VIMIAC16 2025/26/I.

Döntési fák

Előadó: Dr. Hullám Gábor

Mesterséges intelligencia előadássorozat

Az előadás diái az AIMA könyvre épülve (http://aima.cs.berkeley.edu) készültek a University of California, Berkeley mesterséges intelligencia kurzusának anyagainak felhasználásával (http://ai.berkeley.edu).

These slides are based on the AIMA book (http://aima.cs.berkeley.edu) and were adapted from the AI course material of University of California, Berkeley (http://ai.berkeley.edu).

Problémamegoldás

- Keresés
- Kényszer kielégítés
- Keresés ellenséges térben
- Heurisztikák

Tudásreprezentáció következtetés

- Ontológia
- Logika
- Való-színűségi hálók
- Bizony-talanság kezelése

Döntések

- Hasznosság
- Döntési hálók

Megerősítéses tanulás

- Szekvenciális döntések
- Markov döntési folyamat

Felügyelt tanulás

- Regressziós modellek
- Döntési fák
- Neurális hálók
- Együttes modellek

Tanulás minták alapján

- A következőkben olyan módszereket vizsgálunk, amikor minták, **mintapéldák alapján akarjuk kialakítani a döntési rendszerünket**.

- Nagyon gyakran a számítógép tanulásához rendelkezésünkre áll egy csomó mintapélda, és ez hordozza az információt.
- Nincsenek előre kialakított szabályaink a feladatra, csak a mintáink.

Tanulás minták alapján

- Persze ilyenkor is valamilyen struktúrában igyekszünk felhasználni a minták hordozta tudást.

- Tehát a módszerünk – az, hogy hogyan használjuk fel a mintákat – bizonyos mértékben problémafüggetlen.

- A hétköznapi életben nagyon gyakori a minta alapján való tanulás, míg a klasszikus (számítógépes) megoldásokban a humán fejlesztő szabályokat alkotott, és ezeket programozta be.

Tanulás vs. analitikus tervezés

- Egy algoritmust, pl. egy döntést (ágensfüggvényt) alapvetően két módon lehet megvalósítani:

(1) **Analitikus tervezés**:

- Begyűjteni az analitikus (fizikai, kémiai stb. összefüggésekből felépített) modelleket az adott problémára.
- Megtervezni analitikusan a konkrét mechanizmust, és azt algoritmusként implementálni.

- (2) **Tanulás:**
  - Megtervezni a tanulás mechanizmusát, és azt algoritmusként implementálni.
  - Majd alkalmazásával megtanulni vele a minták alapján a döntés tényleges mechanizmusát, és azt algoritmusként implementálni.

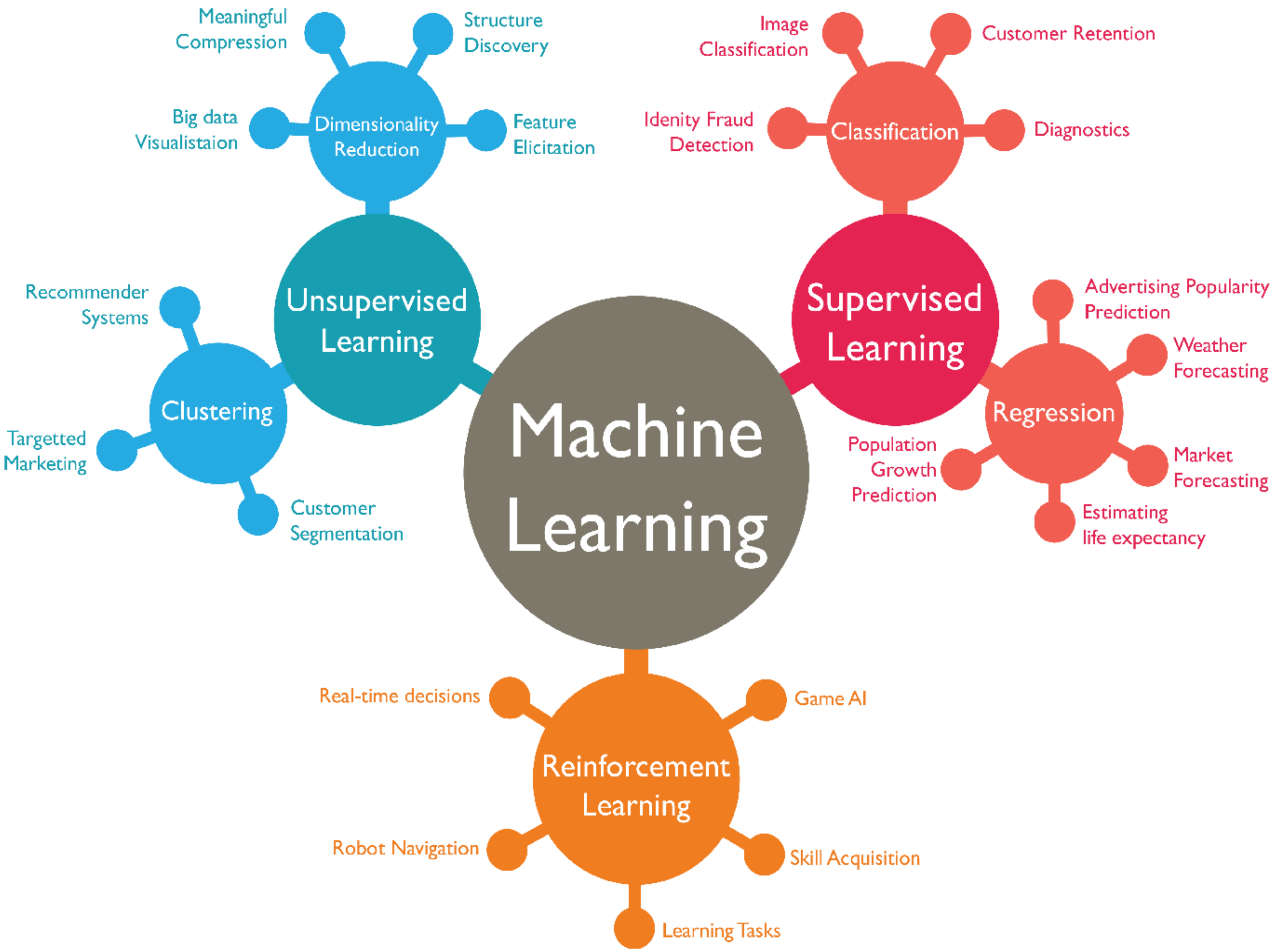
- A következőkben a (2)-vel foglalkozunk

Gépi tanulás főbb fajtái

- **felügyelt tanulás** egy esetnél mind a bemenetet, mind a kimenetet észlelni tudjuk (bemeneti minta + kívánt válasz)
- **megerősítéses tanulás** az ágens az általa végrehajtott tevékenység csak bizonyos értékelését kapja meg, esetleg nem is minden lépésben (jutalom, büntetés, **megerősítés**)
- **felügyelet nélküli tanulás** semmilyen információ sem áll rendelkezésünkre a helyes kimenetről (az észlelések közötti összefüggések tanulása)

Gépi tanulás főbb fajtái

- **féligellenőrzött tanulás** a tanításra használt esetek egy részénél mind a bemenetet, mind a kimenetet észlelni tudjuk (bemeneti minta + kívánt válasz), a másik – tipikusan nagyobb – részénél csak a bemeneti leírás ismert



Wahid, Abdul. “Big Data and Machine Learning for Businesses.” LinkedIn SlideShare, April 26, 2017. https://www.slideshare.net/awahid/big-data-and-machine-learning-for-businesses.

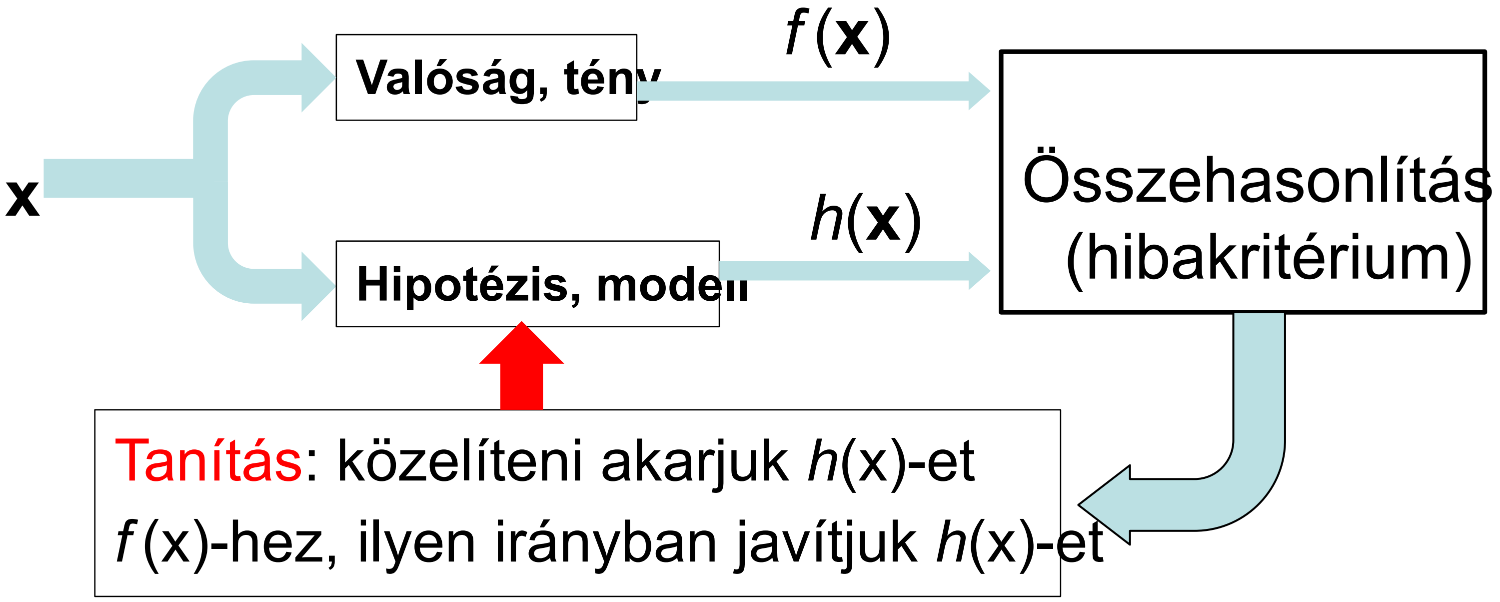
Felügyelt tanulás

tanulási példa: $(\mathbf{x}, f(\mathbf{x}))$ adatpár, ahol $f(\mathbf{x})$ ismeretlen

tanulás célja: $f(\mathbf{x})$ értelmes közelítése egy $h(\mathbf{x})$ hipotézissel

$h(\mathbf{x}) = f(\mathbf{x})$, **x** – ismert példákon gyakran teljes pontosság

