Информатика

Биологический институт
Национальный исследовательский
Томский государственный университет

Лекция 3

Кодирование текста

Цвет в биологии и информатике

Дмитрий Владимирович Курбатский

старший преподаватель каф. ихтиологии и гидробиологии, научный сотрудник ЛМБ БИ ТГУ, магистр биологии

- Зоологический музей (к. 123)
- Компьютерный класс (к. 028)

Главный корпус

- Группа ВКонтатике «Курсы "Информатика" и "Информационные технологии"»:

  vk.com/i_it_bi_tsu

- Персональный раздел:

  zoo.tsu.ru/kdv

- Рейтинг на сайте Professorrating.ru

Блок 1

Символы и строки

Символы

- Символ – не только буква или цифра!
- 1..4 байт
- тип *Char*

Пробелы и не только

- неразрывный пробел °
- табулятор →
- конец абзаца ¶
- разрывы
  - строки ↲
  - колонки
  - страницы
  - раздела
- мягкий перенос ¬

Типы символов

- управляющие
  - пробелы
  - разрывы
  - переключатели
- буквенно-цифровые
  - латинские (английские)
  - национальные
  - диакритики
- знаки пунктуации
  - стандартные
  - расширенные
- математические и иные символы
- псевдографика
- пиктограммы и идеограммы

Переносимый набор символов

- NUL должен быть символом, где все биты установлены в 0.
- Коды десятичных цифр 0—9 должны идти в возрастающем порядке, причём коды двух соседних цифр должны отличаться на единицу.
  - 30h 31h … 39h
- Коды всех символов должны быть представимы одним байтом.
- Коды символов должны быть неотрицательными.
- Всего – 103 символа.

Управляющие символы

- Телетайп

Номер		Английское название	Русское название	Escape
	0	NULL	пустой символ	\0
	1	START OF HEADING	начало заголовка	
	2	START OF TEXT	начало текста	
	3	END OF TEXT	конец текста	
	4	END OF TRANSMISSION	конец передачи	
	5	ENQUIRY	запрос	
	6	ACKNOWLEDGE	подтверждение	
	7	BELL	звуковой сигнал	\a
	8	BACKSPACE	возврат на шаг	\b
	9	CHARACTER TABULATION (horizontal tabulation)	горизонтальная табуляция	**\t**
0A		LINE FEED (LF)	перевод строки	**\n**
0B		LINE TABULATION (vertical tabulation)	вертикальная табуляция	\v
0C		FORM FEED	смена страницы	\f
0D		CARRIAGE RETURN (CR)	возврат каретки	**\r**
0E		SHIFT OUT (locking-shift one)	режим национальных символов	
0F		SHIFT IN (locking-shift zero)	режим обычного ASCII	

CR и LF

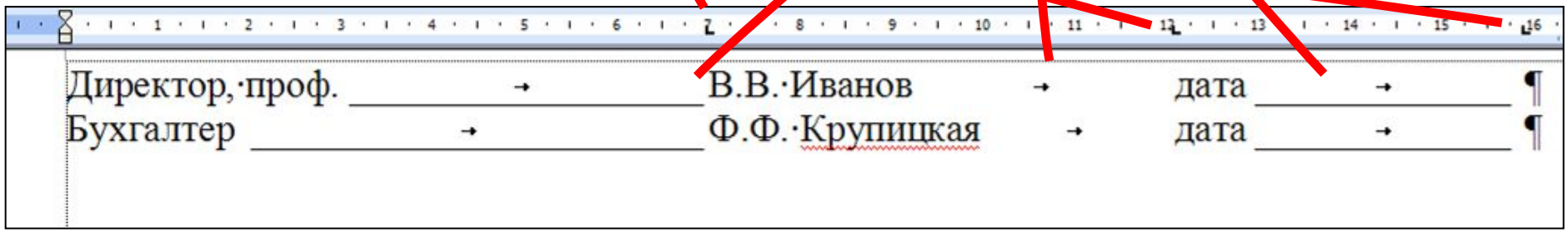
Перевод строки

- **LF** (\n, *0Ah*) - используется в Multics, **UNIX**, UNIX-подобных операционных системах (**GNU/Linux**, AIX, Xenix, **Mac OS X**, **FreeBSD** и др.), BeOS, Amiga UNIX, RISC OS и др.
- **CR** (\r, *0Dh*) используется в 8-битовых машинах Commodore, машинах TRS-80, Apple II, системах **Mac OS** до версии 9 и OS-9.
- **CR+LF** (\r\n, *0D0Ah*) используется в DEC RT-11 и большинстве других ранних не-UNIX- и не-IBM-систем, а также в CP/M, MP/M, **MS-DOS**, OS/2, **Microsoft Windows**, Symbian OS, протоколах Интернет.
- В HTML:
  -
  - <p>

¶

Табуляция

- Символ табуляции
- Позиция табуляции



Набираете вы текст в Ворде...

Номер	Английское название	Русское название
10	DATA LINK ESCAPE	???
11	DEVICE CONTROL ONE	1-й код управления устройством
12	DEVICE CONTROL TWO	2-й код управления устройством
13	DEVICE CONTROL THREE	3-й код управления устройством
14	DEVICE CONTROL FOUR	4-й код управления устройством
15	NEGATIVE ACKNOWLEDGE	отрицательное подтверждение
16	SYNCHRONOUS IDLE	пустой символ для синхронного режима передачи
17	END TRANSMISSION BLOCK	конец блока передаваемых данных
18	CANCEL	отмена
19	END OF MEDIUM	конец носителя
1A	SUBSTITUTE	символ замены
1B	ESCAPE	Альтернативный регистр №2 (АР2)
1C	INFORMATION SEPARATOR FOUR (file separator)	разделитель данных № 4 (разделитель файлов)
1D	INFORMATION SEPARATOR THREE (group separator)	разделитель данных № 3 (разделитель групп)
1E	INFORMATION SEPARATOR TWO (record separator)	разделитель данных № 2 (разделитель записей)
1F	INFORMATION SEPARATOR ONE (unit separator)	разделитель данных № 1 (разделитель полей)
7F	DELETE	удаление

Управляющие символы Unicode

- 034F, COMBINING GRAPHEME JOINER. Объединить символы, стоящие слева и справа (создать лигатуру).
- 200B, ZERO-WIDTH SPACE, пробел нулевой ширины. При выравнивании по ширине может расширяться.
- 200C, ZERO WIDTH NON-JOINER. Запрещает образование лигатур.
- 200D, ZERO WIDTH JOINER. Разрешает образование лигатур.
- 200E, LEFT-TO-RIGHT MARK. Писать слева направо.
- 200F, RIGHT-TO-LEFT MARK. Писать справа налево.
- 2028, LINE SEPARATOR, разделитель строк. Разделяет строки текста, но не абзацы.
- 2029, PARAGRAPH SEPARATOR, разделитель абзацев. Разделяет абзацы текста.
- 202A, LEFT-TO-RIGHT EMBEDDING. Начало текста, написанного слева направо, внутри текста, написанного справа налево.

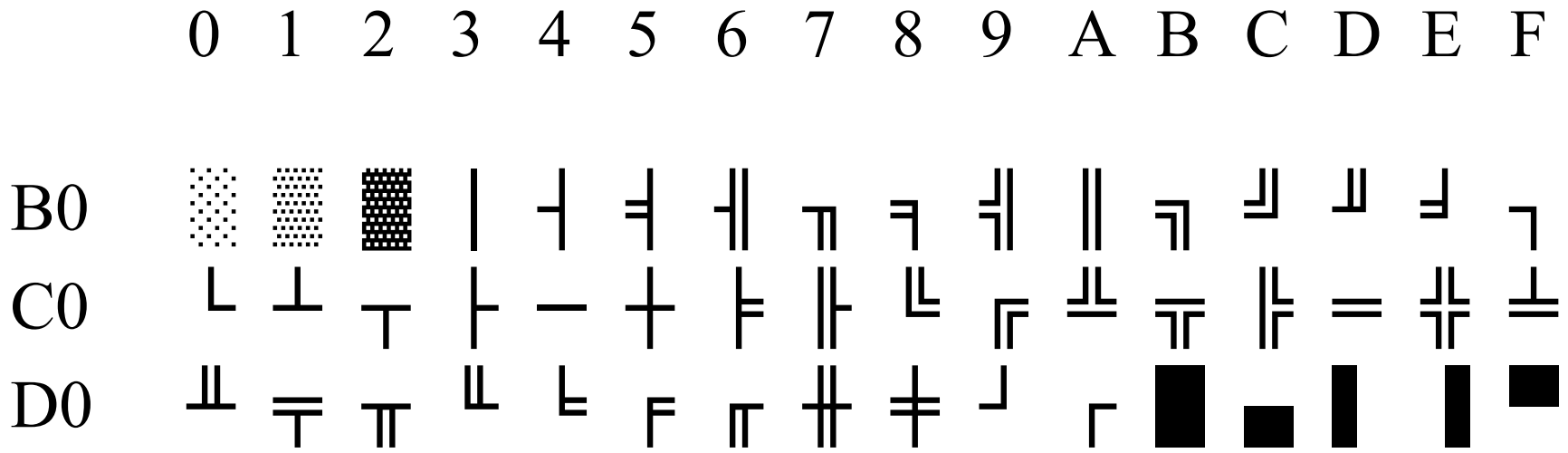
Управляющие символы Unicode

- 202B, RIGHT-TO-LEFT EMBEDDING. Начало текста, написанного справа налево, внутри текста, написанного слева направо.
- 202C, POP DIRECTIONAL FORMATTING. Конец вставленного текста с другим направлением.
- 202D, LEFT-TO-RIGHT OVERRIDE.
- 202E, RIGHT-TO-LEFT OVERRIDE.
- 2060, WORD JOINER, соединитель слов.
- FE01 … FE0F, VARIATION SELECTOR-1…16, выбор варианта начертания № 1 … № 16.
- FEFF, ZERO WIDTH NO-BREAK SPACE / BYTE ORDER MARK, неразрывный пробел нулевой ширины / индикатор порядка байтов. Этот символ используется для указания того, что данный файл записан в UTF-16 или UTF-32 с определённым порядком байтов (поскольку символа FFFE нет, а в UTF-8 байты FE и FF не используются). Использование этого символа в качестве неразрывного пробела нулевой ширины не рекомендуется; для этого есть символ U+2060 (word joiner).
- FFFD, REPLACEMENT CHARACTER, заменяющий символ. Используется, когда значение символа неизвестно или не может быть выражено в Уникоде (см. также символ 1A).
- E0100 … E01EF, VARIATION SELECTOR-17…256, выбор варианта начертания № 17 … № 256.

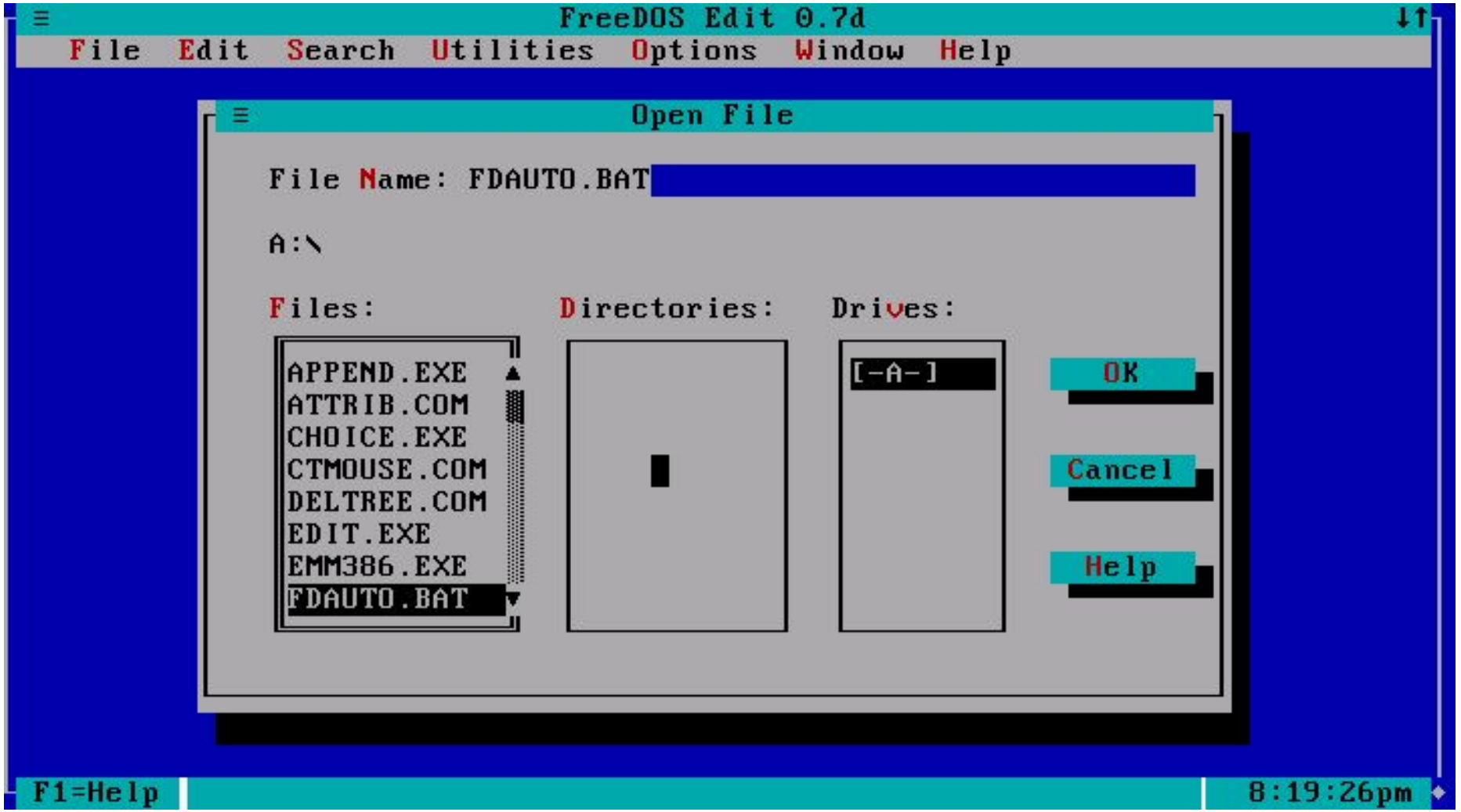
Типогра́фика

- – это художественное оформление текста посредством набора и вёрстки.
- MS Word – это немного другое…

Псевдографика



Псевдографика



Символы пунктуации

- апостроф ’ '
- скобки [ ], ( ), { }
- двоеточие :
- запятая ,
- тире –, -, —, —
- многоточие …, . . .
- восклицательный знак !
- точка .
- дефис -
- дефис-минус -
- вопросительный знак ? (см. также ¿)
- кавычки “ ”, " ", « », ‘ ’
- точка с запятой ;

Дефис – нет пробелов.
Тире – есть пробелы (кроме **числовых диапазонов**).

Всякие чёрточки

- Дефис
  - *hyphen*
  - кое-что
  - аудио- и видеофайлы
- Короткое тире
  - *N dash*
  - 2009–2016 гг.
- Длинное тире
  - *M dash*
  - Это — пример тире.
  - в мае — июле
  - теорема Гаусса — Остроградского
- Минус
  - *minus sign*
  - 1 − 2 = −1

CTRL + N.MINUS

(который справа)

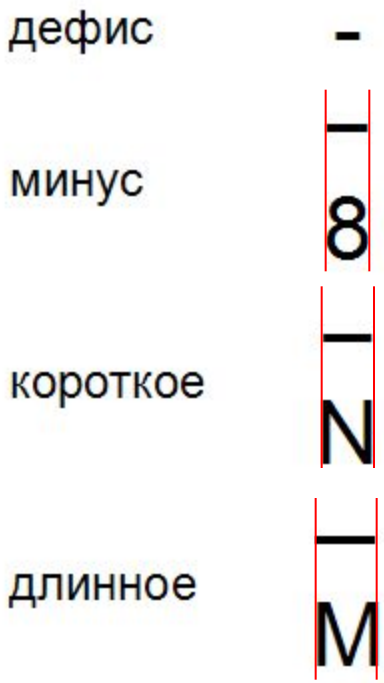
-	дефисоминус
‐	дефис
−	минус
‒	цифровая чёрточка
	символ переноса
⁃	чёрточный буллит
–	короткое тире
—	длинное тире
―	горизонтальная черта

Про тире у Лебедева.

