

AROUND THE WORLD IN 80 DOGS

A RUFF GUIDE TO THE
DATABASE CHARACTER SET



AROUND THE WORLD IN 80 DOGS

A RUFF GUIDE TO THE
DATABASE CHARACTER SET









dog

A close-up photograph of a brown dog's face, looking slightly to the right. The dog has a dark brown coat with some lighter fur on its chest. In the background, a wooden table is visible with a large glass mug of beer in the foreground. The mug has a crest with a lion and the word "ENGLAND" on it. A small white dog is partially visible behind the main dog.

dog



- - . . - - - -



dog



• • •

To any blind people offended by that joke, you know where to find me!

7/79



dog



[64 6f 67]

→ 🐕 80dogs.txt

Hund	hondur	köpeği	tz' i'
Ha'DIBaH	hualp	mbula	ukudla
abwo	huan	mbwa	welpe
alabai	hund	mbwagh	xolo
anjing	imba	mbwá	zwin
ashun	imbua	mbwene	āso
aso	imbwe	mbu	āšun
cane	inu	njau	šuo
cão	it	njoka	אַבָּא
cain	jindo	pa	כלב
caine	kelb	perro	كلب
câini	khuy	pes	ଫୁତ୍ତା
chien	klèb	pies	ഇംഗ്ലീഷ്
chó	koer	qen	犬
cyn	koira	qeni	狗
dog	kotta	sag	戌
emba	køter	sagol	ଜୀବିତ
ghaddu	kutya	sobaka	ଜୀବିତ
gom	kutta	suns	ପାତା
hond	köpek	szczeniak	🐶



- » Before Unicode
- » Unicode
- » Oracle CharacterSet

0 1 2 3 4 5 6 7 8 9 a b c d e f

0x 1 2 3 4 5 6 7 8 9 0 # @ : > √

1x ſ / S T U V W X Y Z ≠ , % γ \ №

2x – J K L M N O P Q R ! # *] ; Δ

3x & A B C D E F G H I ? . □ [< □?

→ encode.py ascii < 80dogs.txt

✓ ascii: 53 good dogs

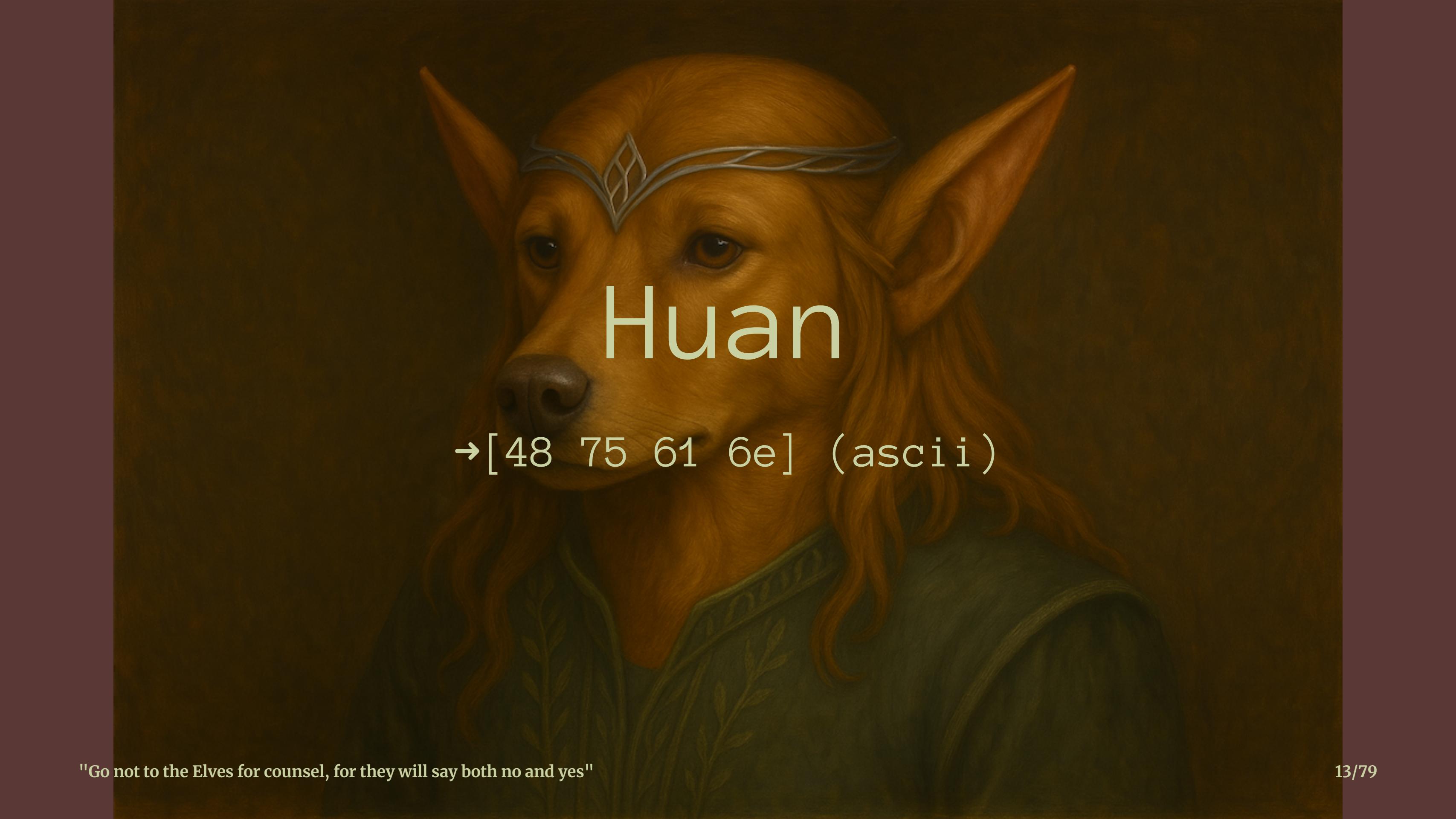
Hund Ha'DIBaH abwo alabai anjing ashun aso cane chien cyn dog emba
ghaddu gom hond hondur hualp huan hund imba imbua imbwe inu it
jindo kelb khuy koer koira kotta kutya kutta mbula mbwa mbwagh mbwene
njau njoka pa perro pes pies qen qeni sag sagol sobaka suns
szczeniak ukudla welpe xolo zwin

