Information Retrieval Evaluation

Nicola Ferro

@frrncl

Intelligent Interactive Information Access (IIIA) Hub
Department of Information Engineering
University of Padua

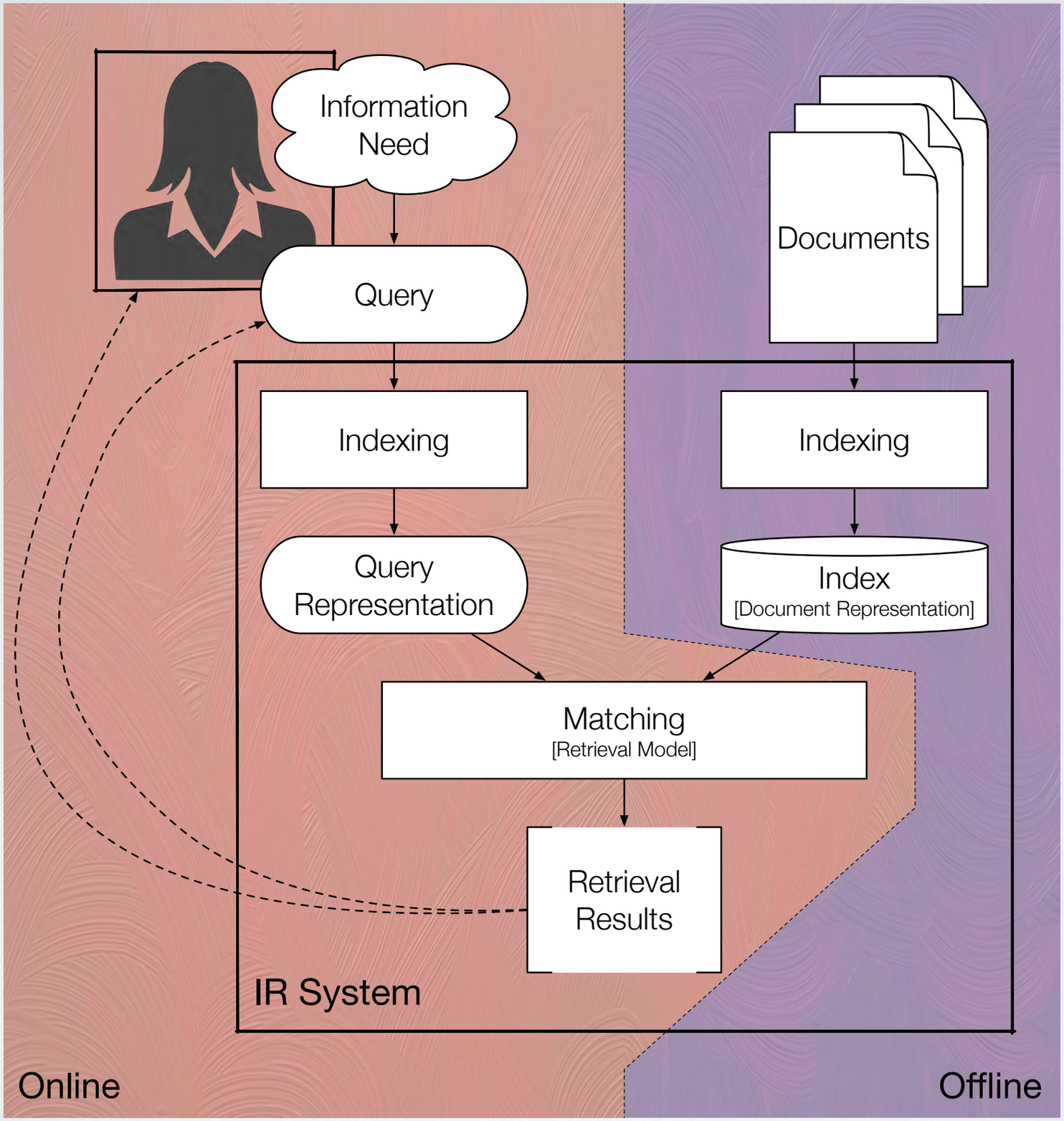
ESSIR 2022, 13th European Summer School in Information Retrieval
18-22 July 2022, Lisbon, Portugal

Outline

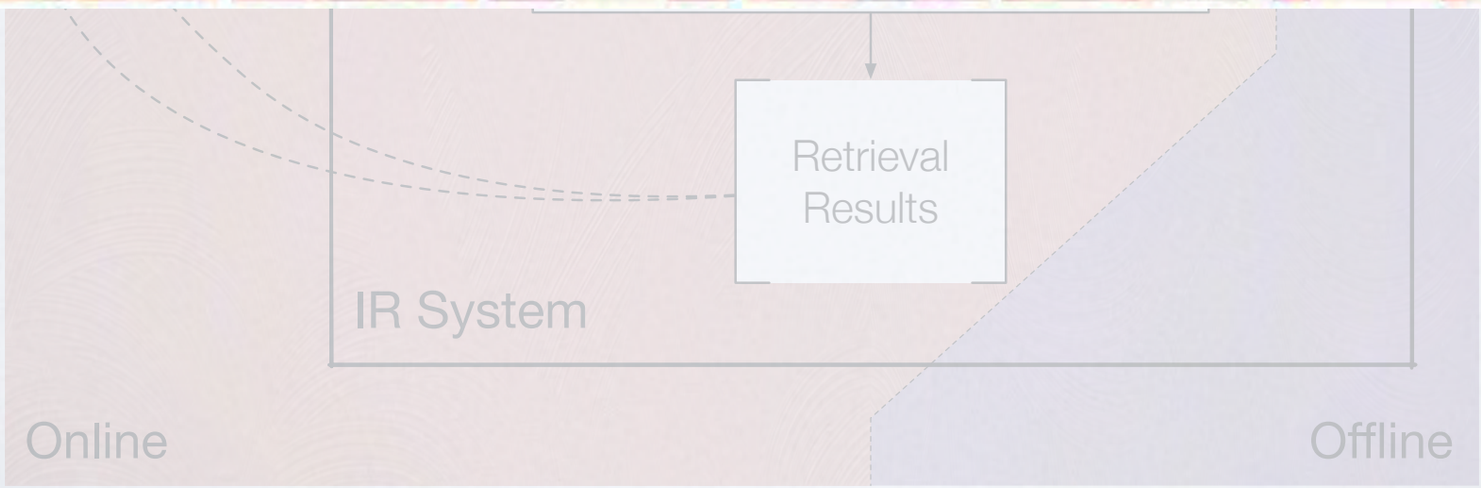
- Basics of Experimental Evaluation
- Relevance Assessment
- Evaluation Measures
- Statistical Hypothesis Testing
- Foundations of Measurement
- Reproducibility

Evaluation Basics

Our Goal


Information Need
Query
Documents
Indexing
Indexing
Query Representation
Index
[Document Representation]
Matching
[Retrieval Model]
Retrieval Results
IR System
Online
Offline

Our Goal


Information Need
Documents
Retrieval Results
IR System
Online
Offline

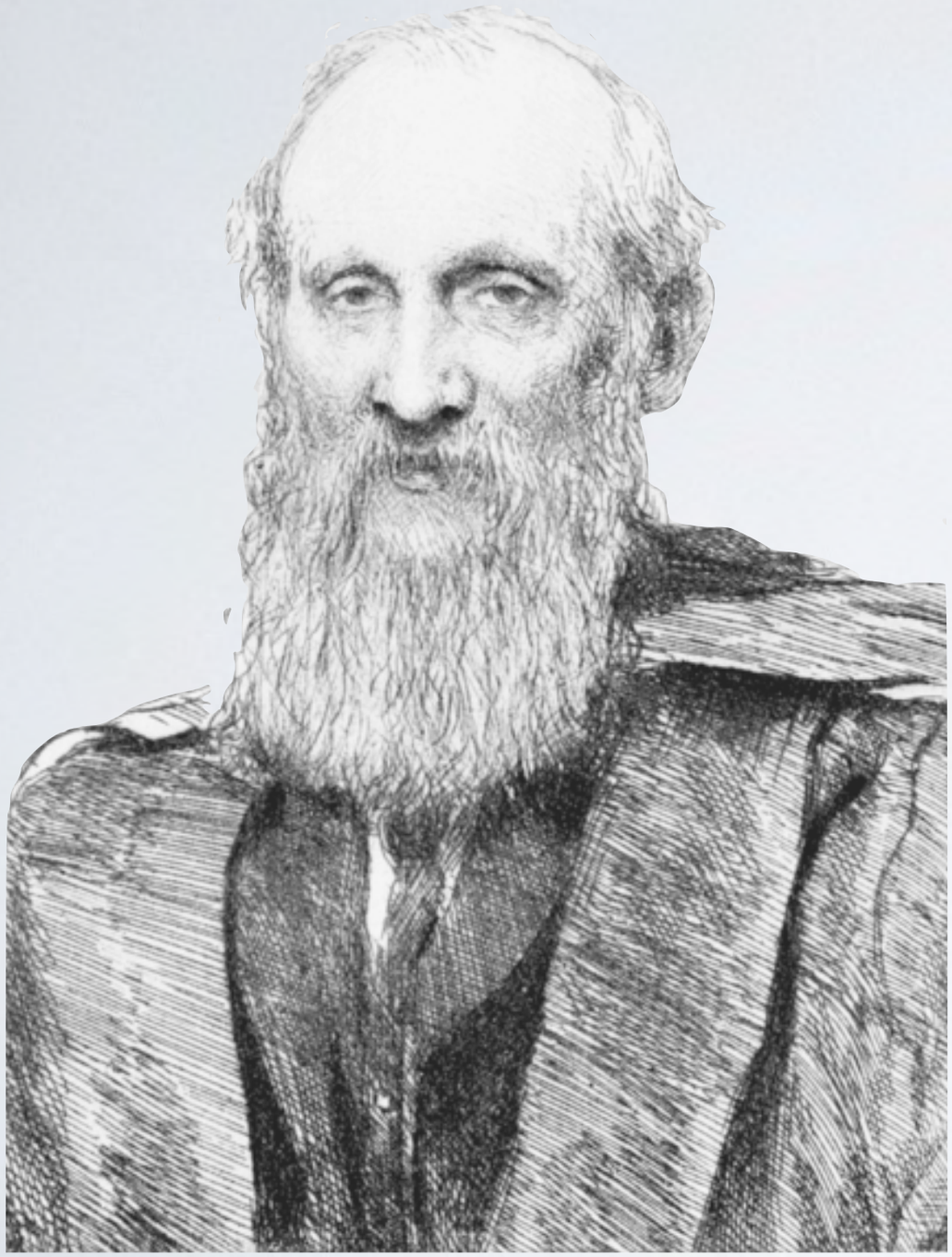
Deeply Rooted…

Experimentation

Deeply Rooted…

Experimentation

Why Evaluation?



"To measure is to know"

"If you cannot measure it, you cannot improve it"

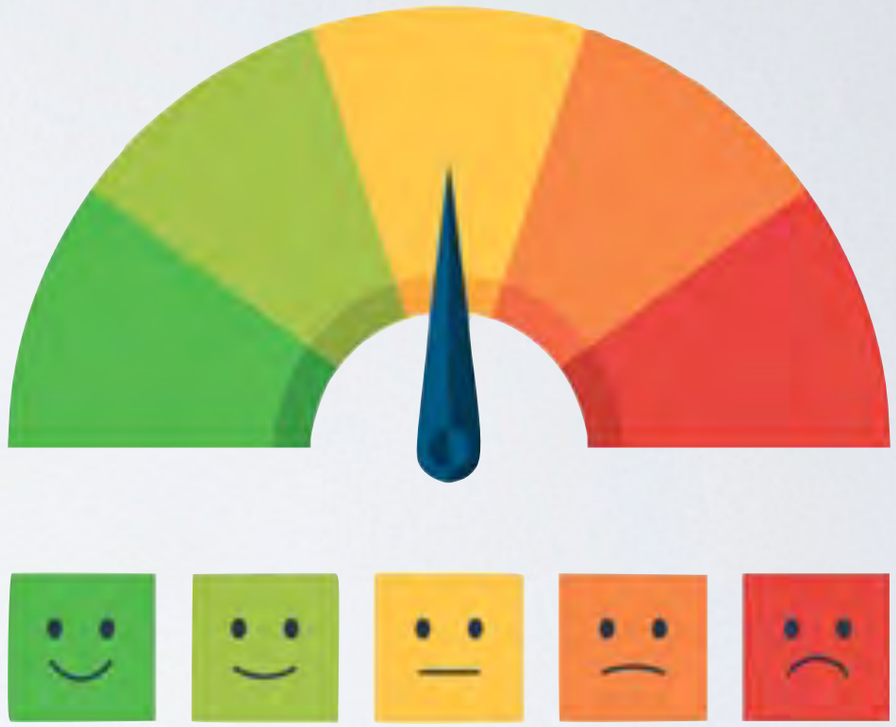
Lord William Thompson,
first Baron Kelvin (1824-1907)

What to Evaluate?

Efficiency

VS

Effectiveness



Evaluation in Action

(c) IBM Corporation. http://www.youtube.com/watch?v=3G2H3DZ8rNc

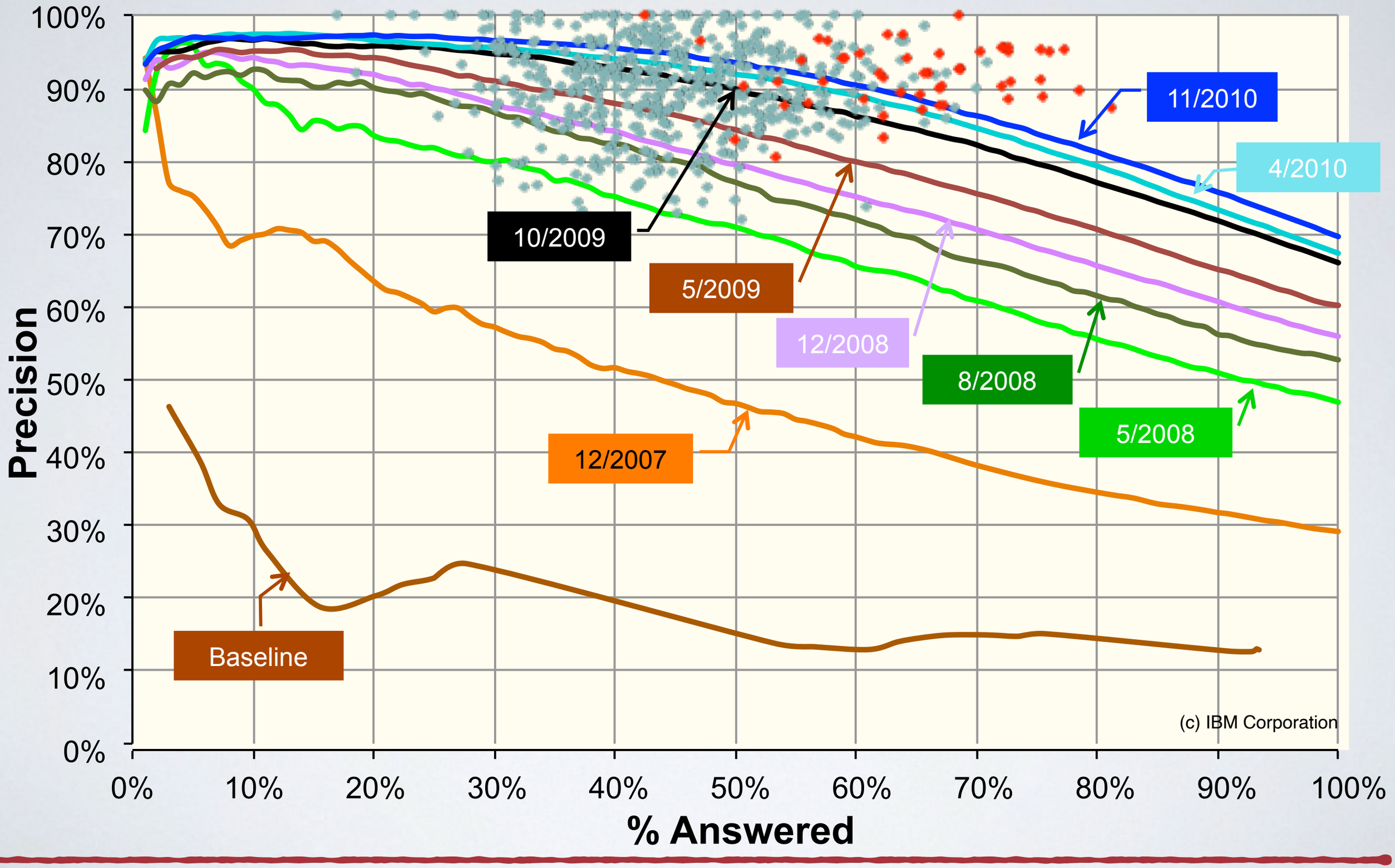
IBM Watson:
Deep QA Project

Evaluation in Action

(c) IBM Corporation. http://www.youtube.com/watch?v=3G2H3DZ8rNc

IBM Watson:
Deep QA Project

Evaluation in Action

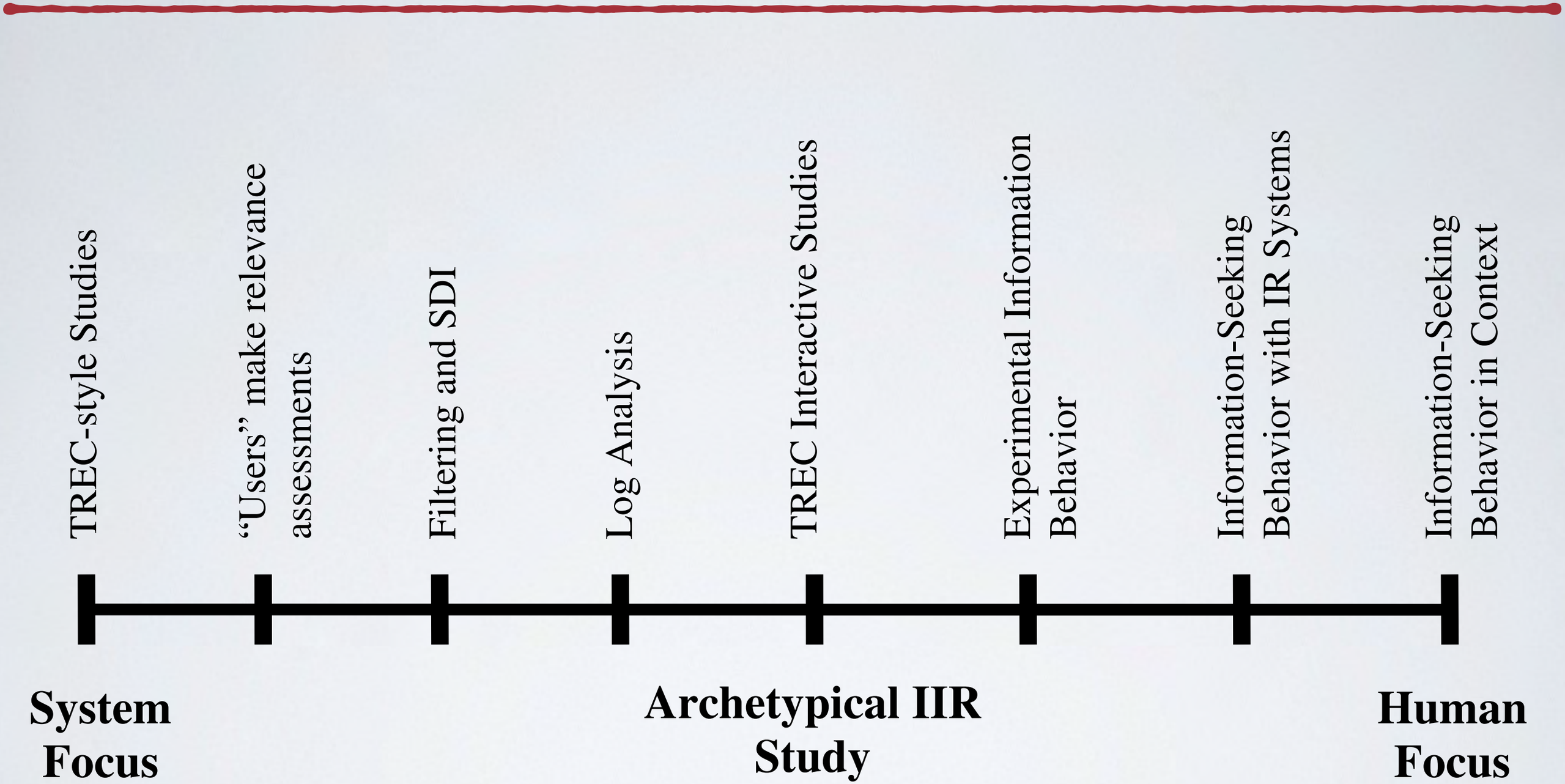


Critical Issues in Evaluation

- It must be scientifically **valid**
  - valid methodology, measures, and statistics
  - large-scale enough to be statistically valid
  - must be “repeatable” if possible
- It must be **realistic** for the applications that will be using the information retrieval systems
  - **task** and use cases
- It must be **understandable** to your audience/client

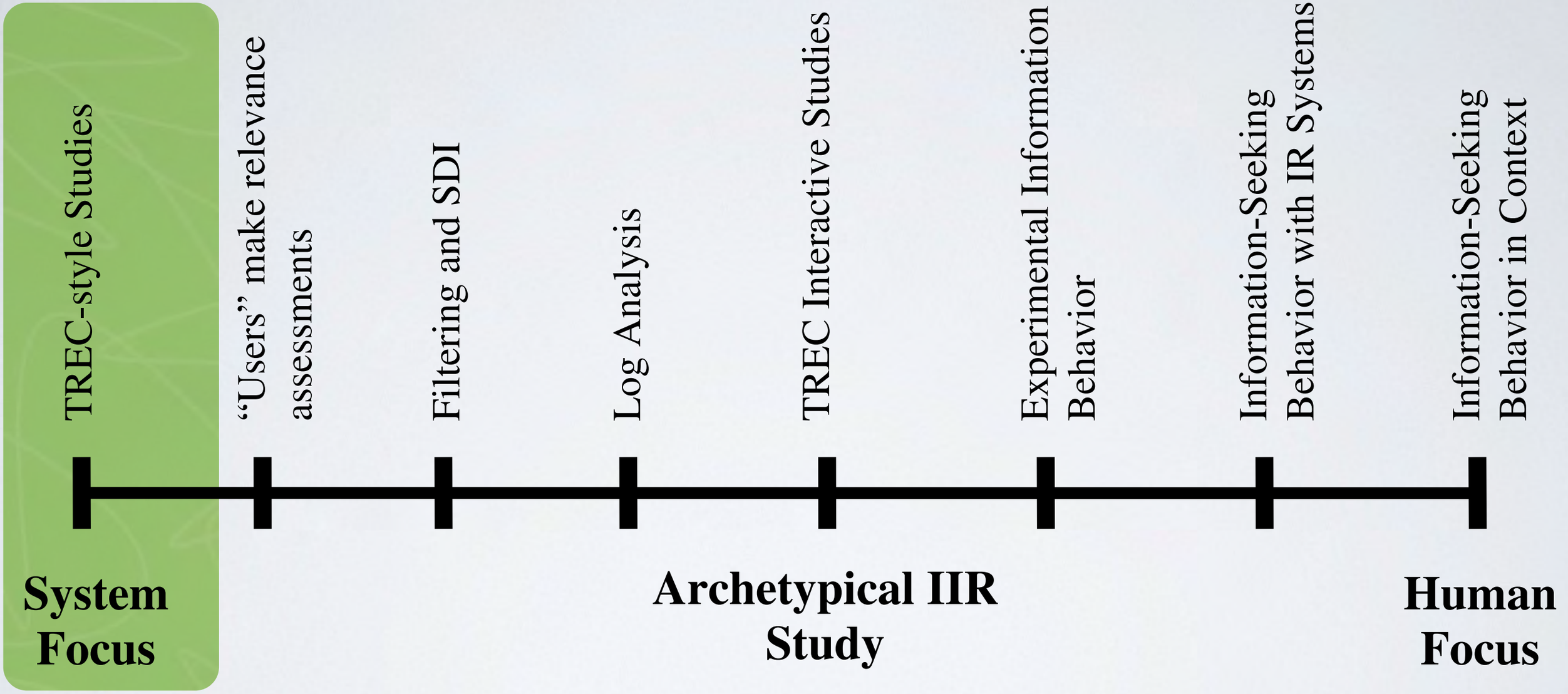
Harman, D. K. (2011). *Information Retrieval Evaluation*. Morgan & Claypool Publishers, USA.

Evaluation Spectrum



Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. Foundations and Trends in Information Retrieval (FnTIR), 3(1-2), 1-224.

Evaluation Spectrum



Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. Foundations and Trends in Information Retrieval (FnTIR), 3(1-2), 1-224.

How Does Experimental Evaluation Work

- **Cranfield Paradigm** by Cyril W. Cleverdon
  - Dates back to mid 1960s
- Makes use of **experimental collections**
  - **documents** (corpora)
  - **topics**, which are a surrogate for information needs
  - **relevance judgments** (binary or graded) also called relevance assessment or ground-truth (or qrels)
- Ensures **comparability** and **repeatability** of the experiments

Cyril W. Cleverdon

Cleverdon, C. W. (1962). *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*. Aslib Cranfield Research Project, College of Aeronautics, Cranfield, UK.

Cleverdon, C. W. (1997). *The Cranfield Tests on Index Languages Devices. In Spärck Jones, K. and Willett, P., editors, Readings in Information Retrieval*, pages 47–60. Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA.

Some Document Collections

- Historical
  - **CACM**: 3,024 abstracts from the Communications of the ACM
    http://ir.dcs.gla.ac.uk/resources/test_collections/cacm/
- Mid-nineties
  - **TIPSTER**: 528,155 documents (news articles, US government reports, …), Disks 4 and 5 excluding Congressional Record subcollection
    https://catalog.ldc.upenn.edu/LDC93T3A
- Early 2000s
  - **WT10g**: 1,692,096 Web pages crawled in 2001
    http://ir.dcs.gla.ac.uk/test_collections/wt10g.html
  - **GOV2**: 25,205,179 Web pages crawled fro .gov sites in early 2004
    http://ir.dcs.gla.ac.uk/test_collections/gov2-summary.htm
  - **CLEF Multilingual Corpus**: 4,883,227 multilingual news articles corpus in 13 languages (Bulgarian, Dutch, English, Farsi, Finnish, French, German, Hungarian, Italian, Portuguese, Spanish) gathered in 1994, 1995 and 2002. Topics in 28 different languages (Bengali, Bulgarian, Chinese, Czech, Dutch, English, Farsi, Finnish, French, German, Greek, Hindi, Hungarian, Indonesian, Italian, Japanese, Marathi, Norwegian, Oromo Polish, Portuguese, Russian, Slovenian, Spanish, Swedish, Tamil, Telugu, Thai)
- Today
  - **ClueWeb 2009**: 1,040,809,705 Web pages in 10 languages crawled between January and February 2009
    https://lemurproject.org/clueweb09/
  - **ClueWeb 2012**: 733,019,372 English Web pages crawled between February 10, 2012 and May 10, 2012
    https://lemurproject.org/clueweb12/
  - **The New York Times Annotated Corpus**: 1,855,658 news articles from January 1987 through December 2007 from New York Times
    https://catalog.ldc.upenn.edu/LDC2008T19
  - **TREC Washington Post Corpus**: 595,037 news articles and blog posts from January 2012 through August 2017 from Washington Post
    https://trec.nist.gov/data/wapost/
  - **MS MARCO**: 3.2 million English documents, 8.8 million passages, 1 million questions
    https://microsoft.github.io/msmarco/

Evaluation with Test Collections in a Nutshell

Topic | Run | Assessed Run | Weighted Assessed Run | Measure Score

Relevance Assessment

Assessed Run	Weighted Assessed Run
Highly Relevant	3
Not Relevant	0
Partially Relevant	1
Fairly Relevant	2
Not Relevant	0
Not Relevant	0

$DCG = 4.6309$

Sanderson, M. (2010). Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval (FnTIR),* 4(4):247–375.

