Evaluation of Library Services

Min-Yen KAN

Why Evaluation?

- Run as a business, need to justify costs and expenditure
- Quantitative data analysis necessitated by evolution into automated and digital libraries
- Need benchmarks to evaluate effectiveness of library

Quantitative metrics

- Circulation per capita
- Library visits per capita
- Program attendance per capita
- Turnover rate
- Registration as % of population

- *Output measures for public libraries*
Zweizig and Rodger (1982)

Evaluation types

- Macroevaluation
  - Quantitative, comparable statistics
  - Degree of exposure
- Microevaluation
  - Diagnostic
  - Gives rationale for performance
- Materials-Based / Use-based
  - Evaluate the items' suitability

Exposure

- Axiom
  - The more a book in a library is exposed, the more effective the library.
- Defining “an exposure” as a simple count
  - Pros
    - Easy; can handle different levels of granularity
  - Cons
    - 5 × 1 day borrowing is five times more exposure than 1 × 5 day borrowing
    - Shorter circulation would increase counts

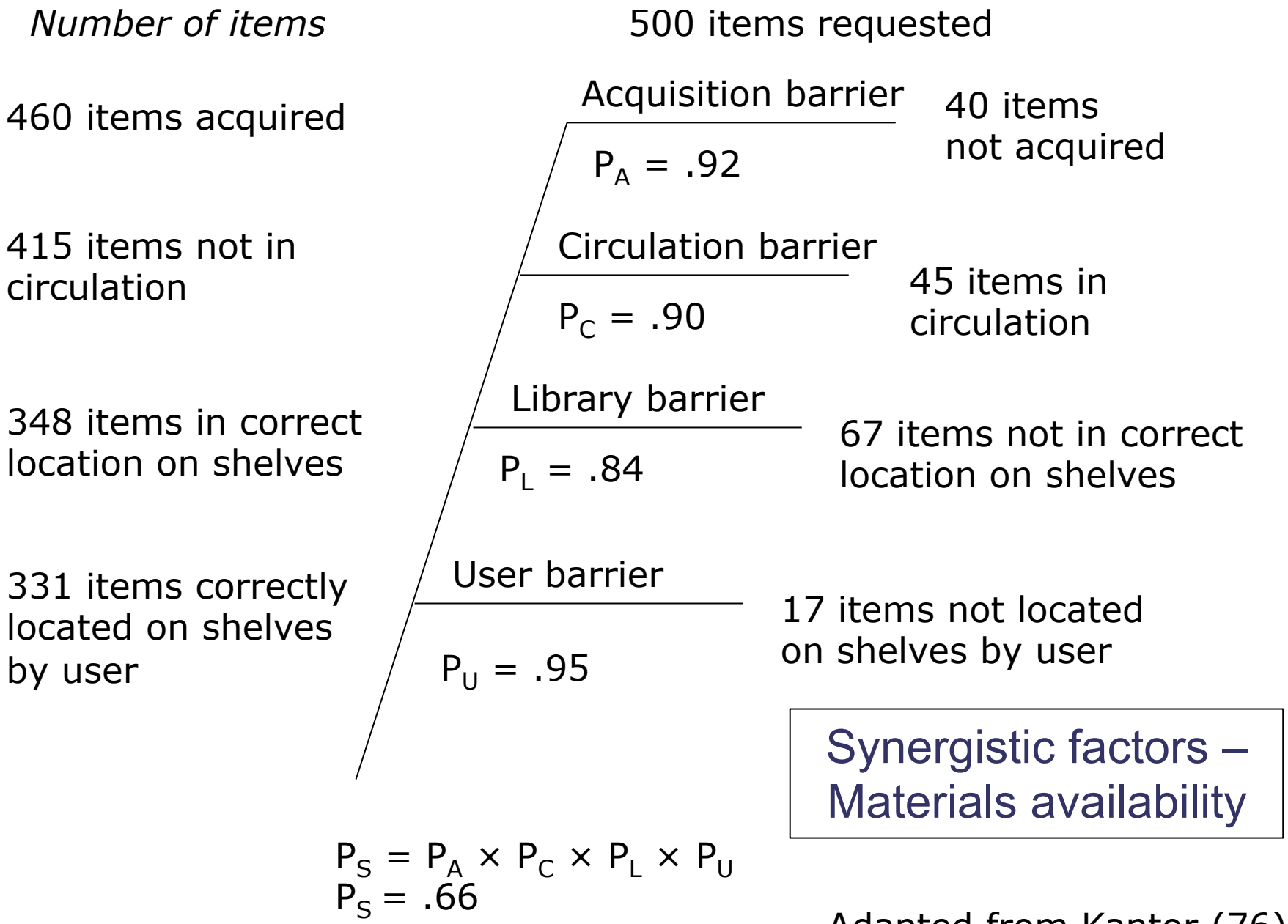
More exact ways to quantify exposure

- Item-use days: Meier (61)
  - A book borrowed for five days may not be used at all
- Effective user hours: De Prospo *et al.* (73)
  - Sample users in library

What about ways to quantify exposure in the digital library?

Bang for the buck?

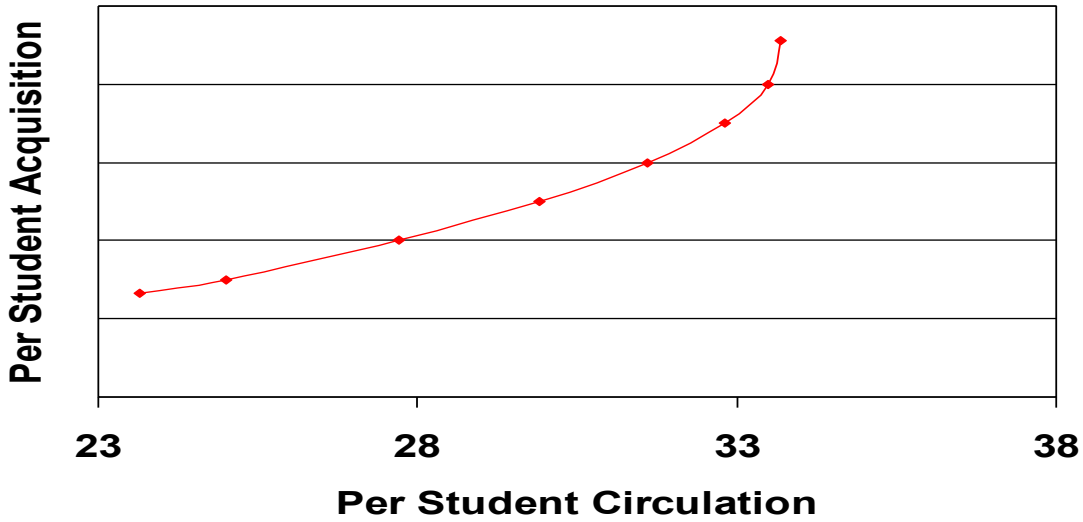
_____,
the greater the exposure.


Number of items
500 items requested
460 items acquired
Acquisition barrier
40 items not acquired
P_A = .92