Всякие чёрточки

А ещё есть:

- макрон
- перечёркивание
- подчёркивание
- верхняя черта
- тонкая граница
- толстая граница
- национальные значки



[Почитать на Хабре.]

Кавычки

- Клавиатурные "клавиатурные"

- Французские кавычки («ёлочки») «ёлочки»
- Немецкие кавычки («лапки») „лапки“
- Английские двойные кавычки “английские двойные”
- Английские одиночные кавычки ‘английские одиночные’
- Польские кавычки „польские кавычки”
- Шведские обратные кавычки »шведские«
- Китайский 『引號』
- Японский 「こんばんは」

- Правильно:

«„Цыганы“ мои не продаются вовсе», — сетовал Пушкин.

- Неправильно:

««Цыганы» мои не продаются вовсе», — сетовал Пушкин.

Ещё символы

- амперсанд &
- коммерческое at @
- звёздочка, астериск*
- косая черта, слэш, дробь *I*, */*
- обратная косая черта, обратный слэш \
- маркер списка, буллит •
- циркумфлекс ^
- крестик †, ‡
- градус °
- штрих ′, ″, ‴

23° 12′ с.ш.

но

+7 °C

- перевёрнутый восклицательный знак ¡
- перевёрнутый вопросительный знак ¿

Ещё символы

- октоторп, решётка, хэш #
- знак номера №
- знак деления ÷
- порядковый индикатор º, ª
- процент, промилле, миллионная доля %, ‰, ‱
- абзац ¶
- знак параграфа §
- тильда ~
- подчёркивание _
- вертикальная черта ¦, |

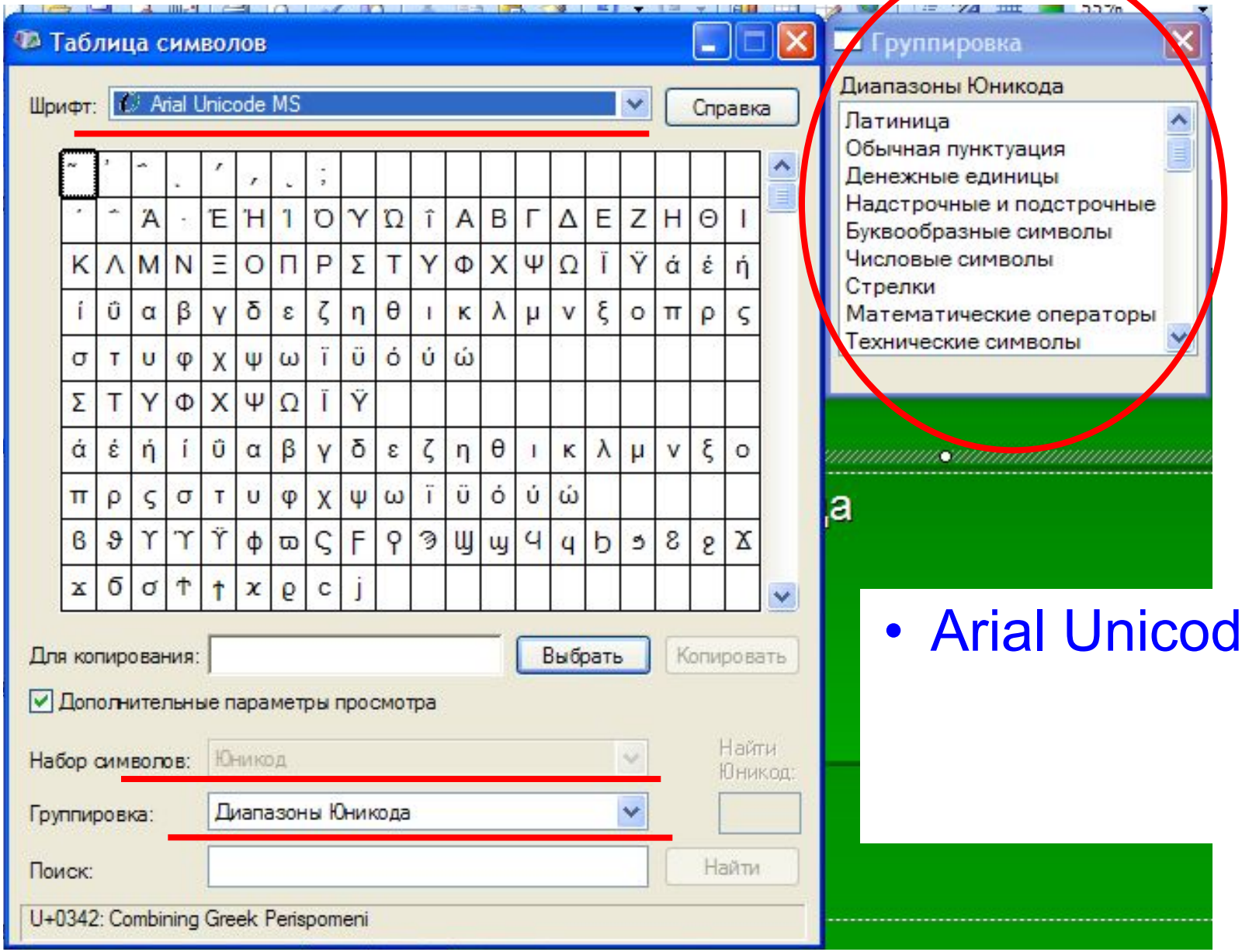
Интеллектуальная собственность

- знак охраны авторского права (©)
- знак правовой охраны товарного знака (®)
- символ знака обслуживания (℠)
- знак охраны смежных прав для фонограммы (℗)
- товарный знак(™)
- знак копилефта 

Способы вставки символов в текст

- Таблица символов
- *Вставка => Символ*
- ALT + код на дополнительной (справа) клавиатуре
- ALT + X после кода (в MS Word)
  - эта же комбинация после символа преобразует его в числовой код
- COMPOSE + код (Linux)
- ALT + код (MacOS; код другой, чем в Windows)

Таблица символов



- Arial Unicode MS

☝ Студенту на заметку

Символ alt-код

- ' Alt+39
- - Alt+45
- – Alt+0150
- — Alt+0151
- ˆ Alt+0136
- ¦ Alt+0166
- ¨ Alt+0168
- ¯ Alt+0175
- ´ Alt+0180
- ¸ Alt+0184
- ¿ Alt+168
- ˜ Alt+0152
- ‘ Alt+0145
- ’ Alt+0146
- “ Alt+0147
- ” Alt+0148
- „ Alt+0132
- ‹ Alt+0139
- › Alt+0155
- ± Alt+241
- « Alt+174
- » Alt+175
- × Alt+0215
- ÷ Alt+246
- √ Alt+251
- ∩ Alt+239
- ≈ Alt+247
- ≡ Alt+240
- ≤ Alt+243
- ≥ Alt+242

☝ Студенту на заметку

- ↑ Alt+24
- → Alt+26
- ↓ Alt+25
- ← Alt+27
- ↔ Alt+29
- ¢ Alt+155
- £ Alt+156
- ¤ Alt+0164
- ¥ Alt+157
- § Alt+21
- © Alt+0169
- ¬ Alt+170
- ® Alt+0174
- °Alt+248
- µ Alt+230
- ¶ Alt+20
- ·Alt+250
- † Alt+0134
- ‡ Alt+0135
- •Alt+249
- … Alt+0133
- ‰ Alt+0137
- ♠ Alt+6
- ♣ Alt+5
- ♥ Alt+3
- ♦ Alt+4
- € Alt+0128
- ¼ Alt+172
- ½ Alt+171
- ¾ Alt+0190
- ²Alt+253
- ³Alt+0179

☝ Студенту на заметку

- ∞ Alt+236
- ™ Alt+0153
- α Alt+224
- Γ Alt+226
- δ Alt+235
- εAlt+238
- Θ Alt+233
- π Alt+227
- Σ Alt+228
- σ Alt+229
- τAlt+231
- Φ Alt+232
- φ Alt+237
- Ω Alt+234

° Alt + 0176

± Alt + 0177

² Alt + 0178

³ Alt + 0179

Не забывая про NumLock и раскладку.

Диакритики

- По месту начертания
  - надстрочные
  - подстрочные
  - внутристрочные.
- По способу начертания
  - свободно приставляемые к основному знаку
  - требующие изменить и его форму.
- По фонетико-орфографическому значению
  - имеющие фонетическое значение
    - придающие букве новое звуковое значение
    - уточняющие варианты произношения какого-либо звука
    - указывающие на то, что буква сохраняет своё стандартное значение
    - просодические знаки
      - знаки долготы и краткости гласных
      - знаки музыкальных тонов
      - знаки ударения

Диакритики

  - имеющие только орфографическое значение
    - позволяющие избегать омографию
    - использующиеся по традиции
  - знаки иероглифического значения
    - указывающие на сокращённое или условное написание
    - знаки, указывающие на применение букв для других целей
- По формальному статусу
  - знаки, с помощью которых образуются новые буквы алфавита
  - знаки, сочетания букв с которыми не считается отдельной буквой
- По обязательности использования
  - знаки, отсутствие которых делает текст орфографически неверным, а иногда и нечитаемым
  - знаки, используемые только в особых обстоятельствах: в книгах для начального обучения чтению, в священных текстах, в редких словах с неоднозначным чтением и т. п.

Диакритики

знак ударения,
код ALT + 0769

акут	´	*sześć*
двойной акут	˝	Felhőszakadás
гравис	`	кля̀твопреступле́ние
двойной гравис	̏	
кратка (бреве)	˘	паўночна-*ўсходні*
перевёрнутая кратка	̑	
гачек / háček	ˇ	
седиль / cédille	¸	Française
циркумфлекс	ˆ	limba română
умлаут	¨	schön

Диакритики

точка .

крючок / dấu hỏi ̉

рожок / dấu móc ̛

макрон ̄

огонэк / nosinė ̨

кружок / kroužek ̊

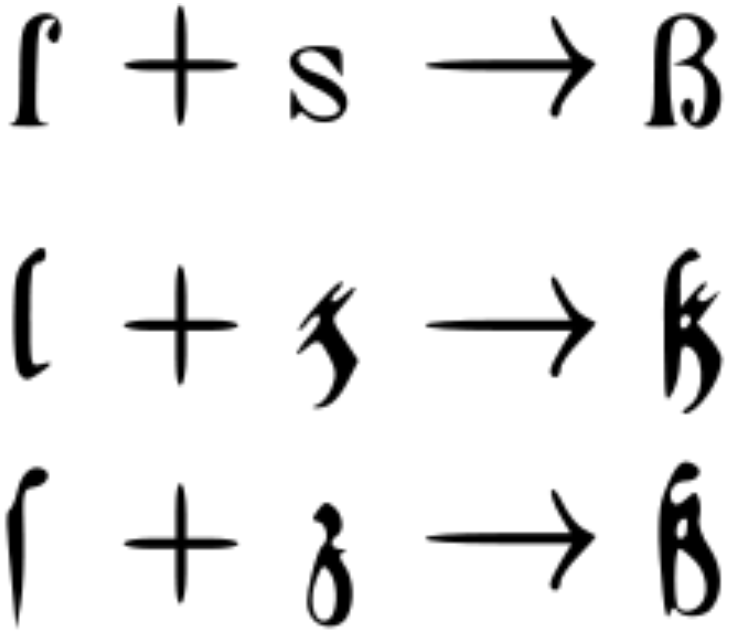
густое придыхание / дасия ̔

тонкое придыхание / псили ̓

Tiếng Việt, còn gọi tiếng Việt Nam hay Việt ngữ, là ngôn ngữ của người Việt (người Kinh) và là ngôn ngữ chính thức tại Việt Nam. Đây là tiếng mẹ đẻ của khoảng 85% dân cư Việt Nam, cùng với hơn bốn triệu người Việt hải ngoại Tiếng Việt còn là ngôn ngữ thứ hai của các dân tộc thiểu số tại Việt Nam.

Эсцет

- ß
- ~ ss



Шрифты

Serif ~ с засечками
Sans serif ~ без засечек

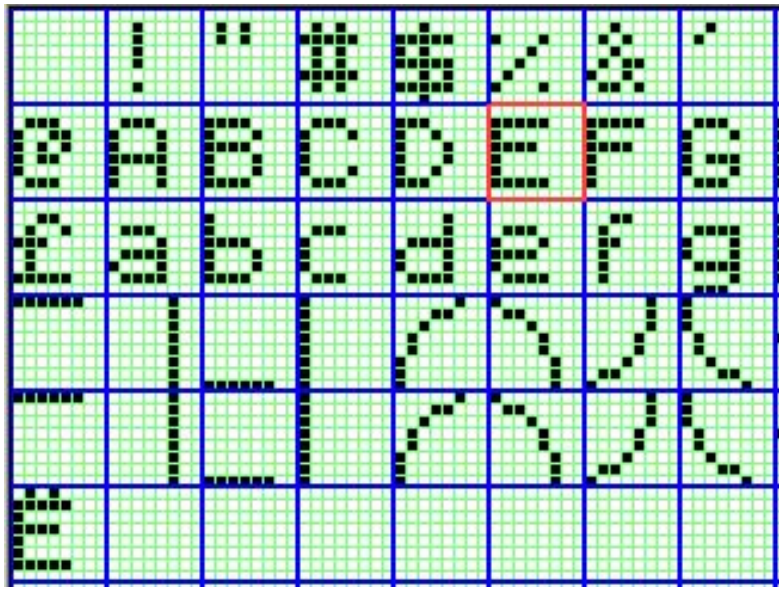
- группа
  - готические
  - с/без засечек
- начертание
  - прямой
  - курсивный;
- насыщенность
  - светлый
  - полужирный
  - жирный
- ширина
  - нормальный
  - узкий
  - широкий
  - шрифт фиксированной ширины;

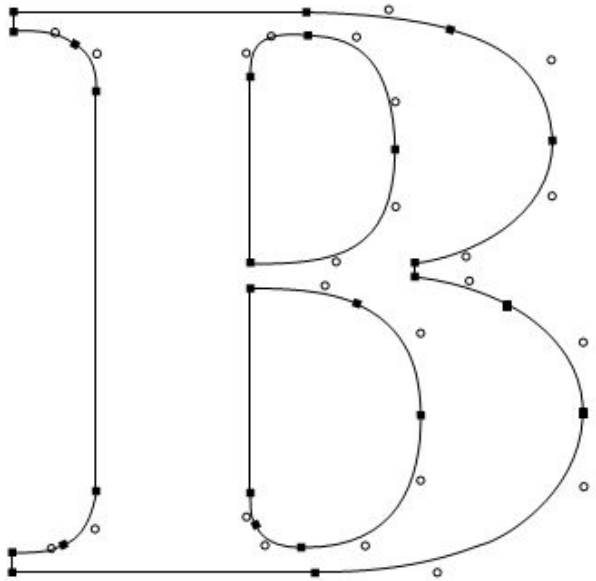
- размер (кегль)
  - 1 пункт = 1/72 дюйма
    - **0.376 мм** (типометрическая система Дидо)
    - **0.3528 мм** (компьютерный, Adobe)
- чёткость
- контраст
- различимость
- удобочитаемость
- ёмкость

Терминология

Компьютерные шрифты

- Тип
  - растровые
  - векторные
    - TrueType
- Ширина символа
  - моноширинные
    - Courier
  - пропорциональные





Моноширные шрифты

```
E|--5---5-5-------------------------|
B|---------------3---3-3---5-3-2--|
G|-------3---3-3---------------------|
D|-----------------------------------|
A|-----------------------------------|
E|-----------------------------------|
```

