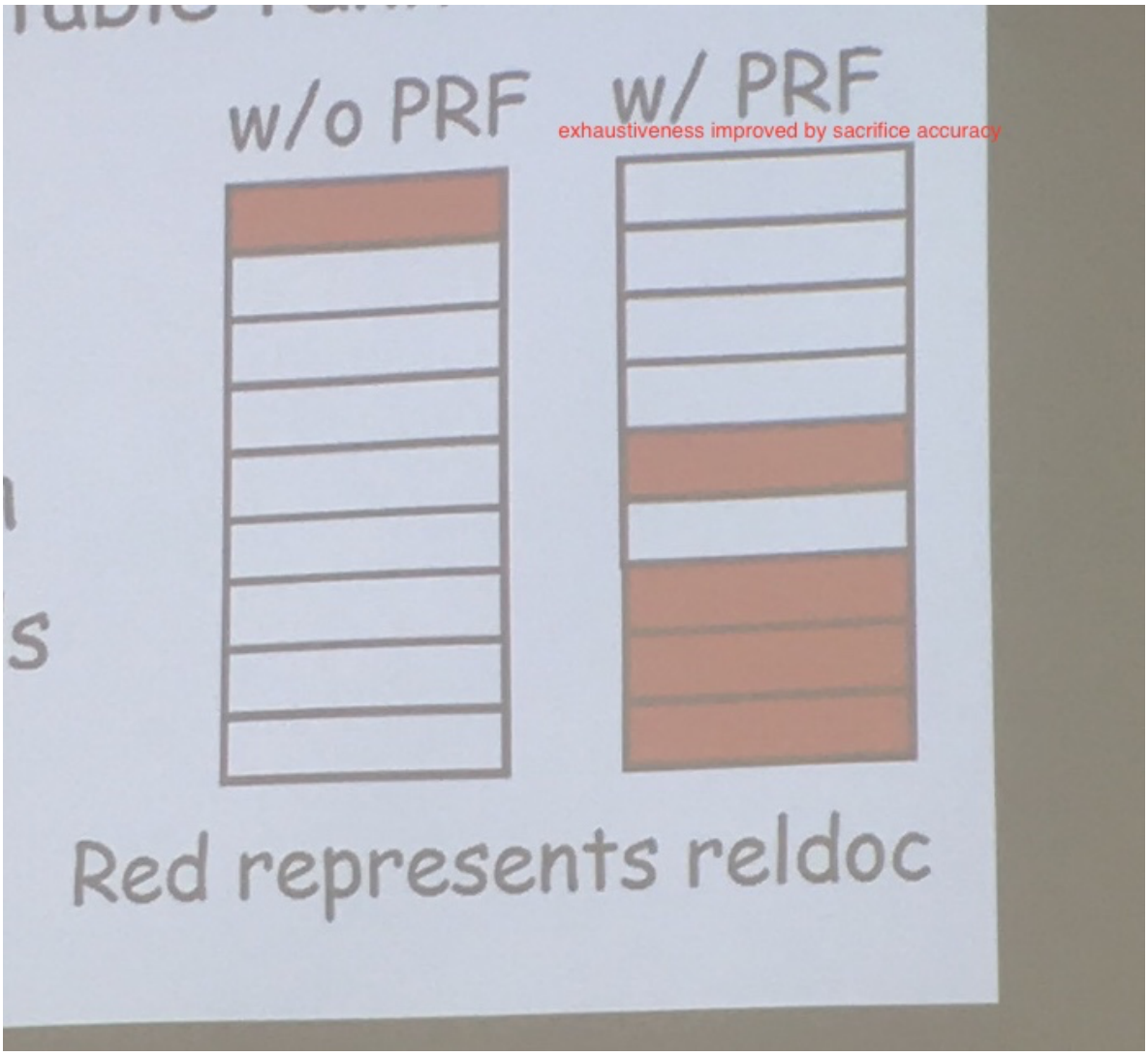
Day 9&10

Properties of PRF (pseudo relevance feedback/ blind feedback/ local feedback)

- To avoid human efforts, the top-D documents are unconditionally regarded as “relevant”
  - Rocchio’s formula works with at least one R, and so we seldom use N in PRF
  - In practice, top-K high-weight terms in top-D docs are added to query vector (K, D range 10-20 by several reports)
- Fully automated: no one can tell if PRF is done
- The rationale of why and in which situation PRF works well has not been clear

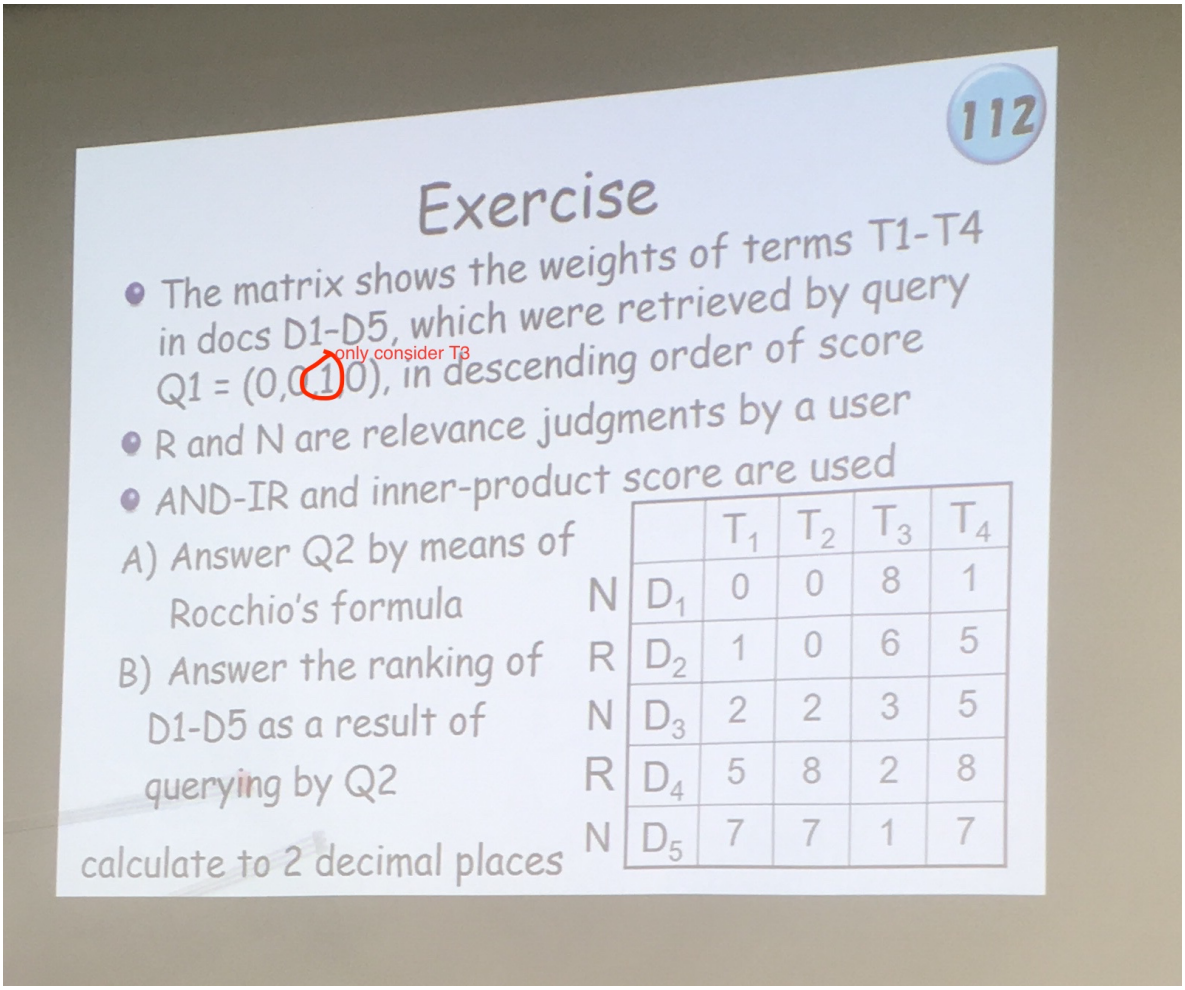
Properties of PRF (con’t)

