Plan for Today's Lecture(s)

- The What and Why of Categories
- Cultural, Individual, and Institutional categories
- Principles for creating categories
- Classical categories... NOT

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

17 October 2013
Lecture 15.1 – The What and Why of Categories

How Do We Understand Each Other?

- It sometimes seems amazing that people can communicate because they organize and name the world in such different ways
- People establish a shared context through negotiated and iterative interactions and by providing assistance when possible
- People apply the recall / precision tradeoff, using more general categories to establish context and more specific ones when precise identification is required

Categories are "Equivalence Classes"

- Categories are sets or groups of resources or abstract entities that are treated the same
- This almost never means that every instance of the category is identical
- It only means that for some purpose we treat them in the same way

Categories are Essential

- Categories are involved whenever we perceive, communicate, analyze, predict, classify – or otherwise attempt to make sense of our experiences
- All human languages and cultures divide the physical and experiential "worlds" into categories
- Some of this "sense-making" is discovering categories and some is creating them, but these aspects are often interconnected

Categories are Created

- Categories can be created in many different ways according to many different types of principles
- The equivalence classes that are created depend on which aspects or roles or meanings of the category members are treated as relevant

Categories are Models

- Defining categories as equivalence classes in this way should remind you of this definition:

  Models are simplified descriptions of a subject that abstract from its complexity to emphasize some features or characteristics while intentionally de-emphasizing others

- Categories are COGNITIVE and LINGUISTIC MODELS for applying prior knowledge

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

17 October 2013
Lecture 15.2 – Category Categories,
or Contexts for Categorization

Category Categories

- Cultural Categorization
- Individual Categorization
- Institutional Categorization

Cultural Categories

- Embodied in culture and language, developing slowly and typically changing slowly
- Many have a perceptual or sensorimotor origin based on natural boundaries or discontinuities in perception and experience
- Learned implicitly through development via parent-child interactions, language, and experience
- Informal and contextual acquisition makes cultural categories flexible, creative, generative, and biased

Cultural Categories

- When properties or behaviors co-occur in predictable ways, it is useful to have category words that name them => emergence of language
- Formal education can teach cultural categories but the non-formal mechanisms often dominate or interfere
- Thinking and research about categorization is primarily about cultural categories

Linguistic Relativity

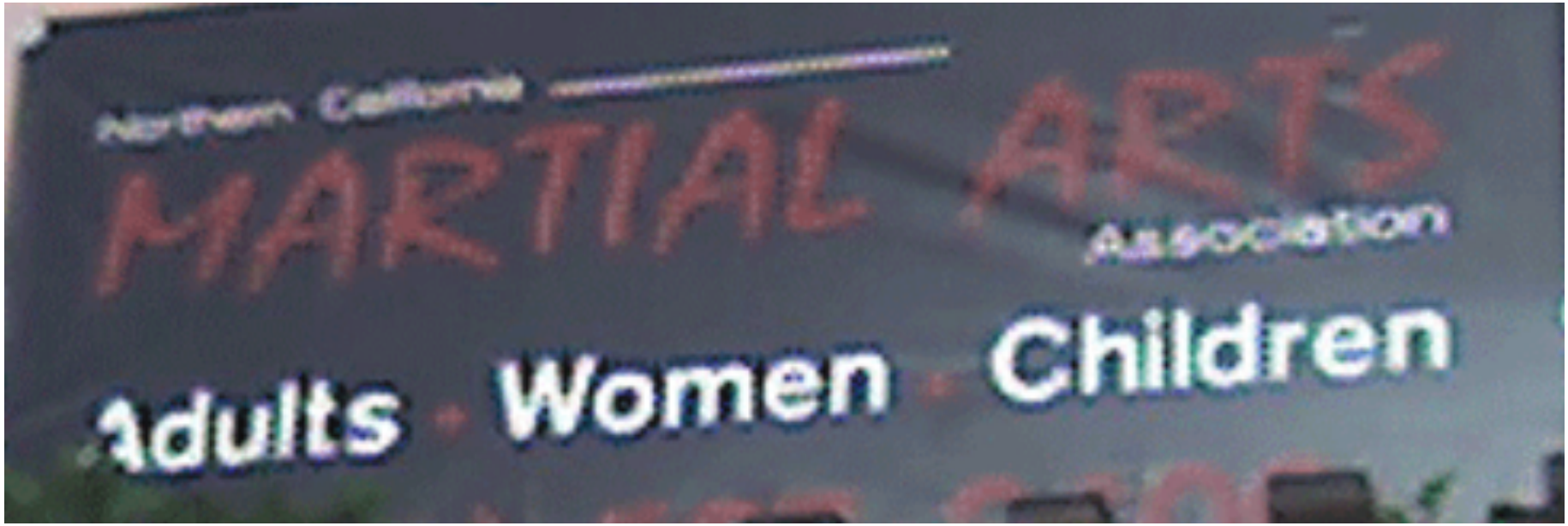
- Languages differ a great deal in the words they contain (which concepts are lexicalized)
- They also differ in more fundamental ways by requiring speakers or writers to attend to details about the world or aspects of experience that another language allows them to ignore
- But this doesn't mean that people can't understand concepts that are not lexicalized

