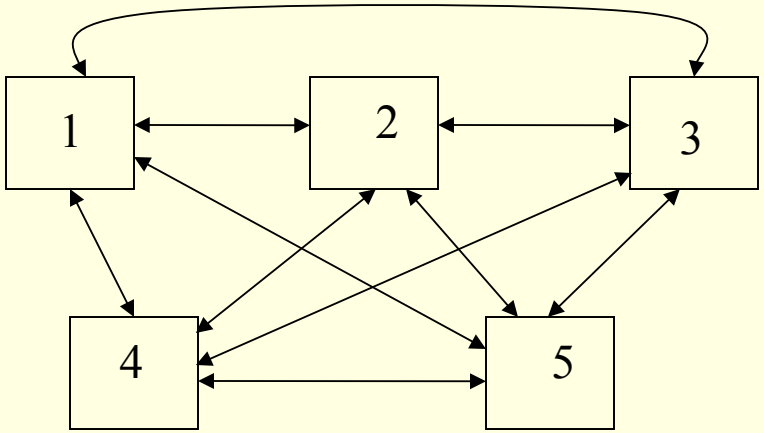
Lecture 4. Information Distance

Textbook, Sect. 8.3, 8.4 (3rd ed.) and Bennett, Gacs, Li, Vitanyi, Zurek, IEEE Trans-IT 44:4(1998), 1407:1423; Li, Badger, Chen, Kwong,, Kearney, Zhang : Bioinformatics, 17:2(2001), 149-154; Li, Chen, Li, Ma, Vitanyi, IEEE Trans-IT 50:12(2004), 3250-3264; Cilibrasi and Vitanyi, IEEE Trans-IT 51:4(2005), 1523-1545; Cilibrasi and Vitanyi, IEEE Trans Knowledge Data Engin 19:3(2007), 370-383.

- In classical Newton world, we use length to measure distance: 10 miles, 2 km
- In the modern information world, what measure do we use to measure the distances between
  - Two documents?
  - Two genomes?
  - Two computer virus?
  - Two junk emails?
  - Two (possibly copied) programs?
  - Two pictures?
  - Two internet homepages?
- They share one common feature: they all contain information, represented by a sequence of bits.

The Problem:

Given: Literal objects (binary files)



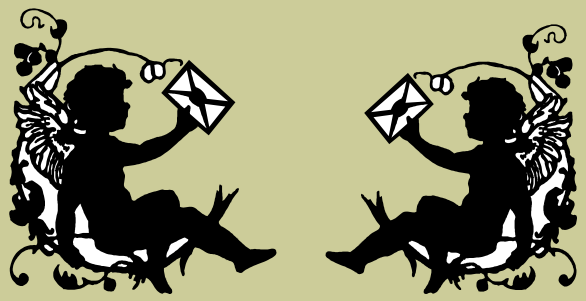
Determine: "Similarity" Distance Matrix (distances between every pair)

Applications: Clustering, Classification, Evolutionary trees of Internet documents, computer programs, chain letters, genomes, languages, texts, music pieces, ocr, ……

We are interested in a general theory of information distance.

The classical approach does not work

- For all the distances we know: Euclidean distance, Hamming distance, edit distance, none is proper. For example, they do not reflect our intuition on:



Austria

Byelorussia

- But from where shall we start?
- We will start from first principles and make no more assumptions. We wish to derive a general theory of information distance.

Admissible distance

Definition. D is an admissible distance if it satisfies:

- Symmetric, $D(x,y)=D(y,x)$
- $D(x,y) > 0$, for $x \neq y$, and $D(x,x)=0$,
- (up to an additive logarithmic term)
- Density requirements: $|\{y : D(x,y)<d\}| \leq 2^d$, or Normalize scaling by: $\sum_y 2^{-D(x,y)} \leq 1$
- D is upper semicomputable. That is,
- $\{d: D(x,y) \leq d$ and d is rational$\}$ is r.e.

Information Distance

- Information Distance (Li, Vitanyi, 96; Bennett,Gacs,Li,Vitanyi,Zurek, 98)

$$E(x,y) = \min \{ |p|: p(x)=y \ \& \ p(y)=x\}$$

Binary program for a Universal Computer
(Lisp, Java, C, Universal Turing Machine)

Theorem (i) $E(x,y) = \max \{C(x|y),C(y|x)\}$ (up to log term)

Kolmogorov complexity of x given y, defined as length of shortest binary ptogram that outputs x on input y.

(ii) $E(x,y) \leq D(x,y)$

(iii) $E(x,y)$ is an admissible distance and in fact a metric

) $E(x,y)$ is lower semicomputable

The fundamental theorem

Theorem (i). $E(x,y) = \max\{ C(x|y), C(y|x) \}$.

Remark. The theorem is counterintuitive! Note that all these theorems are up to an additive $O(\log C(x,y))$ term.

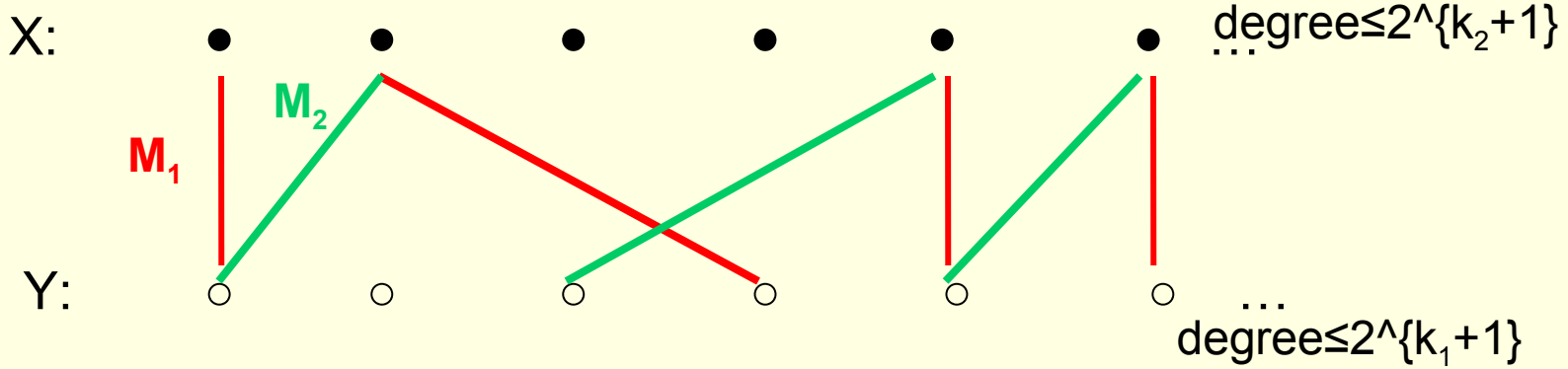
Proof. By the definition of $E(x,y)$, it is obvious that $E(x,y) \geq \max\{C(x|y), C(y|x)\}$. We now prove the difficult part: $E(x,y) \leq \max\{C(x|y), C(y|x)\}$.

$E(x,y) \leq \max\{C(x|y), C(y|x)\}$.

Proof. Define graph G={X U Y, E}, and let k_1=C(x|y), k_2=C(y|x), assuming $k_1 \leq k_2$

- where X={0,1}*x{0}
- and Y={0,1}*x{1}

- E={{u,v}: u in X, v in Y, C(u|v)≤k_1, C(v|u)≤k_2}



- We can partition E into at most 2^{k_2+2} matchings {M_i}.
  - For each (u,v) in E, node u has most 2^{k_2+1} edges hence it belongs to at most 2^{k_2+1} matchings, similarly node v belongs to at most 2^{k_1+1} matchings. Thus, edge (u,v) can be put in an unused matching M_i.
- Program P: has k_2,i, where M_i contains edge (x,y)
  - Generate the set of matchings {M_i} (by enumeration using k_2)
  - From M_i,x ➔ y, from M_i,y ➔ x. QED

Universality

Theorem (ii). For every admissible distance D, up to a small additive term, we have

for all x,y, $E(x,y) \leq D(x,y)$ (universality)

Comments: E(x,y) is optimal information distance – it discovers all effective similarities

Proof. Let ***D*** be the class of admissible distances we have defined. For some D(.,.) in ***D***, let D(x,y)=d, Define S(x)={z: D(x,z)≤d}. S(x) is r.e., yεS(x) and $|S(x)| \leq 2^d$. Thus for every y in this set, C(y|x)≤d+O(log d). Since D(x,y) is symmetric, we also derive C(x|y) ≤ d+O(log d). By the fundamental theorem, up to additive log d :

$$E(x,y) = \max\{C(x|y),C(y|x)\} \leq D(x,y)$$

Using prefix complexity we can replace additive log d by a constant. QED

Theorem (iii). E(x,y) is an admissible distance and metric

Proof. Obviously (up to some constant or logarithmic term), $E(x,y)=E(y,x)$; $E(x,x)=0$; $E(x,y)>0$ for $y \neq x$;

Triangle inequality:

$$E(x,y) = \max\{C(x|y), C(y|x)\}$$
$$\leq \max\{C(x|z)+C(z|y), C(y|z)+C(z|x)\}$$
$$\leq \max\{C(x|z),C(z|x)\}+\max\{C(z|y),C(y|z)\}$$
$$= E(x,z)+E(z,y)$$

Density: $|\{y : E(x,y)<d\}|<2^d$ (because there are only this many programs of length d.

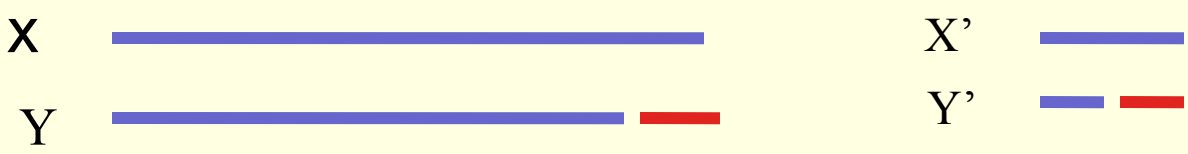
Upper semicomputable: $\{d: E(x,y)\leq d, d \text{ rational}\}$ is r.e.

QED

Normalizing

- Information distance measures the absolute information distance between two objects. However when we compare “big” objects which contain a lot of information and “small” objects which contain much less information, we need to compare their “relative” shared information.
- Examples: E. coli has 5 million base pairs. H. Influenza has 1.8 million base pairs. They are sister species. Their information distance would be larger than H. influenza with the trivial sequence which contains no base pair and no information.
- Thus we need to normalize the information distance by $d(x,y)=E(x,y)/\max\{C(x),C(y)\}$.
- Project: try other types of normalization.

Continued



E(x,y)=E(x',y') =

But x and y are much more similar than x' and y'

So, we Normalize:

Li Badger Chen Kwong Kearney Zhang 01
Li Vitanyi 01/02
Li Chen Li Ma Vitanyi 04
Cilibrasi, Vitanyi, de Wolf 04
Cilibrasi, Vitanyi 05

$$d(x,y) = \frac{E(x,y)}{\text{Max}\{C(x),C(y)\}}$$

Normalized Information Distance (NID)

The "Similarity metric"

Normalized Information Distance

Definition. We normalize E(x,y) to define the normalized information distance:

$$d(x,y)=E(x,y)/\max\{C(x),C(y)\}$$
$$=\max\{C(x|y,C(y|x)\}/\max\{C(x),C(y)\}$$

- The new measure still has the following properties:
  - Triangle inequality (to be proved)
  - symmetric;
  - $d(x,y)\geq 0$;
  - Hence it is a metric again!
  - But it is not r.e. any more.

Theorem. d(x,y) satisfies triangle inequality

Proof. Let $M_{xy}=\max\{C(x),C(y)\}$ We need to show:

$$E(x,y)/M_{xy} \le E(x,z)/M_{xz} + E(z,y)/M_{zy}, \text{ that is:}$$

$$\max\{C(x|y),C(y|x)\}/M_{xy} \le \max\{C(x|z),C(z|x)\}/M_{xz} + \max\{C(z|y),C(y|z)\}/M_{zy}$$

Case 1. Let $C(z) \le C(x), C(y)$. Consider

$$\max\{C(x|y),C(y|x)\} \le \max\{C(x|z)+C(z|y),\ C(y|z)+C(z|x)\}$$
$$\le \max\{C(x|z),C(z|x)\} + \max\{C(z|y),C(y|z)\}\ .$$

Then divide both sides by M_{xy}, and replace M_{xy} on the right by M_{xz} or M_{zy}.