415 items not in circulation
Circulation barrier
45 items in circulation
P_C = .90
348 items in correct location on shelves
Library barrier
67 items not in correct location on shelves
P_L = .84
331 items correctly located on shelves by user
User barrier
17 items not located on shelves by user
P_U = .95
Synergistic factors – Materials availability
$P_S = P_A \times P_C \times P_L \times P_U$
P_S = .66
Adapted from Kantor (76)

Effectiveness as Circulation

- Need a minimal size to function at all
- The larger the collection the better…

  … to a point



- From Hodowanec (78)

Macroevaluation

- In general, more exact measures require aggregating *sampling*, which tend towards microevaluation
  - So it's a continuum after all

- Administrators use a battery of measures; not a single one, to measure effectiveness – Spray (76)

Microevaluation

Drilling down to the individual needs level

- The more concrete the need, the easier to evaluate
- Failure is harder to measure than success
  - Case 1: Got a sub-optimal resource
  - Case 2: Got some material but not all

Material-centered collection evaluation

What's the purpose…

… of the collection

- Who's the readership – academic, public?

… of the evaluation

- Document change in demand?
- Justify funding?
- Select areas to weed materials?
- Adjust shelving/organization?

Material-based evaluations

- Checklist
  - Use standard reference bibliographies to check against
- Citation
  - Use an initial seed of resources to search for resources that cite and are cited by them

Are these methods really distinct?