Evaluation with Test Collections in a Nutshell

Topic

Run

Assessed Run

Weighted Assessed Run

Measure Score

Relevance Assessment

Assessed Run	Weighted Assessed Run
Highly Relevant	3
Not Relevant	0
Partially Relevant	1
Fairly Relevant	2
Not Relevant	0
Not Relevant	0

$DCG = 4.6309$

Sanderson, M. (2010). Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval (FnTIR),* 4(4):247–375.

Example of Topic

```
<?xml version="1.0" encoding="UTF-8"?>
<topic>
    <identifier>41</identifier>

    <title lang="en">Pesticides in Baby Food</title>
    <title lang="fr">Des pesticides dans la nourriture pour bébés</title>
    <title lang="it">Pesticidi negli alimenti per bambini</title>
    <title lang="ru">Пестициды в детском питании</title>
    <title lang="zh">嬰兒食品中含有殺蟲劑</title>
    <title lang="ja">ベビーフード中の病虫害防除剤</title>
    <title lang="th">ยาฆ่าแมลง ใน อาหาร เด็กอ่อน</title>
    <title lang="so">Sunta cayayaanka ee Cuntada Ilmaha</title>
    <title lang="sw">Dawa za kuulia wadudu katika Chakula cha Mtoto</title>

    <description lang="en">Find reports on pesticides in baby food.</description>
    <description lang="fr">
        Rechercher des documents sur les pesticides dans la nourriture pour bébés.
    </description>
    <description lang="it">
        Trova documenti che parlano dei pesticidi negli alimenti per bambini.
    </description>
    <description lang="ru">Найти статьи о пестицидах в детском питании</description>
    <description lang="zh">查詢有關嬰兒食品中含有殺蟲劑的報導。</description>
    <description lang="ja">ベビーフード中の病虫害防除剤に関する記事を探したい。</description>
    <description lang="th">หา รายงาน ที่ เกี่ยวข้อง กับ ยาฆ่าแมลง ใน อาหาร เด็กอ่อน</description>
    <description lang="so">Hel wargelinada sunta cayayaanka ee cuntada ilmaha.</description>
    <description lang="sw">
        Pata ripoti kuhusu dawa za kuulia wadudu katika chakula cha mtoto.
    </description>

    <narrative lang="en">
        Relevant documents give information on the discovery of pesticides in baby food.
        They report on different brands, supermarkets, and companies selling baby food
        which contains pesticides. They also discuss measures against the contamination
        of baby food by pesticides.
    </narrative>
    <narrative lang="fr">
        Les documents pertinents informent sur la découverte de pesticides dans la
        nourriture pour bébés. Ils contiennent des informations sur les différentes
        marques, les supermarchés et les firmes ayant mis en vente de la nourriture pour
        bébés renfermant des pesticides. Ils relatent également les mesures prises contre
        la contamination de la nourriture pour bébés par les pesticides.
    </narrative>
    <narrative lang="it">
        I documenti rilevanti forniscono informazioni sulla scoperta di pesticidi nei
        cibi per bambini. Riportano i diversi marchi, i supermercati e le ditte che hanno
        venduto alimenti per bambini con i pesticidi. Sono anche rilevanti i documenti
        che discutono le misure contro la contaminazione degli alimenti per bambini con
        i pesticidi.
    </narrative>
</topic>
```

- Topics consists of:
  - **title**: a brief statement expressing the information need. It resembles the typical search engine query
  - **description**: more detailed formulation of the information need
  - **narrative**: instructions for assessors on when to consider a document relevant
- Typical experimental collections make use of **50 topics**

Evaluation with Test Collections in a Nutshell

Topic

Run

Relevance Assessment

Assessed Run

Highly Relevant
Not Relevant
Partially Relevant
Fairly Relevant
Not Relevant
Not Relevant

Weighted Assessed Run

3
0
1
2
0
0

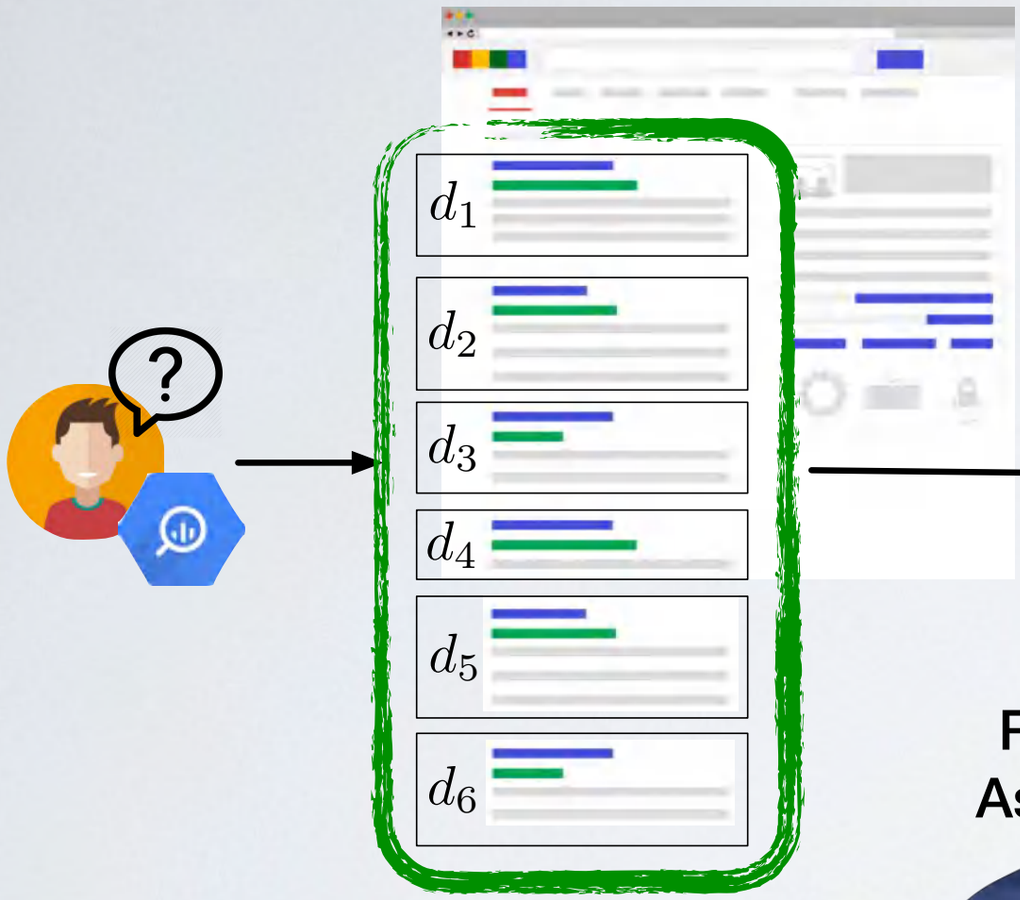
Measure Score

$$DCG = 4.6309$$

Sanderson, M. (2010). Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval (FnTIR),* 4(4):247–375.

Evaluation with Test Collections in a Nutshell

Topic | Run | Assessed Run | Weighted Assessed Run | Measure Score



Relevance Assessment

Assessed Run	Weighted Assessed Run
Highly Relevant	3
Not Relevant	0
Partially Relevant	1
Fairly Relevant	2
Not Relevant	0
Not Relevant	0

$$DCG = 4.6309$$

Sanderson, M. (2010). Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval (FnTIR),* 4(4):247–375.

Example of Relevant Document

<DOC>
<DOCNO> LA091294-0164 </DOCNO>
<DOCID> 078484 </DOCID>
<SOURCE>
<P>
Los Angeles Times
</P>
</SOURCE>
<DATE>
<P>
September 12, 1994, Monday, Home Edition
</P>
</DATE>
<SECTION>
<P>
Part A; Page 19; Column 1; Metro Desk
</P>
</SECTION>
<LENGTH>
<P>
1502 words
</P>
</LENGTH>
<HEADLINE>
<P>
ALAR PANIC SHOWS POWER OF MEDIA TO TRIGGER FEAR; HEALTH: '60 MINUTES' BROADCAST CREATED SCARE AT TIME WHEN THE INDUSTRY WAS ALREADY MOVING AWAY FROM USE OF THE CHEMICAL.
</P>
</HEADLINE>
<BYLINE>
<P>
By DAVID SHAW, TIMES STAFF WRITER
</P>
</BYLINE>
<TEXT>
<P>
The warning -- uttered on prime-time network television, against a backdrop of a giant apple marked with a skull and crossbones -- was indeed ominous:
</P>
<P>
"The most potent cancer-causing agent in our food supply is a substance sprayed on apples to keep them on the trees longer and make them look better."
</P>
<P>
So said Ed Bradley on "60 Minutes" on Feb. 26, 1989, and so began a nationwide panic, fed by other media, which quickly followed "60 Minutes" 's lead with their own stories on the killer chemical daminozide (better known by its trade name -- Alar).
</P>
<P>
Young children, the media reported, were especially vulnerable, because they tended to drink a lot of apple juice and because their digestive and immune systems were not fully developed.
</P>
<P>
But at the time of the "60 Minutes" broadcast -- which was viewed by an estimated 40 million Americans -- industry was already moving away from Alar, and the nation's three major baby-food makers said they were using non-Alar apples.
</P>
<P>
The Natural Resources Defense Council, an activist environmental group, wanted the U.S. Environmental Protection Agency to ban Alar and several other pesticides. When the group produced a report ("Intolerable Risk: Pesticides in Our Children's Food") condemning Alar, Newsweek rushed into print with the story before the report was officially released.
</P>
<P>
The report, Newsweek said, would "almost certainly generate frightening headlines."
</P>
<P>
Indeed it did -- three weeks later, right after "60 Minutes" aired its story. News media throughout the country -- effectively manipulated by the Natural Resources Defense Council and aided by public appeals and congressional testimony from that well-known molecular biologist Meryl Streep -- almost made it seem that one bite of an Alar-treated apple or one swig of juice made from Alar-treated apples would mean instant death.
</P>
<P>
Coverage of the Alar scare was "outrageous . . . completely alarmist," says Marla Cone, who writes about the environment for the Los Angeles Times.
</P>
<P>
But Alar was a made-to-order media story. It had apples, kids and cancer. "A lot of (media) people were suckered in," Cone says.
</P>
<P>
The media coverage produced a nationwide hysteria.
</P>
<P>
School boards in Los Angeles, New York, Chicago, Atlanta and many other cities banned apples and apple products from their cafeterias. Some parents raced after their children's school buses to yank apples from their lunch boxes. Supermarkets came under intense pressure to remove apples form their shelves. Uniroyal, the manufacturer of Alar, pulled the product off the market. Sales of apples plummeted, forcing many farmers to dump their crops or give them away -- costing the industry more than $100 million, according to economists' estimates.
</P>
<P>
Reaction to the Alar scare "set a troubling precedent," the Washington Post editorialized several weeks later. "A complicated scientific issue was allowed to be decided not by officials charged with protecting the public, on the basis of hard evidence, but by a frightened public acting on incomplete and often erroneous press reports."
</P>
<P>
The EPA had expressed concern about the safety of Alar for many years before Newsweek and "60 Minutes" jumped on the story. But the agency had decided that test results were either flawed, contradictory or insufficiently conclusive to warrant an immediate ban and formal action was delayed, pending hearings in late 1990.
</P>
<P>
David Gelber, the producer of the "60 Minutes" Alar story and now the executive producer for ABC's "Peter Jennings Reporting," says EPA and scientific criticism of Alar convinced him the story was worth doing at the time.
</P>
<P>
Dr. John A. Moore, then acting administrator of the EPA, said on "60 Minutes" that Alar "should come off the market" because of what he had earlier described as "an inescapable and direct correlation" between exposure to Alar and "the development of life-threatening tumors."
</P>
<P>
"The public had a right to know that was their view," Gelber says.
</P>
<P>
Al Meyerhoff, senior attorney for the Natural Resources Defense Fund, also defends the Alar story.
</P>
<P>
He says the apple industry launched a "concerted disinformation campaign" in an effort to persuade the media and consumers alike that Alar was not

Example of Relevant Document

<DOC>
<DOCNO> LA041694-0248 </DOCNO>
<DOCID> 032882 </DOCID>
<SOURCE>
<P>
Los Angeles Times
</P>
</SOURCE>
<DATE>
<P>
April 16, 1994, Saturday, Home Edition
</P>
</DATE>
<SECTION>
<P>
Business; Part D; Page 1; Column 2; Financial Desk
</P>
</SECTION>
<LENGTH>
<P>
1135 words
</P>
</LENGTH>
<HEADLINE>
<P>
WARNING FLAG OVER GATT; ENVIRONMENTALISTS SAY PACT WILL WEAKEN U.S. SAFEGUARDS
</P>
</HEADLINE>
<BYLINE>
<P>
By MELISSA HEALY, TIMES STAFF WRITER
</P>
</BYLINE>
<TEXT>
<P>
The environmental community, which recently split deeply over support for the North American Free Trade Agreement, is issuing warnings about the new world trade agreement in a newly unified voice.
</P>
<P>
The environmentalists charge that the comprehensive General Agreement on Tariffs and Trade accord signed Friday in Morocco will erode the United States' -- and California's -- ability to enforce its environmental strictures on everything from recycling to pesticide use to air pollution.
</P>
<P>
"If this . . . is enacted," said Barbara Dudley, executive director of Greenpeace USA, "over two decades of environmental protection could be severely weakened."
</P>
<P>
Nowhere is that more true than in the area of food safety, some environmentalists argue.
</P>
<P>
The United States, with some of the world's most restrictive regulation of pesticides, prohibits the entry of food products with detectable traces of about 40 chemicals -- substances used by many of its trading partners and listed as allowable by the standard-setting organization of the new agreement.
</P>
<P>
But if U.S. Customs Service inspectors begin turning away food imports that bear traces of these chemicals under the new agreement, America's trading partners are almost certain to cry foul, environmentalists warned. If their challenges stand, one activist said, U.S. pesticide protections could topple, one after another.
</P>
<P>
Supporters of the accord acknowledged that the letter of the new agreement may indeed put the United States, with its strong environmental protections, on the defensive. That is because some U.S. laws championed by environmentalists, as well as state laws, do not appear to be based on undisputed scientific evidence demonstrating that a regulation will improve the public's health or mitigate a known environmental hazard.
</P>
<P>
Supporters of the agreement also argue that the United States' market power, as well as a growing appreciation for the environmental ethic among America's trading partners, will cause virtually any challenge to U.S. environmental laws to fail.
</P>
<P>
"The practical implementation of trade agreements is often more politically sensitive and realistic than the sheer language of the treaty," said William K. Reilly, administrator of the U.S. Environmental Protection Agency in the George Bush Administration and an avowed free trader.
</P>
<P>
Other free traders cited the power and transparency of the processes by which U.S. laws are passed and regulations are made. In a challenge before the international trade deliberative body, U.S. defenders could produce reams of scientific data, risk assessments and economic analysis to prove that an American environmental standard was established not to keep foreign products out, but to benefit the public's health.
</P>
<P>
"In a fair fight," said Linda Fisher, a Washington-based trade attorney with the Los Angeles law firm Latham & Watkins, "the United States will win."
</P>
<P>
But environmental activists are not comforted by those assurances. They say that the fine print of the new agreement would allow a trading partner to argue before the trade court in Geneva, Switzerland, that Washington's environmental laws -- or those of individual U.S. states -- constitute an unfair barrier to the entry of that country's exports.
</P>
<P>
If the United States regulates more strictly than its trading partners the fuel emissions or stipulates the fuel efficiency of cars sold or operated within its borders (which it does), or prohibits food products that bear traces of certain pesticides it considers hazardous (which it does), a trading partner, in principle, can challenge the federal stricture under the trade agreement.
</P>
<P>
Environmentalists said that under the trade agreement, trading partners could target laws such as the one that grew out of California's Proposition 65, which requires a cautionary label on any product that would expose its user to a carcinogen or a chemical that could be harmful to a developing fetus or pregnant woman.
</P>
<P>
Other California state regulations that could be challenged go beyond federal government requirements by making manufacturers of agricultural chemicals furnish the state with data on the chemical's possible effects on human reproduction, water pollution, exposed workers and endangered species.
</P>
<P>
Another California stricture that could come under attack is one that has required wine manufacturers to sponsor efforts to warn consumers of the possible dangers posed by the lead in the foil that covers wine corks. Finally ... A World Trade Pact
</P>
<P>
Ministers from 124 nations ended seven years of complex negotiations Friday and formally concluded the Uruguay Round of the General Agreement on Tariffs and Trade talks in Marrakesh, Morroco. Following are the main elements of the 26,000-page, 385-pound global world trade pact:
</P>
<P>
MARKET ACCESS - This is the backbone of the act. Countries pledge to cut tariffs on industrial and farm goods by an average of about 37%. The United States and European Union agree to trim tariffs between them by one half.
</P>
<P>
SERVICES - For the first time, rules will govern annual trade in services such as banking, insurance and travel, as well as the movement of labor. The United States reserves the right to deny other countries favorable access to the lucrative U.S. financial services market, but will hold off for at least 18 months. Washington has threatened to challenge EC curbs on audio-visual goods.
</P>

Example of Not Relevant Document

<DOC>
<DOCNO> LA072894-0040 </DOCNO>
<DOCID> 064623 </DOCID>
<SOURCE>
<P>
Los Angeles Times
</P>
</SOURCE>
<DATE>
<P>
July 28, 1994, Thursday, Home Edition
</P>
</DATE>
<SECTION>
<P>
Food; Part H; Page 24; Column 1
</P>
</SECTION>
<LENGTH>
<P>
3275 words
</P>
</LENGTH>
<HEADLINE>
<P>
NEW VIETNAMESE CUISINE; MARKETS; THE BEST OF LITTLE SAIGON: MADE IN AMERICA
</P>
</HEADLINE>
<BYLINE>
<P>
By LINDA BURUM
</P>
</BYLINE>
<TEXT>
<P>
* Little Saigon Supermarket, 9822 Bolsa Ave., Westminster, (714) 531-7272, 9 a.m. to 9 p.m. daily.
</P>
<P>
The powerful whirling helicopter blades kicked up thick clouds of dust as Twan Ngo pushed her youngest child into a hovering American chopper. It was April 27, 1975, three days before the fall of Saigon and the last time Ngo would see her suburban Saigon home.
</P>
<P>
Many of the 165,000 Vietnamese who came to the United States in the first post-war wave of immigrants were relocated in Northeastern or Midwestern cities. At first the Ngos found themselves in Green Bay, Wis., where, as you might expect, they found little in the way of Vietnamese food.
</P>
<P>
Eventually, along with many fellow transplanted Vietnamese, the Ngos wound up in Orange County near the neighborhood centered on Bolsa Avenue -- and running from Westminster through Garden Grove to Santa Ana -- that came to be known as Little Saigon. Lured by the warmer climate and a growing, cohesive Vietnamese community, the Ngos found many ties to their culture here, one of the strongest being the ready availability of Vietnamese foods and ingredients.
</P>
<P>
"You won't find most of these vegetables in Wisconsin," Ngo says, picking over the greens in the produce section of the bright, ultra-modern Little Saigon Supermarket on Bolsa Ave. By "these vegetables," she means the dozens of Vietnamese specialties such as bac ha , the pale green spongy stems that go into sour soup, or muop huong , a squash resembling a large zucchini.
</P>
<P>
*
</P>
<P>
In recent years, specialty produce farming has turned into a livelihood for a number of local Vietnamese farmers, evolving hand in hand with a now-sizable Vietnamese food industry. Many of these California-grown or -made goods are distributed nationwide to serve the million or so Vietnamese who have come to the United States over the years.
</P>
<P>
At the center of this commercial activity is Orange County, where Vietnamese residents officially increased 271% in the last decade (the Vietnamese-American Political Action Committee contends the number is nearly twice what the census reports). The sheer size of this local customer base opened up a lucrative market for prepared foods as well as shelf-stable items; in much smaller Vietnamese communities, it probably wouldn't be economic to market these perishables.
</P>
<P>
But at Little Saigon Supermarket, Ngo chooses from an increasing range of Vietnamese-style processed meats, all sorts of fresh noodles, herbs, pickled vegetables and soy products (such as fresh tofu and soy milk) that are impossible to import. The selection has made Vietnamese eating here as close to authentic as it gets outside of Vietnam itself.
</P>
<P>
As the selection shows, Little Saigon Supermarket owner David Tran knows the Vietnamese food business inside and out. Tran came to the Little Saigon area when it was still in its awkward growth stages -- a mere stretch of bean and strawberry fields, flower warehouses and bottling plants, and just a few Vietnamese stores and businesses.
</P>
<P>
*
</P>
<P>
He had left Vietnam in '78 on a crowded boat and ended up temporarily in Escondido, learning English and studying to be an electronic technician. Tran found it impossible to support his family on a technician's wages, but having been a businessman (a bicycle wholesaler, to be exact) in Vietnam, he perceived a growing need for Vietnamese wholesale food distribution. He formed Delta Food Company and began to supply Orange County's escalating number of Vietnamese restaurants and markets with wholesale produce and other ingredients used for Vietnamese and Chinese cooking.
</P>
<P>
As time went on, local Asian farmers began to produce small quantities of specialty herbs and vegetables. "Some would ask him to distribute them," says Denise Tran, who is David's sister-in-law and also vice-president of the market.
</P>
<P>
The 1975 Vietnam trade embargo meant an end to Vietnamese imports. Producers in other Asian countries, particularly Thailand, began putting Vietnamese-language labels on foods they had in common with Vietnam, such as fish paste, dry rice noodles and curry powder, and exporting them to the US for the expatriate market. In the beginning, Tran stocked a lot of these items, but they weren't created specifically for the Vietnamese palate, and many Vietnamese cooks have never ceased to regard them as mere substitutes.
</P>
<P>
Now, however, Vietnamese in this country are no longer dependent on imports. They're producing their own extravagant assortments of Vietnamese-style sauces, spice blends, pickled fish, fresh rice papers, deli foods, sweets, beverages and baked goods, all made in America.
</P>
<P>
*
</P>
<P>
There's as much competition among the various brands of these foods as there is between American breakfast cereal or coffee companies. You see five or six styles of curry powder and at least that many of the fresh hot chile sauce called tuong ot toi (not to mention many other varieties of hot sauce). Four companies manufacture the skinny fresh rice-noodle cakes called banh hoi and at least half a dozen make the bologna-like Vietnamese sausages chua lua and cha bi. The UPC bar-coded packages of fish dumpling paste in the freezer case also come from several companies.
</P>
<P>
In addition, Little Saigon Supermarket stocks prepared deli foods such as cakes, vegetable-filled rice noodle rolls, taro balls, mung bean desserts and other tidbits, all from bakeries and snack shops in the area. These are displayed on a huge table in the middle of the store and their packaging is still rather basic: foam trays enveloped in shrink wrap or "clam shell" boxes designed for take-out foods. But the manufactured foods' packaging is starting to get more sophisticated.
</P>
<P>
Now there are even convenience products such as spice packets to flavor duck soup, mixes for Vietnamese rice rolls and a Hamburger Helper-like mix to season bo kho , the Vietnamese beef stew. So far though, no Vietnamese TV dinners have become part of this burgeoning industry's repertoire. Shopping List
</P>
<P>
It's not possible to cover all the locally made Vietnamese products here but this will give you a little taste of what's out there. SOUP SEASONING MIXES
</P>
<P>
Noodle soups are a daily food in Vietnam, as popular as breakfast cereals or hamburgers in this country. Over there, few people bother to make them at home,

Where Do Topics and Collections Come From?

- Topics
  - Collection exploration
  - Logs
  - Observation of real users
- Collections
  - Opportunistic, e.g. my email or university Web site
  - Constructed, e.g. tweets with #hashtag or results of a query to a search engine
  - Naturalistic, e.g. large Web crawl, a year of news, a month of tweets

Soboroff, I. (2017). Building Test Collections: An Interactive Guide for Students and Others Without Their Own Evaluation Conference Series. In Kando, N., Sakai, T., Joho, H., Li, H., de Vries, A. P., and White, R. W., editors, *Proc. 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017)*, pages 1407–1410. ACM Press, New York, USA.

Evaluation with Test Collections in a Nutshell

Topic | Run | Assessed Run | Weighted Assessed Run | Measure Score

Relevance Assessment

Assessed Run	Weighted Assessed Run
Highly Relevant	3
Not Relevant	0
Partially Relevant	1
Fairly Relevant	2
Not Relevant	0
Not Relevant	0

$DCG = 4.6309$

Sanderson, M. (2010). Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval (FnTIR),* 4(4):247–375.

Evaluation with Test Collections in a Nutshell

Topic | Run | Assessed Run | Weighted Assessed Run | Measure Score

Assessed Run	Weighted Assessed Run
Highly Relevant	3
Not Relevant	0
Partially Relevant	1
Fairly Relevant	2
Not Relevant	0
Not Relevant	0

$$DCG = 4.6309$$

Relevance Assessment

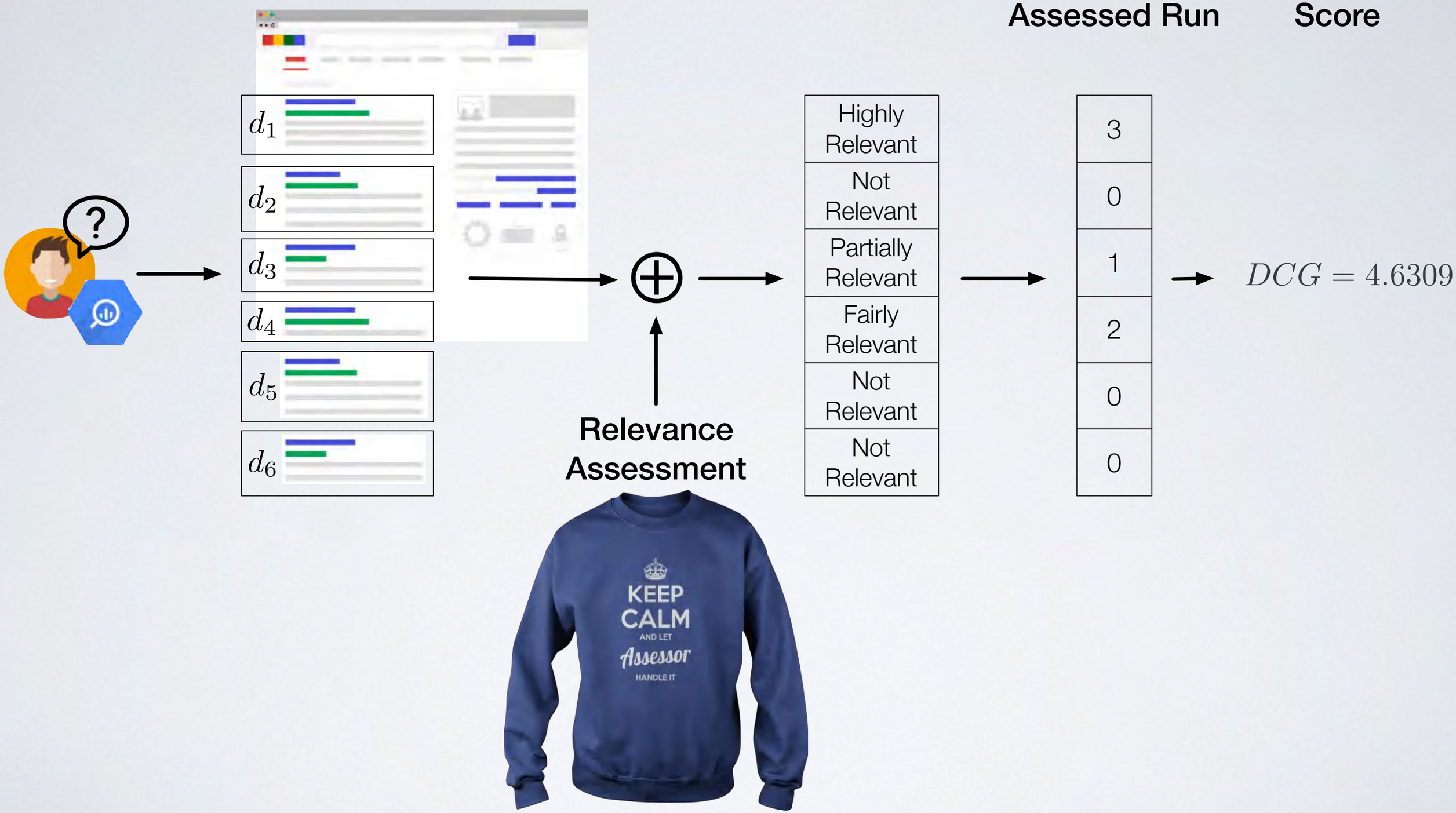
Sanderson, M. (2010). Test Collection Based Evaluation of Information [illegible] *Trends in Information Retrieval (FnTIR),* 4(4):247–375.

