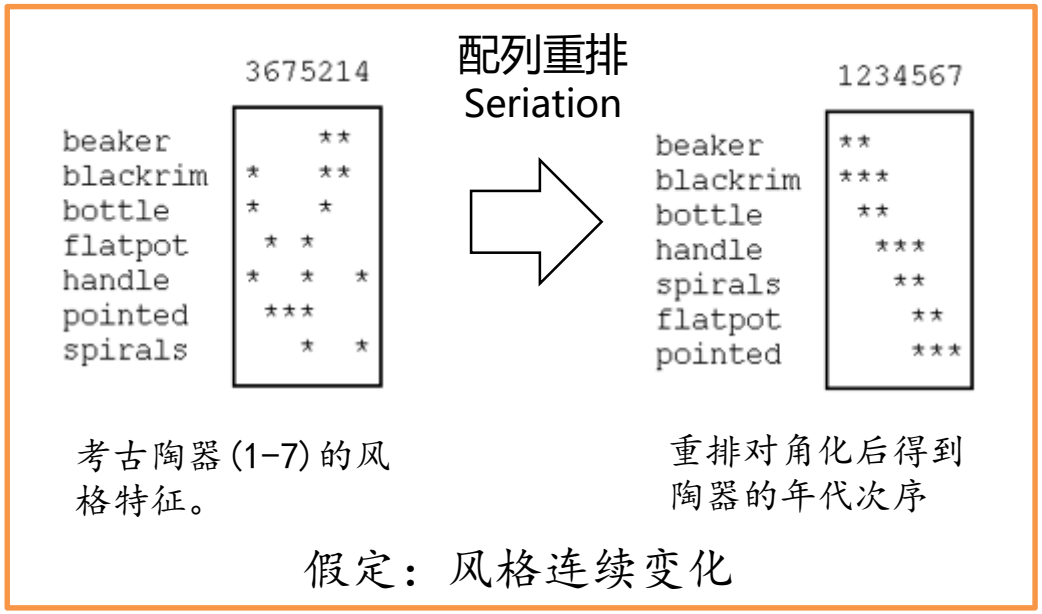
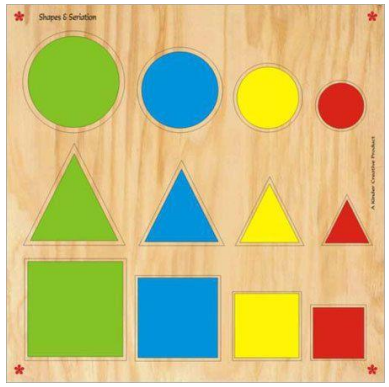
第二十讲 配列