- How do people compile bibliographies in the first place?

Collection Mapping

- Idea: Build the collection in parts
  - Prioritize and budget specific subjects
    - Shrink, grow, keep constant
  - Evaluate subjects according to specific use
    - Which courses it serves, what are each courses' needs

To think about:

- Which of these approaches are **micro** and which are **macro**?

Use Factors

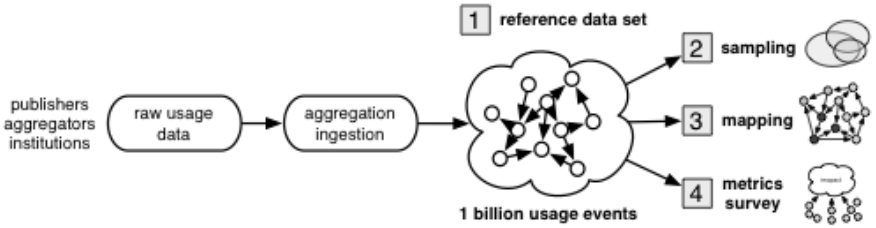
- Age
- Language
- Subject
- Shelf Arrangement
- Quality
- Expected Use
  - Popularity
  - **Information Chain** placement

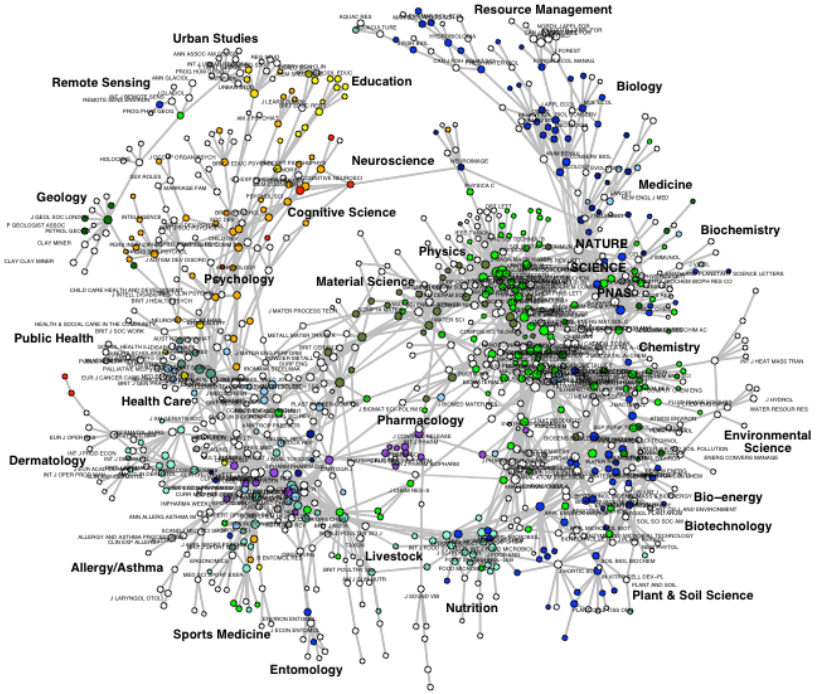
Use-based evaluation

- Physical Library
  - Slips
  - Circulation records
  - Table Counting
- Digital Library
  - Download counts
  - Citation counts (in scholarly works)

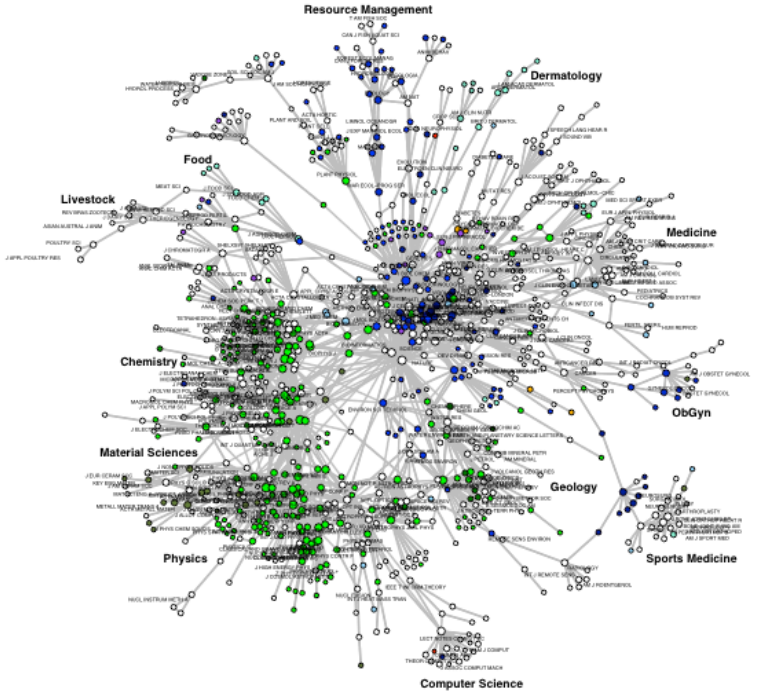
How do these two relate to each other?

