College of Information Studies
University of Maryland Hornbake Library Building College Park, MD 20742-4345

Description

Week 5

LBSC 671

Creating Information Infrastructures

Types of “Metadata”

- Descriptive
  - Content, creation process, relationships
- Technical
  - Format, system requirements
- Usage
  - Display, derivative works
- Administrative
  - Acquisition, authentication, access rights
- Preservation
  - Media migration

Adapted from Introduction to Metadata, Getty Information Institute (2000)

Five “Levels” of Metadata

- Framework
  - Functional Requirements for Bibliographic Records (FRBR)
- Schema (“Data Fields and Structure”) 
  - Dublin Core
- Guidelines (“Data Content and Values”) 
  - Resource Description and Access (RDA)
  - Library of Congress Subject Headings (LCSH)
- Representation (abstract “Data Format”)
  - Resource Description Framework (RDF)
- Serialization (“Data Format”) 
  - RDF in eXtensible Markup Language (RDF/XML)

Adapted from Elings and Waibel, *First Monday*, (12)3, 2007

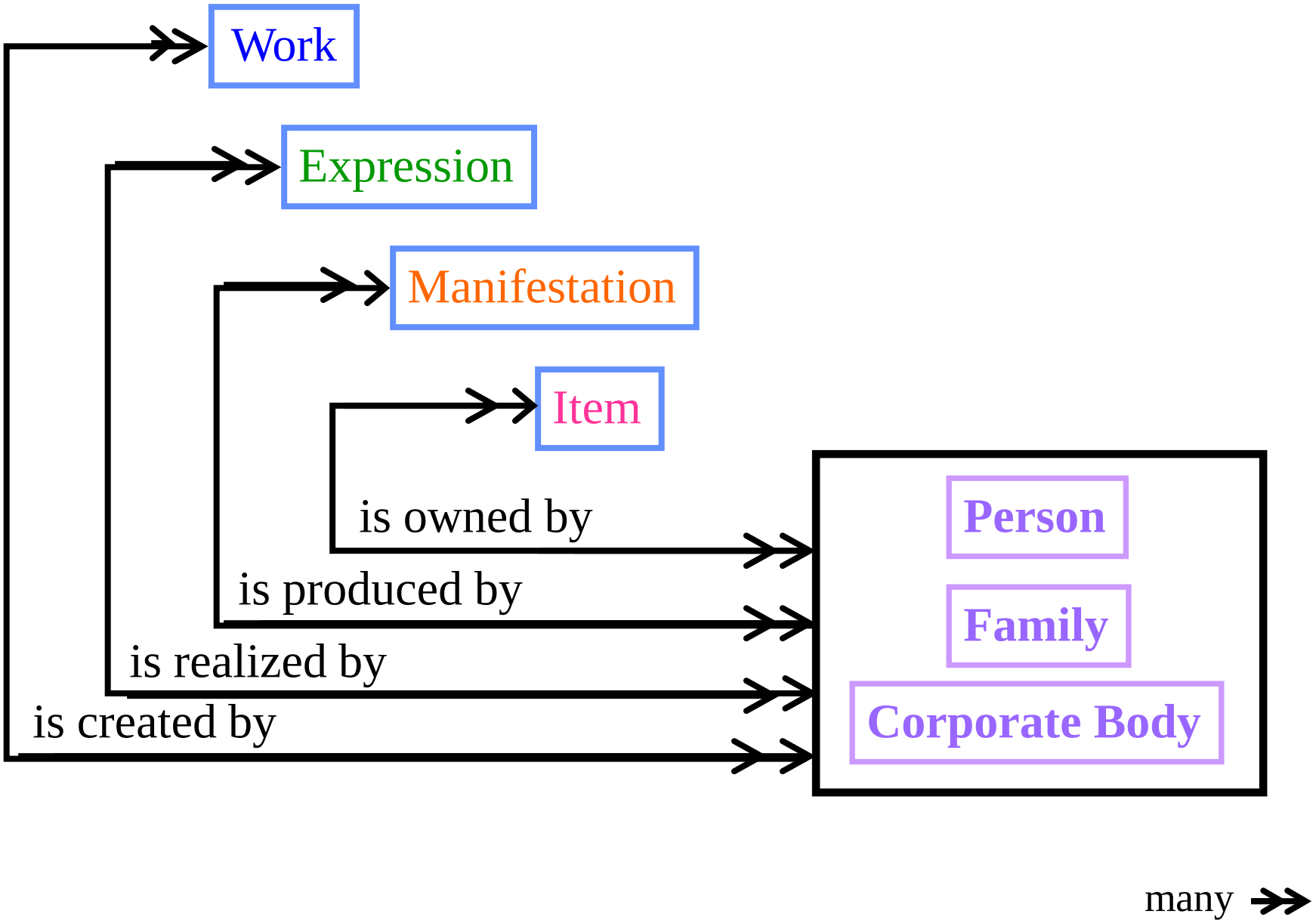
Fostering Consistency

- Content Standards
  - Resource Description and Access (RDA)
  - Describing Archives: a Content Standard (DACS)

- Authority Control
  - Subject Authority
  - Name authority

FRBR Entity Types

- Subject-Only Entities
  - (abstract) Concepts
  - (tangible) Objects
  - (any kind of) Places
  - Events
- Subject or Responsibility Entities
  - Persons
  - (any kind of) “Corporate” Bodies
  - Families (technically, only in FRAD)
- Product Entities
  - Works, Expressions, Manifestations, Items


Work
Expression
Manifestation
Item
is owned by
is produced by
is realized by
is created by
Person
Family
Corporate Body
many

Work

- The idea or impression in the mind of its creator
  - Completely abstract, no physical form

- What all forms, presentations, publications, or performances of a work have in common
  - *Romeo & Juliet*
  - Homer's *Odyssey*
  - Debussy's *Syrinx*

Expression (Realization)

- A work formulated into an ordered presentation
- When a work takes a *form*
  - Can be notational, aural, kinetic, etc.
- Excludes aspects of form not integral to the work
  - Font, layout, etc. (with some exceptions)
- Attributes: Form, Language