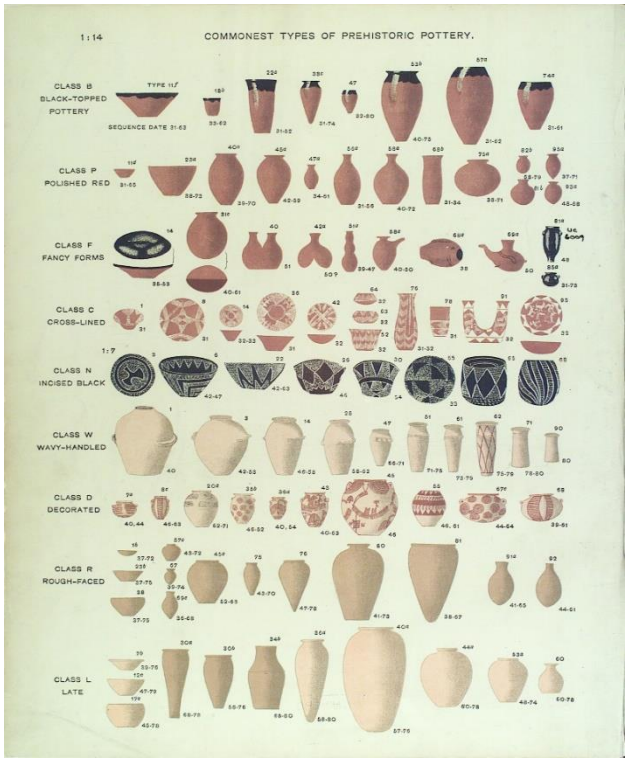
2024.5.22



配列（Seriation）

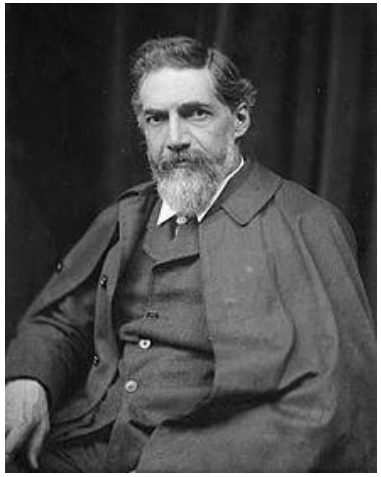
Seriation（排序、配列）方法将物体进行置换排序，使相似的物体彼此靠近。





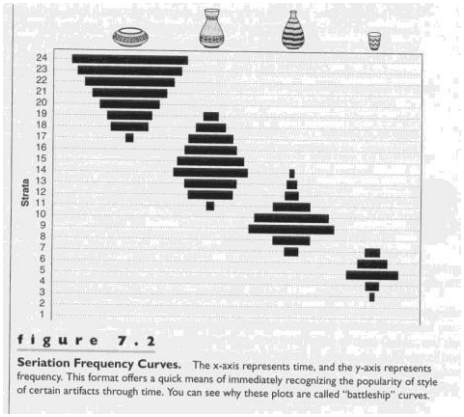
Petrie和考古学年代排序

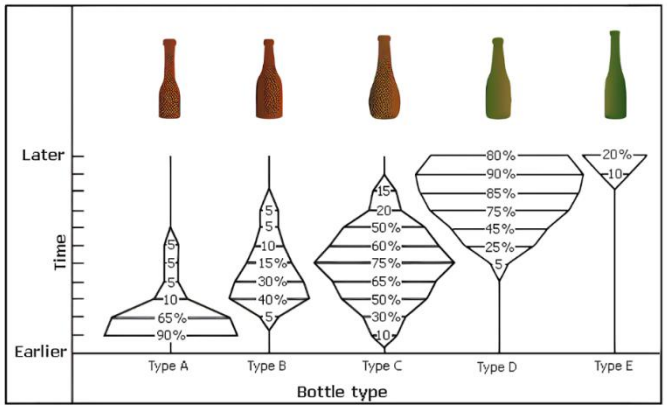
Flinders Petrie（弗林德斯·皮特里1853–1942），英国埃及学家，在考古学和文物保存方面首创了成体系的方法，包括配列seriation方法。中国考古学奠基人夏鼐深受其影响。



By linking styles of pottery with periods, Petrie was the first to use seriation in Egyptology, a new method for establishing the chronology of a site。

Petrie, F. W. M. (1899). Sequences in prehistoric remains. *Journal of the Anthropological Institute* **29**:295–301





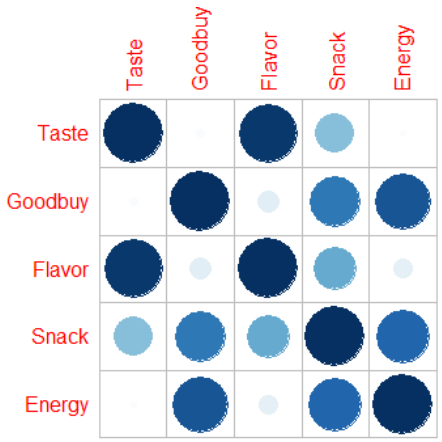
Corrplot 中的排序

R函数corrplot将相关系数（或一般相似度矩阵）可视化。左下图没有经过重排，可视化效果较差。

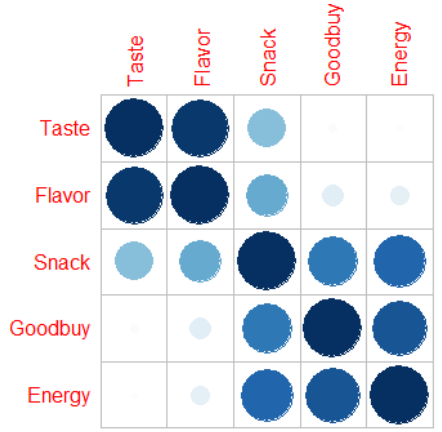
经过恰当的变量重排，使相关性大的变量彼此靠近，我们可得到右下图 - 有序、整洁。