Example of Ground-truth (trec_eval format)

Topic ID	Fixed	Document ID	Judgement
41	0	LA050394-0237	0
41	0	LA112394-0177	0
41	0	LA091294-0164	1
41	0	LA040594-0187	0
41	0	LA041694-0248	1
		…	
42	0	LA031694-0234	0
42	0	LA040494-0111	0
42	0	LA081794-0171	1

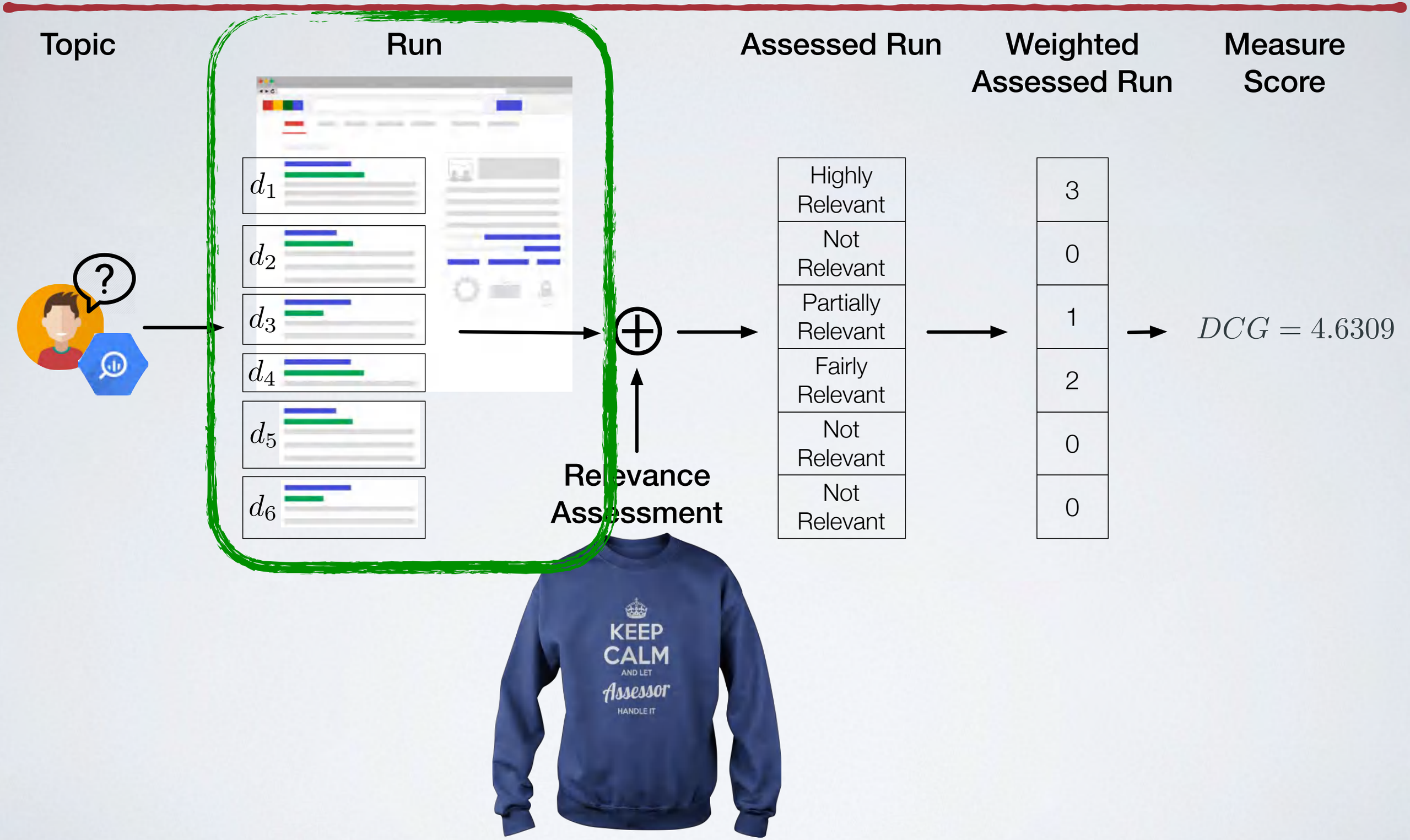
- Relevance judgements (qrels) are textual files whose field are separated by tab or space
- Typically, for each topic there are 300-700 judgement documents and the number of judged document vary from topic to topic

Evaluation with Test Collections in a Nutshell



Sanderson, M. (2010). Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval (FnTIR),* 4(4):247–375.

Evaluation with Test Collections in a Nutshell



Sanderson, M. (2010). Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval (FnTIR),* 4(4):247–375.

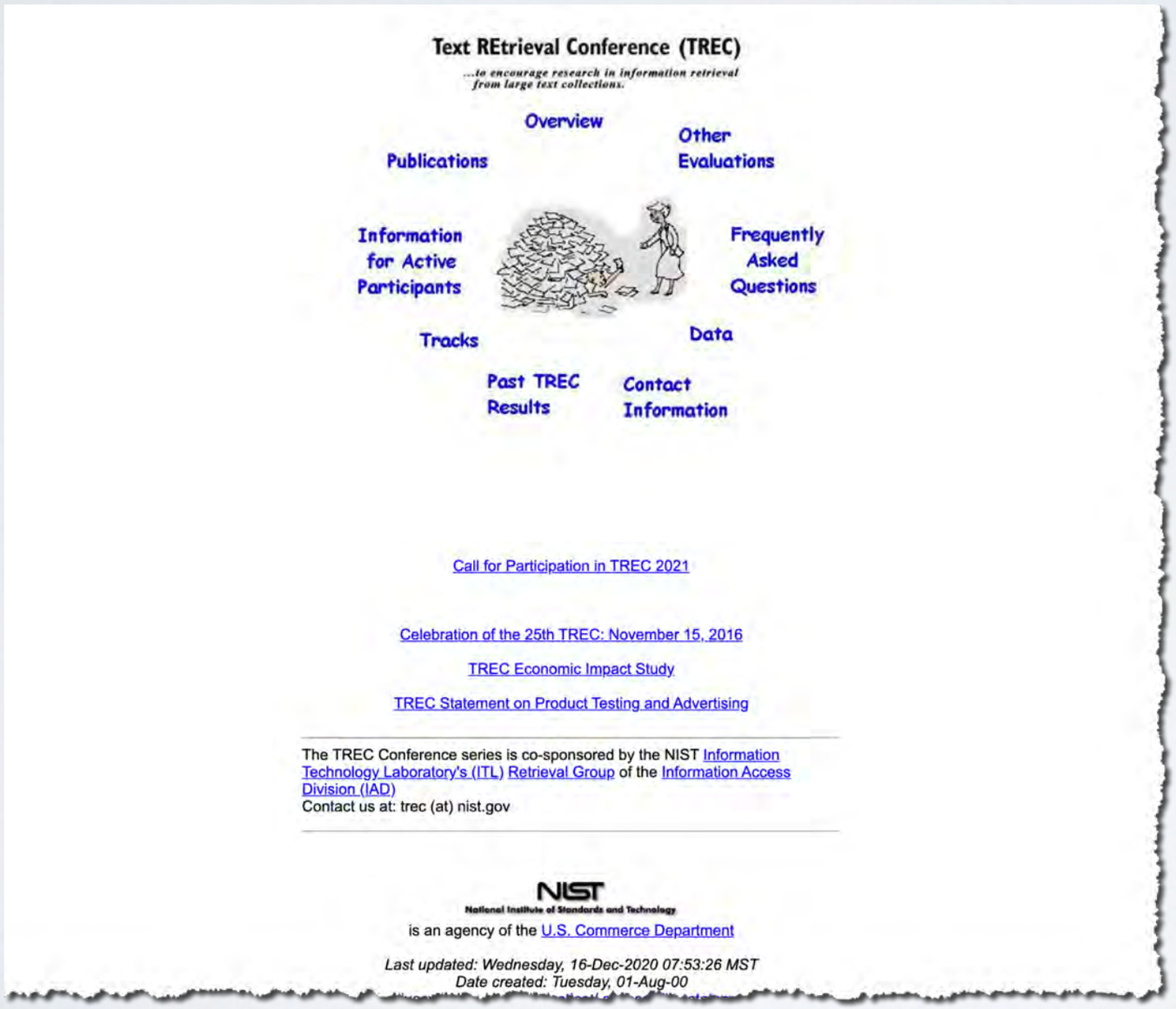
Example of Run (trec_eval format)

Topic ID	Fixed	Document ID	Rank	Score	Run ID
41	Q0	LA050394-0237	1	0.6342	updrun
41	Q0	LA091294-0164	2	0.4278	updrun
41	Q0	LA040594-0187	3	0.4278	updrun
41	Q0	LA041694-0248	4	0.3197	updrun
41	Q0	LA102394-0113	5	0.3005	updrun
		…			
42	Q0	LA081794-0171	1	0.7687	updrun
42	Q0	LA031694-0235	2	0.7011	updrun
42	Q0	LA031694-0234	3	0.6950	updrun

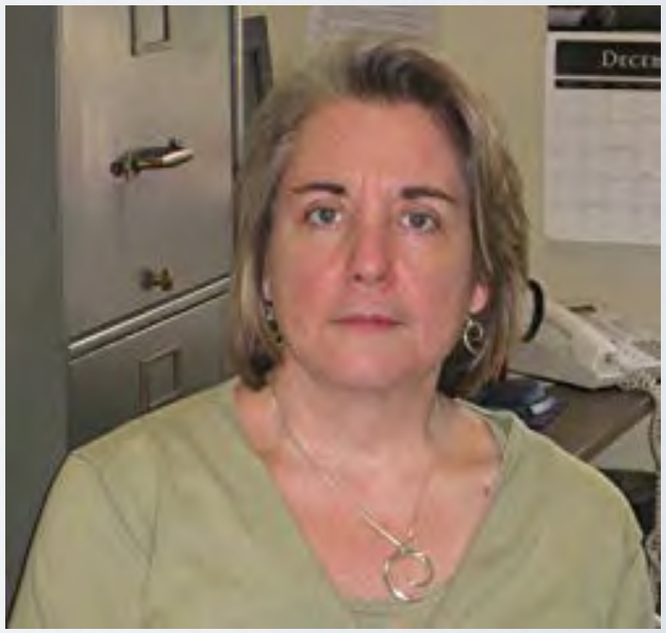
- Runs are textual files whose field are separated by tab or space
- Typically, there are 50 topics and 1,000 documents are retrieved for each topic (i.e. 50,000 lines)

Large-scale Evaluation Initiatives: TREC

- **TREC** (Text REtrieval Conference), USA, since 1992
  - https://trec.nist.gov/



Donna Harman



Ellen M. Voorhees

Harman, D. K. and Voorhees, E. M., editors (2005). *TREC. Experiment and Evaluation in Information Retrieval*. MIT Press, USA.

Large-scale Evaluation Initiatives: NTCIR

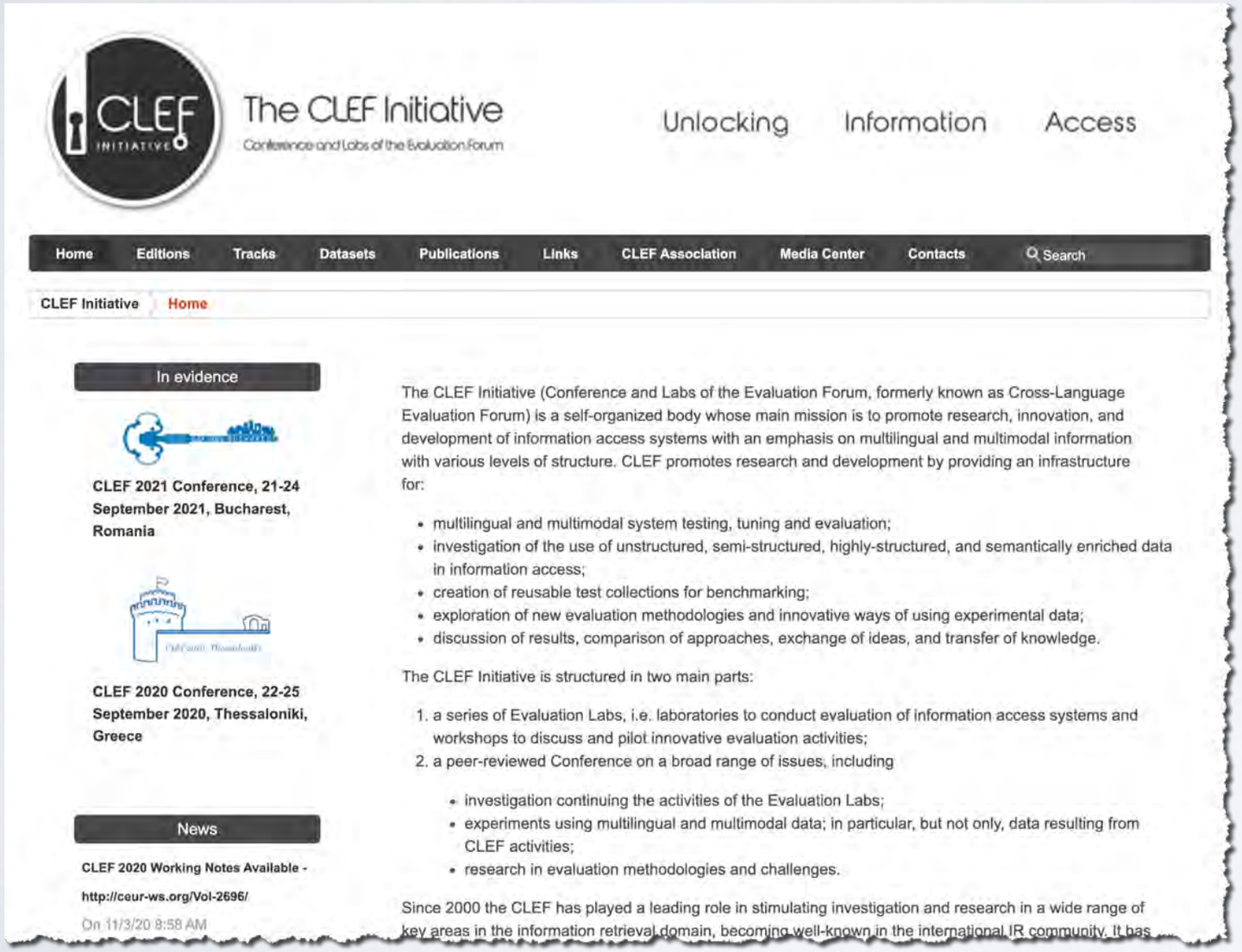
- **NTCIR** (NII Testbeds and Community for Information access Research), Japan, since 1999
  - http://research.nii.ac.jp/ntcir/index-en.html

Noriko Kando

Sakai, T., Oard, D. W., and Kando, N., editors (2021). *Evaluating Information Retrieval and Access Tasks – NTCIR's Legacy of Research Impact*, volume 43 of *The Information Retrieval Series*. Springer International Publishing, Germany.

Large-scale Evaluation Initiatives: CLEF

- **CLEF** (Conference and Labs of the Evaluation Forum), Europe, since 2000
  - http://www.clef-initiative.eu/



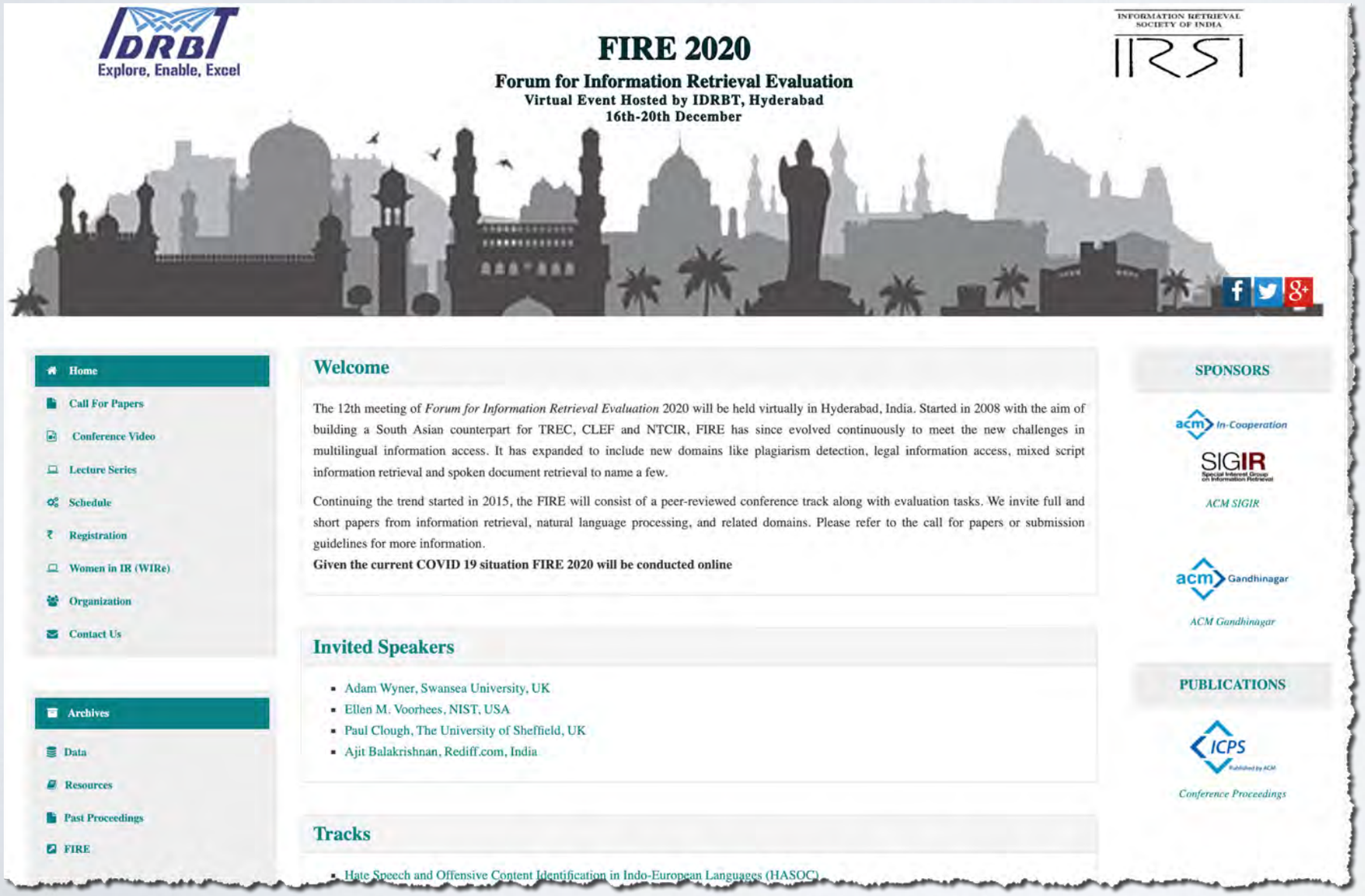
Carol Ann Peters

Nicola Ferro

Ferro, N. and Peters, C., editors (2019). *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*. Springer International Publishing, Germany.

Large-scale Evaluation Initiatives: FIRE

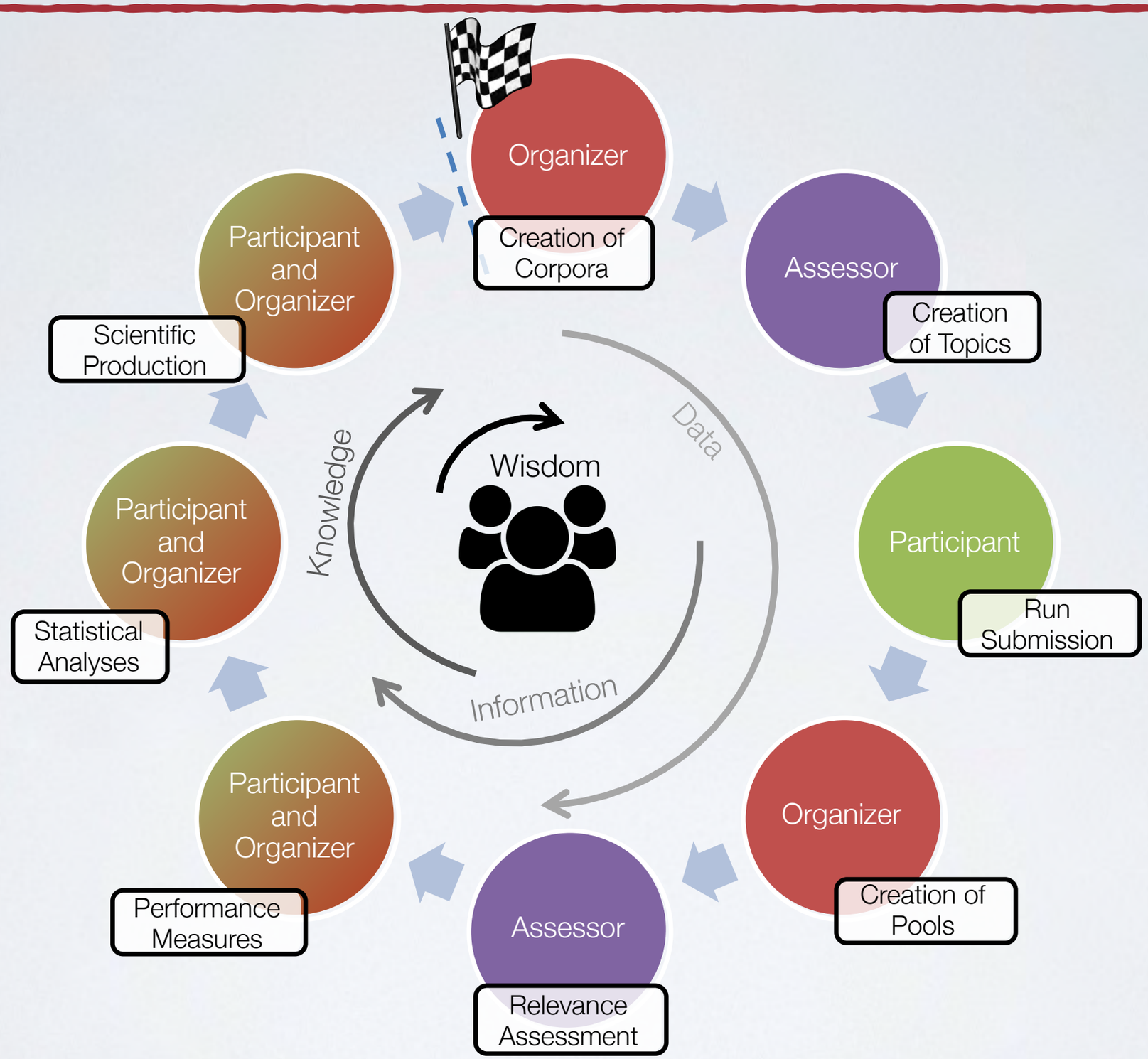
- **FIRE** (Forum for Information Retrieval Evaluation), India, since 2008
  - http://fire.irsi.res.in/



Mandar Mitra

Prasenjit Majumder

Evaluation Initiatives: Typical Cycle



Dussin, M. and Ferro, N. (2009). Managing the Knowledge Creation Process of Large-Scale Evaluation Campaigns. In Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., and Tsakonas, G., editors, Proc. *13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009)*, pages 63–74. LNCS 5714, Springer, Germany.

The “Ideal Test Collection” Today

- 😄 **Corpora** → (not historical) corpora are typically OK
  - < 500 documents, no real value
  - 1-2,000 documents, minimally acceptable
  - > 10,000 documents, actually needed
- 🥶 **Topics** → typical size is still 50 topics
  - < 75 topics, no real value
  - 250 topics, minimally acceptable
  - > 1,000 topics, actually needed
- 🤯 **Relevance Judgements** → binary is still most common option, diversity only recently
  - multi-graded (highly and fairly relevant)
  - types (novel, stimulating, …)
  - need for **pooling** (still open research issue)

Karen Spärck Jones

C. J. “Keith” van Rijsbergen

Spärck Jones, K. and van Rijsbergen, C. J. (1975). *Report on the need for and provision of an ‘ideal’ information retrieval test collection*. British Library Research and Development Report 5266, University Computer Laboratory, Cambridge.

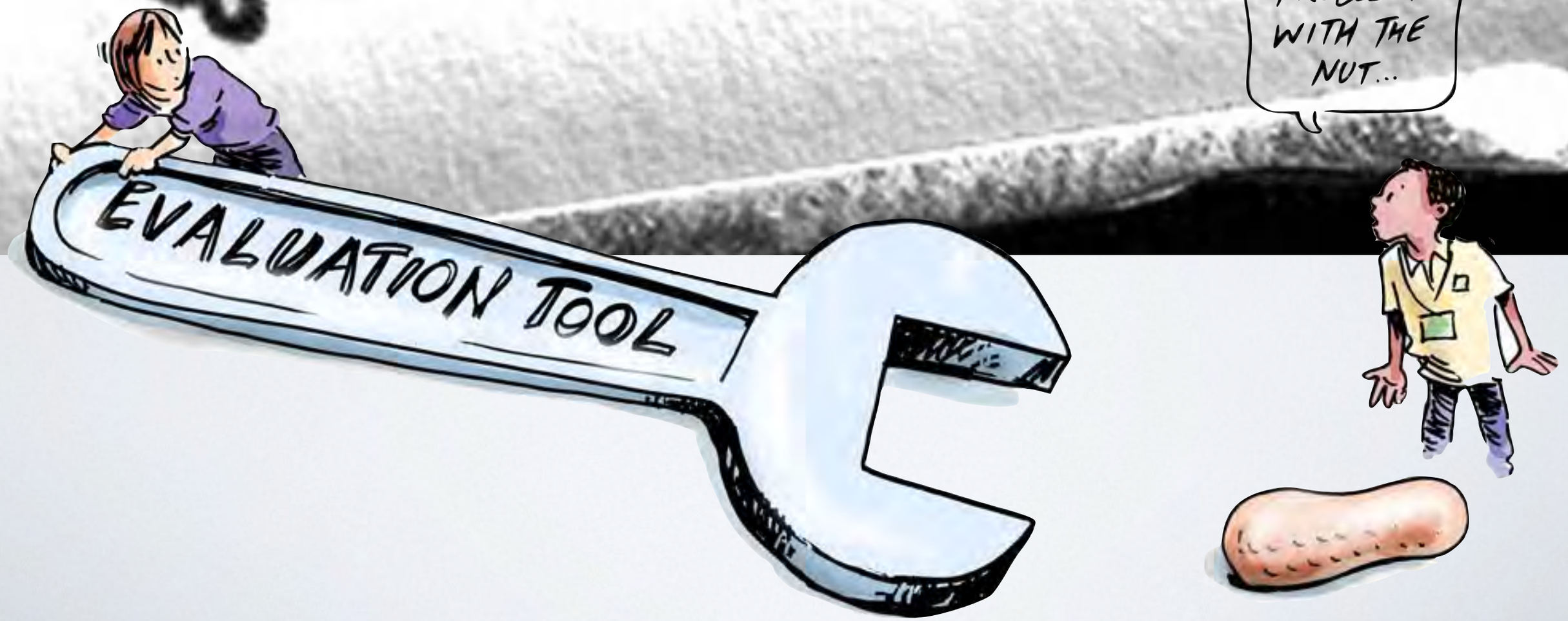
How Valuable is Evaluation?