- PRF often changes the distribution of reldocs in a ranked document list
  - W/o PRF: a few high-rank reldocs
  - W/PRF: quite a few acceptable-rank reldocs (in the middle of the list)
- A possible explanation is:
  - In the 1st stage a short query by a user leads to precision-oriented search
  - In the 2nd stage (relevance feedback as) query expansion leads to recall-oriented search (eg. Type query in google it suggests key words -> query expansion)



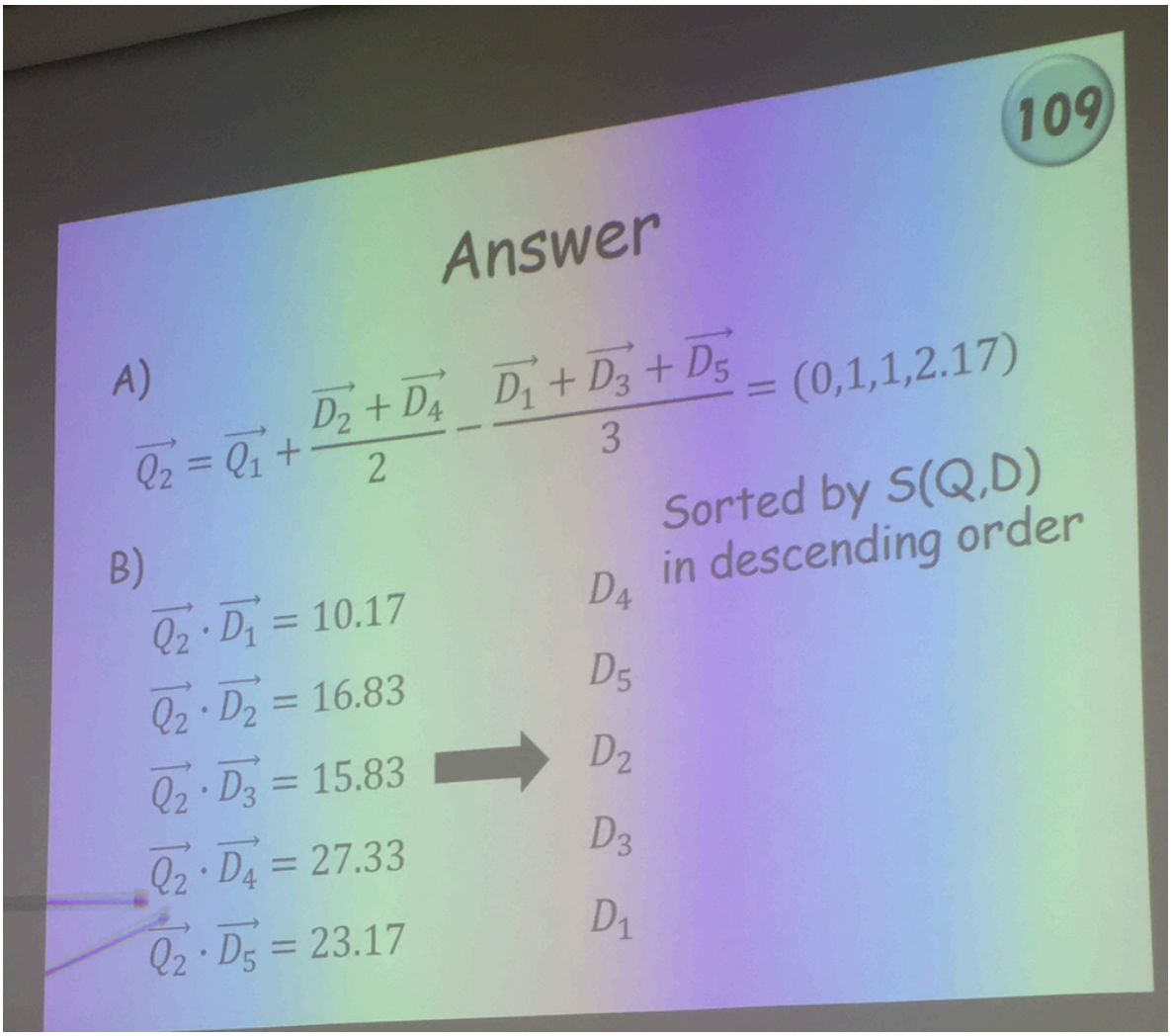
Pros and Cons for RF

- Pros
  - Query expansion (QE) adds elements for synonyms to (and removes irrelevant words from) a query vector, to increase converge
  - That's why RF is popular in tech domain such as patent and research databank
- Cons
  - If user is content with one relic, there is no reason to perform additional search
  - QE can add extraneous terms that would shift the topic of a document



A) avg(R_d2,R_d4) and avg(N_d1,N_d3,N_d5)
R = (3,4,4,6.5)
N = (3,3,4,4.33)
Q2 = Q1+R1-N1 = (0,0,1,0) + (3,4,4,6.5) - (3,3,4,4.33) = (0,1,1,2.17)

B) calculate inner product by using Q2 and each doc
D1 = 1
D2 =



Question:
What these acronyms stand for?

- RCR: responsible conduct of research
- QRP: questionable research practice
- FFP: three major scientific misconducts
  - Fabrication: to make up for deception
  - Falsification: to make false by multilation and addition
  - Plagiarism: to steal idea and pass it off as your own

To be a scientific research

- Valid: well grounded explanation
- Objective: observable evidence
- Reproducible: reach at the same conclusion irrespective of the person and situation

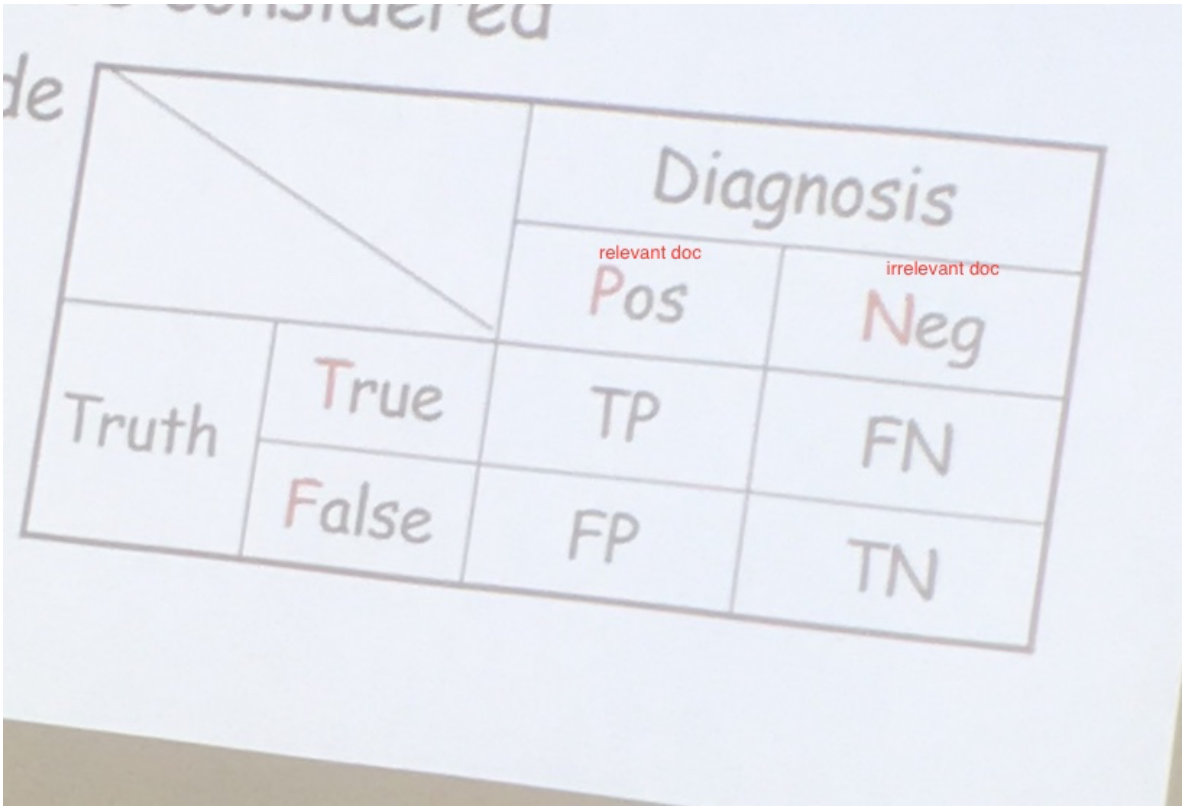
What and how to evaluate achievement for IR research

