I 202: Information Organization & Retrieval Fall 2025

Categorization, Part 1

Pyramid Game Rules

Player A: guess the category
Player B: give clues without using words from the category name

https://www.youtube.com/watch?v=UJMx_YfZiW0

WHY IS THE CONTENT IN THIS COURSE IMPORTANT?

What we are studying about Info Org:

How Built

Information organization choices are influenced by cognitive, cultural, technical, and historical factors.

Why it Matters

Information organization choices have deep social, technical, and ethical consequences.

Course Topics in Organized into One Slide

Data / Information
Collections
Categories
- Types of categories
- Cognitive / language aspects
    - Naming / Lexical similarity
- Structure
    - Hierarchical / Taxonomy
    - Faceted
    - Overlapping / Clustering
    - Network / Ontology
- Use in Navigation & Search
    - Information Architecture
    - Faceted Navigation

Technology Support for Info Org
- Identifiers
- Metadata
- Markup
- Schema / Databases
- Search Ranking / Evaluation
- Automated category creation
- Automated similarity

Social / Ethical Aspects
- Cultural Bias
- Intellectual Property
- Standards Process

This week's focus

Data / Information Collections

Categories

- **Types of categories**
- **Cognitive / language aspects**
  - Naming / Lexical similarity
- Structure
  - Hierarchical / Taxonomy
  - Faceted
  - Overlapping / Clustering
  - Network / Ontology
- Use in Navigation & Search
  - Information Architecture
  - Faceted Navigation

Technology Support for Info Org

- Identifiers
- Metadata
- Markup
- Schema / Databases
- Search Ranking / Evaluation
- Automated category creation
- Automated similarity

Social / Ethical Aspects

- Cultural Bias
- Intellectual Property
- Standards Process

Today's Outline

Who Creates Categories

Principles for Categorization

Classical vs Cognitive Categories

Intro to Categories

- Categories are sets or groups of resources that are assigned a common label
- Categories are equivalence classes of sets or groups of things or abstract entities that we treat the same.
- Categories are involved whenever we perceive, communicate, analyze, predict, classify
- All human languages and cultures divide the physical and experiential "worlds" into categories

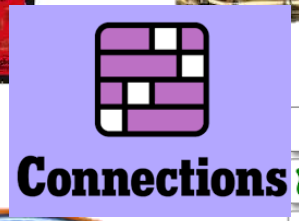
Categories From The Pyramid Game

- **Car Brands**: Fiat, Toyota, Ford, Audi
- **Spanish Words**: “hola”, “Juan”, “despacito”
- **Why You Wear a Tuxedo**: going to a wedding, going to an awards show, performing as a magician, pretending to be a penguin
- **What a Clock Says**: “The time is 10am”, “I have a minute hand and a second hand”, “I am sitting on your wrist with a band around me”
- **Things You Solve**: puzzle, problem, for x
- **Things You Climb**: mountain, tall ladder, tree, society

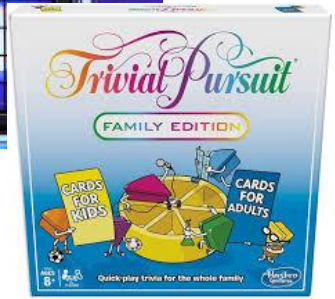
What are the characteristics of these category types?

Which Games / Game Shows are about Categories?

Categories



Individual information / facts / data



LET'S PLAY FAMILY FEUD

Family Feud Gameplay

- Players in two teams are shown a category name
- They are asked to guess the most likely category members based on responses from a survey of 100 people.
- A team gets control if they guess the higher-ranking member
- A team wins if they guess all of the top-ranked members before getting three wrong.

Example Family Feud Q/A

pollev.com/I202

1. Name a House You Never Want to Be In

2. Name Something Associated with Vampires

Example Family Feud Q/A

1. Name a House You Never Want to Be In

Haunted House (27)
Jail/Big House (11)
Doghouse (8)
Drug House (7)
Small House (7)
Glass House (6)
Cat House (5)
Outhouse (5)

2. Name Something Associated with Vampires

Twilight (33)
Blood/Bloodsucker (29)
Garlic (9)
Bat (7)
Cape (7)
Dracula (5)
Fangs (4)
Halloween (4)

https://parade.com/1188030/marynliles/family-feud-questions/

Family Feud Categorization

The “right answers” are based on a response to a survey. What might we assume about how people choose what to say?