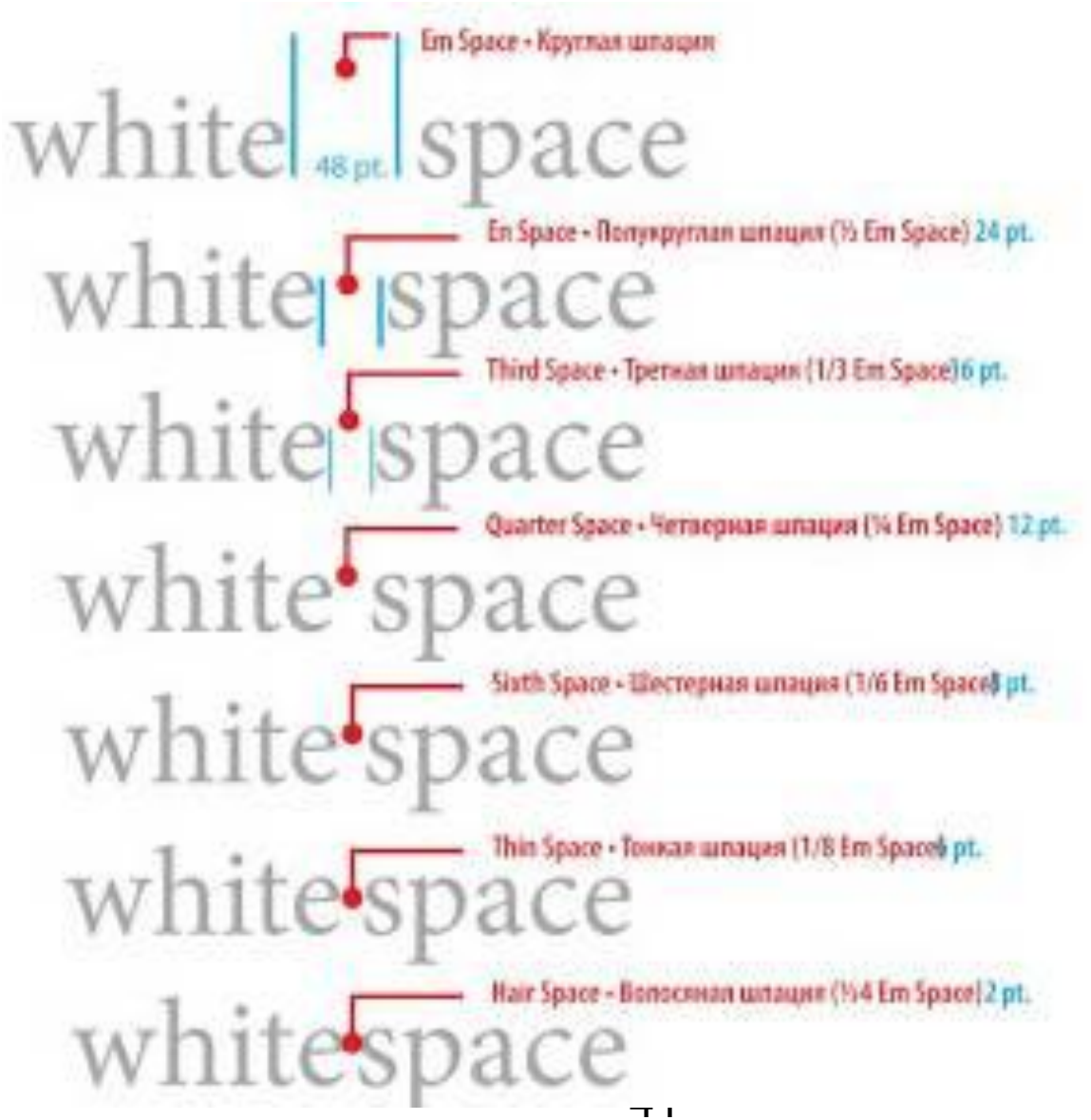
Типогра́фика

Меняются:

- гарнитура
- кегль
- длина строк и расстояния между ними (интерлиньяж)
- изменение пробелов между буквами (кернинг)
- и группами букв (трэкинг)
- отступы и выступы

Кошмарная пустота пробелов…

- ***Em space*** (Круглая шпация, *Em quad*, Mutton, Кегельная шпация) – пробельный элемент, высота и ширина которого равна кеглю (иначе букве M).
- ***En space*** (Полукруглая шпация, Полукегельная, *En quad*) - ширина равна половине *Em space*, то есть половине кегля данного шрифта (примерно равна букве N или n).
- ***Third Space*** (Третная шпация) - ширина 1/3 кегля.
- ***Quarter Space*** (Четверная шпация) - ширина 1/4 кегля.
- ***Sixth Space*** (Шестерная шпация) - ширина 1/6 кегля.
- ***Thin space*** (Тонкая шпация) - ширина составляет 1/8 кегля.
- ***Hair Space*** (Волосяная шпация (1/24 *Em Space*)) - ширина составляет 1/24 кегля.



Пустота продолжается…

- ***Figure Space*** - имеет такую же ширину, что и цифры в данном шрифте, и предназначен для набора таблиц. Неразрывный.
- ***Punctuation Space*** - ширина равна ширине точки.
- ***Zero-width Space*** - показывает места, в которых можно разрывать строку, не добавляя знак переноса; ширина его нулевая. Применяется в языках, в которых пробелов нет.
- ***Narrow No-break Space*** - узкий неразрывный пробел.
- ***Medium Mathematical Space*** - узкий пробел, применяемый в математических формулах.
- ***Word Joiner*** - аналогичен *Zero-width Space*, но неразрывный.
- ***Ideographic Space*** - используется в восточных языках, равняется ширине одного иероглифа.

Неразрывность

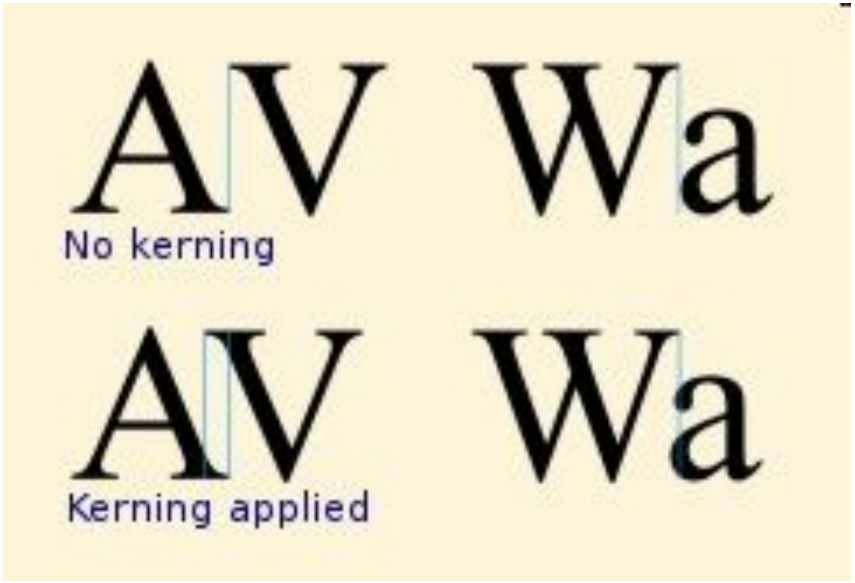
- Неразрывный пробел
  - * *
  - CTRL+SHIFT+SPACE (в MS Word)
  - требуется
    - при инициалах, в сокращениях
    - при числах перед и после единиц измерений, или знаками номера, и т.п.
    - перед тире
    - в числах как разделитель
    - после предлогов и союзов
- Неразрывный дефис
- Мягкий перенос
  - **
  - ¬

Трэкинг и кернинг

разреженный набор +50
нормальный набор 0
плотный набор -50

- разрядка
- вгонка
- выгонка

Tracking:
VAST. V A S T .
Kerning:
VAST. VAST.



Хинтинг

abcfgop AO *abcfgop*
abcfgop AO ***abcfgop***

Кодировки текста

- Управляющий символ (*ucm.*)
- Однобайтовая кодировка
- Двубайтовая (и более) кодировка
- Смешанная кодировка

ASCII

- *American Standard Code for Information Interchange*
- 8 (7) бит
- Цифры 0..9 представляются своими двоичными значениями, перед которыми стоит 0011b
  - «5» ~ 0011 0101b ~ 35h ~ символ № 53
- Буквы A-Z верхнего и нижнего регистров различаются в своём представлении только одним битом; представляются своими порядковыми номерами в алфавите, перед которыми стоит 100b (верхний регистр) или 110b (нижний регистр).
  - «F» ~ 0100 0110b ~ 46h ~ № 70
- Основа для многих последующих и национальных кодировок.
- Используются программно и аппаратно.

8-битные кодировки

- KOI-8
  - уЯЕЫШ ЕЭЈ ЬФЙИ НСЗЛЙИ ЖТБОГХЪУЛЙИ ВХМПЛ, ДБ ЧЩРЕК ЮБА
- CP1251
  - Съешь ещё этих мягких французских булок, да выпей чаю
- ISO866
  - ‘кЃим Ґйс нвЁе ¬пЈЄЁе да жг§бЄЁе Ўг«®Є, ¤ ўлЇҐ© з о
  - ▒к╔им ╔йс нвЁе ╡ пё╚Ёе да═╩ жг╦ б╚Ёе ╟г╙ ╚, ╔═ ╔л╞╔ ╘ з═о
- MacCyrillic
  - ‘ъешь ещЮ этих мЯгких французских булок, да выпей чаю
- ISO 8859-1
  - Ñúåøü åù¸ ýòèõ ìÿãêèõ ôðàíöóçñêèõ áóëîê, ää âûïåé ÷àþ
- EBCDIC

Юникод *Unicode*

- универсальный набор символов (UCS, *universal character set*)
  и
- семейство кодировок (UTF, *Unicode transformation format*)

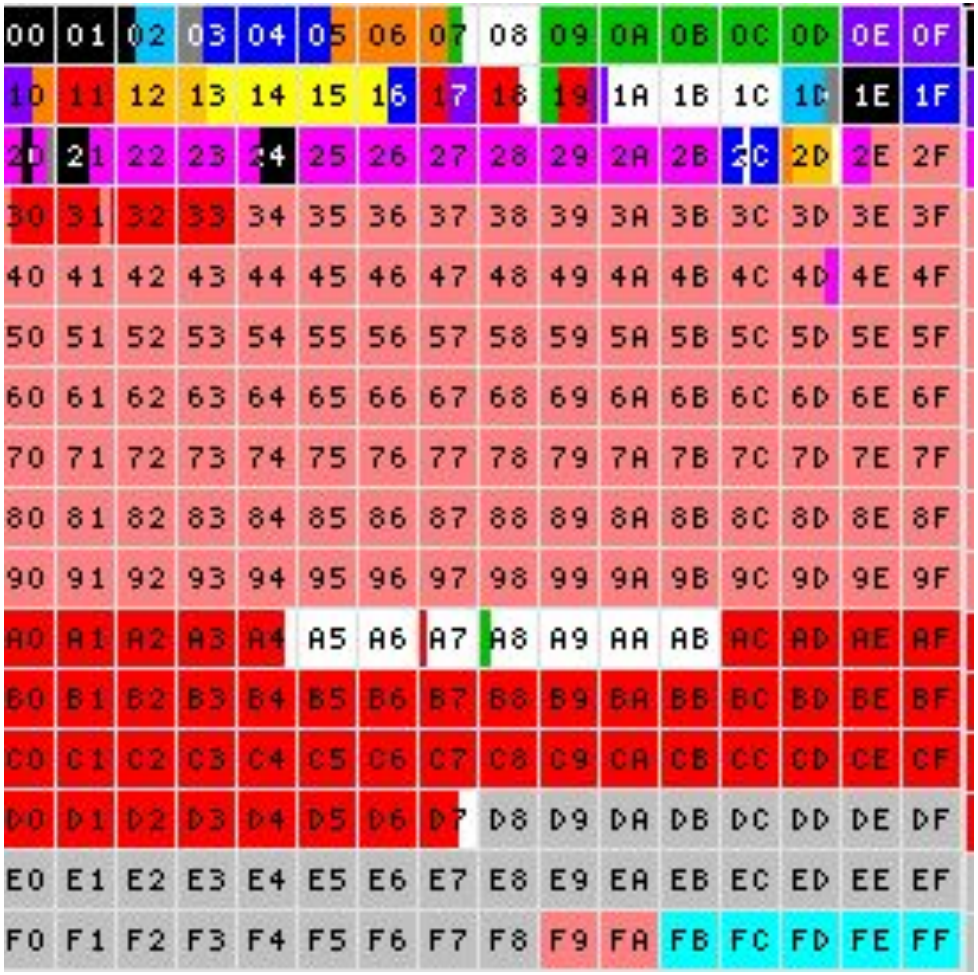
- 1-6 байт
  - UTF-8
  - UTF-16
  - UTF-32
- всего – 1 112 064 (= $2^{20} + 2^{16} - 2^{11}$) символов
  - хотя можно было бы и 2^{32}
- до 1FFFFFh

- символы:
  - базовые (base characters)
  - модифицирующие (combining characters)
  - селекторы варианта начертания (variation selectors)

- обозначения: U+04F0

Структура Юникода

- Плоскости (17 по 2^{16} символов)
  - Плоскость 0 (0000—FFFF): Базовая многоязыковая плоскость (Basic Multilingual Plane, BMP)
  - Плоскость 1 (10000—1FFFF): Дополнительная многоязыковая плоскость (Supplementary Multilingual Plane, SMP)
  - Плоскость 2 (20000—2FFFF): Дополнительная иероглифическая плоскость (Supplementary Ideographic Plane, SIP)
  - Плоскость 3 (30000—3FFFF): Третичная иероглифическая плоскость (Tertiary Ideographic Plane, TIP)
  - Плоскости 4—13 (40000—DFFFF) не используются
  - Плоскость 14 (E0000—EFFFF): Дополнительная плоскость особого назначения (Supplementary Special-purpose Plane, SSP)
  - Плоскость 15 (F0000—FFFFF) используется как дополнительная область-A для частного использования (Supplementary Private Use Area-A, SPUA-A)
  - Плоскость 16 (100000—10FFFF) используется как дополнительная область-B для частного использования (Supplementary Private Use Area-B, SPUA-B)



Базовая плоскость UNICODE

Чёрный — расширенный латинский алфавит;
Голубой — лингвистические символы международного фонетического алфавита IPA;
Синий — другие европейские алфавиты;
Оранжевый — письменности Ближнего Востока;
Светло-оранжевый — письменности Африки;
Зелёный — письменности Южной Азии;
Фиолетовый — письменности Юго-восточной Азии;
Красный — письменности Восточной Азии;
Розовый — унифицированные китайско-японско-корейские символы;
Жёлтый — письменности аборигенов Северной Америки;
Пурпурный — символы;
Тёмно-серый — диакритики;
Светло-серый — суррогатные пары UTF-16 и области для частного использования;
Циан — другие знаки;
Белый — не используется.