- In medical science
  - In vitro: inside test tube
  - In vivo: inside experimental animal

    - Clinical test: targeting “human subject” (recently “participant” is politically correct)
- In IR
    - Module-by-module basis: indexing, stemming, searching and sorting algorithms
    - System as a black-box
    - “Participants” as hypothetical users

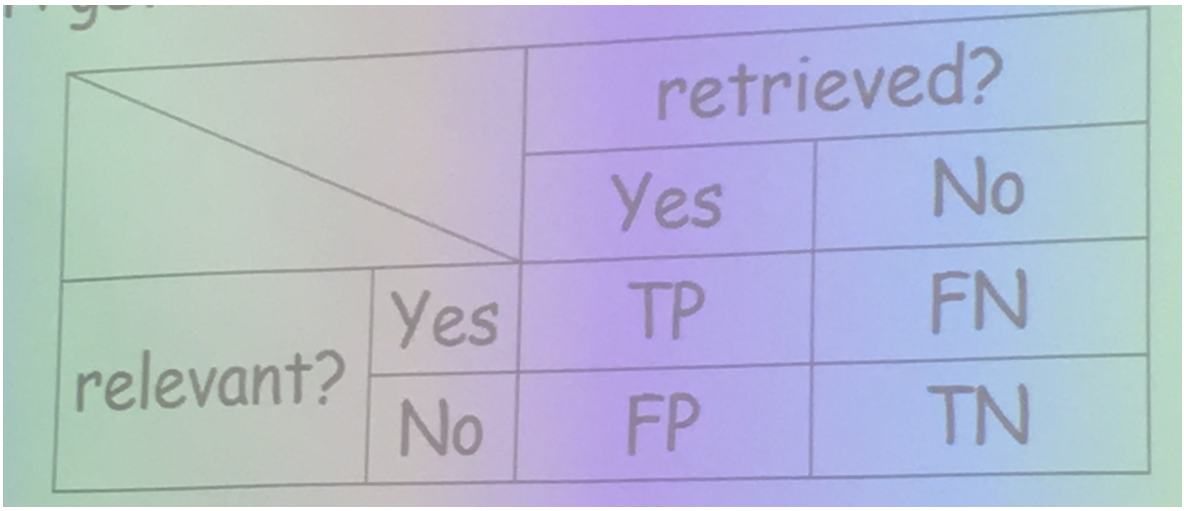
Successes and failures

- The result of a medical test is positive or negative, but the diagnosis can be incorrect
- Diagnosis is pos or neg for each of which is either true or false, resulting in 2 success and failure types can be considered
- Table in right side is a contingency table, in which each cell denotes # of trials in each category



Contingency table

- As with medcial diagnosis, positive(P) means something you cannot miss it, whereas negative(N) must be ignored (let them leave)
- F denotes cases where P are N were errors
- FP: get irrelevant one, FN: miss relevant one



Test collection

- **Search topic:** description for information need (often called just “topic”)
  - Query must be made manually or automatically depending on the purpose (topic is not Q)
- **Document collection:** contain target reldocs
- **Relevance judgment:** rel-docs for each topic
  - Exhaustive judgement is impractical for TB order data, and so “pooling” is often used
- Others: utility programs (e.g. scoring)

Pooling in relevance judgement

- **Hypothesis:**
  - Many effective and heterogeneous IR system are available
  - Union of documents (“pool of documents”) retrieved by one of the systems should cover relevance documents
- human assessors judge only the doc pool to reduce workload
- But, where are such many IR systems?

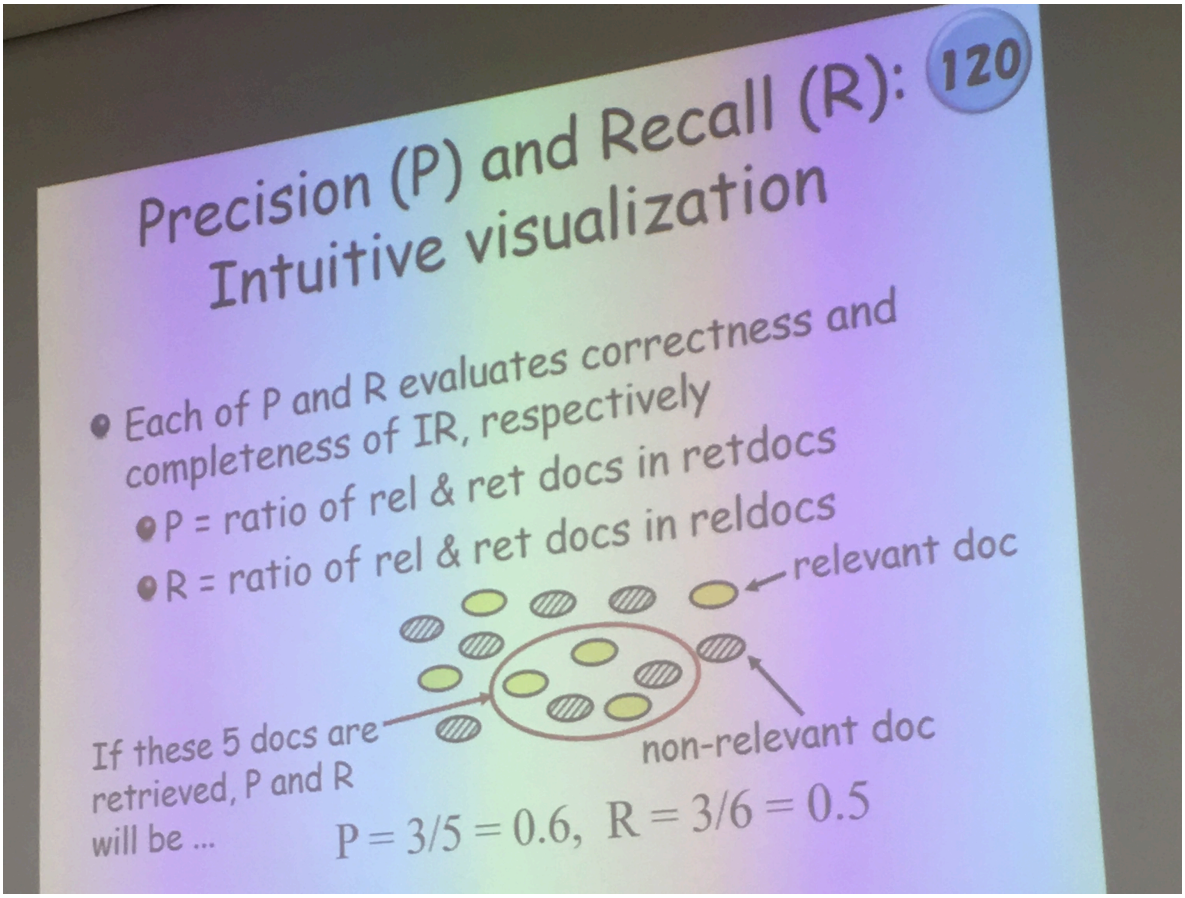
Business model for pooling

- An organizing team that wants to produce test collection arranges a competition-style workshop
- Win-win b/w organizing team and participants
  - Organizer needs as many ranked documents as possible, for efficient pooling
  - Researchers who want a huge dataset for their research would participate in the workshop
- Famous international evaluation workshop
  - Us: TREC (NIST), JP: NTCIR (NII), AND EU: CLEF

