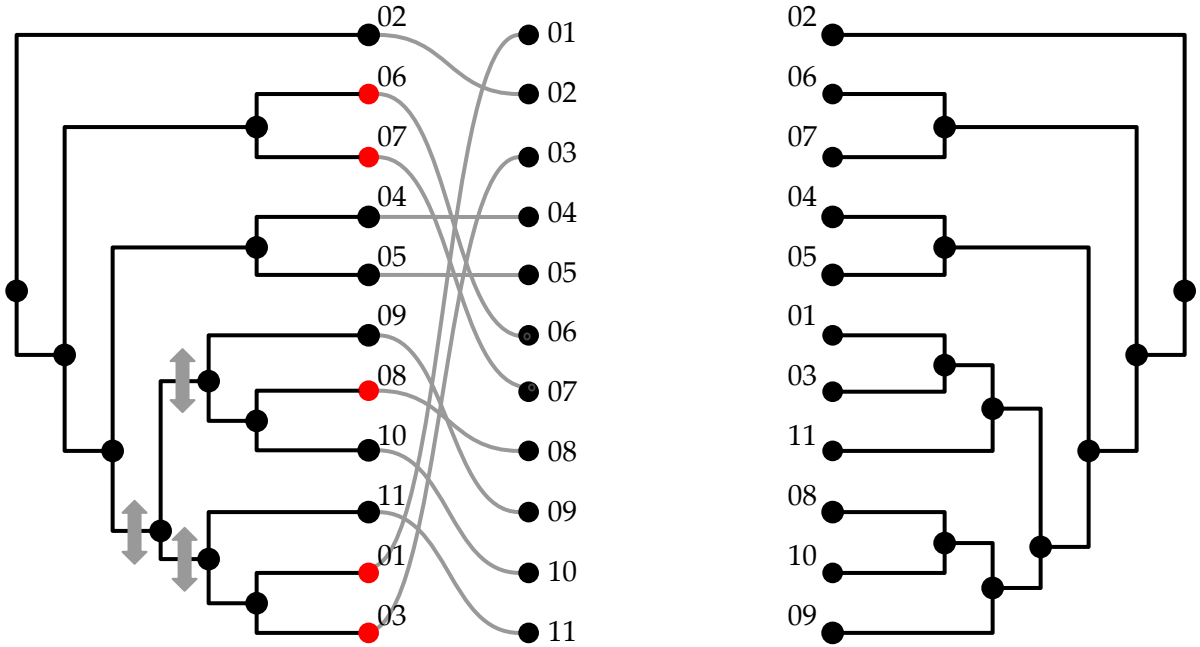


5 leaves to delete (optimal)

The two criteria are not equivalent!



The two criteria are not equivalent!



order 1

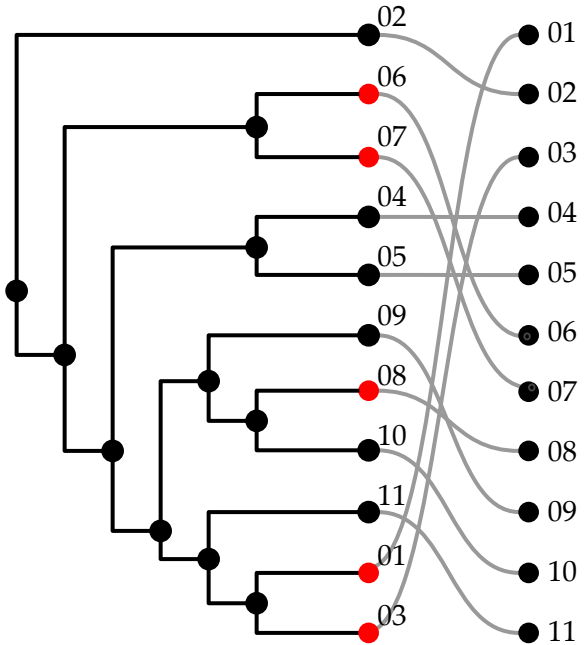
22 crossings

5 leaves to delete (optimal)

chronological order

order 2 of the same tree

The two criteria are not equivalent!

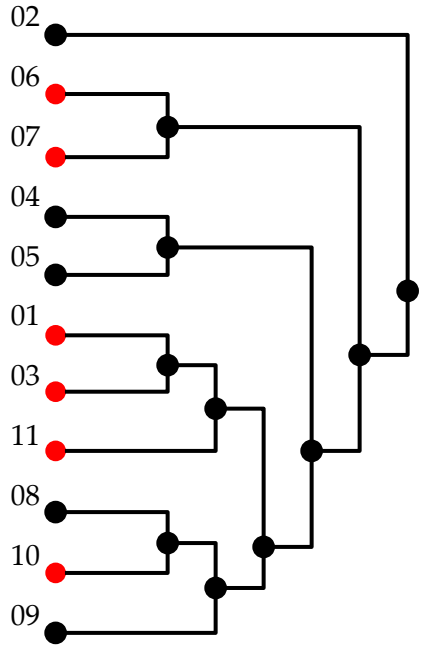


order 1

22 crossings

5 leaves to delete (optimal)