MESUR project





Usage



Citation

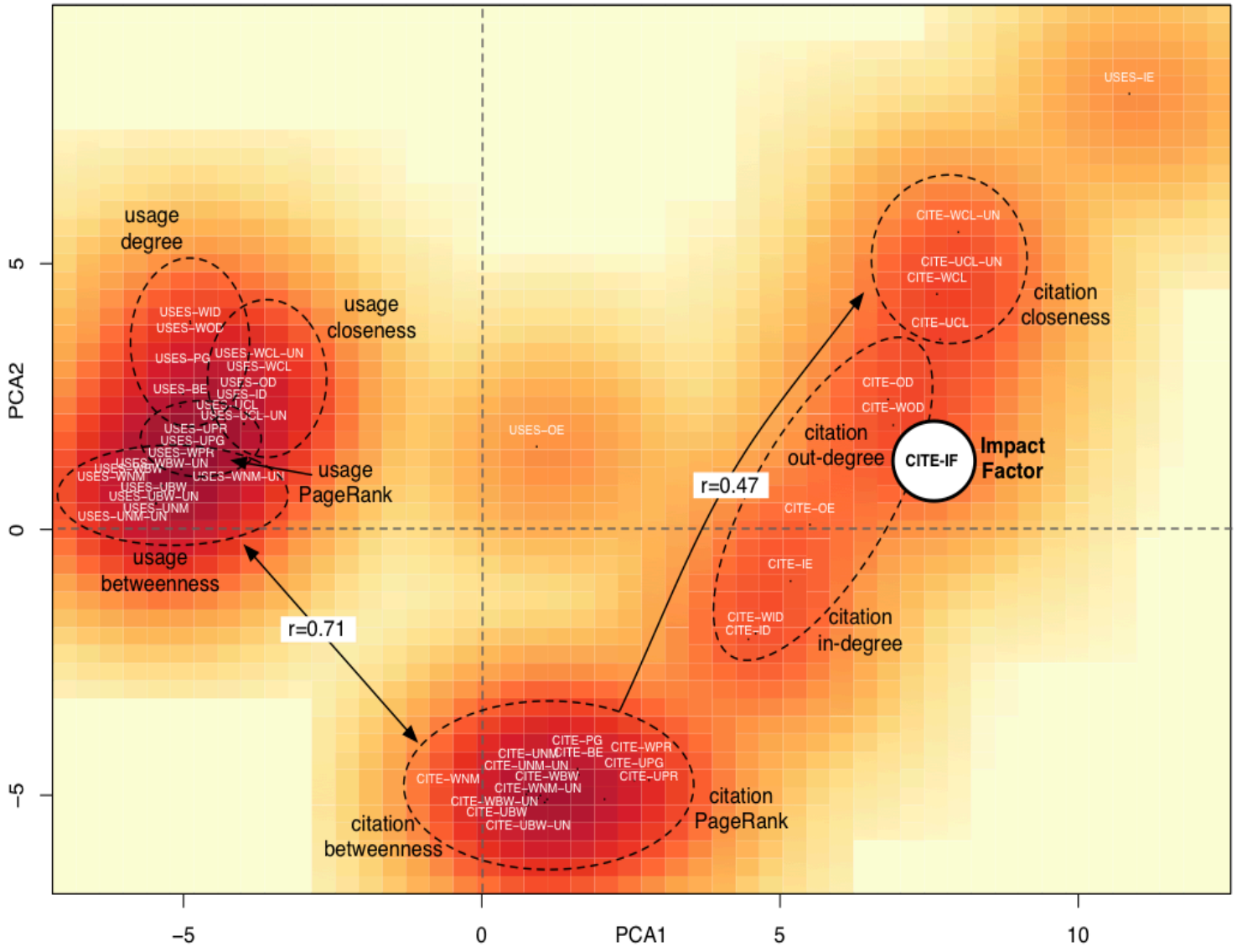
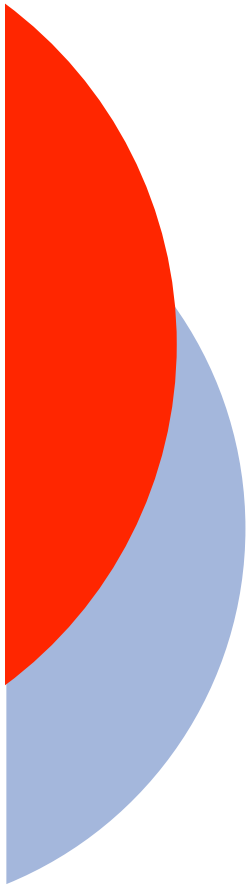