Linguistic Relativity in Reykjavík

The Whorfian Hypothesis

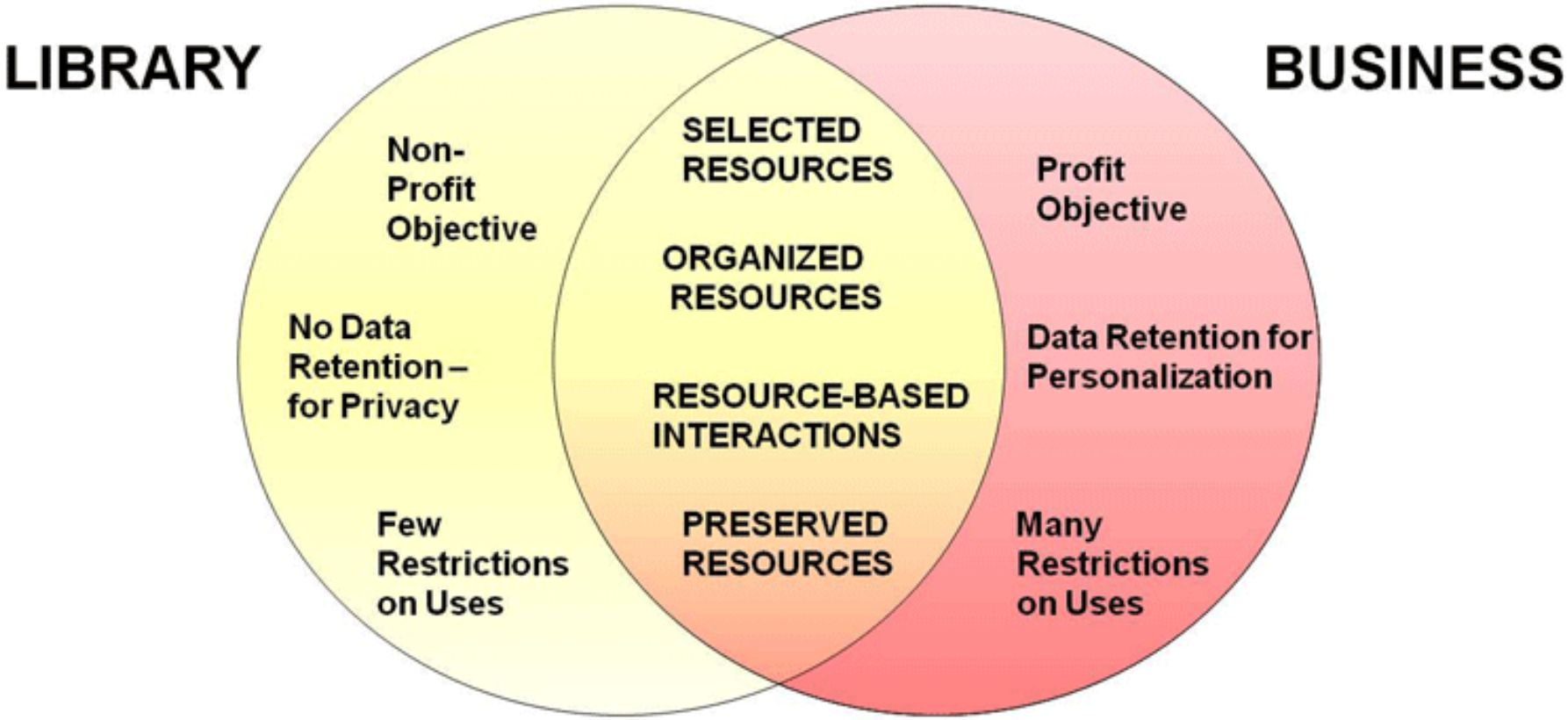
- Benjamin Whorf (mid 20th century) is widely credited and miscredited with the hypothesis that "language shapes thought" - that the categories embodied in languages shape or even constrain how people perceive and understand the world
- Cognitive science professor Lera Boraditsky has done some ingenious experimental tests of the Whorfian hypothesis - see http://www-psych.stanford.edu/~lera/papers/wsj.pdf for a very readable summary...