$h(\mathbf{x'}) \cong f(\mathbf{x'})$, **x' – a tanulás közben még nem látott esetek** (**általánosító képesség**)



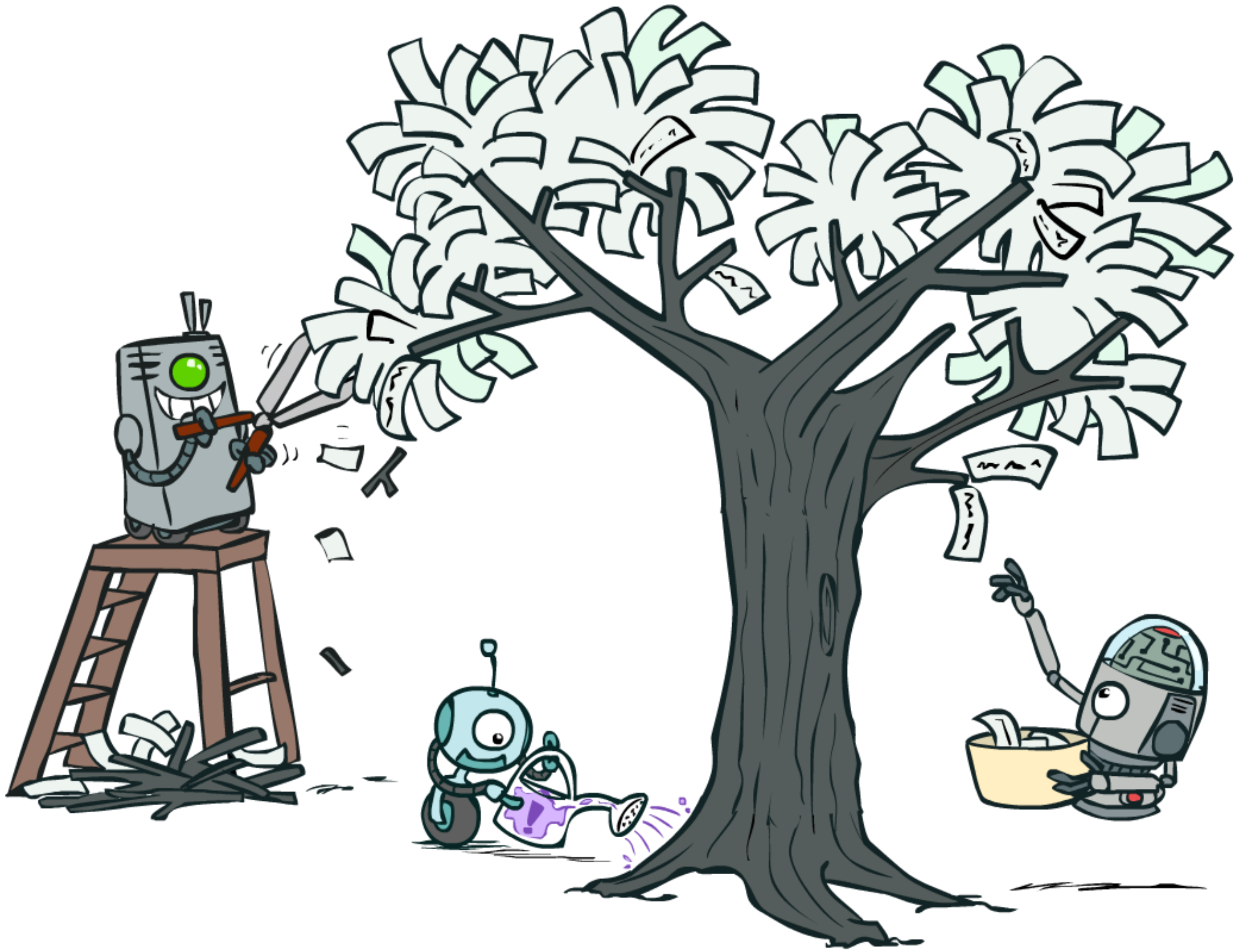
Feladat: *az ismeretlen **f**-re vonatkozó példák egy halmaza alapján (**tanítóhalmaz**), adjon meg egy olyan **h** függvényt (**hipotézist**), amely tulajdonságaiban jól közelíti az **f**-et (amit **teszthalmazon** verifikálunk).*

Sokféle tanítható eszközt (=modellt) hoztak létre:

- neurális hálók
- Bayes-hálók
- kernelgépek
- döntési fák
- legközelebbi szomszéd osztályozók
- stb. stb.

A következőkben először a döntési fákkal foglalkozunk – működésük az emberi gondolkodás számára jóval jobban áttekinthető, mint pl. a mesterséges neurális hálóké. (white box ⇔ black box)
Vannak regressziós döntési fák is, de mi az osztályozással foglalkozunk.

Döntési fák



Példaprobléma: döntés, hogy egy adott étteremben várjunk-e asztalra?
Cél: előállítani a döntés logikai függvényét

A problémát tulajdonságokkal vagy **attribútumokkal** jellemezzük.

1. *Alternatíva*: van-e a környéken más megfelelő étterem. (I/N)
2. *Bár*: van-e az étteremben kényelmes bár, ahol várhatunk. (I/N)
3. *Pén/Szom*: igaz pénteken- és szombatonként. (I/N)
4. *Éhes*: éhesek vagyunk-e. (I/N)
5. *Kuncsaft*: hányan vannak már benn (*Senki*, *Néhány* és *Tele*).
6. *Ár*: mennyire drága (*Olcsó, Közepes, Drága*)
7. *Eső*: esik-e kint (I/N)
8. *Foglalás*: foglalni kell-e előzetesen az asztalt (I/N)
9. *Típus*: az étterem típusa (*Francia, Olasz, Thai v. Burger*)
10. *BecsültVár*: a pincér által bemondott várakozási idő
    (0-10 perc, 10-30, 30-60, >60 perc)

A **döntés eredményét** a döntési ágensfüggvény adja meg:

VárniFog = *f(Alternativa*, *Bár*, *Pén/Szom*, *Éhes*, *Kuncsaft*, *Ár*,
Eső, *Foglalás*, *Típus*, *BecsültVár*, …)

Az eddigi tapasztalataink - tanítóminták

Pl.	Attribútumok										Cél
	Altern	*Bár*	*Pént*	*Éhes*	*Kuncs*	*Ár*	*Eső*	*Fogl*	*Típus*	*Becs*	*VárniFog*
X_1	*Igen*	*Nem*	*Nem*	*Igen*	*Néhány*	*Drága*	*Nem*	*Igen*	*Francia*	*0–10*	*Igen*
X_2	*Igen*	*Nem*	*Nem*	*Igen*	*Tele*	*Olcsó*	*Nem*	*Nem*	*Thai*	*30–60*	*Nem*
X_3	*Nem*	*Igen*	*Nem*	*Nem*	*Néhány*	*Olcsó*	*Nem*	*Nem*	*Burger*	*0–10*	*Igen*
X_4	*Igen*	*Nem*	*Igen*	*Igen*	*Tele*	*Olcsó*	*Nem*	*Nem*	*Thai*	*10–30*	*Igen*
X_5	*Igen*	*Nem*	*Igen*	*Nem*	*Tele*	*Drága*	*Nem*	*Igen*	*Francia*	*>60*	*Nem*
X_6	*Nem*	*Igen*	*Nem*	*Igen*	*Néhány*	*Közep*	*Igen*	*Igen*	*Olasz*	*0–10*	*Igen*
X_7	*Nem*	*Igen*	*Nem*	*Nem*	*Senki*	*Olcsó*	*Igen*	*Nem*	*Burger*	*0–10*	*Nem*
X_8	*Nem*	*Nem*	*Nem*	*Igen*	*Néhány*	*Közep*	*Igen*	*Igen*	*Thai*	*0–10*	*Igen*
X_9	*Nem*	*Igen*	*Igen*	*Nem*	*Tele*	*Olcsó*	*Igen*	*Nem*	*Burger*	*>60*	*Nem*
X_{10}	*Igen*	*Igen*	*Igen*	*Igen*	*Tele*	*Drága*	*Nem*	*Igen*	*Olasz*	*10–30*	*Nem*
X_{11}	*Nem*	*Nem*	*Nem*	*Nem*	*Senki*	*Olcsó*	*Nem*	*Nem*	*Thai*	*0–10*	*Nem*
X_{12}	*Igen*	*Igen*	*Igen*	*Igen*	*Tele*	*Olcsó*	*Nem*	*Nem*	*Burger*	*30–60*	*Igen*

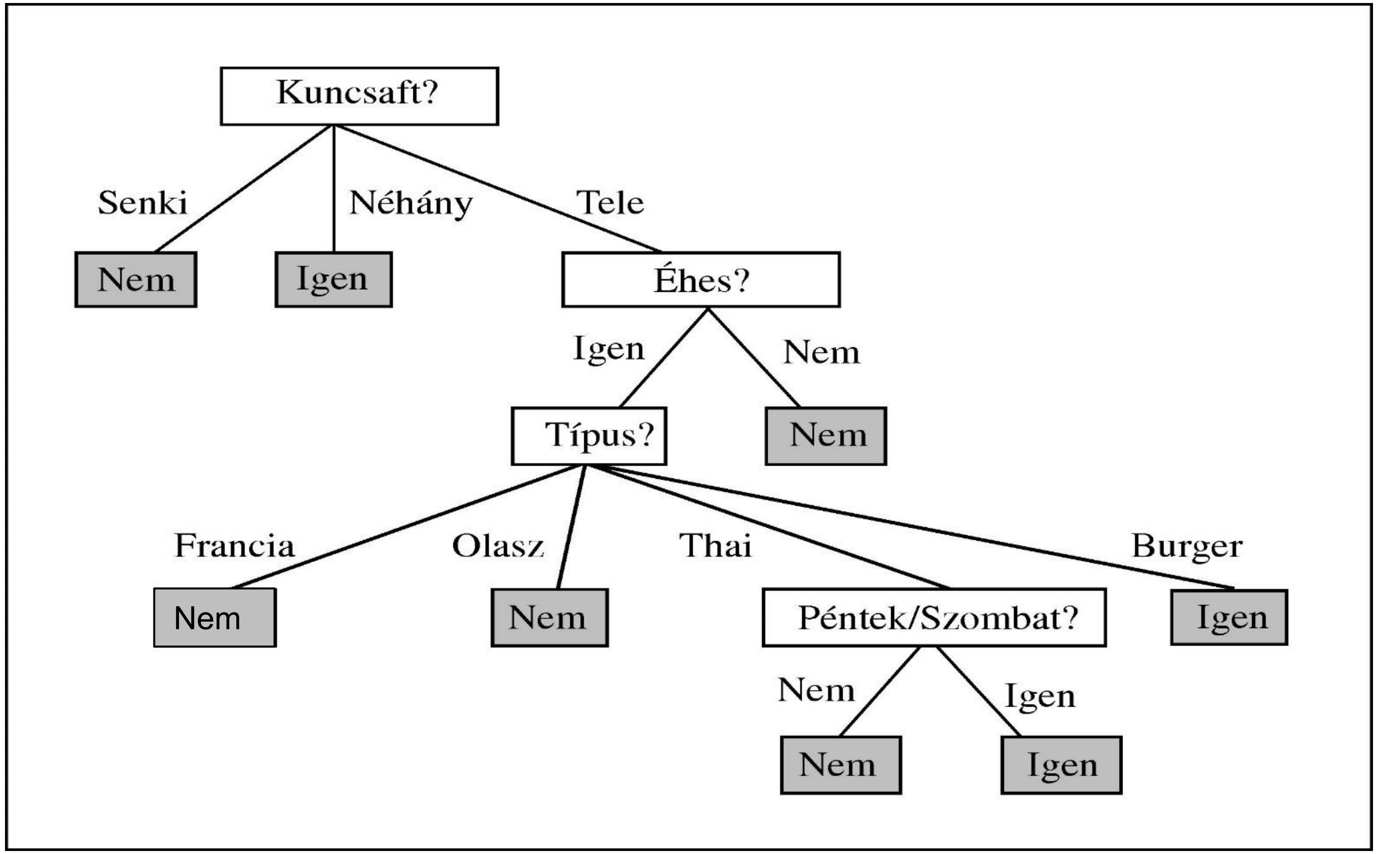
Mi benne a viselkedésünk mintája? Van-e egyáltalán egy konzisztens - minden tanítópéldának megfelelő – viselkedési mintánk?
(9216 lehetséges példa van, 12 alapján tanulunk)