Case 2. Let $C(z)\ge C(x)\ge C(y)$. By symmetry of information theorem, we know $C(x)-C(x|z) = C(z)-C(z|x)$, since $C(z) \ge C(x)$, we obtain $C(z|x) \ge C(x|z)$.

Similarly, $C(z|y)\ge C(y|z)$. Thus we only need to prove

$$C(x|y)/C(x) \le C(z|x)/C(z) + C(z|y)/C(z) \qquad (1)$$

We know

$$C(x|y)/C(x) \le [\ C(x|z) + C(z|y)\ ]/C(x) \qquad (2)$$

The lefthand ≤1. Let $\Delta=C(z)-C(x) = C(z|x)-C(x|z)$. Add Δ to righthand side of (2) to the nominator and denominator, so that the righthand sides of (1) and (2) are the same. If the righthand of (2) size was >1, then although this decreases the righthand side of (2), it is still greater than 1, hence (1) holds. If the righthand side of (2) was <1, then adding Δ only increases it further, hence (1) again holds.

QED

Practical concerns

- d(x,y) is not computable, hence we replace C(x) by Compress(x) (shorthand: Comp(x))

$$d(x,y) = \frac{Comp(xy)-min\{Comp(x),Comp(y)\}}{max\{Comp(x),Comp(y)\}}$$

- Note: $max\{C(x|y),C(y|x)\} = max\{ C(xy)-C(y), C(xy)-C(x)\}$
  $= C(xy) - min\{C(x),C(y)\}$

Approximating C(x),C(xy) – a side story

- The ability to approximate C*(xy)* gives the accuracy of *d(x,y).* Let's look at compressing genomes.
- DNAs are over alphabet {A,C,G,T}. Trivial algorithm gives 2 bits per base.
- But all commercial software like "compress", "compact", "pkzip", "arj" give > 2 bits/base
- There are DNA compression programs GenCompress and DNACompress.
- Converted GenCompress to 26 letter alphabet for English documents. But bzip2 and PPMZ also fine.

Compression experiments on DNA sequences

Sequence	Length	Unix Compress	Arith-2	Biocompress-2	GenCompress	Improvement
PANMTPAC	100314	2.12	1.87	1.88	1.86	10.26%
MPOMTCG	186609	2.2	1.97	1.94	1.9	54.45%
CHNTXX	155844	2.19	1.93	1.62	1.61	0.68%
HUMGHCSA	66495	2.19	1.94	1.31	1.1	29.25%
HUMHBB	73308	2.2	1.92	1.88	1.82	46.75%
HUMHDABCD	58864	2.21	1.94	1.88	1.82	46.99%
HUMDYSTROP	38770	2.23	1.92	1.93	1.92	4.88%
HUMHPRTB	56737	2.2	1.93	1.91	1.85	64.24%
VACCG	191737	2.14	1.9	1.76	1.76	0.00%

Bit per base. Without compression it is 2 bits per base,

100*[C(x)-C(x|y)]/C(xy) of the 7 Genomes --- Experiments on Symmetry of Information:

- We computed C(x)-C(x|y) on the following 7 species of bacteria ranging from 1.6 to 4.6 million base pairs
  - Archaea: *A. fulgidus, P. abyssi, P. horikoshii*
  - Bacteria: *E. coli, H. influenzae, H. pylori 26695, H. pylori strain J99*.
- Observe the approximate symmetry in this [C(x)-C(x|y)]/C(xy)*100 table.

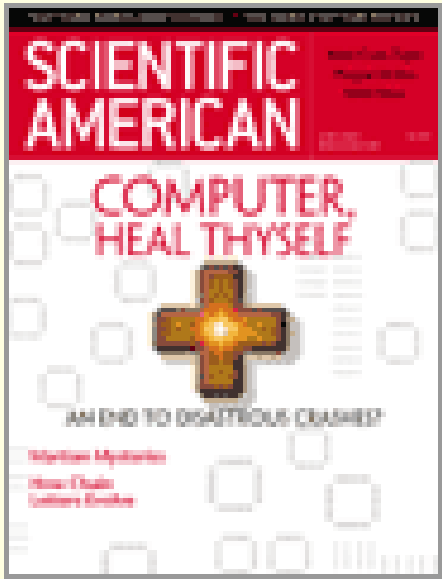
Sequence	A. fulgidus	P. abyssi	P. horikoshii	E. coli	H. influenzae	H. pylori-1	H. pylori-2
A. fulgidus		0.018326	0.01955	-0.00055	-0.0024	-0.00177	-0.00226
P. abyssi	0.023072		0.797546	0.000089	0.000988	0.000812	0.000705
P. horikoshii	0.023055	0.794383		-0.00039	0.000617	0.000109	-0.00011
E. coli	0.000373	-0.00108	-0.00054		0.04876	0.00816	0.008371
H. influenzae	0.000274	0.000145	-4.4E-05	0.049059		0.018303	0.017776
H. pylori-1	-0.00222	0.000307	-0.00014	0.009068	0.016523		43.06986
H. pylori-2	-0.00131	-0.00078	-6.2E-05	0.009796	0.01968	43.17104	

Applications of information distance

- Evolutionary history of chain letters
- Whole genome phylogeny
- Data mining and time series classification
- Plagiarism detection
- Clustering music, languages etc.
- Google distance --- meaning inference

Application 1. Chain letter evolution

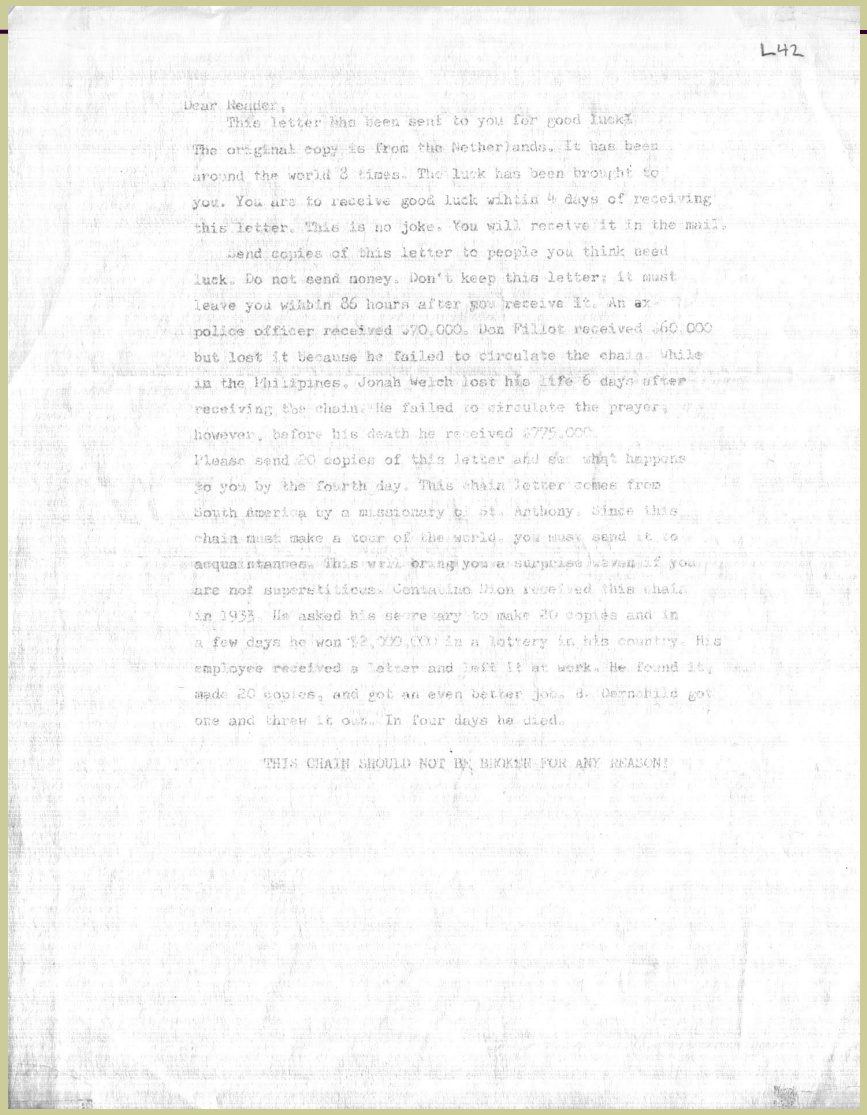
- Charles Bennett collected 33 copies of chain letters that were apparently from the same origin during 1980—1997.
- Li, Ma, Bennett were interested in reconstructing the evolutionary history of these chain letters.
- Because these chain letters are readable, they provide a perfect tool for classroom teaching of phylogeny methods and test for such methods.
- *Scientific American*: Jun. 2003



A sample letter:

Trust in the Lord with all your heart and he will acknowledge and He will light the way. This Prayer has been sent to you for good luck. The original copy is from the Netherlands. It has been around the world nine times. The luck has been brought to you. You are to receive good luck within four days of receiving this letter. This is nojoke. You will receive it in the mail. Send copies of this letter to people you think need good luck. Do not send money. Do not keep this letter. It must leave your hands within ninety six hours after you receive it. An RAF officer received $70,000. Don Elliott received $50,000 and lost it because he broke the chain. While in the Phillipines, General Welch lost his life six days after he received this letter. He failed to circulate the Prayer. However, before his death, he received $775,000. Please send twenty copies and see what happens to you on the fourth day. This chain comes from Venzuela and was written by Sol Anthony De Cadif, a missionary from South America. Since this chain must make a tour of the world. you must make twenty copies identical to this one and send it to your friends, parents, and acquantances. After a few days you will get a surprise. This is true, even if you are not superstitious. Take note of the following. Constantine Diaz received the chain in 1953. He asked his secretary to make twenty copies and send them. A few days later he won a lottery for two million dollars in his country. Carlo Craduit, and office employee, received the chain. He forgot it and in a few days lost his job. He found the chain and sent it to twenty people. Five days later he got an even better job. Dolin Moirchild received the chain and not believing in it, threw it away. Nine days later he died. For no reason what so ever should this chain be broken

*A very pale letter reveals evolutionary path: ((copy)*mutate)**



L42

Dear Reader,

This letter has been sent to you for good luck. The original copy is from the Netherlands. It has been around the world 9 times. The luck has been brought to you. You are to receive good luck within 4 days of receiving this letter. This is no joke. You will receive it in the mail.

Send copies of this letter to people you think need luck. Do not send money. Don't keep this letter; it must leave you within 96 hours after you receive it. An ex-police officer received $70,000. Don Elliot received $60,000 but lost it because he failed to circulate the chain. While in the Philipines, Jonah Welch lost his life 6 days after receiving the chain. He failed to circulate the prayer, however, before his death he received $775,000.

Please send 20 copies of this letter and see what happens to you by the fourth day. This chain letter comes from South America by a missionary to St. Anthony. Since this chain must make a tour of the world, you must send it to acquaintances. This will bring you a surprise even if you are not superstitious. Contadino Dion received this chain in 1933. He asked his secretary to make 20 copies and in a few days he won $2,000,000 in a lottery in his country. His employee received a letter and left it at work. He found it, made 20 copies, and got an even better job. S. Demchik got one and threw it out. In four days he died.

THIS CHAIN SHOULD NOT BE BROKEN FOR ANY REASON!