✗ ascii: 27 bad dogs

cão cain câine câini chó klèb køter köpek köpeğî mbwá mbu tz'í'
āso āšun šuo كلب كلب كُفَّا ՚ابو كلب 犬 狗 𠙴 ՚ééchqâ'í



✓ ascii: 240 chars encoded in 240 bytes, 1.00 bytes per char



Huan

→ [48 75 61 6e] (ascii)



Ha'Dibah

→ [48 61 27 44 49 62 61 48] (ascii)

ascii	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
0x	<nul>	<soh>	<stx>	<etx>	<eot>	<enq>	<ack>	<bel>	<bs>	<tab>	<lf>	<vt>	<ff>	<cr>	<so>	<si>
1x	<dle>	<dc1>	<dc2>	<dc3>	<dc4>	<nak>	<syn>	<etb>	<can>		<sub>	<esc>	<fs>	<gs>	<rs>	<us>
2x	!	"	#	\$	%	&	'	()	*	+	,	-	.	/	
3x	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{	^	^	}	~

"WOOF"→[01010111 01001111 01001111 01000110]

"woof"→[01110111 01101111 01101111 01100110]

"HAU"→[1001000 1000001 1010101]

"hau"→[1101000 1100001 1110101]

iso-8859

Standard	Alias	Region	Year
iso-8859-1	latin1	Western Europe	1987
iso-8859-2	latin2	Central & Eastern Europe	1987
iso-8859-3	latin3	South Europe	1988
iso-8859-4	latin4	North Europe	1988
iso-8859-5	cyrillic		1988
iso-8859-6	arabic		1987
iso-8859-7	greek		1987
iso-8859-8	hebrew		1988
iso-8859-9	latin5	Turkish	1989
iso-8859-10	latin6	Nordic	1992
iso-8859-11	thai		2001
iso-8859-13	latin7	Baltics	1998
iso-8859-14	latin8	Celtic	1998
iso-8859-15	latin9	Western Europe (Improved)	1999
iso-8859-16	latin10	South-Eastern Europe	2001

→ encode.py ascii latin1 < 80dogs.txt

✓ ascii ✓ latin1: 53 good dogs

Ha'DIBaH Hund abwo alabai anjing ashun aso cane chien cyn dog emba
ghaddu gom hond hondur hualp huan hund imba imbua imbwe inu it
jindo kelb khuy koer koira kotta kutta kutya mbula mbwa mbwagh mbwene
njau njoka pa perro pes pies qen qeni sag sagol sobaka suns
szczeniak ukudla welpe xolo zwin

✗ ascii ✓ latin1: 9 dogs were bad, now good

chó câin câine câini cão klèb köpek koter mbwá

✗ ascii ✗ latin1: 18 bad dogs

köpegi mbu tz'i' āso āšun l'ééchqá'í šuo كلب אבָן קֹטָן ຊາ
𠂇 犬 狗 犃 🐶

✓ ascii: 53 good dogs, 240 chars encoded in 240 bytes, 1.00 bytes per char

✓ latin1: 62 good dogs, 278 chars encoded in 278 bytes, 1.00 bytes per char



O meu cão não tem nariz. Como é que ele cheira? Terrível!