Important properties for test collection

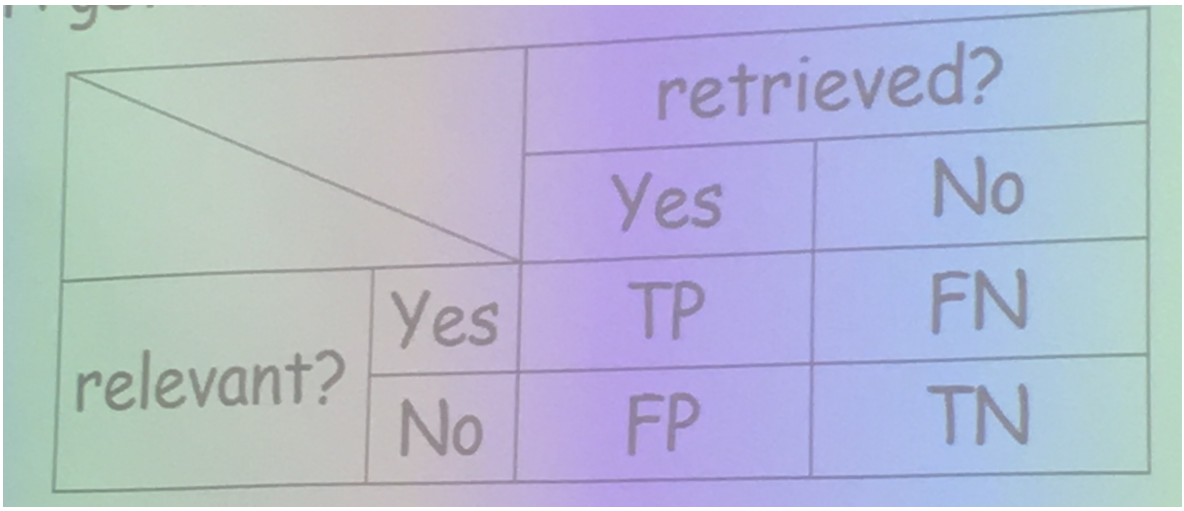
- Stability for the test
  - Subtle changes of participating systems should result in negligible

changes
    - Stability usually increases by more topics
- Reusability of the collection
    - Make it sure that your test collection is not in favour of participating systems due to the pooling
- Reasonable amount of data
    - Only “big” or “huge” is not always virtue
    - Scientific research has been done with an appropriate size of controlled sample

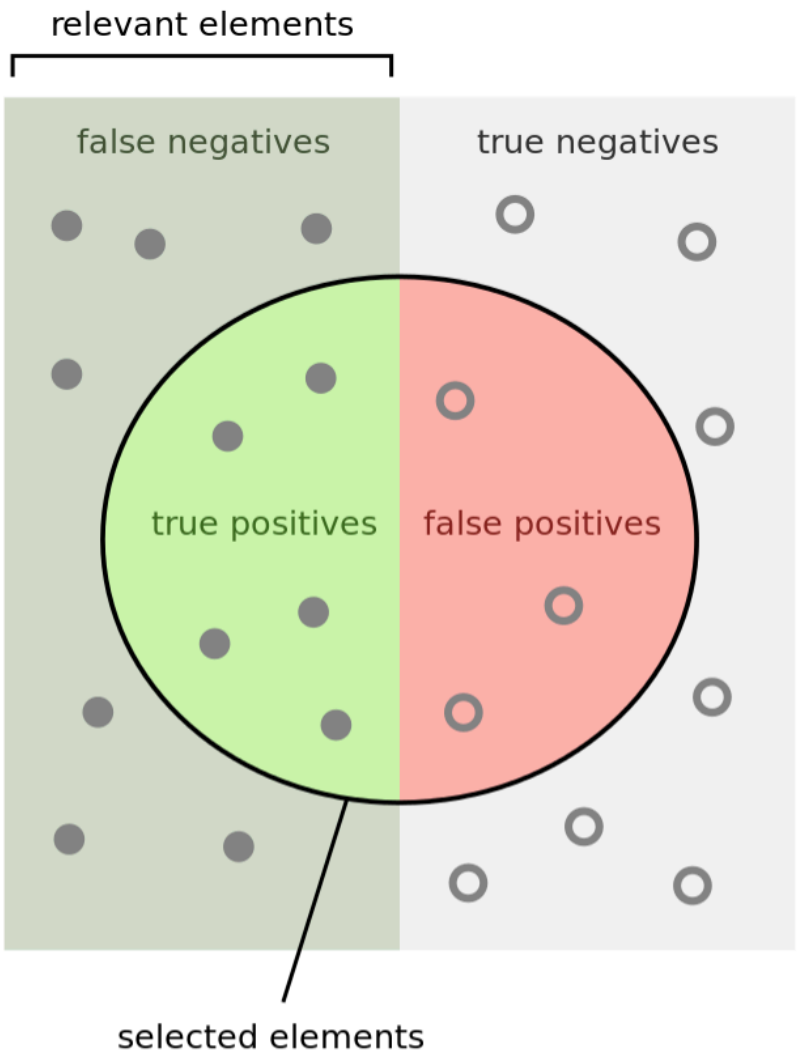


P and R: Formal Explanation

- Contingency table below was obtained for a single query (I.e. 1000 topics result in 1000 contingency tables)
- Answer how to calculate P and R w/ this table



P = TP / (TP+FP)
R = TP / (TP+FN)

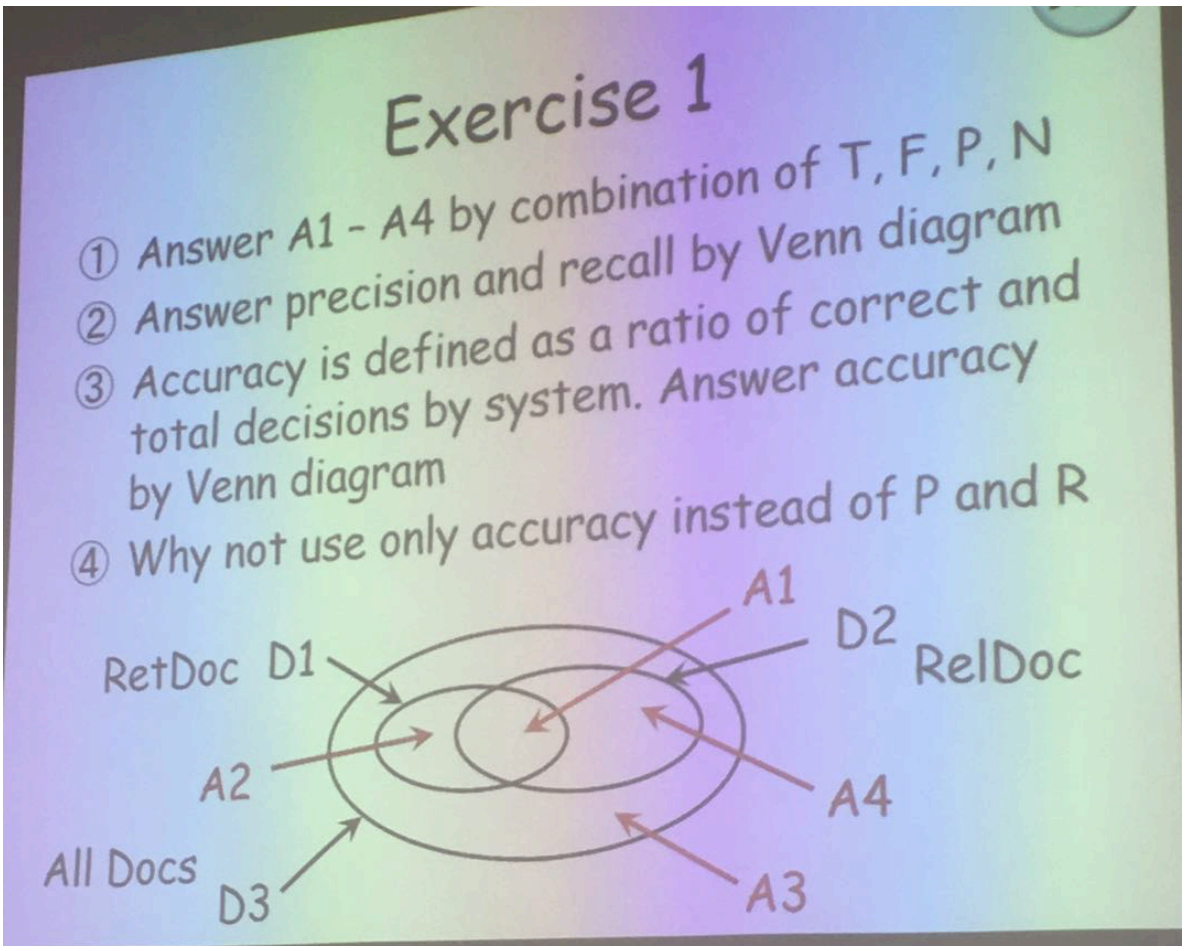


How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{selected elements}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$