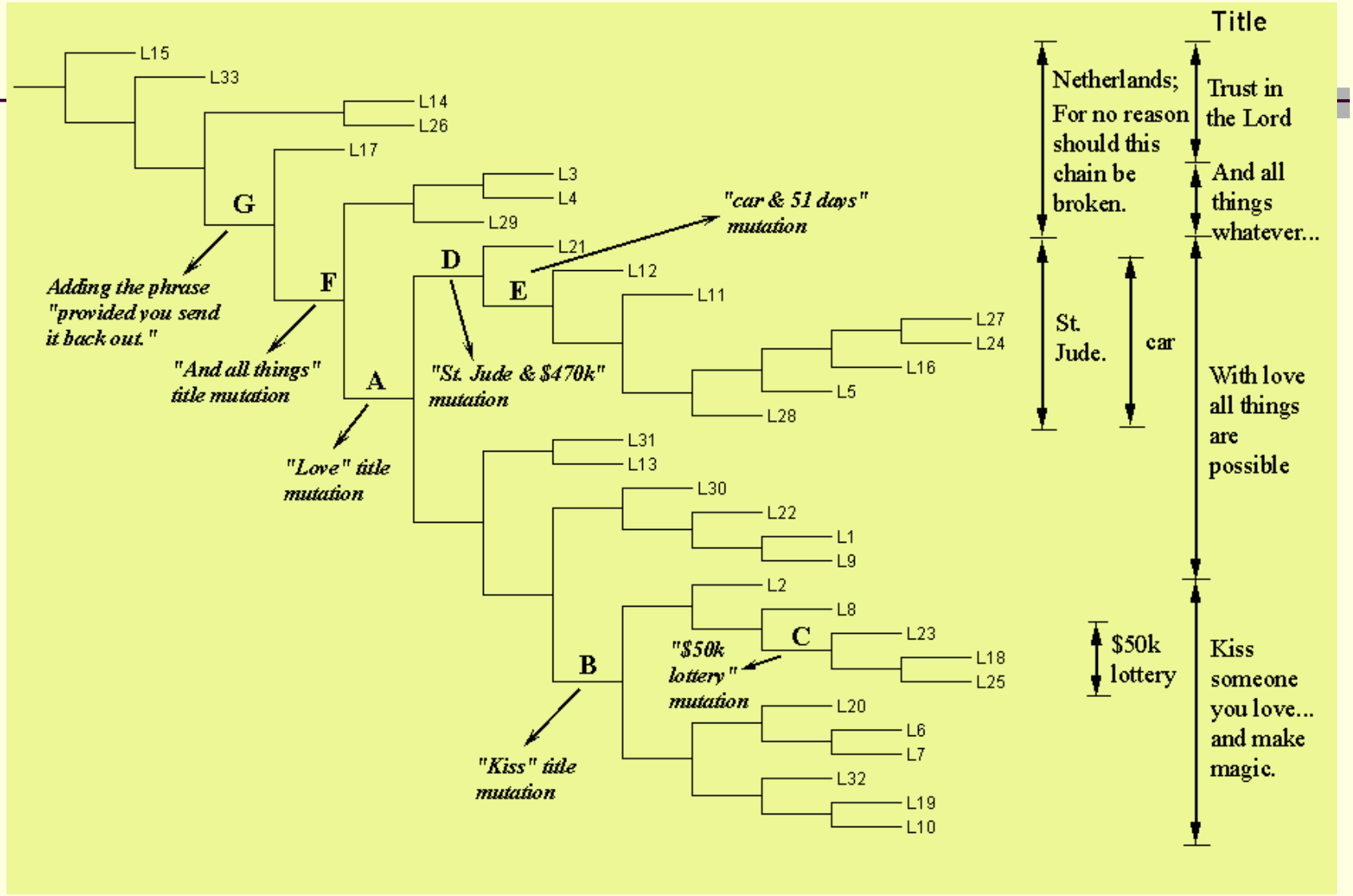
A typical chain letter input file:

with love all things are possible
this paper has been sent to you for good luck. the original is in new england. it has been around the world nine times. the luck has been sent to you. you will receive good luck within four days of receiving this letter. provided, in turn, you send it on. this is no joke. you will receive good luck in the mail. send no money. send copies to people you think need good luck. do not send money as faith has no price. do not keep this letter. It must leave your hands within 96 hours. an r.a.f. (royal air force) officer received $470,000. joe elliot received $40,000 and lost them because he broke the chain. while in the philippines, george welch lost his wife 51 days after he received the letter. however before her death he received $7,755,000. please, send twenty copies and see what happens in four days. the chain comes from venezuela and was written by saul anthony de grou, a missionary from south america. since this letter must tour the world, you must make twenty copies and send them to friends and associates. after a few days you will get a surprise. this is true even if you are not superstitious. do note the following: constantine dias received the chain in 1953. he asked his secretary to make twenty copies and send them. a few days later, he won a lottery of two million dollars. carlo daddit, an office employee, received the letter and forgot it had to leave his hands within 96 hours. he lost his job. later, after finding the letter again, he mailed twenty copies; a few days later he got a better job. dalan fairchild received the letter, and not believing, threw the letter away, nine days later he died. in 1987, the letter was received by a young woman in california, it was very faded and barely readable. she promised herself she would retype the letter and send it on, but she put it aside to do it later. she was plagued with various problems including expensive car repairs, the letter did not leave her hands in 96 hours. she finally typed the letter as promised and got a new car. remember, send no money. do not ignore this. it works.
st. jude

Reconstructing History of Chain Letters

- For each pair of chain letters (*x, y*) we computed *d(x,y)* by GenCompress, hence a distance matrix.
- Using standard phylogeny program to construct their evolutionary history based on the d(x,y) distance matrix.
- The resulting tree is a perfect phylogeny: distinct features are all grouped together.

Phylogeny of 33 Chain Letters



Answers a question in VanArsdale study: “Love” title appeared earlier than “Kiss” title

Application 2. Evolution of Species

- Traditional methods infers evolutionary history for a single gene, using:
  - Max. likelihood: multiple alignment, assumes statistical evolutionary models, computes the most likely tree.
  - Max. parsimony: multiple alignment, then finds the best tree, minimizing cost.
  - Distance-based methods: multiple alignment, NJ; Quartet methods, Fitch-Margoliash method.
- Problem: different gene trees, horizontally transferred genes, do not handle genome level events.

Whole Genome Phylogeny

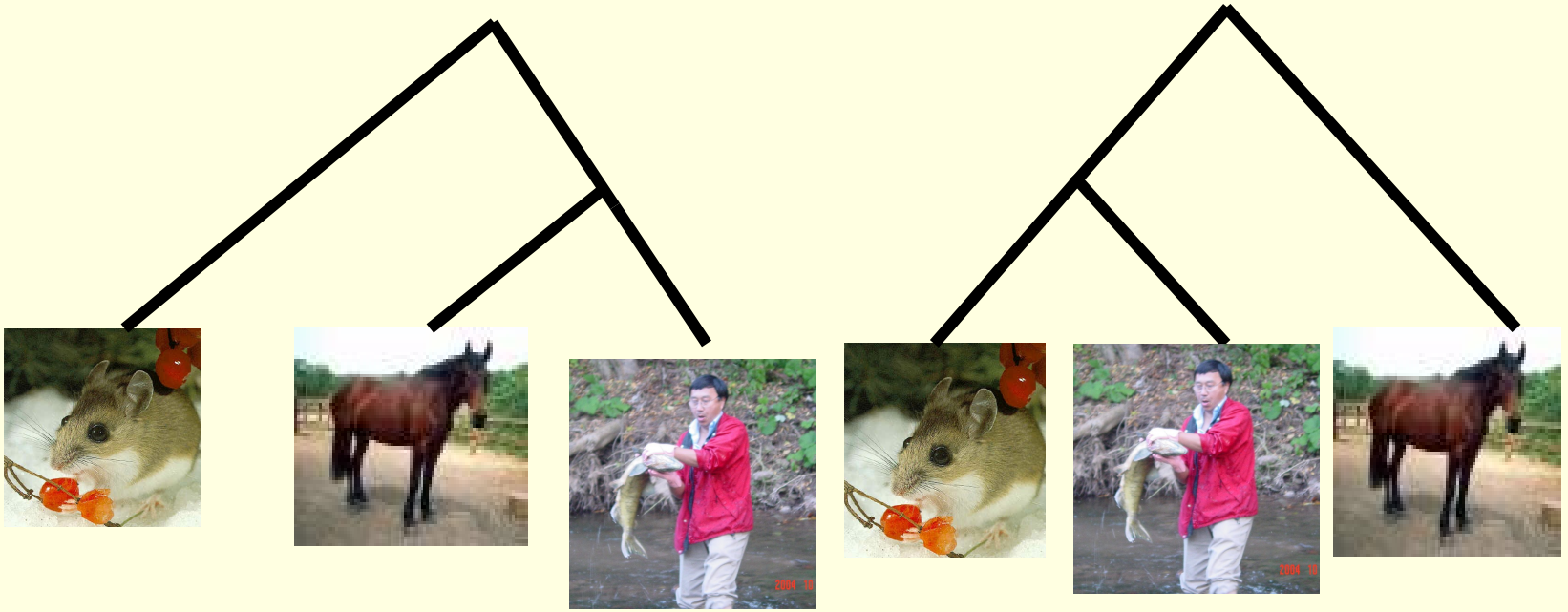
Li, Badger, Chen, Kwong, Kearney, Zhang, Bioinformatics, 2001 (sum measure); Li, Chen, Li, Ma, Vitanyi, IEEE Trans IT 2004 (max measure)

- Our method enables a whole genome phylogeny method, for the first time, in its true sense.
- Prior work: Snel, Bork, Huynen: compare gene contents. Boore, Brown: gene order. Sankoff, Pevzner, Kececioglu: reversal/translocation
- Our method
  - Uses all the information in the genome.
  - No need of evolutionary model – universal.
  - No need of multiple alignment
  - Gene contents, gene order, reversal/translocation, are all special cases.

Eutherian Orders:

- It has been a disputed issue which of the two groups of placental mammals are closer: Primates, Ferungulates, Rodents.
- In mtDNA, 6 proteins say primates closer to ferungulates; 6 proteins say primates closer to rodents.
- Hasegawa's group concatenated 12 mtDNA proteins from: rat, house mouse, grey seal, harbor seal, cat, white rhino, horse, finback whale, blue whale, cow, gibbon, gorilla, human, chimpanzee, pygmy chimpanzee, orangutan, sumatran orangutan, with opossum, wallaroo, platypus as out group, 1998, using max likelihood method in MOLPHY.

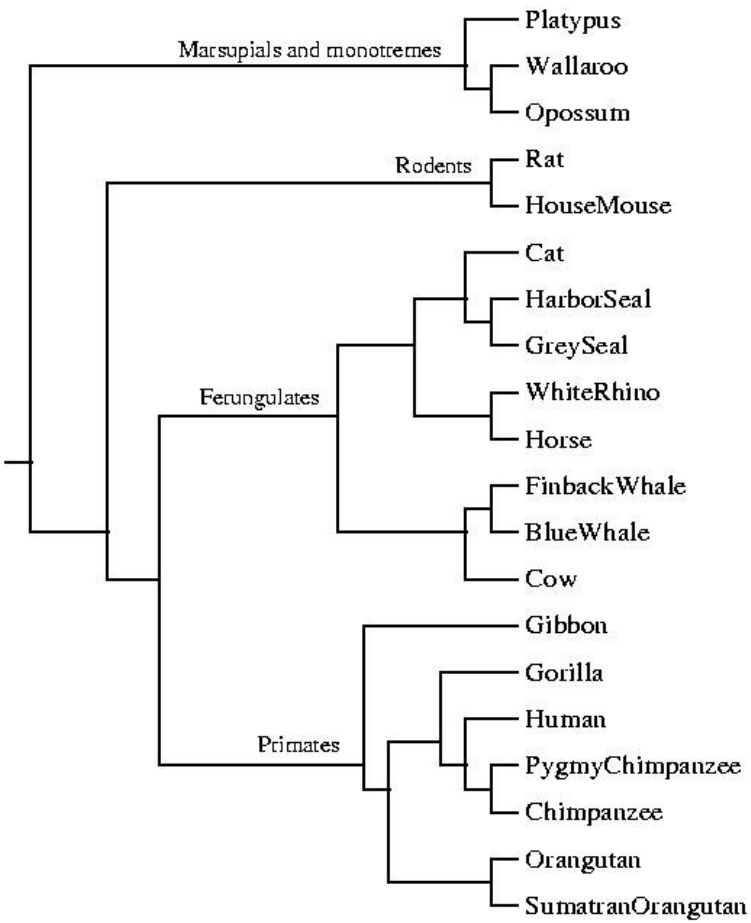
Who is our closer relative?



