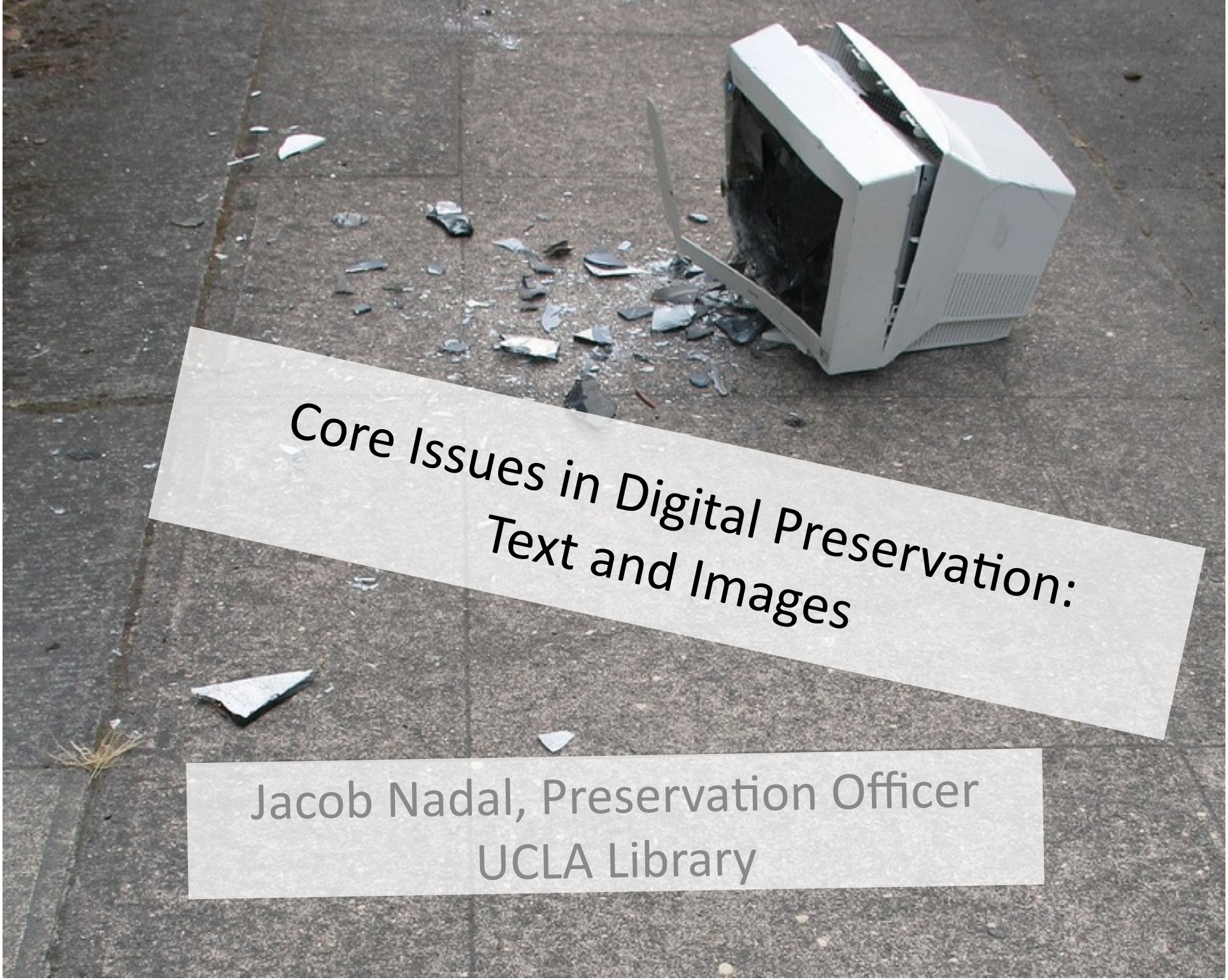
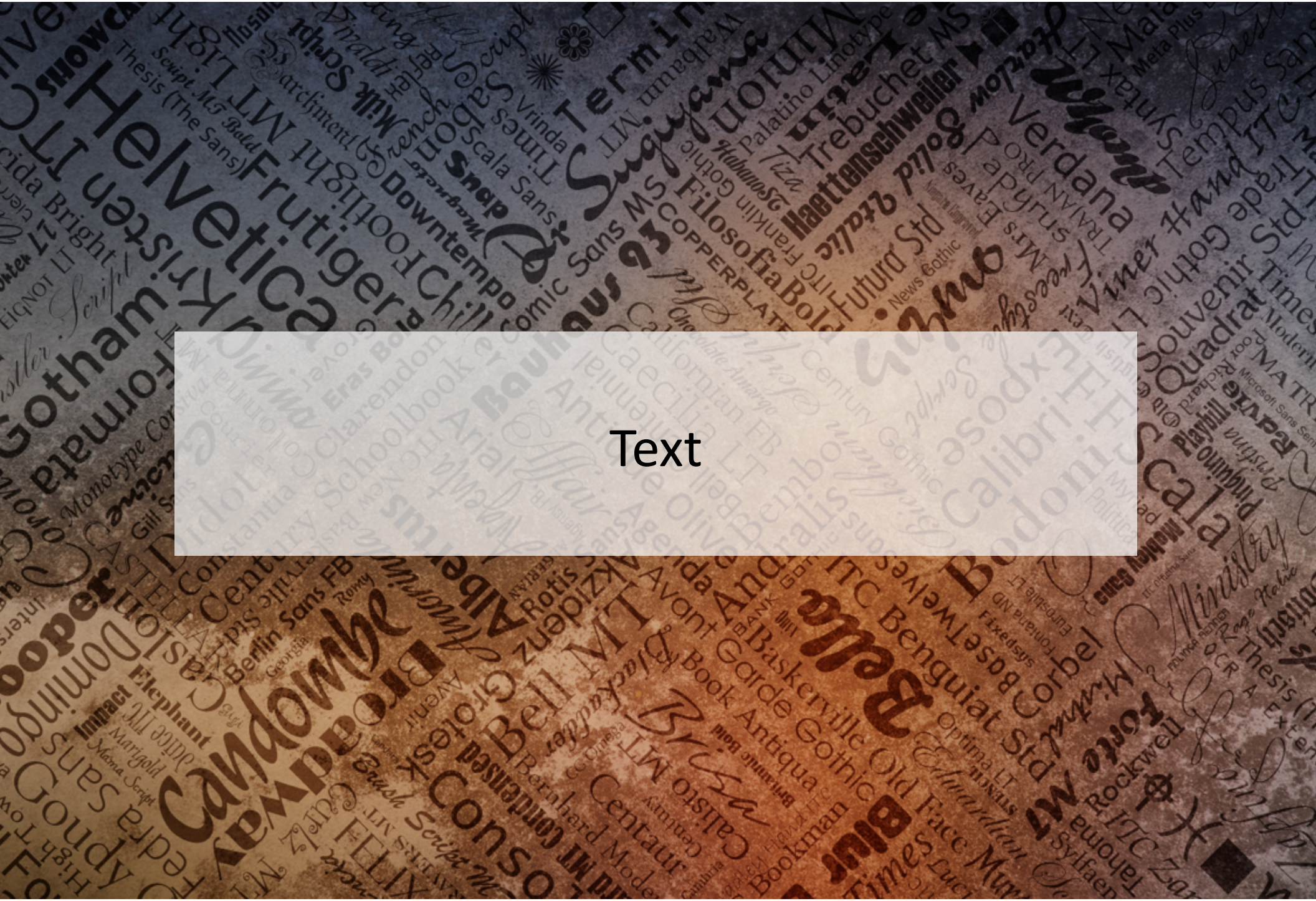