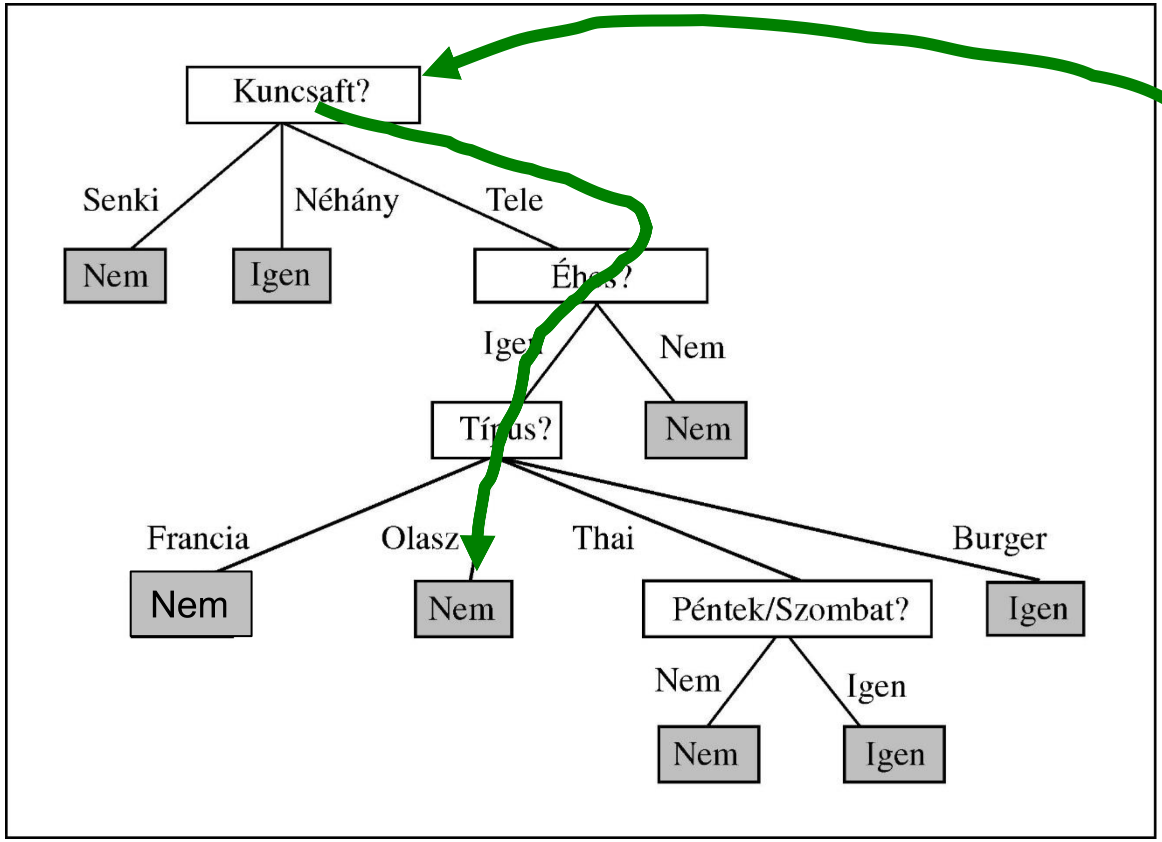
Döntési fák tanulása

döntési fa = egy logikai függvény
bemenete: egy tulajdonsághalmazzal leírt objektum vagy szituáció

kimenete: egy igen/nem „döntés/cselekvés"

belső csomópont = valamelyik tulajdonság tesztje
él = a teszt lehetséges értéke
levél = logikai érték, amelyet akkor kell kiadni, ha ezt a levelet elértük





Egy konkrét eset:

Kuncsaft = Tele és
Éhes és
Típus = Olasz

Várni_Fog = (Kuncsaft = Néhány) V
((Kuncsaft = Tele) ∧ Éhes ∧ (Típus = Thai) ∧ Pént/Szomb) V
((Kuncsaft = Tele) ∧ Éhes ∧ (Típus = Burger))

Döntési fák kifejezőképessége

teljes - **minden (ítélet)logikai függvény felírható döntési faként**.
(az igazságtábla minden sora = a fa egy bejárása,
de így az igazságtábla mérete = a fa mérete = exponenciális!)

Döntési fák kialakítása példák alapján

példa: (attribútumok értékei, célpredikátum értéke)

példa **besorolása**: a célpredikátum értéke

pozitív/negatív példa: a célpredikátum értéke igaz / hamis

tanító halmaz: a teljes példahalmaz

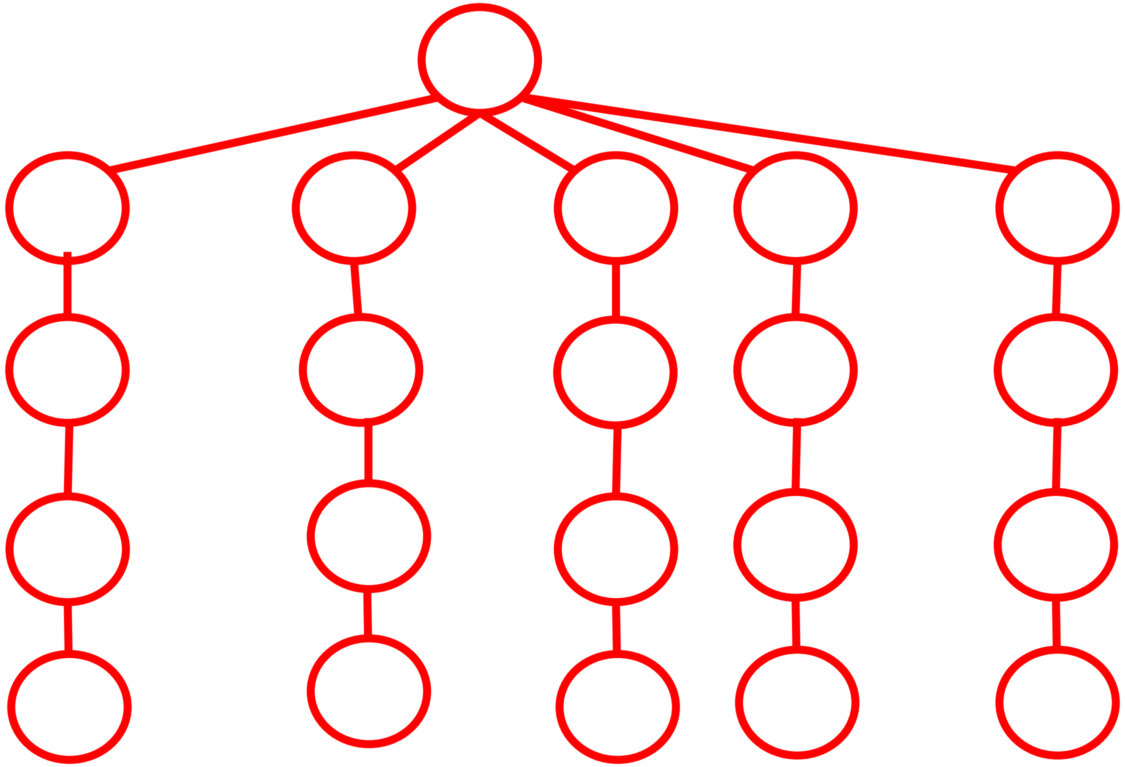
triviális fa, könnyű - mindegyik példához egy önálló bejárási út

a levél a példa besorolását adja.

a fa egyszerűen memorizálja a példát, nem alakít ki jellegzetes mintákat

a példákból **nem általánosít** (nem tud ismeretlen példákra extrapolálni)