Eutherian Orders ...

- We use complete mtDNA genome of exactly the same species.
- We computed *d(x,y)* for each pair of species, and used Neighbor Joining in Molphy package (and our own hypercleaning).
- We constructed exactly the same tree. Confirming Primates and Ferungulates are closer than Rodents.

Evolutionary Tree of Mammals:

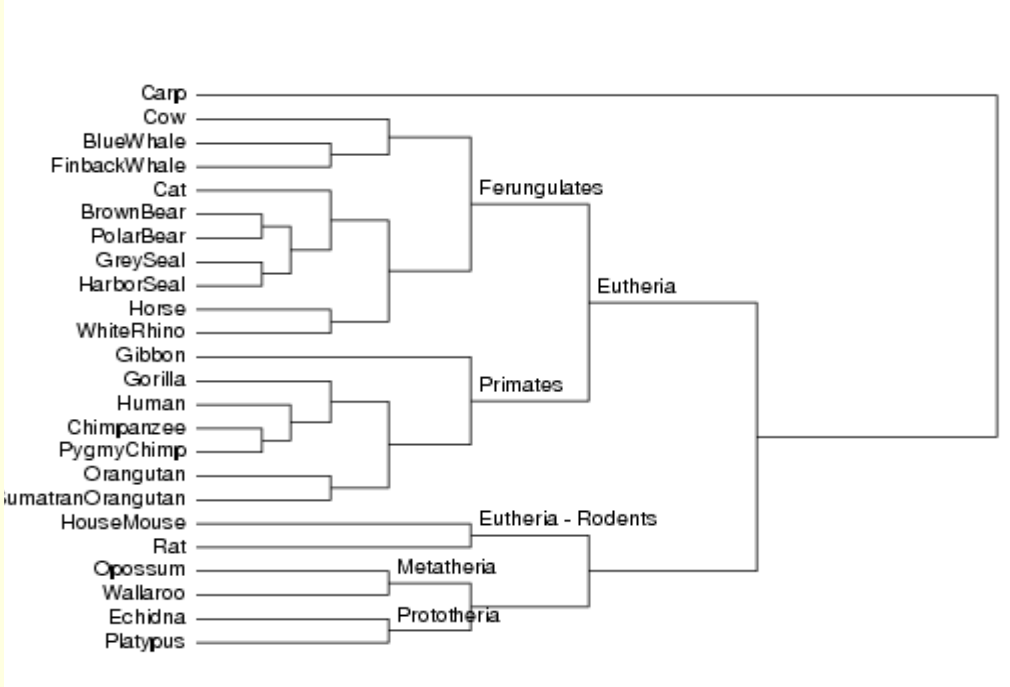


NCD Matrix 24 Species (mtDNA).

Diagonal elements about 0. Distances between primates ca 0.6.

	BlueWhale	BrownBear	Carp	Cat	Chimpanzee	Cow	Echidna	FinWhale	Gibbon	Gorilla	GreySeal	HarborSeal	Horse	HouseMouse	Human	Opossum	Orangutan	Platypus	PolarBear	PygmyChimp	Rat	SumOrang	Wallaroo	WhiteRhino
BlueWhale	0.005	0.906	0.943	0.897	0.925	0.885	0.936	0.616	0.928	0.931	0.901	0.898	0.896	0.936	0.930	0.936	0.928	0.939	0.907	0.930	0.927	0.939	0.925	0.902
BrownBear	0.906	0.002	0.943	0.887	0.935	0.906	0.944	0.915	0.939	0.940	0.875	0.872	0.910	0.934	0.930	0.936	0.938	0.937	0.269	0.940	0.935	0.936	0.925	0.915
Carp	0.943	0.943	0.006	0.946	0.954	0.947	0.955	0.952	0.951	0.957	0.949	0.950	0.952	0.956	0.946	0.956	0.955	0.954	0.945	0.960	0.950	0.955	0.942	0.960
Cat	0.897	0.887	0.946	0.003	0.926	0.897	0.942	0.905	0.928	0.931	0.870	0.872	0.885	0.919	0.922	0.933	0.932	0.931	0.885	0.939	0.920	0.934	0.919	0.897
Chimpanzee	0.925	0.935	0.954	0.926	0.006	0.926	0.948	0.926	0.849	0.731	0.925	0.922	0.921	0.943	0.667	0.943	0.841	0.946	0.931	0.441	0.933	0.835	0.934	0.930
Cow	0.885	0.906	0.947	0.897	0.926	0.006	0.936	0.885	0.931	0.927	0.890	0.888	0.895	0.925	0.920	0.931	0.930	0.929	0.905	0.931	0.921	0.930	0.925	0.899
Echidna	0.936	0.944	0.955	0.942	0.948	0.936	0.005	0.936	0.947	0.947	0.940	0.937	0.942	0.941	0.939	0.936	0.947	0.855	0.935	0.949	0.941	0.947	0.939	0.948
EchnatWhale	0.616	0.915	0.952	0.905	0.926	0.885	0.936	0.005	0.930	0.931	0.911	0.908	0.901	0.935	0.922	0.936	0.933	0.934	0.910	0.932	0.928	0.932	0.927	0.902
Gibbon	0.928	0.939	0.951	0.928	0.849	0.931	0.947	0.930	0.005	0.859	0.932	0.930	0.927	0.948	0.844	0.951	0.872	0.952	0.936	0.854	0.939	0.868	0.935	0.939
Gorilla	0.931	0.940	0.957	0.931	0.731	0.927	0.947	0.931	0.859	0.006	0.927	0.929	0.924	0.944	0.737	0.944	0.835	0.943	0.928	0.732	0.938	0.836	0.934	0.929
GreySeal	0.901	0.875	0.949	0.870	0.925	0.890	0.940	0.911	0.932	0.927	0.003	0.899	0.888	0.924	0.922	0.935	0.931	0.936	0.863	0.939	0.922	0.930	0.910	0.898
HarborSeal	0.898	0.872	0.950	0.872	0.922	0.888	0.937	0.908	0.930	0.929	0.899	0.004	0.888	0.922	0.922	0.935	0.932	0.937	0.860	0.930	0.922	0.928	0.919	0.900
Horse	0.896	0.910	0.952	0.885	0.921	0.895	0.942	0.901	0.927	0.924	0.888	0.888	0.005	0.928	0.913	0.937	0.925	0.936	0.903	0.925	0.912	0.924	0.924	0.848
HouseMouse	0.936	0.934	0.956	0.919	0.943	0.925	0.941	0.935	0.948	0.944	0.924	0.922	0.928	0.006	0.932	0.925	0.944	0.930	0.924	0.942	0.860	0.945	0.931	0.928
Human	0.930	0.930	0.946	0.922	0.667	0.920	0.939	0.922	0.844	0.737	0.922	0.922	0.913	0.932	0.005	0.949	0.854	0.949	0.931	0.681	0.938	0.826	0.934	0.929
Opossum	0.936	0.936	0.956	0.935	0.943	0.931	0.936	0.936	0.951	0.944	0.935	0.935	0.937	0.925	0.949	0.006	0.960	0.938	0.939	0.954	0.941	0.960	0.898	0.952
Orangutan	0.928	0.938	0.955	0.932	0.841	0.930	0.947	0.933	0.872	0.835	0.931	0.932	0.923	0.944	0.854	0.960	0.006	0.954	0.933	0.845	0.943	0.885	0.945	0.934
Platypus	0.939	0.937	0.954	0.931	0.946	0.929	0.855	0.934	0.952	0.943	0.936	0.937	0.936	0.930	0.949	0.938	0.954	0.003	0.932	0.948	0.937	0.949	0.933	0.948
PolarBear	0.907	0.269	0.945	0.885	0.931	0.905	0.935	0.910	0.936	0.928	0.863	0.860	0.903	0.924	0.931	0.939	0.933	0.932	0.002	0.942	0.940	0.936	0.927	0.917
PygmyChimp	0.930	0.940	0.960	0.939	0.441	0.931	0.949	0.932	0.854	0.732	0.939	0.930	0.925	0.942	0.681	0.954	0.845	0.948	0.942	0.007	0.935	0.838	0.931	0.939
Rat	0.927	0.935	0.950	0.920	0.933	0.921	0.941	0.928	0.939	0.938	0.922	0.922	0.912	0.860	0.938	0.941	0.943	0.937	0.940	0.935	0.006	0.939	0.922	0.922
SumOrangutan	0.939	0.936	0.955	0.934	0.835	0.930	0.947	0.932	0.868	0.836	0.930	0.928	0.924	0.945	0.826	0.960	0.885	0.949	0.936	0.838	0.939	0.007	0.942	0.937
Wallaroo	0.925	0.925	0.942	0.919	0.934	0.925	0.939	0.927	0.935	0.934	0.930	0.919	0.924	0.921	0.934	0.891	0.945	0.930	0.927	0.931	0.922	0.942	0.005	0.935
WhiteRhino	0.902	0.915	0.960	0.897	0.930	0.899	0.948	0.902	0.939	0.939	0.898	0.900	0.848	0.928	0.929	0.952	0.934	0.948	0.917	0.939	0.922	0.937	0.935	0.002

Embedding NCD Matrix in dendrogram (hierarchical clustering) for this Large Phylogeny (no errors it seems)



Therian hypothesis

Versus

Marsupionti hypothesis

Mamals:

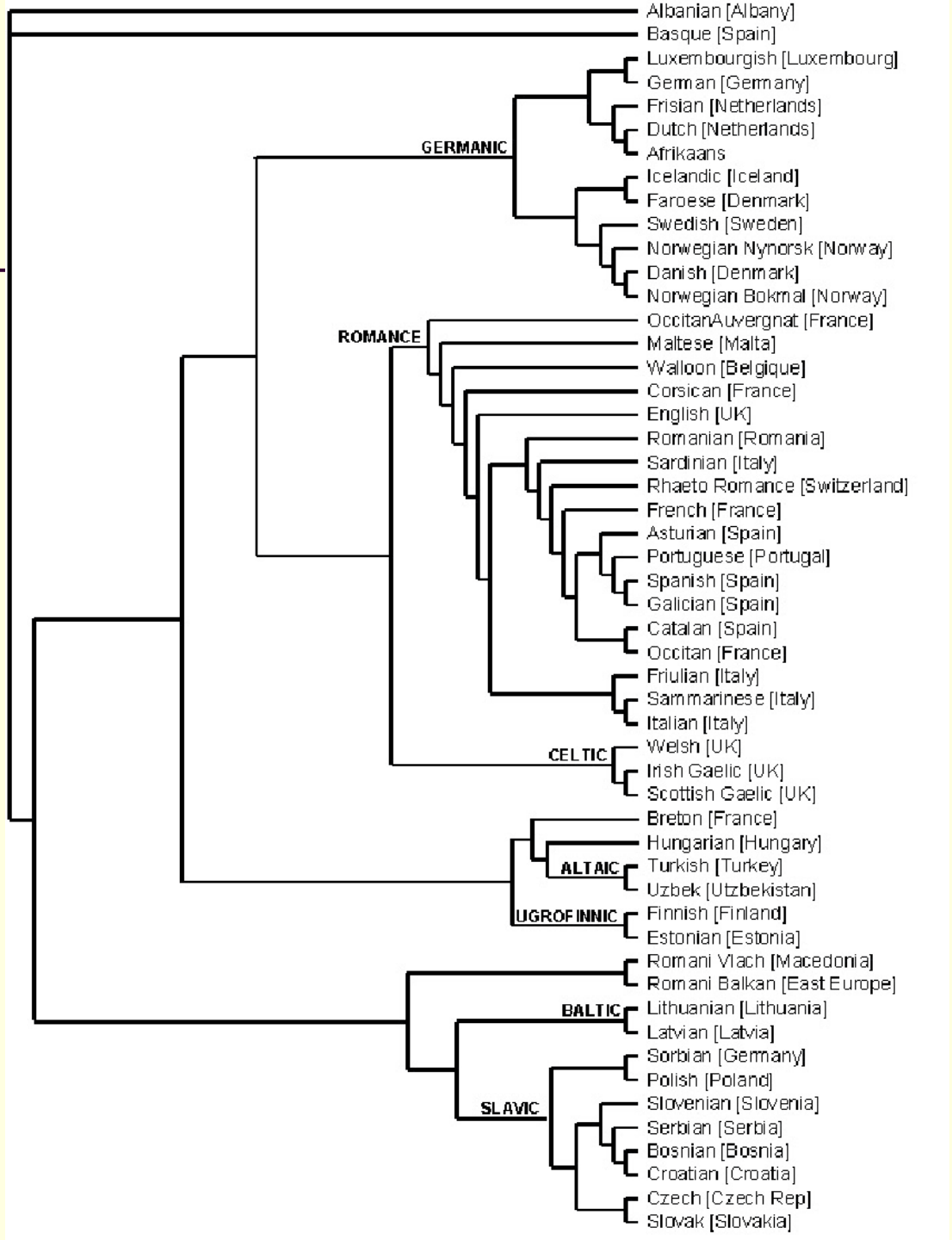
Eutheria
Metatheria
Prototheria

Which pair is closest?