20/79



câine

→ [63 e2 69 6e 65] (latin1)

→ encode.py ascii latin2 < 80dogs.txt

✓ ascii ✓ latin2: 53 dogs were good, still good
Ha'DIBaH Hund abwo alabai anjing ashun aso cane chien cyn dog emba
ghaddu gom hond hondur hualp huan hund imba imbua imbwe inu it
jindo kelb khuy koer koira kotta kutta kutya mbula mbwa mbwagh mbwene
njau njoka pa perro pes pies qen qeni sag sagol sobaka suns
szczeniak ukudla welpe xolo zwin

✗ ascii ✓ latin2: 7 dogs were bad, now good
chó câin câine câini köpek mbwá šuo

✗ ascii ✗ latin2: 20 dogs were bad, still bad
cão klèb köpegi køter mbu tz'i' āso āšun žééchqá'í כלב אבוי
କୁନ୍ତା ଶଂ ମୁଥ ଦୁଇ ହଜାର ପାତା ହଜାର

✓ ascii: 53 good dogs, 240 chars encoded in 240 bytes, 1.00 bytes per char
✓ latin2: 60 good dogs, 269 chars encoded in 269 bytes, 1.00 bytes per char



köpek

→ [6b f6 70 65 6b] (latin2)



Šuo

→ [b9 75 6f] (latin2)

gb2312

→ encode.py ascii gb2312 < 80dogs.txt

✓ ascii ✓ gb2312: 53 good dogs

Ha'DIbaH Hund abwo alabai anjing ashun aso cane chien cyn dog emba
ghaddu gom hond hondur hualp huan hund imba imbua imbwe inu it
jindo kelb khuy koer koira kotta kutta kutya mbula mbwa mbwagh mbwene
njau njoka pa perro pes pies qen qeni sag sagol sobaka suns
szczeniak ukudla welpe xolo zwin

✗ ascii ✓ gb2312: 7 dogs were bad, now good

chó klèb mbwá tz'i' āso 犬 狗

✗ ascii ✗ gb2312: 20 bad dogs

câin câine câini cão köpek köpeği koter mbu āšun l'ééchäq'í šuo אבּו

🐶 କୁତ୍ତା ଶା ମନ୍ଦିର ମନ୍ଦିର କଲବ କଲବ

✓ ascii: 53 good dogs, 240 chars encoded in 240 bytes, 1.00 bytes per char

✓ gb2312: 60 good dogs, 261 chars encoded in 269 bytes, 1.03 bytes per char



狗

→ [b9 b7]





Hund

→ [48 75 6e 64] (gb2312)
→ [48 75 6e 64] (ascii)



→ [?? ?? ?? ??] (gb2312)

Who is this super hero? Sarge? No. Rosemary, the telephone operator? No. Henry, the mild, mannered janitor? Could be...



cane

→ [63 61 6e 65] (gb2312)

L'unica cosa meno italiana della pizza con l'ananas è un cappuccino dopo pranzo.

34/79



狗犬

→ [b9 b7 c8 ae] (gb2312)

?

→[. . b9 b7 . .] (gb2312)



狗

→[.. b9 b7 ..] (gb2312)



肮发

→ [b0 b9 b7 a2] (gb2312)

- » Before Unicode
- » **Unicode**
- » Oracle CharacterSet

Codepoint

$U+0000 \rightarrow U+10ff ff$

Plane		Name
0	U+0000 → U+ffff	Basic Multilingual Plane
1	U+10000 → U+1ffff	Supplementary Multilingual Plane
2	U+20000 → U+2ffff	Supplementary Ideographic Plane
3	U+30000 → U+3ffff	Tertiary Ideographic Plane
4 → 13	U+40000 → U+dffff	Reserved
14	U+e0000 → U+effff	Suppl. Special-Purpose Plane
15	U+f0000 → U+fffff	Private Use Area A
16	U+100000 → U+10ffff	Private Use Area B

Range	Block
U+0000→U+007f	ASCII
U+0080→U+00ff	Latin1
U+0100→U+024f	Latin Extended
U+0250→U+02ff	Phonetic Symbols
U+0300→U+05ff	Greek, Cyrillic, Hebrew
U+0600→U+0fff	Arabic & Indian
U+1000→U+17ff	S & SE Asian
U+1800→U+24ff	E Asian
U+2000→U+27ff	Symbols, Punctuation
U+2800→U+28ff	Braille, Shapes
U+e80→U+9fff	Chinese
U+ac00→U+d7af	Korean
U+d800→U+dbff	High Surrogates
U+dc00→U+dkff	Low Surrogates
U+e000→U+f8ff	Private Use