Core Issues in Digital Preservation: Text and Images

Jacob Nadal, Preservation Officer
UCLA Library


Text

Text

- Digital text encodings have their roots in telegraph codes *(really)*
- ASCII (American Standard Code for Information Interchange) dates from 1968
  - 7-bit code
  - 32 control characters
  - 94 printable characters

USASCII code chart

b_7 b_6 b_5 →	0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
Bits b_4 ↓	b_3 ↓	b_2 ↓	b_1 ↓	Row ↓ Column → 0	1	2	3	4	5	6	7
0	0	0	0	0	NUL	DLE	SP	0	@	P	`	p
0	0	0	1	1	SOH	DC1	!	1	A	Q	a	q
0	0	1	0	2	STX	DC2	"	2	B	R	b	r
0	0	1	1	3	ETX	DC3	#	3	C	S	c	s
0	1	0	0	4	EOT	DC4	$	4	D	T	d	t
0	1	0	1	5	ENQ	NAK	%	5	E	U	e	u
0	1	1	0	6	ACK	SYN	&	6	F	V	f	v
0	1	1	1	7	BEL	ETB	'	7	G	W	g	w
1	0	0	0	8	BS	CAN	(	8	H	X	h	x
1	0	0	1	9	HT	EM	)	9	I	Y	i	y
1	0	1	0	10	LF	SUB	*	:	J	Z	j	z
1	0	1	1	11	VT	ESC	+	;	K	[	k	{
1	1	0	0	12	FF	FS	,	<	L	\	l	|
1	1	0	1	13	CR	GS	-	=	M	]	m	}
1	1	1	0	14	SO	RS	.	>	N	^	n	~
1	1	1	1	15	SI	US	/	?	O	_	o	DEL

Text: UTF-8

- Unicode is an unlimited way of encoding characters
- The **Unicode Transmission Format - 8 bit** (UTF-8) is the most common way to encounter Unicode
  - UTF-8 transmits using 1 to 4 “octets,” 8-bit bytes
  - First 128 of these are US-ASCII, and then there are lots of other things

Text: UTF-8

- Easy to identify
  - Given an unknown text string, a simple search pattern identifies UTF-8 over 99.5% of the time
- Default, native encoding for XML
- Multi-language support

(some of) The UTF-8 Character Set

Unicode code point	character	UTF-8 (hex.)	name
U+0530	ՠ	d4 b0	
U+0531	Ա	d4 b1	ARMENIAN CAPITAL LETTER AYB
U+0532	Բ	d4 b2	ARMENIAN CAPITAL LETTER BEN
U+0533	Գ	d4 b3	ARMENIAN CAPITAL LETTER GIM
U+0534	Դ	d4 b4	ARMENIAN CAPITAL LETTER DA
U+0535	Ե	d4 b5	ARMENIAN CAPITAL LETTER ECH
U+0536	Զ	d4 b6	ARMENIAN CAPITAL LETTER ZA
U+0537	Է	d4 b7	ARMENIAN CAPITAL LETTER EH
U+0538	Ը	d4 b8	ARMENIAN CAPITAL LETTER ET
U+0539	Թ	d4 b9	ARMENIAN CAPITAL LETTER TO
U+053A	Ժ	d4 ba	ARMENIAN CAPITAL LETTER ZHE
U+053B	Ի	d4 bb	ARMENIAN CAPITAL LETTER INI
U+053C	Լ	d4 bc	ARMENIAN CAPITAL LETTER LIWN
U+053D	Խ	d4 bd	ARMENIAN CAPITAL LETTER XEH
U+053E	Ծ	d4 be	ARMENIAN CAPITAL LETTER CA
U+053F	Կ	d4 bf	ARMENIAN CAPITAL LETTER KEN
U+0540	Հ	d5 80	ARMENIAN CAPITAL LETTER HO
U+0541	Ձ	d5 81	ARMENIAN CAPITAL LETTER JA
U+0542	Ղ	d5 82	ARMENIAN CAPITAL LETTER GHAD
U+0543	Ճ	d5 83	ARMENIAN CAPITAL LETTER CHEH
U+0544	Մ	d5 84	ARMENIAN CAPITAL LETTER MEN
U+0545	Յ	d5 85	ARMENIAN CAPITAL LETTER YI
U+0546	Ն	d5 86	ARMENIAN CAPITAL LETTER NOW
U+0547	Շ	d5 87	ARMENIAN CAPITAL LETTER SHA
U+0548	Ո	d5 88	ARMENIAN CAPITAL LETTER VO
U+0549	Չ	d5 89	ARMENIAN CAPITAL LETTER CHA
U+054A	Պ	d5 8a	ARMENIAN CAPITAL LETTER PEH
U+054B	Ջ	d5 8b	ARMENIAN CAPITAL LETTER JHEH
U+054C	Ռ	d5 8c	ARMENIAN CAPITAL LETTER RA
U+054D	Ս	d5 8d	ARMENIAN CAPITAL LETTER SEH
U+054E	Վ	d5 8e	ARMENIAN CAPITAL LETTER VEW
U+054F	Տ	d5 8f	ARMENIAN CAPITAL LETTER TIWN
U+0550	Ր	d5 90	ARMENIAN CAPITAL LETTER REH
U+0551	Ց	d5 91	ARMENIAN CAPITAL LETTER CO
U+0552	Ւ	d5 92	ARMENIAN CAPITAL LETTER YIWN

