

Taxonomi og klassifikationstræer

Af: *Kresten Cæsar Torp og Nicolai Sørensen, Aalborghus Gymnasium*

Man har i flere århundreder inddelt de levende organismer efter deres kropsform og kropsbygning, også kaldet morfologi og anatomi. Det kan man gøre for både dyr, planter og svampe. Hos bakterier er disse sværere at iagttage, så her har man desuden anvendt biokemiske egenskaber, fx at de indeholder bestemte enzymer og kan foretage bestemte reaktioner i deres stofskifte. Morfologisk kriterier bliver i dag i høj grad suppleret med DNA-analyser. Siden Linné's udgivelse af værket *Systema Natura* i 1751 har man gjort dette hierarkisk efter følgende system:

Niveau\Eksempel	Gråsæl <i>Halichoerus grypus</i>	Menneske <i>Homo sapiens</i>	Bøgetræ <i>Fagus silvestris</i>
Domæne	Kerneholdige, <i>Eukaryota</i>		
Rige	Dyreriget, <i>Animalia</i>		Planteriget, <i>Plantae</i>
Division	De bilateralt symmetriske, <i>Bilateria</i>		Karplanter, <i>Tracheophyta</i>
Række	De rygstrengede, <i>Chordata</i>		Frøplanter, <i>Spermatophyta</i>
Klasse	Pattedyr, <i>Mammalia</i>		Tokimbladede, <i>Magnoliopsida</i>
Orden	Rovdyr, <i>Carnivora</i>	Aberne, <i>Primata</i>	Bøgeordenen, <i>Fagales</i>
Familie	Sæler, <i>Phocidae</i>	Menneskeaberne, <i>Hominideae</i>	Bøgefamilien, <i>Fagaceae</i>
Slægt	Øreløse sæler, <i>Halichoerus</i>	Menneske, <i>Homo</i>	Bøgeslægten, <i>Fagus</i>
Art	Gråsæl, <i>grypus</i>	Tænkende menneske, <i>sapiens</i>	Bøg, <i>silvestris</i>

I denne opgave skal I prøve at identificere relevante morfologiske og anatomiske egenskaber, som I kan anvendes til at inddele dyr indenfor dyrerækken Chordater (dem med rygstreng) i dyreklasser (benfisk, krybdyr, fugle og pattedyr). Chordaterne indeholder flere klasser, fx bruskskildede og padder. De tages for overskuelighedens skyld ikke med i denne opgave.

I skal vælge de morfologiske og anatomiske egenskaber, så de kan fungere som en generel procedure for at identificere dyr til ordensniveau.

I forhold til beslutningstræer kaldes egenskaber man inddeler efter også for "features", mens grupper man søger at inddele i kaldes "targets".

Opgave 1. Opstilling af klassifikationstræ

1. Observer dyrene på collagen i bilag 1.
2. Opstil kriterier for at kunne opdele dyrene i dyreklasser. Kriterierne skal være morfologiske eller anatomiske egenskaber. De skal overvejes, så de kan fungere som helt generelle egenskaber for at få chordater man møder identificeret til den rigtige klasse. Det kan være egenskaber, hvor man kan svare ja/nej, fx "vinger tilstede?".
3. Udvælg de tre egenskaber, som I mener er de mest relevante at anvende.
4. Skriv egenskaberne ind i klassifikationstræet, vist i bilag 2.
5. Gå nu videre med opgave 2, hvor I skal teste jeres træ på nye data i form af nye dyr.

Opgave 2. Test af klassifikationskriterier

Nu skal I teste, om jeres model rammer rigtigt. På hver spillebrik er dyrets navn angivet. I kolonne 2-4 skal I nu angive deres score i jeres klassifikationstræ. Ved at sammenligne med scoren kan træets nøjagtighed bestemmes.