例1. 食品问卷调查，5个问题的相关系数，右表。

	Taste	Goodbuy	Flavor	Snack	Energy
Taste	1	0.02	0.96	0.42	0.01
Goodbuy	0.02	1	0.13	0.71	0.85
Flavor	0.96	0.13	1	0.50	0.11
Snack	0.42	0.71	0.50	1	0.79
Energy	0.01	0.85	0.11	0.79	1



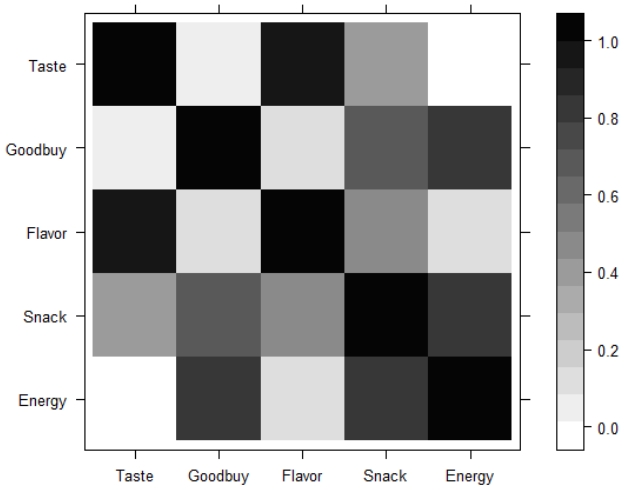
corrplot(r, order="oringial")



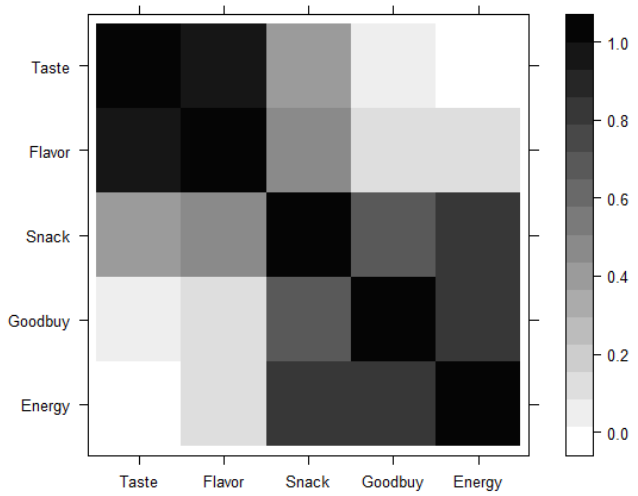
corrplot(r, order="AOE")

R函数：levelplot

相关系数或相似度矩阵更好的可视化方法是热图， 左下图未经重排，右下图是重排之后的效果（R函数：image，levelplot）。



> levelplot(R) #原图



> levelplot(R.sorted)

下面我们主要介绍基于二次优化即矩阵谱分解的排序方法：

- One-mode seriation：对一个下标排序；
- Two-mode seriation：对两个下标排序

单配列（One-mode Seriation）

依据物体的单个属性，比如按尺寸大小（右图），将物体从小到大排列。

单配列问题

One-mode seriation问题： 给定n个物体的相似度矩阵$S = (s_{ij})$或距离矩阵$D = (d_{ij})$，seriation寻找$\{1,2,\dots,n\}$的一个置换π, 使得