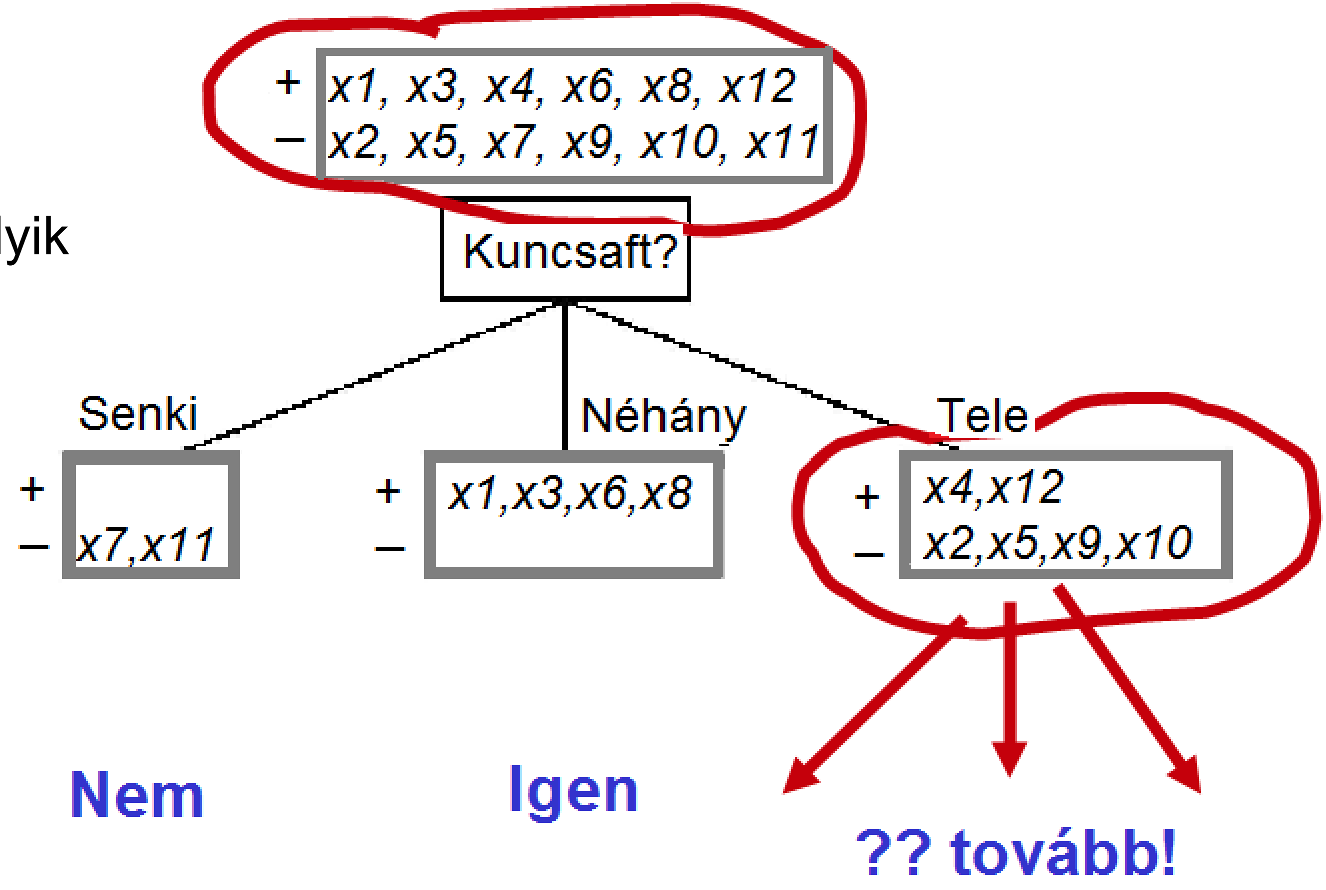
„**igazi**” fa - a jellegzetes minták kinyerése, a módszer képes nagyszámú esetet **tömör** formában leírni, és így az egyelőre ismeretlen példákra is **általánosítani**



A döntési fa építése – általános jelenségek

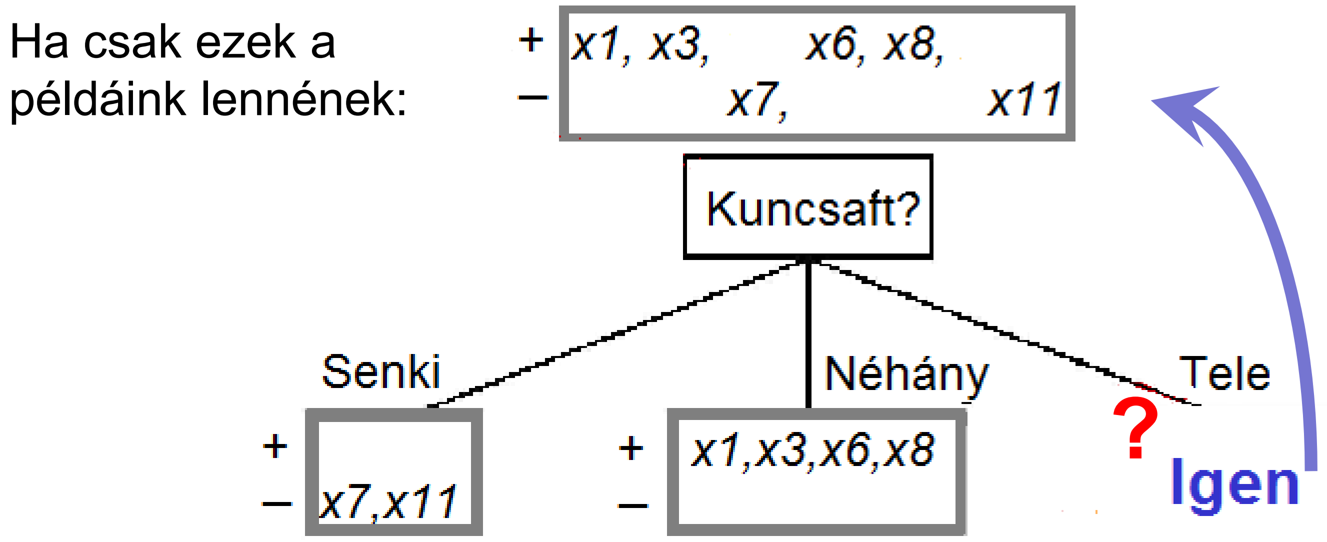
Van néhány pozitív és néhány negatív példánk.

Heurisztika: válasszuk azt az attribútumot, amelyik a legjobban szétválasztja őket. (mohó!)



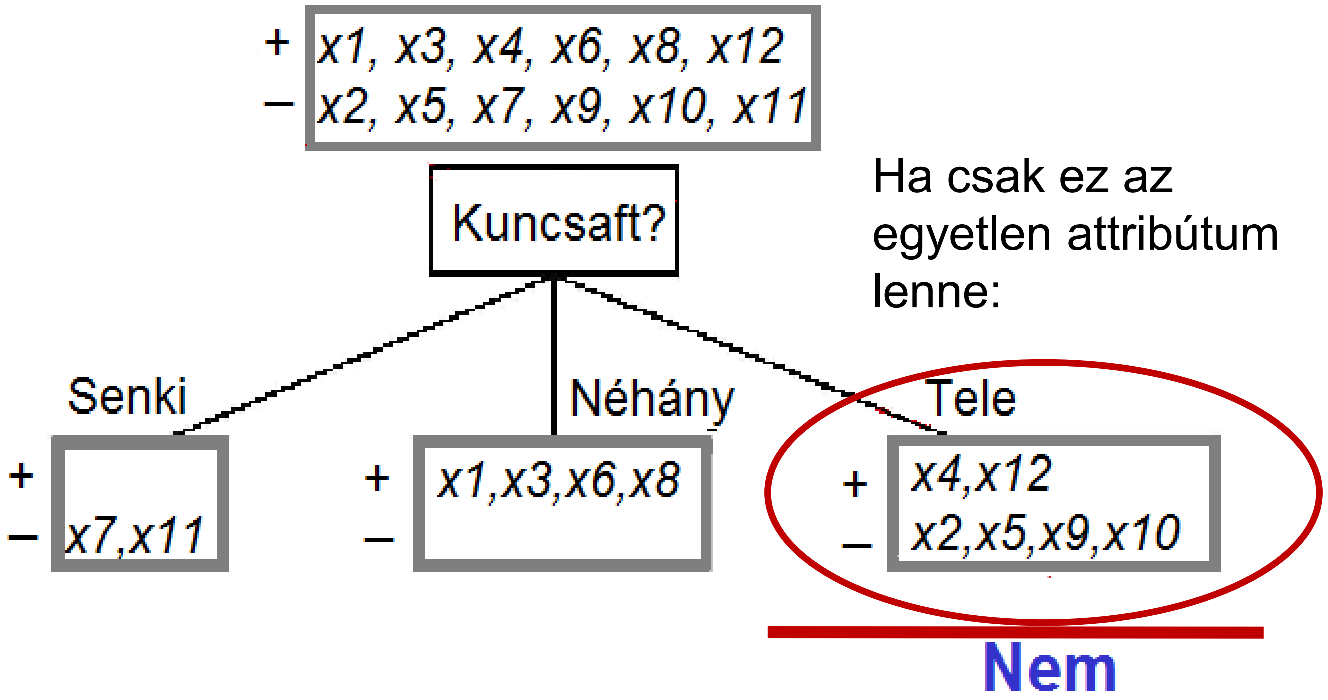
Ha az összes megmaradt eset pozitív, (vagy az összes negatív), akkor készen vagyunk azzal az ággal: a válasz **Igen** vagy **Nem**.

A döntési fa építése – általános jelenségek



Ha nem maradt egyetlen példa sem, ez azt jelenti, hogy ilyen példát nem figyeltünk meg eddig…., de a jövőben mégis jelentkezhet. Ilyenkor valamilyen alapértéket adunk vissza, amelyet rendszerint a csomópont szülőjének többségi válaszából származtatunk.

A döntési fa építése – általános jelenségek



Nem maradt már teszteletlen attribútum, de maradtak még pozitív és negatív példák. Baj van.

Ezeknek a példáknak pontosan megegyezik a leírása, de különböző a besorolásuk.

Néhány adat tehát nem korrekt: a **zaj** torzítja az adatokat – vagy hiányzik még attribútum (infó). **Megoldás?** Pl. a **többségi szavazás** használata.

Döntési fák kialakítása példák alapján

példa: (attribútumok értékei, célpredikátum értéke)
példa **besorolása**: a célpredikátum értéke
pozitív/negatív példa: a célpredikátum értéke igaz / hamis
tanító halmaz: a teljes példahalmaz

A *legkisebb döntési fa* megtalálása - általánosságban nem megoldható

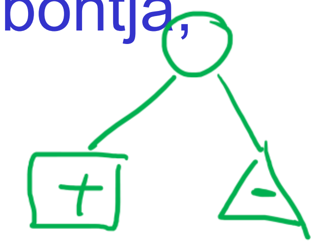
Heurisztika: **mohóság** - egy meglehetősen egyszerű (jó) fa is jó lesz!
Az alapötlet: először a „legfontosabb" attribútumot teszteljük.

„**legfontosabb**" = a **legnagyobb eltérést** okozza példák besorolásában

Elvárás: kisszámú példa alapján korrekt besoroláshoz jussunk: a **bejárási utak rövidek** legyenek, és így az **egész fa kicsi** (tömör) legyen.

Döntési fák kialakítása példák alapján

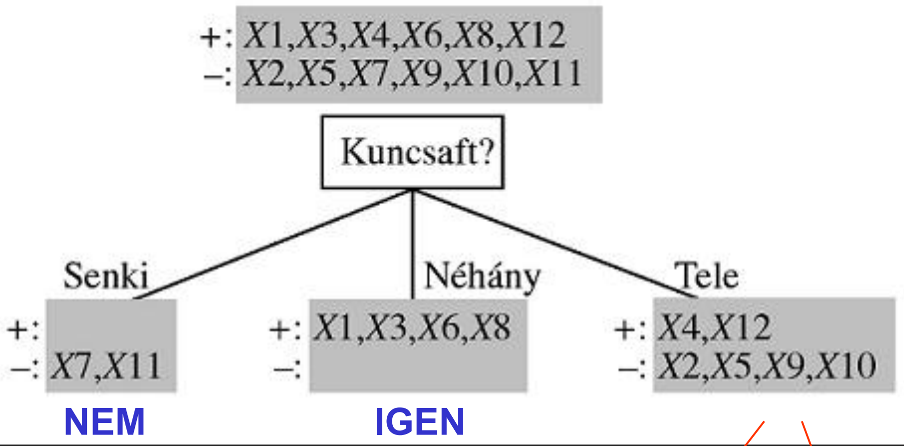
A **tökéletes attribútum** a példákat két csoportra bontja,
az egyikbe **csak pozitív**,
a másikba **csak negatív** példák tartoznak.



