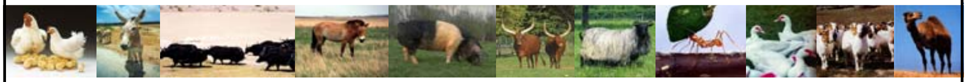
An autonomous agent algorithm to compute the Weitzman genetic diversity for large sets of taxonomic units



Henner Simianer and Manfred Tietze

Institute of Animal Breeding and Genetics
Georg-August University Göttingen
Germany

The problem

→ Weitzman‘s diversity concept (Weitzman 1992, 1993) is widely accepted as a framework to derive priorities for conservation of livestock breeds

→ Weitzman‘s diversity metric D is tedious to calculate

→ Weitzman (1992) presented a recursive algorithm, with N breeds required storage capacity and computing time are ~ proportional to 2^N

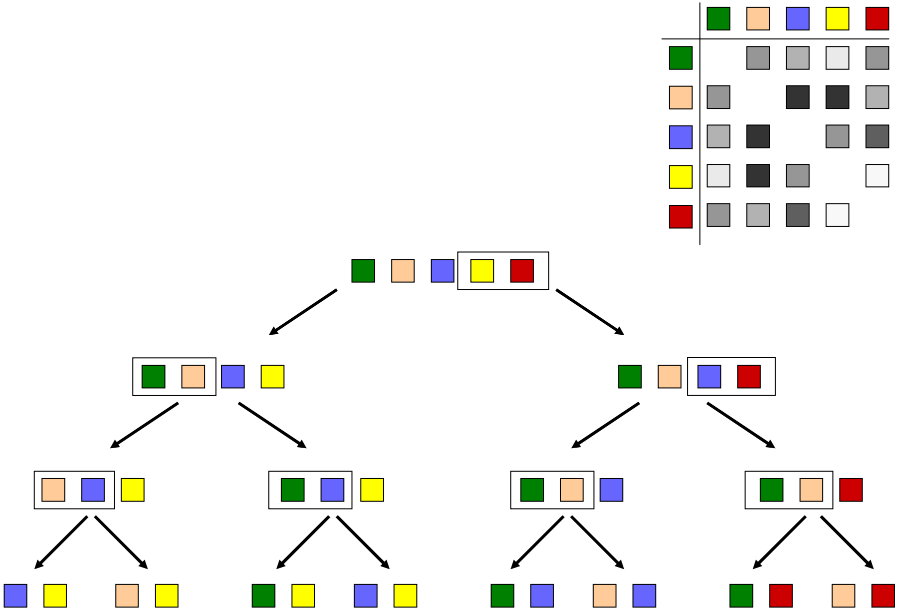
→ exact solutions only published for up to N = 35 breeds

→ suggested approximations (e.g. Thaon d‘Arnoldi et al, 1998) yield a systematic and massive downward bias

The objectives

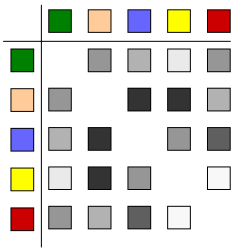
- → Redefine the problem of the calculation of the Weitzman diversity D in a different mathematical context (combinatorial optimisation problem)
- → Present an autonomous agent algorithm to find good solutions (arbitrarily close to the unique exact solution) in linear time
- → Derive a conservative estimate of the magnitude of the estimation error