若i,j相似（s_{ij}较大或d_{ij}较小），则$\pi(i)$，$\pi(j)$之间的距离，比如$|\pi(i) - \pi(j)|$应该较小。

单配列的组合优化框架

将单配列问题设定为二次组合优化问题，极小化目标函数

$$\min \sum s_{ij}[\pi(i) - \pi(j)]^2 \qquad (*)$$

计算复杂度： $n!$

注：其它各种合理的目标设定都是允许的：

$$\min\sum(d_{ij}-|\pi(i)-\pi(j)|)^2，或$$

$$\min\sum\frac{1}{1+d_{ij}}(\pi(i)-\pi(j))^2$$

等等。我们下面将主要讨论（*）

松弛为二次优化

（*）问题中放松π：$\{1,2,\dots,n\}\to\{1,2,\dots,n\}$是置换映射的要求，并改用记号$x$替代$\pi$，假设

$$x：\{1,2,\dots,n\}\to R,$$

记$x_i=x(i)$, 则（*）问题改写为

$$min\sum s_{ij}(x(i)-x(j))^2\triangleq min\sum s_{ij}(x_i-x_j)^2$$

这里需要限制x's, 通常限制$\|\mathbf{x}\|=1$。

记S的第i行或第i列总和$d_i=\sum_j s_{ij}$ (度数),矩阵$D=diag(d_1,\dots.,d_n)$。

改写目标函数

$$\sum_{i,j} s_{ij}(x_i - x_j)^2 \quad = 2\sum_j d_i x_i^2 - 2\sum_{i,j} s_{ij} x_i x_j$$

$$= 2\mathbf{x}^\top D\,\mathbf{x} - 2\mathbf{x}^\top S\mathbf{x} = 2\mathbf{x}^\top (D - S)\,\mathbf{x} \quad \triangleq 2\mathbf{x}^\top L\mathbf{x} \geq 0$$

其中

$$L = D - S$$

称为拉普拉斯(Laplacian)矩阵。显然L是半正定矩阵。

显然当$\mathbf{x} \propto \mathbf{1}$，二次型达到极小值0，为了避免得到这个解（不可排序），我们约束$\mathbf{x}$分量不全相同（约束$\mathbf{x} \perp \mathbf{1}$）。

在上述约束下， 二次型$\mathbf{x}^\top L\mathbf{x}$的极小值为最小非0特征根，并在相应的特征向量处达到极小值。

综合上述讨论，我们有

单向谱配列

假设$S=(s_{ij})$为n个物体的相似度矩阵(即对称矩阵), 假设所有 $s_{ij}\geq 0$，记Laplacian矩阵$L=D-S, D=diag(d_1,\ldots..,d_n), d_i=\sum_j s_{ij}$。约束 $\|\mathbf{x}\|=1$，$\mathbf{x}\perp\mathbf{1}$，则

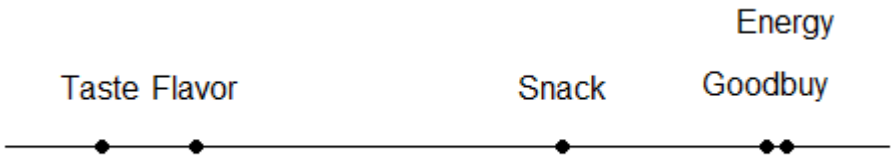
$$\sum s_{ij}(x_i-x_j)^2=2\mathbf{x}^{\mathsf{T}}L\mathbf{x}$$

当$\mathbf{x}$是 L的最小非0特征根对应的特征向量时达到极小。将该特征向量排序即得到谱配列（spectral seriation）

例1（续）L 的最小非0特征根0.807，对应的特征向量：

Taste	Flavor	Snack	Goodbuy	Energy
−0.600	− 0.449	0.131	0.458	0.480

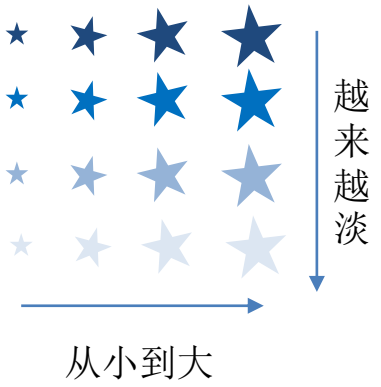
	Taste	Goodbuy	Flavor	Snack	Energy
Taste	1	0.02	0.96	0.42	0.01
Goodbuy	0.02	1	0.13	0.71	0.85
Flavor	0.96	0.13	1	0.50	0.11
Snack	0.42	0.71	0.50	1	0.79
Energy	0.01	0.85	0.11	0.79	1