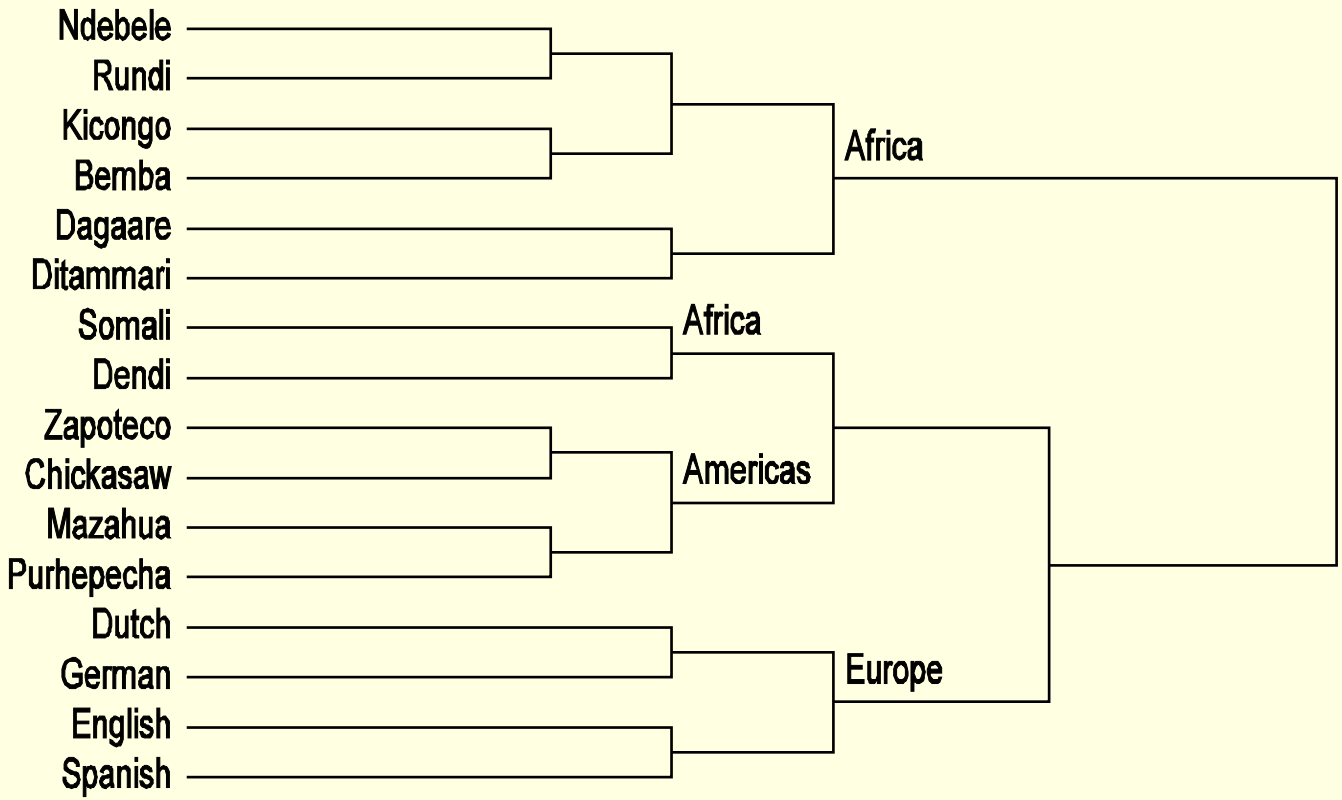
Plagiarism Detection

- The similarity measure also works for checking student program assignments. We have implemented the system SID.
- Our system takes input on the web, strip user comments, unify variables, we openly advertise our methods (unlike other programs) that we check shared information between each pair. It is un-cheatable because it is universal.
- Available at http://genome.cs.uwaterloo.ca/SID

A language tree created using UN's The Universal Declaration Of Human Rights, by three Italian physicists, in *Phy. Rev. Lett., & New Scientist*



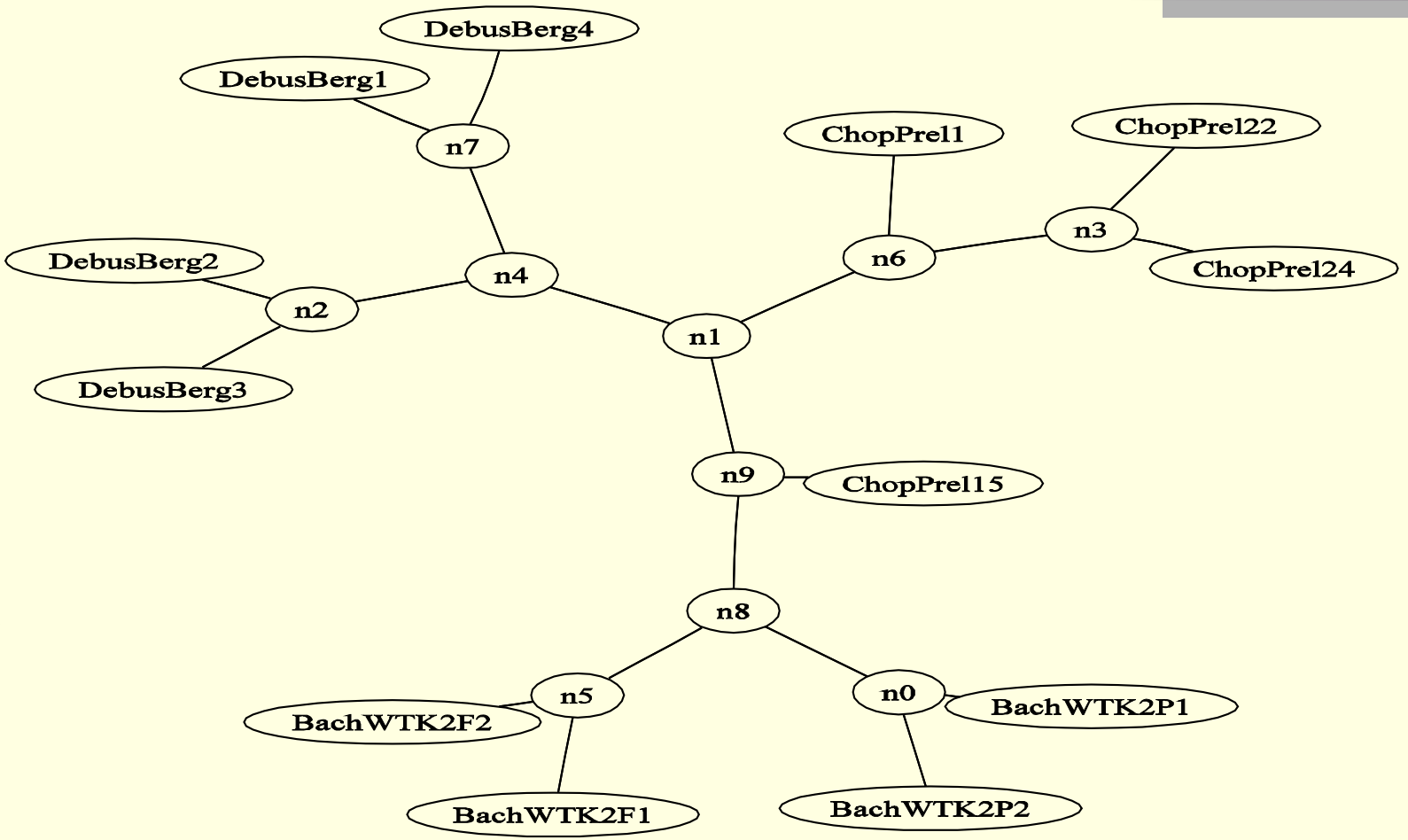
Clustering : Phylogeny of 15 languages: Native American, Native African, Native European Languages



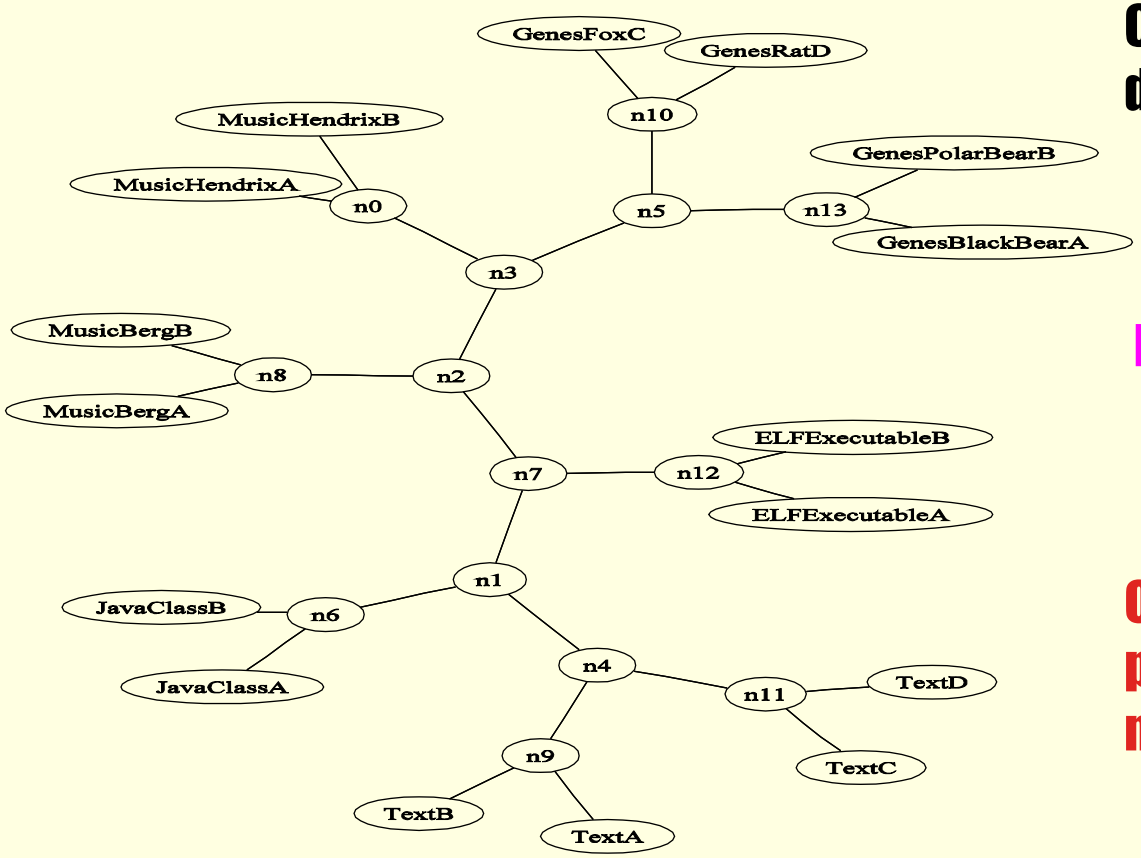
Classifying Music

- By Rudi Cilibrasi, Paul Vitanyi, Ronald de Wolf, reported in *New Scientist, April 2003.*
- They took 12 Jazz, 12 classical, 12 rock music scores. Classified well.
- Potential application in identifying authorship.
- The technique's elegance lies in the fact that it is tone deaf. Rather than looking for features such as common rhythms or harmonies, says Vitanyi, "it simply compresses the files obliviously."