Cultural Categories Don't Have Precise Boundaries



… often causing ambiguity or the "vocabulary problem"

Cultural Category Boundaries Can Be Contentious



Is Google Books a Library or a Business?

Individual Categories

- Created to satisfy ad hoc requirements that emerge from an individual's unique experiences, preferences, and resource collections
- Created intentionally in response to specific organizing requirements
- Can have an imaginative or metaphorical basis that distorts or misinterpret cultural categories
- Collection hobbies often embody peculiar individual categories. *Chacun son goût…*

Individual Categories

- Often have short lifetimes because they are designed to support specific tasks
- When physical resources are involved, individual categories are rarely visible to or shared with others
- With digital resources, increasingly common for individual categories to be intentionally or accidentally revealed
- “Personal information management” is a subfield that focuses on individual organizing systems

Borges' Categorization of Animals

In a certain Chinese encyclopedia "it is written that 'animals are divided into:

(a) belonging to the Emperor, (b) embalmed, (c) tame, (d) suckling pigs, (e) sirens, (f) fabulous, (g) stray dogs, (h) included in the present classification, (i) frenzied, (j) innumerable, (k) drawn with a very fine camelhair brush, (l) et cetera, (m) having just broken the water pitcher, (n) that from a long way off look like flies.'“

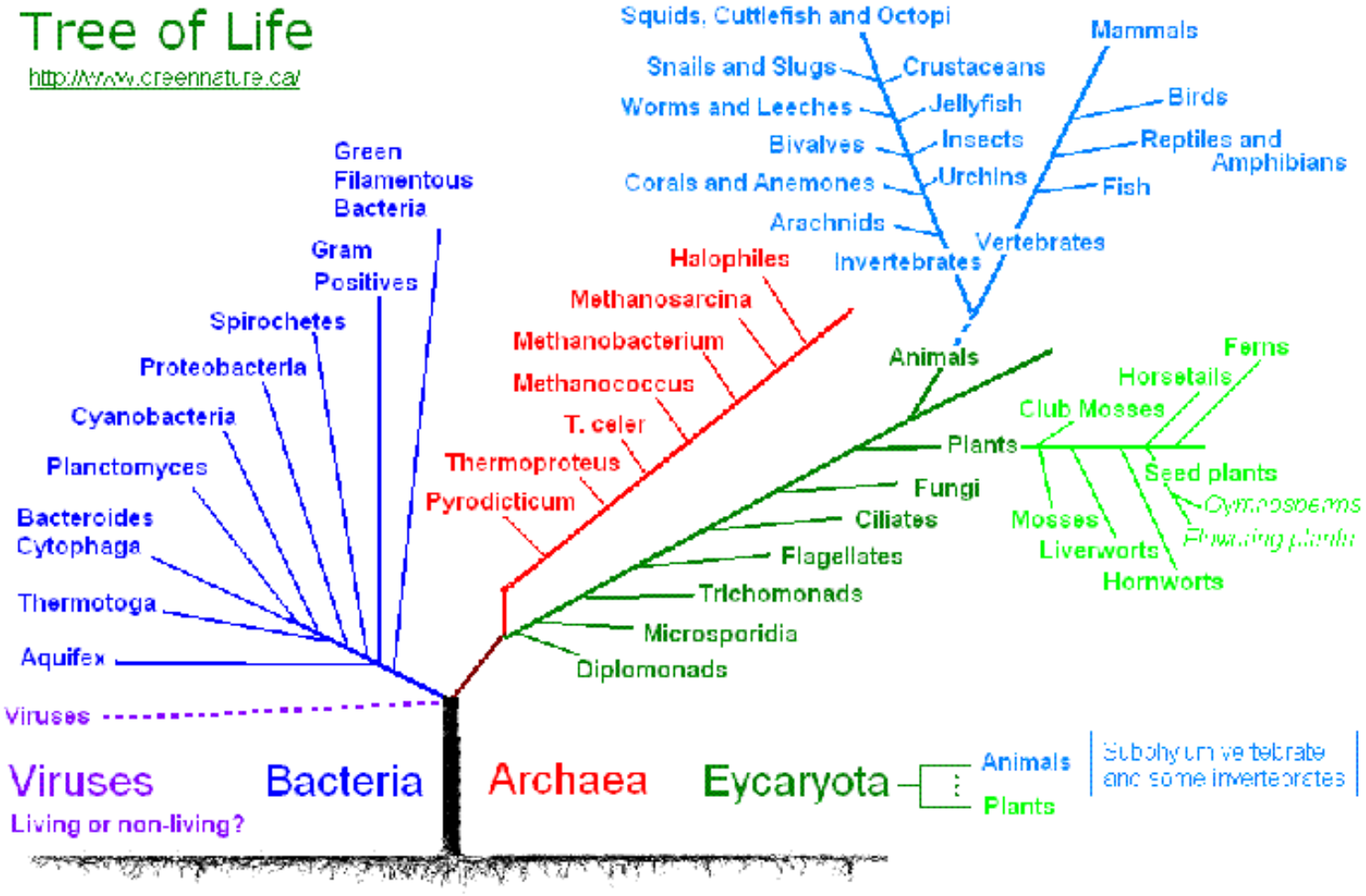
Institutional Categorization

- Explicit construction of a semantic model of a domain to enable more control, robustness, and interoperability than is possible with just the cultural system
- Essential in abstract, information-intensive domains where semantic precision is essential for processes and transactions (especially automated ones)