- The **TREC 2010 Economic Impact** study estimated in about **30 M$** the overall **investment in TREC** by NIST
  - probably much much more if we had a means to estimate also the investment by participants in TREC

- They are the **pillars** for all the subsequent **scientific research** and **technology development**
  - TREC estimated the **return on investment** in the range of **3-5** for each invested dollar

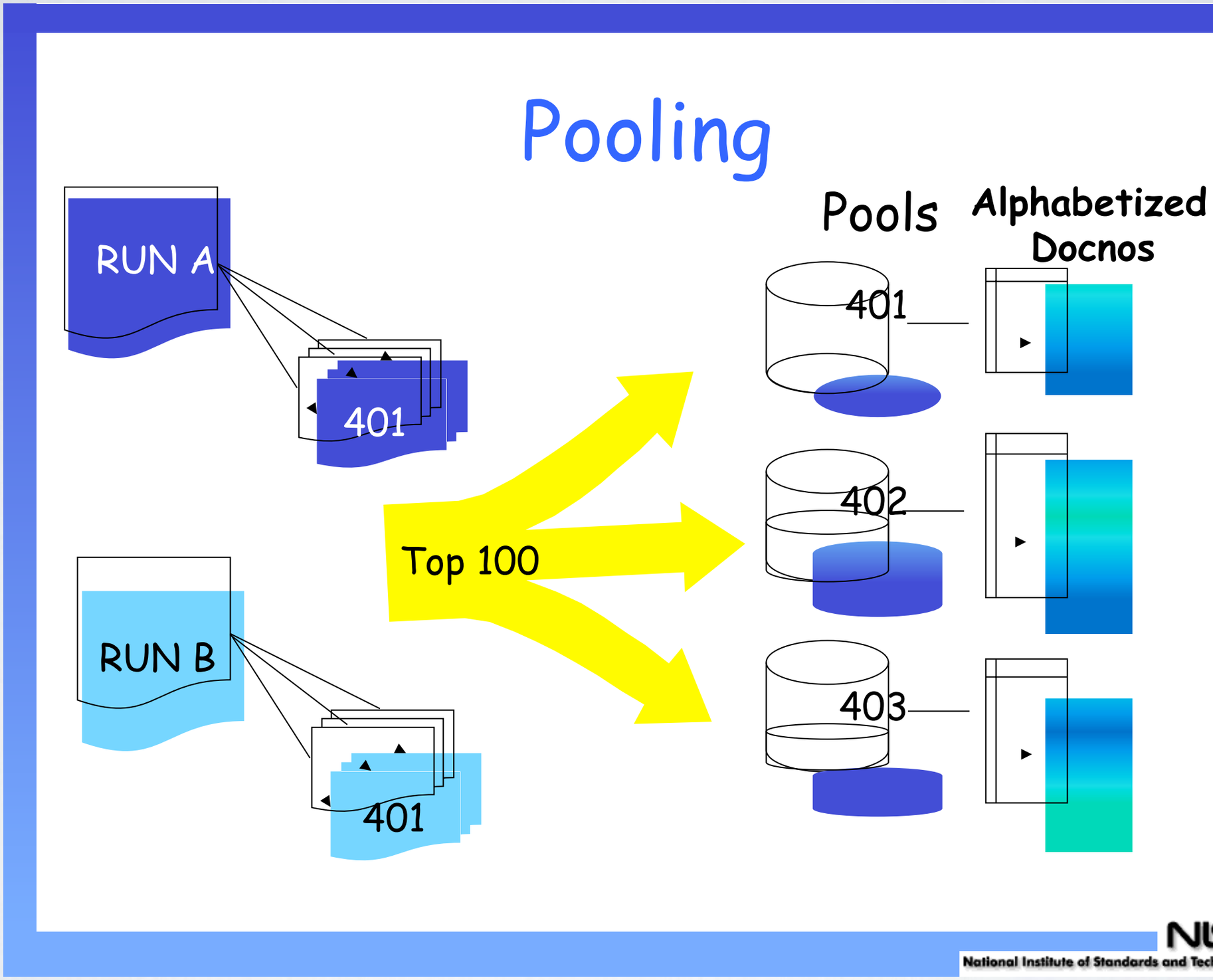
Rowe, B. R., Wood, D. W., Link, A. L., and Simoni, D. A. (2010). *Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program*. RTI Project Number 0211875, RTI International, USA. http: //trec.nist.gov/pubs/2010.economic.impact.pdf.

Questions?



Relevance Assessment

Traditional Depth-k Pools



Harman, D. K. (2013). TREC-Style Evaluations. In Agosti, M., Ferro, N., Forner, P., Müller, H., and Santucci, G., editors, *Information Retrieval Meets Information Visualization – PROMISE Winter School 2012, Revised Tutorial Lectures*, pages 97–115. Lecture Notes in Computer Science (LNCS) 7757, Springer, Heidelberg, Germany.

Relevance Assessment

2 judgments per minute

75 person/days per pool

35,000-75,000 documents per pool

NIST
National Institute of Standards and Technology

Harman, D. K. (2013). TREC-Style Evaluations. In Agosti, M., Ferro, N., Forner, P., Müller, H., and Santucci, G., editors, *Information Retrieval Meets Information Visualization – PROMISE Winter School 2012, Revised Tutorial Lectures*, pages 97–115. Lecture Notes in Computer Science (LNCS) 7757, Springer, Heidelberg, Germany.

Voorhees, E. M. and Harman, D. K. (2001). Overview of TREC 2001. In Voorhees, E. M. and Harman, D. K., editors, *The Tenth Text REtrieval Conference (TREC 2001)*, pages 1–15. NIST, Special Publication 500-250, Washington, USA.

What Makes a Good Pool?

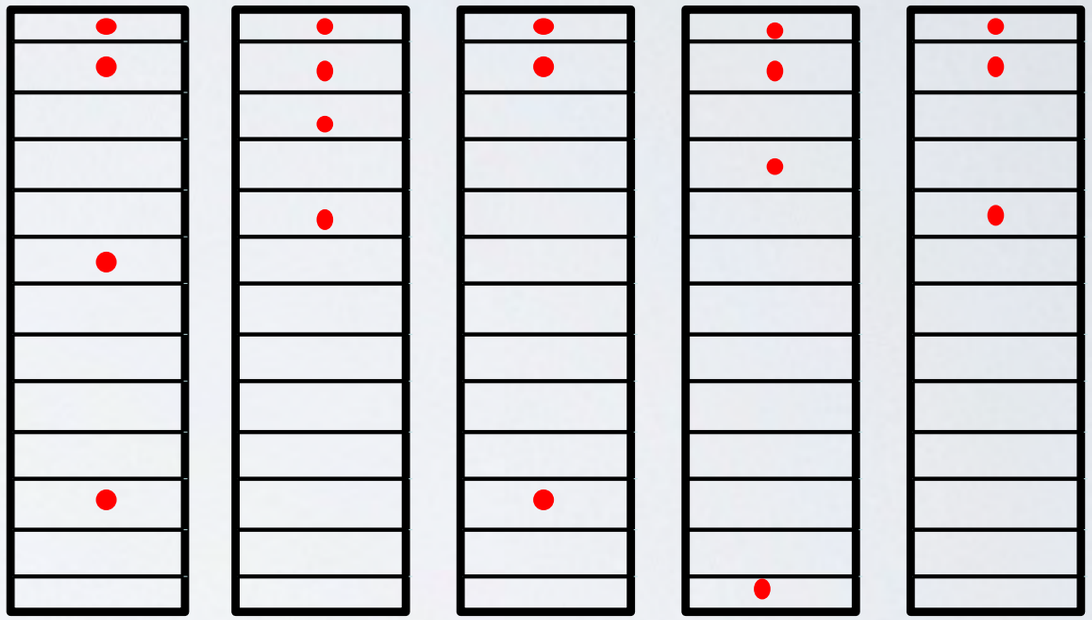
- Depth-*k* pools require large enough *k* and different enough pooled systems in order to produce "**complete**" judgements
  - Not pooled/not assessed documents are typically assumed to be not relevant
  - This is a motivation for organizing large-scale evaluation initiatives
- The objective is not to allow for computing the "exact" value of an evaluation measure but rather to **comparatively assess systems** and detect significant differences in a robust way
- **Leave-one-out tests**: are used to assess the **reusability** of a pool
  - one system/group of systems is removed from the pool
  - all the systems are evaluated using both the original pool and the newly created one
  - the two sets of results are compared by computing the Kendall's τ correlation among the ranking of systems on the original and the new pool and/or the maximum drop in ranking

Justin Zobel

Zobel, J. (1998). How Reliable are the Results of Large-Scale Information Retrieval Experiments. In Croft, W. B., Moffat, A., van Rijsbergen, C. J., Wilkinson, R., and Zobel, J., editors, *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998)*, pages 307–314. ACM Press, New York, USA.

Shallow Pools

- Internal test collections used by commercial search engines have large numbers of topics (ten of thousands), much more than existed in publicly available ones
- The TREC **Million Query** Track created a test collection with **1,755 topics** (using **stratified sampling techniques**)
- 250 topics with 20 judgments per topic are the most cost-effective in terms of minimizing assessor effort and maximizing accuracy in ranking runs
- There might be concerns about how much **reusable** are these test collection
- Need to develop **ad-hoc evaluation measures** aware of the sampling procedures



Carterette, B. A., Pavlu, V., Kanoulas, E., Aslam, J. A., and Allan, J. (2008). Evaluation over Thousands of Queries. In Chua, T.-S., Leong, M.-K., Oard, D. W., and Sebastiani, F., editors, *Proc. 31st Annual Inter- national ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, pages 651–658. ACM Press, New York, USA.

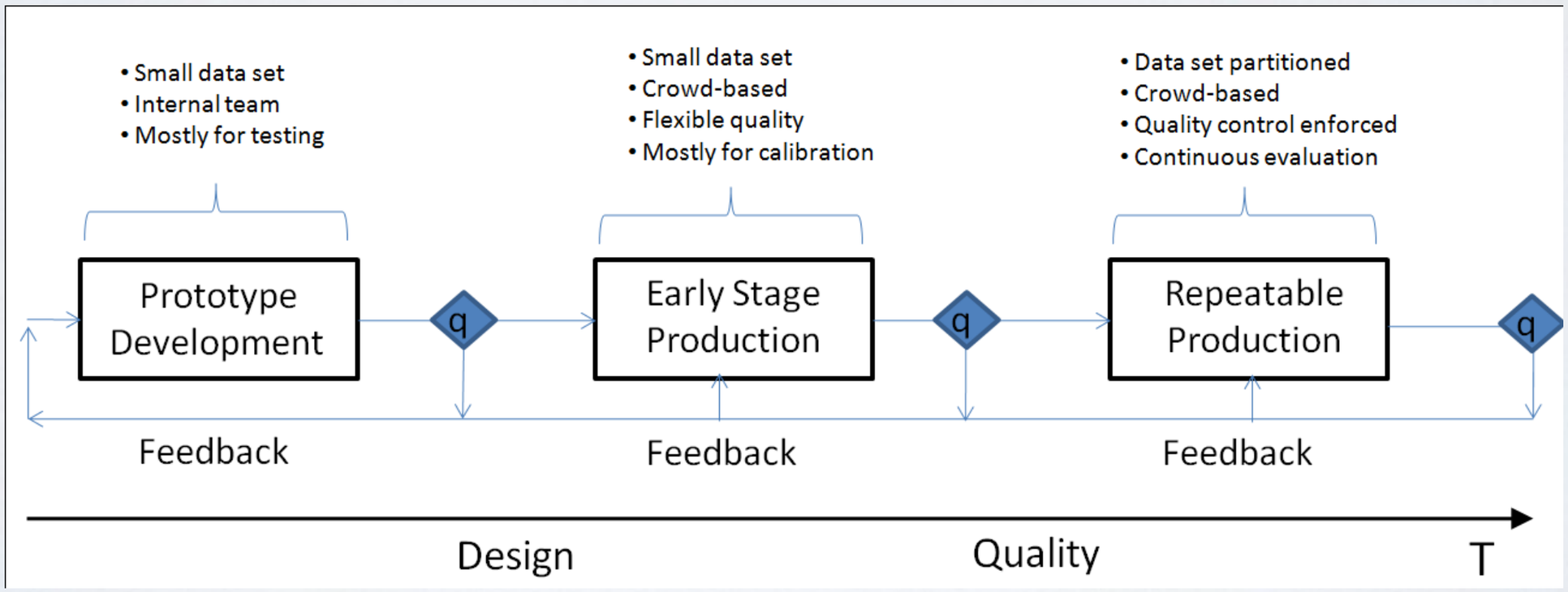
Multi-armed Bandits Pools

- Bandit techniques trade-off between exploiting known good “arms” and exploring to find better arms.
  For collection building, each run is an arm, and reward is finding a relevant doc
- Simulations suggest can get similar-quality collections as pooling but with many fewer judgments
- TREC 2017 Common Core track first attempt to build new collection using bandit technique

Losada, D. E., Parapar, J., and Barreiro, A. (2016). Feeling Lucky? Multi-armed Bandits for Ordering Judgements in Pooling-based Evaluation. In Ossowski, S., editor, *Proc. 2016 ACM Symposium on Applied Computing (SAC 2016)*, pages 1027–1034. ACM Press, New York, USA.
Voorhees, E. M. (2018). On Building Fair and Reusable Test Collections using Bandit Techniques. In Cuzzocrea, A., Allan, J., Paton, N. W., Sri- vastava, D., Agrawal, R., Broder, A., Zaki, M. J., Candan, S., Labrinidis, A., Schuster, A., and Wang, H., editors, *Proc. 27th International Conference on Information and Knowledge Management (CIKM 2018)*, pages 407–416. ACM Press, New York, USA.
Slide courtesy of Ellen M. Voorhees (see her seminar)

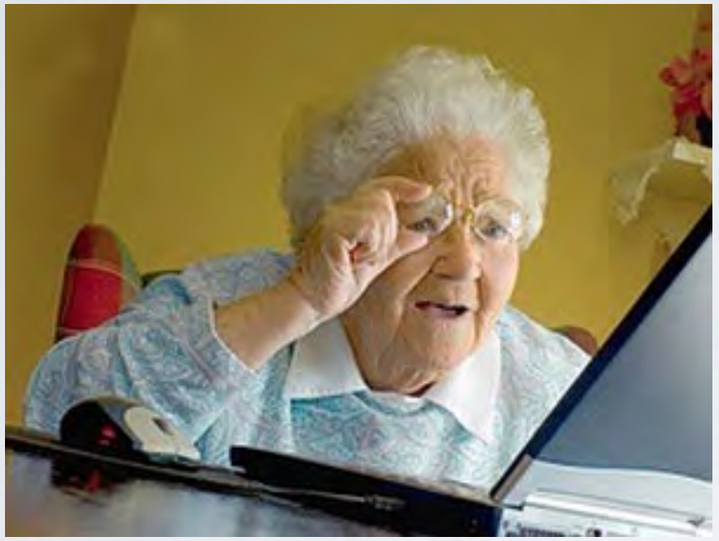
Crowdsourcing

- Large scale, inexpensive, diverse
- Careful design of the task, attention to details, simplicity and usability
- Need assessment of quality of work



Alonso, O. (2014). Evaluation with Respect to Usefulness – Some perspectives from industry. In Ferro, N., editor, *Bridging Between Information Retrieval and Databases - PROMISE Winter School 2013, Revised Tutorial Lectures*, pages 183–192. Lecture Notes in Computer Science (LNCS) 8173, Springer, Heidelberg, Germany.

Crowdsourcing



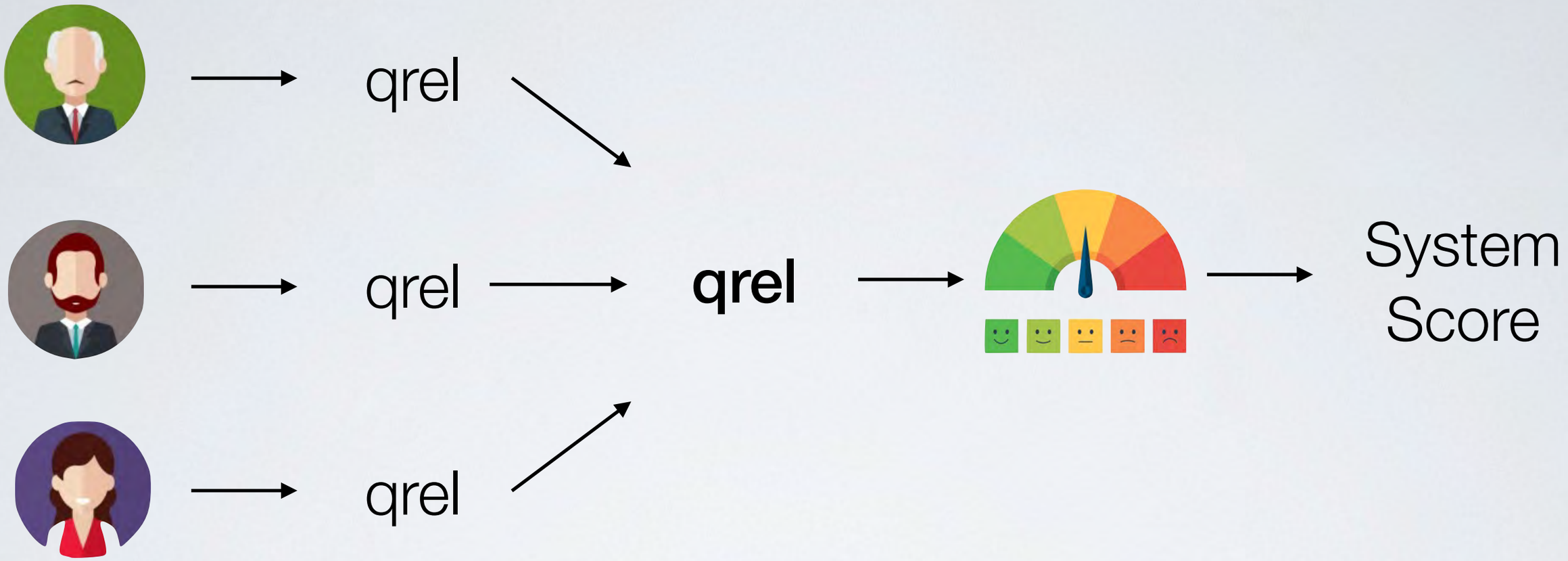
Crowd Assessors

CrowdFlower

amazon mechanical turk

Alonso, O. and Mizzaro, S. (2012). Using Crowdsourcing for TREC Relevance Assessment. *Information Processing & Management*, 48(6):1053– 1066.

Aggregating Relevance Assessments: Upstream Approach



- **Input**: a set of relevance assessments from each assessor
- **Output**: a single set of relevance assessments, from which an evaluation measure is computed

How To Aggregate Relevance Assessments?

Majority Vote (MV)

Expectation Maximization (EM)

Hosseini, M., Cox, I. J., Milić-Frayling, N., Kazai, G., and Vinay, V. (2012). On Aggregating Labels from Multiple Crowd Workers to Infer Relevance of Documents. In Baeza-Yaetes, R., de Vries, A. P., Zaragoza, H., Cambazoglu, B. B., Murdock, V., Lempel, R., and Silvestri, F., editors, Advances in Information Retrieval. *Proc. 32nd European Conference on IR Research (ECIR 2012)*, pages 182–194. Lecture Notes in Computer Science (LNCS) 7224, Springer, Heidelberg, Germany.

What Can Go Wrong in Downstream Approaches?

- Out of 10 relevant documents in a pool, just 1 document has been wrongly labelled as not relevant
  - thus there is a **10% error** with respect to the whole pool
- Run_1 represents the case where the mis-labelled document is retrieved in ranks 1 to 5, while the other runs show what could have happened if it had been correctly labelled
  - P@5, i.e. precision at 5 retrieved documents, passes from 0% to 20%, so a **100% error**
  - AP, i.e. average precision, passes from 7.65% to 14.07%-22.96%, so a **45.61%-66.67% error**

Rank	1	2	3	4	5	6	7	8	9	10	P@5	AP
Run_1	👻	👻	👻	👻	👻	R			R	R	0.0000	0.0765
Run_2					*R*	R			R	R	0.2000	0.1407
Run_3				*R*		R			R	R	0.2000	0.1463
Run_4			*R*			R			R	R	0.2000	0.1556
Run_5		*R*				R			R	R	0.2000	0.1741
Run_6	*R*					R			R	R	0.2000	0.2296

What Can Go Wrong in Downstream Approaches?

- Out of 10 relevant documents in a pool, just 1 document has been wrongly labelled as not relevant
  - thus there is a **10% error** with respect to the whole pool
- Run_1 represents the case where the mis-labelled document is retrieved in ranks 1 to 5, while the other runs show what could have happened if it had been correctly labelled
  - P@5, i.e. precision at 5 retrieved documents, passes from 0% to 20%, so a **100% error**
  - AP, i.e. average precision, passes from 7.65% to 14.07%-22.96%, so a **45.61%-66.67% error**

Mis-labelled documents produce different errors for different runs and measures and these might be greater than the error on the pool itself

Rank	1	2	3	4	5	6	7	8	9	10	P@5	AP
Run_1	👻								R	R	0.0000	0.0765
Run_2					R	R			R	R	0.2000	0.1407
Run_3				R					R	R	0.2000	0.1463
Run_4			R						R	R	0.2000	0.1556
Run_5		R						R	R	R	0.2000	0.1741
Run_6	R							R	R	R	0.2000	0.2296

Aggregating Relevance Assessments: Downstream Approach

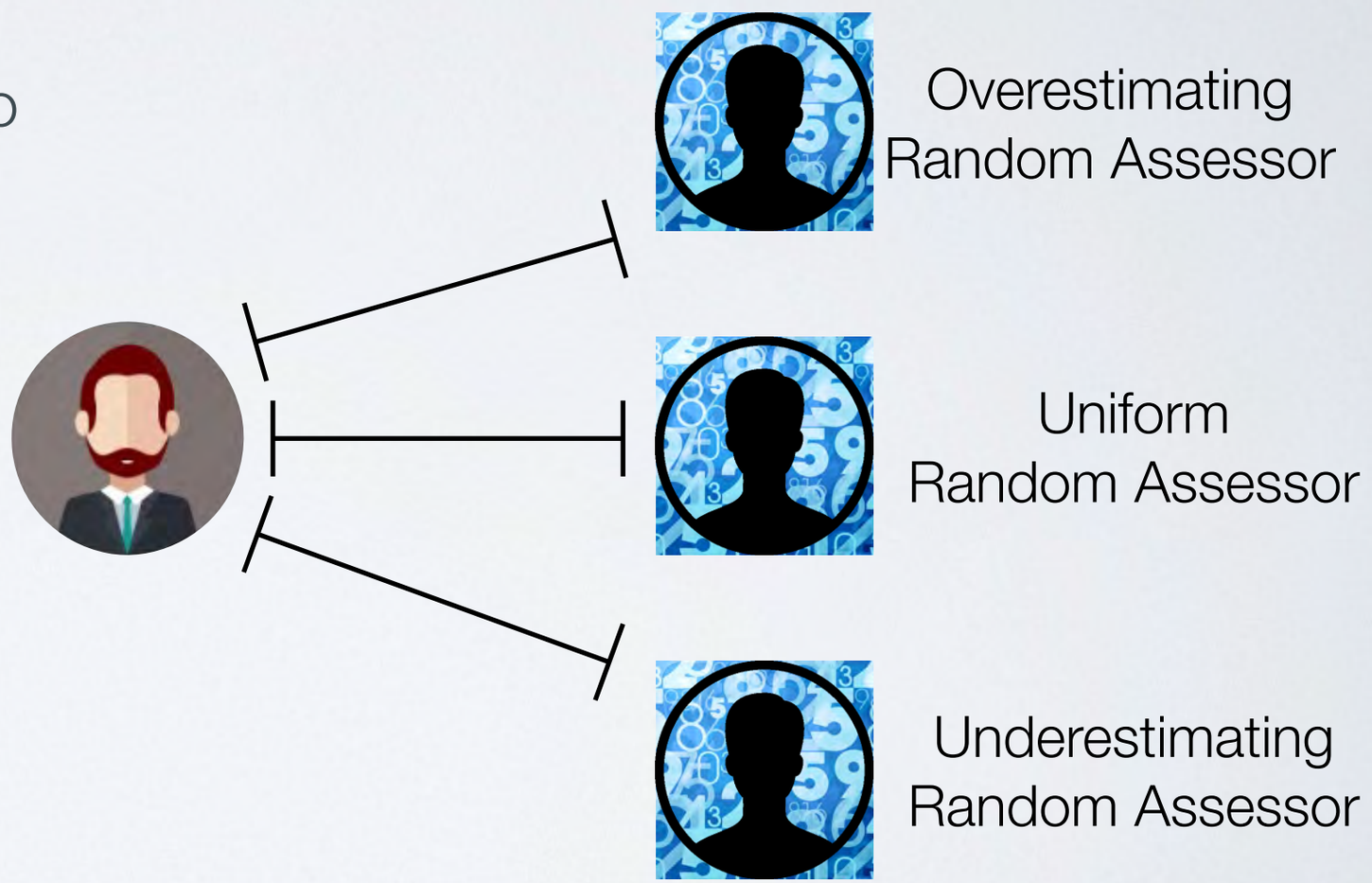
- **Input**: a set of evaluation measures computed according to the relevance judgements of each accessor
- **Output**: an aggregated evaluation measure

Ferrante, M., Ferro, N., and Maistro, M. (2017). AWARE: Exploiting Evaluation Measures to Combine Multiple Assessors. *ACM Transactions on Information Systems (TOIS)*, 36(2):20:1–20:38.

The AWARE Framework

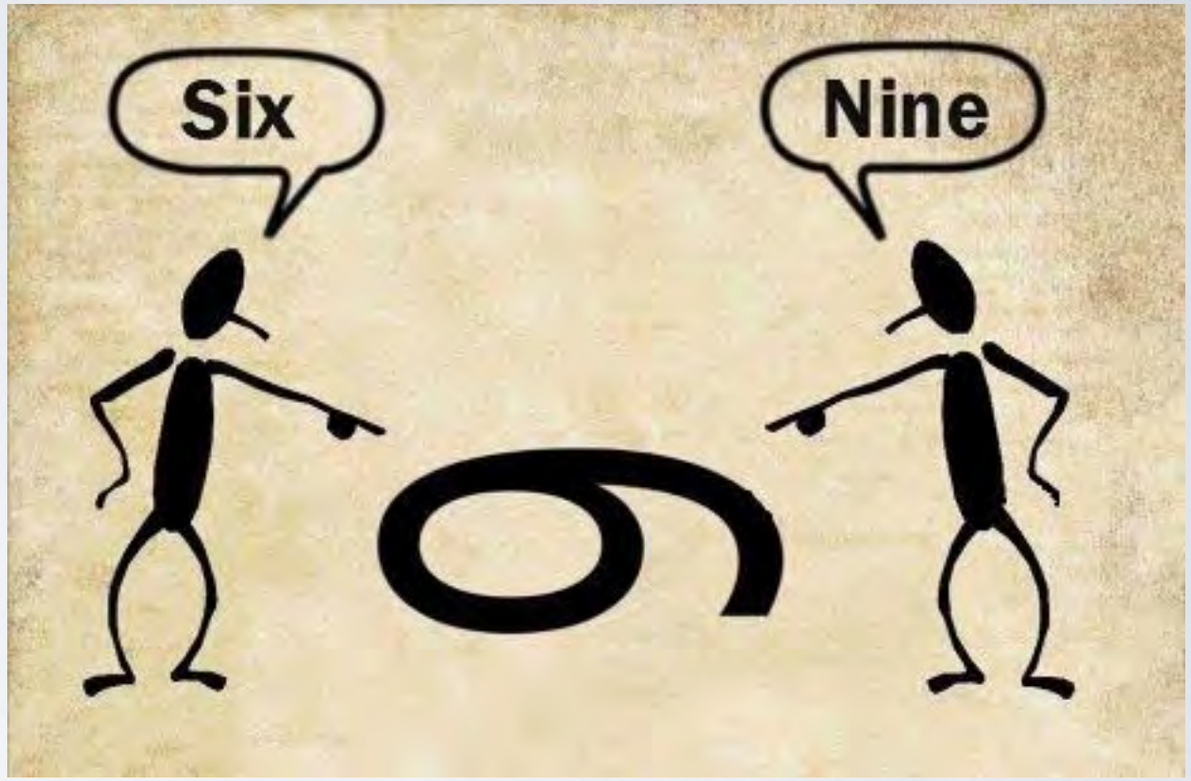
$$\mathbb{E}\Big[\mu(\hat{r}_t)\Big] = \sum_{k=1}^{l} \mathbb{E}\Big[\mu(\hat{r}_t)\Big|W = W_k\Big]\ a_k(t)$$

- We compute the final score of a system as a weighted average of the scores computed according to the relevance judgements of each crowd-assessor
- The challenge is how to estimate the accuracy $a_k(t)$ to be assigned to each assessor
- Unsupervised estimator: the accuracy is proportional to the "distance" from prototypical random assessors



Randomness in Assessment?