按照上述次序重排相关系数矩阵得到P4,P5的图。

双向配列（Two-mode Seriation）

依据物体的两个属性，比如按尺寸大小或接近于圆的程度（右图），将物体按两个属性排列。



双向配列问题

对于列联表或任何丰度矩阵（不是相似度或距离方阵）$A = (a_{ij})$，矩阵的(i,j)元a_{ij}代表行属性i与列属性j之间的联系程度。Two-mode seriation 置换行、列使得对角线附近的值较大。

求解行标号和列标号的置换$u_i = u(i), v_j = v(j)$,若行标 i 与列标 j 联系密切, 即 a_{ij}较大，则$|u_i - v_j|$ 较小，即

$$\sum a_{ij}(u_i - v_j)^2 = \min!$$

二次优化

放松对u,v的要求，假设$\mathbf{u}=(u_i)$为$n\times 1$实数向量，$\mathbf{v}=(v_j)$为$p\times 1$实数向量，$||\mathbf{u}||=||\mathbf{v}||=1$, 双向配列问题设定为求解二次优化问题

$$min\sum a_{ij}(u_i-v_j)^2 \qquad (**)$$

记A的行和列和分别为

$$r_i=\sum_j a_{ij},\ c_j=\sum_i a_{ij},$$
$$D_r=diag(r_1,\dots,r_n), D_c=diag(c_1,\dots,c_p),$$

则容易验证

$$\sum a_{ij}(u_i-v_j)^2=\sum_i r_i u_i^2+\sum_j c_j v_j^2-2\sum_{i,j} a_{ij}\,u_i v_j$$
$$=\mathbf{u}^\top D_r\mathbf{u}+\mathbf{v}^\top D_c\mathbf{v}-2\mathbf{u}^\top A\,\mathbf{v}$$
$$=(\mathbf{u}^\top,\mathbf{v}^\top)\begin{pmatrix}D_r & 0\\ 0 & D_c\end{pmatrix}\begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix}-(\mathbf{u}^\top,\mathbf{v}^\top)\begin{pmatrix}0 & A\\ A^\top & 0\end{pmatrix}\begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix}$$
$$=(\mathbf{u}^\top,\mathbf{v}^\top)\left\{\begin{pmatrix}D_r & 0\\ 0 & D_c\end{pmatrix}-\begin{pmatrix}0 & A\\ A^\top & 0\end{pmatrix}\right\}\begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix}$$

记$S=\begin{pmatrix}0 & A\\ A^{\top} & 0\end{pmatrix}$，其行和组成的对角度数矩阵$D=\begin{pmatrix}D_r & 0\\ 0 & D_c\end{pmatrix}$，

$L=D-S=\begin{pmatrix}D_r & -A\\ -A^{\top} & D_c\end{pmatrix}$ 是Laplacian矩阵。同单向谱配列问题，

L的最小非0 特征根对应的特征向量$\begin{pmatrix}\mathbf{u}_{\min}\\ \mathbf{v}_{\min}\end{pmatrix}$是(**)的最优解。

双向谱配列

假设A $=(a_{ij})$为$n\times p$丰度矩阵, 假设所有 $a_{ij}\geq 0$，记

$$S=\begin{pmatrix}0 & A\\ A^{\top} & 0\end{pmatrix}$$

$D=diag(d_1,\ldots,d_n)$, $d_i=\sum_j s_{ij}$。令Laplacian矩阵 $L=D-S$,

假设$\mathbf{u}=(u_i)$为$n\times 1$实数向量，$\mathbf{v}=(v_j)$为$p\times 1$实数向量，则

$$\sum a_{ij}(u_i-v_j)^2=\mathbf{x}^{\top}L\mathbf{x},\ \mathbf{x}=\begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix}$$

且当$\mathbf{x}$是 L的最小非0特征根对应的特征向量时达到极小。将该特征向量的$\mathbf{u}$， $\mathbf{v}$分别排序即得到双向谱配列。