Manifestation

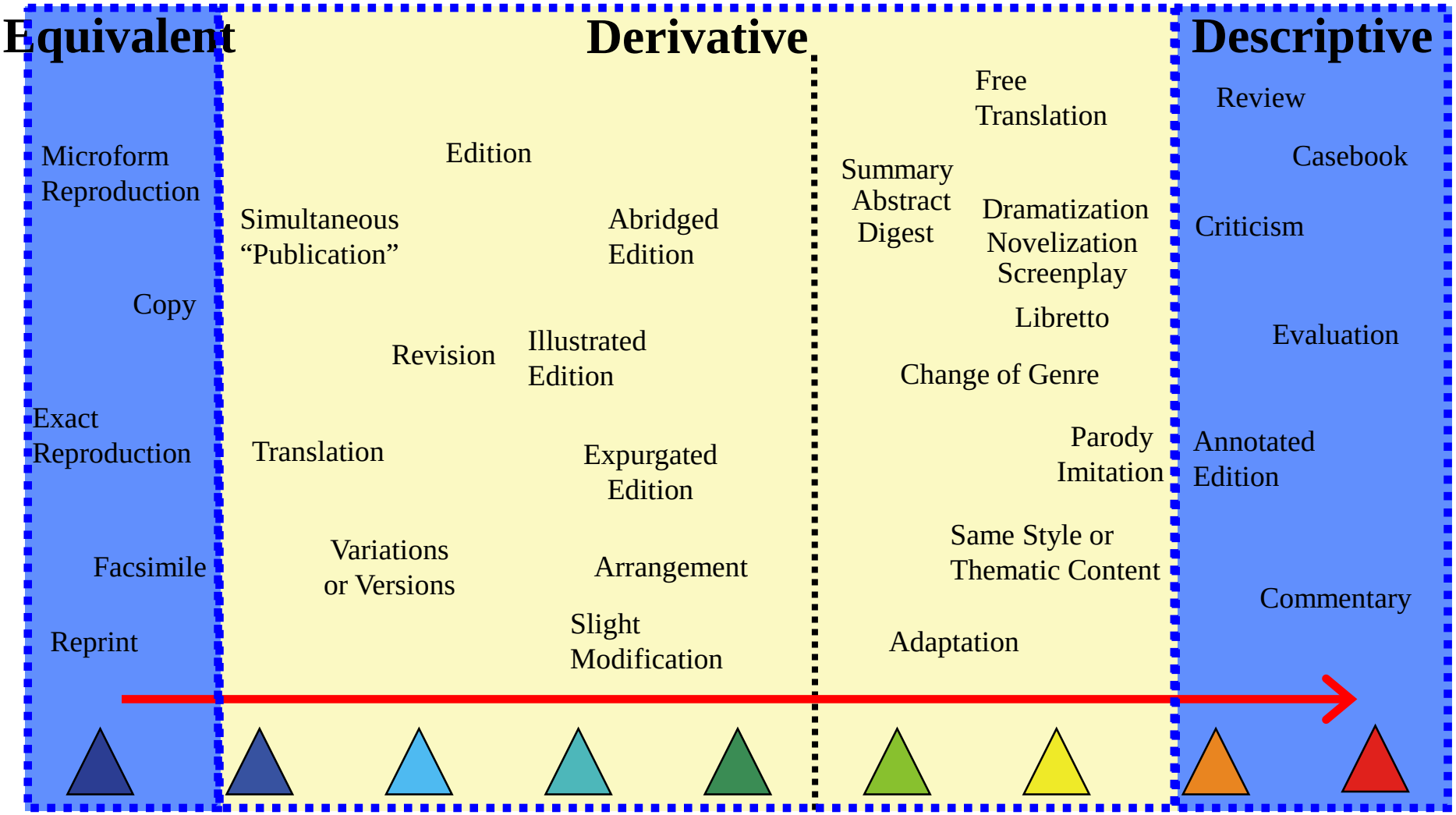
- Physical embodiment of an expression
  - The level usually described via cataloging
- **<u>Set</u>** of physical objects that bear the same:
  - *intellectual content* (expression), and
  - *physical form* (item)
- May have one or many items
  - Mona Lisa, Gone with the Wind, …
- Attributes
  - Format, Physical medium, Manufacturer

Item

- Instance of a manifestation
  - A *thing!*

- Attributes:
  - Owned by, Location, Condition

Family of Works



Original Work - Same Expression

Same Work – New Expression

Cataloging Rules Cut-Off Point

New Work

RDA for Georgia, 2011

FRBR Bibliographic User Tasks

- Find it
  - Search (“to find”)
  - Recognize (“to identify”)
  - Choose (“to select”)
- Serve it
  - Location (“to obtain”)

Resource Description & Access (RDA)

- RDA metadata describes entities *associated with* a resource to help users perform the following tasks:
  - **Find** information on that entity and on resources associated with the entity
  - **Identify**: confirm that the entity described corresponds to the entity sought, or to distinguish between two or more entities with similar names, etc.
  - **Clarify** the relationship between two or more such entities, or to clarify the relationship between the entity described and a name by which that entity is known
  - **Understand** why a particular name or title, or form of name or title, has been chosen as the preferred name or title for the entity

Components of RDA

- "Elements" (Attributes)
  1. Of manifestations and items
  2. Of works and expressions
  3. Of persons and corporate bodies
  4. Of concepts
- Relationships
  5. Among product entities
     - Content entities: work, expression, manifestation, item
  6. Between product and responsibility entities
     - Responsibility entities: person, family, corporate body
  7. Between works and subject entities
     - Subject entities: concepts, objects, places, events

Bibliographic Relationships

- Equivalence: exact (or nearly exact) copies
  - mp3 recording burned from a CD, …
- Derivative: work based on/derived from another
  - Updated edition, adaptation, …
- Descriptive: work that describes another work
  - Criticism, commentary, summary (e.g., Cliffs Notes), …