Unicode code point	character	UTF-8 (hex.)	name
U+0980	ঀ	e0 a6 80	
U+0981	ঁ	e0 a6 81	BENGALI SIGN CAN
U+0982	ং	e0 a6 82	BENGALI SIGN ANU
U+0983	ঃ	e0 a6 83	BENGALI SIGN VISA
U+0984	঄	e0 a6 84	
U+0985	অ	e0 a6 85	BENGALI LETTER A
U+0986	আ	e0 a6 86	BENGALI LETTER AA
U+0987	ই	e0 a6 87	BENGALI LETTER I
U+0988	ঈ	e0 a6 88	BENGALI LETTER II
U+0989	উ	e0 a6 89	BENGALI LETTER U
U+098A	ঊ	e0 a6 8a	BENGALI LETTER UU
U+098B	ঋ	e0 a6 8b	BENGALI LETTER VO
U+098C	ঌ	e0 a6 8c	BENGALI LETTER VO
U+098D	঍	e0 a6 8d	
U+098E	঎	e0 a6 8e	
U+098F	এ	e0 a6 8f	BENGALI LETTER E
U+0990	ঐ	e0 a6 90	BENGALI LETTER AI
U+0991	঑	e0 a6 91	
U+0992	ও	e0 a6 92	
U+0993	ঔ	e0 a6 93	BENGALI LETTER O
U+0994	ক	e0 a6 94	BENGALI LETTER AU
U+0995	খ	e0 a6 95	BENGALI LETTER KA
U+0996	গ	e0 a6 96	BENGALI LETTER KH
U+0997	ঘ	e0 a6 97	BENGALI LETTER GA
U+0998	ঙ	e0 a6 98	BENGALI LETTER GH
U+0999	চ	e0 a6 99	BENGALI LETTER NG
U+099A	ছ	e0 a6 9a	BENGALI LETTER CA
U+099B	জ	e0 a6 9b	BENGALI LETTER CH
U+099C	ঝ	e0 a6 9c	BENGALI LETTER JA
U+099D	ঞ	e0 a6 9d	BENGALI LETTER JH
U+099E	ট	e0 a6 9e	BENGALI LETTER NY
U+099F	ঠ	e0 a6 9f	BENGALI LETTER TT
U+09A0	ড	e0 a6 a0	BENGALI LETTER TT
U+09A1	ঢ	e0 a6 a1	BENGALI LETTER DDA
U+09A2	ণ	e0 a6 a2	BENGALI LETTER DDHA