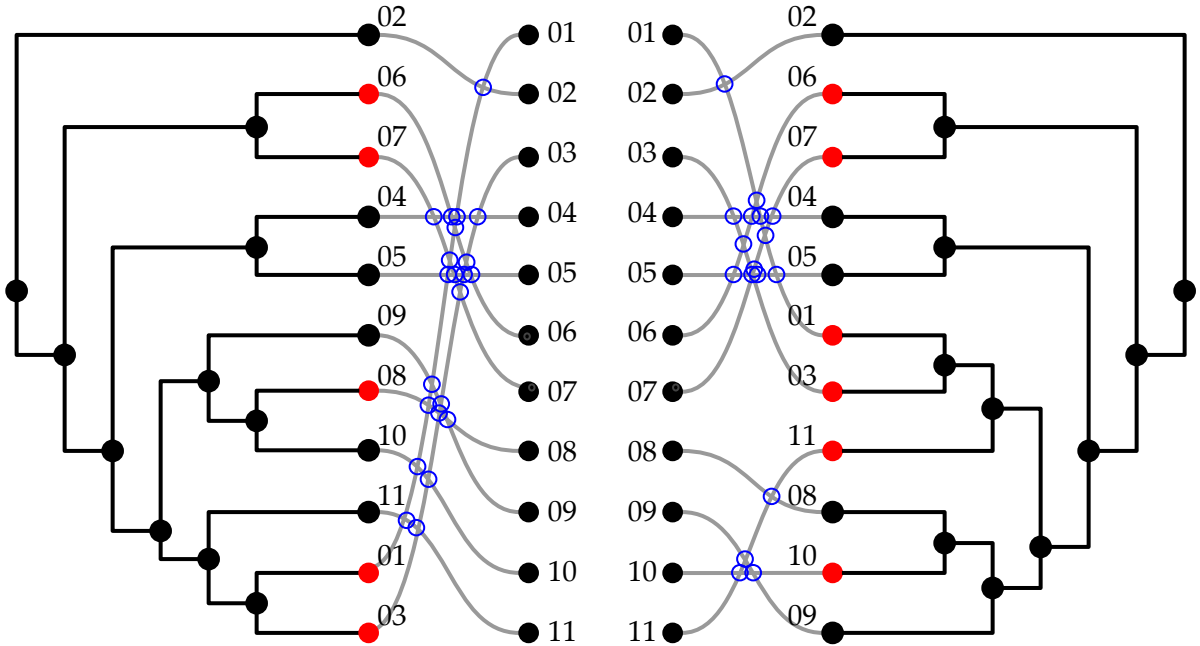
chronological order



order 2 of the same tree

6 leaves to delete

The two criteria are not equivalent!



order 1

22 crossings

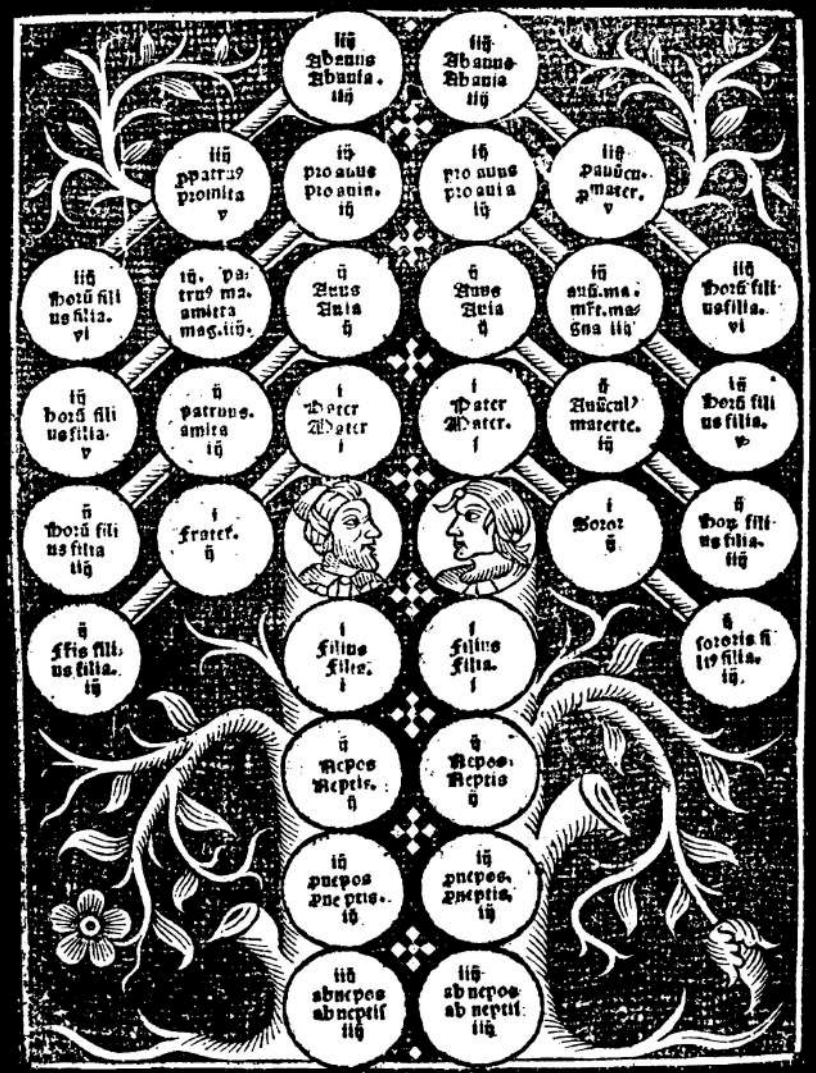
5 leaves to delete (optimal)

chronological order

order 2 of the same tree

17 crossings (optimal)

6 leaves to delete



Finding the optimal order for each criterion

Source: Arbre généalogique (Tholosae, 1542), Bibliothèque municipale de Toulouse, Gallica btv1b10585385x