- Relevance judgments is an intrinsically not deterministic process
  - subjectivity of the notion of relevance
  - inter-assessor agreement and crowd-sourcing
  - variation in assessor own notion of relevance
- Nevertheless, when we use relevance judgements in evaluation measures, we treat them as absolutely exact



Relevance as a Binomial Random Variable

- For each (topic, document) pair, let model relevance as a Binomial Random Variable $X \sim B(1, p)$ which assumes the value 1 with probability p and the value 0 with probability $1 - p$
  - Since $\mathbb{E}[X] = p$, p roughly represents the amount of relevance of a document
- As a consequence, runs become sequences of random variables and evaluation measures become transformations of random variables
- Applications
  - removing the distinction between binary and multi-graded relevance
  - robustness to incomplete information (remember that not assessed documents are assumed to be not relevant)
  - robustness to inter-assessor agreement
  - merging of crowd-assessors
  - …

Ferrante, M., Ferro, N., and Pontarollo, S. (2018). Modelling Randomness in Relevance Judgments and Evaluation Measures. In Pasi, G., Piwowarski, B., Azzopardi, L., and Hanbury, A., editors, Advances in Information Retrieval. *Proc. 40th European Conference on IR Research (ECIR 2018)*, pages 197–209. Lecture Notes in Computer Science (LNCS) 10772, Springer, Heidelberg, Germany.

Ferrante, M., Ferro, N., and Losiouk, E. (2019). Stochastic Relevance for Crowdsourcing. In Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., and Hiemstra, D., editors, *Advances in Information Retrieval. Proc. 41st European Conference on IR Research (ECIR 2019)* – Part I, pages 755–762. Lecture Notes in Computer Science (LNCS) 11437, Springer, Heidelberg, Germany.

Questions?


. . . AND OF COURSE WE'LL ASSESS OUR PROGRESS ALONG THE WAY.
PROGRESS
DilbertCartoonist@gmail.com
WILL YOU BE USING AN ENHANCED ASSESSMENT METHODOLOGY?
I HOPE THAT MEANS SOMETHING. ALL I DID WAS STRING TOGETHER SOME WORDS I HEARD IN THE HALLWAY.
© 2010 Scott Adams, Inc. /Dist. by UFS, Inc.
UM. . . I'LL BE ASSESSING. . . BY MEASURING. . . AND UM. . .
I'D BETTER GET IN ON THIS.
I CAN'T SUPPORT THIS PROJECT UNTIL I SEE YOUR ADVANCED ASSESSMENT METHOD-OLOGY PLAN.
I'LL HAVE IT IN TEN MINUTES, ASSUMING YOU DON'T KNOW WHAT IT'S SUPPOSED TO LOOK LIKE.
VERY GOOD.
www.dilbert.com
I'LL BE IN THE SHOWER TRYING TO WASH MY SOUL.

Evaluation Measures

Evaluation Measures

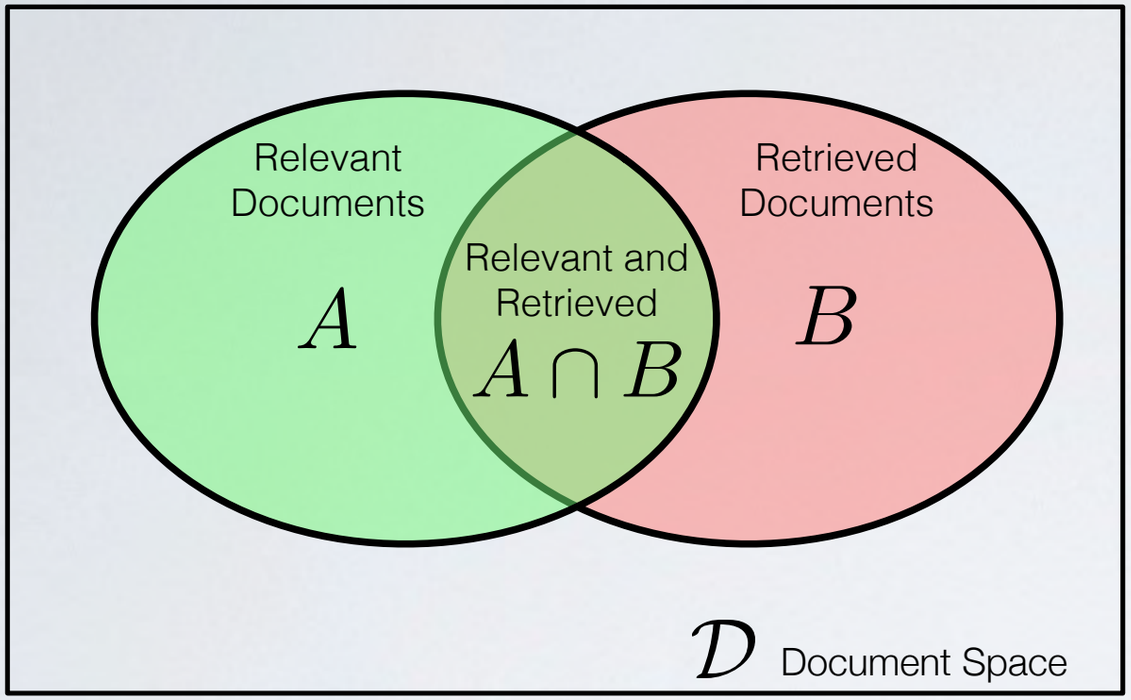
"Measure what is measurable and make measurable what is not"

Galileo Galilei (1564-1642)

A Taxonomy of Evaluation Measures

	Set-Based Retrieval	Rank-Based Retrieval
Binary Relevance	Precision (P) Recall (R) F-measure (F)	Precision at Document Cut-off (P@k) Recall at Document Cut-off (R@k) R-Precision (Rprec) Average Precision (AP) Rank-Biased Precision (RBP) …
Multi-graded Relevance	Not widely agreed generalizations of Precision and Recall	Discounted Cumulated Gain (DCG) …

Set-based Measures: Precision, Recall and F-measure



$$P = \frac{|A \cap B|}{|B|} \qquad R = \frac{|A \cap B|}{|A|}$$

$$F = \frac{2}{\frac{1}{P} + \frac{1}{R}} = 2\frac{P \cdot R}{P + R}$$

- **Precision** is the the proportion of retrieved documents that are actually relevant
- **Recall** is the the proportion of relevant documents actually retrieved
- Together, Precision and Recall measure **retrieval effectiveness**, meant as the ability of a system to retrieve relevant documents while at the same time holding back non-relevant ones
  - maximizing Precision and Recall corresponds to optimal retrieval in the sense of the **Probability Ranking Principle**, i.e. ordering documents by their decreasing probability of being relevant, and creates a tight link between retrieval models and evaluation
- **F-measure** is the harmonic mean of Precision and Recall, summarising them into a single score

van Rijsbergen, C. J. (1974). Foundations of Evaluation. *Journal of Documentation*, 30(4):365–373.
van Rijsbergen, C. J. (1981). Retrieval effectiveness. In Spärck Jones, K., editor, *Information Retrieval Experiment*, pages 32–43. Butterworths, London, United Kingdom.

Set-based Measures: Example

$$P = \frac{4}{10} = 0.40$$

$$R = \frac{4}{8} = 0.50$$

$$F = 2 \cdot \frac{\frac{4}{10} \cdot \frac{4}{8}}{\frac{4}{10} + \frac{4}{8}} = \frac{4}{9} = 0.44$$

Assume

- $|A| = 8$ relevant documents in total
- Lenient mapping to binary relevance degrees

Rank-based Measures: Precision and Recall

- **Precision at Document Cut-off**:

$$P(k) = \frac{1}{k} \sum_{n=1}^{k} r_n$$

where $r_k \in \{0, 1\}$ is the relevance degree of the n-th document

- **Recall at Document Cut-off**:

$$R(k) = \frac{1}{RB} \sum_{n=1}^{k} r_n$$

where $RB = |A|$ is the **recall base**, i.e. the total number of relevant documents

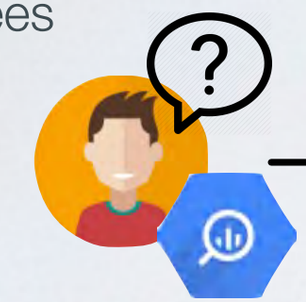
- **Rprec** is Precision computed at the recall base

$$Rprec = P(RB)$$

Rank-based Measures: Example of Precision and Recall

Assume

- $RB = 8$ relevant documents in total
- Lenient mapping to binary relevance degrees



Topic | **Run** | **Assessed Run** | **Binary Weighted Assessed Run**

	Assessed Run	Binary Weighted Assessed Run
1	Highly Relevant	1
2	Not Relevant	0
3	Partially Relevant	1
4	Fairly Relevant	1
5	Not Relevant	0
6	Not Relevant	0
7	Not Relevant	0
8	Fairly Relevant	1
9	Not Relevant	0
10	Not Relevant	0

$$P(5) = \frac{3}{5} = 0.600$$

$$R(5) = \frac{3}{8} = 0.375$$

$$Rprec = P(8) = \frac{4}{8} = 0.500$$

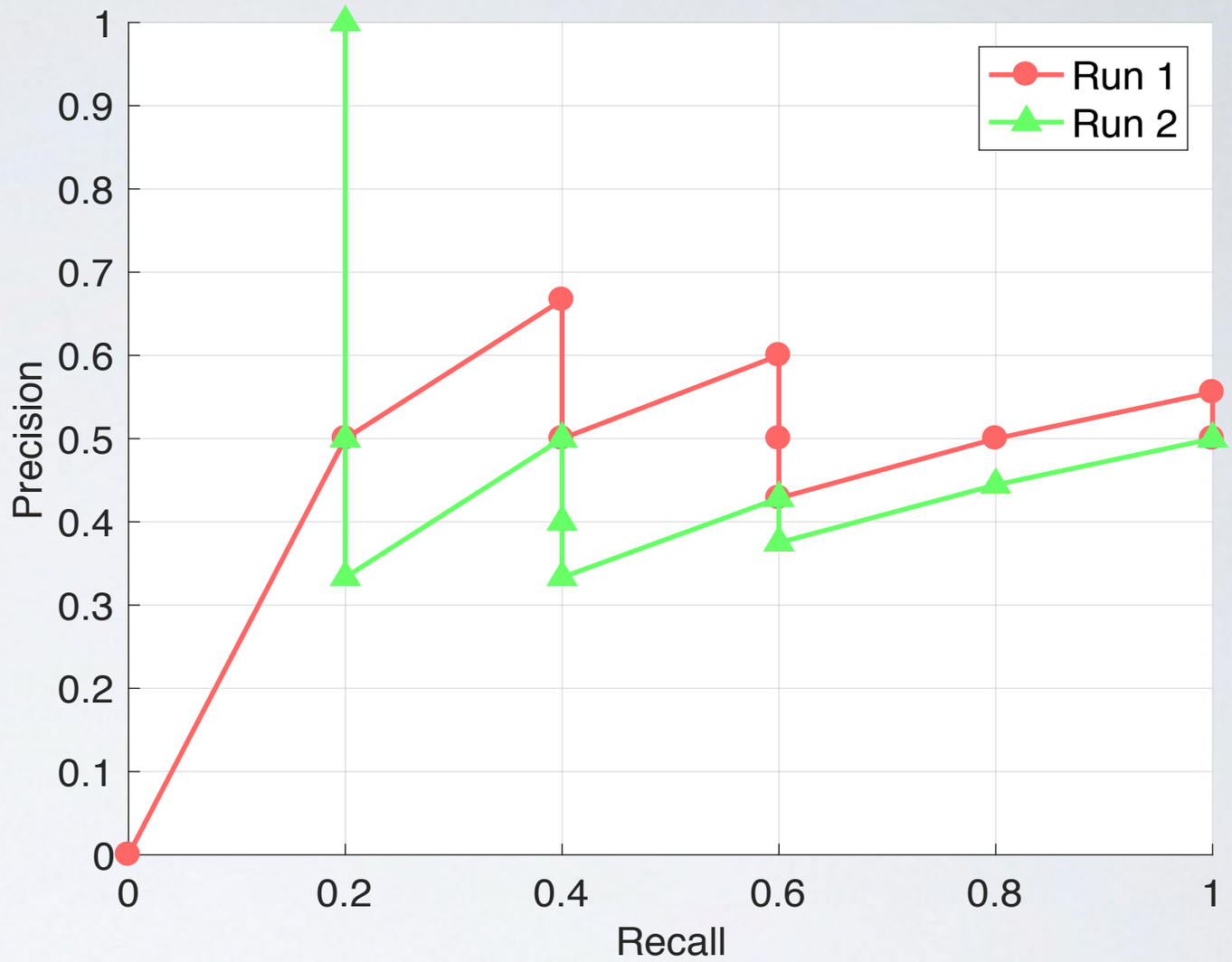
Precision-Recall Curve

Run1

		PS=$0.00 RS=$0.00
1	0	
2	1	PS=$0.50 RS=$0.20
3	1	PS=$0.66 RS=$0.40
4	0	PS=$0.50 RS=$0.40
5	1	PS=$0.60 RS=$0.60
6	0	PS=$0.50 RS=$0.60
7	0	PS=$0.42 RS=$0.60
8	1	PS=$0.50 RS=$0.80
9	1	PS=$0.55 RS=$1.00
10	0	PS=$0.50 RS=$1.00

Run2

		PS=$1.00 RS=$0.20
1	1	
2	0	PS=$0.50 RS=$0.20
3	0	PS=$0.33 RS=$0.20
4	1	PS=$0.50 RS=$0.40
5	0	PS=$0.40 RS=$0.40
6	0	PS=$0.33 RS=$0.40
7	1	PS=$0.42 RS=$0.60
8	0	PS=$0.37 RS=$0.60
9	1	PS=$0.44 RS=$0.80
10	1	PS=$0.50 RS=$1.00



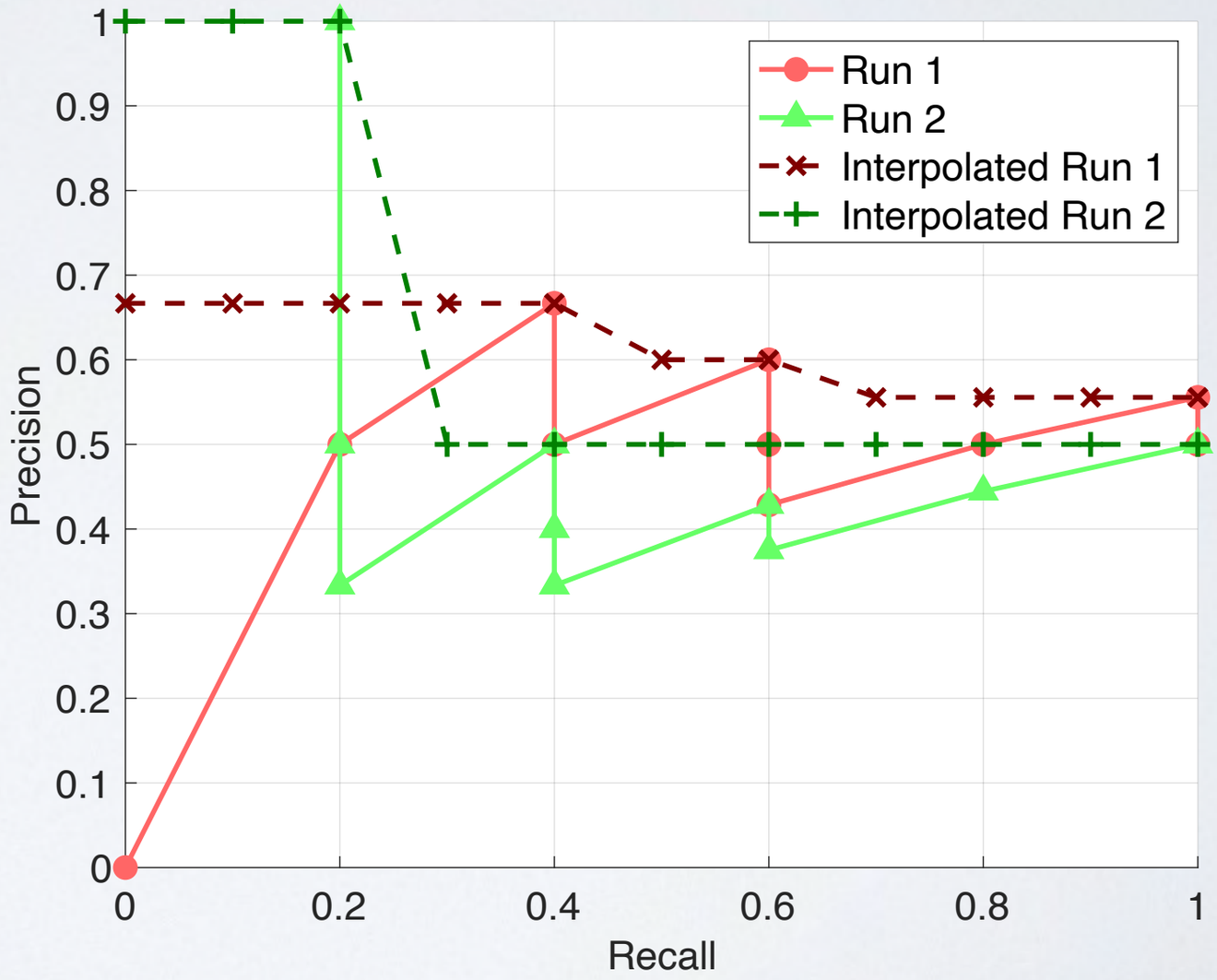
- Assume $RB = 5$ relevant documents in total
- The Precision-Recall curve has a typical saw-tooth shape
  - We may have multiple Precision values for the same Recall value
  - It is difficult to compare runs because they may not have the same Recall values

Interpolated Precision-Recall Curve

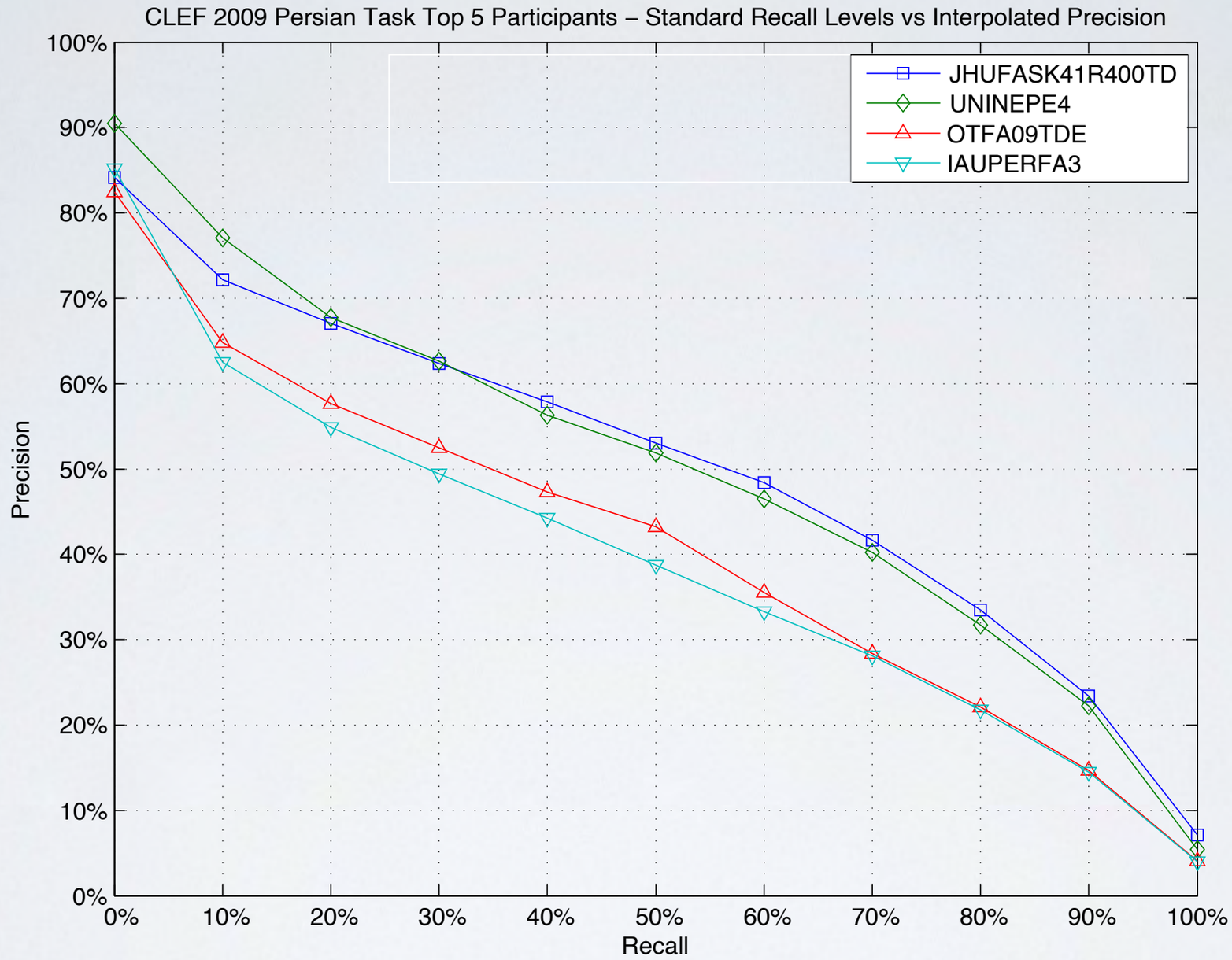
	Run1		Run2	
1	0	**iP = 0.66** PS=$0.00 RS=$0.00	1	**iP = 1.00** PS=$1.00 RS=$0.20
2	1	**iP = 0.66** PS=$0.50 RS=$0.20	0	**iP = 1.00** PS=$0.50 RS=$0.20
3	1	**iP = 0.66** PS=$0.66 RS=$0.40	0	**iP = 1.00** PS=$0.33 RS=$0.20
4	0	**iP = 0.66** PS=$0.50 RS=$0.40	1	**iP = 0.50** PS=$0.50 RS=$0.40
5	1	**iP = 0.60** PS=$0.60 RS=$0.60	0	**iP = 0.50** PS=$0.40 RS=$0.40
6	0	**iP = 0.60** PS=$0.50 RS=$0.60	0	**iP = 0.50** PS=$0.33 RS=$0.40
7	0	**iP = 0.60** PS=$0.42 RS=$0.60	1	**iP = 0.50** PS=$0.42 RS=$0.60
8	1	**iP = 0.55** PS=$0.50 RS=$0.80	0	**iP = 0.50** PS=$0.37 RS=$0.60
9	1	**iP = 0.55** PS=$0.55 RS=$1.00	1	**iP = 0.50** PS=$0.44 RS=$0.80
10	0	**iP = 0.55** PS=$0.50 RS=$1.00	1	**iP = 0.50** PS=$0.50 RS=$1.00

- To interpolate Precision at standard Recall value R_j we use the maximum Precision obtained for any actual Recall value R greater than or equal to R_j

$$iP@R_j = \max_{R \geq R_j} P@R$$



11 points Interpolated Precision-Recall Curve



- Standard Interpolated Precision-Recall curves exhibit a typical inverse relationship among Precision and Recall, indicating a trade-off between these two goals of effectiveness

Cleverdon, C. W. (1972). On the inverse relationship of recall and precision. *Journal of Documentation*, 28(3):195–201.

Buckland, M. and Gey, F. (1994). The relationship between Recall and Precision. *Journal of the American Society for Information Science and Technology (JASIST)*, 45(1):12–19.

Egghe, L. (2008). The measures precision, recall, fallout and miss as a function of the number of retrieved documents and their mutual interrelations. *Information Processing & Management*, 44(2):856–876.

Rank-based Measures: Average Precision

$$AP = \frac{1}{RB} \sum_{k \in \mathcal{R}} P(k) = \frac{1}{RB} \sum_{n=1}^{N} \underbrace{\left(\frac{1}{n} \sum_{m=1}^{n} r_m \right)}_{P(n)} r_n =$$

$$= \underbrace{\frac{rr}{RB}}_{\text{Recall}} \cdot \underbrace{\frac{1}{rr} \sum_{k \in \mathcal{R}} P(k)}_{\text{arithmetic mean of } P(k)}$$

where

- $\mathcal{R}$ is the set of the rank positions of the relevant retrieved documents
- $rr = |\mathcal{R}|$ is the total number of relevant retrieved documents
- N is the total number of retrieved documents, i.e. the length of the run
- The **Mean Average Precision (MAP)** is the mean of AP over a set of topics
  - Differently from the other measures, this mean has its own name since it is the most widely used single number to summarise the whole performance of a system

Chris Buckley

Buckley, C. and Voorhees, E. M. (2005). Retrieval System Evaluation. In Harman, D. K. and Voorhees, E. M., editors, *TREC. Experiment and Evaluation in Information Retrieval*, pages 53–78. MIT Press, Cambridge (MA), USA.