当A是双向中心化的，（**）问题的最优解为奇异值分解
最大奇异值对应的两个特征向量。

引理1. 若$A = (a_{ij})$是$n \times p$双向中心化的矩阵，则对任何$\mathbf{u} \in R^n, \mathbf{v} \in R^p$

$$\mathbf{u}^\top A\, \mathbf{v} = \sum a_{ij}(u_i - v_j)^2$$

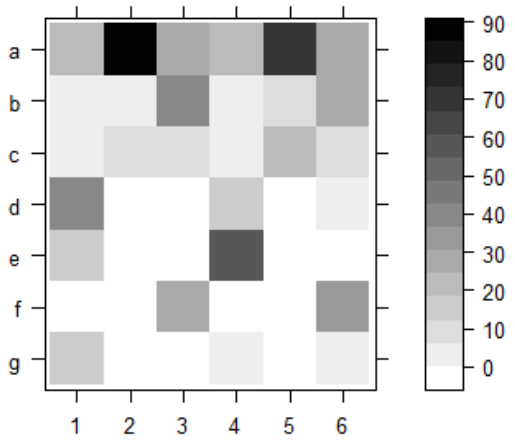
命题1. 对于双向中心化的$n \times p$丰度矩阵$A = (a_{ij})$，（**）等价于极大化问题

$$\max_{\mathbf{u} \in R^n, \mathbf{v} \in R^p} \mathbf{u}^\top A\, \mathbf{v} \ , \big|\big|\mathbf{u}\big|\big| = \big|\big|\mathbf{v}\big|\big| = 1$$

的最优解为A的最大奇异值对应的特征向量。 按照 $\mathbf{u}$，$\mathbf{v}$的次序重新排列A的行和列。

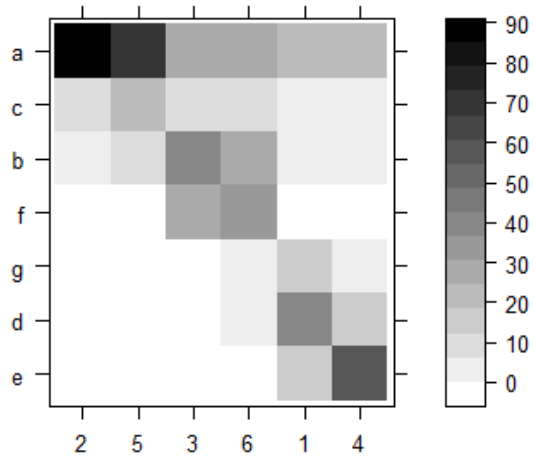
例2. 6个考古地点出土7种燧石，个数统计如下

	a	b	c	d	e	f	g
1	20	3	4	42	18	0	13
2	85	3	12	0	0	0	0
3	26	40	8	0	0	26	0
4	20	1	4	13	58	0	4
5	67	10	23	0	0	0	0
6	26	29	8	3	0	33	1