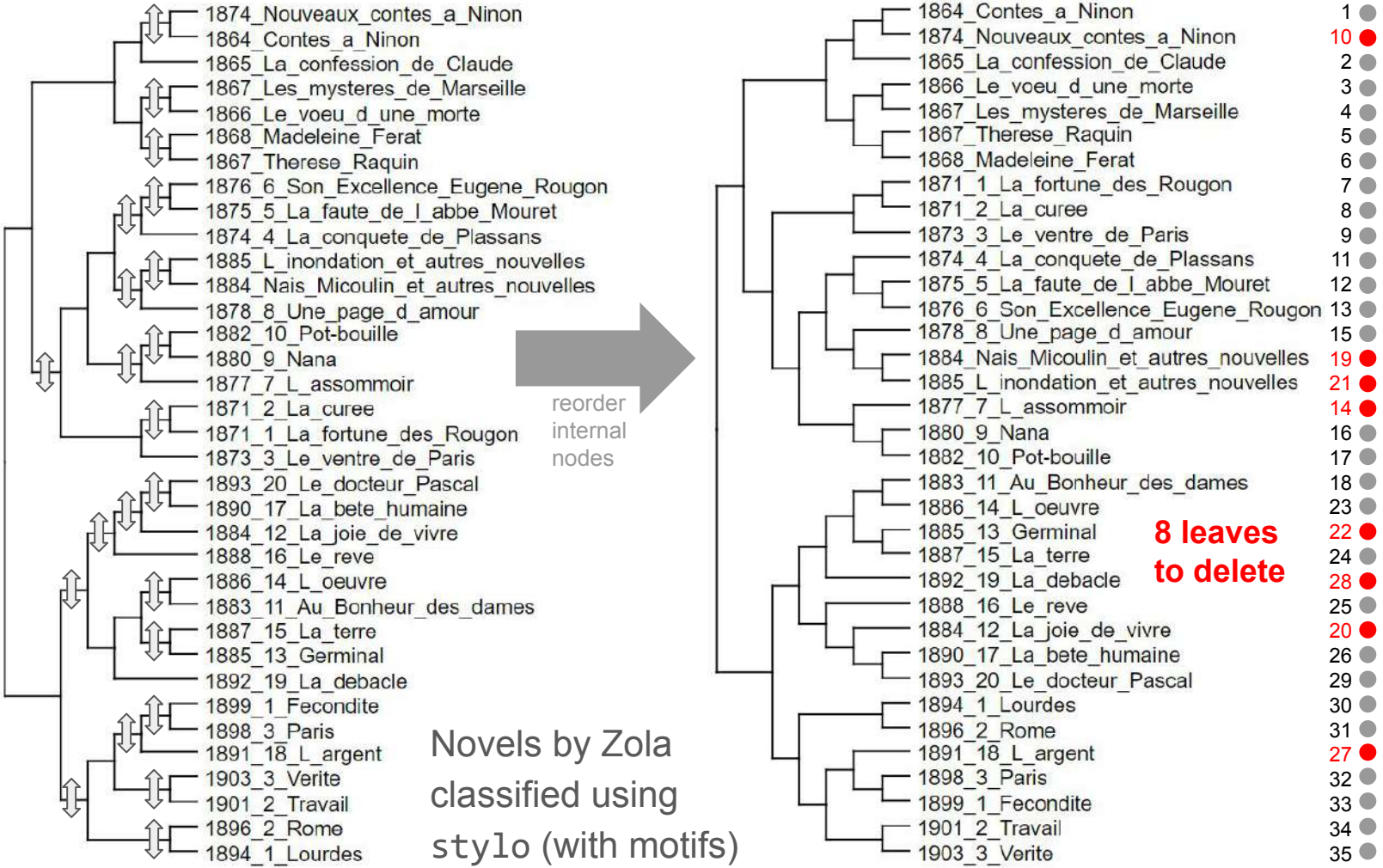
Minimizing the number of leaves to delete / conflicts

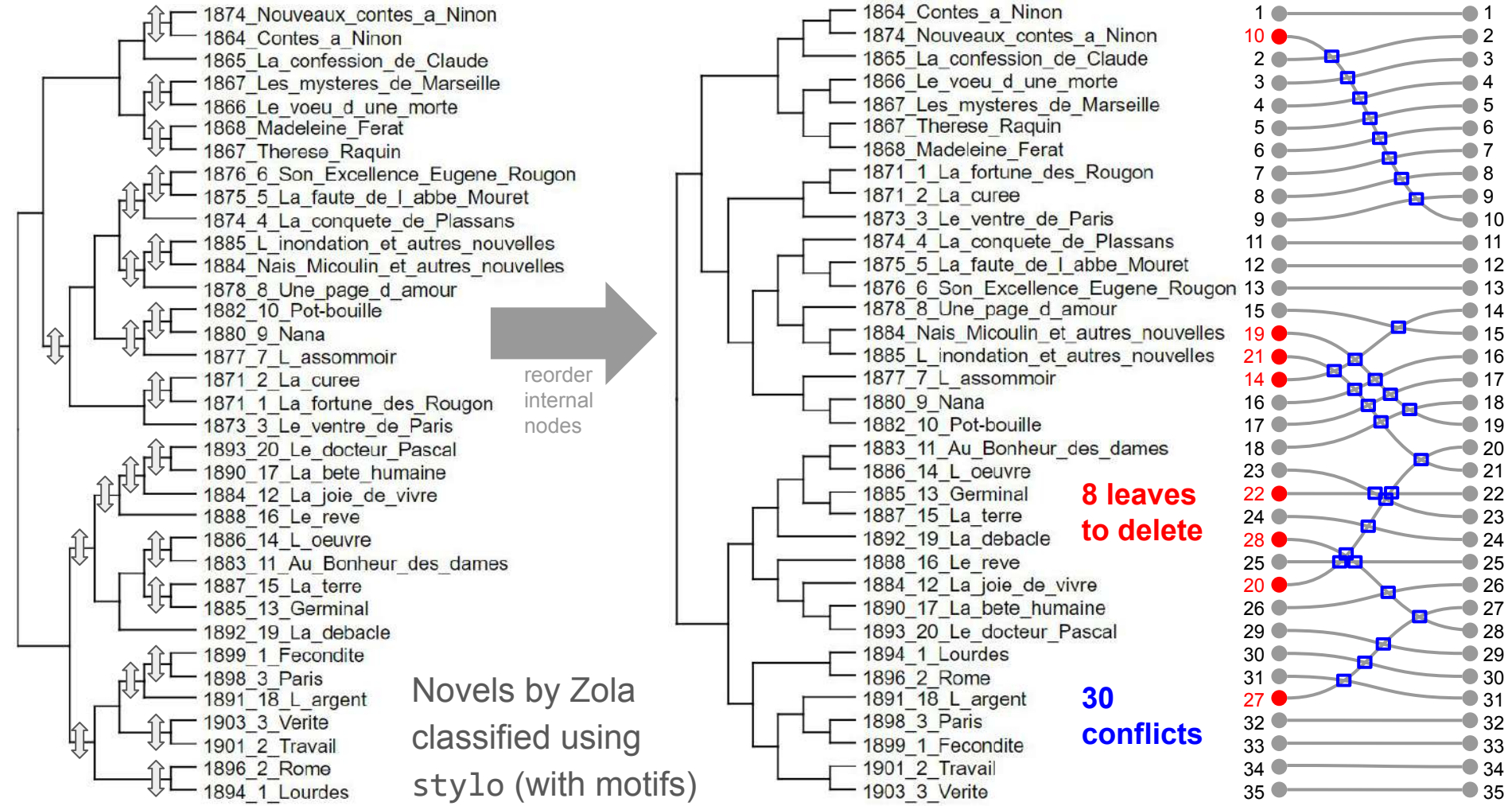
Two algorithms to find an optimal order:

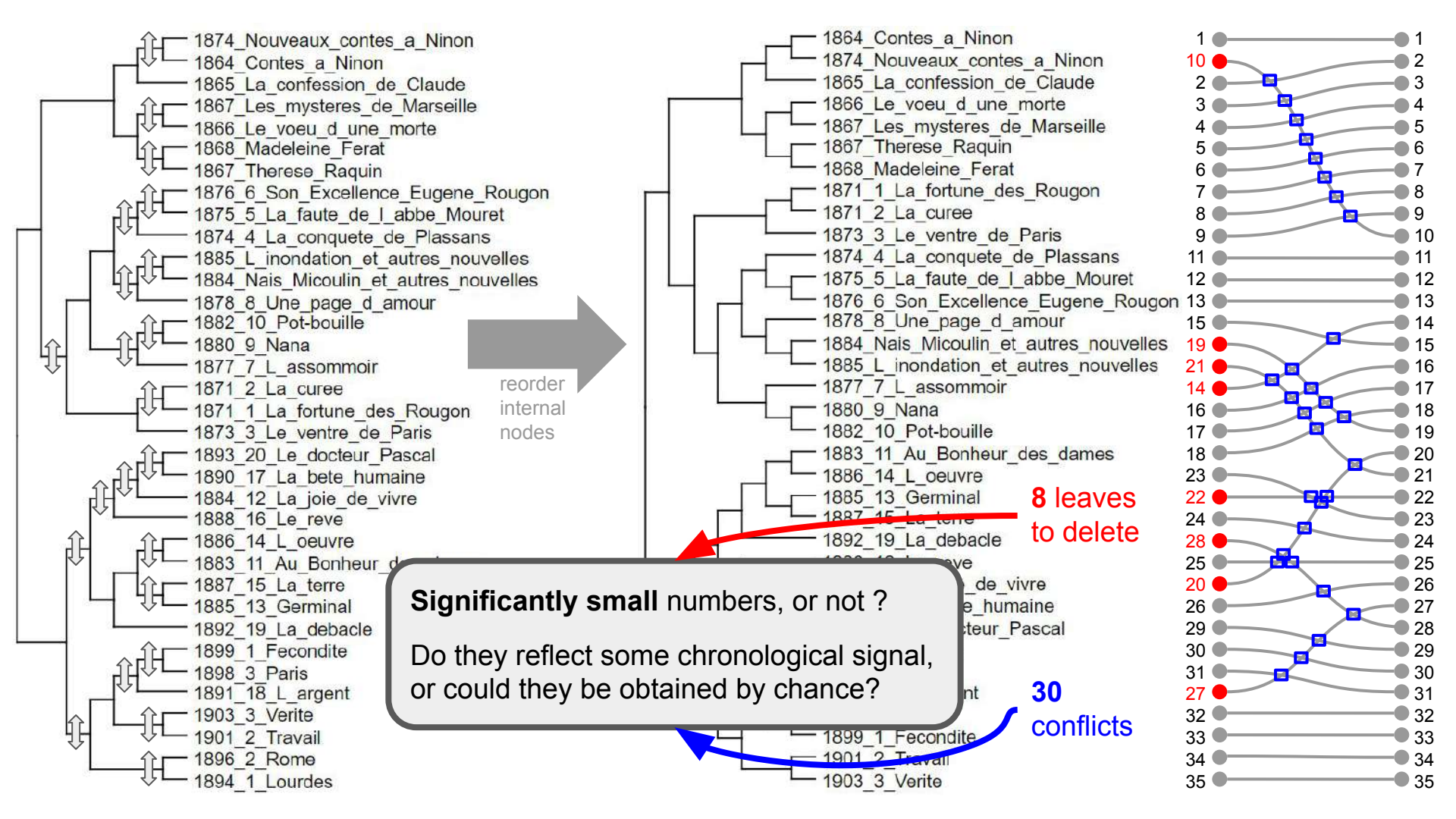
1. minimizing the number of leaves to delete
    - a new dynamic programming algorithm
2. minimizing the number of conflicts
    - an algorithm from bioinformatics: Venkatachalam, Apple, St. John & Gusfield, 2010