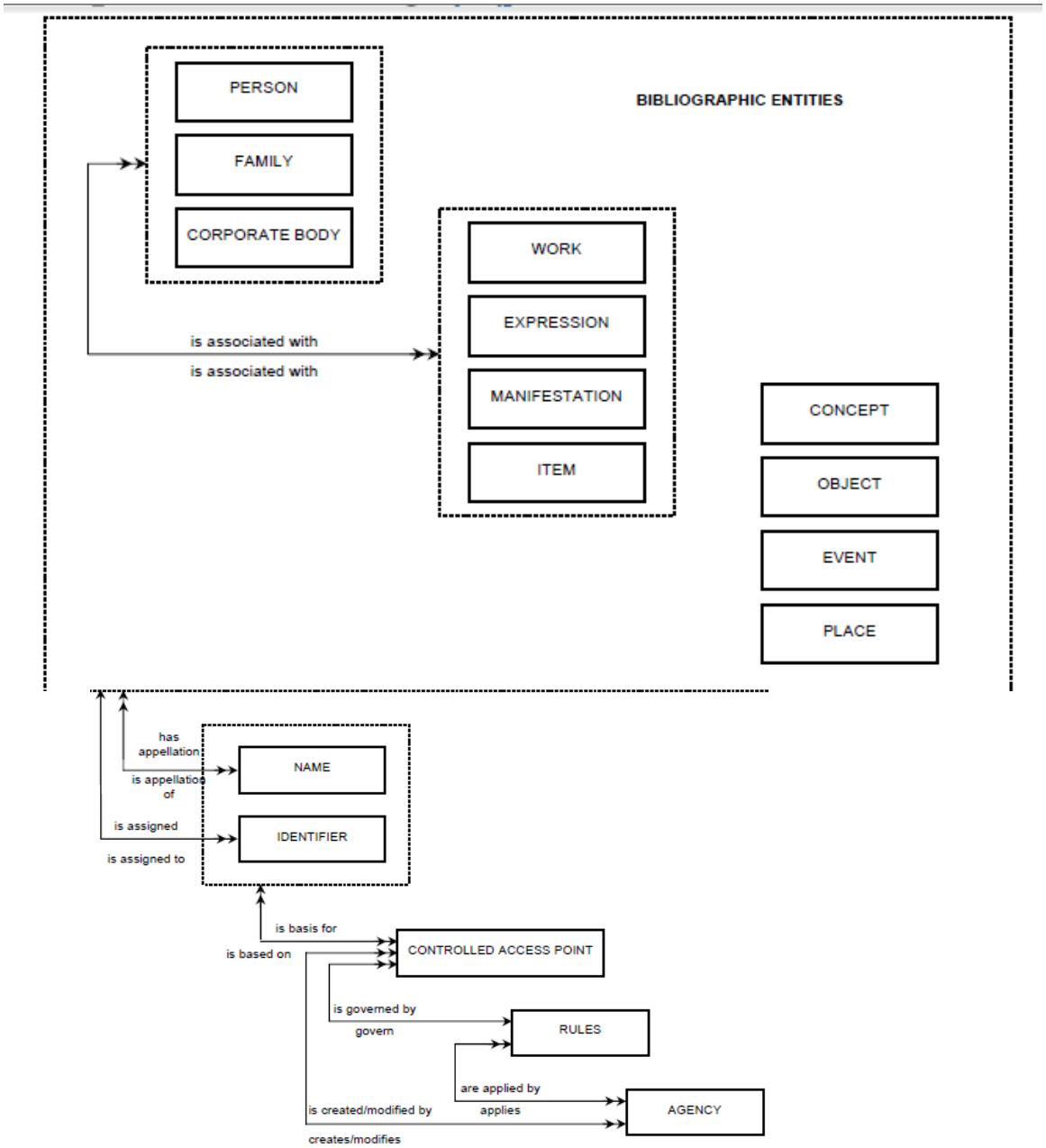
More Bibliographic Relationships

- Whole-part: One work is part of another work
  - Volume in an encyclopedia, chapter in a book, …
- Accompanying: A work meant to go with another work
  - Math workbook w/ textbook, index, documentation, …
- Sequential: Work precedes/continues an existing work
  - Issues of a publication, sequels/prequels, …
- Shared characteristic: Something in common
  - Author, title, language, subject, …