右图：双向中心化后，应用1阶奇异值分解分别重新排列行和列。考古点排序为

2 5 3 6 1 4



R：

```
library(seriation)
seriate(x)
```

后面将考虑欧洲语言相似性问题，这里首先介绍背景：

欧洲语言大多都属于印欧语系。包括罗曼/拉丁、日耳曼、斯拉夫、希腊等语族。里海和黑海北部的雅利安人的南迁和西迁使得欧洲大部分地区以及伊朗、印度的语言有同源性。

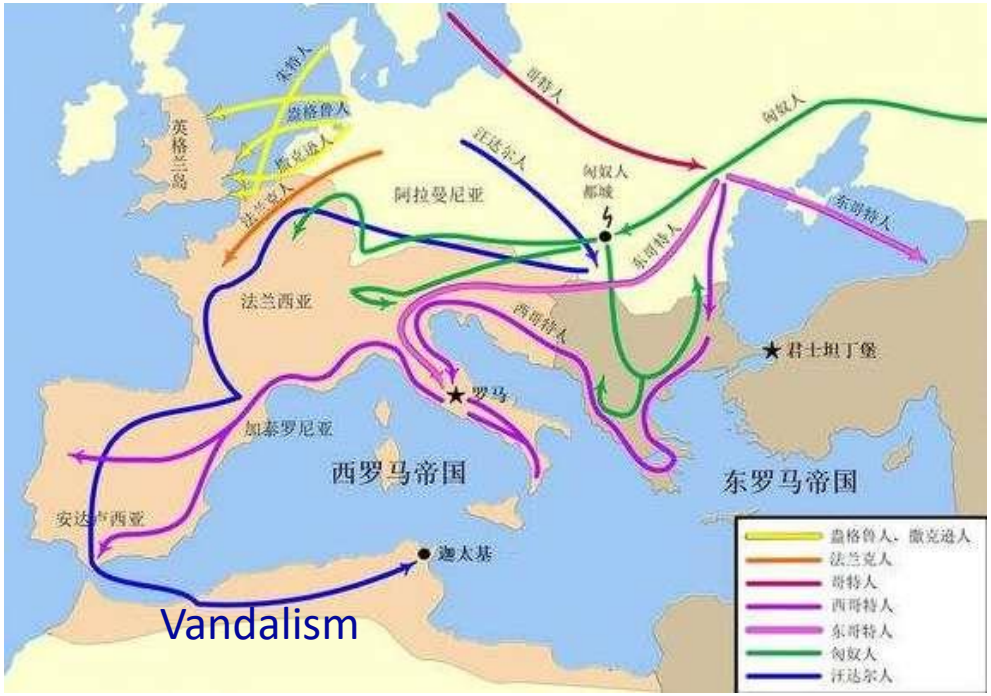
最早的语言、文明、文字出现在地中海东岸的两河流域和地中海南岸尼罗河一带。一般认为，现在通用的罗马字母表源自于BC1200地中海东南岸迦南地区（黎巴嫩）的腓尼基22个字母（**腓尼基字母表，Phoenician Alphabet**）。腓尼基人擅长海上贸易，字母表传到希腊后，希腊人加上了元音字母形成了希腊字母，意大利的伊特鲁里亚人又将希腊字母改造成了罗马字母表。

古典时期，希腊然后是罗马帝国统治了除了北欧和东欧之外的意大利、巴尔干以及亚洲的两河流域和北非的埃及。由于北方斯堪的纳维亚半岛的日耳曼民族南侵，罗马与北方蛮族、东方波斯征战多年。

公元375年，来自亚洲的匈人(Hunnen)阿提拉攻击黑海一带日耳曼民族东哥特人，引发了4-6世纪欧洲民族大迁徙，导致欧亚大陆的种族迁移和融合，以及语言和文化的变化。