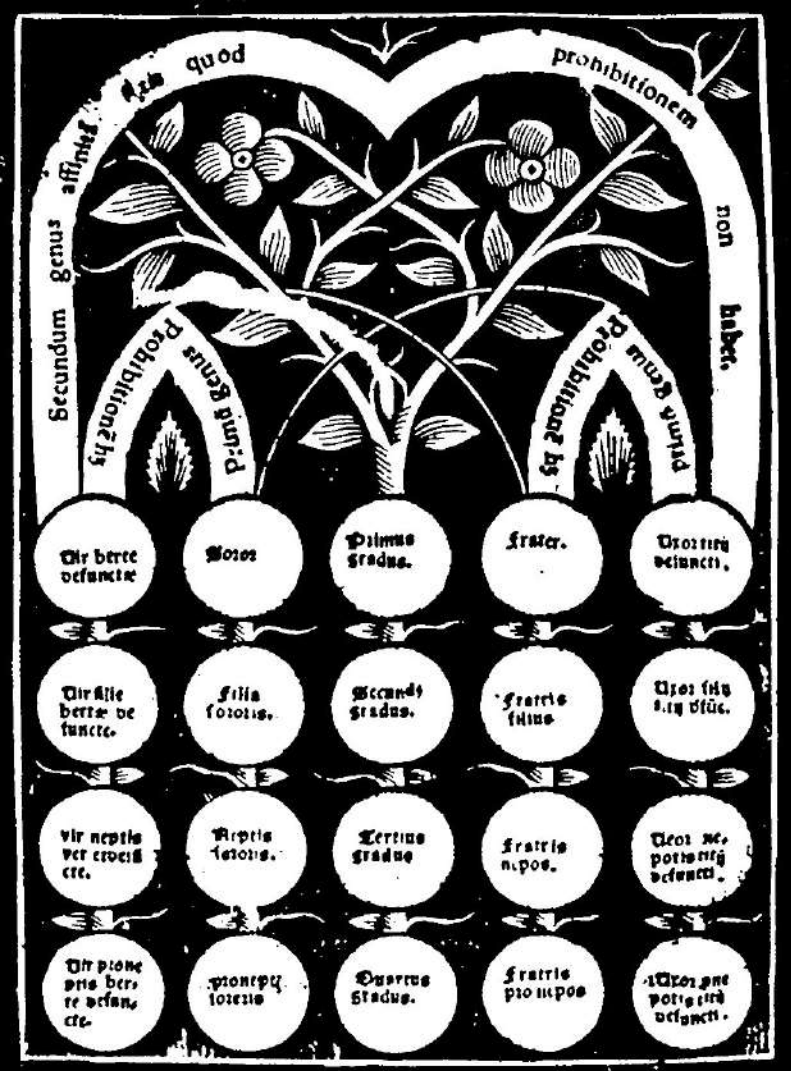
⇒ both quick (polynomial time algorithms) if each node of the tree has a small number of children

⇒ both implemented in Python and available at https://github.com/oseminck/tree_order_evaluation


1874_Nouveaux_contes_a_Ninon
1864_Contes_a_Ninon
1865_La_confession_de_Claude
1867_Les_mysteres_de_Marseille
1866_Le_voeu_d_une_morte
1868_Madeleine_Ferat
1867_Therese_Raquin
1876_6_Son_Excellence_Eugene_Rougon
1875_5_La_faute_de_l_abbe_Mouret
1874_4_La_conquete_de_Plassans
1885_L_inondation_et_autres_nouvelles
1884_Nais_Micoulin_et_autres_nouvelles
1878_8_Une_page_d_amour
1882_10_Pot-bouille
1880_9_Nana
1877_7_L_assommoir
1871_2_La_curee
1871_1_La_fortune_des_Rougon
1873_3_Le_ventre_de_Paris
1893_20_Le_docteur_Pascal
1890_17_La_bete_humaine
1884_12_La_joie_de_vivre
1888_16_Le_reve
1886_14_L_oeuvre
1883_11_Au_Bonheur_des_dames
1887_15_La_terre
1885_13_Germinal
1892_19_La_debacle
1899_1_Fecondite
1898_3_Paris
1891_18_L_argent
1903_3_Verite
1901_2_Travail
1896_2_Rome
1894_1_Lourdes
reorder
internal
nodes
1864_Contes_a_Ninon 1
1874_Nouveaux_contes_a_Ninon 10
1865_La_confession_de_Claude 2
1866_Le_voeu_d_une_morte 3
1867_Les_mysteres_de_Marseille 4
1867_Therese_Raquin 5
1868_Madeleine_Ferat 6
1871_1_La_fortune_des_Rougon 7
1871_2_La_curee 8
1873_3_Le_ventre_de_Paris 9
1874_4_La_conquete_de_Plassans 11
1875_5_La_faute_de_l_abbe_Mouret 12
1876_6_Son_Excellence_Eugene_Rougon 13
1878_8_Une_page_d_amour 15
1884_Nais_Micoulin_et_autres_nouvelles 19
1885_L_inondation_et_autres_nouvelles 21
1877_7_L_assommoir 14
1880_9_Nana 16
1882_10_Pot-bouille 17
1883_11_Au_Bonheur_des_dames 18
1886_14_L_oeuvre 23
1885_13_Germinal 22
1887_15_La_terre 24
1892_19_La_debacle 28
1888_16_Le_reve 25
1884_12_La_joie_de_vivre 20
1890_17_La_bete_humaine 26
1893_20_Le_docteur_Pascal 29
1894_1_Lourdes 30
1896_2_Rome 31
1891_18_L_argent 27
1898_3_Paris 32
1899_1_Fecondite 33
1901_2_Travail 34
1903_3_Verite 35
8 leaves
to delete
Novels by Zola
classified using
stylo (with motifs)