Figure 5: Principal component analysis of Spearman rank-order correlations between 47 preliminary MESUR metrics.



Digital Libraries

IR Evaluation Metrics

Min-Yen KAN

* - Parts of this lecture come from Lilian Tang's lecture material at the Univ. of Surrey

Evaluation Contingency Table

	System says is **relevant**	System says is **irrelevant**
Document is actually **relevant**	TP (True Positive)	FN (False Negative)
Document is actually **irrelevant**	FP (False Positive)	TN (True Negative)

Sensitivity, specificity,
positive and negative predictive value

		Relevant			
		+	-		
Test (System)	+	True Positive (TP)	False Positive (FP)	All with Positive Test TP+FP	*Positive Predictive Value* = TP / (TP+FP)
	-	False Negative (FN)	True Negative (TN)	All with Negative Test FN+TN	*Negative Predictive Value* = TN / (FN+TN)
		All Relevant	All non-relevant	All documents = TP+FP+FN+TN	
		Sensitivity = TP / (TP +FN)	*Specificity* = TN / (FP +TN)	Pre-Test Probability of Relevance = (TP+FN) / (TP+FP+FN+TN) (in this case = *prevalence*)	

Evaluation Metrics

- Precision = Positive Predictive Value

$$\frac{TP}{TP+FP}$$

  - "ratio of the number of relevant documents retrieved over the total number of documents retrieved"
  - how much extra stuff did you get?
- Recall = Sensitivity