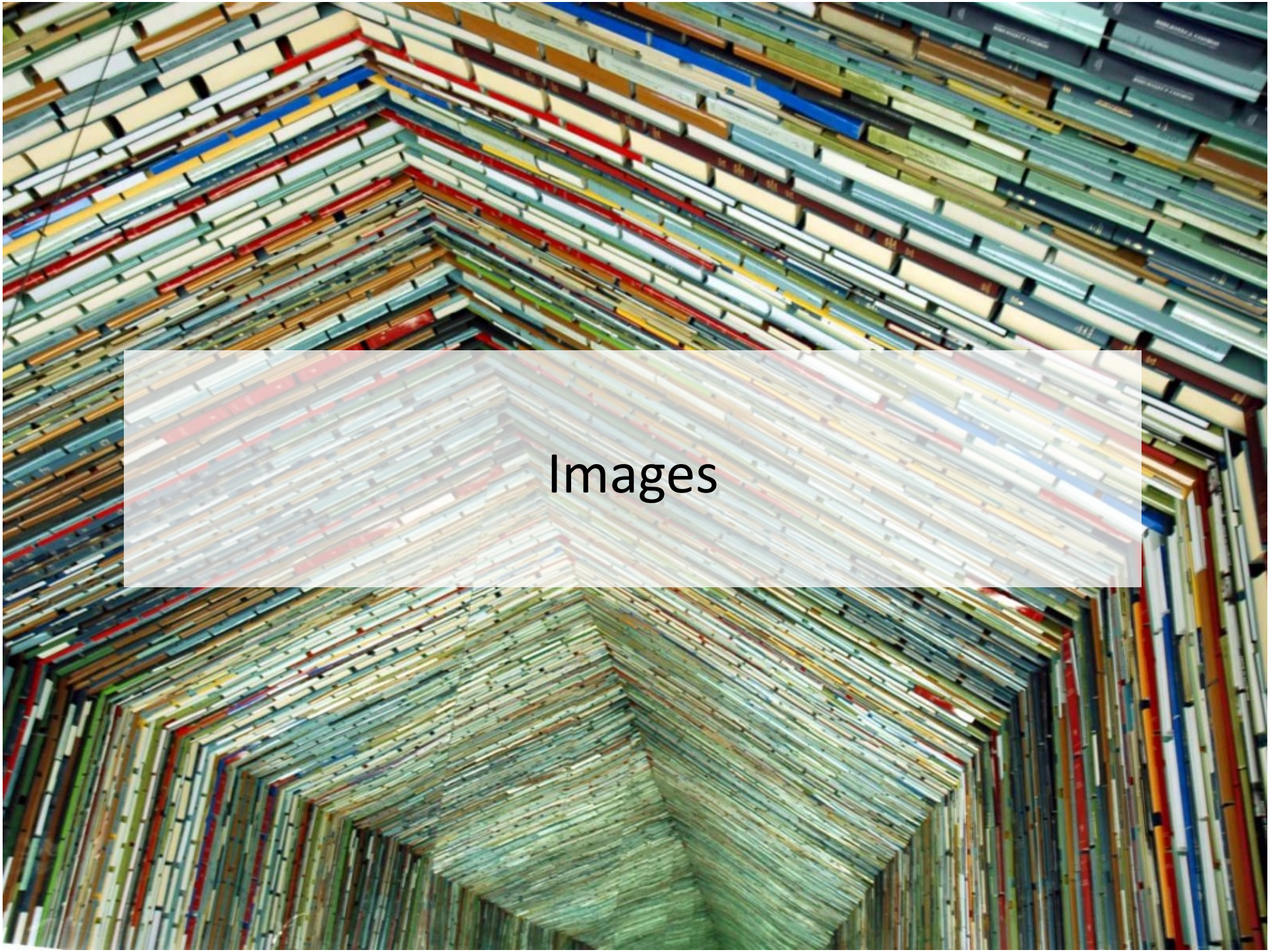
Unicode code point	character	UTF-8 (hex.)	name
U+0E00	ก	e0 b8 80	
U+0E01	ก	e0 b8 81	THAI CHARACTER KO KAI
U+0E02	ข	e0 b8 82	THAI CHARACTER KHO KHAI
U+0E03	ฃ	e0 b8 83	THAI CHARACTER KHO KHUAT
U+0E04	ค	e0 b8 84	THAI CHARACTER KHO KHWAI
U+0E05	ฅ	e0 b8 85	THAI CHARACTER KHO KHON
U+0E06	ฆ	e0 b8 86	THAI CHARACTER KHO RAKHANG
U+0E07	ง	e0 b8 87	THAI CHARACTER NGO NGU
U+0E08	จ	e0 b8 88	THAI CHARACTER CHO CHAN
U+0E09	ฉ	e0 b8 89	THAI CHARACTER CHO CHING
U+0E0A	ช	e0 b8 8a	THAI CHARACTER CHO CHANG
U+0E0B	ซ	e0 b8 8b	THAI CHARACTER SO SO
U+0E0C	ฌ	e0 b8 8c	THAI CHARACTER CHO CHOE
U+0E0D	ญ	e0 b8 8d	THAI CHARACTER YO YING
U+0E0E	ฎ	e0 b8 8e	THAI CHARACTER DO CHADA
U+0E0F	ฏ	e0 b8 8f	THAI CHARACTER TO PATAK
U+0E10	ฐ	e0 b8 90	THAI CHARACTER THO THAN
U+0E11	ฑ	e0 b8 91	THAI CHARACTER THO NANGMONTHO
U+0E12	ฒ	e0 b8 92	THAI CHARACTER THO PHUTHAO
U+0E13	ณ	e0 b8 93	THAI CHARACTER NO NEN
U+0E14	ด	e0 b8 94	THAI CHARACTER DO DEK
U+0E15	ต	e0 b8 95	THAI CHARACTER TO TAO
U+0E16	ถ	e0 b8 96	THAI CHARACTER THO THUNG
U+0E17	ท	e0 b8 97	THAI CHARACTER THO THAHAN
U+0E18	ธ	e0 b8 98	THAI CHARACTER THO THONG
U+0E19	น	e0 b8 99	THAI CHARACTER NO NU
U+0E1A	บ	e0 b8 9a	THAI CHARACTER BO BAIMAI
U+0E1B	ป	e0 b8 9b	THAI CHARACTER PO PLA
U+0E1C	ผ	e0 b8 9c	THAI CHARACTER PHO PHUNG
U+0E1D	ฝ	e0 b8 9d	THAI CHARACTER FO FA
U+0E1E	พ	e0 b8 9e	THAI CHARACTER PHO PHAN
U+0E1F	ฟ	e0 b8 9f	THAI CHARACTER FO FAN
U+0E20	ภ	e0 b8 a0	THAI CHARACTER PHO SAMPHAO
U+0E21	ม	e0 b8 a1	THAI CHARACTER MO MA

Images and Text

- That unicode character set that just scrolled by was, of course, an image.
- Computers don’t read; they encode and decode
- So, digitized books are page images plus text transcriptions plus the metadata that holds all of that together.

Next: Images

TEXT Q&A


Images

TIFF

- Developed by Aldus in 1986, and passed to Adobe.
- Version 6.0 published in 1992 and has no IP restrictions
- TIFF may include compressed parts; **be diligent about using uncompressed TIFF.**
  - LZW (lossless) compression debatable.

JPEG 2000

- Developed in 2000, released as ISO standard with a no-cost license for its core components
- Wavelet-based, so can hold several levels of compression within one file
- Shortage of authoring tools

Digital Negative

- Developed by Adobe to provide a non-proprietary format for RAW camera data
- May be valuable as a digital preservation format for the specific use-case of **born-digital photography**

The Other Image Formats and...

- JPEG (not JPEG2000)
- RAW (Camera sensor data)
- PNG (Portable network graphics)
- PSD (Photoshop document)

… Their Problems

- Compression or size limits (JPEG, PNG)
- Intellectual property / manufacturers proprietary standards (PSD, RAW)

And then there's PDF

- Lots of PDF types, with varying levels of preservability. Currently in version 1.7.
- PDF is (simplistically) a metadata wrapper for text and graphic content.
  - PDF **can** contain almost any media – raster and vector graphics, forms, audio, video, and more
- PDF 1.4 has an off-shoot called PDF/A that is used for archiving

What to put into an image

- Resolution
  - 300 dpi bare minimum, 600 dpi standard, 1200+ for special circumstances
- Bit-Depth (color)
  - 8-bit (256 grays) or 24-bit (256 Reds, 256 Greens, and 256 Blues for 16 million combinations)