Unicode Encodings

» UTF-32

» UTF-16

» UTF-8

→ encode.py utf-32be < 80dogs.txt

✓ utf-32be: 80 good dogs

Ha'DIbaH Hund abwo alabai anjing ashun aso cane chien chó cyn câin
câine câini cão dog emba ghaddu gom hond hondur hualp huan hund
imba imbua imbwe inu it jindo kelb khuy klèb koer koira kotta
kutta kutya köpek köpeği køter mbula mbwa mbwagh mbwene mbwá mbu njau
njoka pa perro pes pies qen qeni sag sagol sobaka suns szczeniak
tz'i' ukudla welpe xolo zwin āso āšun žééchäq'í šuo كلب אבו
କୁଟ୍ଟା ଶ୍ଵର ମନୋ ଦୁଃ ପାତା ହାତା

✓ utf-32be: 80 good dogs, 338 chars encoded in 1352 bytes, 4.00 bytes per char

犬 = U+72ac

- [00 00 72 ac] (utf-32be)
- [ac 72 00 00] (utf-32le)
- [00 00 fe ff 00 00 72 ac] (utf-32be + BOM)
- [ff fe 00 00 ac 72 00 00] (utf-32le + BOM)

» UTF-32

» UTF-16

» UTF-8

→ encode.py utf-32be utf-16be < 80dogs.txt

✓ utf-32be ✓ utf-16be: 80 good dogs

Ha'DIbaH Hund abwo alabai anjing ashun aso cane chien chó cyn câin
câine câini cão dog emba ghaddu gom hond hondur hualp huan hund
imba imbua imbwe inu it jindo kelb khuy klèb koer koira kotta
kutta kutya köpek köpeği køter mbula mbwa mbwagh mbwene mbwá mbu njau
njoka pa perro pes pies qen qeni sag sagol sobaka suns szczeniak
tz'i' ukudla welpe xolo zwin āso āšun žééchäq'í šuo كلب אבו
କୁଟ୍ଟା ଶ୍ଵର ମନୋ ଦୁଃଖ ପାତା ହଜାରି ଚାଲିବାରି