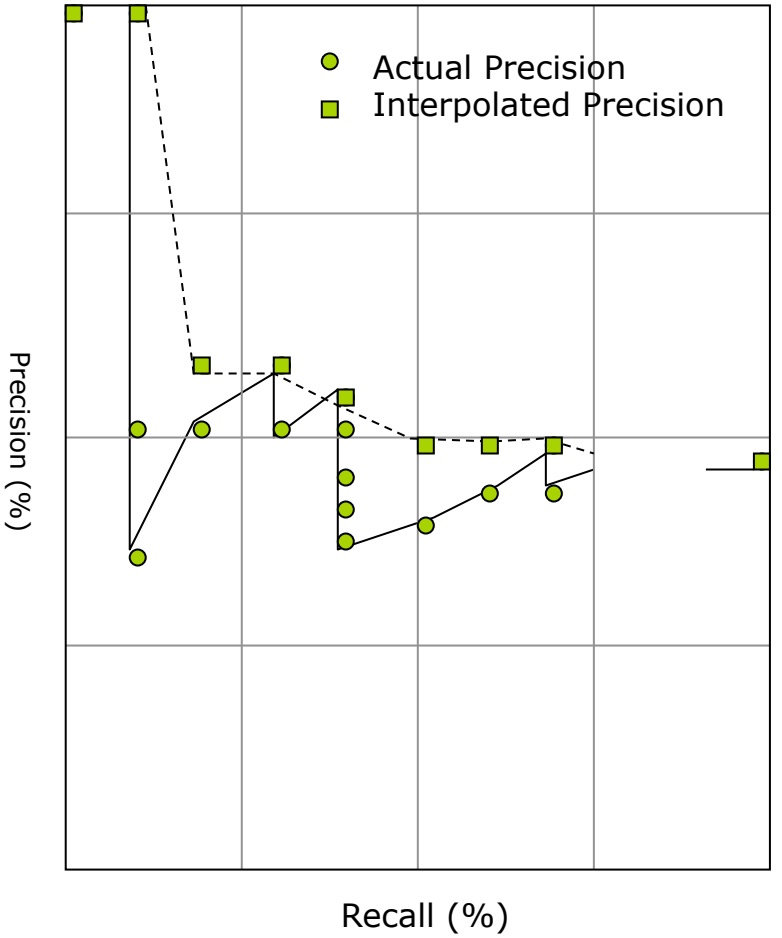
$$\frac{TP}{TP+FN}$$

  - "ratio of relevant documents retrieved for a given query over the number of relevant documents for that query in the database"
  - how much did you miss?

P/R: an example

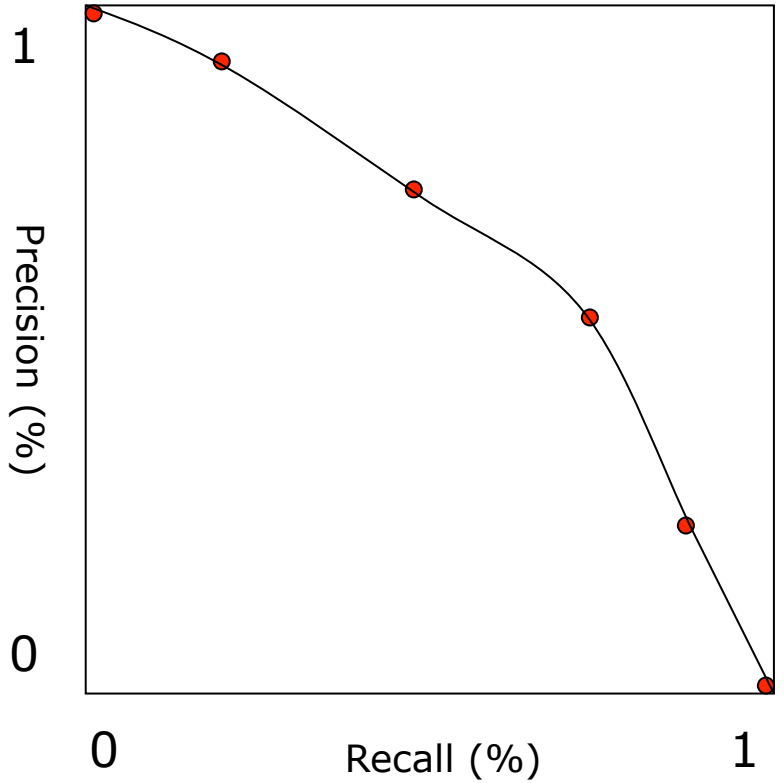
Rank	Decision	$R_{@r}$	$P_{@r}$
1	R	10%	100%
2		10%	50%
3		10%	33%
4	R	20%	50%
5	R	30%	60%
6		30%	50%
7	R	40%	57%
8		40%	50%
9		40%	44%
10		40%	40%
11		40%	36%
12	R	50%	42%
13	R	60%	46%
14	R	70%	50%
…			
22	R	100%	45%