12 Classical Pieces (Bach, Debussy, Chopin)
S(T)=0.95 ---- no errors



Heterogenous Data; Clustering perfect with S(T)=0.95.



Clustering of radically different data.

No features known.

Only our parameter-free method can do this!!

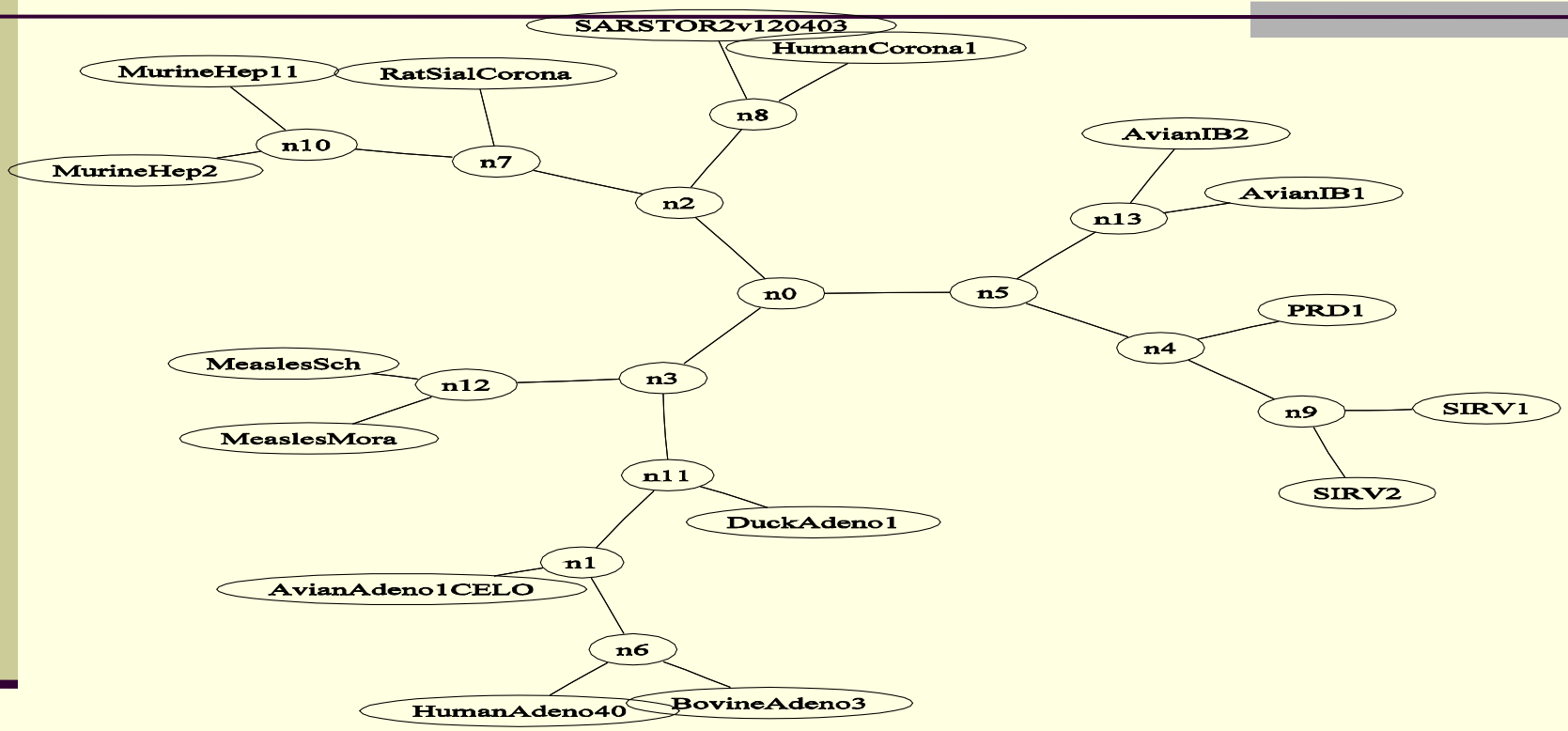
Parameter-Free Data Mining: Keogh, Lonardi, Ratanamahatana, KDD'04

- Time series clustering
  - Compared against 51 different parameter-laden measures from SIGKDD, SIGMOD, ICDM, ICDE, SSDB, VLDB, PKDD, PAKDD, the simple parameter-free shared information method outperformed all --- including HMM, dynamic time warping, etc.
- Anomaly detection

Other applications

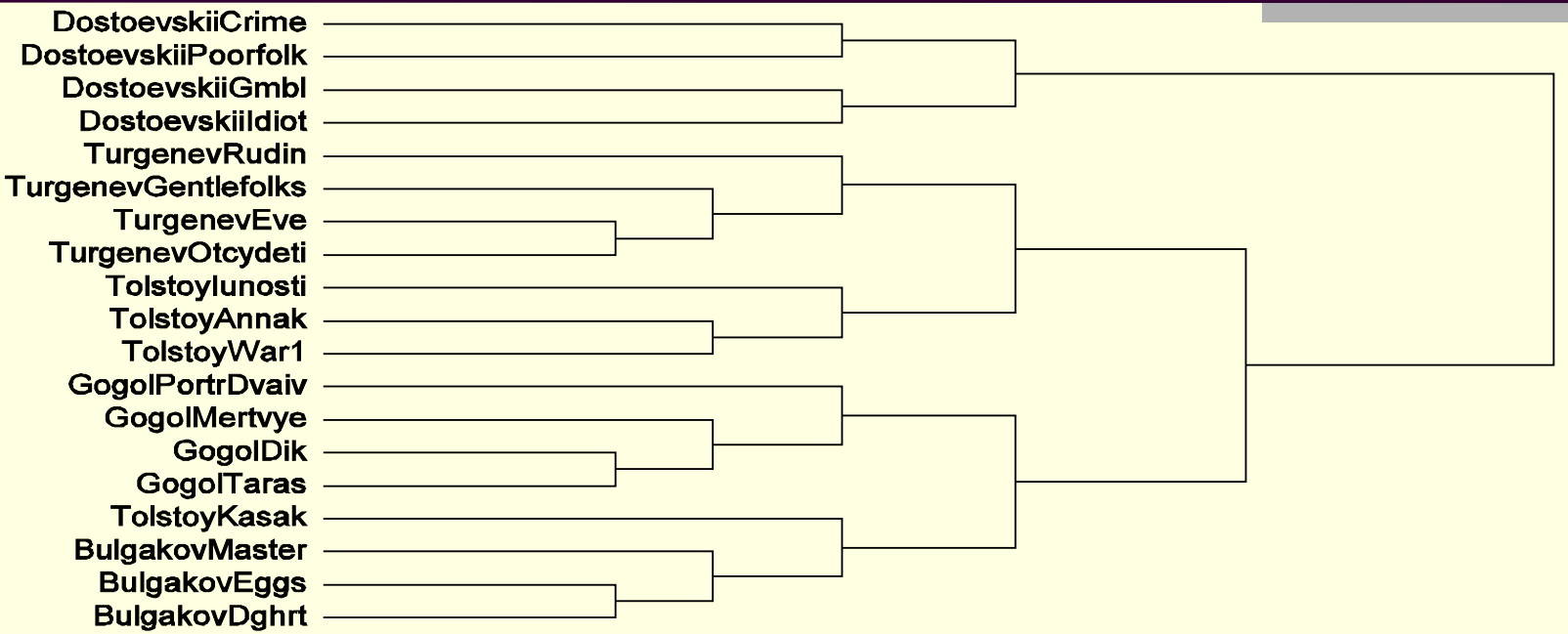
- C. Ane and M.J. Sanderson: Phylogenetic reconstruction
- K. Emanuel, S. Ravela, E. Vivant, C. Risi: Hurricane risk assessment
- Protein sequence classification
- Fetal heart rate detection
- Ortholog detection
- Authorship, topic, domain identification
- Worms and network traffic analysis
- Software engineering

Identifying SARS Virus: S(T)=0.988



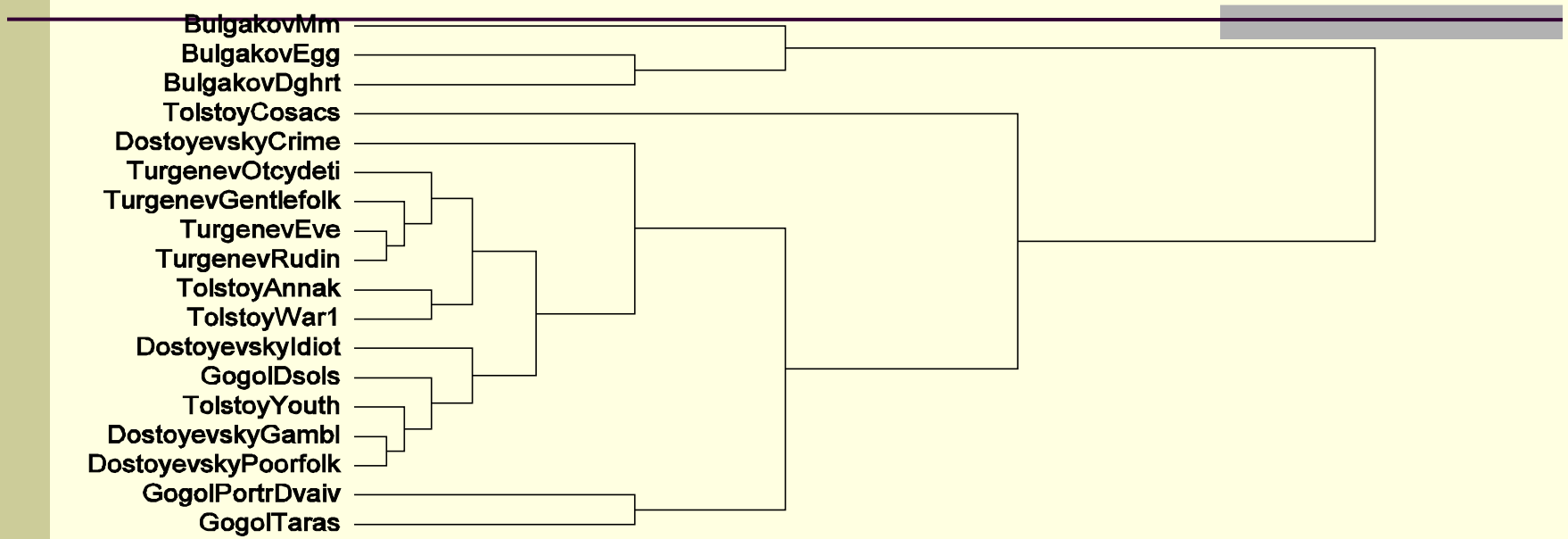
AvianAdeno1CELO.inp: Fowl adenovirus 1; AvianIB1.inp: Avian infectious bronchitis virus (strain Beaudette US); AvianIB2.inp: Avian infectious bronchitis virus (strain Beaudette CK); BovineAdeno3.inp: Bovine adenovirus 3; DuckAdeno1.inp: Duck adenovirus 1; HumanAdeno40.inp: Human adenovirus type 40; **HumanCorona1.inp: Human coronavirus 229E**; MeaslesMora.inp: Measles virus strain Moraten; MeaslesSch.inp: Measles virus strain Schwarz; MurineHep11.inp: Murine hepatitis virus strain ML-11; MurineHep2.inp: Murine hepatitis virus strain 2; PRD1.inp: Enterobacteria phage PRD1; RatSialCorona.inp: Rat sialodacryoadenitis coronavirus; **SARS.inp: SARS TOR2v120403**; SIRV1.inp: Sulfolobus virus SIRV-1; SIRV2.inp: Sulfolobus virus SIRV-2.