- Often the collaborative artifact of many individuals who represent different organizational or business perspectives

Institutional Categorization

- Usually developed via rigorous and formal processes (e.g., in standards organizations or legislative bodies)
- Laws, regulations, and standards often define institutional categories and the rules for working with them
- Require ongoing governance and maintenance because of continuous changes taking place in related cultural and individual systems

Institutional Categorization of Animals (Sorry Borges)



UN Standard Products and Services Codes for "Chicken"

UNv91201

Search Code:

Search Title: %chicken% Search

Return 100 Records (Maximum 800 Records)

#	ID	Name
1	10101601	Live chickens
2	23220000	Chicken processing machinery and equipment
3	50111515	Chicken, minimally processed without additions
4	50111520	Chicken, minimally processed with additions
5	50112010	Chicken, processed without additions
6	50112011	Chicken, processed with additions

Institutional != Unbiased

- Creating institutional categories by more systematic processes than cultural or individual categories does not prevent them from being biased
- Indeed, the goal of institutional categories is often to impose or incentivize biases in interpretation or behavior

"Gross Income" Tax Code Categories

(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) Alimony and separate maintenance payments;

(9) Annuities;

(10) Income from life insurance and endowment contracts;

(11) Pensions;

(12) Income from discharge of indebtedness;

(13) Distributive share of partnership gross income;

(14) Income in respect of a decedent; and

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

Should dividends be taxable income?