Структура Юникода

- Управляющие символы C0 (0000—001F)
- Основная латиница (0020—007F)
- Управляющие символы C1 (0080—009F)
- Дополнительные символы Latin-1 (00A0—00FF)

- Кириллица:
  - U+0400 до U+052F
  - U+2DE0 до U+2DFF
  - U+A640 до U+A69F

UTF-16

- 2 или 4 байта (т.е. 1 или 2 слова)
  - U+0000..U+D7FF и U+E000..U+10FFFF
- метка порядка байтов (*Byte order mark*) U+FEFF
- UCS-2 (*ycm.*)

- Объём:
  - 2^20 – для 2-сложного варианта
    +
  - 2^16 – односложные символы
    -
  - 2^11 – служебный диапазон D800h..DFFFh
    =
  - 1 112 064 символов всего

Суррогатные пары в UTF-16

- U+10437
- 10437h - 10000h = 00437h =

0000 0000 0100 0011 0111b (*20 бит*)
0001h 0037h

- 110110000000000b = D800h (*1-е слово*)
- 110111000000000b = DC00h (*2-е слово*)

- D800h + 0001h = D801h
- DC00h + 0037h = DC37h

- = D801DC37h (*4-байтное слово*)

UTF-8

- 1-6 байт
- самосинхронизирующийся
- может иметь BOM
  - 0xEF 0xBB 0xBF
  - п»ї
- РЎСЛЪРµСЗСНЬ РµС‰С‘ СЌС‚РёС… РјСЏРіРєРёС… С„СЂР°РSС†СґР·СЃРєРёС… Р±СґР»РsРє, РґР° РІС‹РїРµР№ С‡Р°СЂ

UTF-8: самосинхронизация

- (1 байт, 7 бит данных) 0ааа аааа
- (2 байта, 1 байт данных) 110х хххх 10хх хххх
- (3 байта, 2 байта данных) 1110 хххх 10хх хххх 10хх хххх
- (4 байта, ~2.5 байта данных) 1111 0ххх 10хх хххх 10хх хххх 10хх хххх
- (5 байт, 3.25 байт данных) 1111 10хх 10хх хххх 10хх хххх 10хх хххх 10хх хххх
- (6 байт, 31 бит данных) 1111 110х 10хх хххх 10хх хххх 10хх хххх 10хх хххх 10хх хххх

Пример UTF-8

- Код BOM:

  $FEFF_{16}$

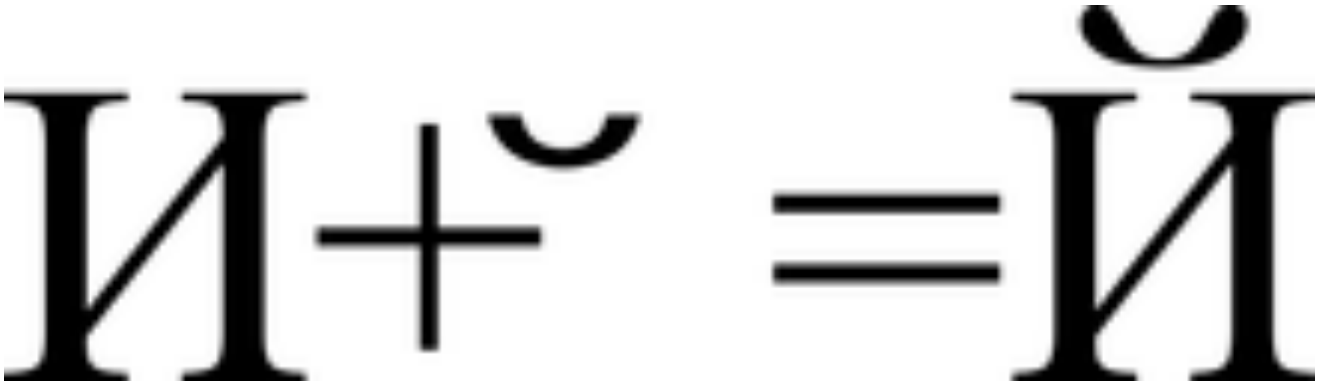
  $1111\ 1110\ 1111\ 1111_2$

	1й байт	*2й байт*	*3й байт*
	1110 xxxx	10xx xxxx	10xx xxxx
bin	1110 1111	1011 1011	1011 1111
hex	EF	BB	BF

Проблемы Юникода

- одинаковые буквы разных языков
- отсутствует вертикальная разметка
- иероглифы одинаковы (CJK-унификация)
- разные правила заглавных букв, цифр
- размер текста, производительность
- несовместимость со старым ПО
- повторы и похожие символы

Формы нормализации



7-битные кодировки

- Используются:
  - символы (A—Z, a—z)
  - цифры (0—9)
  - символы «+», «/», «=»
- Quoted printable
  - Это пример Quoted Printable
    ~
  - =D0=AD=D1=82=D0=BE
    =D0=BF=D1=80=D0=B8=D0=BC=D0=B5=D1=80 Quoted Printable
- Base64
- UTF-7
- UUE

Percent-encoding

- !# $ & ' ( ) * +
- %21%23 %24 %26 %27 %28 %29 %2A %2B

https://ru.wikipedia.org/wiki/%D0%9D%D0%B0%D0%B1%D0%BE%D1%80_%D1%81%D0%B8%D0%BC%D0%B2%D0%BE%D0%BB%D0%BE%D0%B2

=

https://ru.wikipedia.org/wiki/Набор_символов

В HTML

- коды
  - *Фыва ~ Фыва*
- мнемоники
  - *à → à*
  - *α → α*
  - *< → <*
  - *> → >*
  - * → (неразрывный пробел)*

Связанные понятия

- транслит
  - Eto zapis translitom.
  - ГОСТ 7.79-2000
- leet speak
  - Et0 l33t $pea|<.
- локаль
- раскладка клавиатуры
  - типографская раскладка Ильи Бирмана
- пуникод
  - почемужеонинеговорятпорусски
  - b1abfaaepdrnnbgefbaDotcwatmq2g4l
- глиф
- ASCII art
- буквы Н и Я
  - сеть FidoNet


```
(\____/)
(='.'=)
(")_(")
```

Массив символов

0	1	2	3	4	5	6
Длина	*S*	*t*	*r*	*i*	*n*	*g*
6	53h	74h	72h	69h	6Eh	67h

- *Pascal Strings*

Преимущества

- программа в каждый момент времени «знает» о размере строки, и операции добавления символов в конец, копирования и получения размера строки выполняются достаточно быстро;
- строка может содержать любые данные;
- возможно на программном уровне следить за выходом за границы строки при её обработке;
- возможно быстрое выполнение операции вида «взятие N-ого символа с конца строки».

Недостатки

- проблемы с хранением и обработкой символов произвольной длины;
- увеличение затрат на хранение строк — значение «длина строки» также занимает место и в случае большого количества строк маленького размера может существенно увеличить требования алгоритма к оперативной памяти;
- ограничение максимального размера строки (32-битное поле длины даёт 4 294 967 295 байт символов).

Нуль-терминированная строка

- *ASCIIZ*
- *C-strings*
- *zero-terminated*

0	1	2	3	4	5	6
S	*t*	*r*	*i*	*n*	*g*	NUL
53h	74h	72h	69h	6Eh	67h	00h

Преимущества

- отсутствие дополнительной служебной информации о строке (кроме завершающего байта);
- возможность представления строки без создания отдельного типа данных;
- отсутствие ограничения на максимальный размер строки;
- экономное использование памяти;
- возможность использовать алфавит с переменным размером символа.

Недостатки

- долгое выполнение операций получения длины и конкатенации строк;
- отсутствие средств контроля за выходом за пределы строки, например, в случае повреждения завершающего байта;
- невозможность использовать символ завершающего байта в качестве элемента строки.

Операции со строками

Простейшие операции со строками

- получение символа по номеру позиции (индексу);
- конкатенация (соединение) строк.

Производные операции

- получение подстроки по индексам начала и конца;
- проверка вхождения одной строки в другую (поиск подстроки в строке);
- проверка на совпадение строк (с учётом или без учёта регистра символов);
- получение длины строки;
- замена подстроки в строке.

Операции со строками

Операции при трактовке строк как списков

- свёртка;
- отображение одного списка на другой;
- фильтрация списка по критерию.

Более сложные операции

- нахождение минимальной надстроки, содержащей все указанные строки;
- поиск в двух массивах строк совпадающих последовательностей (задача о плагиате).

Возможные задачи для строк на естественном языке

- сравнение на близость указанных строк по заданному критерию;
- определение языка и кодировки текста на основании вероятностей символов и слогов.

Расстояние Левенштейна

- расстояние Дамерау — Левенштейна
- действия:
  - D (*delete*) — удалить
  - I (*insert*) — вставить
  - R (*replace*) — заменить
  - M (*match*) — совпадение
  - обмен

Пример:

M	M	M	I	R	M	R	R
C	O	N		N	E	C	T
C	O	N	E	H	E	A	D

Связанные понятия

- Пустая строка
  - "" != NULL
- Нотация выражений
  - польская: 3 1 + 2 *
  - sudo -V | -h | -l | -L | -p запрос] [-c класс|-] [-a тип_аутентификации] [-u имя_пользователя/#uid] команда
- Маска и формат символов
  - ###.##
  - DD.MM.YYYY HH:MM
- Регулярные выражения
  - /Пункт [АБВ]*\.\n.*>/

Панграмма

- Любя, съешь щипцы, — вздохнёт мэр, — кайф жгуч.
- Шеф взъярён тчк щипцы с эхом гудбай Жюль.
- Эй, жлоб! Где туз? Прячь юных съёмщиц в шкаф.
- Экс-граф? Плюш изъят. Бьём чуждый цен хвощ!
- Эх, чужак! Общий съём цен шляп (юфть) — вдрызг!
- Эх, чужд кайф, сплющь объём вши, грызя цент.
- Чушь: гид вёз кэб цапф, юный жмот съел хрящ.

- Съешь [же] ещё этих мягких французских булок, да выпей чаю.

- The quick brown fox jumps over the lazy dog.
- The five boxing wizards jump quickly.

Lorem ipsum

- Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Блок 2

Цвет в биологии и в информатике

☝ Студенту на заметку

- Сайт «Цветофобия»: igor-bon.narod.ru/
- Основы теории цвета. Система CIE XYZ

У животных

- 4 цвета (+ УФ)
  - птицы
- 3 цвета
  - многие приматы
  - сумчатые
- 2 цвета
  - б.ч. все
- 1 цвет
  - водные млекопитающие
  - ночные
  - хищники