Rank-based Measures: Example of Average Precision

Assume

- $RB = 8$ relevant documents in total
- Lenient mapping to binary relevance degrees

Topic

Run

Assessed Run

1	Highly Relevant
2	Not Relevant
3	Partially Relevant
4	Fairly Relevant
5	Not Relevant
6	Not Relevant
7	Not Relevant
8	Fairly Relevant
9	Not Relevant
10	Not Relevant

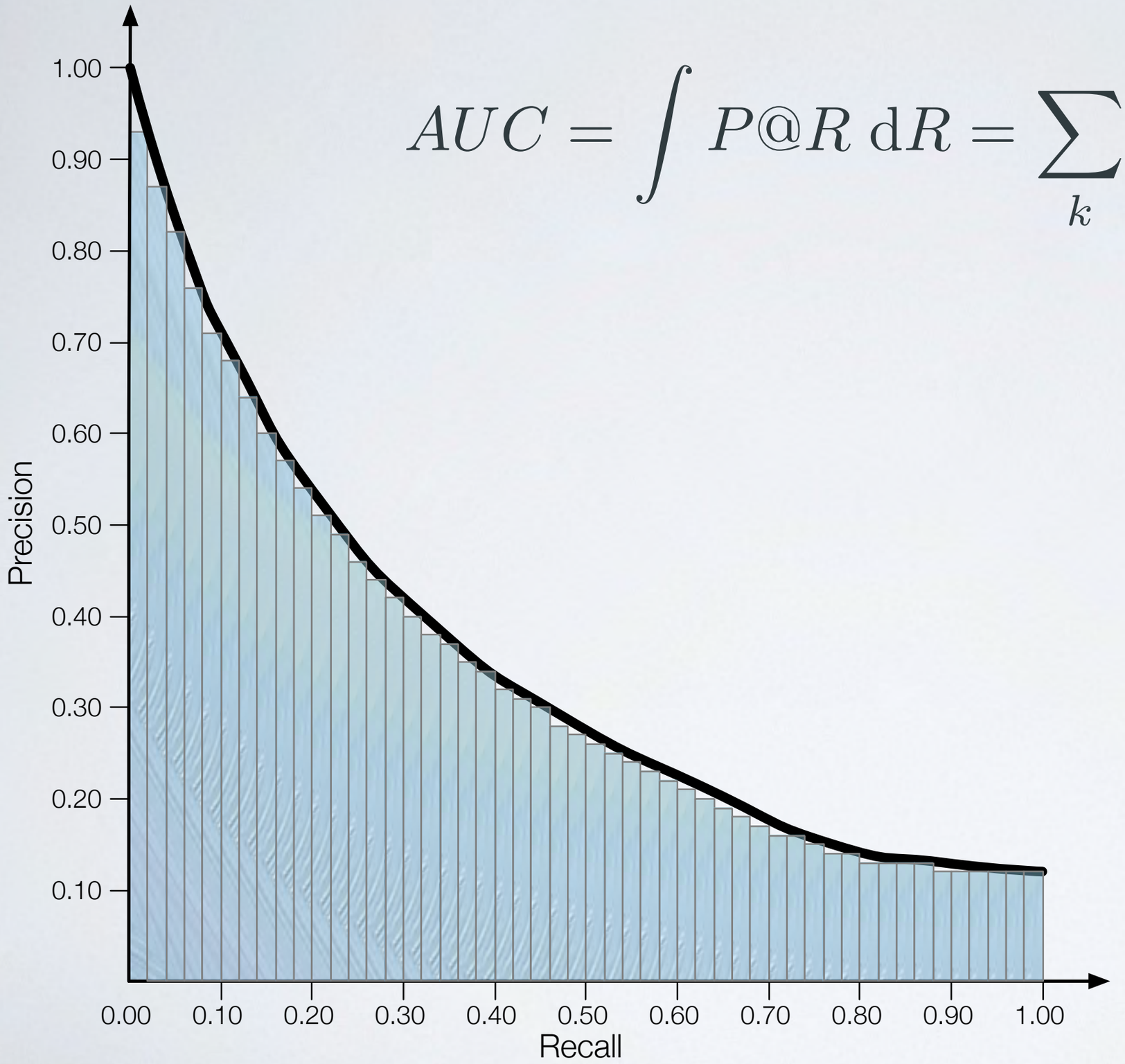
Binary Weighted Assessed Run

1	1
2	0
3	1
4	1
5	0
6	0
7	0
8	1
9	0
10	0

$$AP = \frac{1}{RB}\Big(P(1) + P(3) + P(4) + P(8)\Big)$$

$$= \frac{1}{8}\left(1 + \frac{2}{3} + \frac{3}{4} + \frac{4}{8}\right) = \frac{35}{96} = 0.36$$

Area under the Precision-Recall Curve



$$AUC = \int P@R \, \mathrm{d}R = \sum_k P(k)\Delta R(k)$$

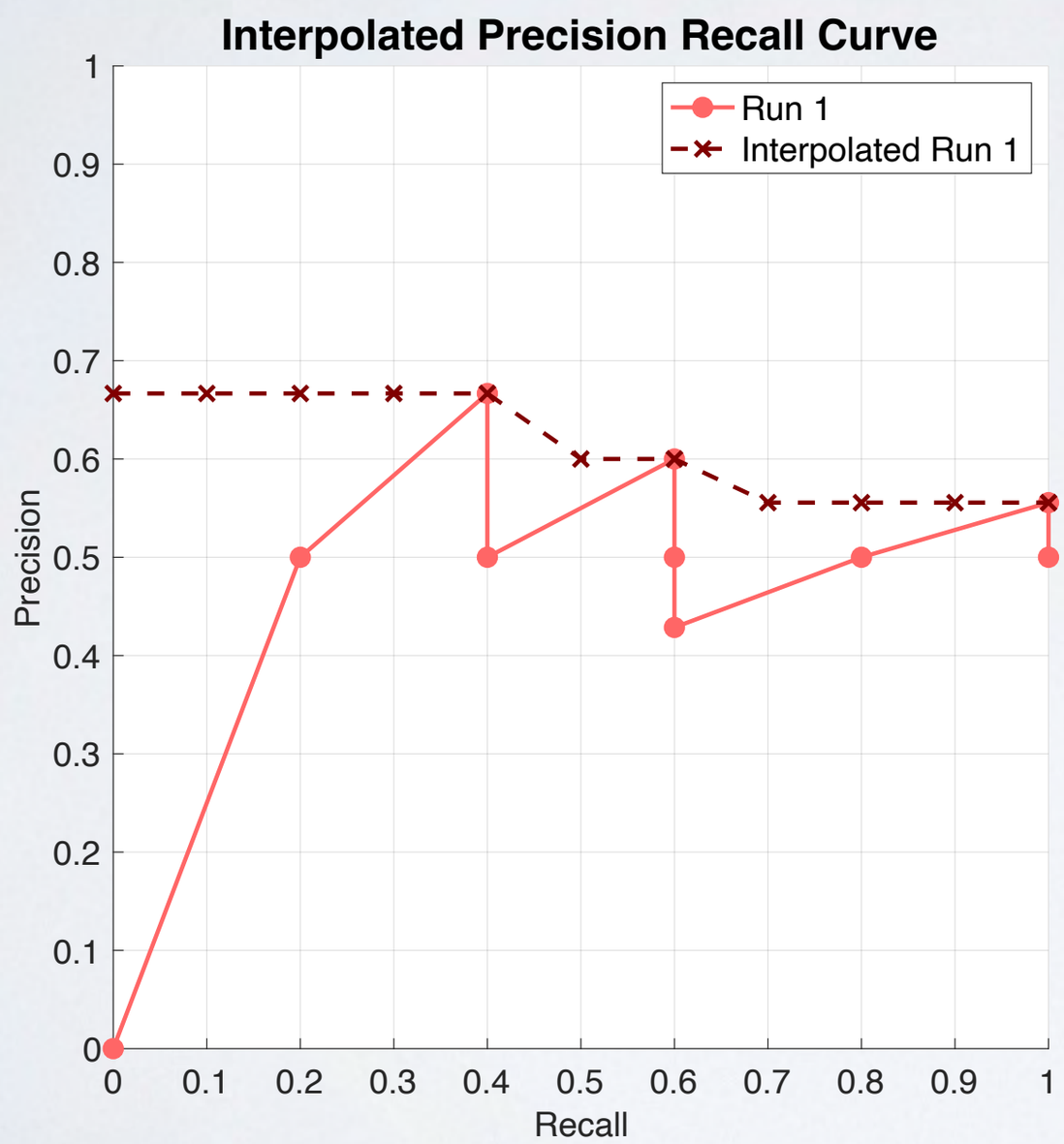
- The **Area Under the** Precision-Recall **Curve (AUC)** is an important indicator of the overall system effectiveness, summarising the trade-off between Precision and Recall

Computing the Area under the Precision-Recall Curve

$$AUC = \sum_{n=1}^{N} P(n)\Big(R(n) - R(n-1)\Big) \quad \text{assuming } R(0) = 0$$

Run1

1	0	PS=$0.00 RS=$0.00
2	1	PS=$0.50 RS=$0.20
3	1	PS=$0.66 RS=$0.40
4	0	PS=$0.50 RS=$0.40
5	1	PS=$0.60 RS=$0.60
6	0	PS=$0.50 RS=$0.60
7	0	PS=$0.42 RS=$0.60
8	1	PS=$0.50 RS=$0.80
9	1	PS=$0.55 RS=$1.00
10	0	PS=$0.50 RS=$1.00



$$\begin{aligned} AUC = \quad & 0.00(0.00 - 0.00) + 0.50(0.20 - 0.00) + \\ & 0.66(0.40 - 0.20) + 0.50(0.40 - 0.40) + \\ & 0.60(0.60 - 0.40) + 0.50(0.60 - 0.60) + \\ & 0.42(0.60 - 0.60) + 0.50(0.80 - 0.60) + \\ & 0.55(1.00 - 0.80) + 0.50(1.00 - 1.00) = \\ = \quad & 0.5620 \end{aligned}$$

Area under the Precision-Recall Curve and Average Precision

$$AUC = \sum_{n=1}^{N} P(n)\Big(R(n) - R(n-1)\Big)$$

- When the n-th document is not relevant, $R(n)$ is equal to $R(n-1)$ and their difference goes to zero
- Therefore, we can sum only on $\mathcal{R}$, i.e. the set of the rank positions of the relevant retrieved documents

$$AUC = \sum_{k \in \mathcal{R}} P(k)\Big(R(k) - R(k-1)\Big)$$

Area under the Precision-Recall Curve and Average Precision

$$AUC = \sum_{k \in \mathcal{R}} P(k) \Big(R(k) - R(k-1) \Big)$$

- Two adjacent rank positions differ just for one relevant document and thus

$$R(k) - R(k-1) = \underbrace{\frac{1}{RB} \sum_{n=1}^{k} r_n}_{R(k)} - \underbrace{\frac{1}{RB} \sum_{n=1}^{k-1} r_n}_{R(k-1)} = \frac{r_k}{RB} = \frac{1}{RB}$$

- Therefore AUC is equal to AP
  - this is one motivation of the importance of AP

$$AUC = \frac{1}{RB} \sum_{k \in \mathcal{R}} P(k) = AP$$

Rank-based Measures: Discounted Cumulated Gain

$$DCG(k) = \begin{cases} \sum_{n=1}^{k} r_k & \text{if } k < b \\ DCG(k-1) + \frac{r_k}{\log_b(k)} & \text{if } k >= b \end{cases} =$$

$$= \sum_{n=1}^{k} \frac{r_k}{\max(1, \log_b(k))}$$

- where the base of the logarithm b indicates the patience of the user in scanning the result list
  - $b = 2$ is an impatient user
  - $b = 10$ is a patient user
- DCG naturally handles multi-graded relevance
- DCG does not depend on the recall base
- DCG is not bounded in [0, 1]

Kalervo Järvelin

Jaana Kekäläinen

Järvelin, K. and Kekäläinen, J. (2002). Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446

Rank-based Measures: Discounted Cumulated Gain

$$DCG(k) = \begin{cases} \sum_{n=1}^{k} r_k & \text{if } k < b \\ DCG(k-1) + \frac{r_k}{\log_b(k)} & \text{if } k >= b \end{cases} =$$

$$= \sum_{n=1}^{k} \frac{r_k}{\max(1, \log_b(k))}$$

- where the base of the logarithm b indicates the patience of the user in scanning the result list
  - $b = 2$ is an impatient user
  - $b = 10$ is a patient user
- DCG naturally handles multi-graded relevance
- DCG does not depend on the recall base
- DCG is not bounded in [0, 1]

Kalervo Järvelin

Jaana Kekäläinen

Järvelin, K. and Kekäläinen, J. (2002). Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446

Rank-based Measures: Discounted Cumulated Gain

$$DCG(k) = \begin{cases} \sum_{n=1}^{k} r_k & \text{if } k < b \\ DCG(k-1) + \frac{r_k}{\log_b(k)} & \text{if } k >= b \end{cases} =$$

$$= \sum_{n=1}^{k} \frac{r_k}{\max(1, \log_b(k))}$$

- where the base of the logarithm b indicates the patience of the user in scanning the result list
  - $b = 2$ is an impatient user
  - $b = 10$ is a patient user
- DCG naturally handles multi-graded relevance
- DCG does not depend on the recall base
- DCG is not bounded in [0, 1]

Kalervo Järvelin

Jaana Kekäläinen

Järvelin, K. and Kekäläinen, J. (2002). Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446

Rank-based Measures: User Models

- Rank-based evaluation measures, implicitly or explicitly, embed a **user model** comprising
  - a **browsing model** that describes how a user interacts with results;
  - a **model of document utility**, describing how a user derives utility from individual relevant documents;
  - a **utility accumulation model** that describes how a user accumulates utility in the course of browsing.
- User models may be more or less **artificial** and may be more or less **correlated** with **actual user behaviour** and preferences
- In the case of DCG
  - a **browsing model**: user steps down the ranked results one-by-one, until s/he reaches the stopping rank k which is picked with a probability proportional to the log of the rank
  - a **model of document utility**: user gains something from each relevant document, proportional to its relevance degree
  - a **utility accumulation model**: user gains from all of the relevant documents from ranks 1 through k

Ben Carterette

Carterette, B. A. (2011). System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation. In Ma, W.-Y., Nie, J.-Y., Baeza-Yaetes, R., Chua, T.-S., and Croft, W. B., editors, *Proc. 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011)*, pages 903–912. ACM Press, New York, USA.

Rank-based Measures: Example of Discounted Cumulated Gain

Assume

- $RB = 8$ relevant documents in total
- An impatient user

	Topic	Run	Assessed Run	Weighted Assessed Run
1			Highly Relevant	3
2			Not Relevant	0
3			Partially Relevant	1
4			Fairly Relevant	2
5			Not Relevant	0
6			Not Relevant	0
7			Not Relevant	0
8			Fairly Relevant	2
9			Not Relevant	0
10			Not Relevant	0

$$DCG = 3 + \frac{1}{\log_2(3)} + \frac{2}{\log_2(4)} + \frac{2}{\log_2(8)} =$$

$$= 5.2976$$

Rank-based Measures: Normalized Discounted Cumulated Gain

- To normalize DCG in [0, 1], you need to compute the ideal run, i.e. the pool sorted in descending order of relevance, which represents the best retrieval possible and the maximum value of DCG

$$nDCG(k) = \frac{DCG(k)}{iDCG(k)}$$

- nDCG is given by the DCG of the run divided by the DCG of the ideal run

Rank-based Measures: Example of Normalized Discounted Cumulated Gain

Assume

- $RB = 8$ relevant documents in total
- An impatient user

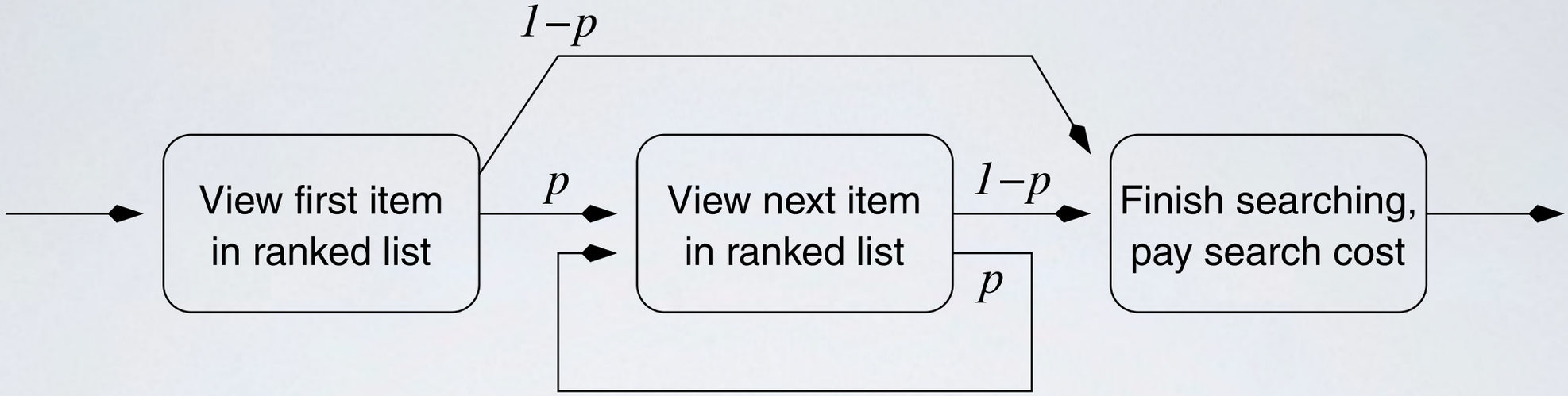
	Assessed Run	Weighted Assessed Run	Weighted Assessed Ideal Run
1	Highly Relevant	3	3
2	Not Relevant	0	3
3	Partially Relevant	1	2
4	Fairly Relevant	2	2
5	Not Relevant	0	2
6	Not Relevant	0	1
7	Not Relevant	0	1
8	Fairly Relevant	2	1
9	Not Relevant	0	0
10	Not Relevant	0	0

$$DCG = 5.2976$$

$$iDCG = 10.1996$$

$$nDCG = 0.5194$$

Rank-based Measures: Rank-Biased Precision



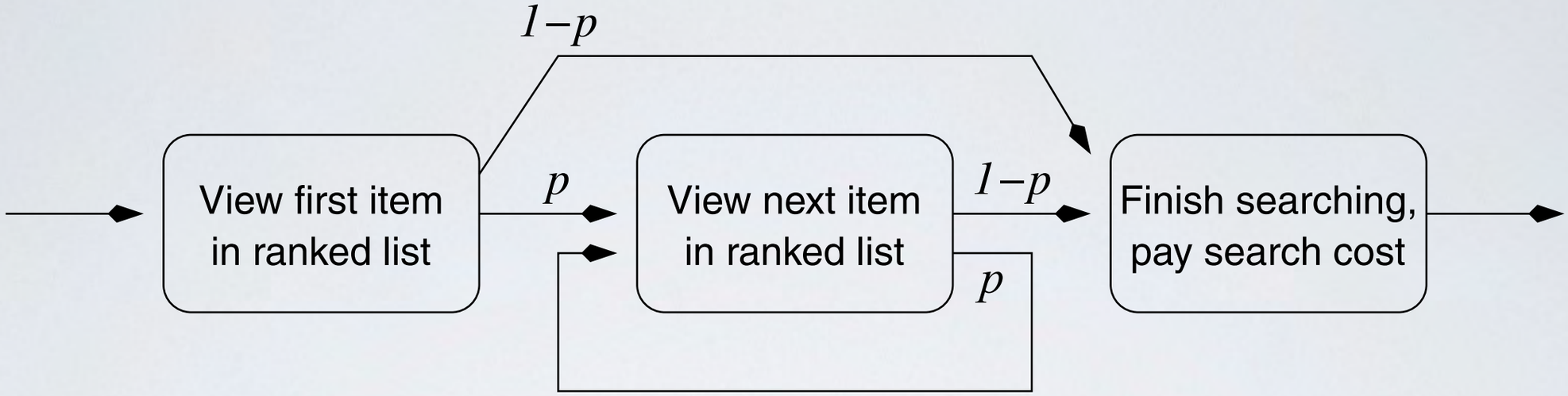
- The user starts from the top ranked document and with probability p, called **persistence**, goes to the next document or with probability $1 - p$ stops
  - typical value for p are: 0.5 for impatient users, 0.8 for patient users, and 0.95 for extremely patient users

$$RBP = (1-p)\sum_{n=1}^{N} p^{n-1} r_n = (1-p)\sum_{k \in \mathcal{R}} p^{k-1}$$

Alistair Moffat

Moffat, A. and Zobel, J. (2008). Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1):2:1–2:27.

Rank-based Measures: Rank-Biased Precision



- The user starts from the top ranked document and with probability p, called **persistence**, goes to the next document or with probability $1 - p$ stops
  - typical value for p are: 0.5 for impatient users, 0.8 for patient users, and 0.95 for extremely patient users

$$RBP = (1-p)\sum_{n=1}^{N} p^{n-1} r_n = (1-p)\sum_{k \in \mathcal{R}} p^{k-1}$$

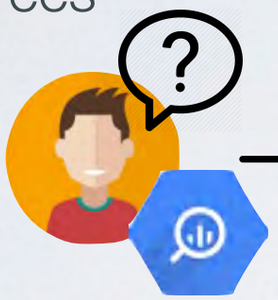
Alistair Moffat

Moffat, A. and Zobel, J. (2008). Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1):2:1–2:27.

Rank-based Measures: Example of Rank-biased Precision

Assume

- $p = 0.8$ a patient user
- Lenient mapping to binary relevance degrees



Topic

Run

	Assessed Run
1	Highly Relevant
2	Not Relevant
3	Partially Relevant
4	Fairly Relevant
5	Not Relevant
6	Not Relevant
7	Not Relevant
8	Fairly Relevant
9	Not Relevant
10	Not Relevant

	Binary Weighted Assessed Run
1	1
2	0
3	1
4	1
5	0
6	0
7	0
8	1
9	0
10	0

$$RBP = (1 - 0.8)\left(0.8^{1-1} + 0.8^{3-1} + 0.8^{4-1} + 0.8^{8-1}\right) = 0.4723$$

Questions?


It works...It doesn't....It works.....
Just a hunch....maybe we do need a better way to measure results.....

Statistical Hypothesis Testing

The Problem: Are They Different?

$$x = \begin{bmatrix} 3.1929 \\ 0.6575 \\ 1.4155 \\ -0.2043 \\ -1.6054 \\ 1.1446 \\ 1.1033 \\ 0.7418 \\ 0.5035 \\ 1.5477 \end{bmatrix} \quad y = \begin{bmatrix} 0.7471 \\ 1.6357 \\ 1.0390 \\ 1.1421 \\ 1.5731 \\ 3.5389 \\ 1.9563 \\ 0.6754 \\ 2.6305 \\ 1.9505 \end{bmatrix} \quad z = \begin{bmatrix} 0.6408 \\ 2.6859 \\ 3.0818 \\ 3.1258 \\ 2.9028 \\ 2.4371 \\ 1.7648 \\ 3.2616 \\ 2.0997 \\ 1.0346 \end{bmatrix}$$

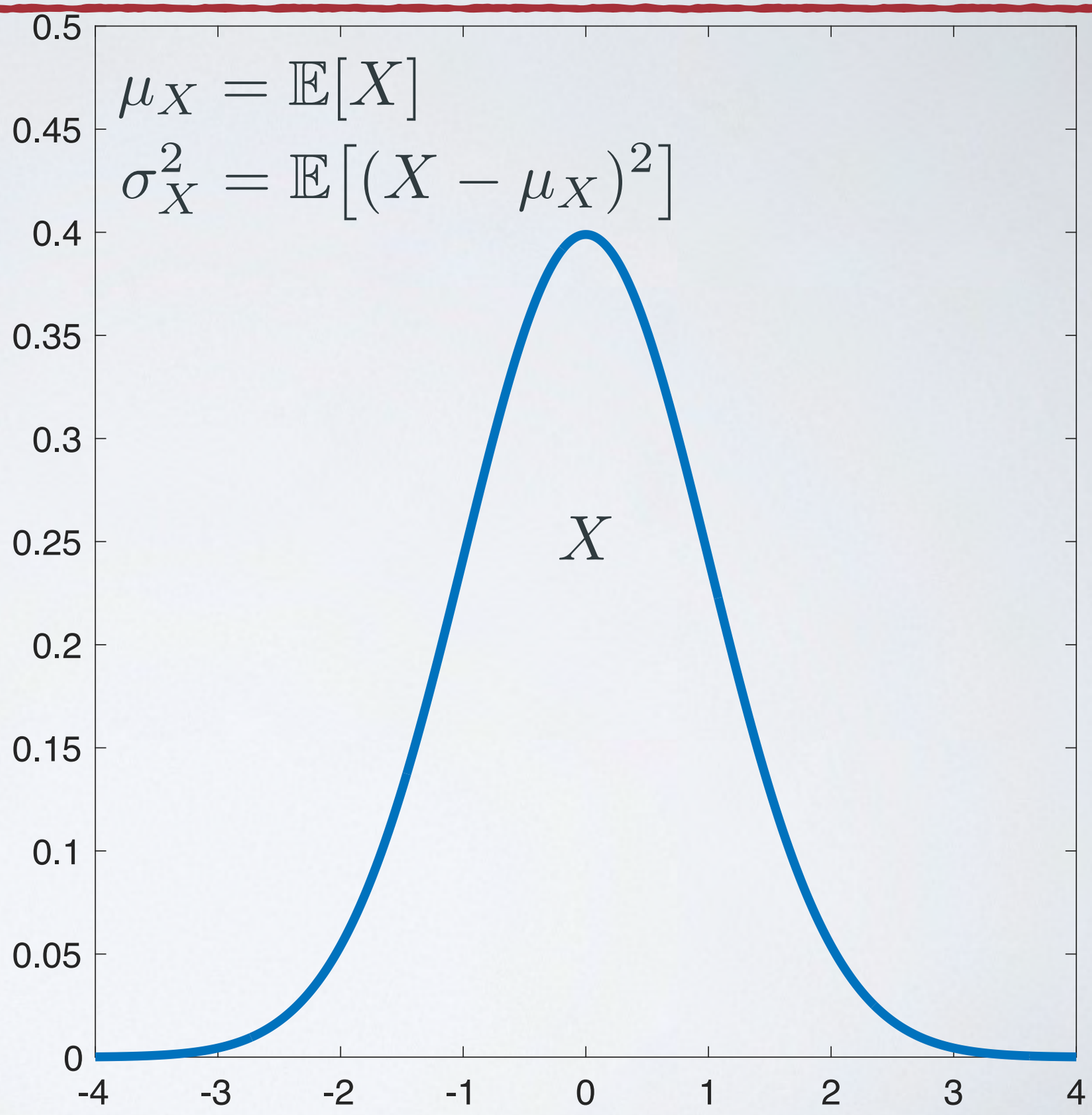
X → $\hat{\mu}_X = 0.8497$, $\hat{\sigma}^2_X = 1.5284$

Y → $\hat{\mu}_Y = 1.6889$, $\hat{\sigma}^2_Y = 0.7890$

Z → $\hat{\mu}_Z = 2.3035$, $\hat{\sigma}^2_Z = 0.8255$

What Did We Do?

$$x = \begin{bmatrix} 3.1929 \\ 0.6575 \\ 1.4155 \\ -0.2043 \\ -1.6054 \\ 1.1446 \\ 1.1033 \\ 0.7418 \\ 0.5035 \\ 1.5477 \end{bmatrix}$$



What Did We Do?

$$x = \begin{bmatrix} 3.1929 \\ 0.6575 \\ 1.4155 \\ -0.2043 \\ -1.6054 \\ 1.1446 \\ 1.1033 \\ 0.7418 \\ 0.5035 \\ 1.5477 \end{bmatrix} \begin{matrix} \longleftarrow X_1 \\ \longleftarrow X_2 \\ \longleftarrow X_3 \\ \longleftarrow X_4 \\ \longleftarrow X_5 \\ \longleftarrow X_6 \\ \longleftarrow X_7 \\ \longleftarrow X_8 \\ \longleftarrow X_9 \\ \longleftarrow X_{10} \end{matrix}$$

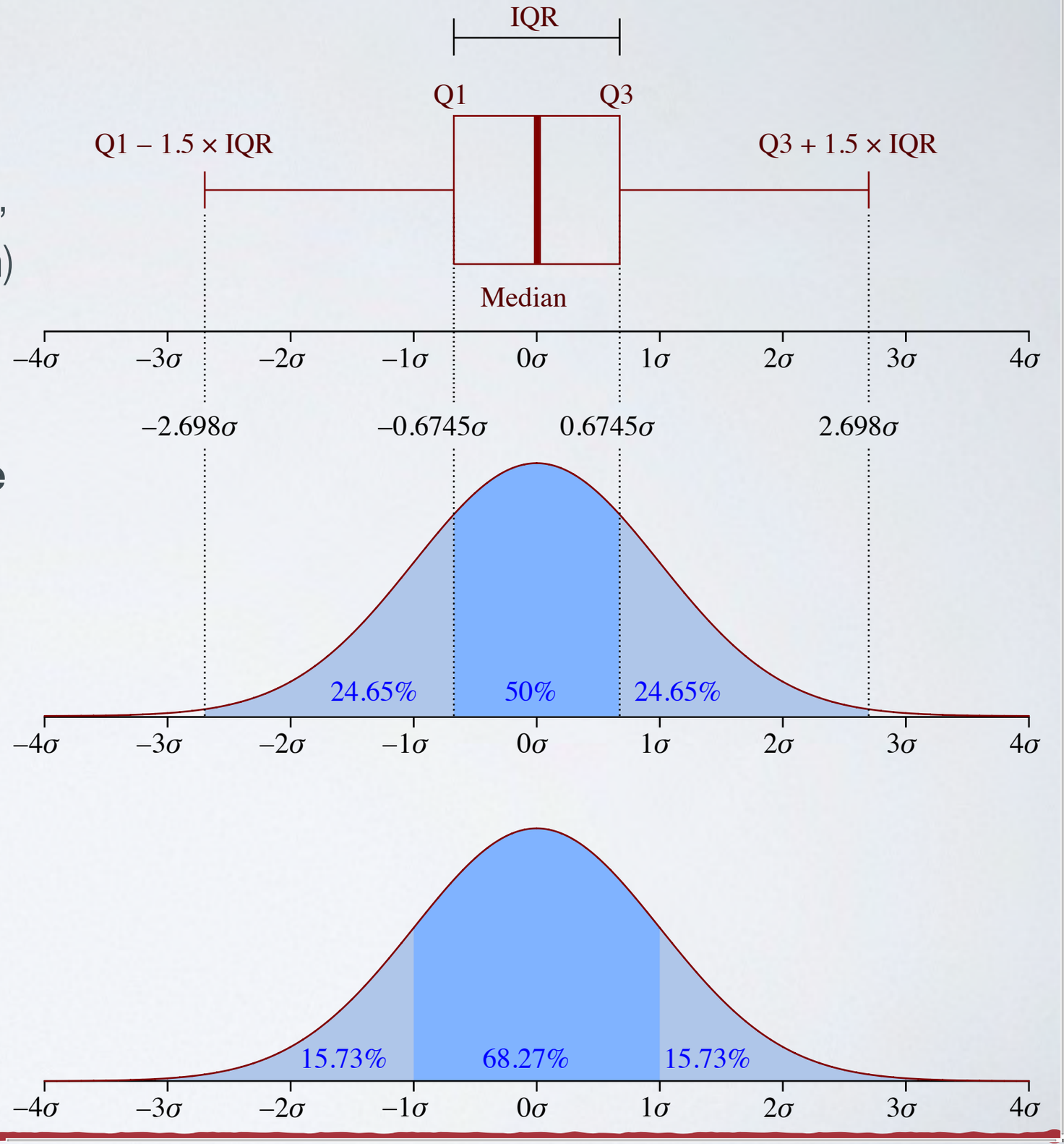
$$\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2_X = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu}_X)^2$$