✓ utf-32be: 80 good dogs, 338 chars encoded in 1352 bytes, 4.00 bytes per char

✓ utf-16be: 80 good dogs, 338 chars encoded in 682 bytes, 2.02 bytes per char

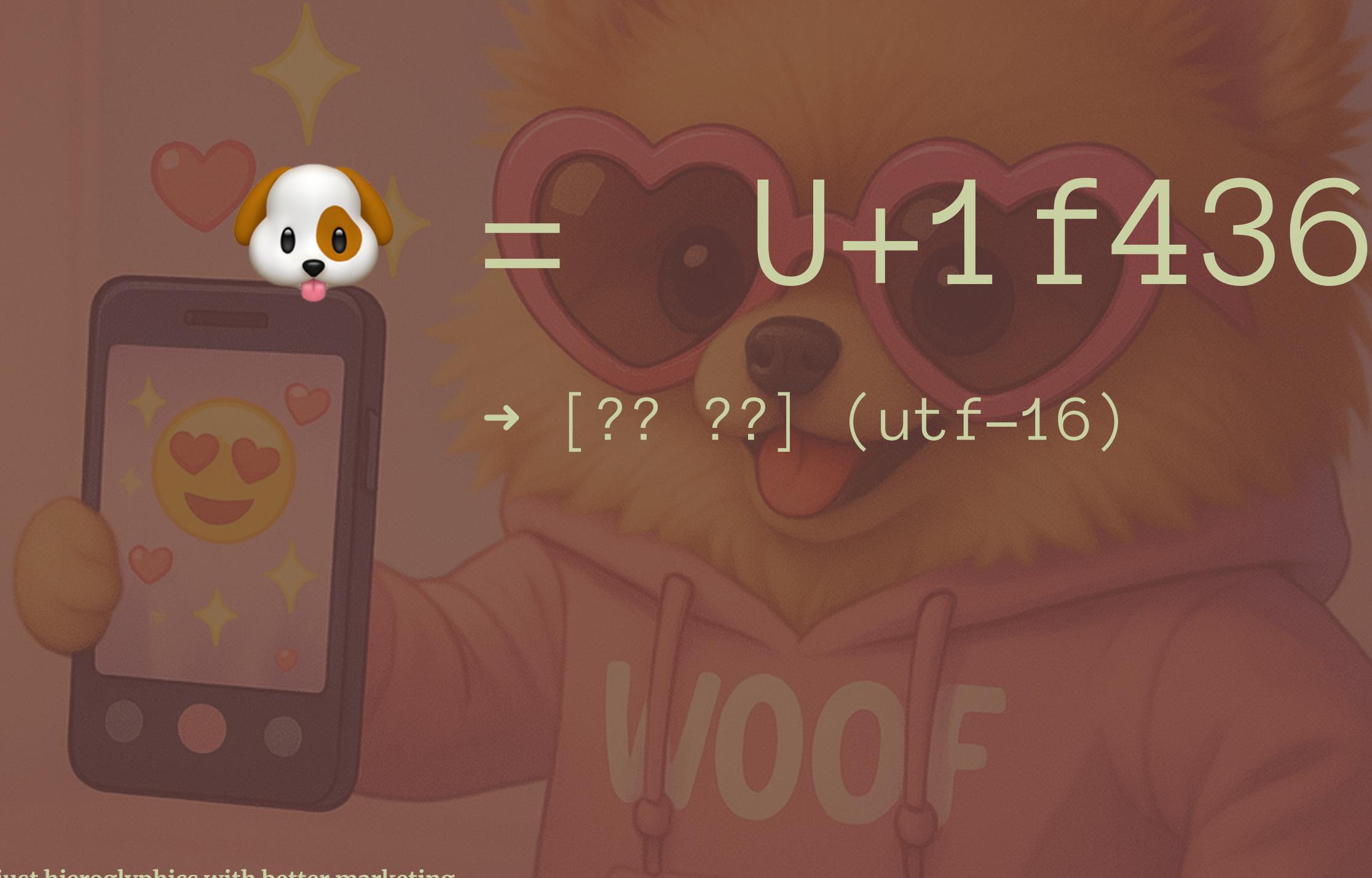
犬 = U+72ac

- [72 ac] (utf-16be)
- [ac 72] (utf-16le)
- [fe ff 72 ac] (utf-16be + BOM)
- [ff fe ac 72] (utf-16le + BOM)



Emojis are just hieroglyphics with better marketing

51/79



= U+1f436

→ [?? ??] (utf-16)

Range	Block
U+0000→U+007f	Latin Basic
U+0080→U+00ff	Latin Supplement
U+0100→U+024f	Latin Extended
U+0250→U+02ff	Phonetics
U+0300→U+05ff	Greek, Cyrillic, Hebrew
U+0600→U+0Fff	Arabic, Indian
U+1000→U+17ff	S & SE Asian
U+1800→U+24ff	E Asian
U+2000→U+27ff	Symbols
U+2800→U+28ff	Braille, Shapes
U+e80→U+9fff	Chinese
U+ac00→U+d7af	Korean
U+d800→U+dbff	High Surrogates
U+dc00→U+dfef	Low Surrogates
U+e000→U+f8ff	Private Use



= U+1f436

$$0x1f436 - 0x10000 = 0xf436$$

= 0b00011101000110110

= [0b00011101] [0b000110110]

= [0x003d] [0x0036]

+ [0xd800] + [0xdc00]

= [0xd83d] [0xdc36]

→ U+d83d U+dc36



= U+1f436

→ U+d83d U+dc36
→ [d8 3d] [dc 36] (utf-16)

Unicode Encodings

» UTF-32

» UTF-16

» UTF-8

```
→ encode.py utf-16be utf-8 < 80dogs.txt
```

✓ utf-16be ✓ utf-8: 80 good dogs

Ha'DIbaH Hund abwo alabai anjing ashun aso cane chien chó cyn câin
câine câini cão dog emba ghaddu gom hond hondur hualp huan hund
imba imbua imbwe inu it jindo kelb khuy klèb koer koira kotta
kutta kutya köpek köpeği køter mbula mbwa mbwagh mbwene mbwá mbu njau
njoka pa perro pes pies qen qeni sag sagol sobaka suns szczeniak
tz'i' ukudla welpe xolo zwin āso āšun Łééchäq'í šuo كلب אבו
କୁଟ୍ଟା ଶ୍ଵର ମନୋ ଦୁଃ ପାତା ହାତି ହାତି

✓ utf-16be: 80 good dogs, 338 chars encoded in 682 bytes, 2.02 bytes per char
✓ utf-8 : 80 good dogs, 338 chars encoded in 413 bytes, 1.22 bytes per char

U+0000→U+007F f

→[0xxxxxxx]

U+0080→U+07ff f

→[110xxxxx 10xxxxxx]

U+0800→U+ffff f

→[1110xxxx 10xxxxxx 10xxxxxx]

U+10000→U+10ffff f

→[11110xxx 10xxxxxx 10xxxxxx 10xxxxxx]

U+00e3 ('ã')

= 00011100011

= 00011 100011

110xxxxx 10xxxxxx

→ [11000011 10100011]

= [c3 a3]

[0xxxxxxx]

[110xxxxx 10xxxxxx]

[1110xxxx 10xxxxxx 10xxxxxx]

[11110xxx 10xxxxxx 10xxxxxx 10xxxxxx]