What Inspires Category Creation?

- Culturally / Cognitively Created Categories
- Individually-Created Categories
- Institutionally-Created Categories
- Mathematically / Scientifically Created Categories
- Computationally-Created Categories

Culturally / Cognitively-Created Categories

- We use tens of thousands of categories that are embodied in our culture and language
- They develop slowly and typically change slowly
- Many have a perceptual or sensorimotor origin based on natural boundaries or discontinuities in perception and experience

More on this in a few slides

Individually-Created Categories

- Created to satisfy ad hoc requirements that emerge from an individual's unique experiences, preferences, and collections
- Created intentionally in response to specific organizing requirements, often short-term ones

INSTITUTIONALLY-CREATED CATEGORIES

- Explicit construction of a category system to enable more control, robustness, and interoperability than is possible with just the culturally-shared system
- Are often the collaborative artifact of many individuals who represent different organizational perspectives
- Usually developed via formal processes (in standards organizations or legislative bodies or via scientific or mathematical discovery) and require ongoing governance and maintenance

Example: Controlled Vocabulary

Example: Categories based on Principles of Geometry

Institutionally Created Category System:
UN Standard Products and Services Codes (UNSPSC)

Search: chicken

A - Raw Materials, Chemicals, Paper, Fuel
- 10000000 - Live Plant and Animal Material and Accessories and Supplies
  - 10100000 - Live animals
    - 10101600 - Birds and fowl
      - 10101601 - Live chickens
    - 10102100 - Birds and fowl hatching eggs
      - 10102101 - Chicken hatching eggs

B - Industrial Equipment & Tools
- 23000000 - Industrial Manufacturing and Processing Machinery and Accessories
  - 23220000 - Chicken processing machinery and equipment

https://www.ungm.org/Public/UNSPSC

Institutionally Created Category System:
"Gross Income" Tax Code Categories

26 U.S. Code § 61 - Gross income defined

(a) General definition

Except as otherwise provided in this subtitle, gross income means all income from whatever source derived, including (but not limited to) the following items:

(1) Compensation for services, including fees, commissions, fringe benefits, and similar items;

(2) Gross income derived from business;

(3) Gains derived from dealings in property;

(4) Interest;

(5) Rents;

(6) Royalties;

(7) Dividends;

(8) Annuities;

(9) Income from life insurance and endowment contracts;

(10) Pensions;

(11) Income from discharge of indebtedness;

(12) Distributive share of partnership gross income;

(13) Income in respect of a decedent; and

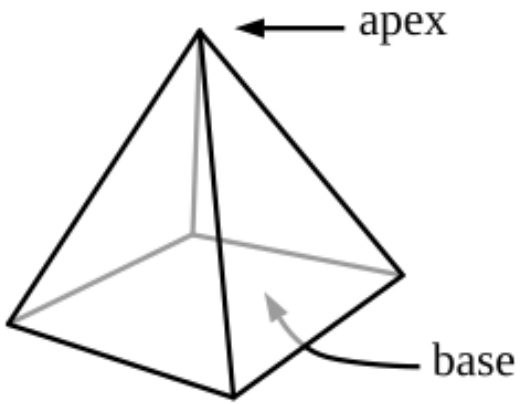
(14) Income from an interest in an estate or trust.

Classifies income types according to taxable or not

https://www.law.cornell.edu/uscode/text/26/61

Institutionally Created Category System:
Definition of the class of shapes that form a Pyramid

A **pyramid** is a polyhedron (a geometric figure) formed by connecting a polygonal base and a point, called the apex. Each base edge and apex form a triangle, called a lateral face. A pyramid is a conic solid with a polygonal base.



https://en.wikipedia.org/wiki/Pyramid_(geometry)

INSTITUTIONALLY-CREATED CATEGORY:

U.S. CENSUS QUESTIONNAIRE, 2020

QUESTIONS ABOUT RACE

➜ **NOTE: Please answer BOTH Question 6 about Hispanic origin and Question 7 about race. For this census, Hispanic origins are not races.**