Russian Authors (in original Cyrillic) S(T)=0.949



I.S. Turgenev, 1818--1883 [Father and Sons, Rudin, On the Eve, A House of Gentlefolk]; **F. Dostoyevsky** 1821--1881 [Crime and Punishment, The Gambler, The Idiot; Poor Folk]; **L.N. Tolstoy** 1828--1910 [Anna Karenina, The Cossacks, Youth, War and Piece]; **N.V. Gogol** 1809--1852 [Dead Souls, Taras Bulba, The Mysterious Portrait, How the Two Ivans Quarrelled]; **M. Bulgakov** 1891--1940 [The Master and Margarita, The Fatefull Eggs, The Heart of a Dog]

Same Russian Texts in English Translation; S(T)=0953



Files start to cluster according to translators!

I.S. Turgenev, 1818--1883 [Father and Sons (**R. Hare**), Rudin (**Garnett, C. Black**), On the Eve (**Garnett, C. Black**), A House of Gentlefolk (**Garnett, C. Black**)]; F. Dostoyevsky 1821--1881 [Crime and Punishment (**Garnett, C. Black**), The Gambler (**C.J. Hogarth**), The Idiot (**E. Martin**); Poor Folk (**C.J. Hogarth**)]; L.N. Tolstoy 1828--1910 [Anna Karenina (**Garnett, C. Black**), The Cossacks (**L. and M. Aylmer**), Youth (**C.J. Hogarth**), War and Piece (**L. and M. Aylmer**)]; N.V. Gogol 1809—1852 [Dead Souls (**C.J. Hogarth**), Taras Bulba ($\approx$ G. Tolstoy, 1860, **B.C. Baskerville**), The Mysterious Portrait + How the Two Ivans Quarrelled ($\approx$ **I.F. Hapgood**]; M. Bulgakov 1891--1940 [The Master and Margarita (**R. Pevear, L. Volokhonsky**),

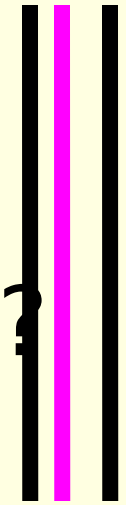
The Fatefull Eggs (**K. Gook-Horujy**), The Heart of a Dog (**M. Glenny**)]

You can use it too!

- CompLearn Toolkit: http://www.complearn.org
- “x” and “y” are literal objects (files);

  What about abstract objects like “home”, “red”, “Socrates”, “chair”, ….?

Or names for literal objects?



The End of Part I

PART II:

Automatic Meaning Discovery Using Google

Cilibrasi, Vitanyi 04/07

Reported in *New Scientist 2005, Slashdot 2005, etc.*

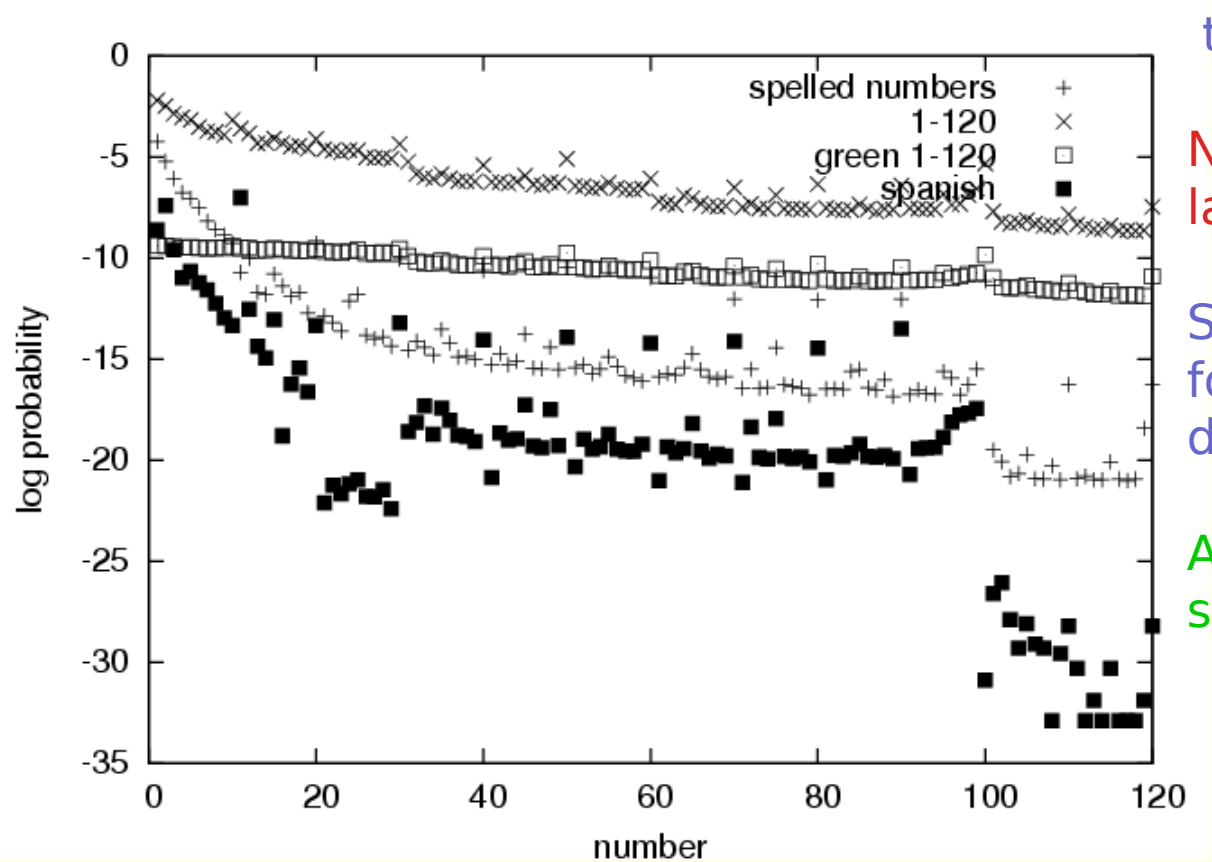
Non-Literal Objects

- Googling for Meaning

- Google distribution:

$$g(x) = \frac{\text{Google page count “x”}}{\text{\# pages indexed}}$$

Numbers versus log-probability



Probability according to Google.

Names in variety of languages and digits.

Same behavior in all formats. Google detects meaning:

All multiples of five stand out.

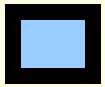
Google Compressor

- Google code length:

$$G(x) = \log \; 1 / g(x)$$

This is the Shannon-Fano code length that has minimum expected code word length w.r.t. $g(x)$.

Hence we can view Google as a Google Compressor.



Normalized Google Distance (NGD)

$$NGD(x,y) = \frac{G(x,y) - \min\{G(x),G(y)\}}{\max\{G(x),G(y)\}}$$

Same formula as NCD, using C = Google compressor

Use the Google counts and the CompLearn Toolkit to apply NGD.

Example

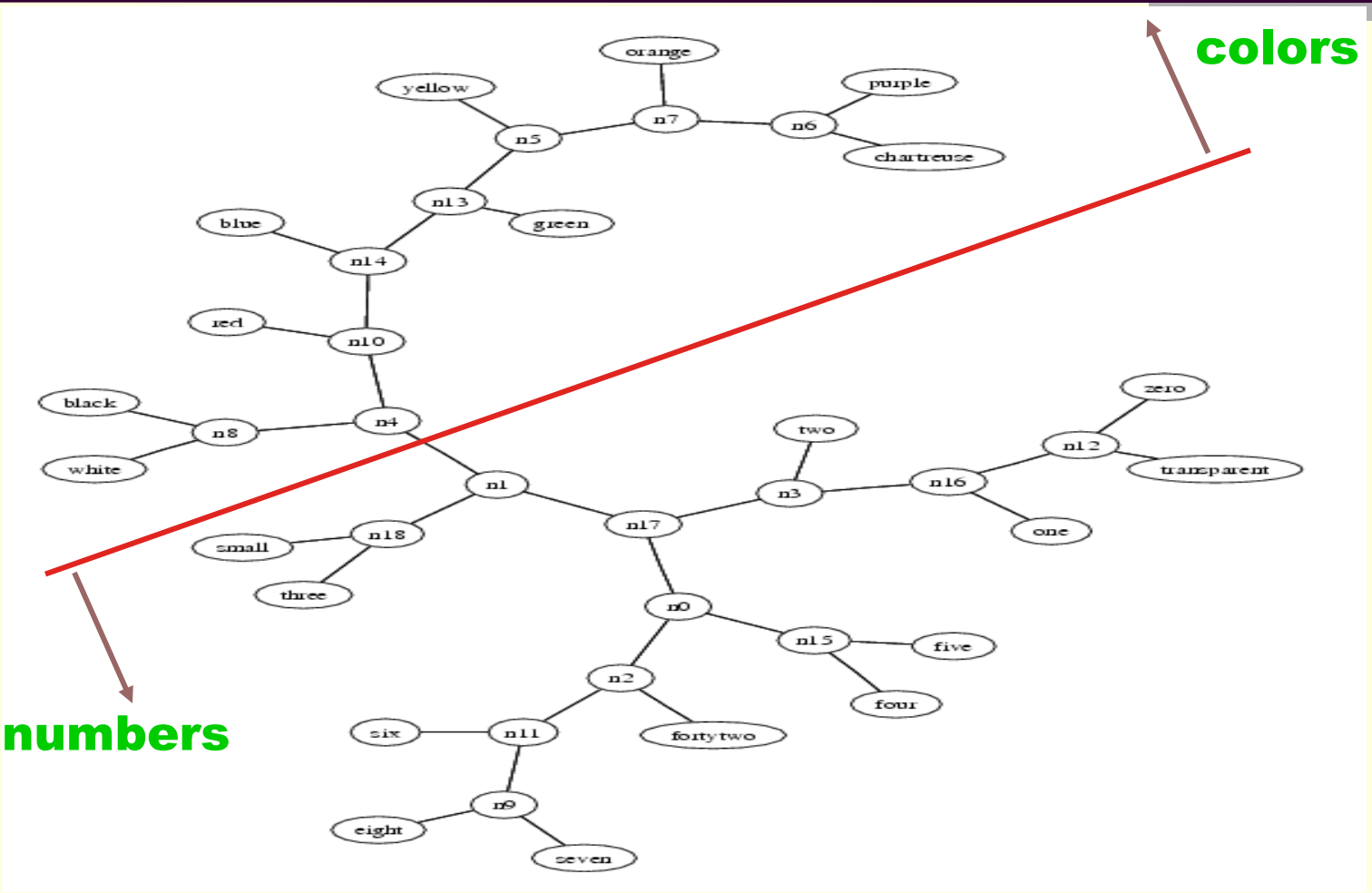
- “horse”: #hits = 46,700,000
- “rider”: #hits = 12,200,000
- “horse” “rider”: #hits = 2,630,000
- #pages indexed: 8,058,044,651