[0xxxxxxx] ASCII

[10xxxxxx] Continuation

[110xxxxx] Lead of 2 byte sequence

[1110xxxx] Lead of 3 byte sequence

[11110xxx] Lead of 4 byte sequence

[...10100011 10100011 11000011 10100011 01010101...]



pies

→ [70 69 65 73] (ascii)
→ [70 69 65 73] (utf-8)



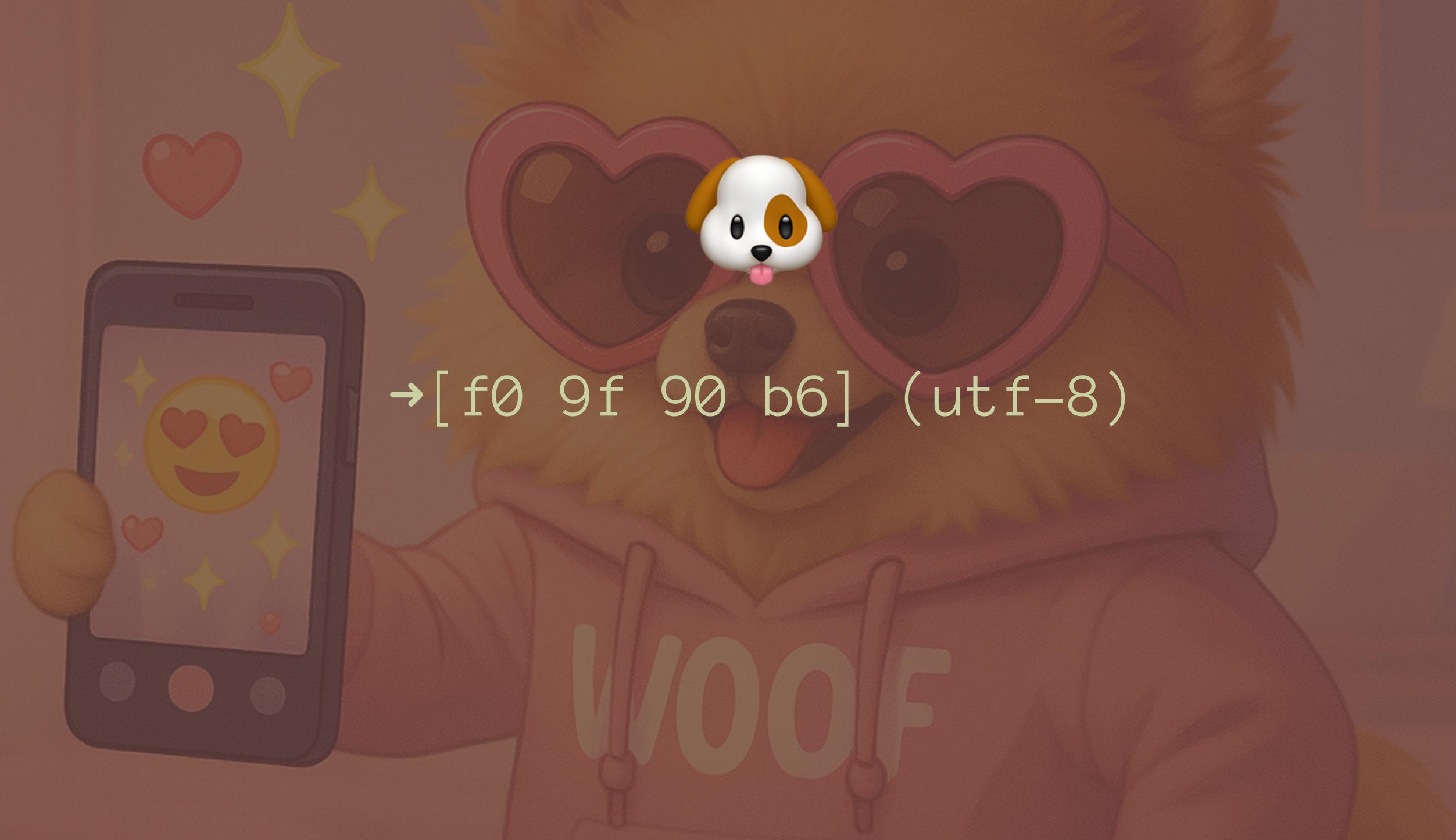
cão

→ [63 c3 a3 6f] (utf-8)