6. Is this person of Hispanic, Latino, or Spanish origin?

- ☐ **No**, not of Hispanic, Latino, or Spanish origin
- ☐ Yes, Mexican, Mexican Am., Chicano
- ☐ Yes, Puerto Rican
- ☐ Yes, Cuban
- ☐ Yes, another Hispanic, Latino, or Spanish origin – *Print, for example, Salvadoran, Dominican, Colombian, Guatemalan, Spaniard, Ecuadorian, etc.*

7. What is this person's race?

Mark ☒ *one or more boxes* ***AND*** *print origins.*

- ☐ White – *Print, for example, German, Irish, English, Italian, Lebanese, Egyptian, etc.*
- ☐ Black or African Am. – *Print, for example, African American, Jamaican, Haitian, Nigerian, Ethiopian, Somali, etc.*
- ☐ American Indian or Alaska Native – *Print name of enrolled or principal tribe(s), for example, Navajo Nation, Blackfeet Tribe, Mayan, Aztec, Native Village of Barrow Inupiat Traditional Government, Nome Eskimo Community, etc.*
- ☐ Chinese
- ☐ Filipino
- ☐ Asian Indian
- ☐ Vietnamese
- ☐ Korean
- ☐ Japanese
- ☐ Native Hawaiian
- ☐ Samoan
- ☐ Chamorro
- ☐ Other Asian – *Print, for example, Pakistani, Cambodian, Hmong, etc.*
- ☐ Other Pacific Islander – *Print, for example, Tongan, Fijian, Marshallese, etc.*
- ☐ Some other race – *Print race or origin.*

For the motivations for the form of these questions, see
https://www.census.gov/newsroom/blogs/random-samplings/2021/08/improvements-to-2020-census-race-hispanic-origin-question-designs.html
For info about this question over time, see
https://www.nytimes.com/interactive/2023/10/16/us/census-race-ethnicity.html

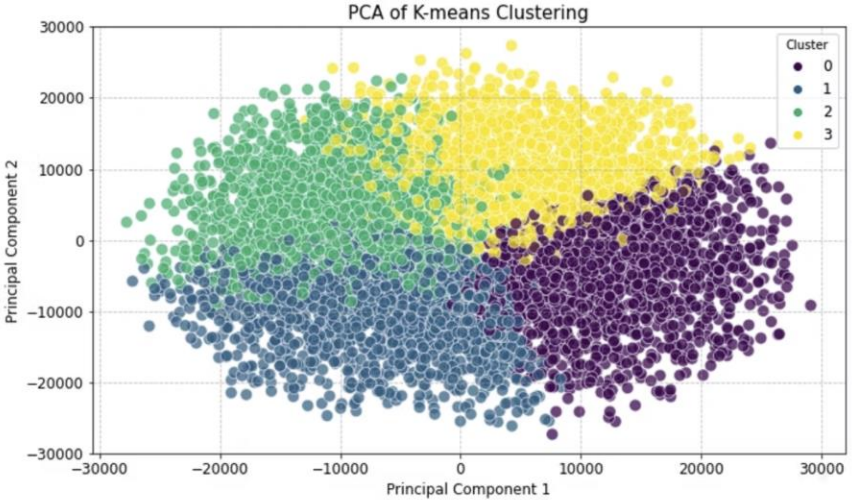
Computationally-Created Categories

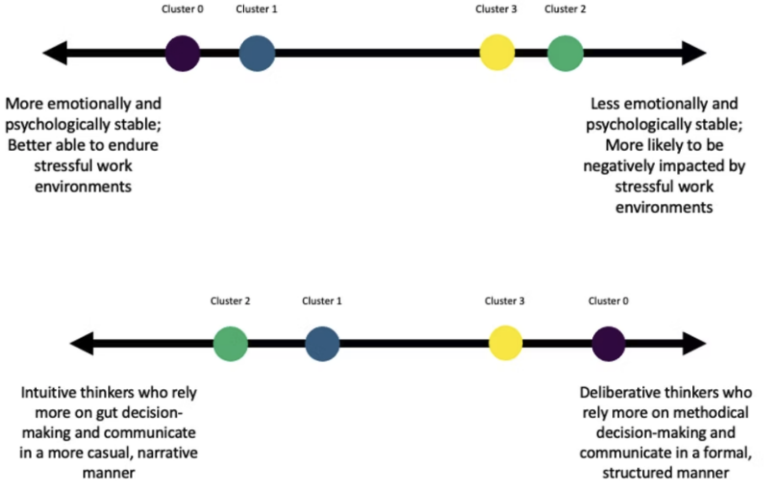
- Created computationally when a collection of resources or resource descriptions is too large for people to think about effectively
- "Supervised' machine learning algorithms can try to find patterns that characterize human-defined categories
- "Unsupervised" algorithms like clustering find correlations among attributes to create categories that may or may not be interpretable by people