Resolution

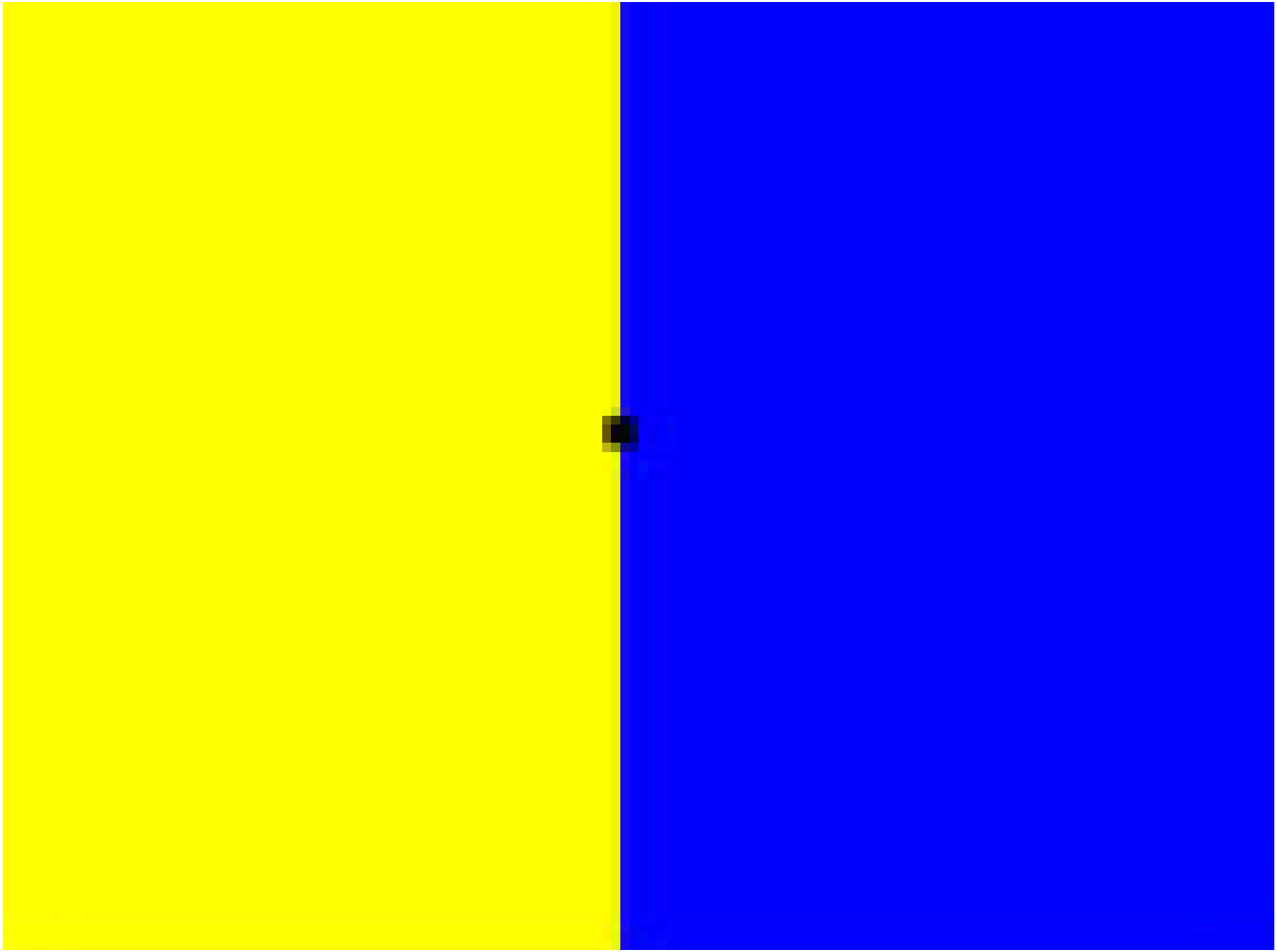
- Scanners
  - Limited by the number of sensors in the scanner's array (top to bottom) and the motion of its motor (left to right)
- Cameras
  - Limited by physical size (H" x W") and sensor density (pixels per inch) of the imaging chip

Color

- Color needs to be calibrated
- The eye, the image sensor, and the image rendering device all have different color sensitivity
- None of these are a perfect match for the source spectra
  - And those vary depending on the type of illumination
- Best practice is to calibrate all devices and **not** edit color on the initial capture
- Create derivatives for each use-case: web delivery (RGB), high-res. display (RGB), print (CMYK), etc.