Dimensions of Variation Across "Categorization Contexts"

- Explicitness – how much awareness is required to use the categorization system
- Effort – how much effort is required
- Precision – "statistical" or formal specification
- Goals – whose purposes are served; individuals, informal groups, formal groups

Dimensions of Variation Across "Categorization Contexts"

- Interoperability – none, between people, between systems
- Reuse – from none to a great deal
- Change – slow, fast, unmanaged, managed
- Feedback – implications for misuse of the categories

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

17 October 2013
Lecture 15.3 – Principles for Creating Categories

Principles for Creating Categories

- Enumeration
- Single Properties
- Multiple Properties
- Family Resemblance
- Similarity
- Theory-Based
- Goal-Derived

Defining Categories by Enumeration

- Simplest way to define a category is by enumerating (listing) its members
- This principle is also called EXTENSIONAL definition; the members of the set are called the EXTENSION
- This principle is easy to understand and implement; the meaning of a category or concept is the specific set of resources associated with it

Defining Categories by Enumeration

- We understand something like "states of the US" as a category by listing all 50 of them
- Enumerative categories enable membership to be unambiguously determined, but at some scale they become impractical or inefficient, and the category either must be sub-divided or be given a definition based on principles other than enumeration

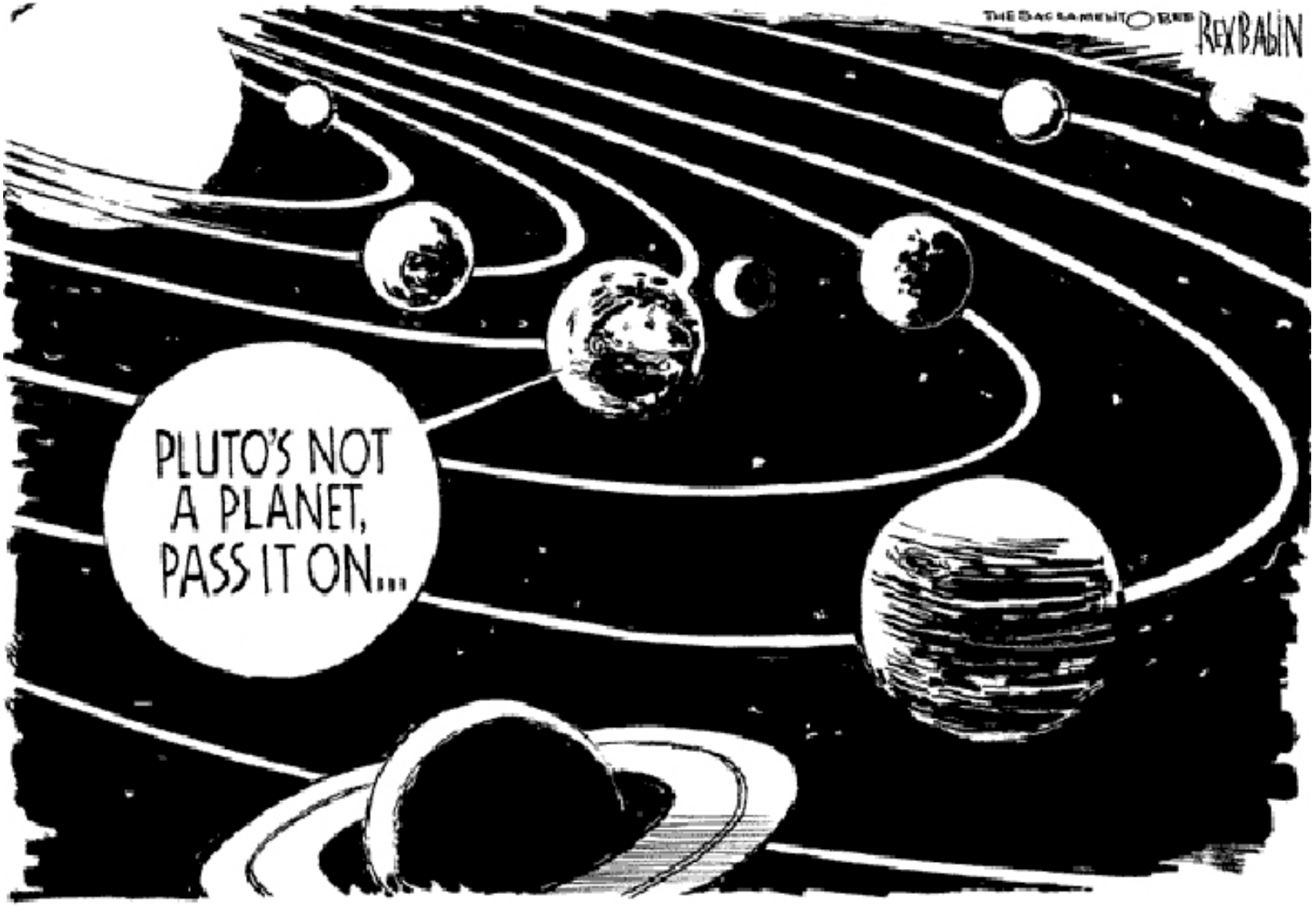
Frege's "Correspondence" Philosophy of Language