Authority Control

- Unify references to the same entity (synonyms)
  - Samuel Clemens, Mark Twain
- Distinguish references to different entities (homonyms)
  - Michael Jordan (basketball), Michael Jordan (computers)
- Establish “access points”
  - Canonical and variant forms, to better support “find it” tasks

Functional Requirements for Authority Data



IFLA, 2013

Some RDA Elements for Products

- Work
  - ID
  - Title
  - Date
  - etc.

- Expression
  - ID
  - Form
  - Date
  - Language
  - etc.

- Manifestation
  - ID
  - Title
  - Statement of responsibility
  - Edition
  - Imprint (place, publisher, date)
  - Form/extent of carrier
  - Terms of availability
  - Mode of access
  - etc.

- Item
  - ID
  - Provenance
  - Location
  - etc.

RDA: Person

- “An individual or an identity established by an individual (either alone or in collaboration with one or more other individuals)”
- Includes fictitious entities
  - Miss Piggy, Snoopy, etc. in scope if presented as having responsibility in some way for a work, expression, manifestation, or item
- Also includes real non-humans
  - Only in US RDA test

RDA Person Examples

```
100 0# $a Miss Piggy.
245 10 $a Miss Piggy’s guide to life / $c
       by Miss Piggy as told to Henry Beard.
700 1# $a Beard, Henry.
```

```
100 0# $a Lassie.
245 1# $a Stories of Hollywood / $c told
       by Lassie.
```

RDA: Language and Script

- Names:
  - USA: In authorized and variant access points, apply the alternative to give a romanized form.
  - For some languages, can also give variant access points in original language/script

- Other elements:
  - If RDA instructions don’t specify language, give element in English

RDA: Preferred Name

- Used as the “authorized” (i.e., canonical) access point
- Choose the form most commonly known
- Variant spellings:
  - Choose the form found on the first resource received
- If individual has more than one identity
  - Construct a preferred name for each identity

RDA: Additions to Preferred Name

- title or other designation associated with person
- date of birth and/or death * ^
- fuller form of name * ^
- period of activity of person * ^
- profession or occupation *
- field of activity of person *

* = if need to distinguish; ^ = option to add even if not needed

RDA: Surnames Indicating Relationships

- Include words, etc., (e.g., Jr., Sr., IV) in preferred name – not just to break conflict

```
100 1# $a Rogers, Roy, $c Jr., $d 1946-
670 ## $a Growing up with Roy and Dale, 1986:
       $b t.p.(Roy Rogers, Jr.) p. 16 (born
        1946)
```

RDA: Terms of Address <u>When Needed</u>

- When the name consists only of the surname
  - (Seuss, Dr.)
- For a married person identified only by a partner's name and a term of address
  - (Davis, Maxwell, Mrs.)
- If part of a phrase consisting of a forename(s) preceded by a term of address
  - (Sam, Cousin)

RDA: Profession or Occupation

- Core:
  - for a person whose name consists of a phrase or appellation not conveying the idea of a person**, or**
  - if needed to distinguish one person from another with the same name

- Overlap with “field of activity”

```
100 1# $a Watt, James $c (Gardener)
```

RDA: Field of Activity of Person

- Field of endevor, area of expertise, etc., in which a person is or was engaged

- Core:
  - For a person whose name consists of a phrase or appellation not conveying the idea of a person, or
  - If needed to distinguish one person from another with the same name

```
100 0# $a Spotted Horse $c (Crow Indian chief)
```

RDA: Associated Date for Person

- Three dates:
  - Date of birth
  - Date of death
  - Period of activity of the person

- Guidelines for <u>probable</u> dates are in RDA 9.3.1

RDA: Associated Place for Person

- Place of birth
- Place of death
- Country associated with the person
- Place of residence

DACS Principles

1. Records in archives possess unique characteristics.
2. The principle of respect des finds is the basis of archival arrangement and description.
3. Arrangement involves identification of groupings within material.
4. Description reflects arrangement.
5. The rules of description apply to all archival materials regardless of form or medium.
6. The principles of archival description apply equally to records created by corporate bodies, individuals, or families.
7. Archival descriptions may be presented at varying levels of detail to produce a variety of outputs.
8. The creators of archival materials, as well as the materials themselves, must be described.