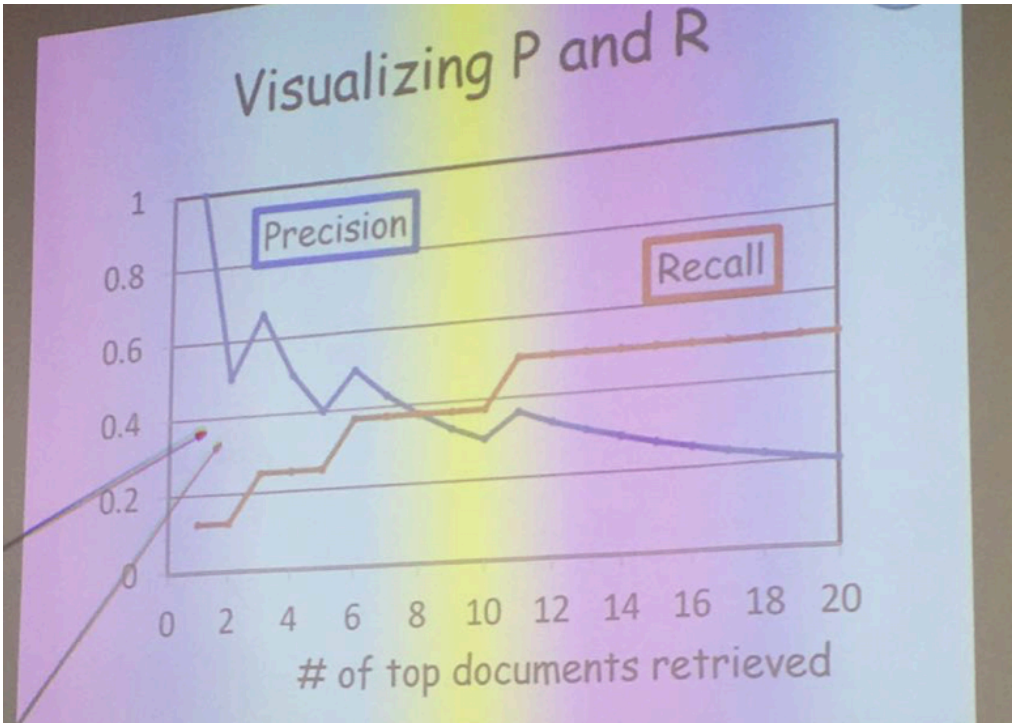
1. A1: TP; A2: FP; A3: TN; A4: FN
2. P = A1/(A1+A2); R = A1/(A1+A4)
3. (A1+A3)/D3
4.

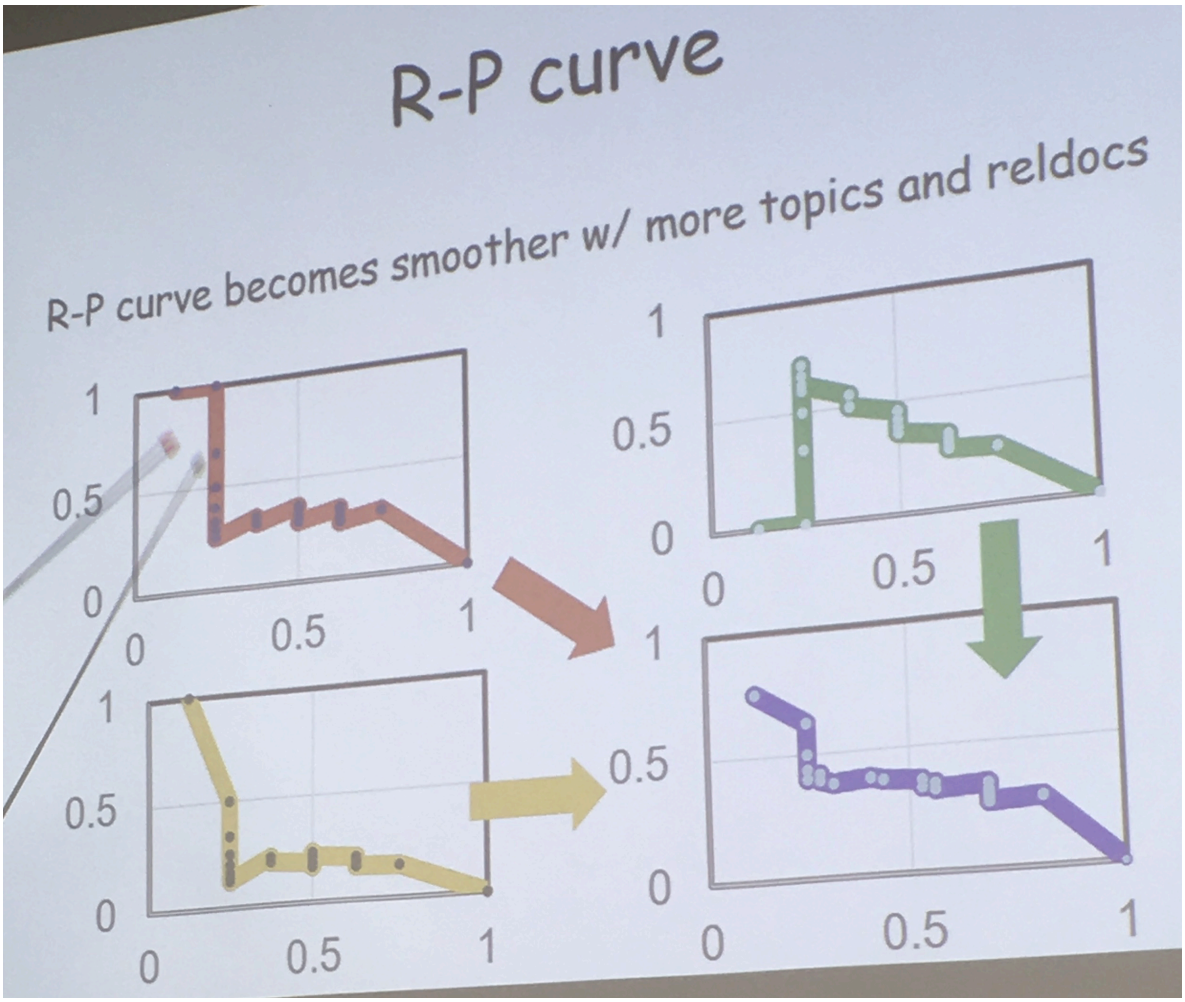
P and R for ranking IR

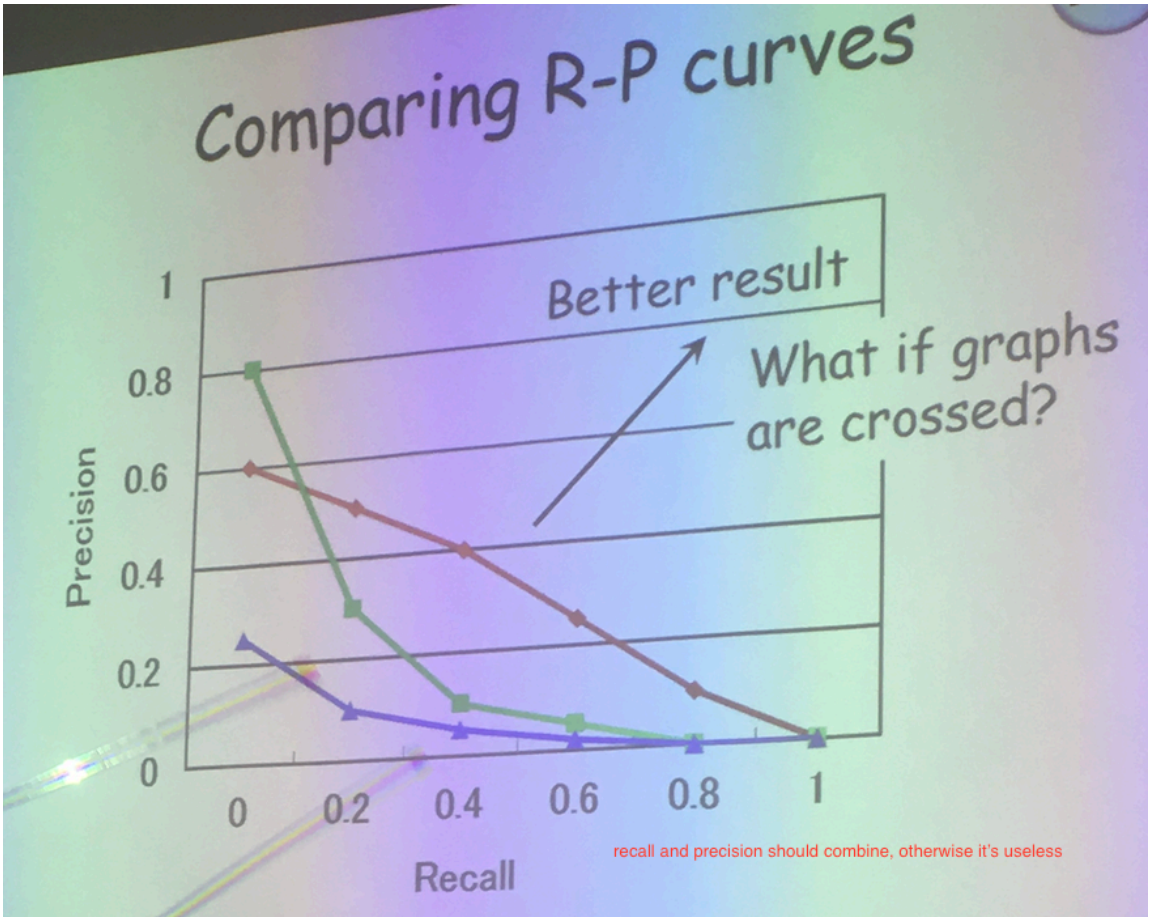
- Difficult to define “retrieved documents”
- P and R are calculated for each fo the top-D documents in the search result