- Gottlob Frege (1848-1925), father of modern logic, investigated the idea of equivalence – how can you tell that two things are the same – in "Uber Sinn und Bedeuting"
- "Sinn" or "sense" (or "intension") – the inner concept that people understand; words have intensions
- "Bedeuting" or "reference" (or "extension") – the thing being referred to; the set of all objects in the world that can be described by the concept; intensions belong to extensions
- But if "meaning" was just based on reference would we need all that complicated ontology and common sense?

If Names Mean What They Refer To...

- Prune == Dried Plum
- Chinese Gooseberry == Kiwi Fruit
- Patagonian Toothfish == Antarctic Cod == Chilean Sea Bass
- Sectarian Conflict == Civil War
- Terrorists == Insurgents == Resistance == Freedom Fighters

Sorry, Pluto



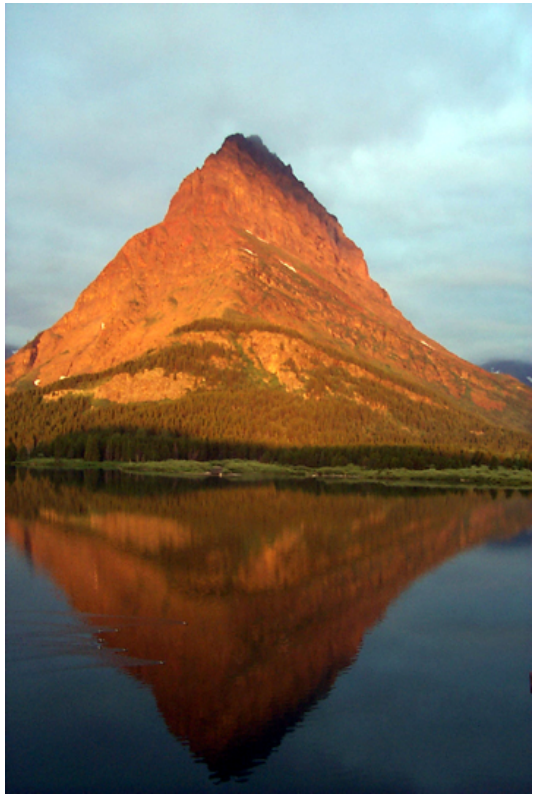
Categories Defined by Single Properties

- Simple category systems can be created using the values of any intrinsic static property, especially when the possible values are discrete and not very numerous
- But you can always define ranges for continuous value properties: low, moderate, high cost
- It is important to use properties that are psychologically or pragmatically relevant for the resource domain

Categories Defined by Single Properties

- Whenever a single property is used to define categories, the choice of property is critical (Gates and Zuckerberg example)
- A single-property category with a large numbers of items in it will lack coherence because differences on other properties will be too apparent