- Bilag 3 indeholder fotos af 24 dyrearter. Klip dem ud som spillebrikker. Har man konserverede dyr til rådighed fra biologisamlingen, kan de også inddrages.
- Anvend nu klassifikationstræet i bilag 2 til at gruppere dem: Foretag for hvert dyr valg ud fra første kriterium, derefter andet og til sidst tredje. Der behøver ikke nødvendigvis lande dyr i alle felter.
- Skriv pointene dyrene opnåede for hvert valg ind i pointskemaet i bilag 4.
- Vurder ud fra hvad I ser, hvordan klassifikationstræet fungerede. Er der dyr I mener ramte forkert, når I sammenligner dem?

Nu skal klassifikationstræet testes.

- Åben en teksteditor, fx *Notesblok*
- Gem filen som en kommasepareret .csv-fil på din computer med et navn og placering, du kan huske.

Tekstfilen kan nu overføres til testprogrammet.

- Åben testprogrammet.

<https://colab.research.google.com/drive/10iS7CscZjpn7B47fRkesTzqFX-U8GNsY?usp=sharing>

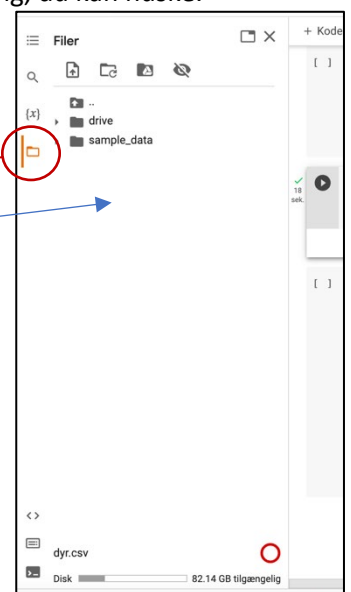
- Åbn filmappen.

- Træk din testfil over i programmets venstre side, og slip den.

- Gå nu trinvis frem. Når du holder musen over et trin, markeret med farvet kasse, kommer der en "play"-knap frem i barren: "[]". Tryk "play", og fortsæt til næste trin. Undervejs er der nogle få indtastninger, men du kan også senere redigere i programmet.

- Du skal undervejs angive en googlekonto, og acceptere, at programmet henter filer fra drev.

- I trin to skal du indsætte stien til din fil. Det gør du ved at højreklikke på din fil, kopiere sti og indsætte den i parentesen. Højreklik og "kopier sti". Sæt ind ved at taste "ctrl+v".



KlassifikationSkabelon.ipynb

Filer

```
#indlæsning af hele datasættet, du skal
# 1) Trække din csv fil ind under filer, det spiller ikke nogen rolle, hvor du gemmer det,
#    men brug gerne dine egne mapper på dit drev
# 2) Kopier stien til filen ved at højreklikke på filen og copy - paste ind nedenfor

data = pd.read_csv("/content/pointskema_KT.csv")

#Vi ser de første 10 rækker
data.head(10)

#DEN SIDSTE KOLONNE SKAL HEDDE "klasse", ellers skal man rette en smule i koden nedenfor
```

FileNotFoundError Traceback (most recent call last)
<ipython-input-13-806bfd1cc2d1> in <module>

- I step 41 indtaster du dine kategorier:

```

+ Kode + Tekst | Kopier til Drev
#Klasse - labels
#INDTAST SELV KLASSE LABELS, DETTE KRÆVER AT MAN SAMMENLIGNER MED FILEN, F.EKS
#I EKSEMPLET MED KREDITVÆRDIGHED SVARER
#0 TIL Ikke krediværdig OG
#1 TIL Krediværdig

class_labels = ['Fisk', 'Krybdyr', 'Fugle', 'Pattedyr']

```

19. I step 42 indtaster du størrelsen af dit trænings sæt.

```

[42] #Vi deler op i trænings - og - test - data
#De første n rækker bliver test - data, de resterende bliver træningsdata

#HER KAN MAN SELV BESLUTTE, HVOR MANGE AF DE FØRSTE RÆKKER
#DER SKAL BRUGES TIL TRÆNINGS DATA. NORMALT BRUGER MAN EN 75% - 80%
#AF DATASÆTTET TIL TRÆNING, OG RESTEN TIL TEST

n = 15 #HER SKAL DU SELV TASTE