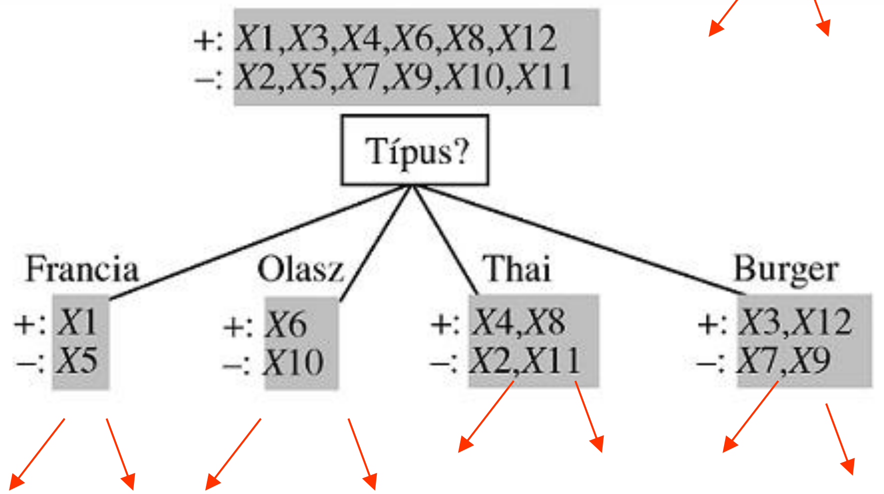
Ezzel be is lehetne fejezni a fa tanítását – de tökéletes attribútum általában nincs!

Olyan attribútumot válasszunk, amely a lehető legmesszebbre jut a pontos besorolás biztosításában.

A *Kuncsaft* attribútum nem tökéletes, de elég jó.
Ha a Kuncsaft tesztelése után döntünk, várhatóan legfeljebb 2-t hibázunk a 12 példából.



A *Típus* attribútum ehhez képest? (vacak…)
Ha a Típus tesztelése után döntünk, várhatóan 6-ot hibázunk a 12 példából. Éppúgy, mint a teszt előtt!



Egy nagymértékben haszontalan attribútum, mint pl. a *Típus* olyan példahalmazokat hoz létre, amelyek nagyjából ugyanolyan arányban tartalmaznak pozitív és negatív példákat, mint az eredeti halmaz.

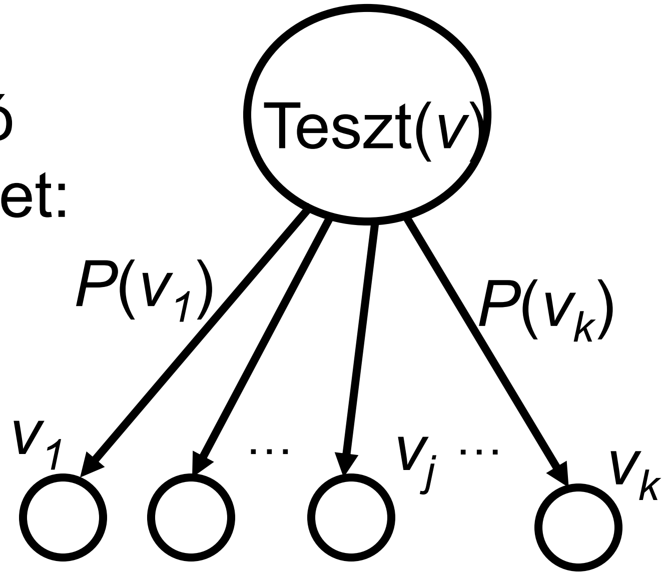
„Elég jó?" „Nagymértékben haszontalan?„

A mérték legyen **maximális**: a **tökéletes** attribútumra,
minimális: olyan attribútumra, aminek egyáltalán **nincs értéke** számunkra.

Egy megfelelő mérték: egy helyzetben az információszükségletünk a
helyes megoldáshoz, az attribútum által adott információ várható értéke, információ átlagos tartalma, entrópia, …

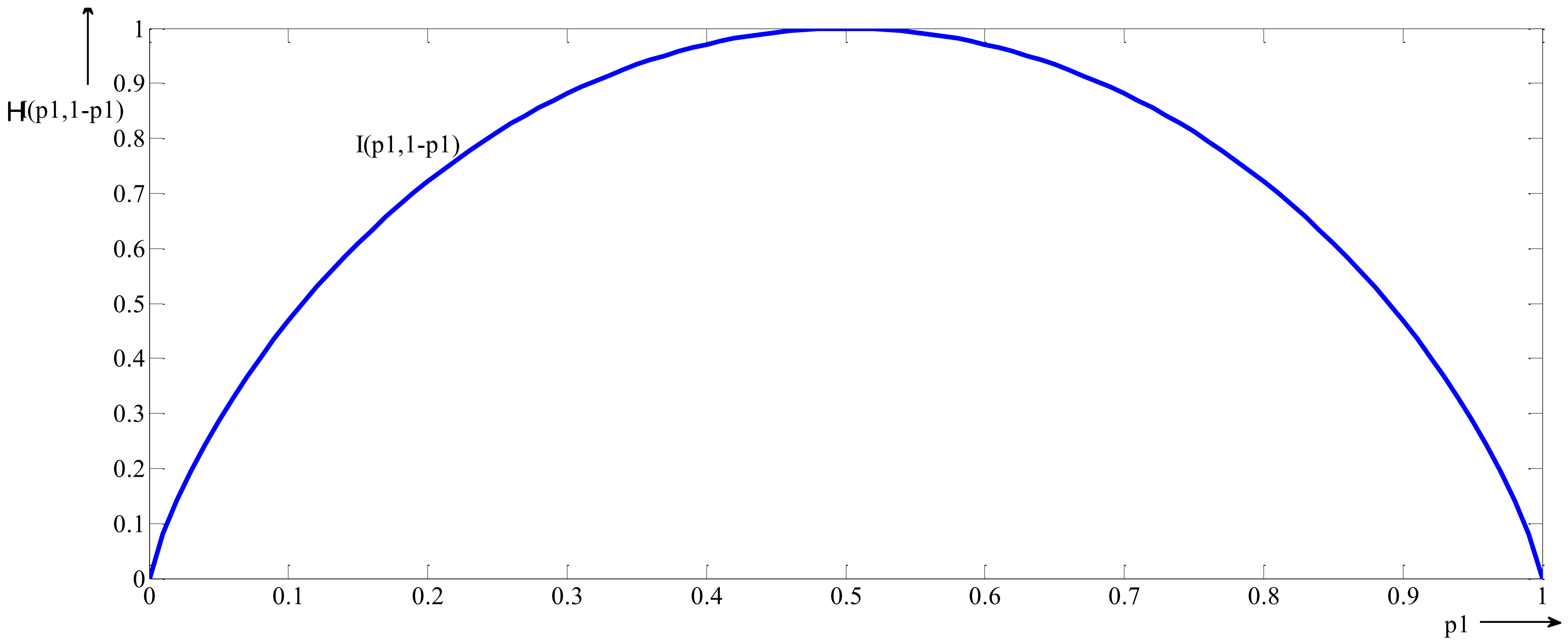
Ha a lehetséges v_j válaszok valószínűsége $P(v_j)$,
akkor az adott kiinduló halmaz esetén a jó döntéshez szükséges információszükséglet:

$$H(P(v_1), P(v_2), \ldots, P(v_k)) = \sum_{j=1}^{k} -P(v_j) \cdot \log_2(P(v_j))$$



pl. Fogadunk 1000 Ft-ba egy szabályos pénzérme dobása esetén, hogy „fej” jön ki.

- Mennyit ér nekünk az infó, ha valaki elárulja, hogy melyik fog kijönni? Mennyi infóra van szükségünk?
- Mennyit ér az infó, ha tudjuk, hogy az érme aszimmetrikus, és 80%-ban „fej” jön ki?
- Mennyit ér nekünk az infó, ha az érme mindkét oldalán „fej” van, tehát 100%-ban ez jön ki?



A döntési fa információtartalma - a tanítóhalmazban található pozitív és negatív példák aránya alapján

A tanítóhalmaz ***p*** pozitív és ***n*** negatív példát tartalmaz. A két válasz: v_1 (pozitív), v_2 (negatív), és a valószínűségük becslése:

$$P(v_1) \cong \frac{p}{p+n}, \qquad P(v_2) \cong \frac{n}{p+n}$$

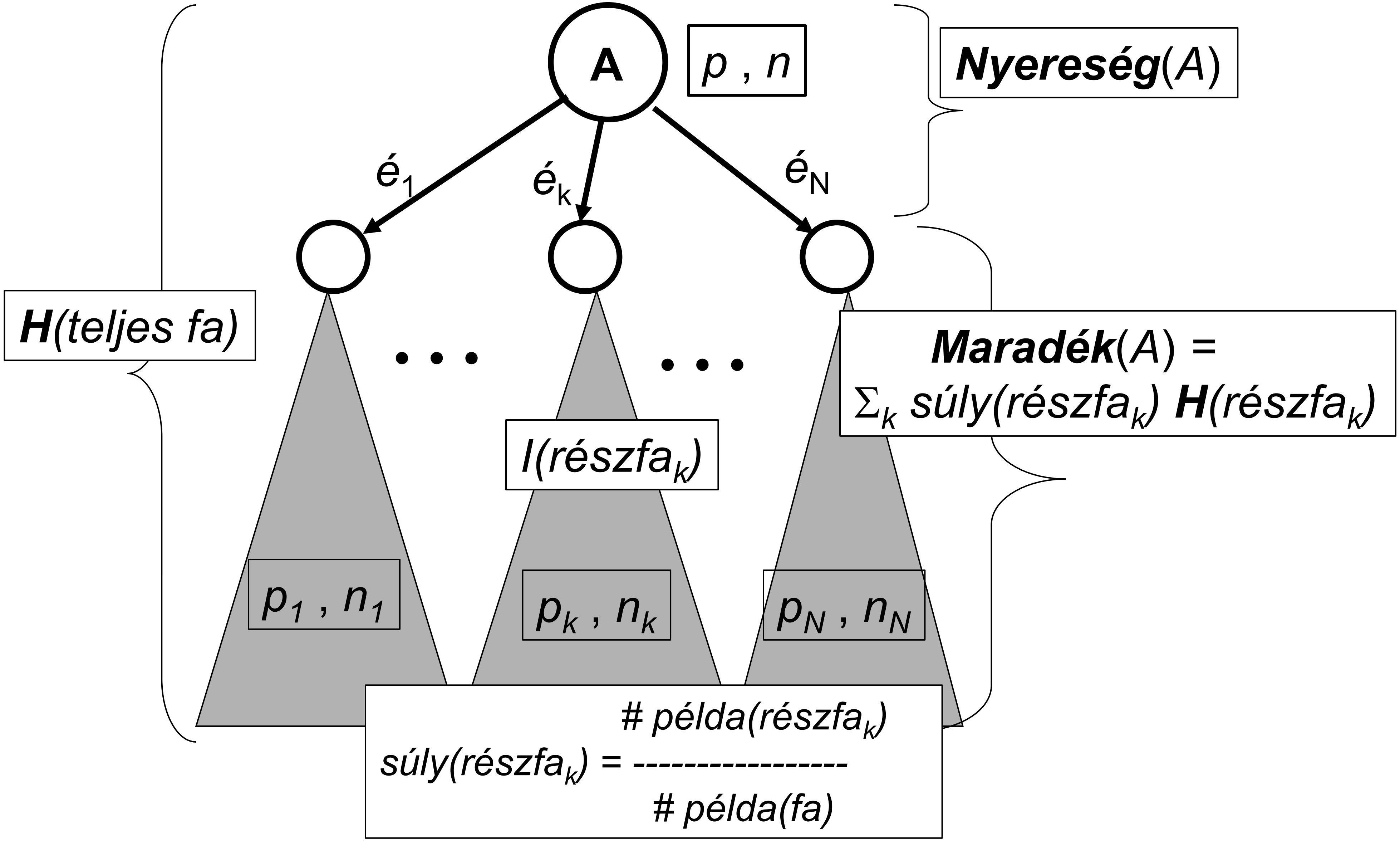
Ekkor a **fa információtartalmának** becslése:

$$H(P(v_1), P(v_2)) \cong H\left(\frac{p}{p+n}, \frac{n}{p+n}\right) = -\frac{p}{p+n} \cdot \log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n} \cdot \log_2\left(\frac{n}{p+n}\right)$$

Hogyan válasszuk ki, hogy melyik attribútumot teszteljük?
Mennyi új információt ad nekünk egy attribútum tesztelése?

Mennyi információra volt szükségünk az attribútum tesztelése **előtt**, és mennyire van **még** szükségünk az attribútum tesztelése **után**? A kettő különbségét nyertük meg a teszttel!

Bármelyik ***A*** attribútum az ***E*** tanító halmazt $\boldsymbol{E}_1, \boldsymbol{E}_2, \ldots, \boldsymbol{E}_N$ részhalmazokra bontja az ***A*** tesztjére adott válaszoknak megfelelően, ha ***A*** tesztje *N* különböző választ ad.



Hogyan válasszunk attribútumot?

$$Maradék(A) = \sum_{k=1}^{N} \frac{p_k + n_k}{p + n} \cdot H\left(\frac{p_k}{p_k + n_k}, \frac{n_k}{p_k + n_k}\right)$$

Az attribútum tesztjének **információnyeresége** az **eredeti** információigény és a **teszt utáni** új információigény **különbsége**:

$$Nyereség(A) = H\left(\frac{p}{p + n}, \frac{n}{p + n}\right) - Maradék(A)$$

Válasszuk a pillanatnyilag legnagyobb nyereséget adó attribútumot!
(**Mohó eljárás**)

Közbevetés: mi az információszükséglet, ha 100%-os az egyik (és 0%-os a másik) eset?

$$H(0,1) = ?$$

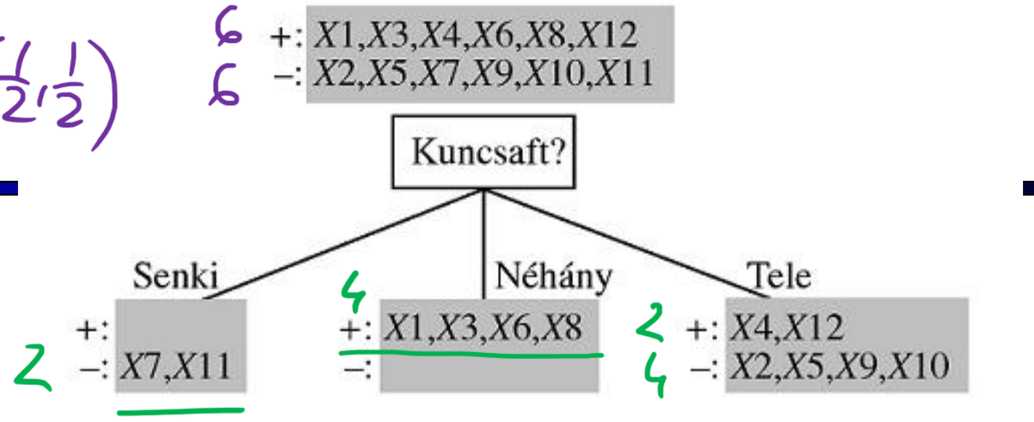
$\log_2()$ helyett használhatunk ln()-t, hiszen csak egy konstanssal való szorzásban különböznek: $\log_2(x) = \log_2(e).\ln(x)$

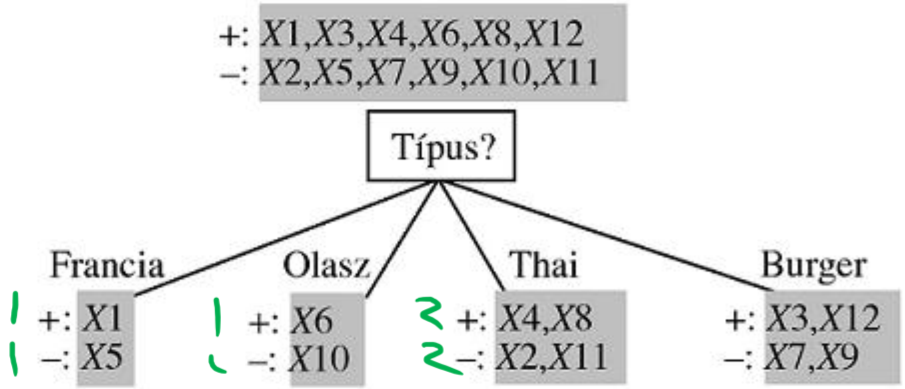
$$0 \cdot \ln(0) = \lim_{x \to 0} \frac{\ln(x)}{1/x}$$

helyette a deriváltak hányadosa vizsgálható (l′Hopital szabály)

$$0 \cdot \ln(0) = \lim_{x \to 0} \frac{\ln(x)}{1/x} = \lim_{x \to 0} \frac{1/x}{-1/x^2} = \lim_{x \to 0} (-x) = 0$$

Nézzük meg a *Kuncsaft* és a *Típus* attribútumokat $I(\frac{1}{2},\frac{1}{2})$





$$Nyereség(Kuncsaft) = \boxed{1} - \left[\boxed{\frac{2}{12}} \cdot \underbrace{H(0,1)}_{} + \frac{4}{12} \cdot H(0,1) + \boxed{\frac{6}{12}} \cdot H\left(\frac{2}{4},\frac{4}{4}\right)\right] \approx 0{,}54 \text{ bit}$$

$+ \; -$ $\quad I(1,0)$ $\quad \frac{2}{6}, \frac{4}{6}$

$1 - 1 = \emptyset$

$$Nyereség(Típus) = 1 - \left[\frac{2}{12} \cdot \underbrace{H\left(\frac{1}{2},\frac{1}{2}\right)}_{1} + \frac{2}{12} \cdot \underbrace{H\left(\frac{1}{2},\frac{1}{2}\right)}_{1} + \frac{4}{12} \cdot \underbrace{H\left(\frac{2}{4},\frac{2}{4}\right)}_{1} + \frac{4}{12} \cdot \underbrace{H\left(\frac{2}{4},\frac{2}{4}\right)}_{1}\right] = 0 \text{ bit}$$

Több attribútumra is kellene a számítás, de a Kuncsaft kimagaslóan jó.

(**Kuncsaft = Tele**) ágon még fennmaradó példák

Példák	**Attribútumok**									**Cél**
	Altern	*Bár*	*Pént*	*Éhes*	*Ár*	*Eső*	*Fogl*	*Típus*	*Becs*	***VárniFog***

−

X_2	*Igen*	*Nem*	*Nem*	*Igen*	*Olcsó*	*Nem*	*Nem*	*Thai*	*30–60*	***Nem***
X_5	*Igen*	*Nem*	*Igen*	*Nem*	*Drága*	*Nem*	*Igen*	*Francia*	*>60*	***Nem***
X_9	*Nem*	*Igen*	*Igen*	*Nem*	*Olcsó*	*Igen*	*Nem*	*Burger*	*>60*	***Nem***
X_{10}	*Igen*	*Igen*	*Igen*	*Igen*	*Drága*	*Nem*	*Igen*	*Olasz*	*10–30*	***Nem***

+

X_4	*Igen*	*Nem*	*Igen*	*Igen*	*Olcsó*	*Nem*	*Nem*	*Thai*	*10–30*	***Igen***
X_{12}	*Igen*	*Igen*	*Igen*	*Igen*	*Olcsó*	*Nem*	*Nem*	*Burger*	*30–60*	***Igen***

Részfában: $p_1 = 2$, $n_1 = 4$, $p_1/(p_1+n_1) = 1/3$, $n_1/(p_1+n_1) = 2/3$

(**Kuncsaft = Tele**) ág

H = H(részfa) = H(2/6, 4/6) = 0.9183 (2 pozitív és 4 negatív példa)

Ha pl. **Alternatíva**-t teszteljük, akkor a nyereség:

$$Ny(Altern) = H - [\underbrace{5/6\ I(2/5, 3/5)}_{\text{Altern = Igen}} + \underbrace{1/6\ I(0, 1)}_{\text{Altern = Nem}}]$$

Altern = Igen	Altern = Nem
$p2 = 2$, $n2 = 3$	$p3 = 0$, $n3 = 1$

6 példa

$$Ny(Altern) = H - [5/6\ H(2/5, 3/5) + \cancel{1/6\ H}(0, 1)]$$

$$= 0.92 - 0.81 \approx .11$$

(**Kuncsaft = Tele**) ág

Ny(Altern) = H – [5/6 H(2/5, 3/5) + 1/6 H(0,1)] = .92 - .81 ≈ .11

Ny(Bár) = H – [1/2 H(1/3, 2/3) + 1/2 H(1/3, 2/3)] = .92 - .92 = 0

Ny(Péntek) = H – [5/6 H(2/5, 3/5) + 1/6 H(0,1)] = .92 - .81 ≈ 0.11

Ny(Éhes) = H – [4/6 H(1/2, 1/2) + 2/6 H(0,1)] = .92 - .66 ≈ 0.25 ***Válasszuk pl. ezt!***