(Single-Level) DACS Elements

Required

- Reference code
- Name+location of repository
- Title
- Date
- Extent
- Name of creator(s)
- Scope and content
- Conditions governing access
- Languages and scripts
- Plus, for “Optimal”
  - Administrative/biographical history
  - Access points

Optional

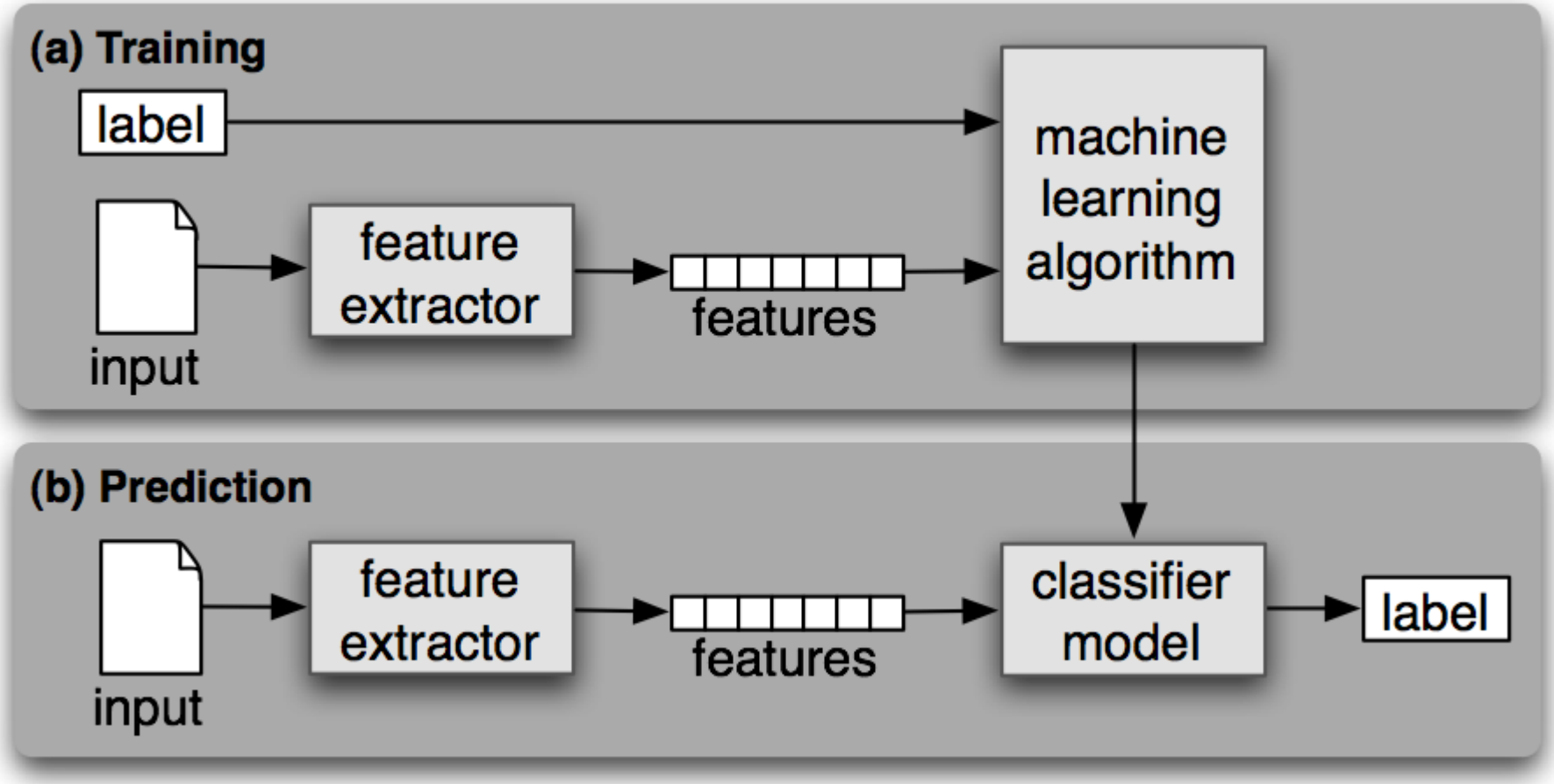
- System of arrangement
- Physical access
- Technical access
- Conditions for reproduction and use
- (other) Finding aids
- Custodial history
- Immediate source of acquisition
- Appraisal, destruction, scheduling
- Accruals (anticipated additions)
- Existence+location of originals
- Existence+location of copies
- Related archival materials
- Publication note
- Notes
- Description control