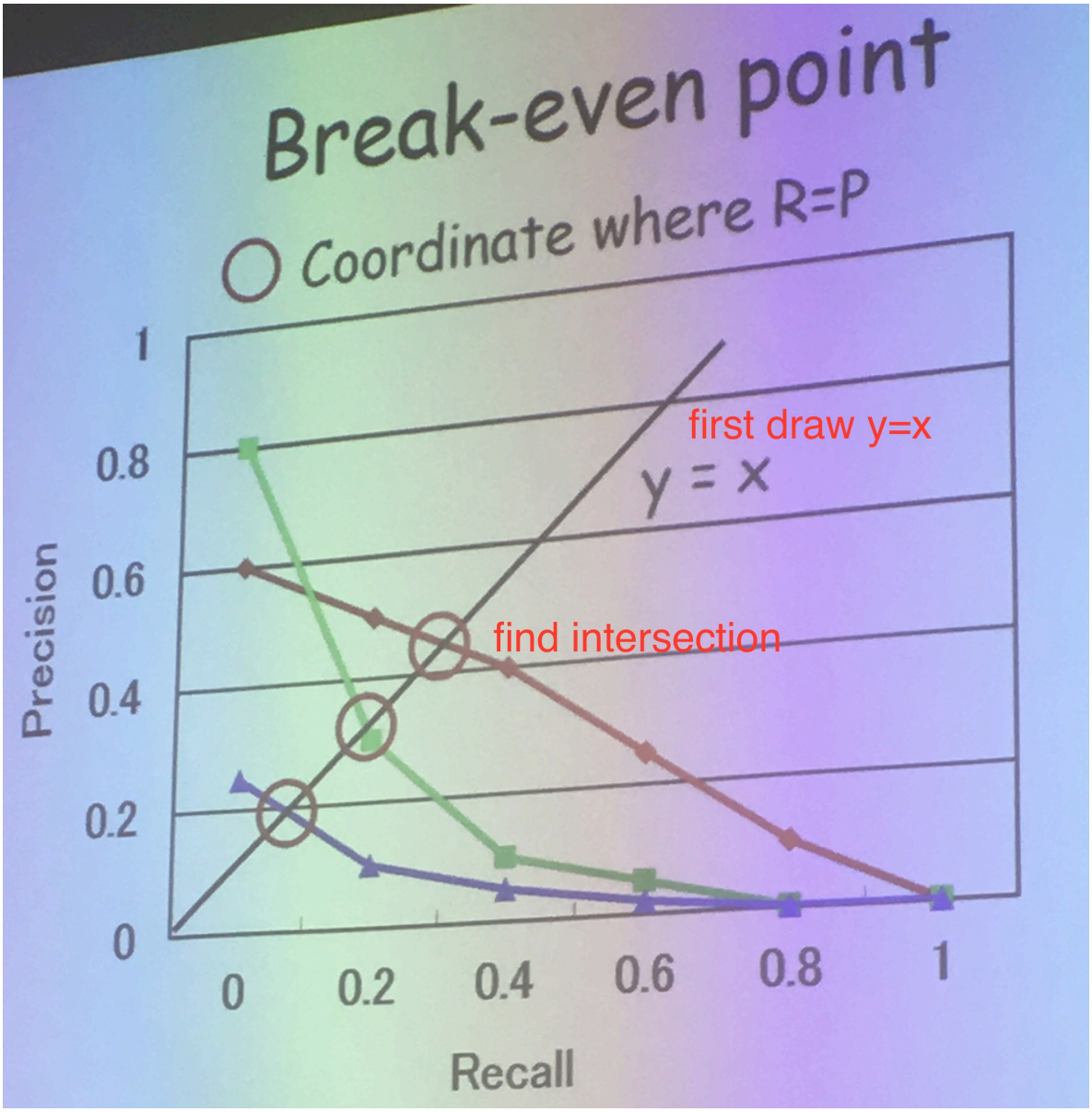
Rank	R/N	Precision	Recall	
1	R	1	0.125	
2	N	0.5	0.125	
3	R	0.667	0.25	
4	N	0.5	0.25	
5	N	0.4	0.25	
6	R	0.5	0.375	
# of relevant docs = 8				

Example for top-3: P = 2/3, R= 2/8


Visualizing P and R
Precision
Recall
1
0.8
0.6
0.4
0.2
0
0 2 4 6 8 10 12 14 16 18 20
# of top documents retrieved