```

20. I step 43 indtaster du dybden af dit træ, fx "3":

```

#print(y_train)

#Vi initialiserer en beslutningstræ - model med max dybde på 1

#DU KAN SELV BESTEMME DYBDEN, MEN PAS PÅ.
#EN ALT FOR STOR DYBDE LEDER TIL OVERFITTING, HVORIMOD OVERFITTING = VI SKELNER 'FOR MEGET'
#EN FOR LILLE DYBDE LEDER TIL UNDERFITTING UNDERFITTING = 'VI SKELNER FOR LIDT'

#PRØV DIG FREM, OG SE HVORDAN TRÆET KLARER SIG PÅ TRÆNINGSDATA

max_dybde = 3
model_tree = DecisionTreeClassifier(max_depth=max_dybde)

#Vi fitter modellen til vores træningsdata
model_tree.fit(X_train, y_train)

```

21. I step 44 får du et overblik over dit beslutningstræ. "Samples" angiver hvor mange arter der er endt i hver kasse. "Value" viser hvilken orden de tilhørte. DU kan her se, hvor mange der ramte rigtigt. Du kan også se, det valg, træet på den baggrund foretog.

22. De to sidste steps viser resultatet af testen. I step 60 får du en relativ værdi for hvor godt træet virker. I sidste step får du en relativ værdi for hvor meget de tre spørgsmål hver for sig var i stand til at forudsige.

```

Name: klasse, dtype: int64

[60] #Nøjagtigheden, dette er et mål for, hvor godt din model klarede sig på test - data, dvs.
#hvor stor en andel kunne modellen forudsige korrekt

accuracy_score(y_test, y_pred)

0.8888888888888888

#Hvor stor procentvis betydning har de enkelte features i forhold til at forudsige klassen
#dette er utroligt interessant, da man selvfølgelig er interesseret i hvilke
#features der har størst betydning

model_tree.feature_importances_

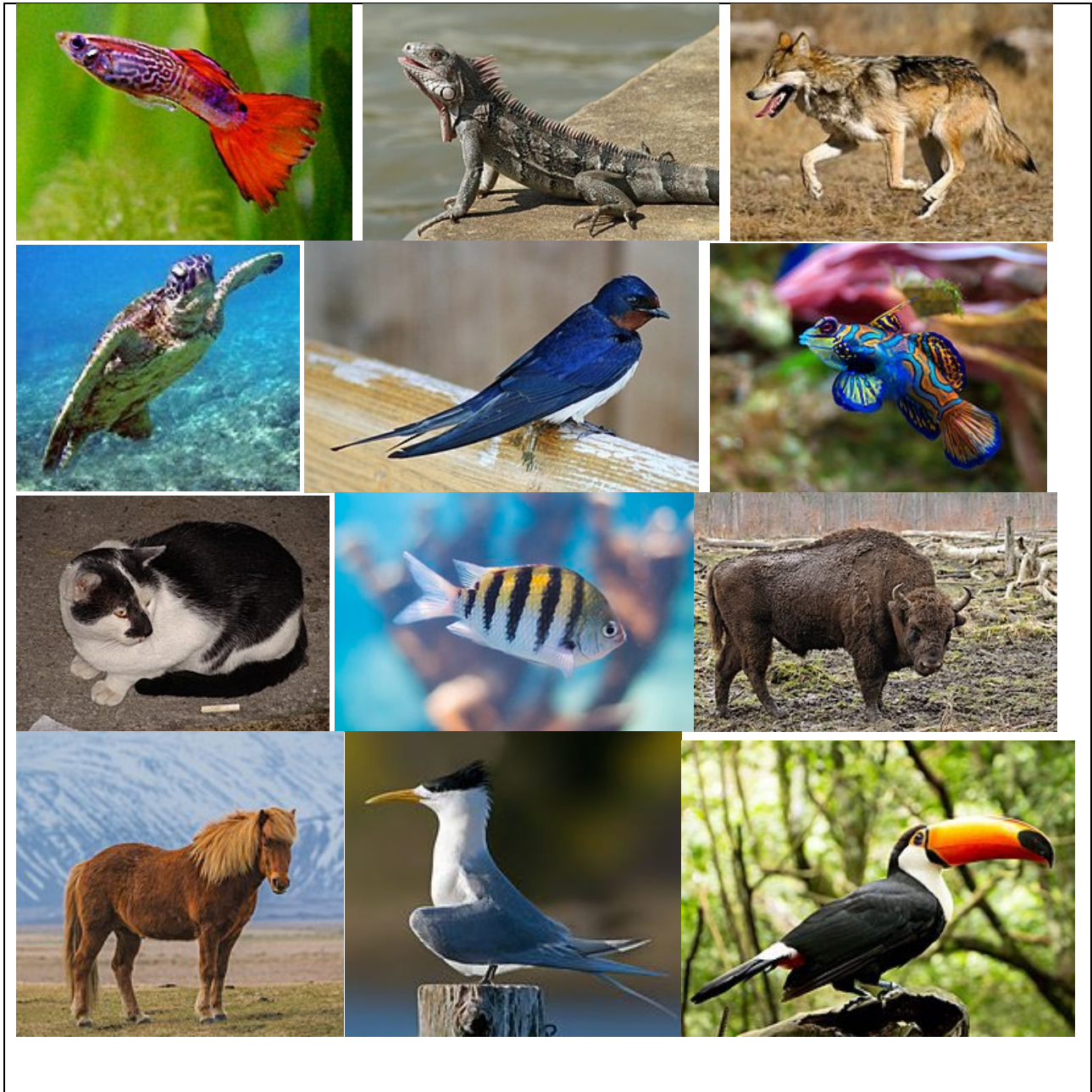
array([0.24901186, 0.36004312, 0.39094502])