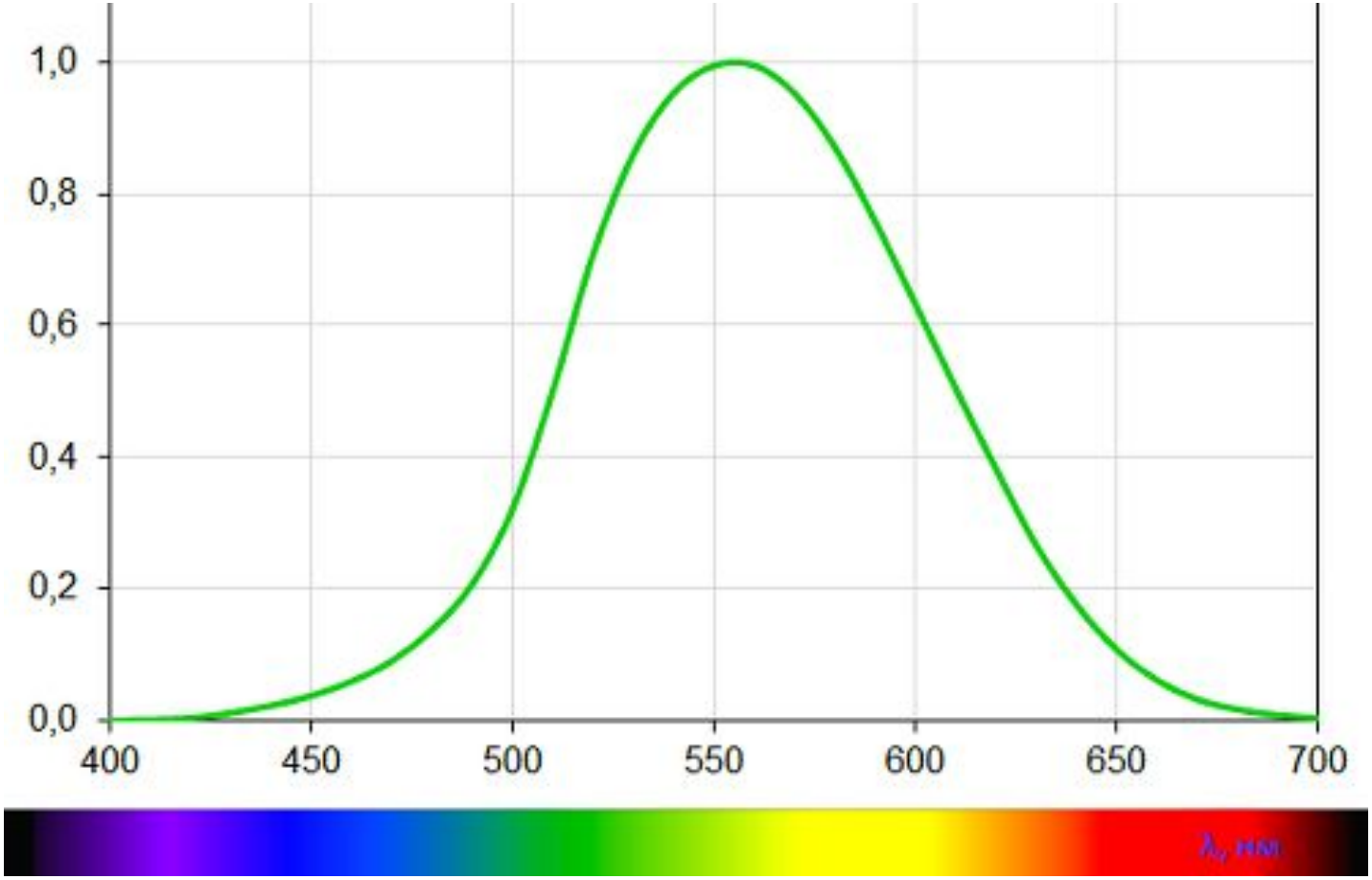
Цветовое восприятие человека

- Колбочки

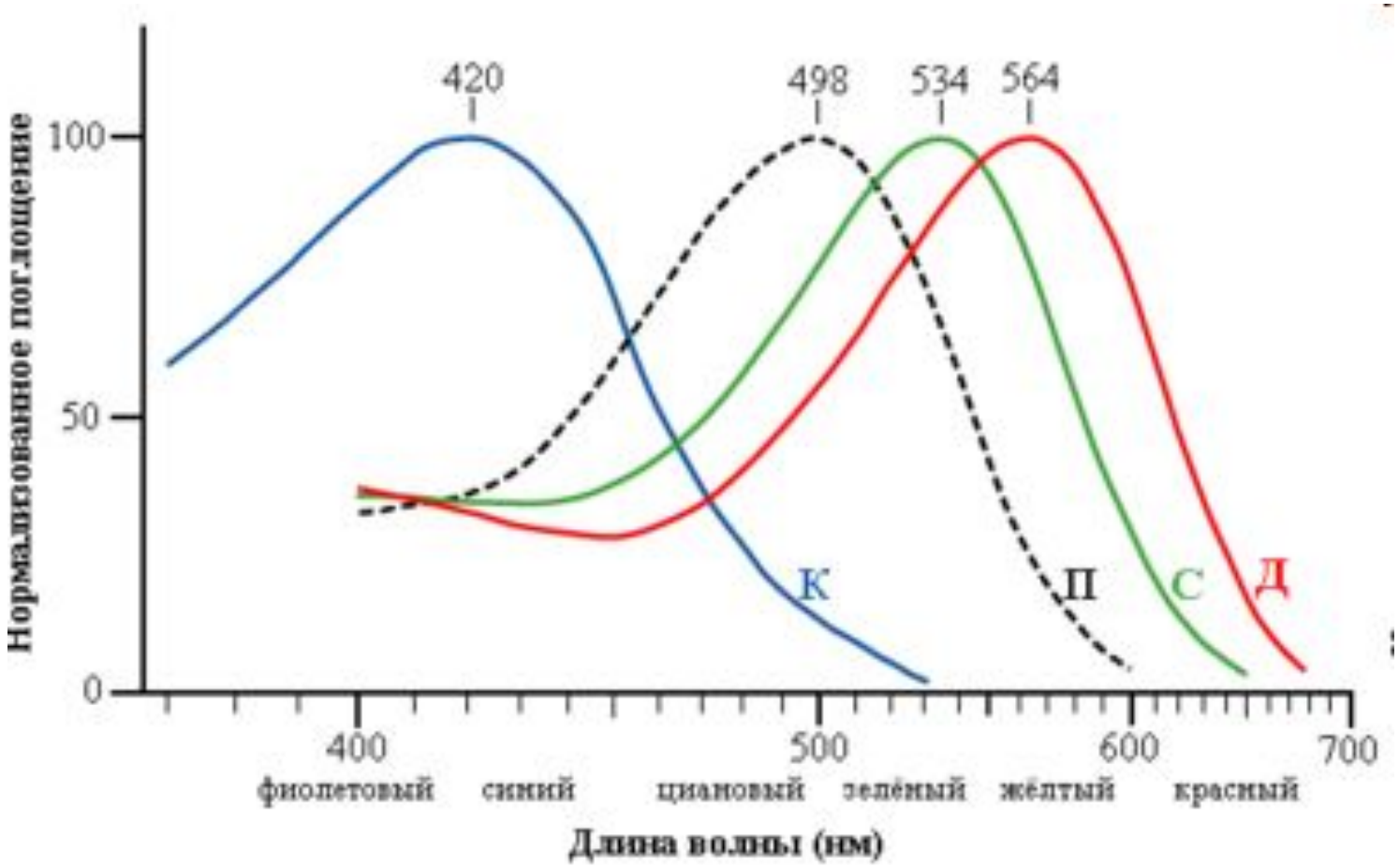
Название Максимум Цвет

- S, β 443 nm синий, фиолетовый
- M, γ 544 nm зелёный
- L, ρ 570 nm красный

Кривая видности



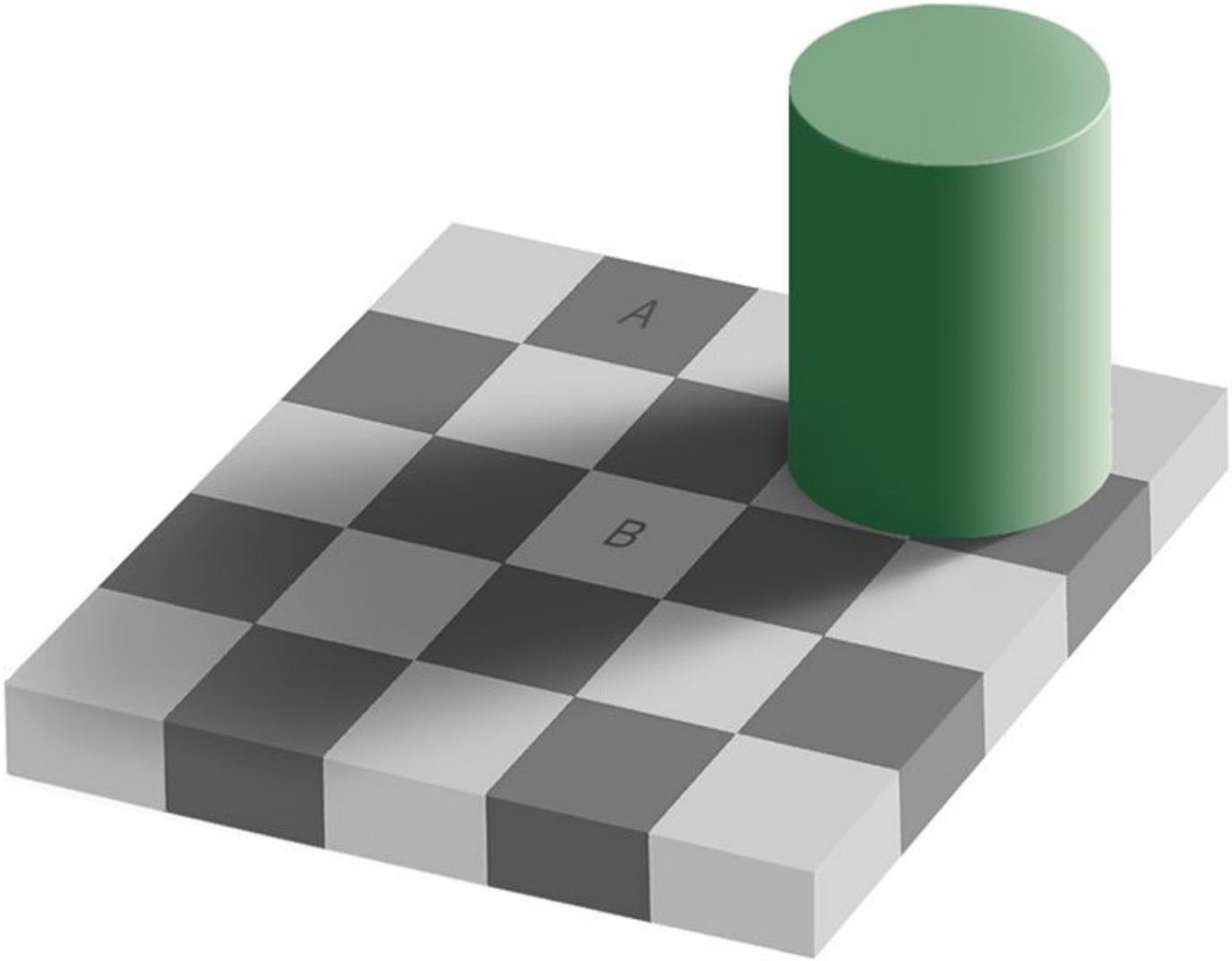
Чувствительность разных типов рецепторов



Связанные понятия

- цветовое пространство
- метамерия
  - источника света
  - наблюдателя
  - объекта
- дополнительные цвета

Пример иллюзии

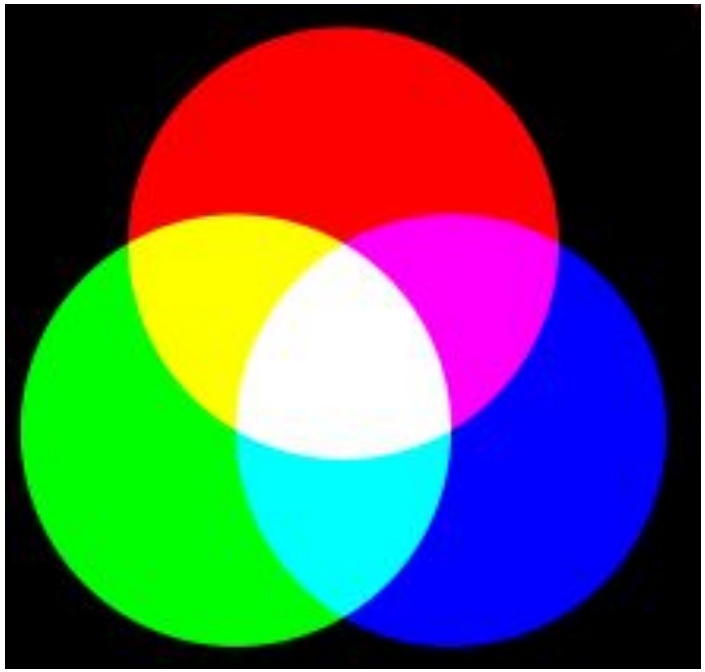


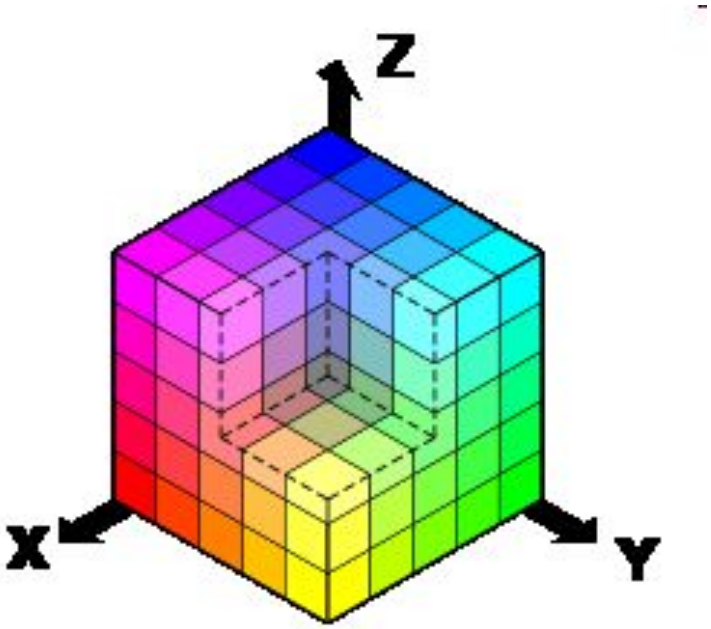
Цветовые модели

- Описание восприятия человеком
  - LMS и XYZ
  - L*a*b* — то же пространство в других координатах
- Аддитивные модели — сложение цветов
  - получение цвета на мониторе (например, RGB)
- Субтрактивные модели
  - полиграфические модели (например, CMYK)
- Модели, не связанные с физикой оборудования, являющиеся стандартом передачи информации
- Математические модели
  - полезные для каких-либо способов цветокоррекции, но не связанные с оборудованием, например HSV
- Табличные модели (*Pantone*)

Модель RGB

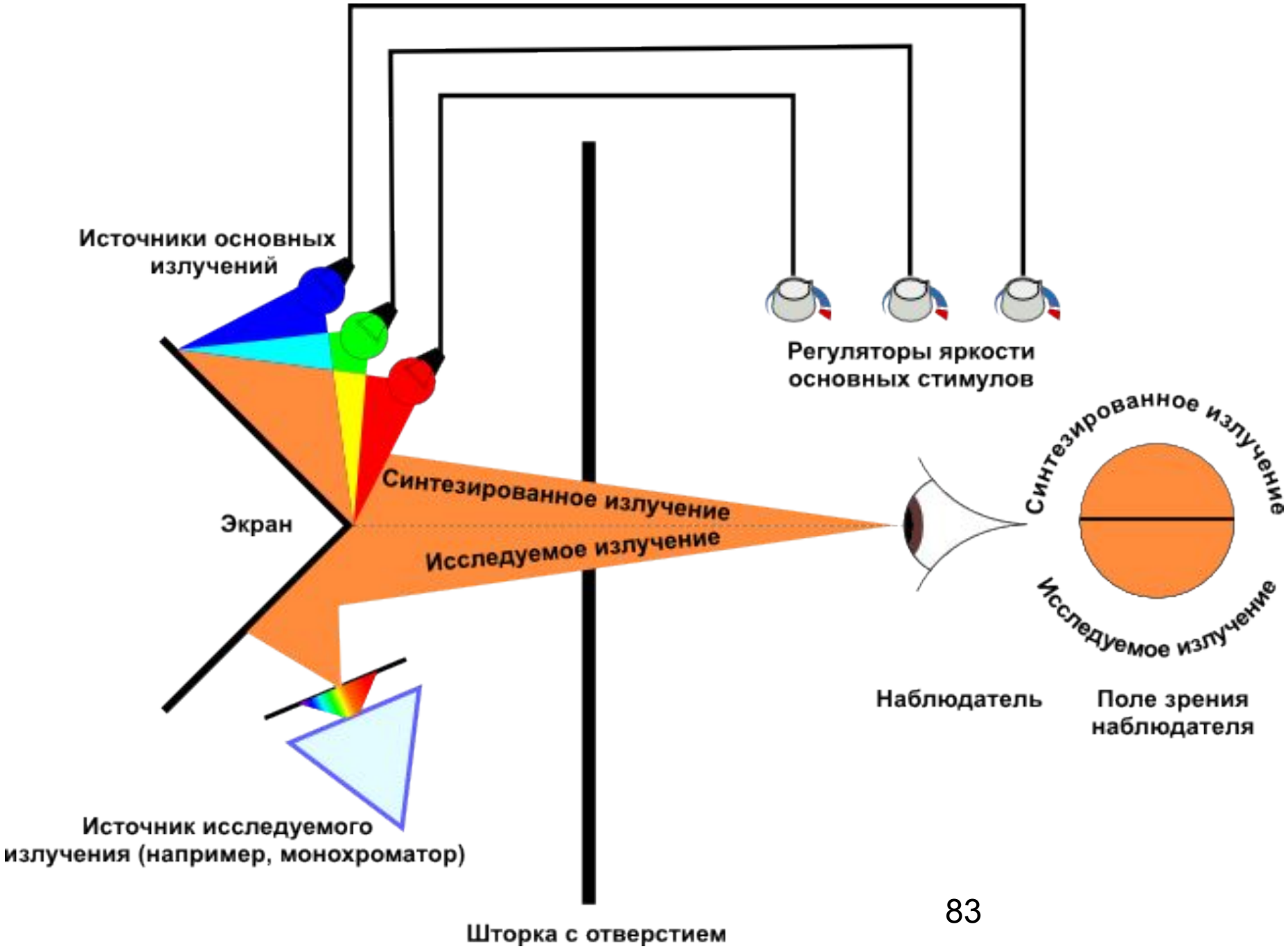
- аддитивная модель
- *Red*, *Green*, *Blue* — красный, зелёный, синий
- самая простая, часто встречается