Don’t trust your eyes

CHROMATIC ADAPTATION

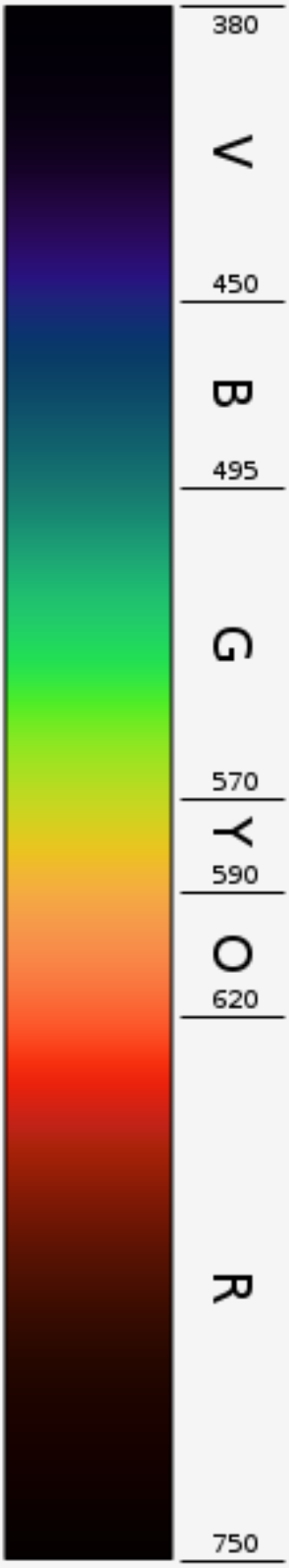


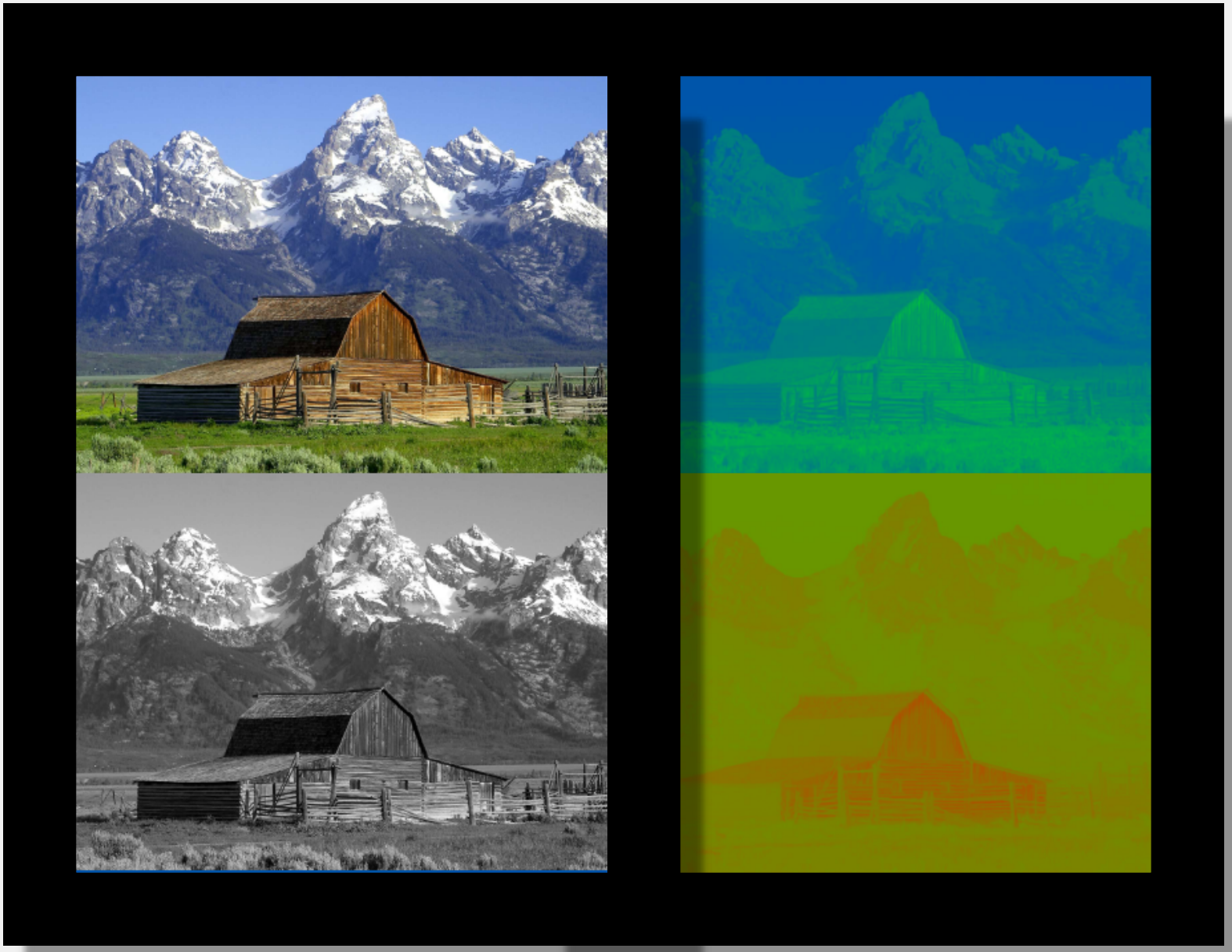
Seeing and Recording and Transmission

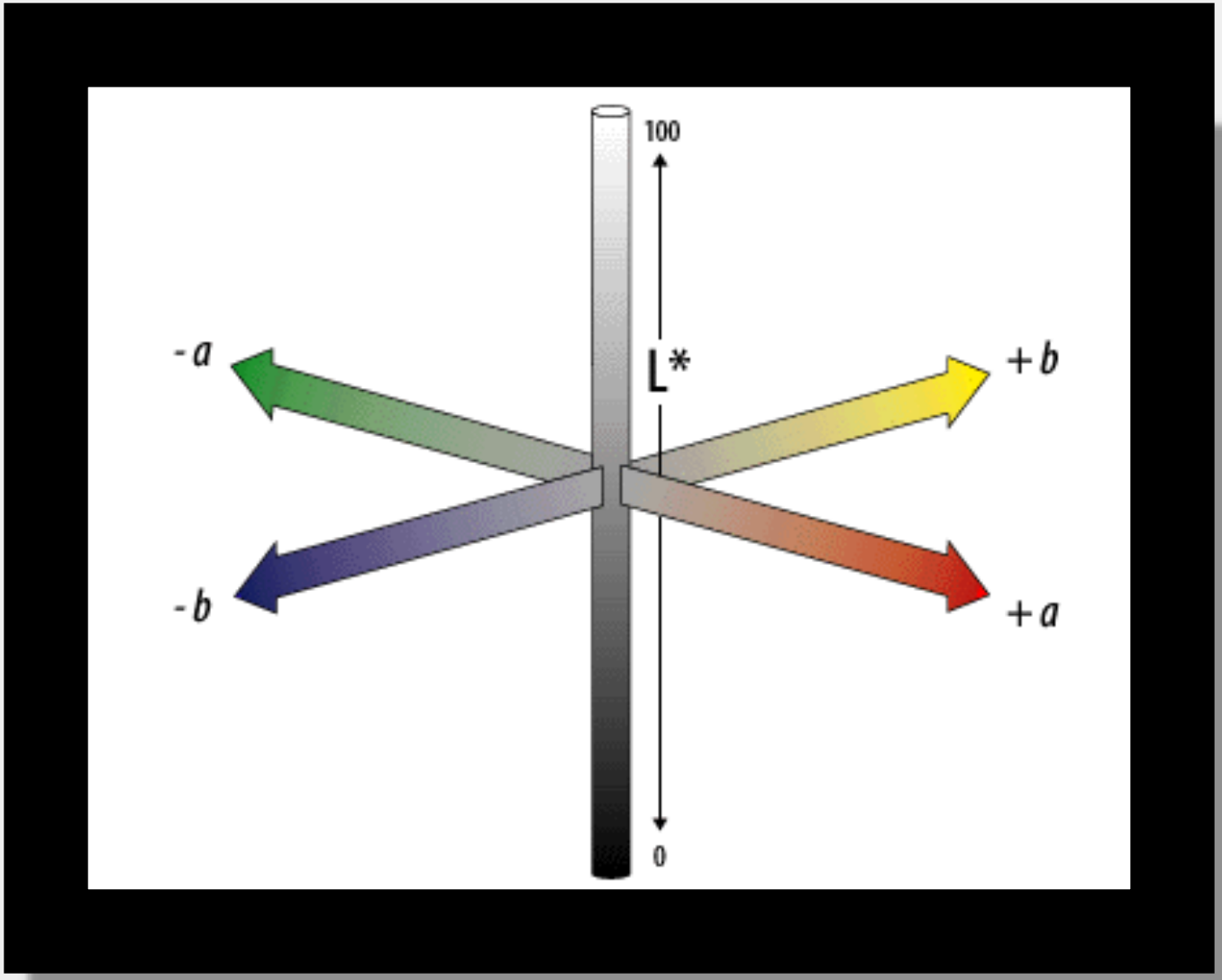
- The eye processes light in two ways
  - Hue and saturation (color shade and depth; cones)
  - Luminance (brightness, like “black & white”; rods).
- Computers and digital imaging devices process light as three color channels: red, green, and blue
  - A fixed amount of data is assigned to each color
  - “24-bit” color has 8 bits worth of R, G, and B (256 levels each; 16.7 million combinations)
- Colors are returned as RGB (digital) or CMYK (print)