```

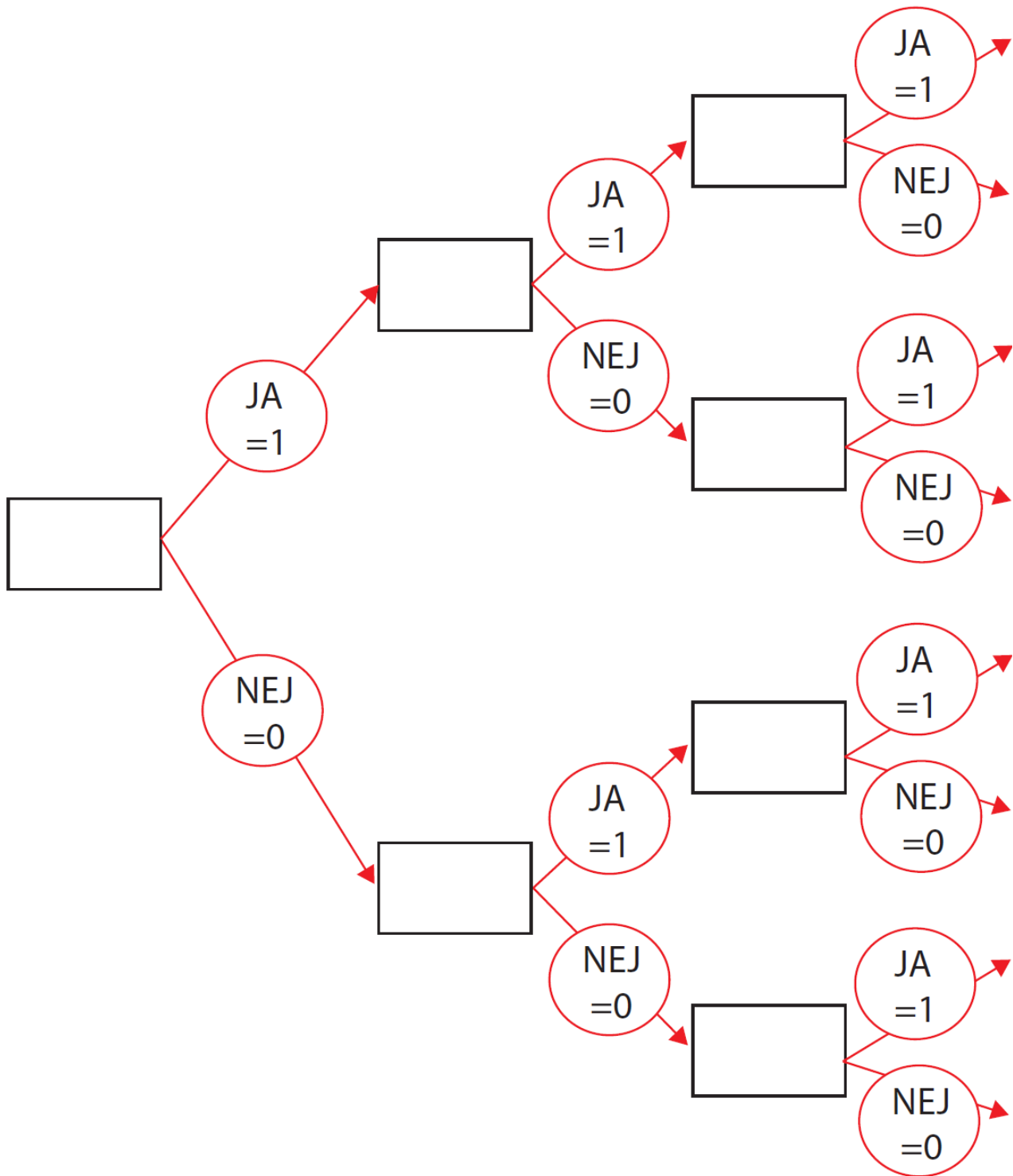
23. Overvej, hvilke forbedringer du kunne foreslå til træet.