Single Property Category

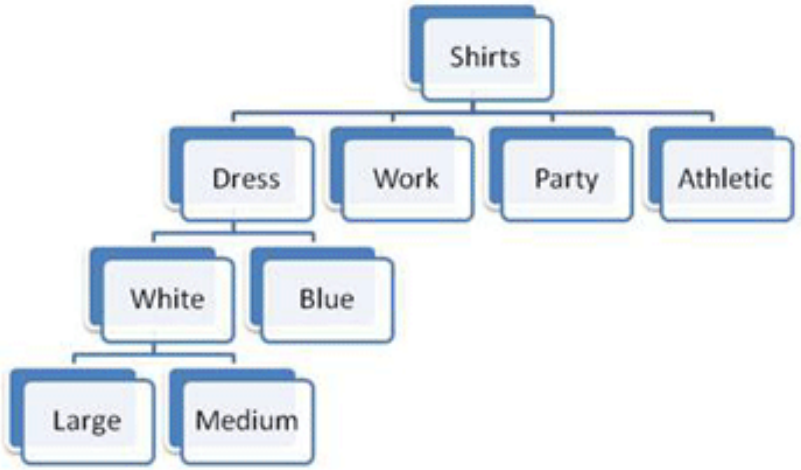


Single Property Category

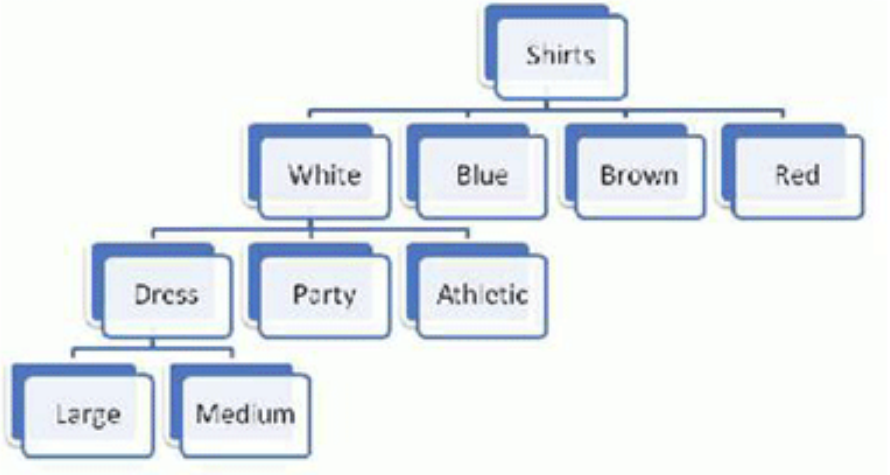
Categories Defined by Multiple Properties

- Categories based on multiple properties form a hierarchy
- Each successive property differentiates the items in a category to create sub-categories
- No more properties are needed if every item is in its own category or if the remaining differences are not important to us
- For information resources, easily perceived properties are much less useful than ABOUTNESS

Property Order Determines the Category Hierarchy



Style – Color - Size



Color – Style - Size

Category Hierarchies

- Categories can be organized into a hierarchy from the most general to the most specific
- My cat can be described as
  - An animal
  - A mammal
  - A cat
  - An American Shorthair
  - Boris

Basic-Level Categories

- In the middle of category hierarchies are those that more "basic" because the within-category differences are smallest and the between-category differences are the largest
- This means that perceptual, cognitive, and motor functions are "sharpest" at this level at identifying and thinking about category membership
- These are the categories when Plato “carves nature at the joints”

Basic-Level Categories

- Members tend to have the same shape
- They tend to "share parts"
- The same motor movements are involved in interacting with them

The Classical View of Categories

- Some categories have clear boundaries defined by a small number of ESSENTIAL or necessary and sufficient properties
- *Necessary* means that every instance must have the property to be in the category
- *Sufficient* means that any instance that has the necessary properties is in the category

The Classical View of Categories

- All members of the category have equal status in the equivalence class
- Example: A prime number is an integer divisible only by itself and 1

Wittgenstein

- Ludwig Wittgenstein (1889-1951) – "philosophy of ordinary language" - first to discuss problems with classical theory
- Dismantles Frege in "Philosophische Untersuchungen"
- Agrees with Frege that where the extensions have fixed characteristics or can be enumerated you can understand words by following the association to their extensions
- But rebuts Frege with argument that there are no fixed extensions for most words

Wittgenstein's Rebuttal to Frege: Meaning is Use

- There may be defining features for typical instances
- But there are no features that are necessary and sufficient for all examples of the category
- Even when features can be identified, they change in different contexts and over time
- Different instances vary substantially in how typical or representative they are of the category even though they share all the required features

Necessary and Sufficient Properties - Not!