Modeling Use of Language

- Normative
  - Observe how people **<u>do</u>** talk or write
    - Somehow, come to understand what they mean each time
  - Create a **<u>theory</u>** that associates language and meaning
  - Interpret language use based on that theory

- Descriptive
  - Observe how people **<u>do</u>** talk or write
    - Someone “trains” us on what they mean each time
  - Use **<u>statistics</u>** to learn how those are associated
  - Reverse the model to guess meaning from what’s said

Supervised Machine Learning



Steven Bird et al., *Natural Language Processing*, 2006

Some Examples of Features

- Topic
  - Counts for each word
- Sentiment
  - Counts for each word
- Human values
  - Counts for each word
- Sentence splitting
  - Ends in one of .!?
  - Next word capitalized
- Part of speech
  - Word ends in –ed, -ing, …
  - Previous word is a, to, …
- Named entity
  - All+only first letters caps
  - Next word is said, went, …
- Gender of person name
  - Last letter

Metadata Extraction: Named Entity "Tagging"

- Machine learning techniques can find:
  - Location
  - Extent
  - Type

- Two types of features are useful
  - Orthography
    - e.g., Paired or non-initial capitalization
  - Trigger words
    - e.g., Mr., Professor, said, …

Your query has finished

Rough'n'Ready GTE

Search

Clear

OR

AND

Topic

Organization

Speaker

Story Jewish-Arab relations : Politics and government : Palestinian Arabs : Middle East : Israel : Terroris

Person

Location

Text

5 stories about: Jewish-Arab relations : Politics and government : Palestinian Arabs : Middle East : Israel : Terr

Jewish-Arab relations : Politics and government : Palestinian Arabs : Middle East : Israel : Terrorism : Pale

Jewish-Arab relations : Israel : Middle East : Middle East peace negotiations : Politics and government : P

male 5

Well as all work during president Clinton's trip to New York tonight and he enjoys the performance of the opera Carmen at Lincoln Center and see the scene there is a lot of Broadway. Now earlier today Mr. Clinton announced that the UN united Nations general assembly that he plans to send a nuclear test ban treaty to the Senate the treaty bans all nuclear test explosions and is regarded as a milestone in the arms control. Two israeli security guards were wounded in an early morning shooting in Jordan a government official says three men and a car opened fire on the guard's car wounding both before Skipping guards were treated at a hospital and released it is real several West Bank villages were sealed by israeli soldiers who search for the islamic militants behind two recent suicide bombings in Jerusalem palestinian leader Yasser Arafat says that he believes those was counsel for the bombing case and abroad.