- Skal nogle features skiftes ud?
- Skal rækkefølgen af features være anderledes?
- Kan et træ med tre features klare opgaven?

Bilag 1. Dyrecollage



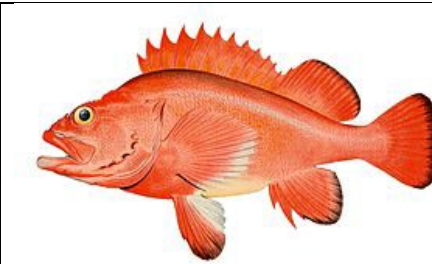
Bilag 2. Klassifikationstræ



Bilag 3. Testdata -Dyrekort (Klip kortene ud)



Kameleon



Rødfisk



Dragefisk



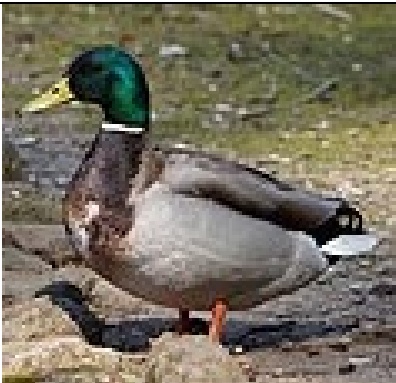
Flodhest



Kongeørn



Kolibri



Gråand



Skildpadde



Tyrannosaurus rex



Flyvefisk



Søhest



Rødspætte



Torsk



Krokodille



Markfirben



Hugorm



Stork



Gråspurv



Pingvin



Gorilla



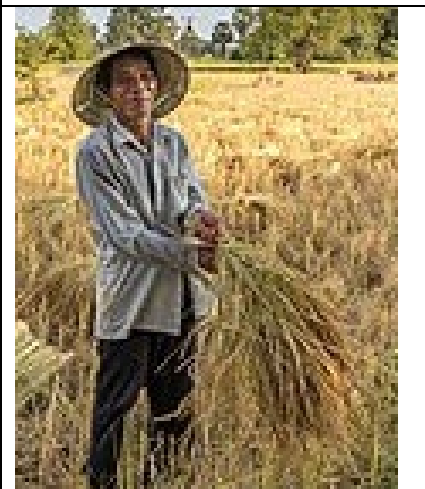
Spækhugger



Elefant



Flagermus



Menneske

Bilag 4. Pointskema

Dyr	Features (navngiv de tre features, og skriv opnåede point ind for hver art)			Klasse
	1:	2:	3:	
Dragefisk				1 Osteichthyes (benfisk)
Elefant				4 Mammalia (pattedyr)
Flagermus				4 Mammalia (pattedyr)
Flodhest				4 Mammalia (pattedyr)
Flyvefisk				1 Osteichthyes (benfisk)
Gorilla				4 Mammalia (pattedyr)
Gråand				3 Aves (fugle)
Gråspurv				3 Aves (fugle)
Hugorm				2 Reptilia (krybdyr)
Kamæleon				2 Reptilia (krybdyr)
Kolibri				3 Aves (fugle)
Kongeørn				3 Aves (fugle)
Krokodille				2 Reptilia (krybdyr)
Markfirben				2 Reptilia (krybdyr)
Menneske				4 Mammalia (pattedyr)
Pingvin				3 Aves (fugle)
Rødfisk				1 Osteichthyes (benfisk)
Rødspætte				1 Osteichthyes (benfisk)
Skildpadde				2 Reptilia (krybdyr)
Søhest				1 Osteichthyes (benfisk)
Spækhugger				4 Mammalia (pattedyr)
Stork				3 Aves (fugle)
Torsk				1 Osteichthyes (benfisk)