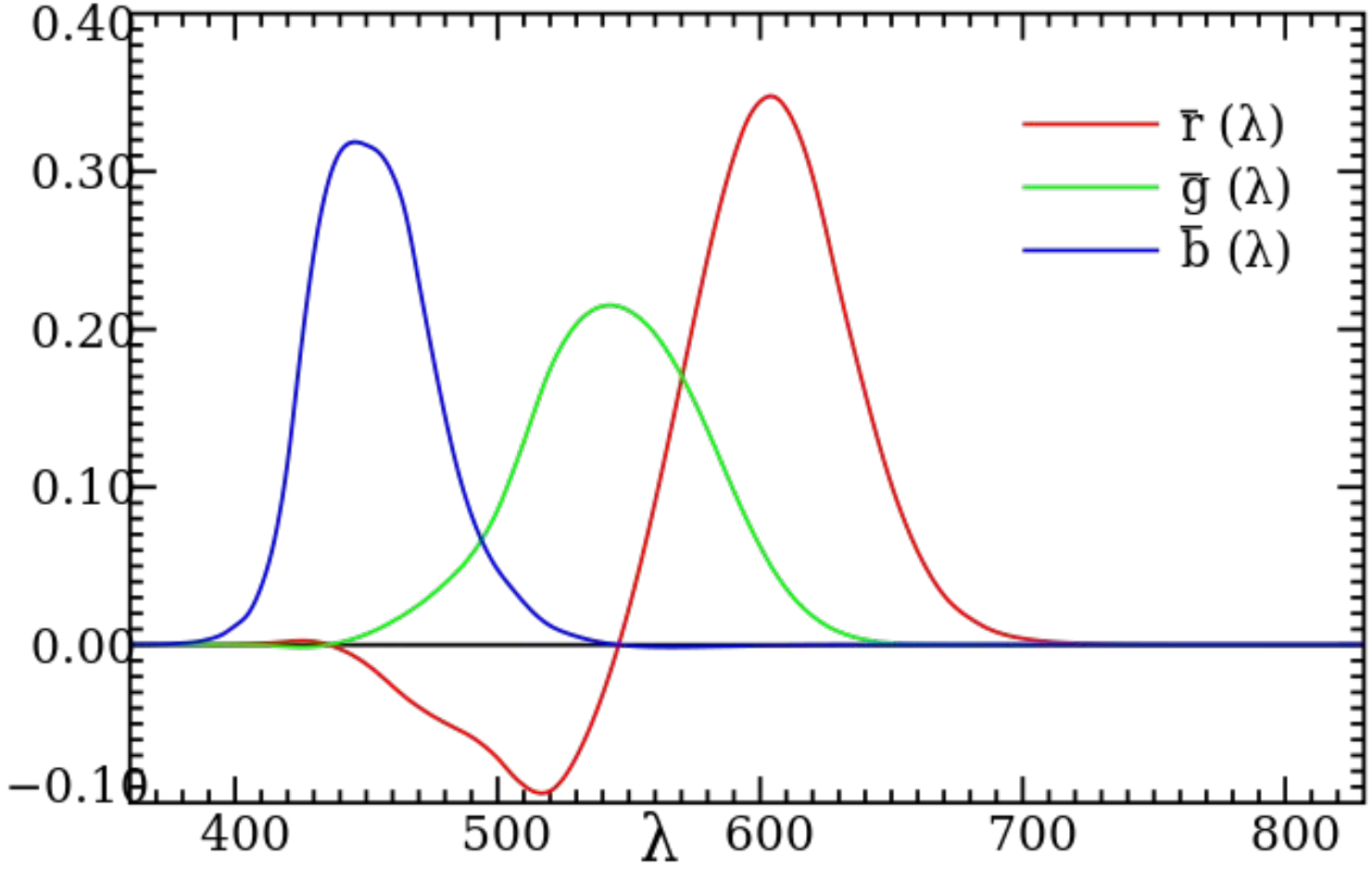


История RGB

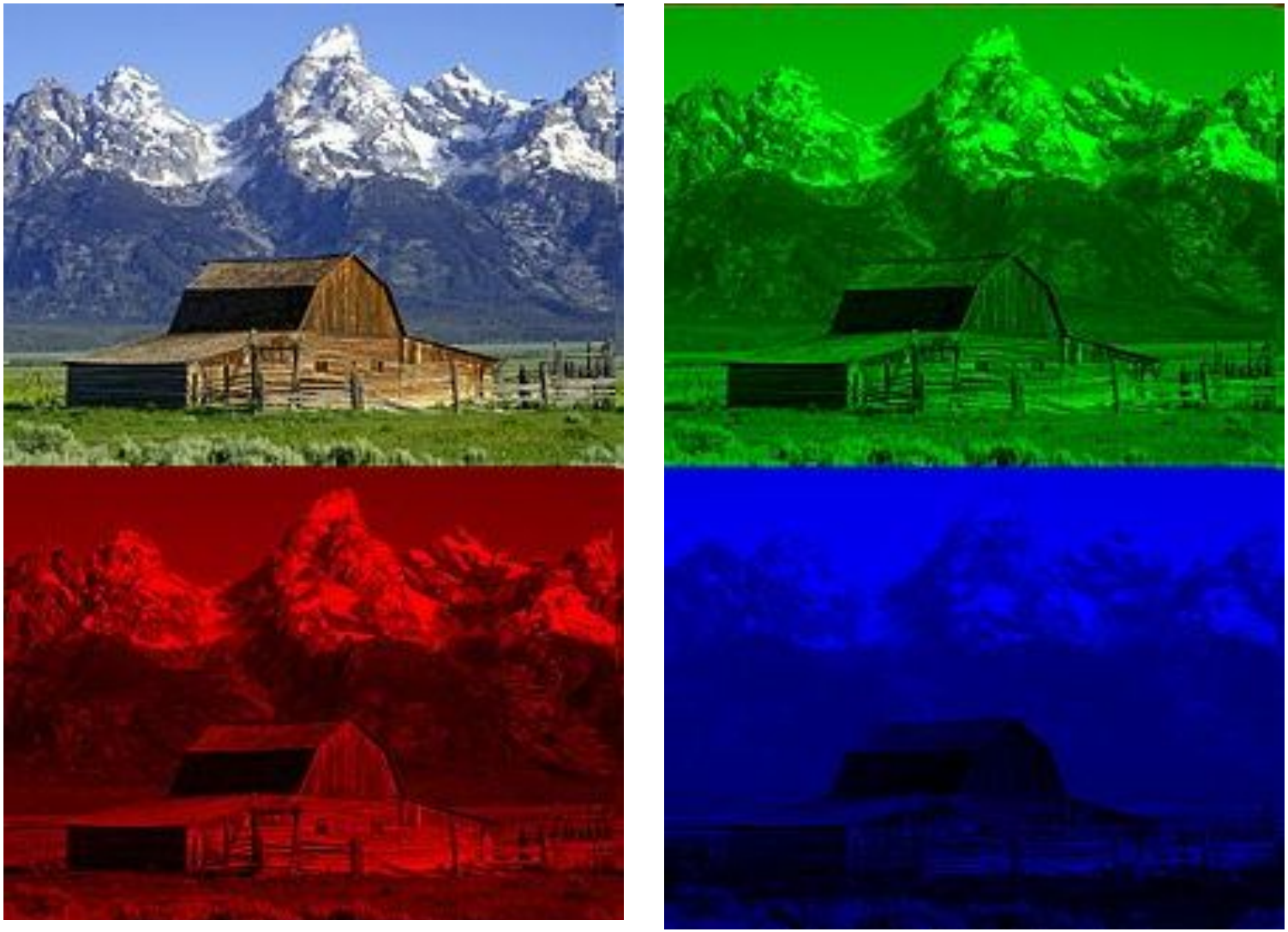
- 20-е годы XX века
- независимо друг от друга Джон Гилд (*John Guild*) и Дэвид Райт (*David Wright*)
- Международная комиссия по освещению (МКО), или CIE — *Commission Internationale de l'Éclairage*.


Источники основных излучений
Регуляторы яркости основных стимулов
Синтезированное излучение
Экран
Исследуемое излучение
Синтезированное излучение
Исследуемое излучение
Наблюдатель
Поле зрения наблюдателя
Источник исследуемого излучения (например, монохроматор)
Шторка с отверстием

Синтез цвета



Пример RGB



Матрица ЖК монитора

Альфа-канал

- яркость = фон + (лицо – фон) * альфа

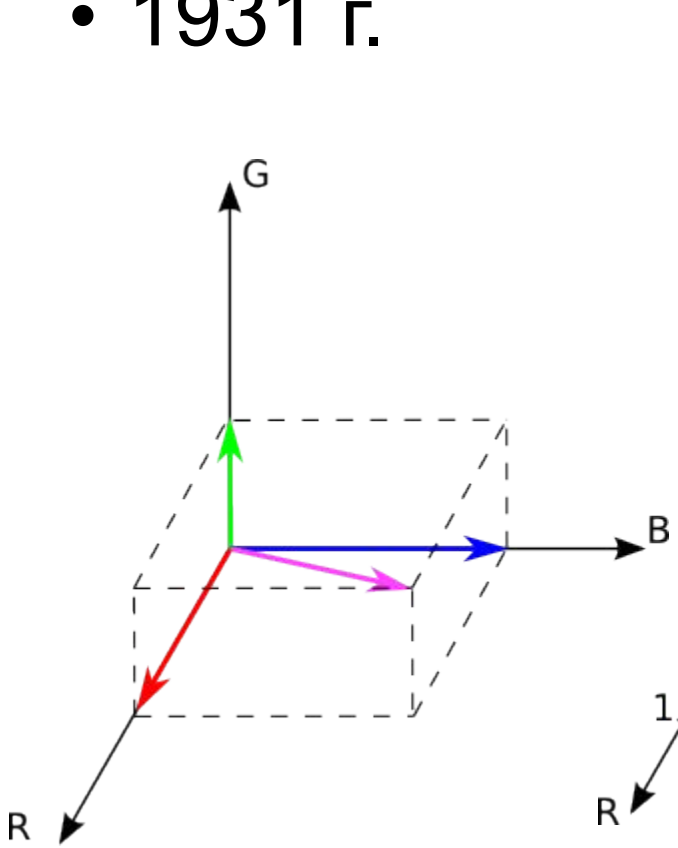
Числовое представление

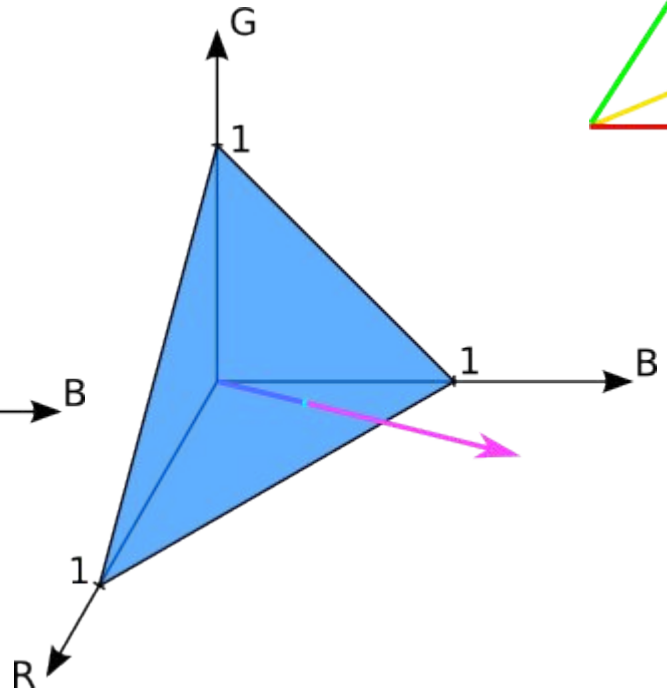
- #RrGgBb
  - HTML
- 0xaabbggrr
  - WinAPI
  - a – прозрачность либо гамма

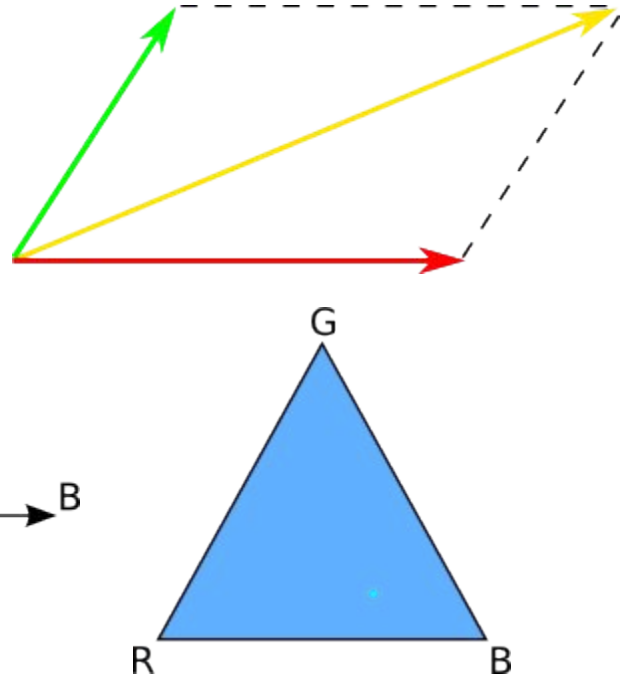
RGBA

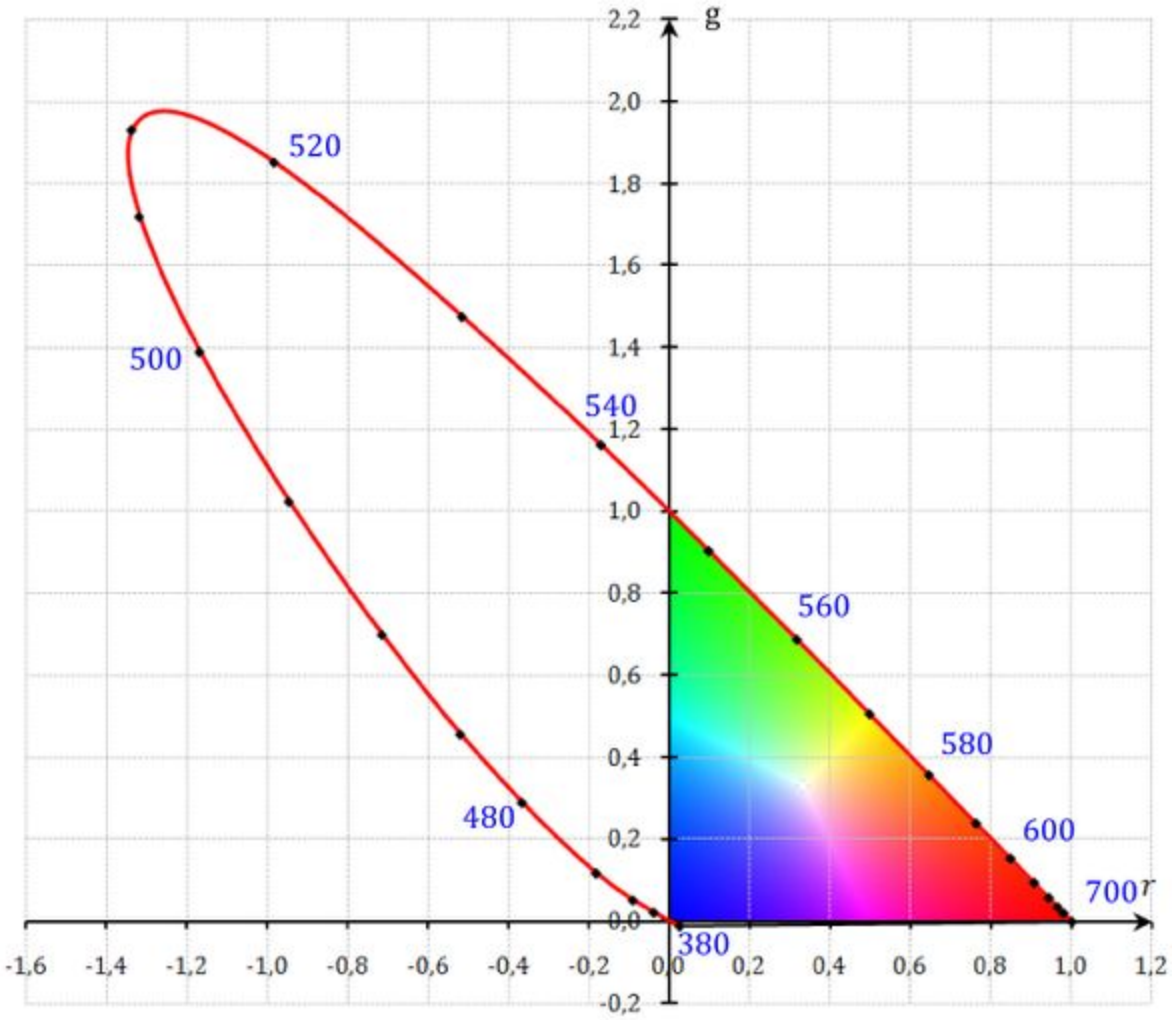
Модель XYZ

- Дин Джадд (*Deane B. Judd*)
- 1931 г.

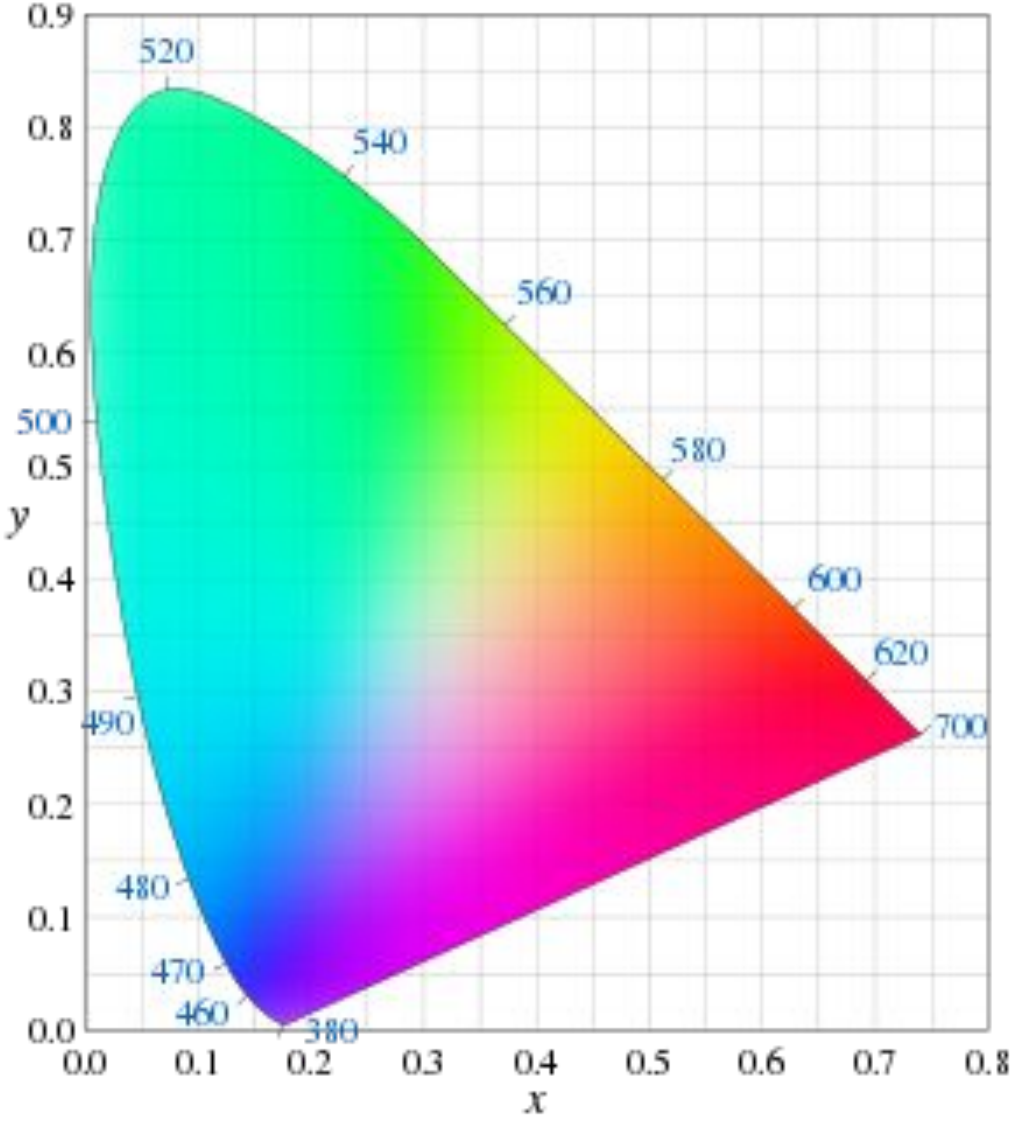


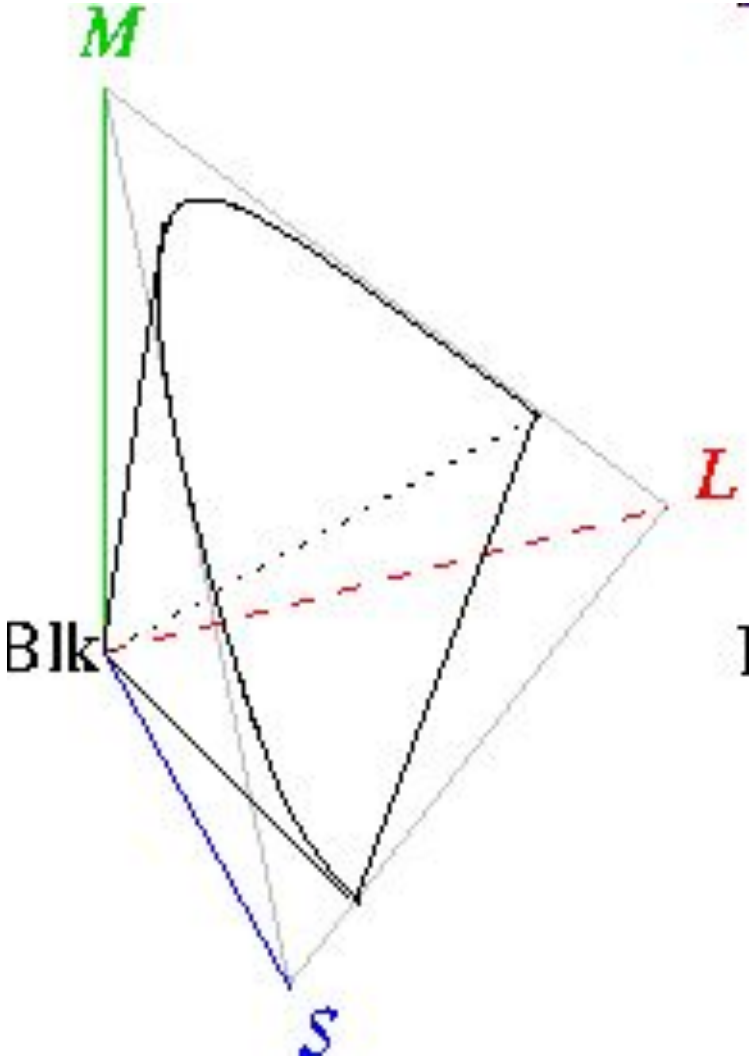





g
r
2,2
2,0
1,8
1,6
1,4
1,2
1,0
0,8
0,6
0,4
0,2
0,0
-0,2
-1,6
-1,4
-1,2
-1,0
-0,8
-0,6
-0,4
-0,2
0,0
0,2
0,4
0,6
0,8
1,0
1,2
520
500
540
560
580
600
700
480
380