- $\{X_1, X_2, \ldots, X_n\}$ is a **random sample** of size n, i.e. a sequence of **independent and identically distributed (i.i.d)** random variables drawn from a distribution
  - x_i is an **observation** (realisation) of the random variable X_i, i.e. the actual value assumed by that random variable in a given trial
- The **sample mean** $\hat{\mu}_X$ and the **sample variance** $\hat{\sigma}^2_X$ are **unbiased estimators** of the **population mean** μ_X and **population variance** σ^2_X

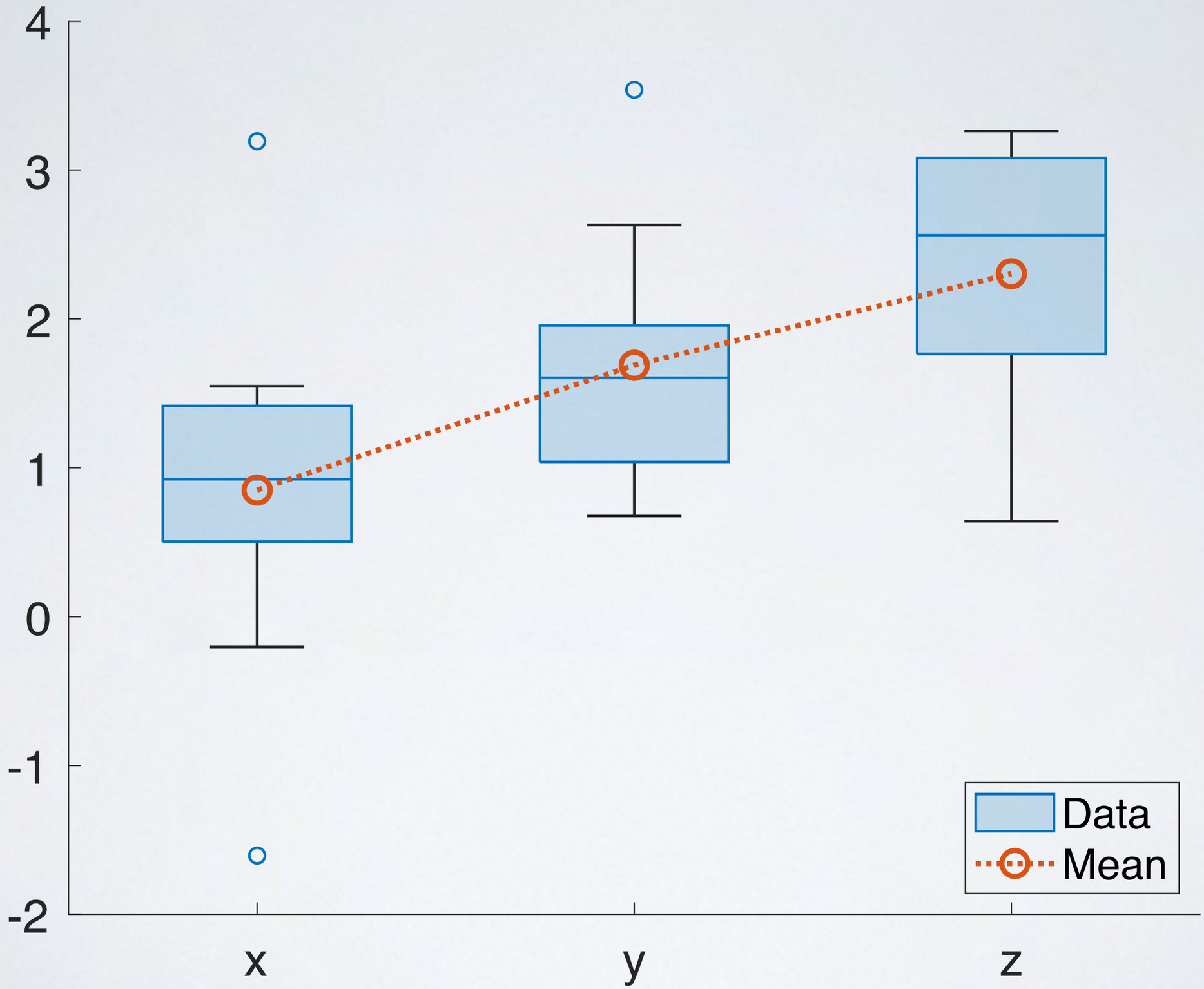
Box Plot

- A **boxplot** is a graphical tool to summarise a distribution of data
- The box shows the **first quartile** (Q1), the **second quartile** (Q2, the **median**) with a line inside the box, and the **third quartile** (Q3)
- The box represents the **Inter-Quartile Range (IQR)**, i.e. the difference Q3-Q1
- The extension of the **whiskers** represents 1.5·IQR
  - they roughly cover ~99% of the data, assuming a normal distribution
- Any data outside the whiskers is considered an **outlier**



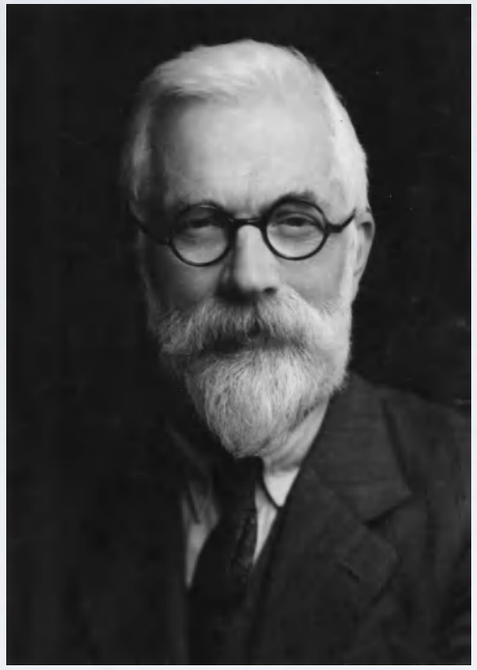
McGill, R., Tukey, J. W., and Larsen, W. A. (1978). Variations of Box Plots. *The American Statistician*, 32(1):12–16.

The Problem: Are They Different?

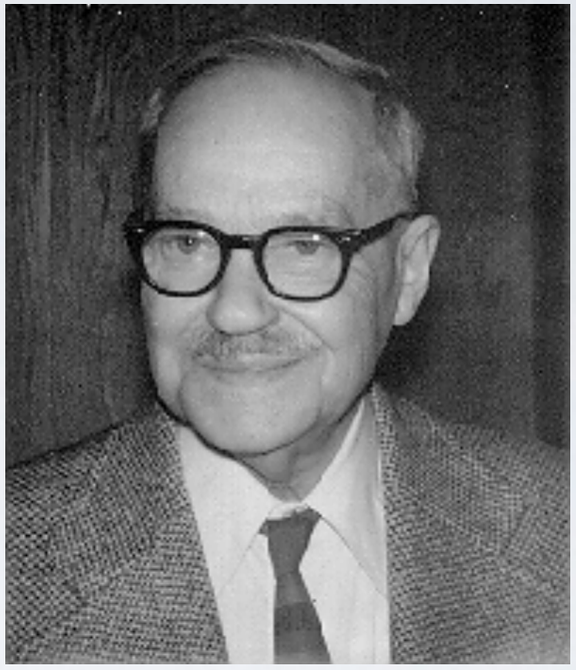


Statistical Hypothesis Testing

William Sealy Gosset "Student"



Ronald Aylmer Fisher



Jerzy Neyman

Egon Sharpe Pearson

- **Statistical hypothesis testing** provides us with a mathematical framework to conduct **statistical inference** from the data
- It compares the so-called **null hypothesis** H_0 against an **alternative hypothesis** H_1 or H_A
- The comparison is **statistically significant** if the data are unlikely to be a realisation of the null hypothesis with respect to a chosen threshold, called **significance level** α. In this case we **reject** the null hypothesis; in the opposite case, we **fail to reject** the null hypothesis

Student (1908). The Probable Error of a Mean. *Biometrika*, 6(1):1–25.
Fisher, R. A. (1925). Statistical Methods for Research Workers. Oliver & Boyd, Edinburgh, UK.
Neyman, J. and Pearson, E. S. (1928). On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference. *Biometrika*, 20A(1/2):175–240.
Lehmann, E. L. (1993). The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two? *Journal of the American Statistical Association*, 88(424):1242–1249.

Formulating the Problem

$$x = \begin{bmatrix} 3.1929 \\ 0.6575 \\ 1.4155 \\ -0.2043 \\ -1.6054 \\ 1.1446 \\ 1.1033 \\ 0.7418 \\ 0.5035 \\ 1.5477 \end{bmatrix}$$

$$\hat{\mu}_X = 0.8497$$

Sample Mean

$$\begin{cases} H_0: & \mu_X = \mu_Y \\ H_1: & \mu_X \neq \mu_Y \end{cases}$$

Random Sampling

$\mu_X - \mu_Y$

Random Sampling

$$y = \begin{bmatrix} 0.7471 \\ 1.6357 \\ 1.0390 \\ 1.1421 \\ 1.5731 \\ 3.5389 \\ 1.9563 \\ 0.6754 \\ 2.6305 \\ 1.9505 \end{bmatrix}$$

$$\hat{\mu}_Y = 1.6889$$

Sample Mean

μ_X Population Mean

μ_Y Population Mean

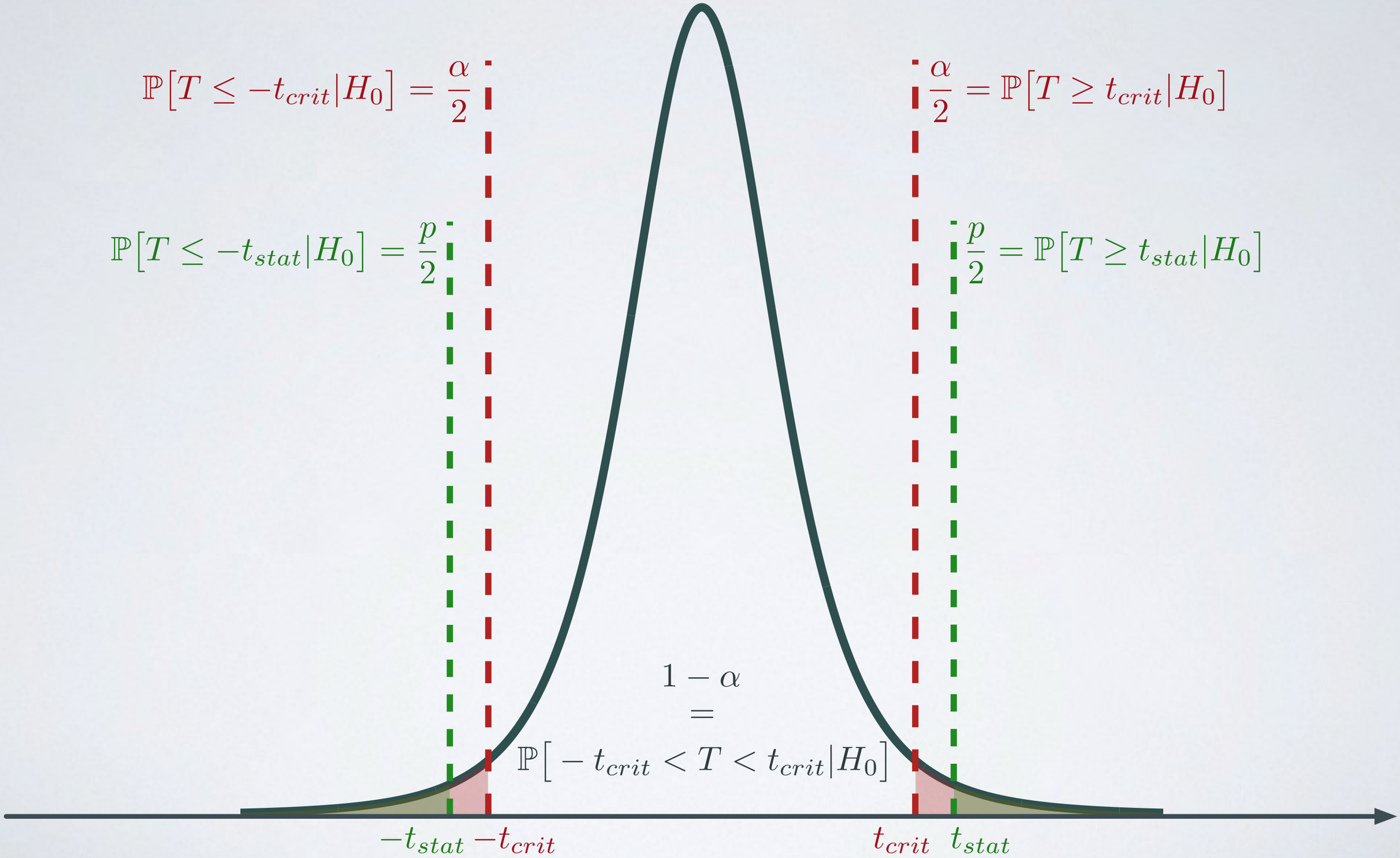
Sakai, T. (2018). *Laboratory Experiments in Information Retrieval*, volume 40 of *The Information Retrieval Series*. Springer Singapore.

Test Statistic

Test statistic distribution T under the null hypothesis H_0

$$\mathbb{P}\big[T \leq -t_{crit} | H_0\big] = \frac{\alpha}{2}$$

$$\mathbb{P}\big[T \leq -t_{stat} | H_0\big] = \frac{p}{2}$$



$$\frac{\alpha}{2} = \mathbb{P}\big[T \geq t_{crit} | H_0\big]$$

$$\frac{p}{2} = \mathbb{P}\big[T \geq t_{stat} | H_0\big]$$

Test Statistic

Test statistic distribution T under the null hypothesis H_0

$$\mathbb{P}\big[T \leq -t_{crit} | H_0\big] = \frac{\alpha}{2}$$

$$\mathbb{P}\big[T \leq -t_{stat} | H_0\big] = \frac{p}{2}$$

$$\frac{\alpha}{2} = \mathbb{P}\big[T \geq t_{crit} | H_0\big]$$

$$\frac{p}{2} = \mathbb{P}\big[T \geq t_{stat} | H_0\big]$$

$$1 - \alpha = \mathbb{P}\big[-t_{crit} < T < t_{crit} | H_0\big]$$

$-t_{stat}$ $-t_{crit}$ t_{crit} t_{stat}

Types of Error

	**We fail to reject** H_0 **[not statistically significant]**	**We reject** H_0 **[statistically significant]**
H_0 **is true** **[e.g. systems are equivalent]**	Correct conclusion [true negative] Probability $1 - \alpha$	**Type I error** [false positive] Probability α
H_0 **is false** **[e.g. systems are not equivalent]**	**Type II Error** [false negative] Probability β	**Power** (Correct conclusion) [true positive] Probability $1 - \beta$

Multiple Comparisons

- Type I errors concern the comparison of 2 samples
  - What happens when you need to compare c samples?
  - They originate $m = \binom{c}{2}$ possible pairs to be compared, i.e. hypotheses to be tested simultaneously
- Performing multiple comparisons increases the Type I error probability, i.e. it is easier the reject the null hypothesis when you should not, since the m pairwise comparisons are **independent**

$$\mathbb{P}(\text{No Type I Error}) = 1 - \alpha$$

$$\mathbb{P}(\text{No Type I Errors}) = \prod_{i=1}^{m}(1-\alpha) = (1-\alpha)^m$$

$$\mathbb{P}(\text{At Least One Type I Error}) = 1 - (1-\alpha)^m$$

General Linear Models (GLM)

$$\text{Data} = \text{Model} + \text{Error}$$

- A GLM explains the variation of a **dependent variable** (**Data**) in terms of a controlled variation of **independent variables** (**Model**) in addition to a **residual** uncontrolled variation (**Error**)
- Regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- ANalysis Of VAriance (ANOVA)

$$y_{ij} = \mu.. + \alpha_j + \varepsilon_{ij}$$

- the above regression model corresponds to the ANOVA one once you add as many x_{ij} predictors as many levels there are in the experimental condition α_j, e.g., by using dummy coding

Rutherford, A. (2011). *ANOVA and ANCOVA. A GLM Approach*. John Wiley & Sons, New York, USA, 2nd edition.

Modelling System Effects (one-way ANOVA)

ANOVA

$$y_{ij} = \underbrace{\mu_{..} + \alpha_j}_{\text{Model}} + \underbrace{\varepsilon_{ij}}_{\text{Error}}$$

	α_1	α_2	$\cdots$	α_q	
τ_1	y_{11}	y_{12}	$\cdots$	y_{1q}	$\mu_{1.}$
τ_2	y_{21}	y_{22}	$\cdots$	y_{2q}	$\mu_{2.}$
$\vdots$	$\vdots$	$\vdots$	y_{ij}	$\vdots$	$\mu_{i.}$
τ_p	y_{p1}	y_{p2}	$\cdots$	y_{pq}	$\mu_{p.}$
	$\mu_{.1}$	$\mu_{.2}$	$\mu_{.j}$	$\mu_{.q}$	$\mu_{..}$

Modelling System Effects (one-way ANOVA)

	α_1	α_2	$\cdots$	α_q	
τ_1	y_{11}	y_{12}	$\cdots$	y_{1q}	$\mu_{1\cdot}$
τ_2	y_{21}	y_{22}	$\cdots$	y_{2q}	$\mu_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	y_{ij}	$\vdots$	$\mu_{i\cdot}$
τ_p	y_{p1}	y_{p2}	$\cdots$	y_{pq}	$\mu_{p\cdot}$
	$\mu_{\cdot 1}$	$\mu_{\cdot 2}$	$\mu_{\cdot j}$	$\mu_{\cdot q}$	$\mu_{\cdot\cdot}$

Estimators

- Grand mean

$$\hat{\mu}_{..} = \frac{1}{pq} \sum_{i=1}^{p} \sum_{j=1}^{q} y_{ij}$$

- Marginal mean of the j-th system

$$\hat{\mu}_{.j} = \frac{1}{p} \sum_{i=1}^{p} y_{ij} \qquad \hat{\alpha}_{.j} = \hat{\mu}_{.j} - \hat{\mu}_{..}$$

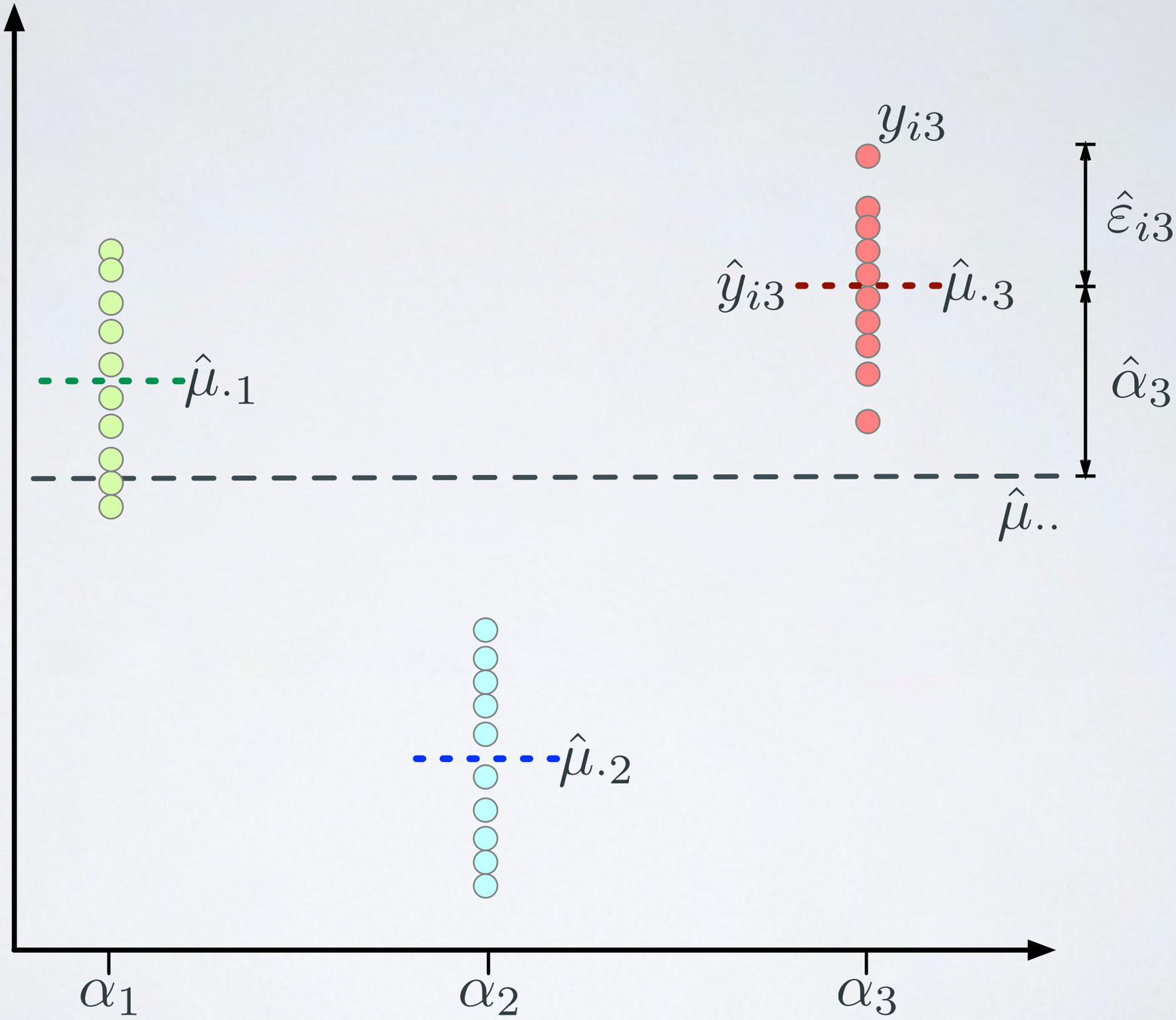
- Predicted score

$$\hat{y}_{ij} = \hat{\mu}_{..} + \hat{\alpha}_{.j} = \hat{\mu}_{.j}$$

- Prediction error

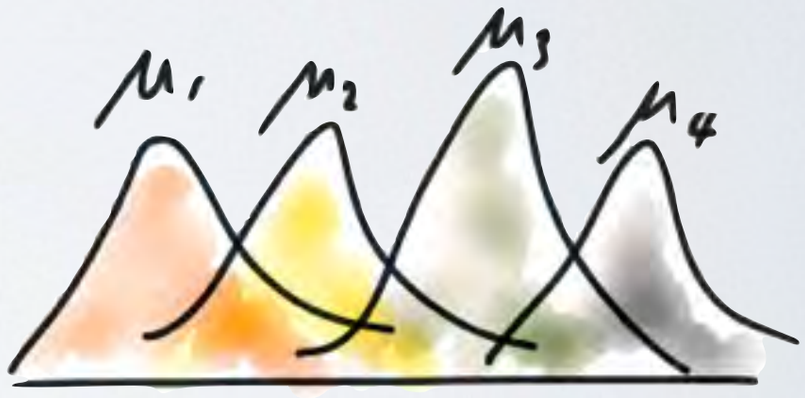
$$\hat{\varepsilon}_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \hat{\mu}_{.j}$$

Estimators


y_{i3}
$\hat{\varepsilon}_{i3}$
$\hat{y}_{i3}$
$\hat{\mu}_{.3}$
$\hat{\alpha}_3$
$\hat{\mu}_{.1}$
$\hat{\mu}_{..}$
$\hat{\mu}_{.2}$
α_1
α_2
α_3

Assessment: ANOVA

- **Analysis of Variance (ANOVA)** was developed by statistician and evolutionary biologist Ronald Fisher (1890-1962)
- It provides a statistical test of whether or not the means of several groups are equal
  - H_0 is the **null hypothesis** that all the means are equal
- It partitions the observed variance in a particular variable into components attributable to different sources of variation



ANOVA

$\mu_1 = \mu_2 = \mu_3 = \mu_4$?

Assessment: Sum of Squares

$$\underbrace{y_{ij} - \hat{\mu}_{..}}_{\text{Total Effect}} = \underbrace{\hat{\mu}_{.j} - \hat{\mu}_{..}}_{\text{System Effect } \hat{\alpha}_j} + \underbrace{y_{ij} - \hat{\mu}_{.j}}_{\text{Error Effect } \hat{\varepsilon}_{ij}}$$

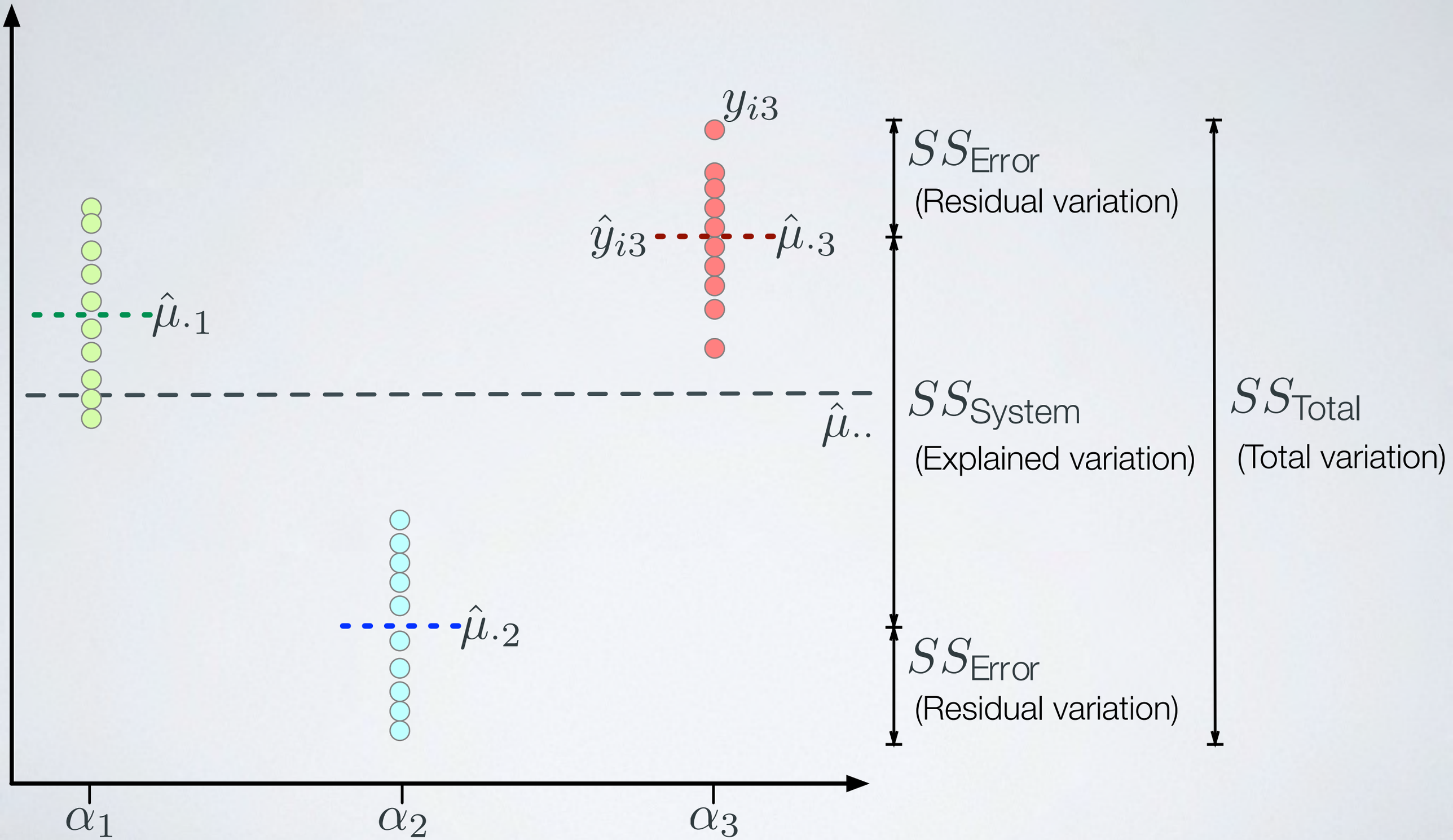
- Sum of squares (SS)

$$SS_{fact} = \sum_{i=1}^{p} \sum_{j=1}^{q} \left([\text{total} \mid \text{system} \mid \text{error}] \text{ effect} \right)^2$$

- Variance break-down

$$SS_{\text{Total}} = SS_{\text{System}} + SS_{\text{Error}}$$

Assessment: Sum of Squares



Assessment: Degrees of Freedom and Mean Squares

- Degrees of Freedom (DF)

$$DF_{\text{Total}} = pq - 1$$

$$DF_{\text{System}} = q - 1$$

$$DF_{\text{Error}} = (pq - 1) - (q - 1) = q(p - 1)$$

- Mean Squares (MS)

$$MS_{fact} = \frac{SS_{fact}}{DF_{fact}}$$

Walker, H. M. (1940). Degrees of freedom. *Journal of Educational Psychology*, 31(4):253–269.