From: Managing Gigabytes



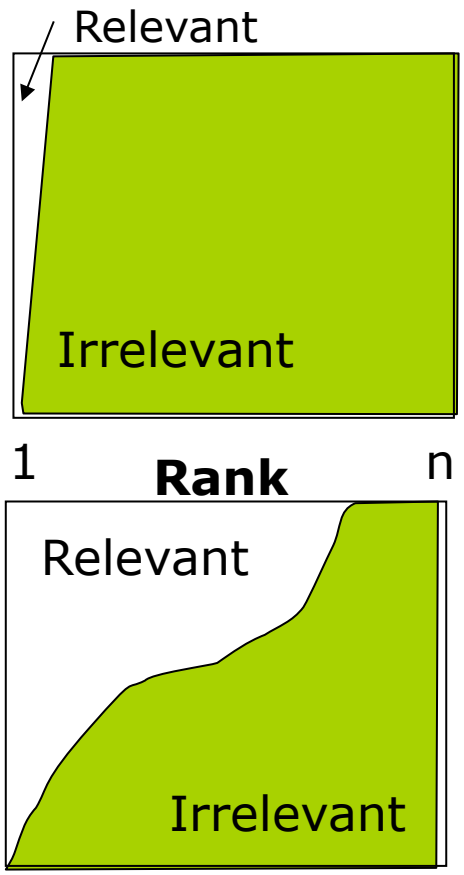
Precision / Recall

- Interpolated precision gives a non-increasing curve
- But it doesn't factor in the size of the corpus
  - Previous example on a corpus of 25 docs = 40% precision
  - On a corpus of 2.5 M docs = also 40%



Factoring in size of a corpus

- Look at how P/R or Sn/Sp varies as a function of rank:
- Choose a number of different ranks and calculate P/R or Sn/Sp
  - Correspond to vertical lines on graphs at right
  - Plot Sn vs. 1-Sp to get points for ROC curve. Interpolate curve.

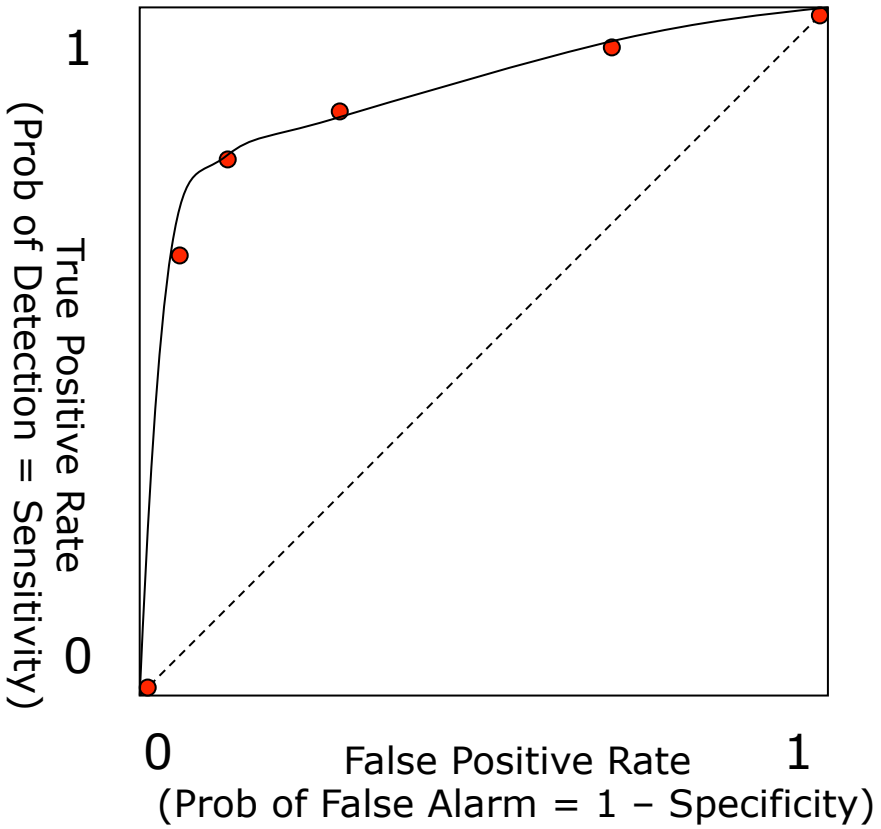


Which of these examples is which from the previous slide?

ROC Curve

Look at the **probability** or **rate** of detection

- What does the diagonal represent?
- How do we compare ROC curves versus each other?



Getting a single number

- 11 pt average
  - Average precision at each .1 interval in recall
- Precision at recall point (% or absolute)
- F Measure
  - Ratio of precision to recall: (e.g., F_3 = weight precision heavier)

$$F_b = \frac{(b^2+1)\,PR}{b^2P + R}$$

- Area under ROC curve (Accuracy)
  - 1 = perfect, .9 excellent, .5 worthless

- What's the difference between these measures?
- Which measures are best suited to which scenarios?