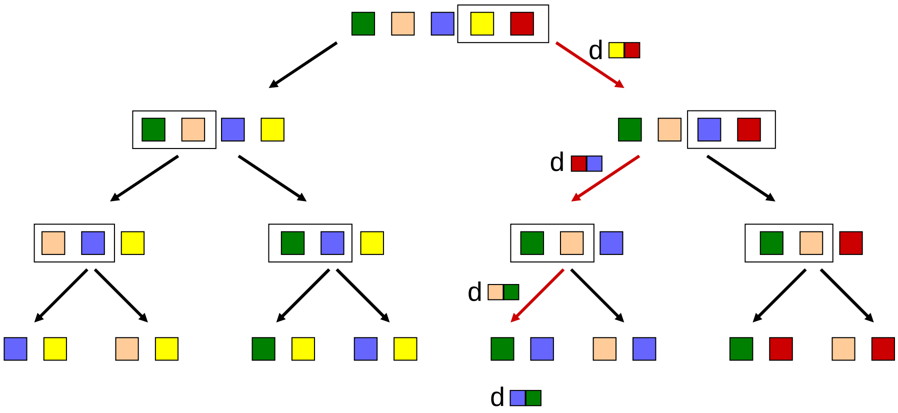
Weitzman‘s recursive algorithm



Weitzman‘s recursive algorithm

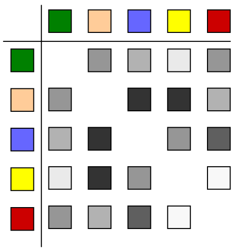
Link h	Representative g	d(g,h)
🟨	🟥	d 🟨🟥
🟥	🟦	d 🟥🟦
🟧	🟩	d 🟧🟩
🟦	🟩	d 🟦🟩
🟩		D = Σd(g,h)





Weitzman‘s recursive algorithm

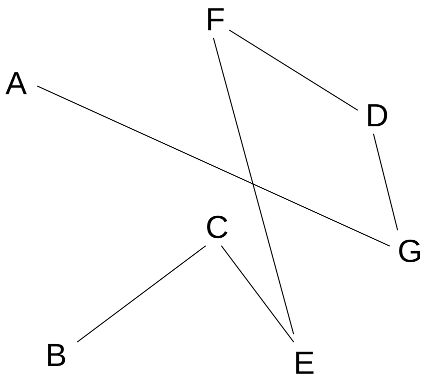
Link h	Representative g	d(g,h)
🟨	🟥	d 🟨🟥
🟥	🟦	d 🟥🟦
🟧	🟩	d 🟧🟩
🟦	🟩	d 🟦🟩
🟩		D = Σd(g,h)



- The sequence of link elements is unique
- If the sequence of link elements is known, D is easily computed from the distance matrix

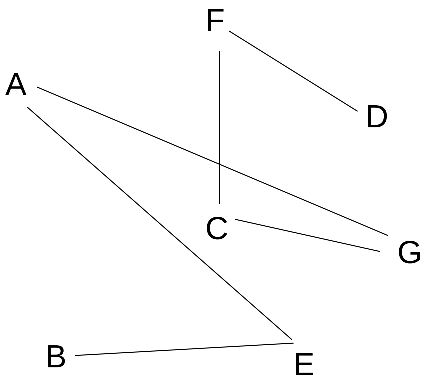
→ Find the sequence of link elements that maximizes D

Mathematically, this problem is in the class of discrete combinatorial optimisation problems (cf. ‚travelling salesman problem‘)