Jewish-Arab relations

Middle East peace negotiations

Middle East

Palestinian self-rule areas

Israel

Politics and government

Arafat, Yasir

Palestinian Arabs

Gender Classification Example

```
>>> classifier.show_most_informative_features(5)
Most Informative Features
last_letter = 'a' female : male = 38.3 : 1.0
last_letter = 'k' male : female = 31.4 : 1.0
last_letter = 'f' male : female = 15.3 : 1.0
last_letter = 'p' male : female = 10.6 : 1.0
last_letter = 'w' male : female = 10.6 : 1.0
```

```
>>> for (tag, guess, name) in sorted(errors):
print 'correct=%-8s guess=%-8s name=%-30s'
correct=female guess=male name=Cindelyn ...
correct=female guess=male name=Katheryn
correct=female guess=male name=Kathryn ...
correct=male guess=female name=Aldrich ...
correct=male guess=female name=Mitch ...
correct=male guess=female name=Rich ...
```

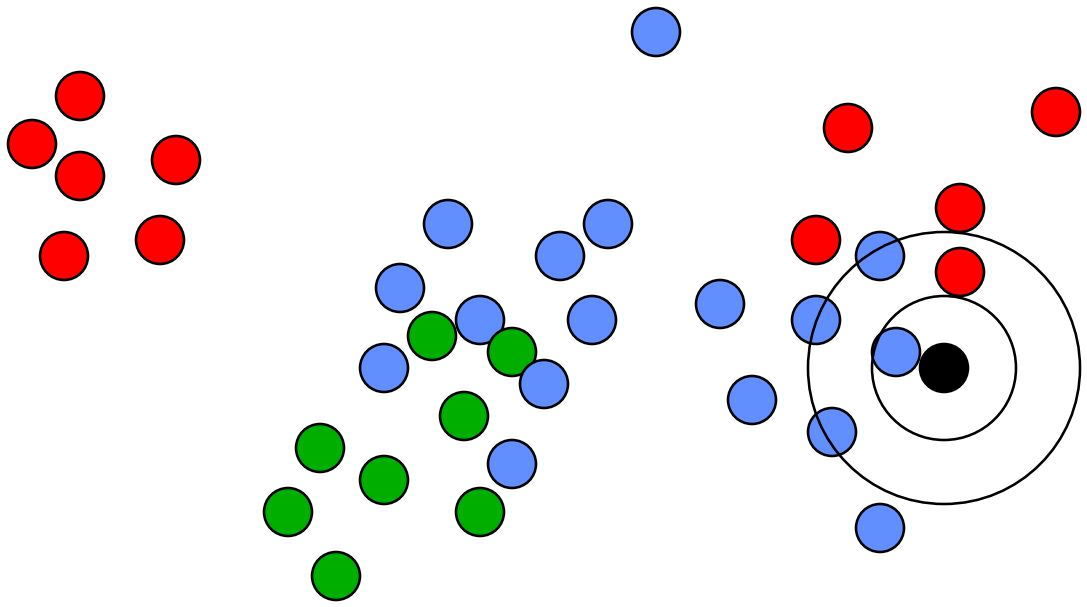
Sentiment Classification Example

```
>>> classifier.show_most_informative_features(5)
Most Informative Features
contains(outstanding) =     True pos : neg = 11.1 : 1.0
contains(seagal) =          True neg : pos =  7.7 : 1.0
contains(wonderfully) =     True pos : neg =  6.8 : 1.0
contains(damon) =           True pos : neg =  5.9 : 1.0
contains(wasted) =          True neg : pos =  5.8 : 1.0
```

Supervised Learning Techniques

- Decision Tree
  - Explainable (near the top)
- Naïve Bayes
  - Efficient training
- Maximum Entropy
  - Good use of limited training data
- k-Nearest-Neighbor
  - Easily extended to multi-class problems

Machine Learning for Classification: The k-Nearest-Neighbor Classifier



Supervised Learning Limitations

- Rare events
  - It can’t learn what it has never seen!
- Overfitting
  - Too much memorization, not enough generalization
- Unrepresentative training data
  - Reported evaluations are often very optimistic
- It doesn’t know what it doesn’t know
  - So it always guesses some answer
- Unbalanced “class frequency”
  - Consider this when deciding what’s good enough

Before You Go!

- On a sheet of paper (no names), answer the following question:

  What was the muddiest point in today’s class?