Tyrannosaurus				2 Reptilia (krybdyr)

Resultatskemaet gemmes som en kommasepareret .csv-fil efter nedenstående format (semikolonsepareret fil accepteres også):

```

4lemmer,vinger,pels,klasse
0,0,0,1
1,0,1,4
1,1,1,4
1,0,0,4
0,1,0,1
1,0,1,4
1,1,0,3
1,1,0,3
0,0,0,2
1,0,0,2
1,1,0,3
1,1,0,3
1,0,0,2
1,0,0,2
1,0,1,4
1,1,0,3
0,0,0,1
0,0,0,1
1,0,0,2
0,0,0,1
0,0,0,4
1,1,0,3
0,0,0,1
1,0,0,2

```


Kilder til fotos:

Alle fotos stammer fra: [Creative Commons — Attribution 2.0 Generic — CC BY 2.0](https://creativecommons.org/licenses/by/2.0/)

Søhest: Vassil

Rødspætte: 4028mdk09

Torsk: Wilhelm Thomas Fiege

Krokodille: Bobisbob

Hugorm: Zwentibold

Markfirben: Quartl

Stork: W Schulenburg

Gråspurv: Zeynel Cebeci

Pingvin: Sander van der Wel

Gorilla: Brocken Inaglory

Spækhugger: Mlewan

Elefant: Charles J. Sharp

Menneske: Basile Morin

Flagermus: PD-USGov, exact author unknown

Guppy: Jdiemer

Sergeantfisk: Matthew T Rader

Grøn iguan: Rjcastillo

Havskilpadde: Brocken Inaglory

Landsvale: Malene Thyssen

Mexikansk ulv: Jim Clark

Kat: Fernando Losada Rodríguez

Flyvefisk: Tara Casazza

Flodhest: P. Brundel

Kongeørn: Jarkko Järvinen

Kolibri: US Gov

Gråand: Commonists

Skilpadde: Jonathan Zander

Tyrannosaurus rex: Nobu Tamura

Kameleon: Heionlein

Ichtyosaurus: Red Natters

Dragefisk: Jens Petersen

Rødfisk: Rocco Leandre Aguilera

Sergeantfisk: Mathew Rader

Mandarinfisk: Luc Viatour

Bison: Michael Gäbler

Hest: Eatcha

Terne: J. Harrison

Tukan: Julio Cesar Lopes