Example: Computationally-created Categories via Cluster Analysis

Goal: determine personality trait groupings for doctors and nurses

1. Analyzed reddit posts by physicians; assigned scores for ~200 personality characteristics
2. k-Means Clustering of authors' posts according to these characteristics
3. Performed PCA to visualize the clusters
4. Analyzed the meaning of the clusters





https://www.receptiviti.com/post/harnessing-psycholinguistic-clusters-to-transform-people-insights-into-action

(Some) Principles for Intentionally Creating Categories

- Enumeration — Car Brands
- Single Property — Things that can be Climbed
- Multiple Properties — Expensive & Brittle Things
- Goal-based — Things you take on a camping trip
- Theory-based — Definition of a Triangle

Principle:
Defining Categories by Enumeration

Car Brands

The simplest way to define a category is by enumerating (listing) its members

The meaning of a category or concept is NOT a property; it is simply the specific set of resources

This principle is easy to understand and implement

To learn it, you have to memorize its members

16 German States

29 Indian States

Andhra Pradesh	AP
Arunachal Pradesh	AR
Assam	AS
Bihar	BR
Chhattisgarh	CT
Goa	GA
Gujarat	GJ
Haryana	HR
Himachal Pradesh	HP
Jammu and Kashmir	JK
Jharkhand	JH
Karnataka	KA
Kerala	KL
Madhya Pradesh	MP
Maharashtra	MH
Manipur	MN
Meghalaya	ML
Mizoram	MZ
Nagaland	NL
Odisha	OR
Punjab	PB
Rajasthan	RJ
Sikkim	SK
Tamil Nadu	TN
Telangana	TG
Tripura	TR
Uttar Pradesh	UP
Uttarakhand	UT
West Bengal	WB

Principle:
Defining Categories by A Single Property

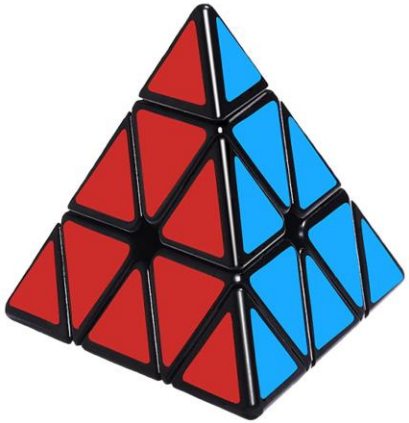
Climbable Things

Tall Things

Use only the values of any single property

Intrinsic static properties are often the easiest ones to use (color, size, shape…)

Single Property Category: "Pyramid- Shaped"



Principle:
Defining Categories by Multiple Properties

Things That Are Expensive & Breakable

Items being categorized can be described using observed "separable" or "combining" properties

If using one attribute only, these all go into the same category

But if using 3 attributes, they do not

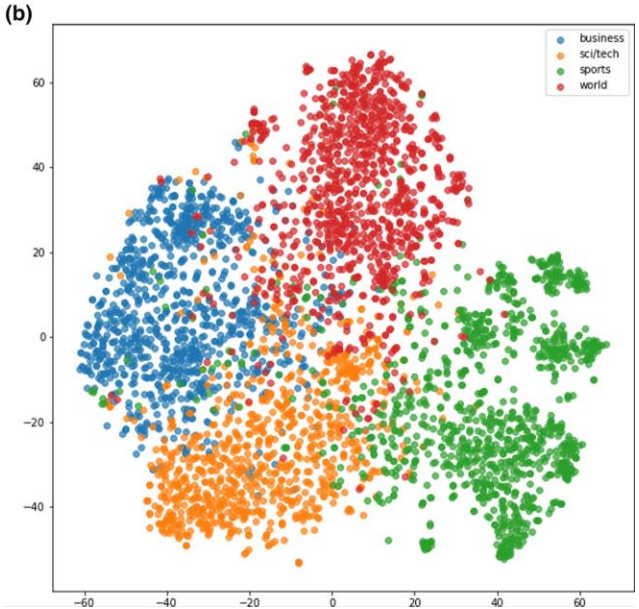
Pyramid Shape?	Human Made?	Location?
Yes	Yes	Egypt
Yes	No	Glacier Park
Yes	Yes	UC San Diego

Principle:
Defining Categories by a set of Rules

Example: Triangles are all and only 3-sided closed polygons

Example: In the game of baseball, a foul ball is a batted ball that lands or remains in foul territory, which is the area outside the foul lines extended from home plate past first and third base.