Ny(Ár) = H – [4/6 H(1/2, 1/2) + 2/6 H(0,1)] = .92 - .66 ≈ 0.25

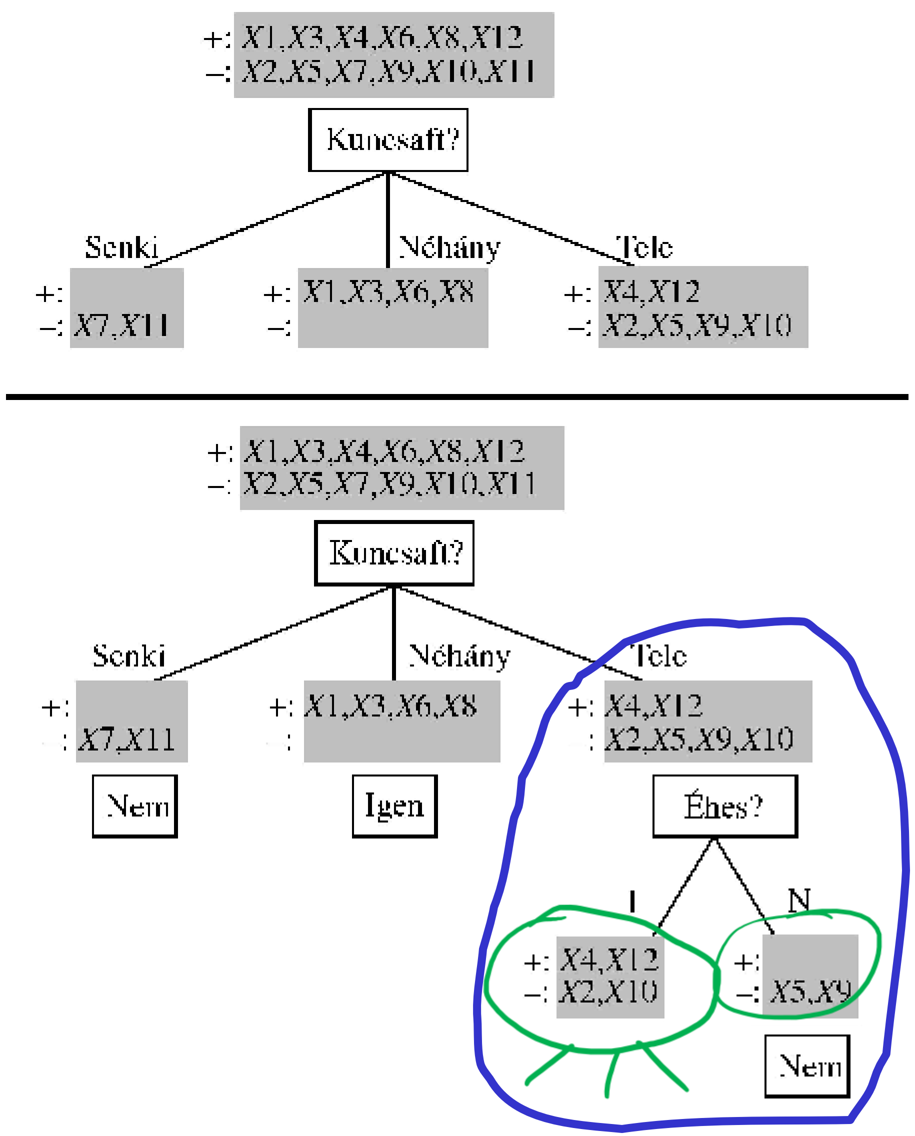
Ny(Eső) = H – [5/6 H(2/5, 3/5) + 1/6 H(0,1)] = .92 - .81 ≈ 0.11

Ny(Foglalt) = H – [2/6 H(0,1) + 4/6 H(1/2,1/2)] = .92 - .66 ≈ 0.25

Ny(Típus) = H –
[2/6 H(1/2, 1/2) + 1/6 H(0,1) + 2/6 H(1/2, 1/2) + 1/6 H(0,1)] = .92 - .66 ≈ 0.25

Ny(Becs) = H –
[2/6 H(1/2, 1/2) + 2/6 H(0,1) + 2/6 H(1/2, 1/2)] = .92 - .66 ≈ 0.25

(**Kuncsaft = Tele**) ÉS **Teszt(Éhes?)**



(**Kuncsaft = Tele**) ÉS **(Éhes?=Igen)** ágon még fennmaradó példák

Példa	**Attribútumok**								**Cél**
	Alt	*Bár*	*Pént*	*Ár*	*Eső*	*Fogl*	*Típus*	*Becs*	***VárniFog***
X_2	*Igen*	*Nem*	*Nem*	*Olcsó*	*Nem*	*Nem*	*Thai*	*30–60*	***Nem***
X_{10}	*Igen*	*Igen*	*Igen*	*Drága*	*Nem*	*Igen*	*Olasz*	*10–30*	***Nem***
X_4	*Igen*	*Nem*	*Igen*	*Olcsó*	*Nem*	*Nem*	*Thai*	*10–30*	***Igen***
X_{12}	*Igen*	*Igen*	*Igen*	*Olcsó*	*Nem*	*Nem*	*Burger*	*30–60*	***Igen***

Részfában: $p_1 = 2$, $n_1 = 2$, $p_1/(p_1+n_1) = ½$, $n_1/(p_1+n_1) = ½$

Információszükséglet a következő teszt előtt ezen az ágon: 1 bit!

Különböző attribútumok nyeresége a (**Kuncsaft = Tele**) ÉS **(Éhes?=Igen)**

Ny(Alt) = H – [1/1 H(1/2, 1/2) + 0] = 1 - 1 = 0

Ny(Bár) = H – [1/2 H(1/2, 1/2) + 1/2 H(1/2, 1/2)] = 1 - 1 = 0

Ny(Péntek) = H – [1/4 H(0,1) + 3/4 H(1/3,2/3)] = 1 - .92 ≈ .08

Ny(Ár) = H – [1/4 H(0,1) + 3/4 H(1/3,2/3)] = 1 - .92 ≈ .08

Ny(Eső) = H – [1/1 H(1/2, 1/2) + 0] = 1 - 1 = 0

Ny(Foglalt) = H – [1/4 H(0,1) + 3/4 H(1/3,2/3)] = 1 - .92 ≈ .08

Ny(Típus) = H – [1/4 H(0,1) + 1/4 H(0,1) + 1/2 H(1/2, 1/2)] = .5 ***Válasszuk pl. ezt!***

Ny(Becs) = H – [1/2 H(1/2, 1/2) + 1/2 H(1/2, 1/2)] = 1 - 1 = 0

(**Kuncsaft = Tele**) ÉS **(Éhes?=Igen) ÉS** (**Típus = Thai**)
ágon még fennmaradó példák

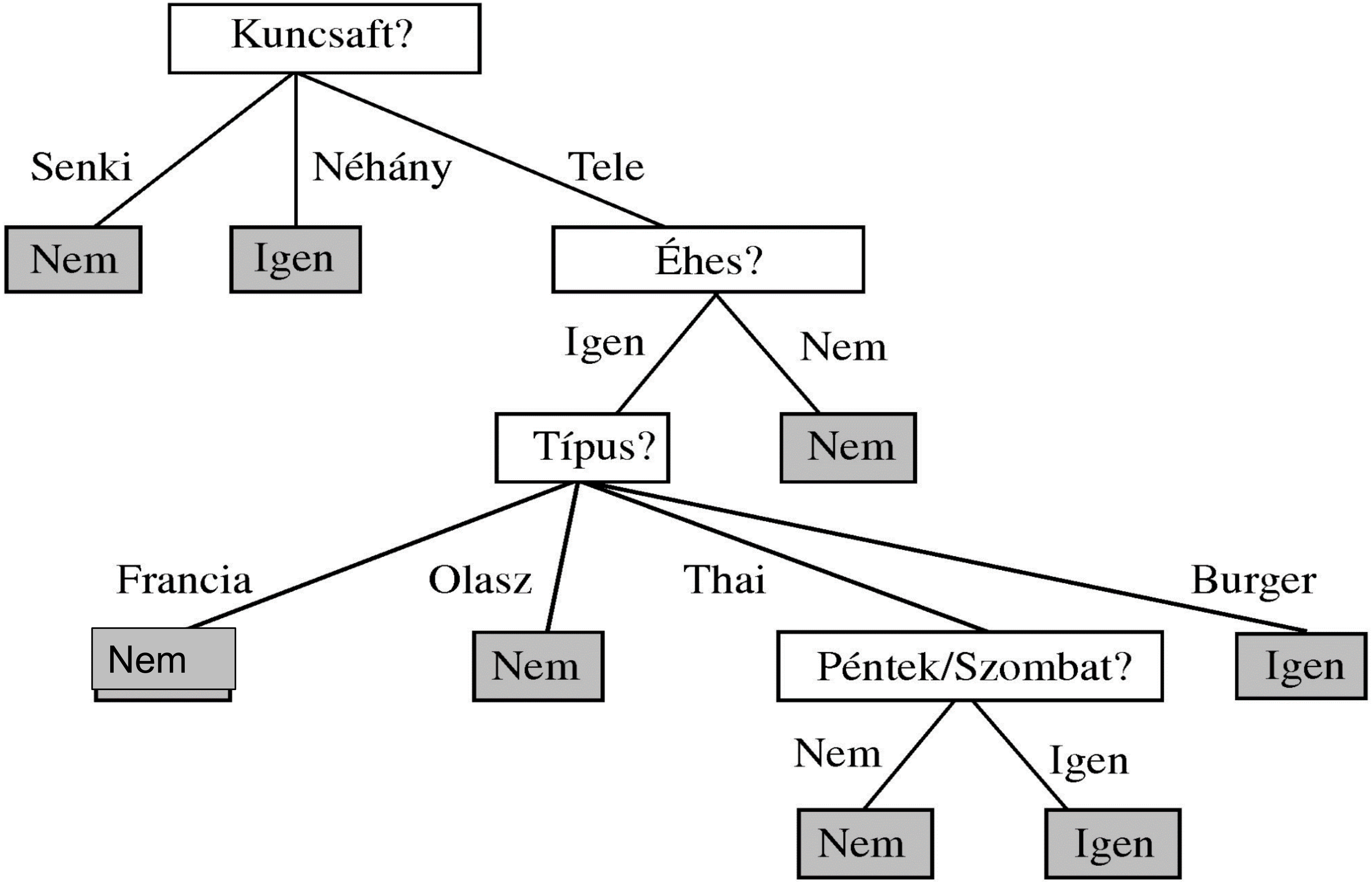
Példa			**Attribútumok**					**Cél**
	Alt	*Bár*	*Pént/Szo*	*Ár*	*Eső*	*Fogl*	*Becs*	***VárniFog***
$\boldsymbol{X_2}$	*Igen*	*Nem*	*Nem*	*Olcsó*	*Nem*	*Nem*	*30–60*	***Nem***
$\boldsymbol{X_4}$	*Igen*	*Nem*	*Igen*	*Olcsó*	*Nem*	*Nem*	*10–30*	***Igen***

A többi nem jó jellemző

Részfában: $p_1 = 1$, $n_1 = 1$, $p_1/(p_1+n_1) = 1/2$, $n_1/(p_1+n_1) = 1/2$

És ez ugyanúgy marad pl. a *Bár* = *Nem*, vagy *Eső* = *Nem*, vagy *Ár* = *Olcsó*, stb. mentén (azaz ezekben az irányokban nem érdemes építeni a fát).