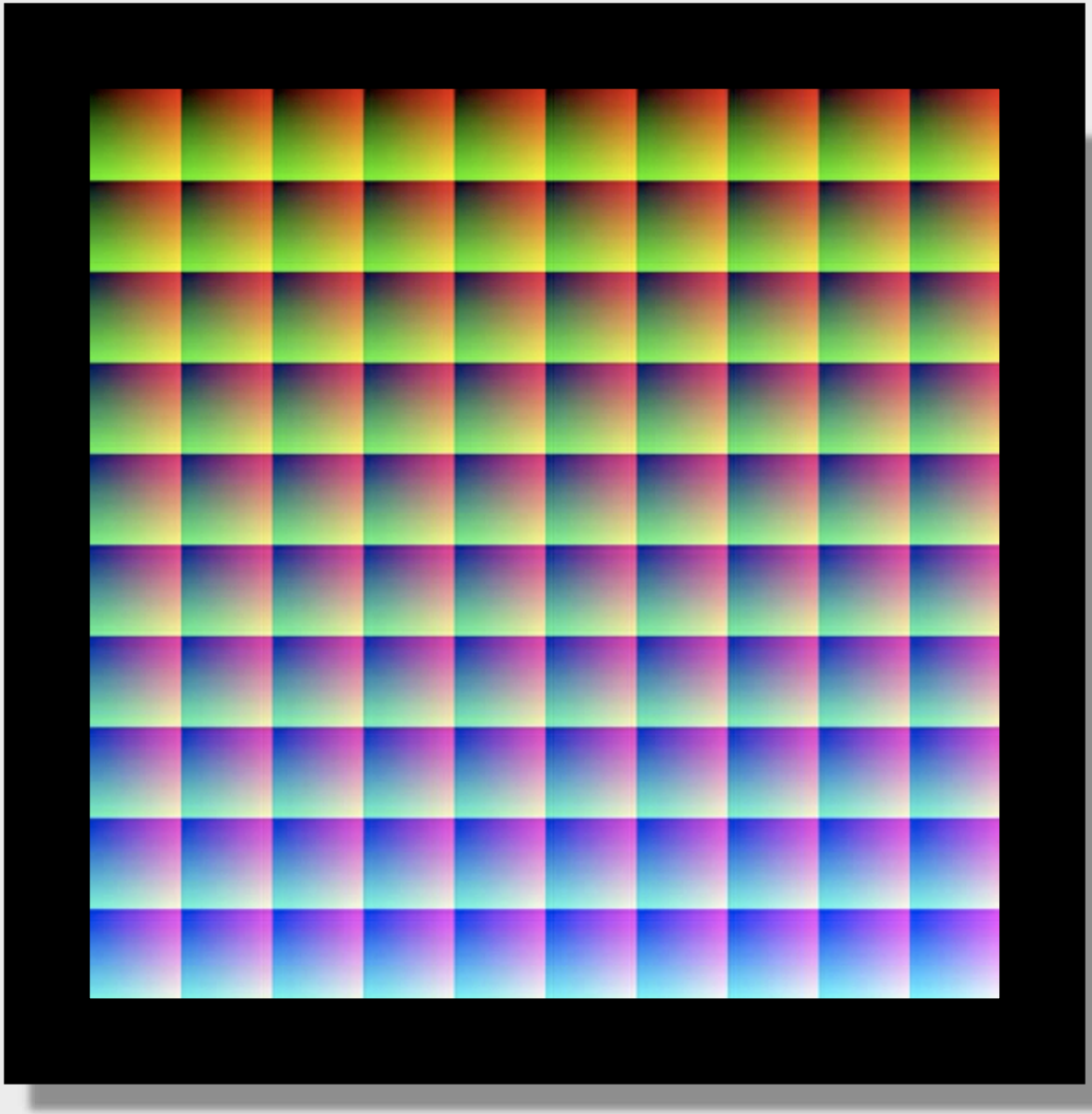
Multi-spectral imaging

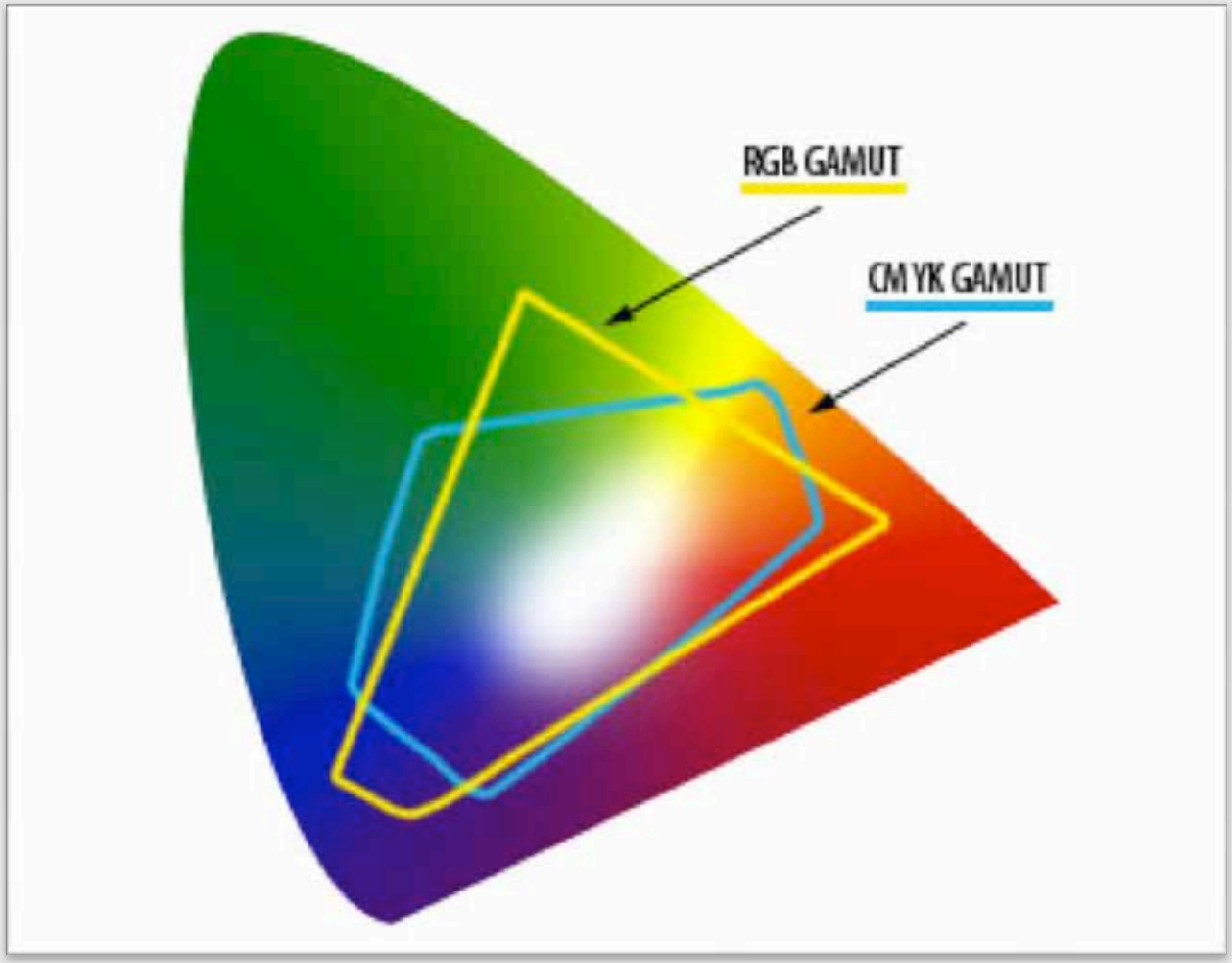
- Light is radiation. Our visible spectrum ranges from 390 to 750 nanometers.
  - Immediately below (longer freq.) is infrared, which we encounter as heat, above is ultraviolet
- Under different types of radiation, media reflect, refract, fluoresce in different ways.
  - Infrared, Ultraviolet, X-radiation, Polarization, and more can produce different imaging effects
  - More image capture in more spectra means more complete digital representation
- But mostly, we just need the visible spectrum.


380
V
450
B
495
G
570
Y
590
O
620
R
750




100
L*
0
-a
+b
-b
+a




RGB GAMUT
CMYK GAMUT

Starting Background Color

Note how much your eye adjusted, and how quickly.

Ending Background Color

25

Seeing and Record… Transmissio…

- The eye processes light in two …
  - Hue and saturation (color shade …