B
C
E
F
D
G
A

→ D_1

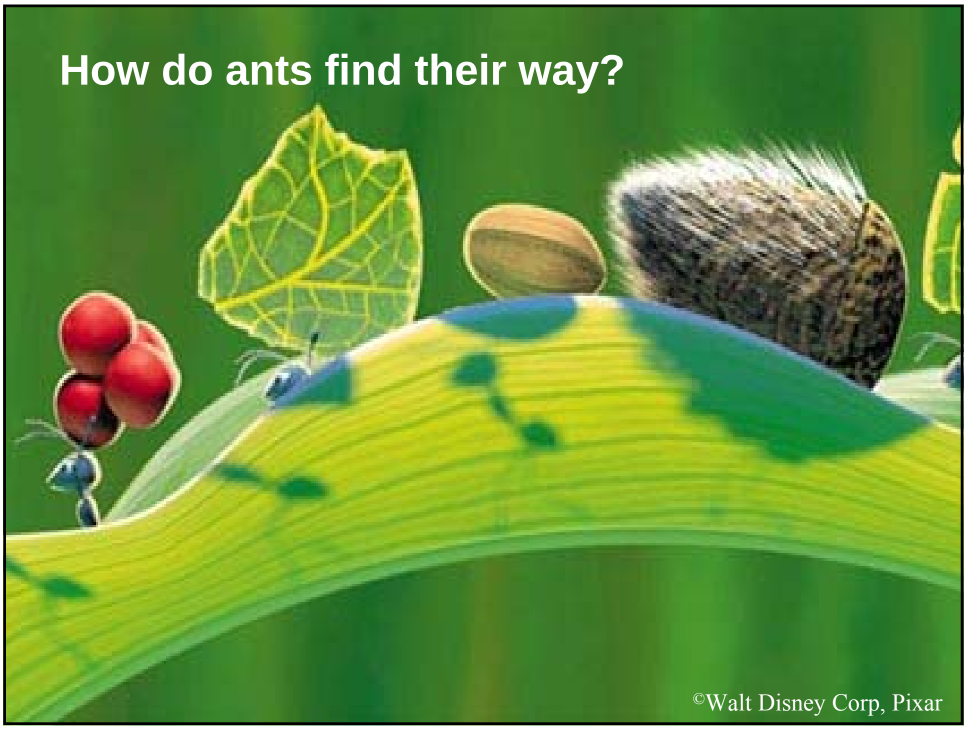


B
E
A
G
C
F
D

→ D_2

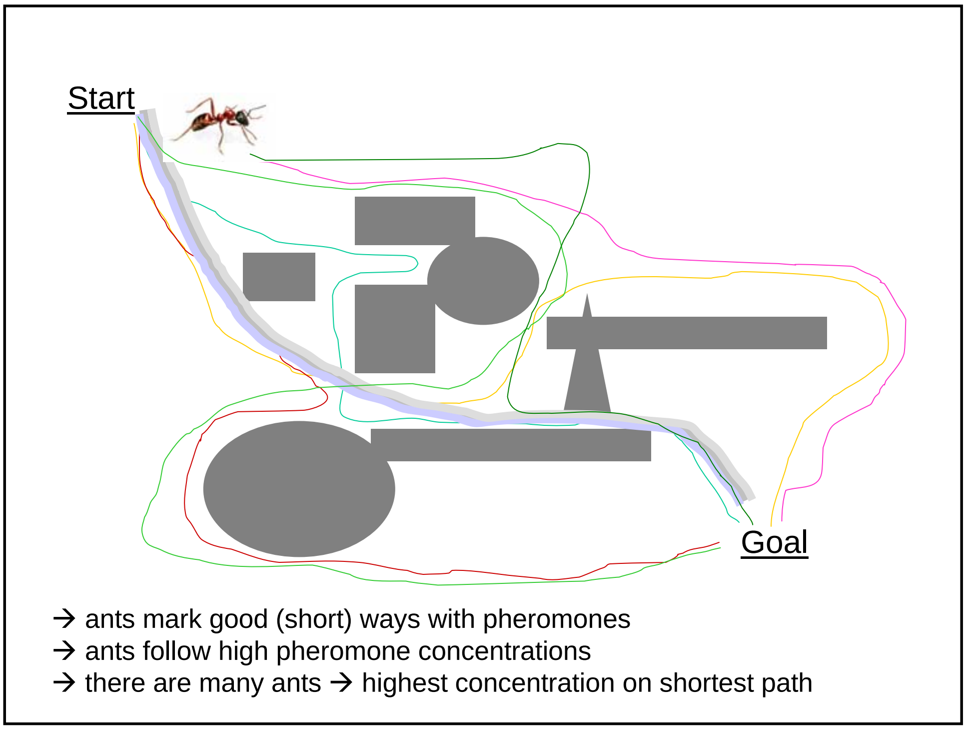
N! (>> 2^N) possible orders

Standard problem of mathematics, lots of efficient algorithms with known good properties

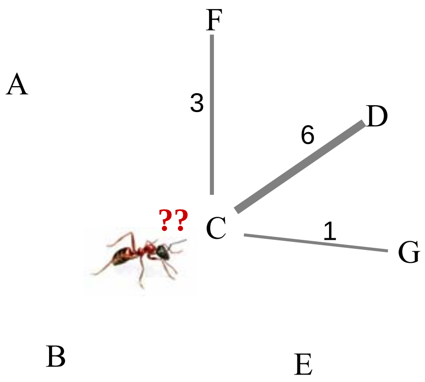


How do ants find their way?