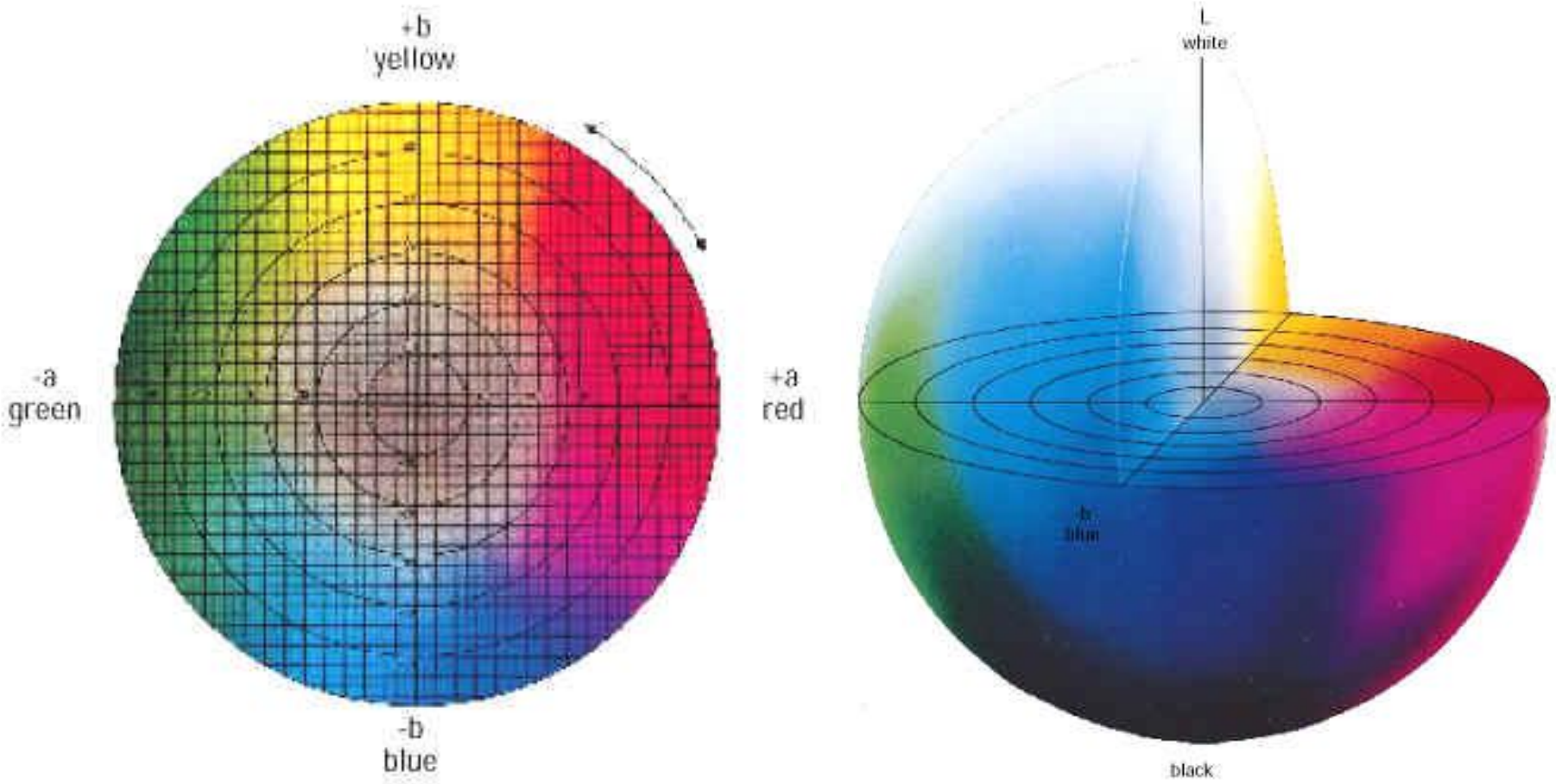
XYZ



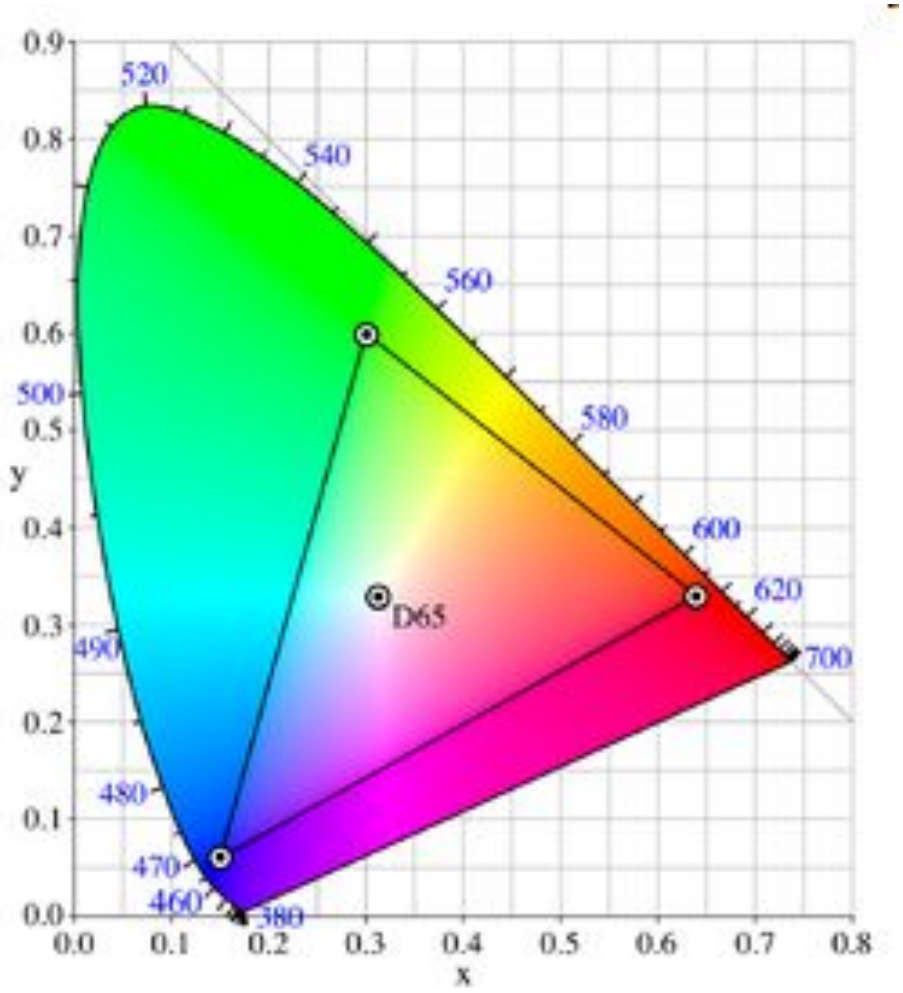


- Y – яркость
- Z – колбочки S
- X >= 0

L*a*b*

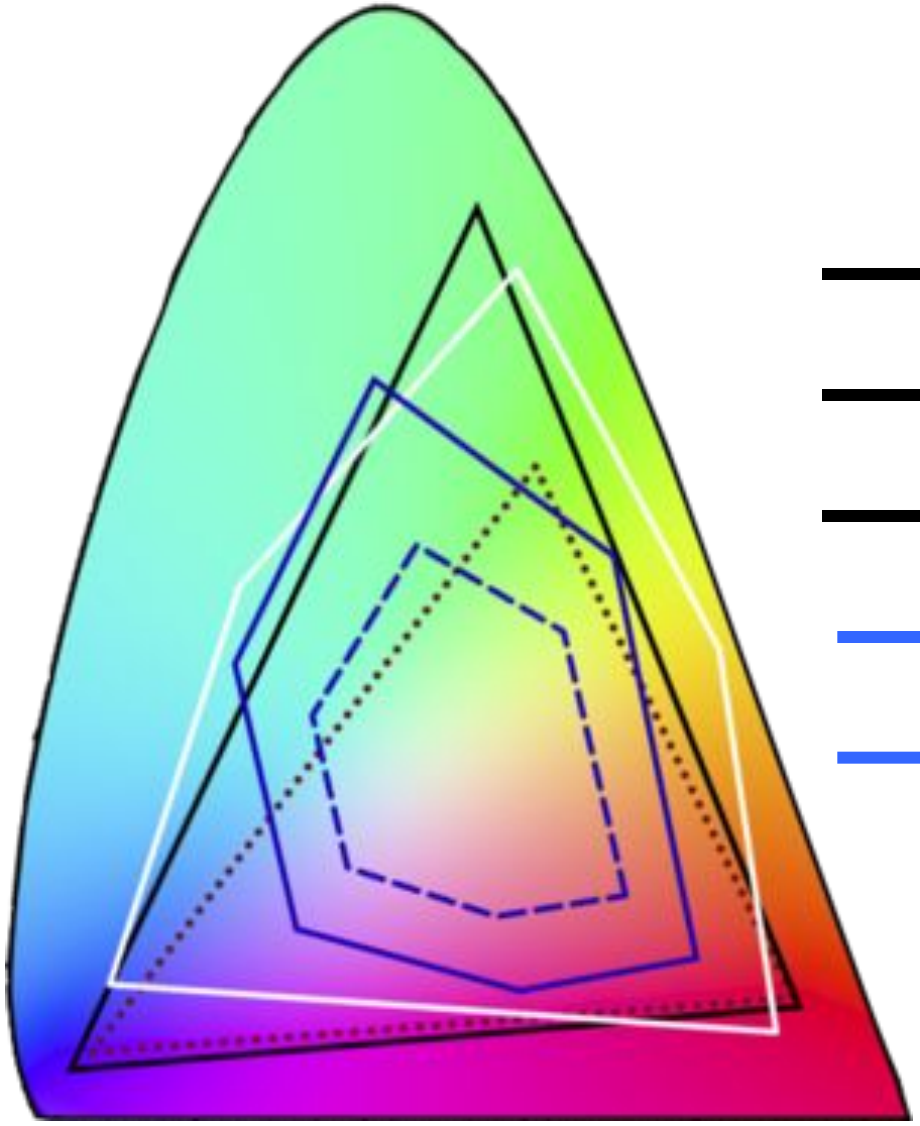


sRGB



недостаток

Проблемы цветопередачи



фотоэмульсия

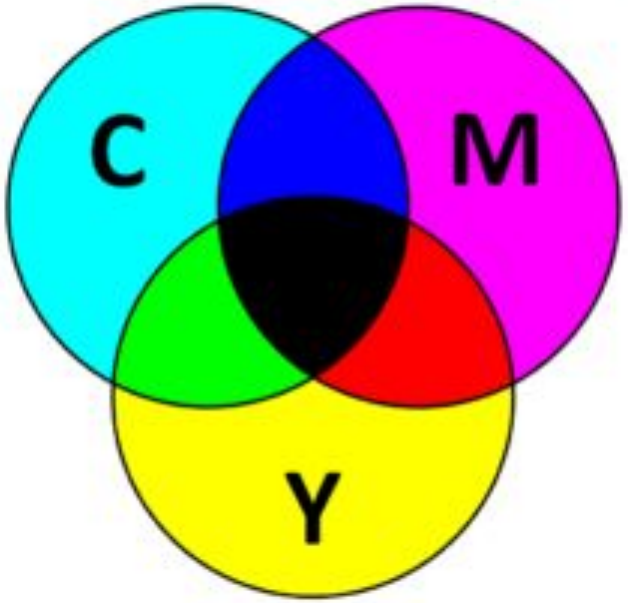
монитор

Adobe RGB

офсетная печать

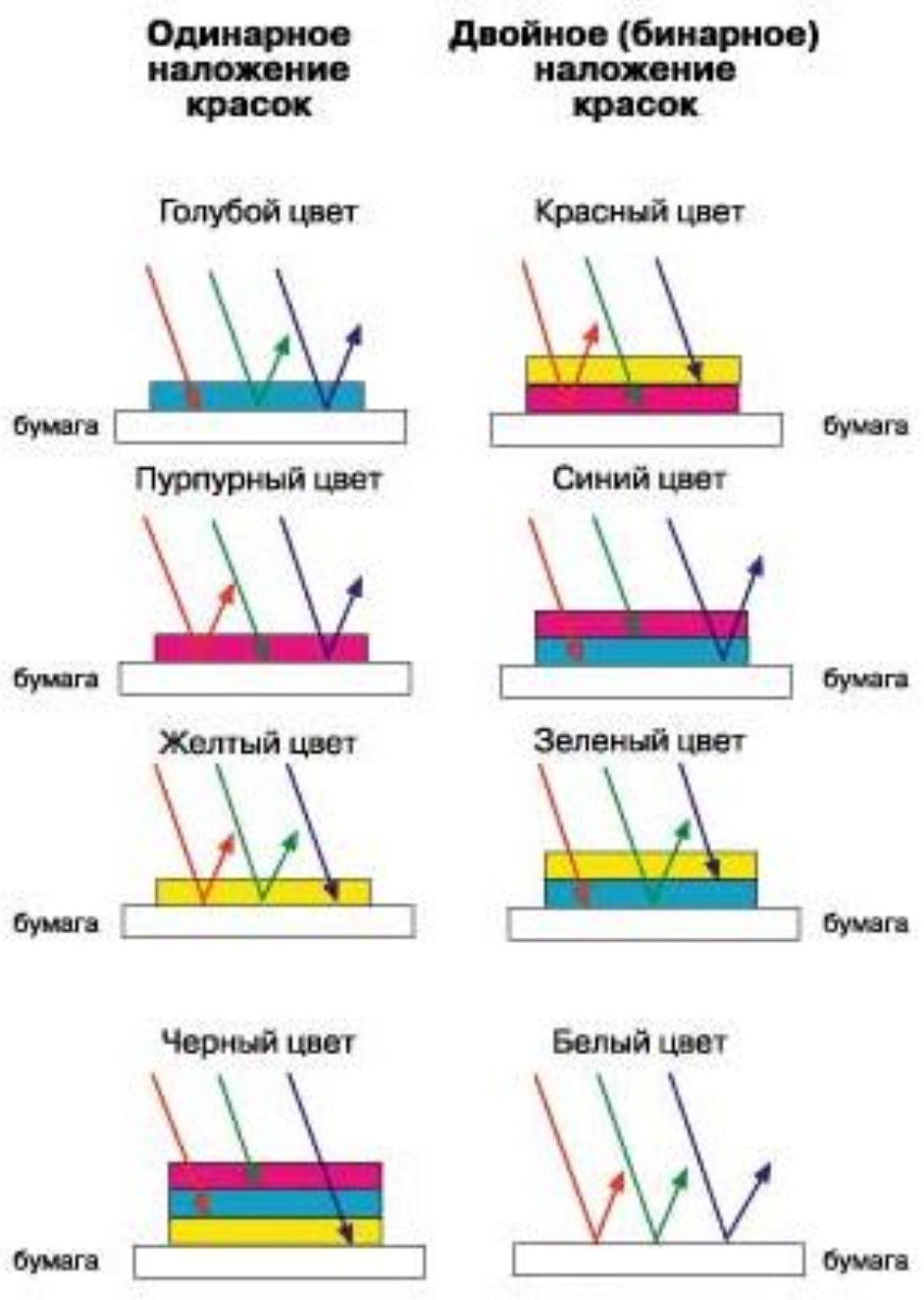
бытовой принтер

CMYK

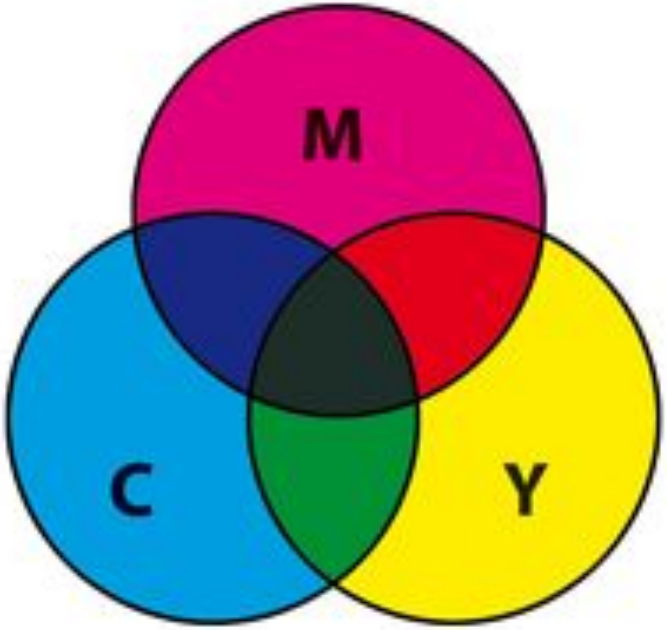


- субтрактивная модель
- 4 цвета
  - голубой *Cyan*
  - сиреневый *Magenta*
  - жёлтый *Yellow*
  - (чёрный) *Key color*
- триадная печать
- часто применяется для печати

Печать красками

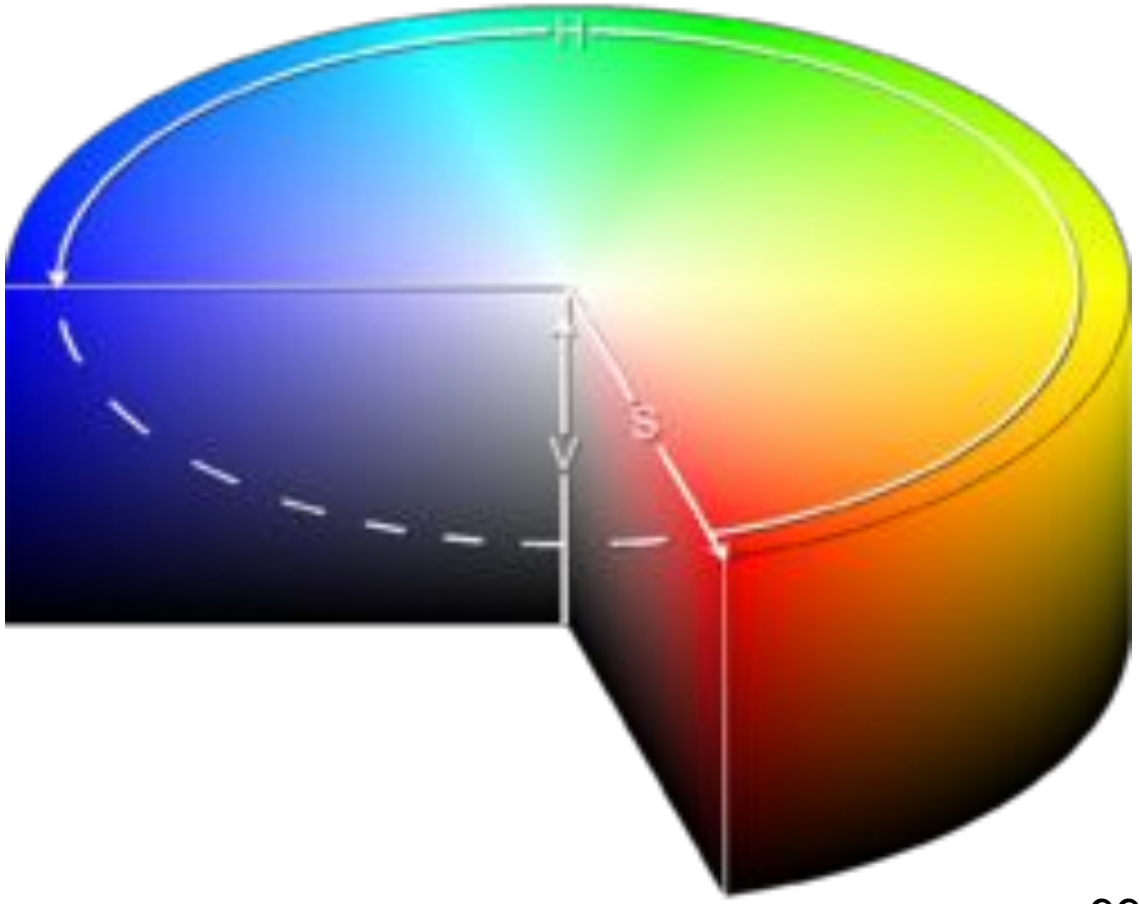


Почему – K?

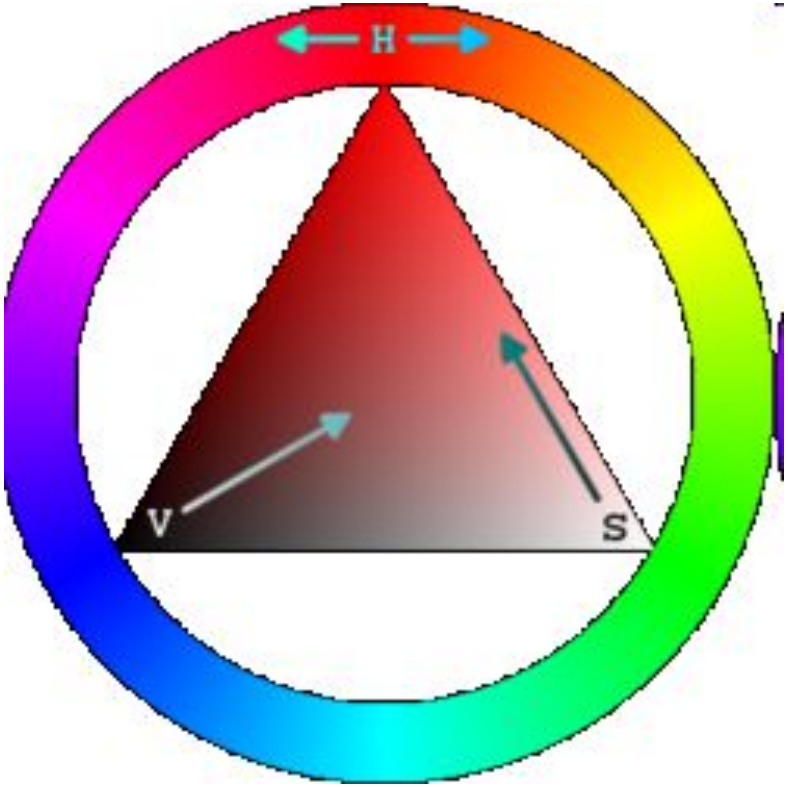


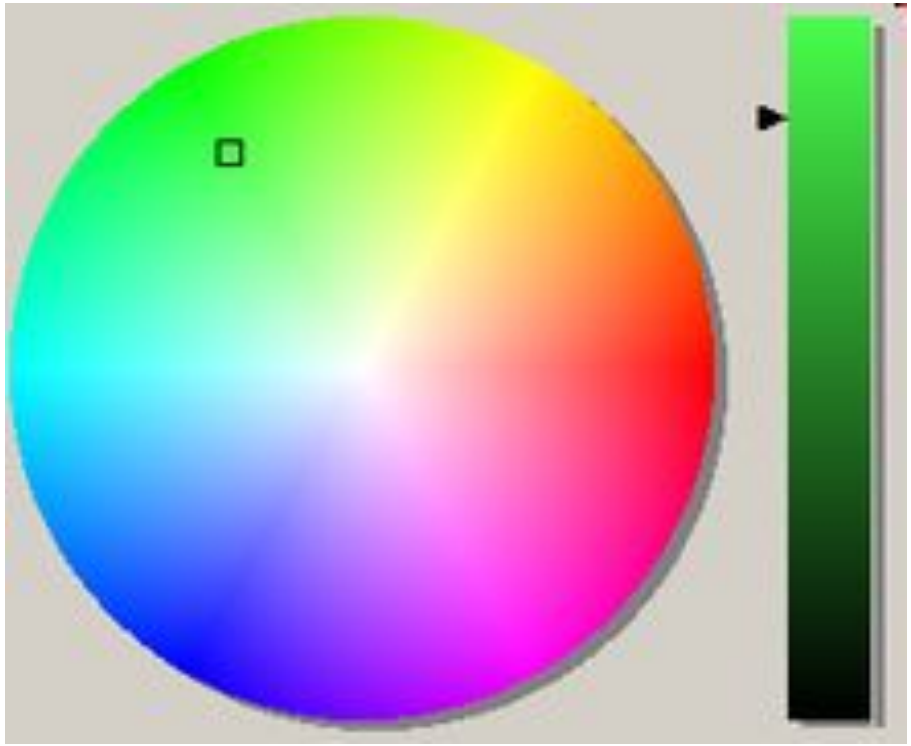
- *Hue* — цветовой тон
  - например, красный, зелёный или сине-голубой
  - варьируется в пределах 0—360°, однако иногда приводится к диапазону 0—100 или 0—1.
- *Saturation* — насыщенность