  - Luminance (brightness, like "black…
- Computers and digital imaging … light as three color channels: re…
  - A fixed amount of data is assigned…
  - "24-bit" color has 8 bits worth of … levels each; 16.7 million combinat…
- Colors are returned as RGB (digi… (print)

26

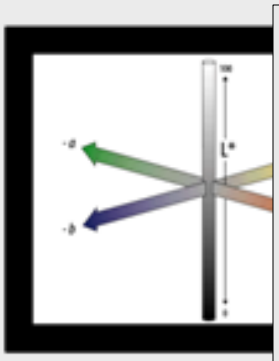
Multi-spectral in…

- Light is radiation. Our visible sp… from 390 to 750 nanometers.
  - Immediately below (longer freq.) … encounter as heat, above is ultrav…
- Under different types of radiati… refract, fluoresce in different wa…
  - Infrared, Ultraviolet, X-radiation, … more can produce different imagi…
  - More image capture in more spec… complete digital representation
- But mostly, we just need the vis…

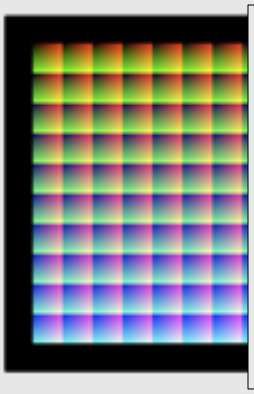
27

28

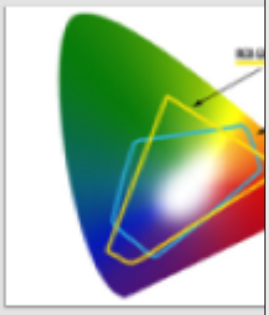
29



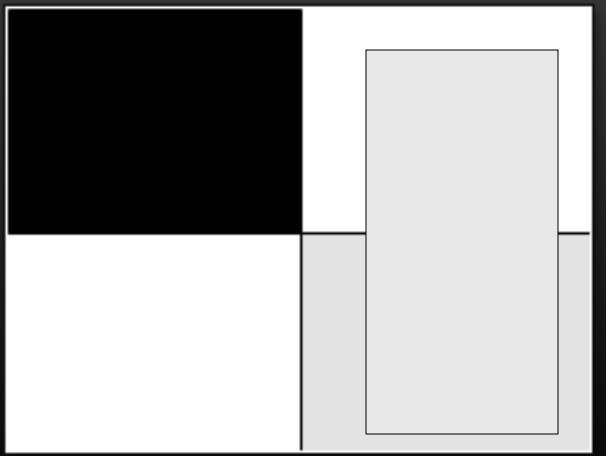
30



31



32



http://www.jacobnadal.com/247

IMAGE Q&A