Principle: Defining Categories By Multiple Shared Statistical Properties



Example: Document clustering with BERT vectors

Subakti, A., Murfi, H. & Hariadi, N. The performance of BERT as data representation of text clustering. *J Big Data* **9**, 15 (2022). https://doi.org/10.1186/s40537-022-00564-9

Principle: Defining Categories By Goals

Things to Bring Camping Why You Wear a Tuxedo

The members of each of these categories have few or no discernable properties in common

These are unlikely to be lexicalized because of their ad hoc-ness and context-dependence

Classical vs Cognitive Science-Based Categories

Classical View of Categories

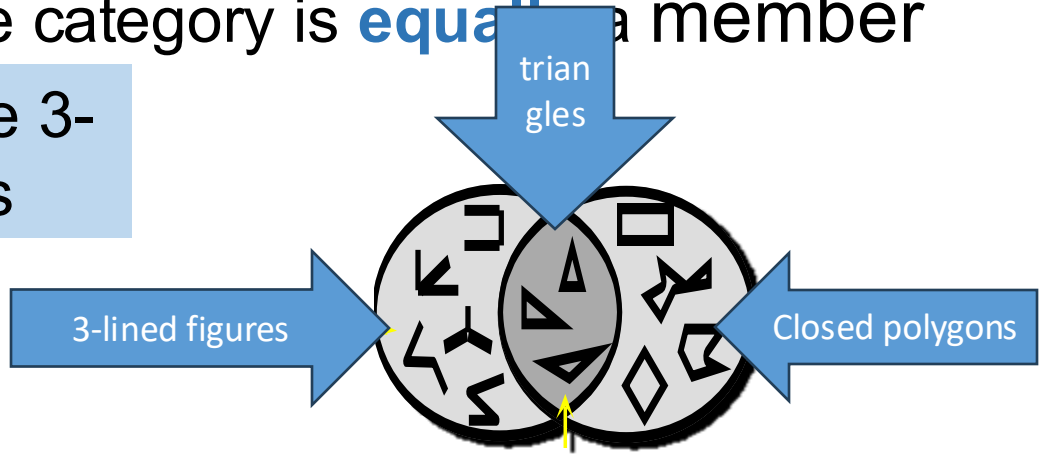
- Dates to Plato and Aristotle
- Platonic Ideal: There are two parallel worlds; one of ideal forms that reflect a higher truth; the other is physical reality
- These truths are discovered through logical reflection

Plato from Raphael's *The School of Athens* (1509–1511)

Classic View of Categories

1. Categories are defined by a list of properties shared by **all elements** in a category (necessary & sufficient)
2. Category membership is **binary** (in or out)
3. Because membership is defined by rules, **every** member in the category is **equa** a member

Example: triangles are 3-sided closed polygons



Classic View of Categories

1. Categories are defined by a list of properties shared by **all elements** in a category (necessary & sufficient)
2. Category membership is **binary** (in or out)
3. Because membership is defined by rules, **every** member in the category is **equally** a member

Example: define the category: “birds”

Everything that has feathers?

Everything that flies?

Every living thing with feathers that flies?

Name a category whose membership can be defined by a few precise rules

Classic View of Categories

1. Categories are defined by a list of properties shared by **all elements** in a category (necessary & sufficient)
2. Category membership is **binary** (in or out)
3. Because membership is defined by rules, **every** member in the category is **equally** a member

This view is

- Conceptually simple
- Straightforward for programs to implement
- Rules are very elaborate for real categories

But it is RARELY how people really categorize!