  - варьируется в пределах 0—100 или 0—1
  - иногда называют чистотой цвета
  - чем ближе этот параметр к нулю, тем ближе цвет к нейтральному серому.
- *Value* (значение цвета) или *Brightness* — яркость
  - задаётся в пределах 0—100 и 0—1.

Пространство HSV



Варианты интерфейса



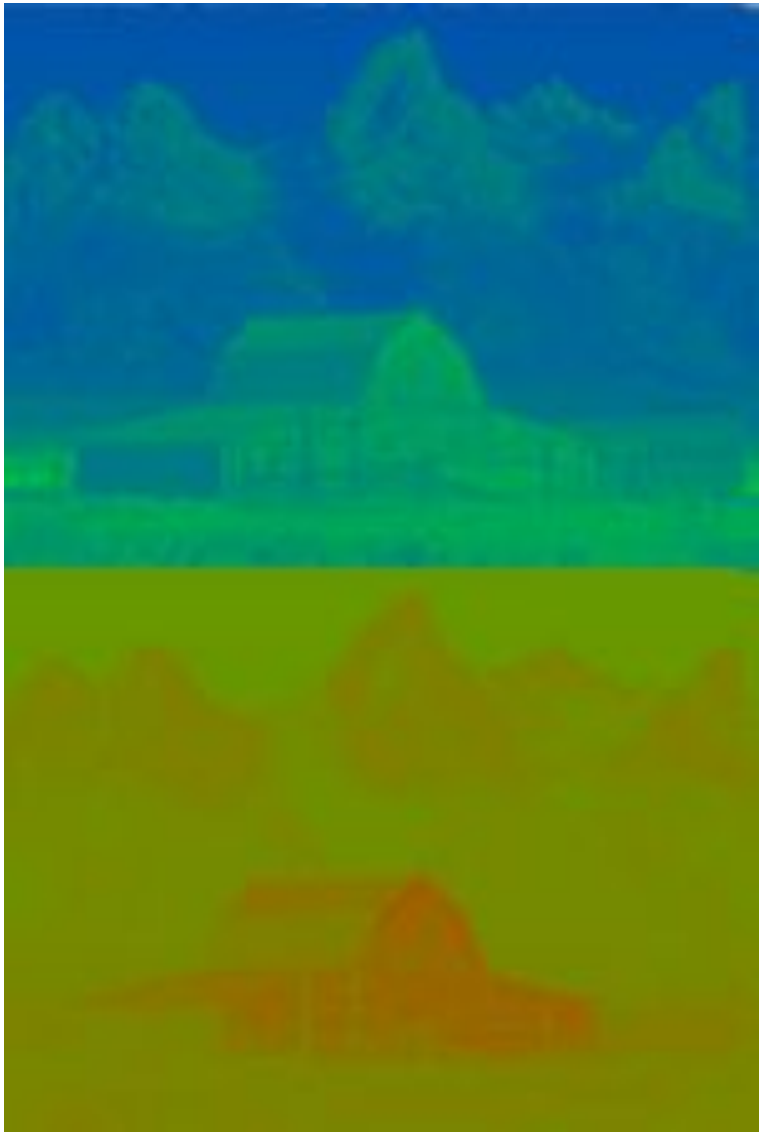


Пример HSV

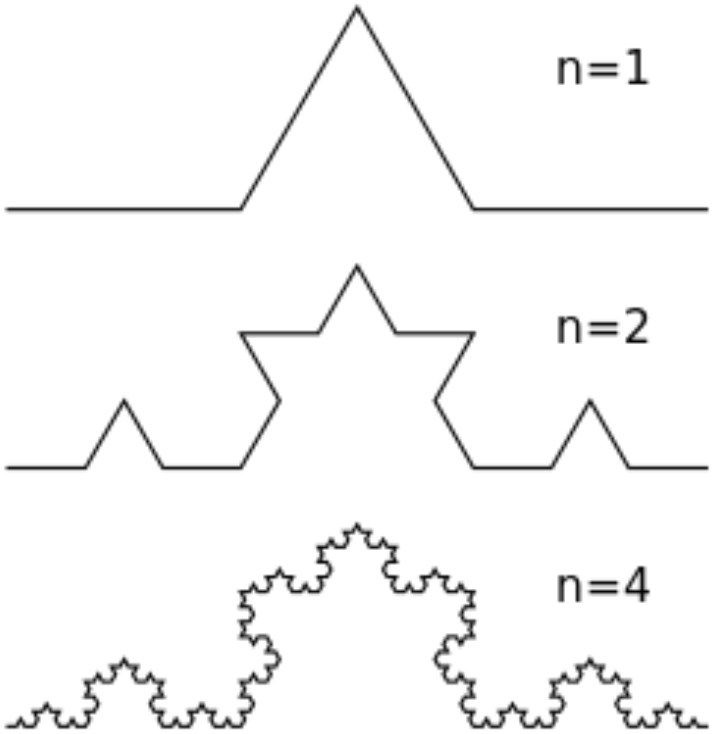
YUV

- 3 компоненты — яркость (Y) и две цветоразностных (U и V)
- YPbPr – аналоговый сигнал
- YCbCr – цифровой сигнал
  - R = Y + 1.13983 * (V - 128)
  - G = Y - 0.39465 * (U - 128) - 0.58060 * (V - 128)
  - B = Y + 2.03211 * (U - 128)

Пример YUV



P.S. Фракталы



Снежинка Коха

$$i = \sqrt{-1}$$

$$c = x + i \cdot y$$

$$Z_0 = 0$$

$$\begin{aligned} Z_1 &= Z_0^2 + c \\ &= x + iy \\ Z_2 &= Z_1^2 + c \\ &= (x + iy)^2 + x + iy \\ &= x^2 + 2ixy - y^2 + x + iy \\ &= x^2 - y^2 + x + (2xy + y)i \end{aligned}$$

$$Z_3 = Z_2^2 + c = \ldots$$

Построение множества Мандельброта