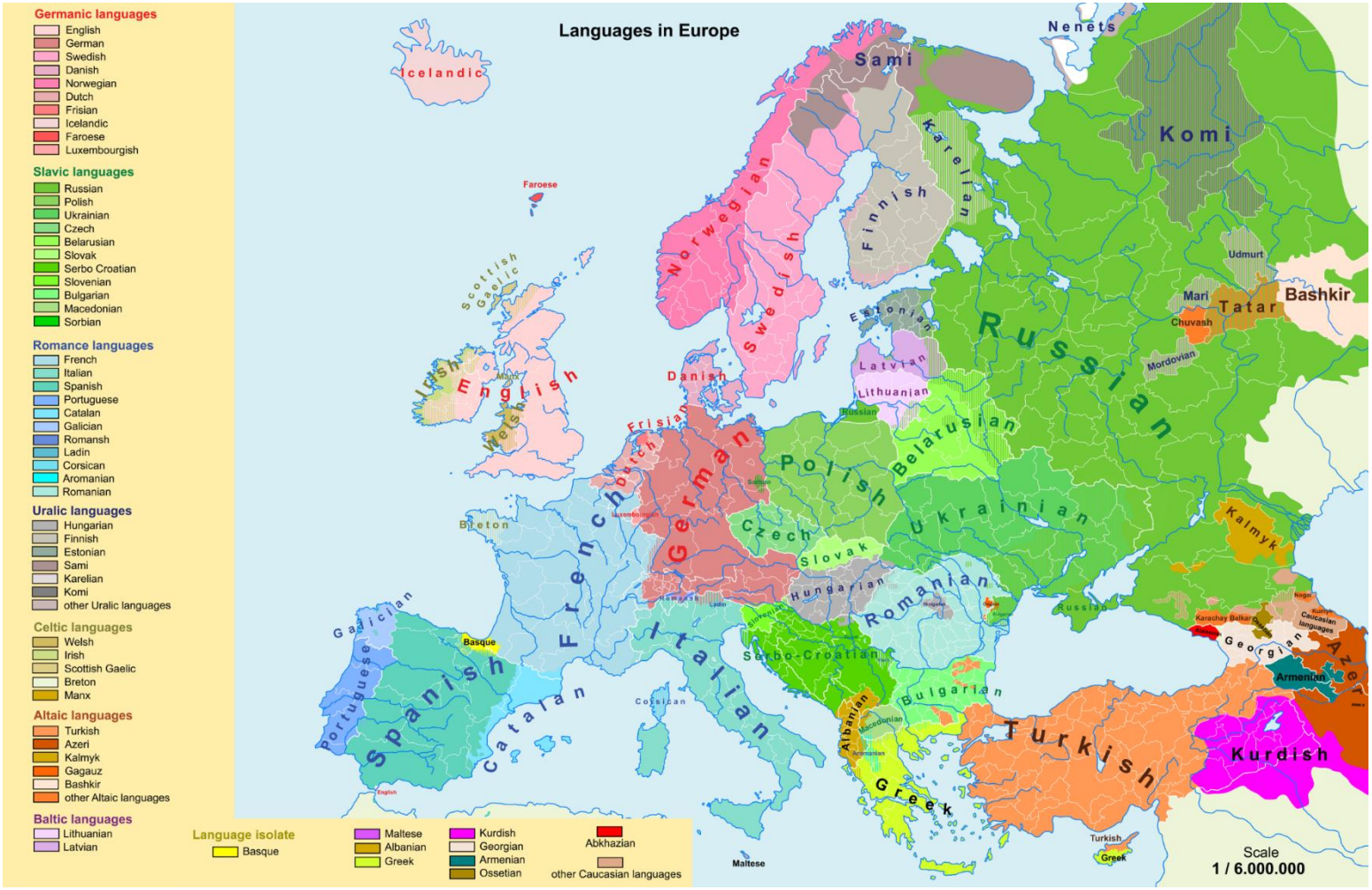
4-6世纪欧洲民族大迁徙：欧洲日尔曼“蛮族人”包括东哥特人、西哥特人、汪达尔人、勃艮第人、伦巴底人、法兰克人等被匈人驱赶被迫西迁，导致西罗马帝国灭亡，中世纪开始，因此形成了若干现代欧洲国家的原型和语言特征：

法兰克人驱逐西哥特人建立了法兰克王国（法国、德国、意大利）、
西哥特人驱逐汪达尔人建立了西哥特王国（西班牙）、
汪达尔人驱逐罗马人建立了北非汪达尔-阿兰王国（突尼斯、迦太基）、
东哥特人驱赶罗马人建立了东哥特王国（意大利）、
盎格鲁-撒克逊人进入不列颠岛驱逐凯尔特人建立了七王国。



Vandalism

欧洲语言地图


Languages in Europe
Germanic languages
English
German
Swedish
Danish
Norwegian
Dutch
Frisian
Icelandic
Faroese
Luxembourgish
Slavic languages
Russian
Polish
Ukrainian
Czech
Belarusian
Slovak
Serbo Croatian
Slovenian
Bulgarian
Macedonian
Sorbian
Romance languages
French
Italian
Spanish
Portuguese
Catalan
Galician
Romansh
Ladin
Corsican
Aromanian
Romanian
Uralic languages
Hungarian
Finnish
Estonian
Sami
Karelian
Komi
other Uralic languages
Celtic languages
Welsh
Irish
Scottish Gaelic
Breton
Manx
Altaic languages
Turkish
Azeri
Kalmyk
Gagauz
Bashkir
other Altaic languages
Baltic languages
Lithuanian
Latvian
Language isolate
Basque
Maltese
Albanian
Greek
Kurdish
Georgian
Armenian
Ossetian
Abkhazian
other Caucasian languages
Scale
1 / 6.000.000