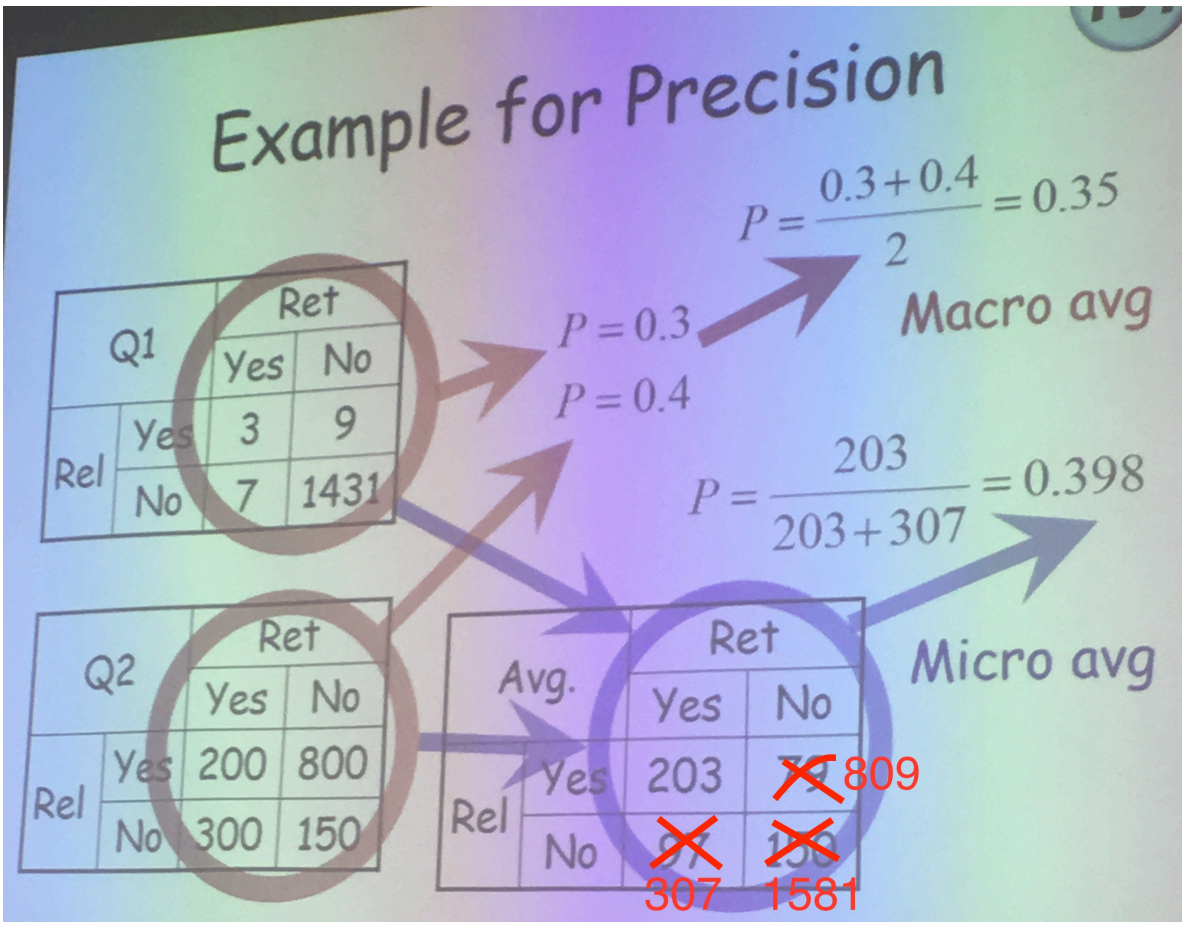

R-P curve
R-P curve becomes smoother w/ more topics and reldocs
1
0.5
0
0 0.5 1


Comparing R-P curves
Better result
What if graphs are crossed?
Precision
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
Recall
recall and precision should combine, otherwise it's useless



Averaging P (or R) over topics

- R-P curve is visible but indecisive as a test
- IR result for topics Q1 & Q2 (1450 docs)
  - Q1: TP = 3, FP = 7, FN = 9, TN = 1431
  - Q2: TP = 200, FP = 300, FN = 800, TN = 150
- When we "average" P or R over the 2 topics, which average is more fair?
  - A) calculate avg. of P (or R) independently then average them over the topics
    - Q1: P = 3 / (3+7) = 0.3 R = 3 / (3+9) = 0.25
    - Q2: P = 200 / (200+300) =0.4 R = 200 / (200+800) = 0.2
  - B) summing up contingency tables of 2 topics then calculate P (or R) by that new table



Exercise

A) complete the table below using our example

	Q1	Q2	Macro	Micro
P	0.3	0.4	0.35	0.398
R	0.25	0.2	0.225	0.201

B) answer the reason for the difference b/w macro and micro averages

C) answer the implication of observation in B) from an IR evaluation point of view

- A) See below
- B) Whereas macro average equally treats each topic wrt the ratio of TP and TP+FP (or TP+FN), micro avg is influenced by the absolute numbers of TP, FP, and FN
- C) Assuming each topic corresponds to a hypothetical user, macro avg becomes large for a system that saves as many users as possible

	Q1	Q2	Macro	Micro
P	0.3	0.4	0.35	0.398
R	0.25	0.2	0.275	0.201

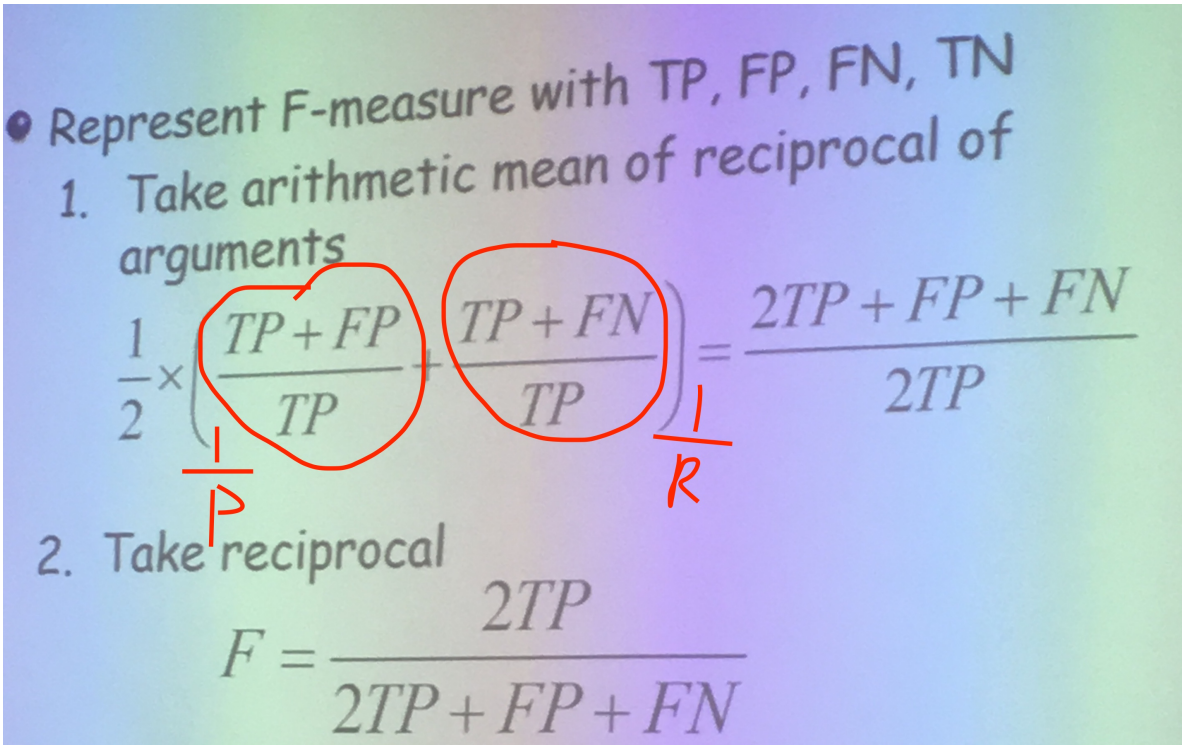
Quiz

- When you walked from A to B (one way: 100km) at an average of 10kmh, and walked on the same way backward at an average of 5kmh, what is the average speed in one-round trip?
  - 200 / (100/10 + 100/5) = 6.67kmh

F-measure (F1)

- A harmonic mean of P and R
  - Reciprocal of arithmetic mean between reciprocals of P and R: ranges within [0,1]
  - The larger, the better
- Optionally, parametric constant to change the weight of either value
  - F1 is a case where P/R have equal weight

$$F = \frac{1}{\left(\frac{1}{R} + \frac{1}{P}\right) \times \frac{1}{2}} = \frac{2 \times R \times P}{R + P}$$



Average Precision (AP)

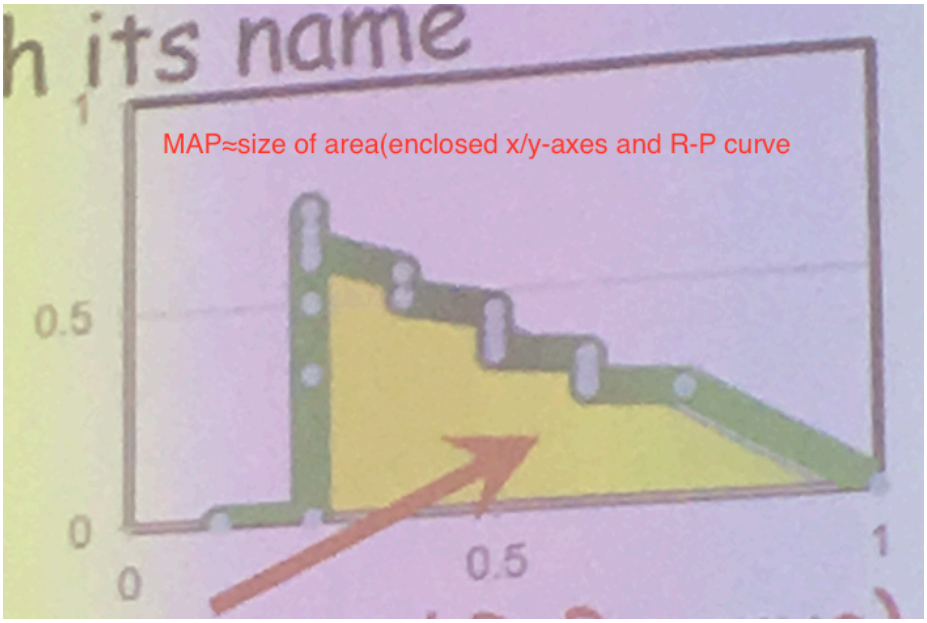
- An arithmetic mean of precision values at which a relevant document was found
- No score for relevant docs missed, only consider found relevant docs