狗

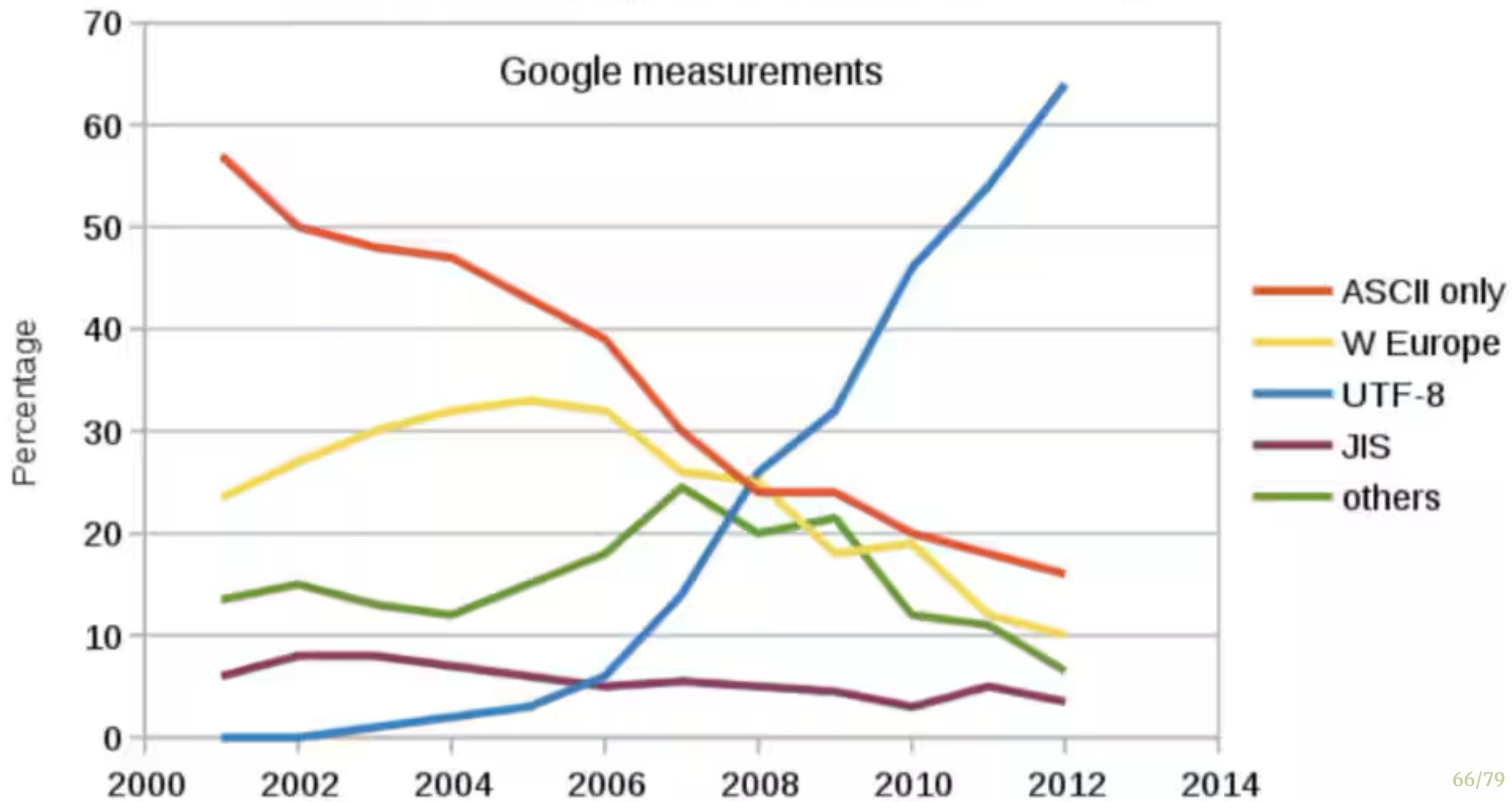
→[e7 8b 97] (utf-8)



→ [f0 9f 90 b6] (utf-8)

Encoding	Good Dogs	Bytes per Char
ascii	53	1.00
latin1	62	1.00
latin2	60	1.00
gb2312	60	1.03
utf-8	80	1.22
utf-16	80	2.02
utf-32	80	4.00

Share of web pages with different encodings



- » Before Unicode
- » Unicode
- » Oracle CharacterSet

```
create database piesek  
character set we8iso8859p1  
national character set al16utf16
```

...

```
create table dogs (name      nchar(2),  
                  language char(10 char));
```

```
insert into dogs (name, language)  
values (n'🐶', 'Emoji');
```


we8iso8859p1
al16utf16

we 8 iso8859p1
al 16 utf16

Database Version	CharacterSet	National CharacterSet
6	we8iso8859p1	
7	we8iso8859p2, zhs16cgb231280	
8i	utf8	utf8
9i	al32utf8	al16utf16

UTF8 (\neq UTF-8)

U+0000→U+007f

→ [0xxxxxxx]

U+0080→U+07ff

→ [110xxxxx 10xxxxxx]