- Wittgenstein's Counterexample is "Game"
  - No common properties are shared by all games
  - Some involve competition, others are cooperative
  - Some involve physical skill, others more mental skills, others luck
  - Some require equipment, others don't
  - Some involve teams, others are solitary
- No fixed category boundary - we can extend game category to include video games, networked games, word games, mind games, …

Family Resemblances

- Instances of a category often share many features, but some instances might have properties that are not widely shared
- These widely shared though not universal properties produce FAMILY RESEMBLANCES among the members
- Categories based on family resemblance in $20,000 Pyramid game

Family Resemblance

Chen Family

Allan Bradley
Power Supply Product Family

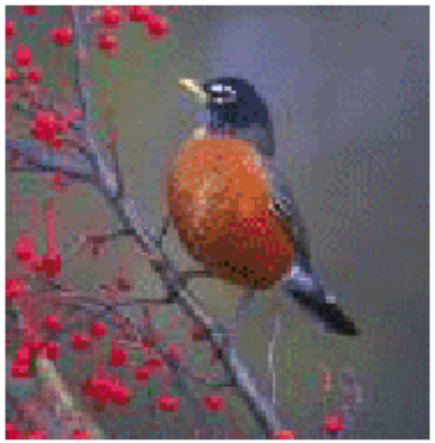
Characteristic Features

- Perceived degree of category membership has to do with which features help define the category
- Members usually do not have ALL the necessary features, but have some subset
- Those members that have more of the central features are seen as more central members

Gradience in Category Membership

- Not all members of a category are equally good examples
- The perceived centrality or typicality of category membership depends on the extent to which the most characteristic properties are shared
- "Someone says to me 'Show the children a game.' I teach them gaming with dice, and the other says 'I didn't mean that kind of game.' " (Wittgenstein)

Gradience in Category Membership



Which bird is most typical of the category?

Similarity

- It is easy to say that we can create categories on the basis of resource similarity
- But some people say that similarity is a meaningless and "mostly vacuous" concept because there must always be some unstated set of properties, and if we identify the properties and how they are used, a "similarity" mechanism has nothing to do

Theory-Based Categories

- Sometimes a category is defined because of a theory of causation or capability that explains why some collection of things go together
- A good theory can "trump" similarity or family resemblance based on surface properties
- Example: The category of "computer" is based on being able to compute, not on appearance

Goal-Derived Categories

- Imagine you are in a city about to be hit by a hurricane and you have 10 minutes to take any of your possessions and evacuate. How do you define the category of what to take?
- You might have a set of properties, but they are not likely to be ones that are easy to define: irreplaceable, priceless, sentimentally valuable, etc.
- They have no discernable properties in common
- Barsalou, Lawrence W. "Ad hoe categories." *Memory & cognition* 11, no. 3 (1983): 211-227.

Goal-Derived Retail Category?

Summary: Why Study Categorization?

- Categorization is central to how we organize information and the world, and categories are involved whenever we communicate, analyze, predict, or classify
- Whenever we design data structures, programming language class hierarchies, user or application interfaces, ...
- Categorization is much messier than our computer systems and applications would like
- But understanding how people (and each of us) categorize can help us design better systems and interfaces

Readings for Next Lecture

- TDO 7 through 7.2
- Tavris, Carol. How Psychiatry Went Crazy. Wall Street Journal, 17 May 2013
- Hoyt, Clark. Semantic minefields. New York Times, 15 May 2010.
- Prewitt, Kenneth. Fix the Census' Archaic Racial Categories. New York Times, 21 August 2013.
- Rosenthal, Arnon, Len Seligman, and Scott Renner. "From semantic integration to semantics management: case studies and a way forward." ACM SIGMOD Record 33, no. 4 (2004): 44-50.