NGD(horse,rider) = 0.443

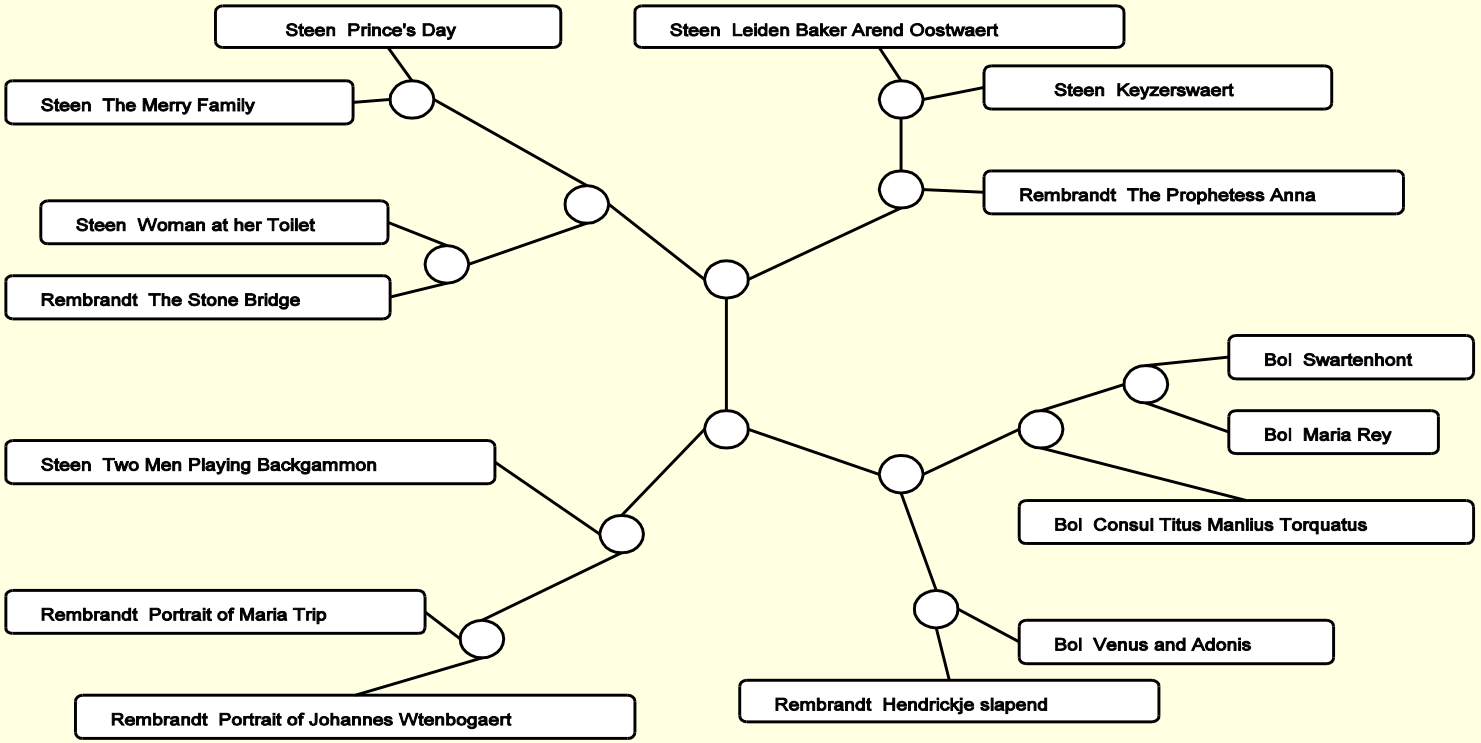
Theoretically+empirically: scale-invariant

Colors and Numbers—The Names!

Hierarchical Clustering



Hierarchical Clustering of 17th Century Dutch Painters, Paintings given by name, without painter's name.



Hendrickje slapend, Portrait of Maria Trip, Portrait of Johannes Wtenbogaert, The Stone Bridge, The Prophetess Anna, Leiden Baker Arend Oostwaert, Keyzerswaert, Two Men Playing Backgammon, Woman at her Toilet, Prince's Day, The Merry Family, Maria Rey, Consul Titus Manlius Torquatus, Swartenhont, Venus and Adonis

Next: Binary Classification

- Here we use the NGD
  for a Support Vector Machine (SVM)
  binary classification learner
  (we could also use a neural
  network)

Setup:

Anchor terms, positive/negative examples,

Test set → Accuracy

Using NGD in SVM (Support Vector Machines) to learn concepts (binary classification)

Training Data

Positive Training	(22 cases)			
avalanche	bomb threat	broken leg	burglary	car collision
death threat	fire	flood	gas leak	heart attack
hurricane	landslide	murder	overdose	pneumonia
rape	roof collapse	sinking ship	stroke	tornado
train wreck	trapped miners			
Negative Training	(25 cases)			
arthritis	broken dishwasher	broken toe	cat in tree	contempt of court
dandruff	delayed train	dizziness	drunkenness	enumeration
flat tire	frog	headache	leaky faucet	littering
missing dog	paper cut	practical joke	rain	roof leak
sore throat	sunset	truancy	vagrancy	vulgarity
Anchors	(6 dimensions)			
crime	happy	help	safe	urgent
wash				

Testing Results

	Positive tests	Negative tests
Positive Predictions	assault, coma, electrocution, heat stroke, homicide, looting, meningitis, robbery, suicide	menopause, prank call, pregnancy, traffic jam
Negative Predictions	sprained ankle	acne, annoying sister, campfire, desk, mayday, meal
Accuracy	15/20 = 75.00%	

Example:

Emergencies

Example: Religious Terms

Training Data

Positive Training	(22 cases)			
Allah	Catholic	Christian	Dalai Lama	God
Jerry Falwell	Jesus	John the Baptist	Mother Theresa	Muhammad
Saint Jude	The Pope	Zeus	bible	church
crucifix	devout	holy	prayer	rabbi
religion	sacred			
Negative Training	(23 cases)			
Abraham Lincoln	Ben Franklin	Bill Clinton	Einstein	George Washington
Jimmy Carter	John Kennedy	Michael Moore	atheist	dictionary
encyclopedia	evolution	helmet	internet	materialistic
minus	money	mouse	science	secular
seven	telephone	walking		
Anchors	(6 dimensions)			
evil	follower	history	rational	scripture
spirit				

Testing Results

	Positive tests	Negative tests
Positive Predictions	altar, blessing, communion, heaven, sacrament, testament, vatican	earth, shepherd
Negative Predictions	angel	Aristotle, Bertrand Russell, Greenspan, John, Newton, Nietzsche, Plato, Socrates, air, bicycle, car, fire, five, man, monitor, water, whistle
Accuracy	24/27 = 88.89%	

Example: Classifying Prime Numbers

Training Data

Positive Training	(21 cases)			
11	13	17	19	2
23	29	3	31	37
41	43	47	5	53
59	61	67	7	71
73				
Negative Training	(22 cases)			
10	12	14	15	16
18	20	21	22	24
25	26	27	28	30
32	33	34	4	6
8	9			
Anchors	(5 dimensions)			
composite	number	orange	prime	record

Testing Results

	Positive tests	Negative tests
Positive Predictions	101, 103, 107, 109, 79, 83, 89, 91, 97	110
Negative Predictions		36, 38, 40, 42, 44, 45, 46, 48, 49
Accuracy	18/19 = 94.74%	

Actually, 91 is not A prime. This is a false positive. So Accuracy is 17/19= 89,47%

Example: Electrical Terms

Training Data

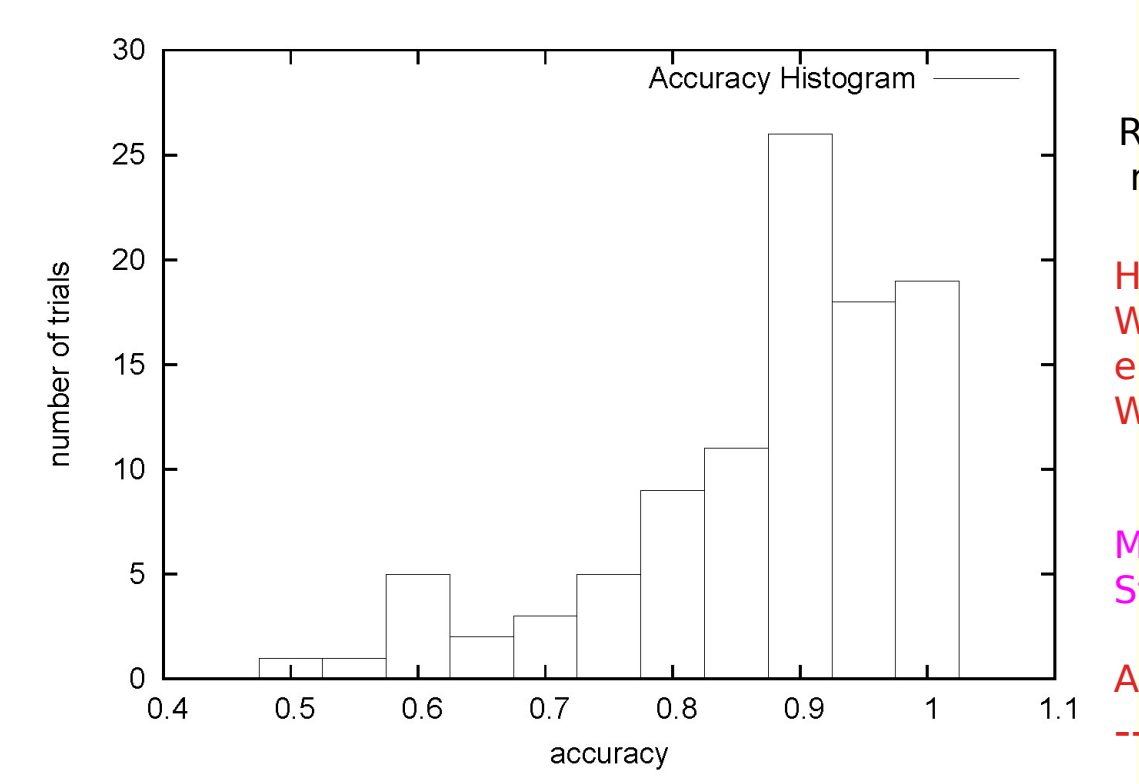
Positive Training	(58 cases)			
Cottrell precipitator	Van de Graaff generator	Wimshurst machine	aerial	antenna
attenuator	ballast	battery	bimetallic strip	board
brush	capacitance	capacitor	circuit	condenser
control board	control panel	distributer	electric battery	electric cell
electric circuit	electrical circuit	electrical condenser	electrical device	electrical distribut
electrical fuse	electrical relay	electrograph	electrostatic generator	electrostatic machi
filter	flasher	fuse	inductance	inductor
instrument panel	jack	light ballast	load	plug
precipitator	reactor	rectifier	relay	resistance
security	security measures	security system	solar array	solar battery
solar panel	spark arrester	spark plug	sparking plug	suppresser
transmitting aerial	transponder	zapper		
Negative Training	(55 cases)			
Andes	Burnett	Diana	DuPonts	Friesland
Gibbs	Hickman	Icarus	Lorraine	Madeira
Quakeress	Southernwood	Waltham	Washington	adventures
affecting	aggrieving	attractiveness	bearer	boll
capitals	concluding	constantly	conviction	damming
deeper	definitions	dimension	discounting	distinctness
exclamation	faking	helplessness	humidly	hurling
introduces	kappa	maims	marine	moderately
monster	parenthesis	pinches	predication	prospect
repudiate	retry	royalty	shopkeepers	soap
sob	swifter	teared	thrashes	tuples
Anchors	(6 dimensions)			
bumbled	distributor	premeditation	resistor	suppressor
swimmers				

Testing Results

	Positive tests	Negative tests
Positive Predictions	cell, male plug, panel, transducer, transformer	
Negative Predictions		Boswellizes, appointer, enforceable, greatness, planet
Accuracy	10/10 = 100.00%	

Comparison with WordNet Semantics

http://www.cogsci.princeton.edu/~wn



NGD-SVM Classifier on 100 randomly selected WordNet Categories

Randomly selected positive, negative and test sets

Histogram gives accuracy With respect to PhD experts entered knowledge in the WordNet Database

Mean Accuracy is 0.8725
Standard deviation is 0.1169

Accuracy almost always > 75%
--Automatically

Next: Translation Using NGD

Problem:

Given starting vocabulary	
English	**Spanish**
tooth	diente
joy	alegria
tree	arbol
electricity	electricidad
table	tabla
money	dinero
sound	sonido
music	musica
Unknown-permutation vocabulary	
plant	bailar
car	hablar
dance	amigo
speak	coche
friend	planta

Translation:

Predicted (optimal) permutation	
English	**Spanish**
plant	planta
car	coche
dance	bailar
speak	hablar
friend	amigo