U+0800→U+ffff

→ [1110xxxx 10xxxxxx 10xxxxxx]

U+10000→U+10ffff

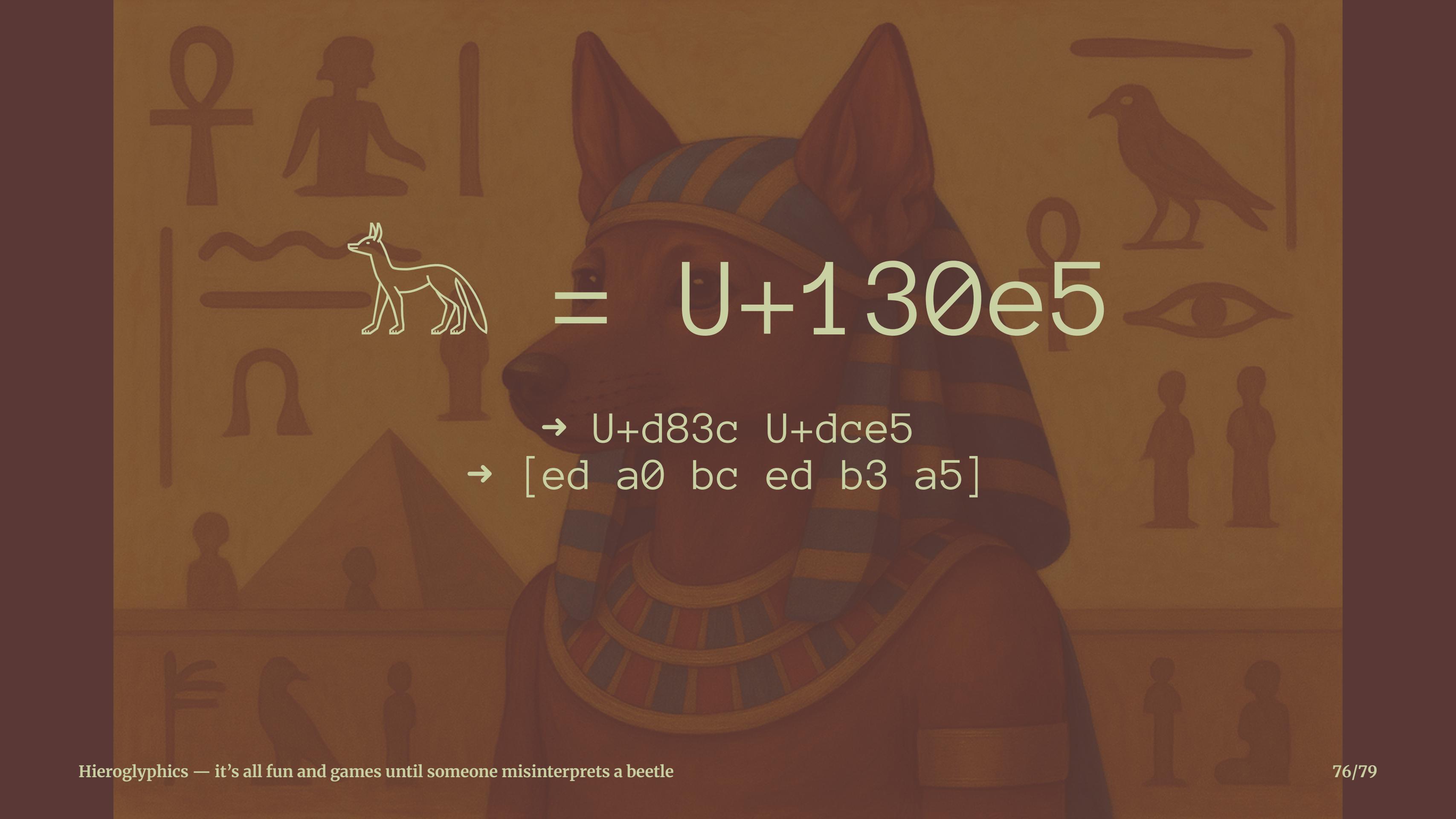
→ [11110*** 10***** 10***** 10*****]



Hieroglyphics — it's all fun and games until someone misinterprets a beetle



= U+130e5



= U+130e5

→ U+d83c U+dce5
→ [ed a0 bc ed b3 a5]

In Summary

- » CharacterSet: AL32UTF8 (not UTF8)
- » National CharacterSet: AL16UTF16 (not UTF8)

[54 48 41 4e 4b 53] (ascii/latin1/gb2312/utf-8)

[11110000 10011111 10011001 10001111] (utf-8)

[54 48 41 4e 4b 53] (ascii/latin1/gb2312/utf-8)

→ THANKS

[11110000 10011111 10011001 10001111] (utf-8)

→ U+1f64f → 