A kialakított döntési fa



Ha pl. a Pént/Szomb helyett a Becs attribútumot választjuk – ugyanúgy **konzisztens** a kialakuló fa mindegyik példánkkal, **DE máshogy általánosít**!

Mikor álljunk le a döntési fa növesztésével?

Alapvetően 2 stratégia:

1. ***Korai leállás***

(A levelekkel mérhetjük, hogy mekkorára nőtt a fa, mennyire komplex.)

Döntési fa nyesése (*pruning*)

(1) Beállított max. mélység (mint a mélységkorlátozott keresésnél), ha elérjük, akkor ezen a szinten mindenképpen leállunk, nem fejtjük ki tovább az adott ágat.

(2) Vizsgáljuk meg az elvégzett teszteket, ismerjük fel a nem releváns (semmitmondó) felhasznált attribútumot, és az adott ágat ott fejezzük be (metsszük vissza a fát).

Döntési fa nyesése

2. ***Engedjük túlnőni, majd visszavágjuk, visszanyessük***

   Első lehetőség

   Ismerjük fel ha nem releváns attribútumot használtunk, és az adott ágat ne fejtsük ki tovább, sőt messük le a kialakított részfát.

   Irreleváns attribútum: példahalmaz kettévágásánál a kapott részhalmazok kb. azonos arányban tartalmaznak pozitív és negatív példát, mint az eredeti halmaz.
   Az információ nyereség ilyenkor közel 0.

   Az információ nyereség hiánya az irrelevancia jó jelzése.
   Milyen nagy legyen az információ nyereség, hogy egy attribútumot mégis felhasználjunk a példahalmaz megosztására?

   Statisztikai tesztet végzünk: **feltesszük, hogy nincs jellegzetes minta** = ún. **nullhipotézis**.

Statisztikai teszt – fordítva gondolkodunk, feltesszük, hogy nincs a halmazban jellegzetes minta = ún. **nullhipotézis**. Utána ezt megpróbáljuk cáfolni –vagy legalább belátni,hogy nagyon valószínűtlen!

Ha a nullhipotézis igaz, akkor elvileg 0 lenne az információnyereség, az ettől való eltérés csak a véletlen műve.

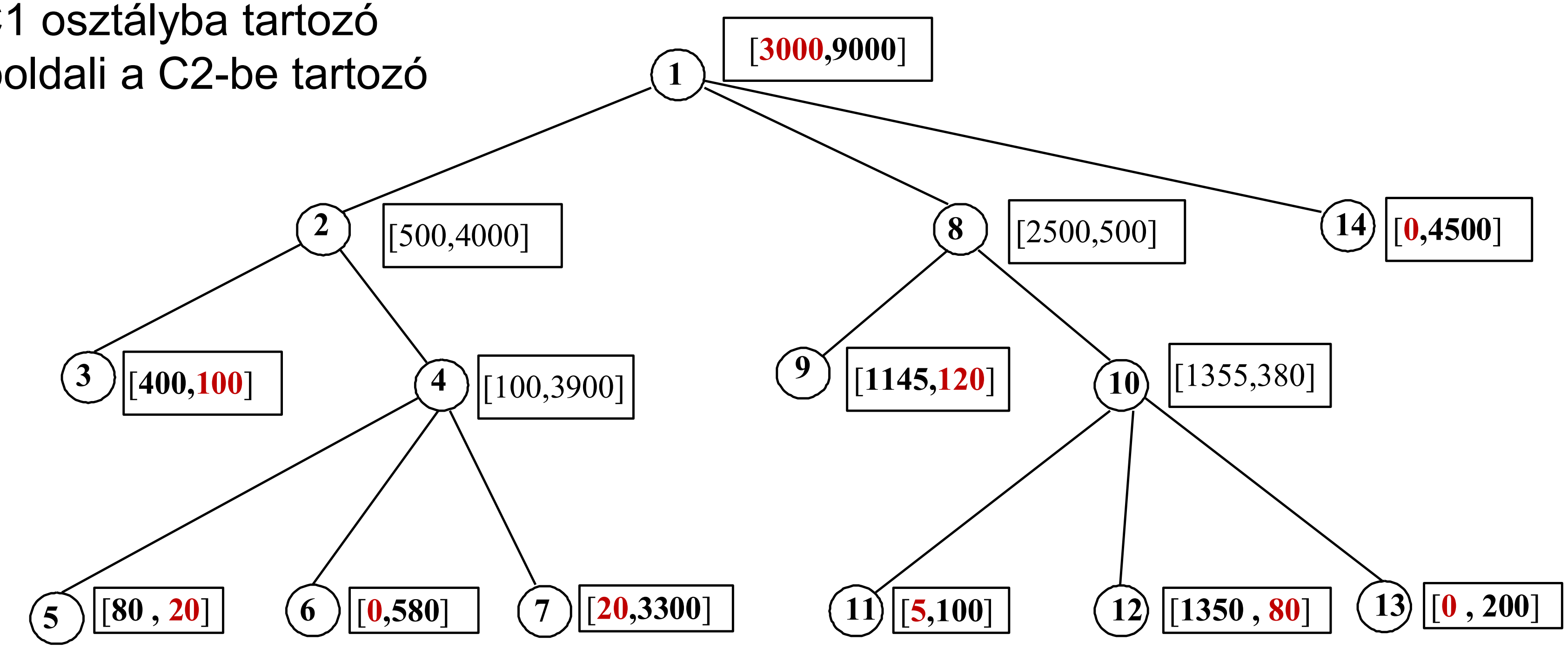
Ha az eltérés **nagy** (jóval nagyobb, mint 0) ⇒ **nem valószínű**, hogy az észlelt eltérés véletlen ingadozásból következik ⇒
a nullhipotézis nem valószínű: **vélhetőleg lényeges minta** van az adatokban.
(Valamilyen valószínűséggel cáfolni tudjuk a nulla hipotézist.)

Döntési fa nyesése: *második lehetőség*

Hibaarány – komplexitás kompromisszumon alapuló metszés

Egy kétosztályos (bináris) osztályozási példa kapcsán

(Baloldali szám a C1 osztályba tartozó minták száma, jobboldali a C2-be tartozó minták száma)



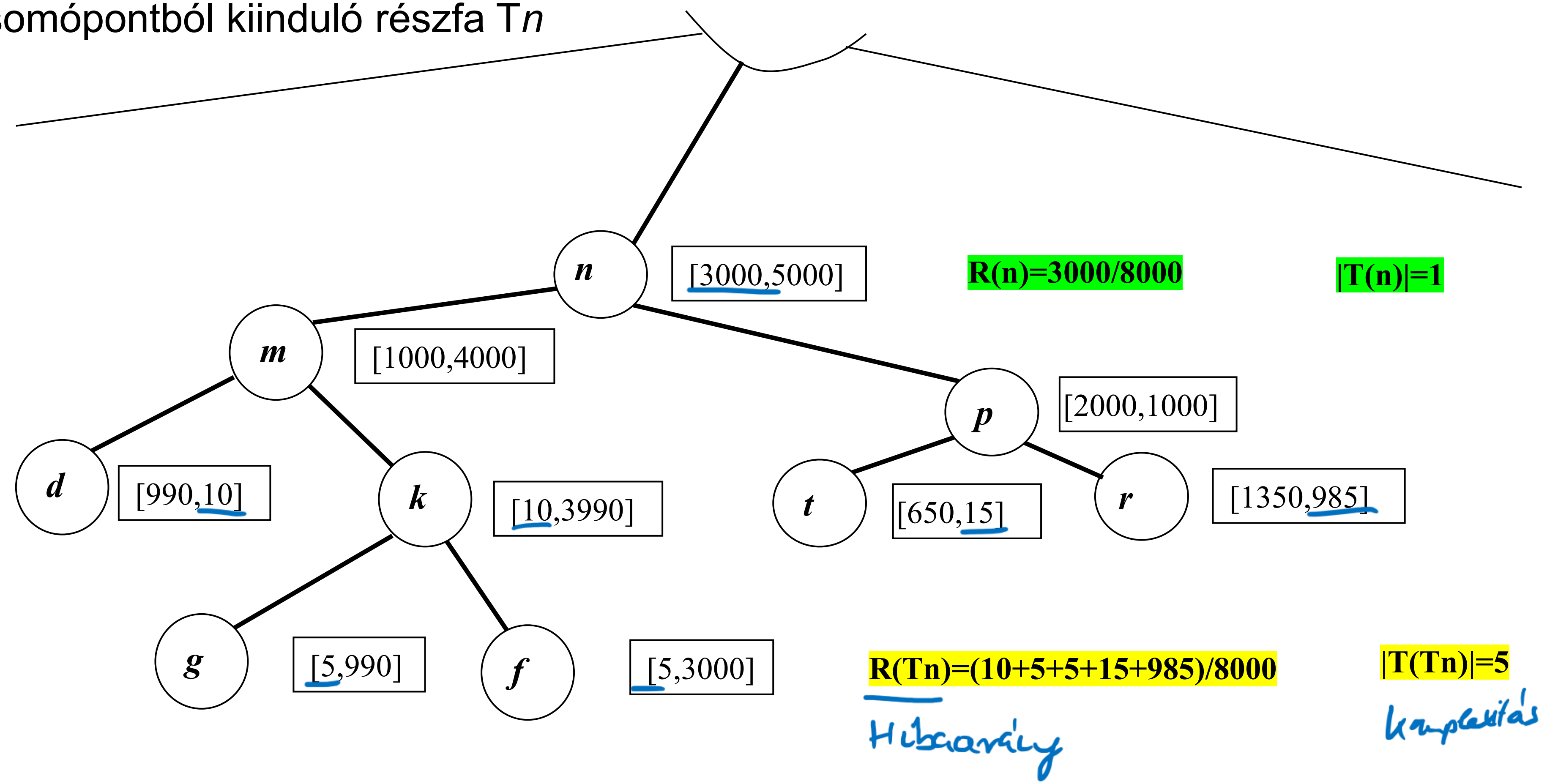
Komplexitás – legyen a levelek száma (lehetne más is), a példánkban:
a gyökérnél 1, a kifejtett fánál 9

Hibaarány:
a gyökérnél 3000/12000 (hiszen a jobb válaszunk C2 lenne),
a kifejtett fánál: (100+20+0+20+120+5+80+0+0)/12000

Hibaarány – komplexitás kompromisszumon alapuló metszés

Számozzuk a csomópontokat, jelölés:

- az *n*-dik csomópontot önmagában (mintha levél lenne) jelölje *n*
- az *n*-edik csomópontból kiinduló részfa T*n*



Hogyan hasonlítsuk össze a komplexitást a hibával?!?