©Walt Disney Corp, Pixar

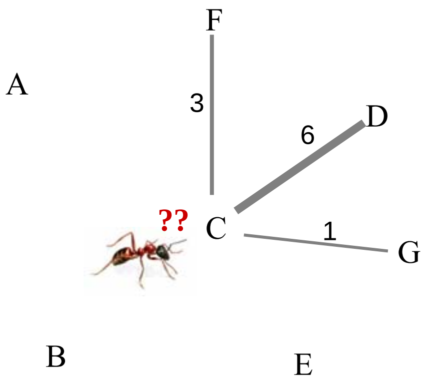


→ ants mark good (short) ways with pheromones
→ ants follow high pheromone concentrations
→ there are many ants → highest concentration on shortest path



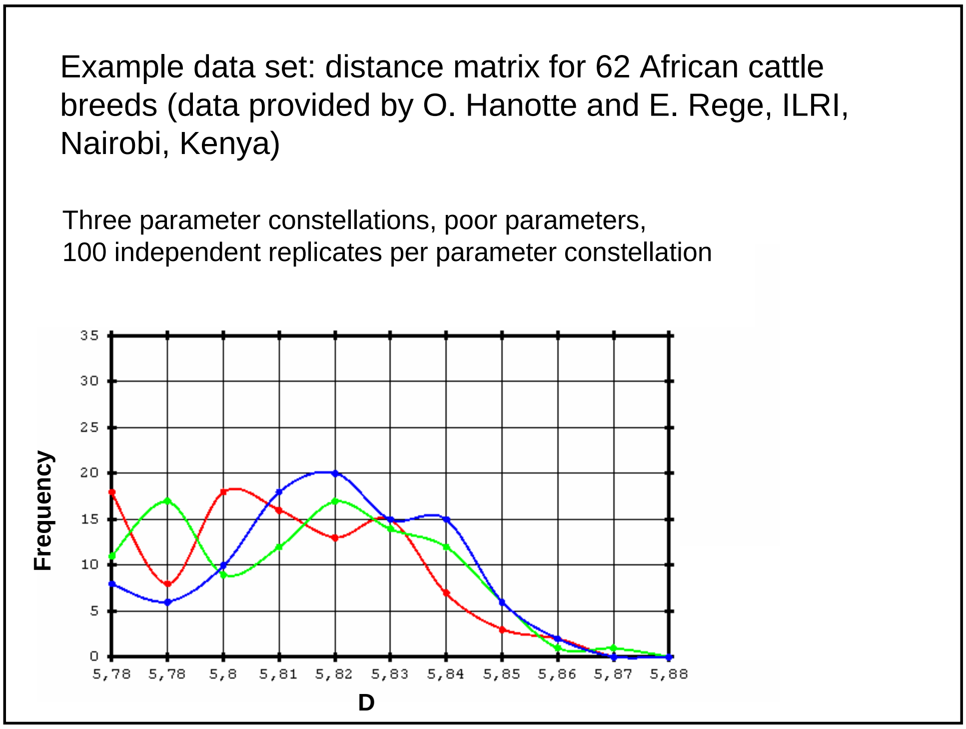
An ,ant algorithm‘ to find the best way:

1. Send many (100) ants on random paths
2. identify the best (,largest D‘) path, mark this path with a pheromone proportional to its quality
3. Send again 100 ants, let them decide according to pheromone concentration



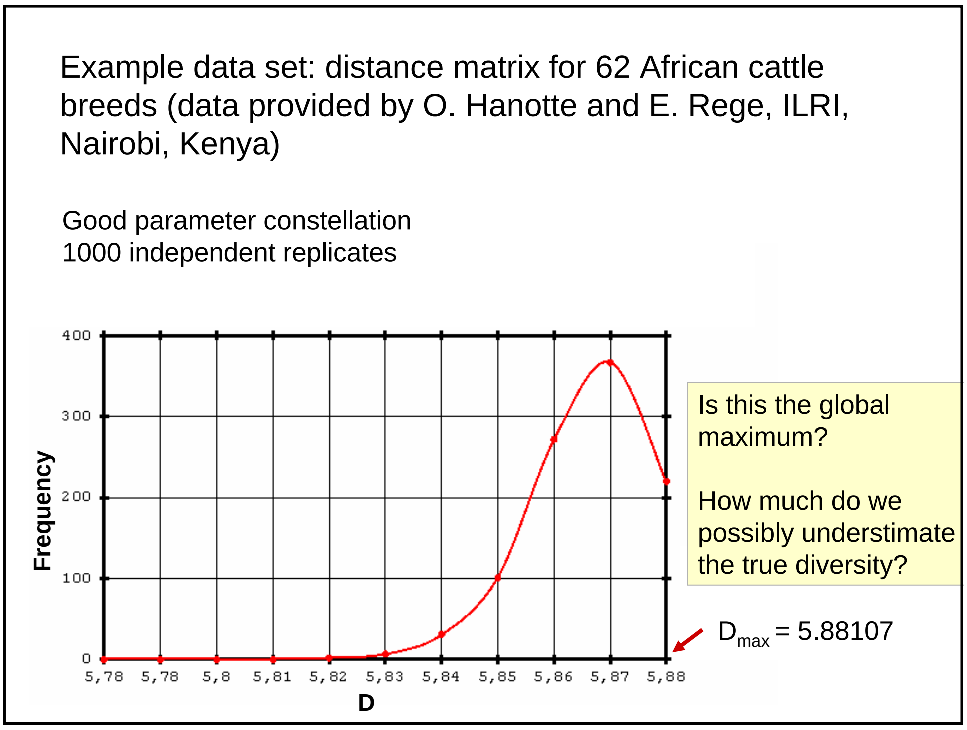
An ,ant algorithm‘ to find the best way:

1. Send many (100) ants on random paths
2. identify the best (,largest D‘) path, mark this path with a pheromone proportional to its quality
3. Send again 100 ants, let them decide according to pheromone concentration, then mark again
   ... pheromone concentration, pheromone degradation, independent restarts, parameter optimization



Example data set: distance matrix for 62 African cattle breeds (data provided by O. Hanotte and E. Rege, ILRI, Nairobi, Kenya)

Three parameter constellations, poor parameters,
100 independent replicates per parameter constellation



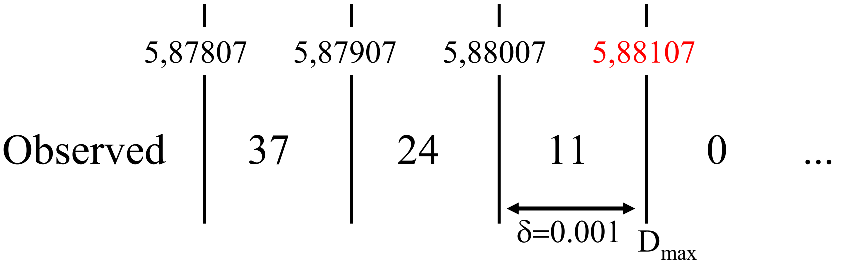
Example data set: distance matrix for 62 African cattle breeds (data provided by O. Hanotte and E. Rege, ILRI, Nairobi, Kenya)

Good parameter constellation
1000 independent replicates

Is this the global maximum?

How much do we possibly understimate the true diversity?

D_{max} = 5.88107



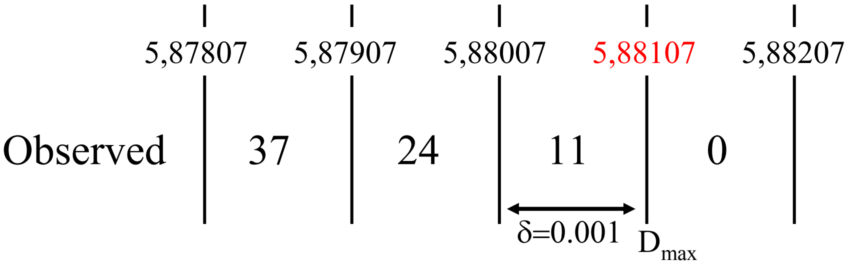
5,87807 5,87907 5,88007 5,88107

Observed 37 24 11 0 ...