1874_Nouveaux_contes_a_Ninon
1864_Contes_a_Ninon
1865_La_confession_de_Claude
1867_Les_mysteres_de_Marseille
1866_Le_voeu_d_une_morte
1868_Madeleine_Ferat
1867_Therese_Raquin
1876_6_Son_Excellence_Eugene_Rougon
1875_5_La_faute_de_l_abbe_Mouret
1874_4_La_conquete_de_Plassans
1885_L_inondation_et_autres_nouvelles
1884_Nais_Micoulin_et_autres_nouvelles
1878_8_Une_page_d_amour
1882_10_Pot-bouille
1880_9_Nana
1877_7_L_assommoir
1871_2_La_curee
1871_1_La_fortune_des_Rougon
1873_3_Le_ventre_de_Paris
1893_20_Le_docteur_Pascal
1890_17_La_bete_humaine
1884_12_La_joie_de_vivre
1888_16_Le_reve
1886_14_L_oeuvre
1883_11_Au_Bonheur_des_dames
1887_15_La_terre
1885_13_Germinal
1892_19_La_debacle
1899_1_Fecondite
1898_3_Paris
1891_18_L_argent
1903_3_Verite
1901_2_Travail
1896_2_Rome
1894_1_Lourdes
reorder
internal
nodes
Novels by Zola
classified using
stylo (with motifs)
1864_Contes_a_Ninon
1874_Nouveaux_contes_a_Ninon
1865_La_confession_de_Claude
1866_Le_voeu_d_une_morte
1867_Les_mysteres_de_Marseille
1867_Therese_Raquin
1868_Madeleine_Ferat
1871_1_La_fortune_des_Rougon
1871_2_La_curee
1873_3_Le_ventre_de_Paris
1874_4_La_conquete_de_Plassans
1875_5_La_faute_de_l_abbe_Mouret
1876_6_Son_Excellence_Eugene_Rougon
1878_8_Une_page_d_amour
1884_Nais_Micoulin_et_autres_nouvelles
1885_L_inondation_et_autres_nouvelles
1877_7_L_assommoir
1880_9_Nana
1882_10_Pot-bouille
1883_11_Au_Bonheur_des_dames
1886_14_L_oeuvre
1885_13_Germinal
1887_15_La_terre
1892_19_La_debacle
1888_16_Le_reve
1884_12_La_joie_de_vivre
1890_17_La_bete_humaine
1893_20_Le_docteur_Pascal
1894_1_Lourdes
1896_2_Rome
1891_18_L_argent
1898_3_Paris
1899_1_Fecondite
1901_2_Travail
1903_3_Verite
8 leaves
to delete
30
conflicts
1 1
10 2
2 3
3 4
4 5
5 6
6 7
7 8
8 9
9 10
11 11
12 12
13 13
15 14
19 15
21 16
14 17
16 18
17 19
18 20
23 21
22 22
24 23
28 24
25 25
20 26
26 27
29 28
30 29
31 30
27 31
32 32
33 33
34 34
35 35


1874_Nouveaux_contes_a_Ninon
1864_Contes_a_Ninon
1865_La_confession_de_Claude
1867_Les_mysteres_de_Marseille
1866_Le_voeu_d_une_morte
1868_Madeleine_Ferat
1867_Therese_Raquin
1876_6_Son_Excellence_Eugene_Rougon
1875_5_La_faute_de_l_abbe_Mouret
1874_4_La_conquete_de_Plassans
1885_L_inondation_et_autres_nouvelles
1884_Nais_Micoulin_et_autres_nouvelles
1878_8_Une_page_d_amour
1882_10_Pot-bouille
1880_9_Nana
1877_7_L_assommoir
1871_2_La_curee
1871_1_La_fortune_des_Rougon
1873_3_Le_ventre_de_Paris
1893_20_Le_docteur_Pascal
1890_17_La_bete_humaine
1884_12_La_joie_de_vivre
1888_16_Le_reve
1886_14_L_oeuvre
1887_15_La_terre
1885_13_Germinal
1892_19_La_debacle
1899_1_Fecondite
1898_3_Paris
1891_18_L_argent
1903_3_Verite
1901_2_Travail
1896_2_Rome
1894_1_Lourdes
reorder
internal
nodes
1864_Contes_a_Ninon
1874_Nouveaux_contes_a_Ninon
1865_La_confession_de_Claude
1866_Le_voeu_d_une_morte
1867_Les_mysteres_de_Marseille
1867_Therese_Raquin
1868_Madeleine_Ferat
1871_1_La_fortune_des_Rougon
1871_2_La_curee
1873_3_Le_ventre_de_Paris
1874_4_La_conquete_de_Plassans
1875_5_La_faute_de_l_abbe_Mouret
1876_6_Son_Excellence_Eugene_Rougon
1878_8_Une_page_d_amour
1884_Nais_Micoulin_et_autres_nouvelles
1885_L_inondation_et_autres_nouvelles
1877_7_L_assommoir
1880_9_Nana
1882_10_Pot-bouille
1883_11_Au_Bonheur_des_dames
1886_14_L_oeuvre
1885_13_Germinal
1892_19_La_debacle
1899_1_Fecondite
1901_2_Travail
1903_3_Verite
8 leaves
to delete
30
conflicts
Significantly small numbers, or not ?
Do they reflect some chronological signal,
or could they be obtained by chance?
1 1
10 2
2 3
3 4
4 5
5 6
6 7
7 8
8 9
9 10
11 11
12 12
13 13
15 14
19 15
21 16
14 17
16 18
17 19
18 20
23 21
22 22
24 23
28 24
25 25
20 26
26 27
29 28
30 29
31 30
27 31
32 32
33 33
34 34
35 35



Evaluating the significance of the obtained results

Source: Arbre généalogique (Tholosae, 1542), Bibliothèque municipale de Toulouse, Gallica btv1b10585387t

Our second sub-problem

Is our dendrogram really **consistent** with the chronology or not?

Our second sub-problem

Is our dendrogram really **consistent** with the chronology or not?