Name Some Games

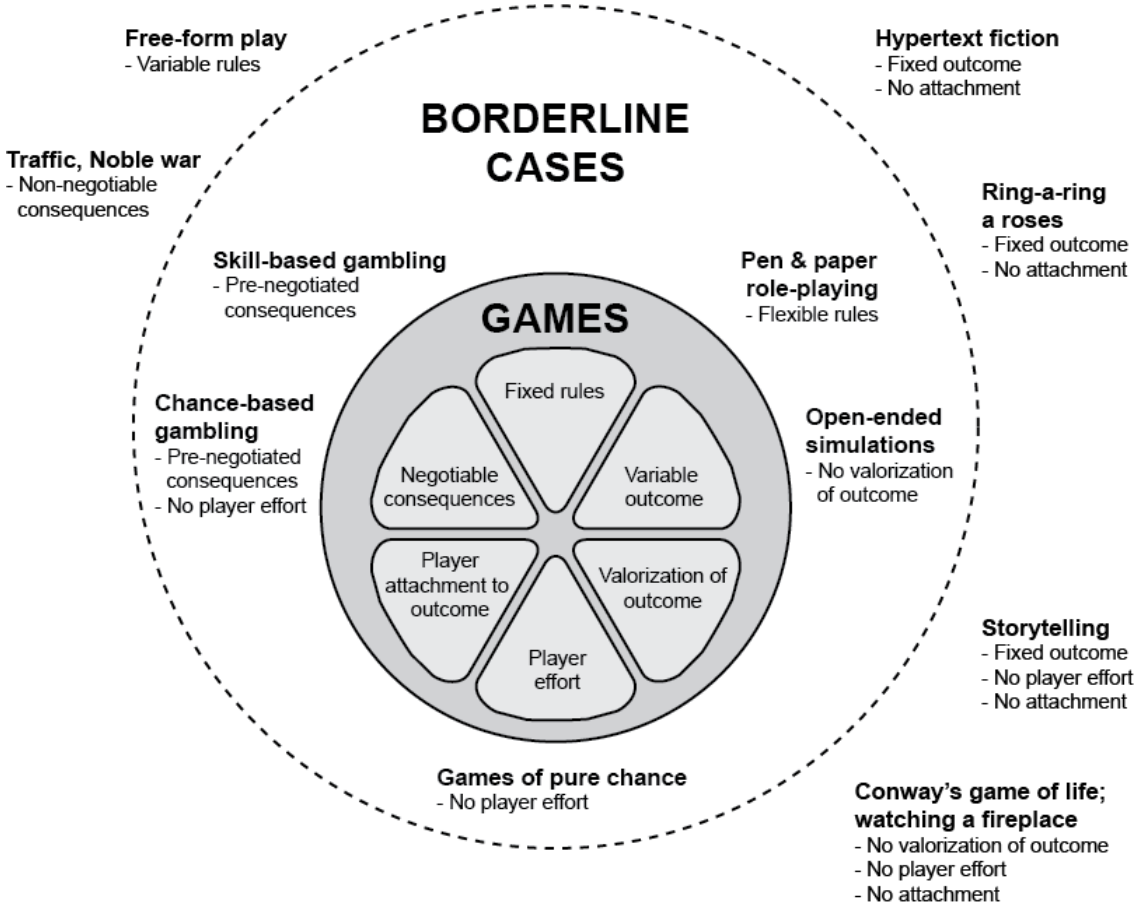
What Are **Necessary and Sufficient** Conditions for Something to be a Game?

A list of properties shared by **all elements?**

Wittgenstein's Famous Example of "Game"

- No common properties are shared by all games
    - Competition: card games, ball games, Olympic games
    - Developmental play: Children's games
    - Luck: dice games
    - Skill: chess
- No fixed boundary; can be extended to new games
    - Video games
- Alternative to Classical Categorization theory:
    - Concepts related by Family Resemblances

NOT GAMES



https://www.jesperjuul.net/ludologist/2015/06/10/what-is-a-game-redux/

Implications of Wittgenstein's Example

- There may be **defining features** for **typical members**
- But there are **no features** that are **necessary and sufficient** for **all examples** of the category
- Different members **vary substantially** in how typical or representative they are

Problems with the Classical View

It does not reflect how people categorize:

- People do **not** rely on abstract definitions or lists of shared properties (Rosch 1973)

  Example: Are curtains furniture?
- Some members are **more typical** than others

  Example: Chicken as a bird vs eagle as a bird
- At least some aspects of categorization seem to reflect the human body and mind

  Examples: Color categories, emotion categories

Next Time

- Cognitive / linguistic influence on categorization, cont.
- The institutionally created category of **genre**