δ=0.001 D_{max}

H_o: no further realisations in the interval $[D_{max}, \infty]$

H_a: further realisations in the interval $[D_{max}, \infty]$



5,87807 5,87907 5,88007 5,88107 5,88207

Observed 37 24 11 0

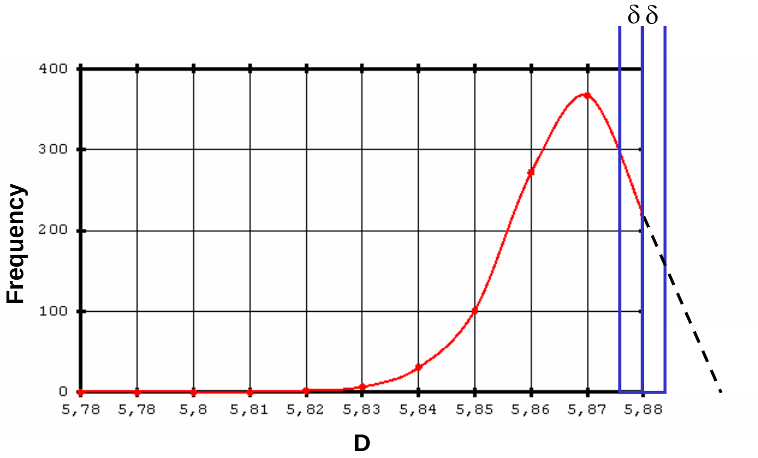
δ=0.001 D_{max}

H_o: no further realisations in the interval $[D_{max}, \infty]$

H_a: further realisations in the interval $[D_{max}, \infty]$

H_o: no further realisations in the interval $[D_{max}, D_{max}+ \delta]$

H_a: further realisations in the interval $[D_{max}, D_{max}+ \delta]$



5,87807 5,87907 5,88007 5,88107 5,88207

Observed	37	24	11	$N_{obs} = 0$

11/24 = 0,4583

0,4583 x 11
= 5,0417 = N_{exp}

$$X^2 = \frac{(N_{obs} - N_{exp})^2}{N_{exp}} = \frac{(0 - N_{exp})^2}{N_{exp}}$$

$$= N_{exp} \sim \chi^2_{1df} \begin{cases} \alpha = 0.05 & \rightarrow 3.84 \\ \alpha = 0.01 & \rightarrow 6.63 \\ \alpha = 0.001 & \rightarrow 10.83 \end{cases}$$

$N_{exp} = 5.0417 \rightarrow \alpha = 0.024$

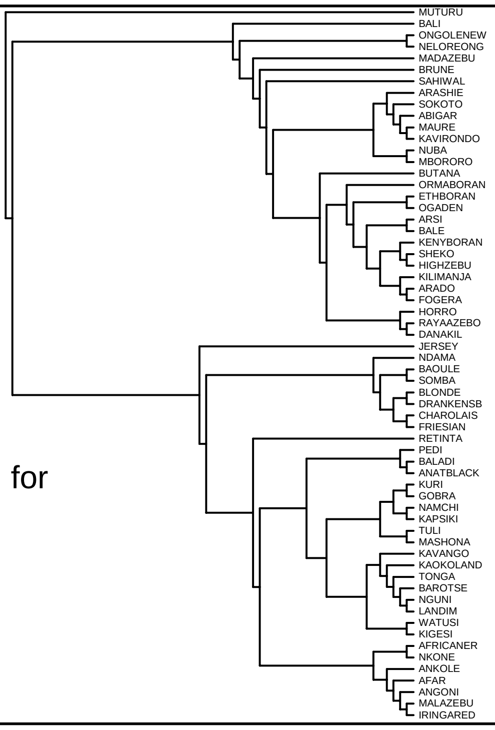
5,87807 5,87907 5,88007 5,88107 5,88207

Observed 37 24 11 $N_{obs} = 0$

→ Probability is 2.4 per cent, that the true value is in the interval [5,88107; 5,88207] but was by chance not observed

max. error = 100 x δ/D_{max} = 0.017 per cent

More replicates → reduce $\alpha|\delta$ or
→ reduce $\delta|\alpha$



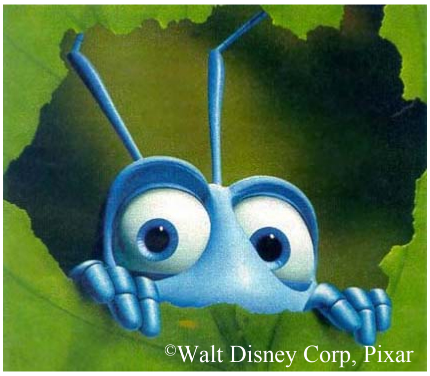
Maximum likelihood tree for
62 African cattle breeds

Estimated Weitzman diversity
D = 5.88107

Conclusions

- → Calculation of Weitzman's diversity redefined as a discrete combinatorial optimisation problem
- → An ant (autonomous agent) algorithm is suggested to find good solutions in linear time
- → Arguments are presented to estimate the possible bias, which is small
- → In all data sets where the exact solution is known, the same result was obtained with the ant algorithm
- → Size of the optimisation problem is no longer prohibitive (used to calculate D for up to 800 units)

©Walt Disney Corp, Pixar



©Walt Disney Corp, Pixar

Any questions?