⇒ could **the same values** be obtained **by chance**, without chronological signal?

Our second sub-problem

Is our dendrogram really **consistent** with the chronology or not?

⇒ could **the same values** be obtained **by chance**, without chronological signal?

Method to estimate some “p-value” of the result:

1. Generate 10 000 random orders of the leaves
2. Compute the smallest number of leaves to delete / conflicts for each order
3. Count how many random orders get a result as low as the chronological order

Our second sub-problem

Is our dendrogram really **consistent** with the chronology or not?

⇒ could **the same values** be obtained **by chance**, without chronological signal?

Method to estimate some “p-value” of the result:

1. Generate 10 000 random orders of the leaves
2. Compute the smallest number of leaves to delete / conflicts for each order
3. Count how many random orders get a result as low as the chronological order

Example: For 0.2% of the random orders, the number of leaves to delete is as low as for the chronological order ⇒ probably not obtained by chance
⇒ significantly consistent with the chronology

In practice

Provide the tree in the Newick parenthesis format, with leaves labeled according to the order:

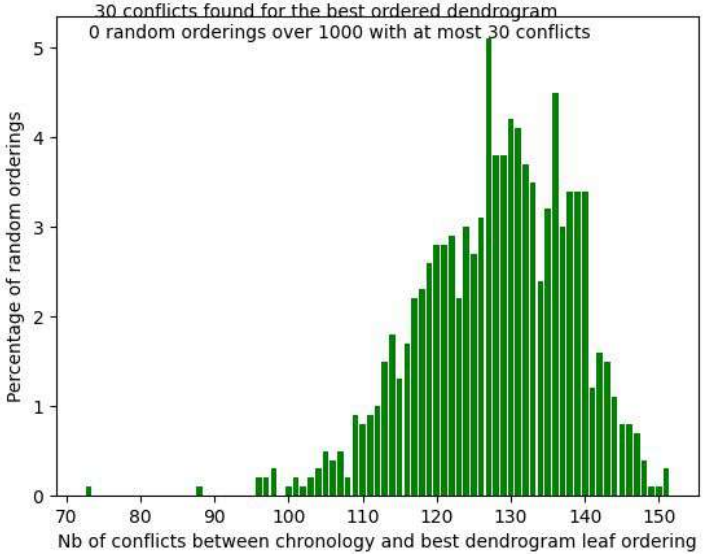
(((((1874a_Nouveaux_contes_a_Ninon,1864_Contes_a_Ninon),1865_La_confession_de_Claude),((1867_Les_mysteres_de_Marseille,1866_Le_voeu_d_une_morte),(1868_Madeleine_Ferat,1867_Therese_Raquin))),(((((1876_6_Son_Excellence_Eugene_Rougon,1875_5_La_faute_de_l_abbe_Mouret),1874b_4_La_conquete_de_Plassans),((1885a_L_inondation_et_autres_nouvelles,1884a_Nais_Micoulin_et_autres_nouvelles),1878_8_Une_page_d_amour)),((1882_10_Pot-bouille,1880_9_Nana),1877_7_L_assommoir)),((1871_2_La_curee,1871_1_La_fortune_des_Rougon),1873_3_Le_ventre_de_Paris))),(((((1893_20_Le_docteur_Pascal,1890_17_La_bete_humaine),1884b_12_La_joie_de_vivre),1888_16_Le_reve),(((1886_14_L_oeuvre,1883_11_Au_Bonheur_des_dames),(1887_15_La_terre,1885b_13_Germinal)),1892_19_La_debacle)),((((1899_1_Fecondite,1898_3_Paris),1891_18_L_argent),(1903_3_Verite,1901_2_Travail)),(1896_2_Rome,1894_1_Lourdes))));

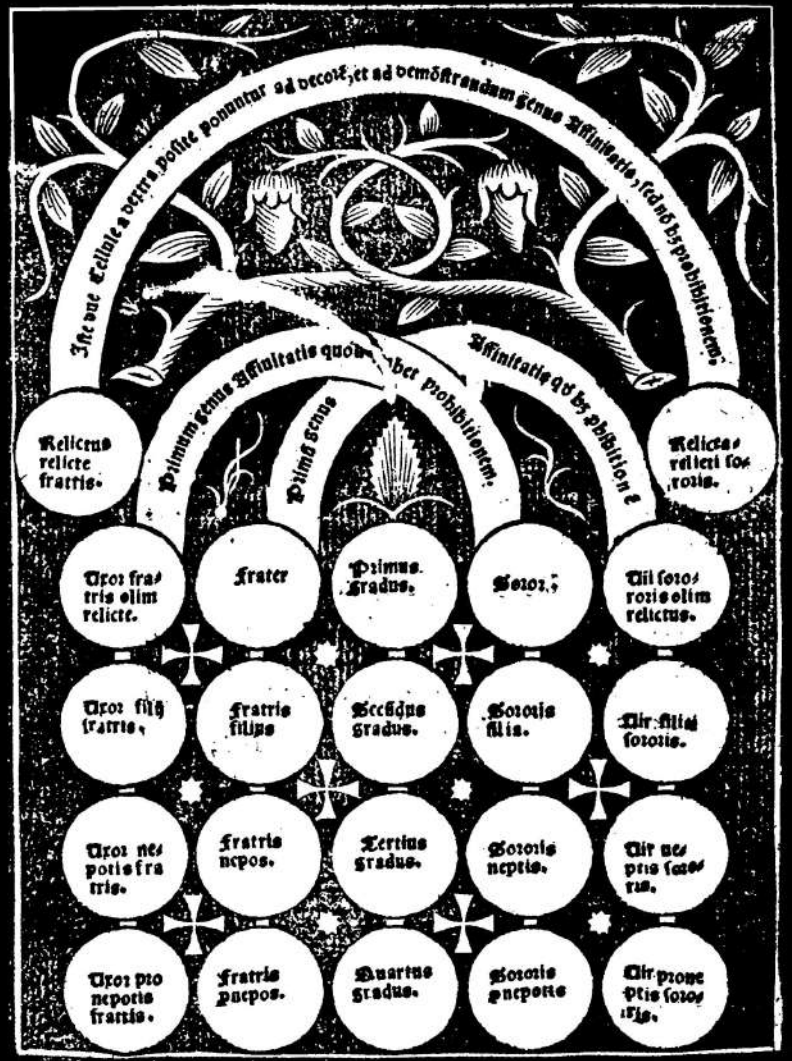
In practice

Provide the tree in the Newick parenthesis format, with leaves labeled according to the order:

(((((1874a_Nouveaux_contes_a_Ninon,1864_Contes_a_Ninon),1865_La_confession_de_Claude),((1867_Les_mysteres_de_Marseille,1866_Le_voeu_d_une_morte),(1868_Madeleine_Ferat,1867_Therese_Raquin))),(((((1876_6_Son_Excellence_Eugene_Rougon,1875_5_La_faute_de_l_abbe_Mouret),1874b_4_La_conquete_de_Plassans),((1885a_L_inondation_et_autres_nouvelles,1884a_Nais_Micoulin_et_autres_nouvelles),1878_8_Une_page_d_amour)),((1882_10_Pot-bouille,1880_9_Nana),1877_7_L_assommoir)),((1871_2_La_curee,1871_1_La_fortune_des_Rougon),1873_3_Le_ventre_de_Paris))),(((((1893_20_Le_docteur_Pascal,1890_17_La_bete_humaine),1884b_12_La_joie_de_vivre),1888_16_Le_reve),(((1886_14_L_oeuvre,1883_11_Au_Bonheur_des_dames),(1887_15_La_terre,1885b_13_Germinal)),1892_19_La_debacle)),((((1899_1_Fecondite,1898_3_Paris),1891_18_L_argent),(1903_3_Verite,1901_2_Travail)),(1896_2_Rome,1894_1_Lourdes))));

Get the criteria, the optimal leaf order for each and the results of the random order simulation:





Conclusion and perspectives

Source: Arbre généalogique (Tholosae, 1542), Bibliothèque municipale de Toulouse, Gallica btv1b10585386c

Conclusion & perspectives

What we provide:

- **two criteria** to evaluate whether **a tree is consistent with an order** on the leaves, and whether it could be caused by chance;
- a **practical tool** in Python to find an **optimal order on the leaves of a tree** to best reflect some given order on the leaves.

⇒ a **new tool and method for textual data analysis**?

Conclusion & perspectives

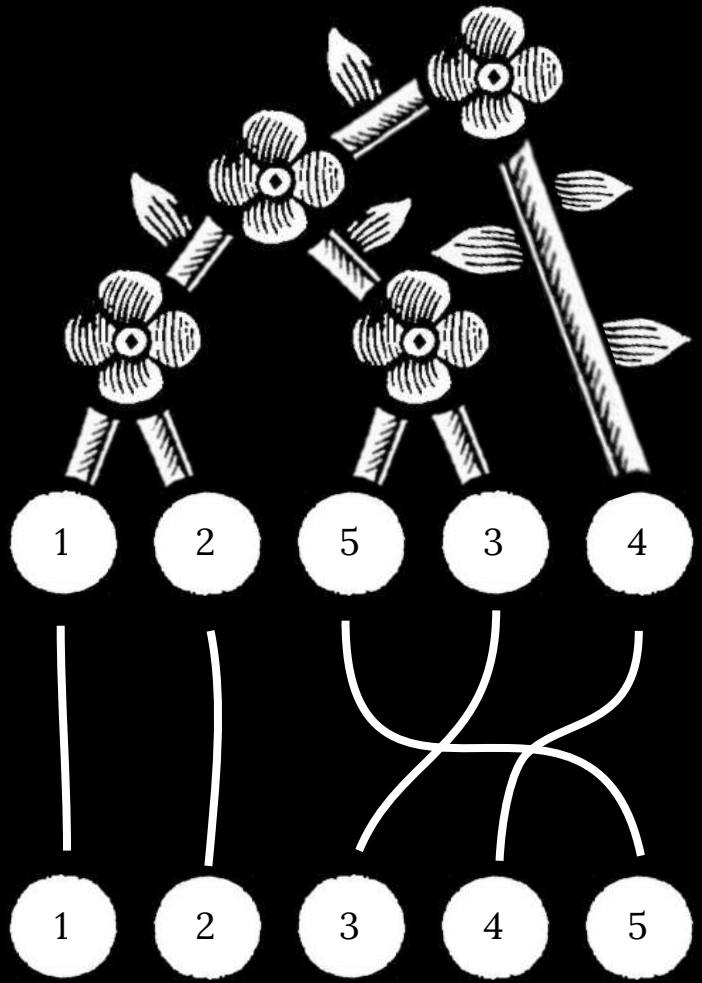
What we provide:

- **two criteria** to evaluate whether **a tree is consistent with an order** on the leaves, and whether it could be caused by chance;
- a **practical tool** in Python to find an **optimal order on the leaves of a tree** to best reflect some given order on the leaves.

⇒ a **new tool and method for textual data analysis**?

What we are still investigating:

- a more direct way to **measure the "chronological signal"** in textual data
- **algorithmics aspects** of the problem: NP-hardness, practical algorithms...
- a **mathematical formula** to evaluate whether the number of leaves to delete or number of conflicts is significantly low or not



Thank you for your attention!

➠ https://github.com/oseminck/tree_order_evaluation

Work supported by the French government under the management of the Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, references ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and ANR-16-IDEX-0003 (I-Site Future, programme “Cité des dames, créatrices dans la cité”).