Rank	R/N	Precision	Recall
1	**R**	**1**	0.125
2	N	0.5	0.125
3	**R**	**0.667**	0.25
4	N	0.5	0.25
5	N	0.4	0.25
6	**R**	**0.5**	0.375
# of relevant docs = 8			

AP = (1+2/3+3/6)/8=0.27
Divided by 8 not 3

Mean average precision (MAP)

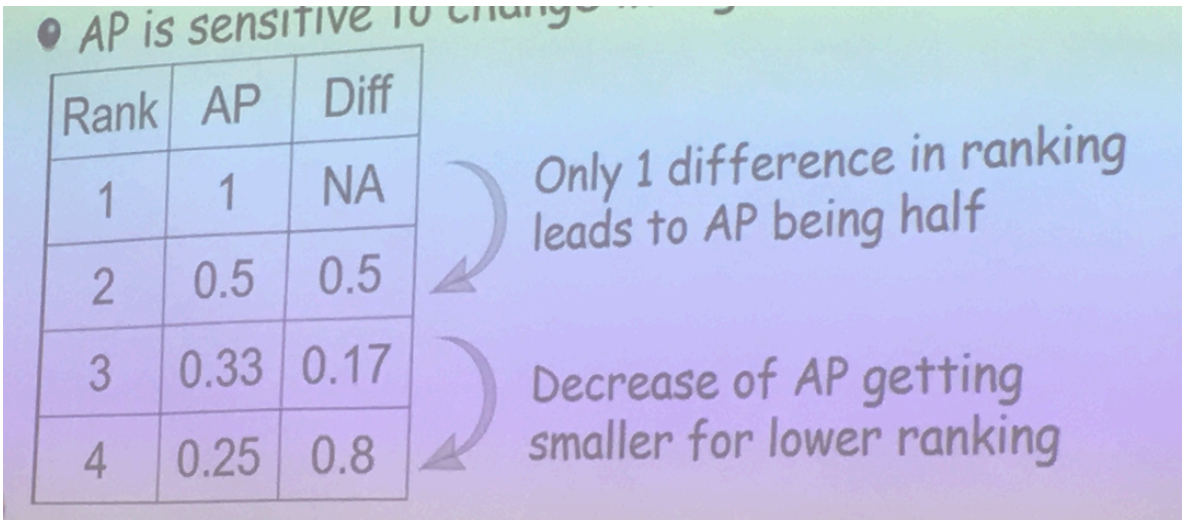
- Macro avg of APs over all the topics
  - Can be used to evaluate IR system per se
  - (P and R should be used together)
- Should not be confused with its name
  - MAP represents P and R
  - Why? **No score for a missed reldoc**



- MAP value itself is not important
  - Comparison is always important

A property of AP

- When #reldocs is inherently small, #data points in R-P curve also small
- Suppose a topic has only a single reldoc, what is AP for each document ranking
- AP is sensitive to change in higher ranking



How to conform achievement

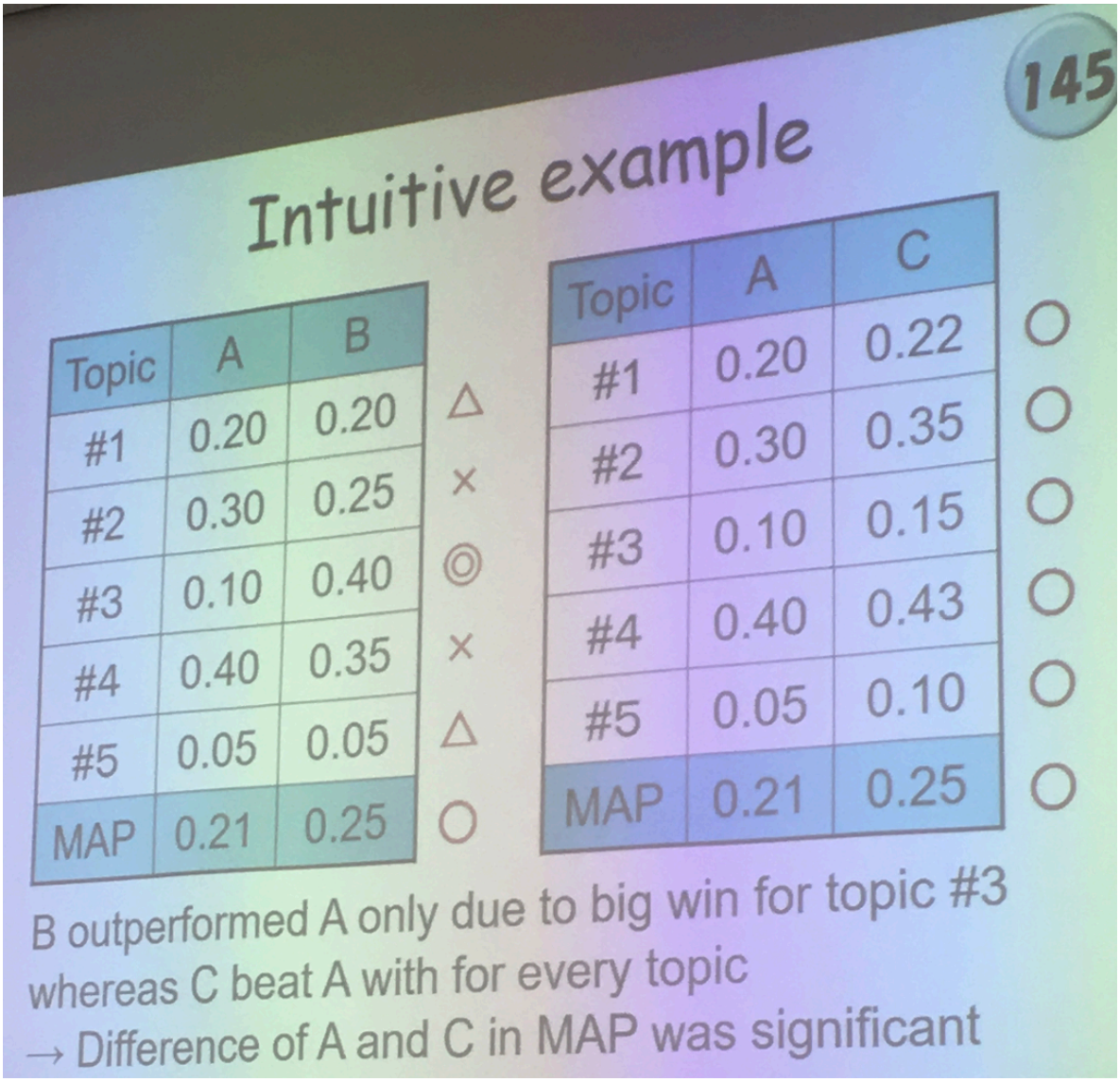
- map(S): MAP for system S
  - MAP(S) usually ranges in 0.2-0.4, partially due to selection of discriminative topics
- Improvement of MAP for systems A and B
  - map(A) = 0.24, map(B) = 0.35 -> 0.11
- But difference does not tell us everything
  - 0.6 -> 0.65 (improved 0.125 errors)

0.4 errors remained -> 0.05 out of 0.4 was improved
  - 0.9 -> 0.95 (improved 0.5 errors)

0.1 errors remained -> 0.05 out of 0.1 was improved

Example of useful analysis

- Statistical testing: to make sure if difference of score b/w two systems is significant
  - In IR, t-test has often been used
  - Two-tailed paired t-test, where each sample (item) is AP for each of the search topics
- Ablation test: to evaluate the contribution of each component, by iteration of the IR trials
  - In each iteration, only a single component is disconnected, and so the reduction of MAP is the contribution of that component



Looking at what happened inside your system

- Automatic analysis tells you numerical result and probability of events, but not “so what?”

    - Produce “nice-looking” tables and charts is not the end but start the real analysis
    - Identify and modify problems is up to you
- Collect many logs on a module-by-module basis to trace reasoning of system
    - Ask to yourself “why that happened?” Until you cannot answer anymore
    - See also success cases (maybe by chance?)

Example scenario of manual investigation

- Why that happened?: query contained “salsa”
- So what?: salsa is polysemous and so does not match to terms in relevant documents
- Why not prepared for it?: yes, we used PRF
- Why it did not work?: insufficiently query expansion (QE) due to documents being short
- Now what?: need to change our strategy
- What’s next?: trying to use query logs for QE