Assessment: F-test

- Let us assume to have a random sample of size $n = pq$ (**independence**) from q **normally-distributed** random variables with **same variance** (homoskedasticity)

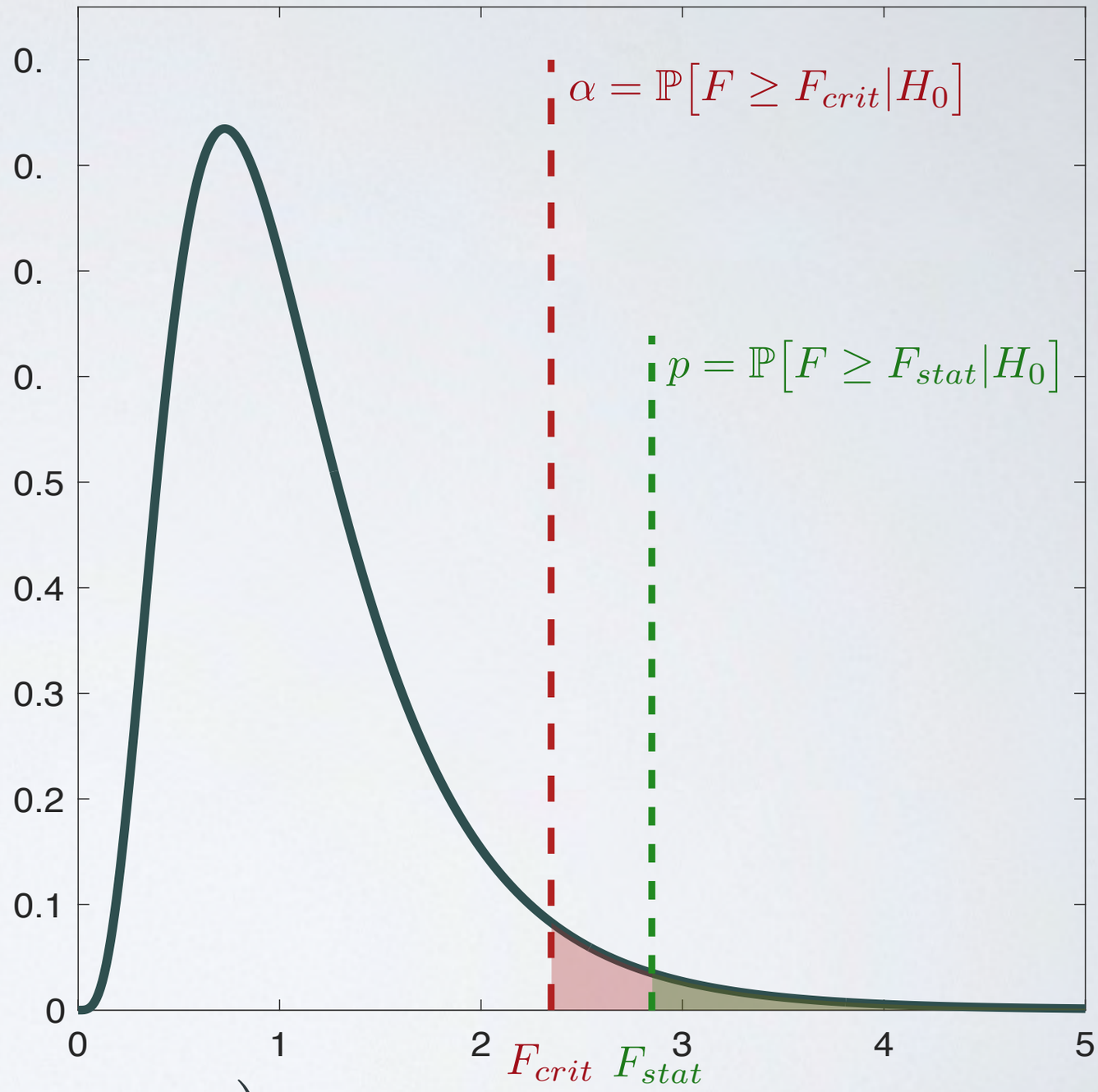
$$Y_{ij} \sim \mathcal{N}(\mu_j, \sigma^2)$$

- Under the null hypothesis

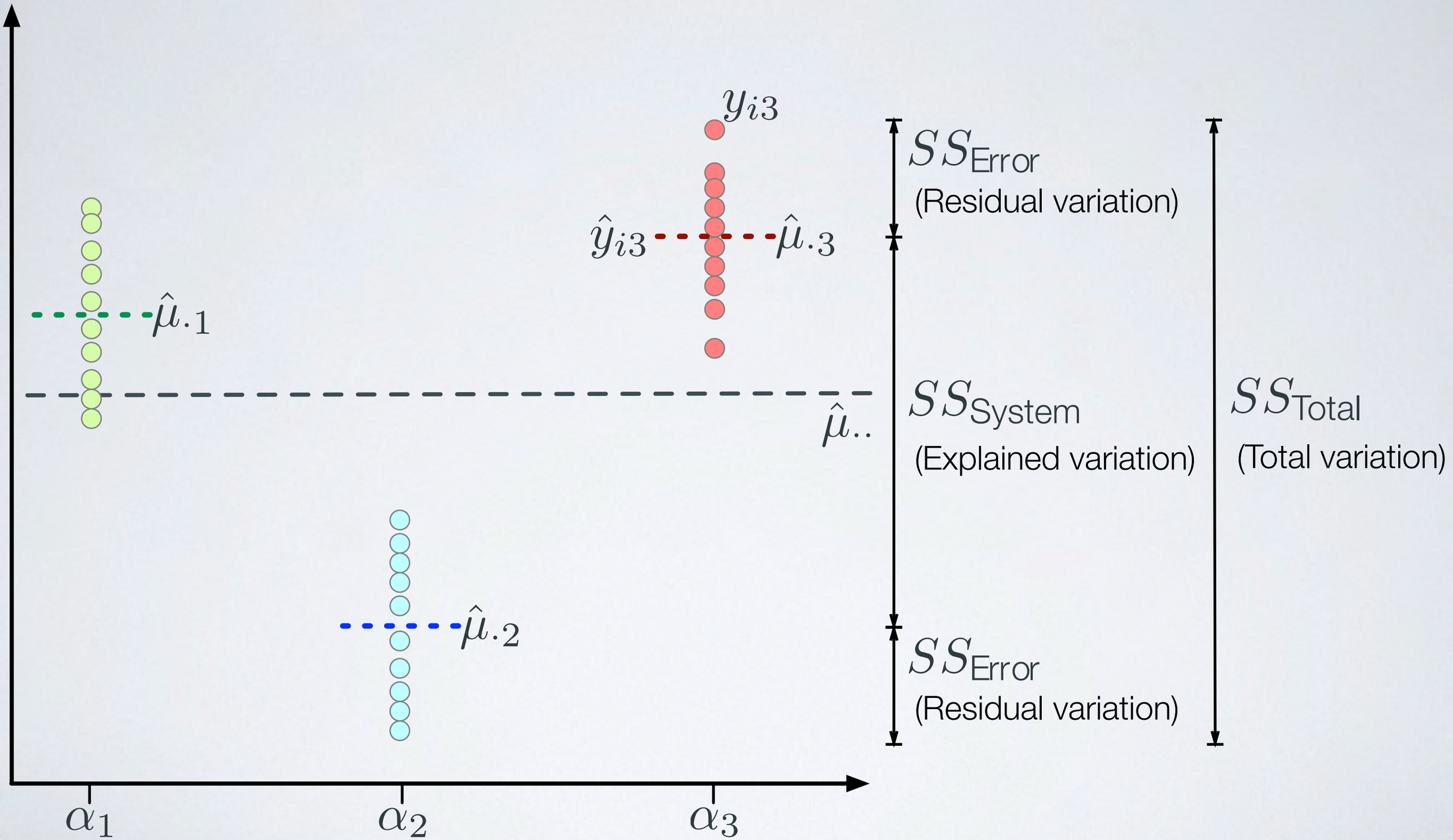
$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_q$$

the tests statistic is

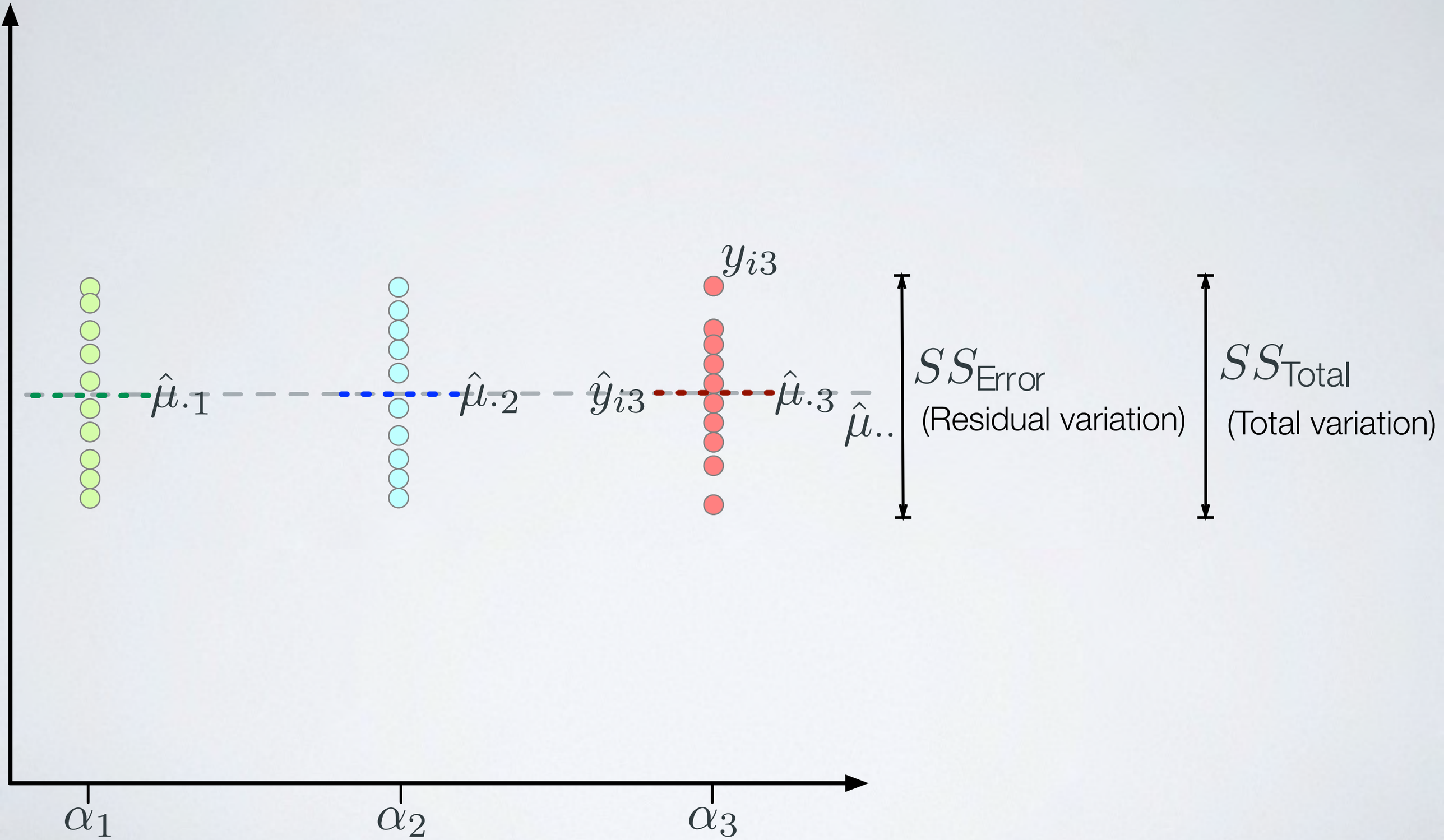
$$F_{stat} = \frac{MS_{\text{System}}}{MS_{\text{Error}}} \sim F_{\left(DF_{\text{System}}, DF_{\text{Error}}\right)}$$



Assessment: F-test under H_0



Assessment: F-test under H_0



Tukey HSD Test

- The **Tukey Honestly Significant Difference (HSD)** test creates confidence intervals $|t|$ for all pairwise differences between factor levels, while controlling the family error rate

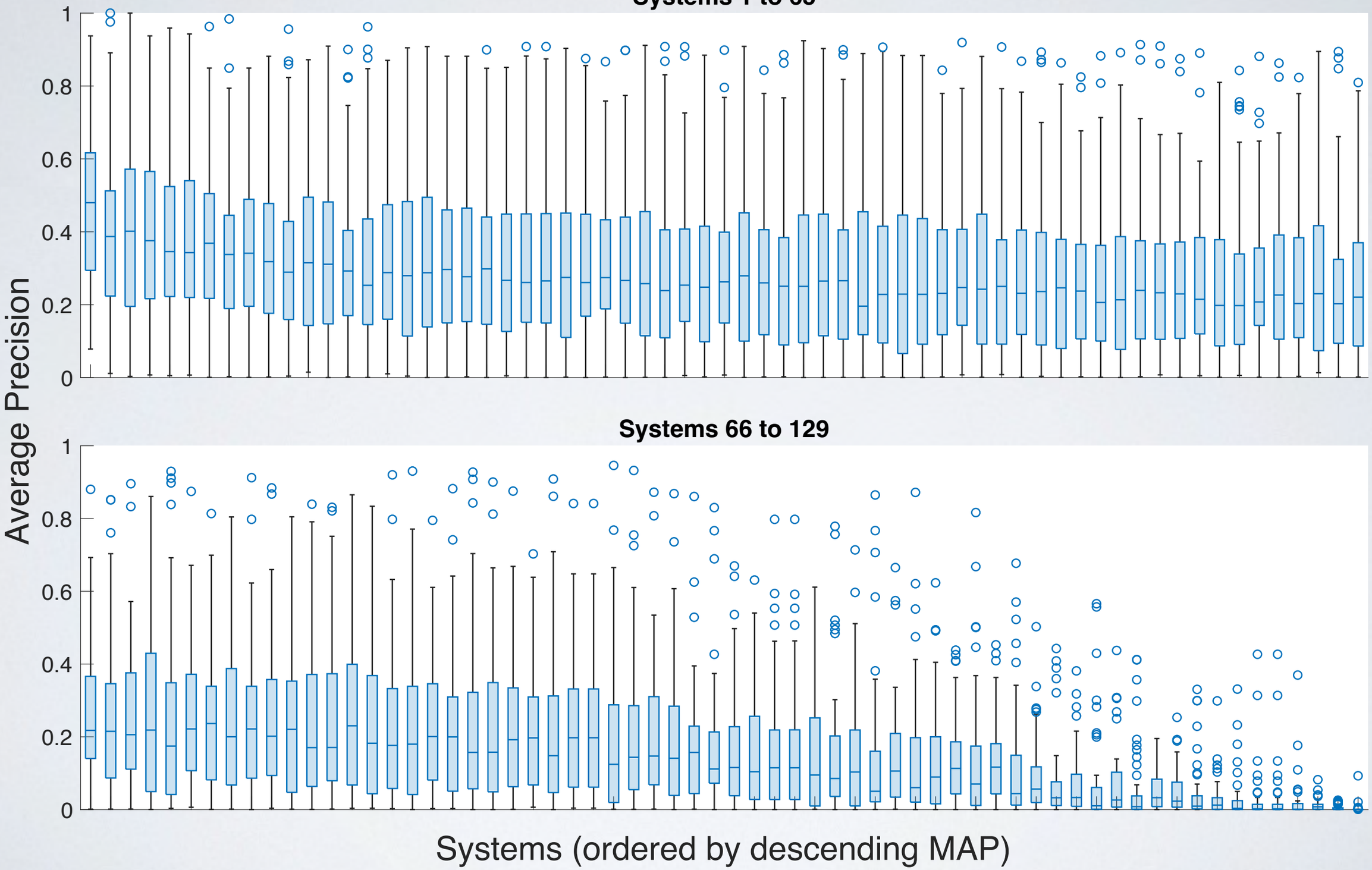
$$|t| = \frac{|\hat{\mu}_{.u} - \hat{\mu}_{.v}|}{\sqrt{\frac{2MS_{\text{Error}}}{p}}} > \frac{1}{\sqrt{2}} \mathrm{q}_{\alpha,q,q(p-1)}$$

  - $\hat{\mu}_{.u}$ and $\hat{\mu}_{.v}$ are the marginal means of the two factor levels, i.e. the two systems to be compared
  - $\mathrm{q}_{\alpha,q,q(p-1)}$ is the upper $100 * (1 - \alpha)$-th percentile of the **studentized range distribution**, i.e. the distribution of the range of samples drawn from a normal distribution, considering q systems to compare using p topics

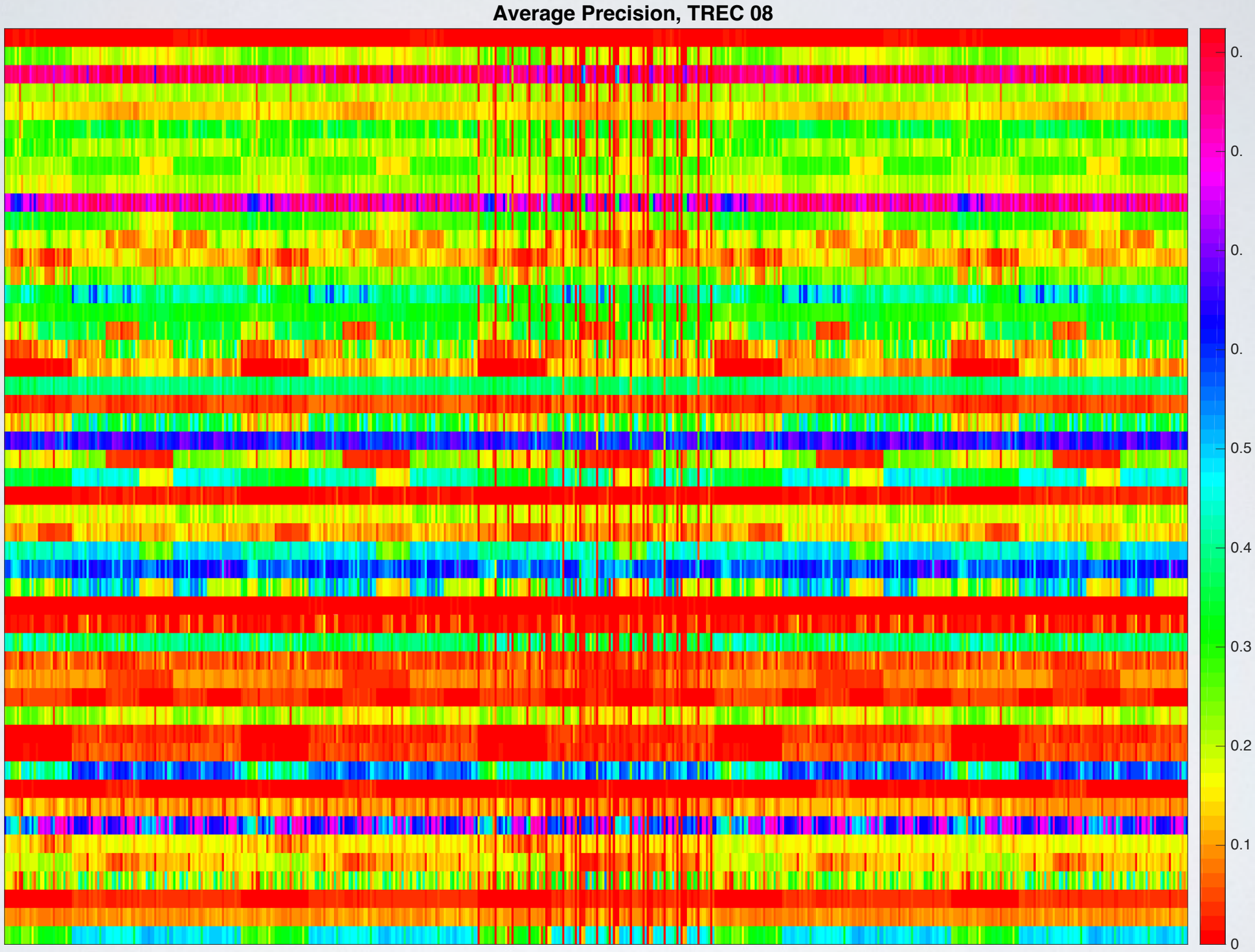
John Wilder Tukey

Tukey, J. W. (1949). Comparing Individual Means in the Analysis of Vari- ance. Biometrics, 5(2):99–114.
Newman, D. (1939). The Distribution of Range in Samples from a Normal Population, Expressed in Terms of an Independent Estimate of Standard Deviation. *Biometrika*, 31(2):20–30.
Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. John Wiley & Sons, USA.

Example: Analysis of TREC 8 Ad Hoc



Example: Analysis of TREC 8 Ad Hoc



	α_1	α_2	$\cdots$	α_q	
τ_1	y_{11}	y_{12}	$\cdots$	y_{1q}	$\mu_{1\cdot}$
τ_2	y_{21}	y_{22}	$\cdots$	y_{2q}	$\mu_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	y_{ij}	$\vdots$	$\mu_{i\cdot}$
τ_p	y_{p1}	y_{p2}	$\cdots$	y_{pq}	$\mu_{p\cdot}$
	$\mu_{\cdot 1}$	$\mu_{\cdot 2}$	$\mu_{\cdot j}$	$\mu_{\cdot q}$	$\mu_{\cdot\cdot}$

Example: Analysis of TREC 8 Ad Hoc

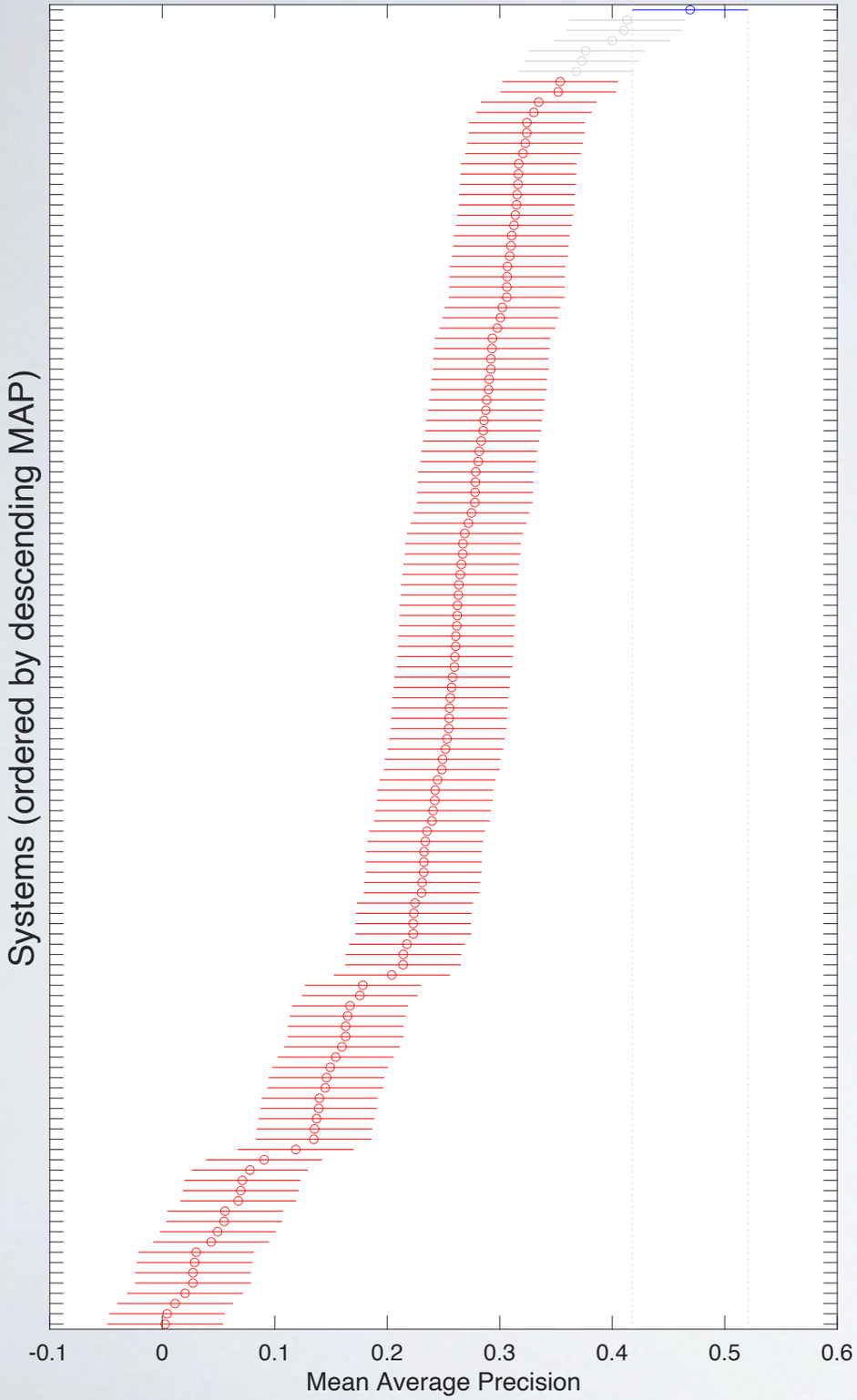
$$y_{ij} = \mu_{..} + \underbrace{\tau_i}_{\text{Topic Effect}} + \underbrace{\alpha_j}_{\text{System Effect}} + \varepsilon_{ij}$$

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	167.9974	49	3.4285	251.9463	0	0.6559
System	60.0299	128	0.4690	34.4635	0	0.3991
Error	85.3502	6272	0.0136			
Total	313.3375	6449				

- Topic is a statistically significant and large-size effect
- IR system is a statistically significant and large(medium)-size effect, quite smaller than topic

Banks, D., Over, P., and Zhang, N.-F. (1999). Blind Men and Elephants: Six Approaches to TREC data. *Information Retrieval*, 1(1-2):7–34.
Tague-Sutcliffe, J. M. and Blustein, J. (1994). A Statistical Analysis of the TREC-3 Data. In Harman, D. K., editor, *The Third Text REtrieval Conference (TREC-3)*, pages 385–398. National Institute of Standards and Technology (NIST), Special Publication 500-225, Washington, USA.

Example: Analysis of TREC 8 Ad Hoc



- 129 systems amount to 8,256 system pairs to be compared
- According to Tukey HSD Test 3,425 system pairs are significantly different
- The top group consists of 7 systems

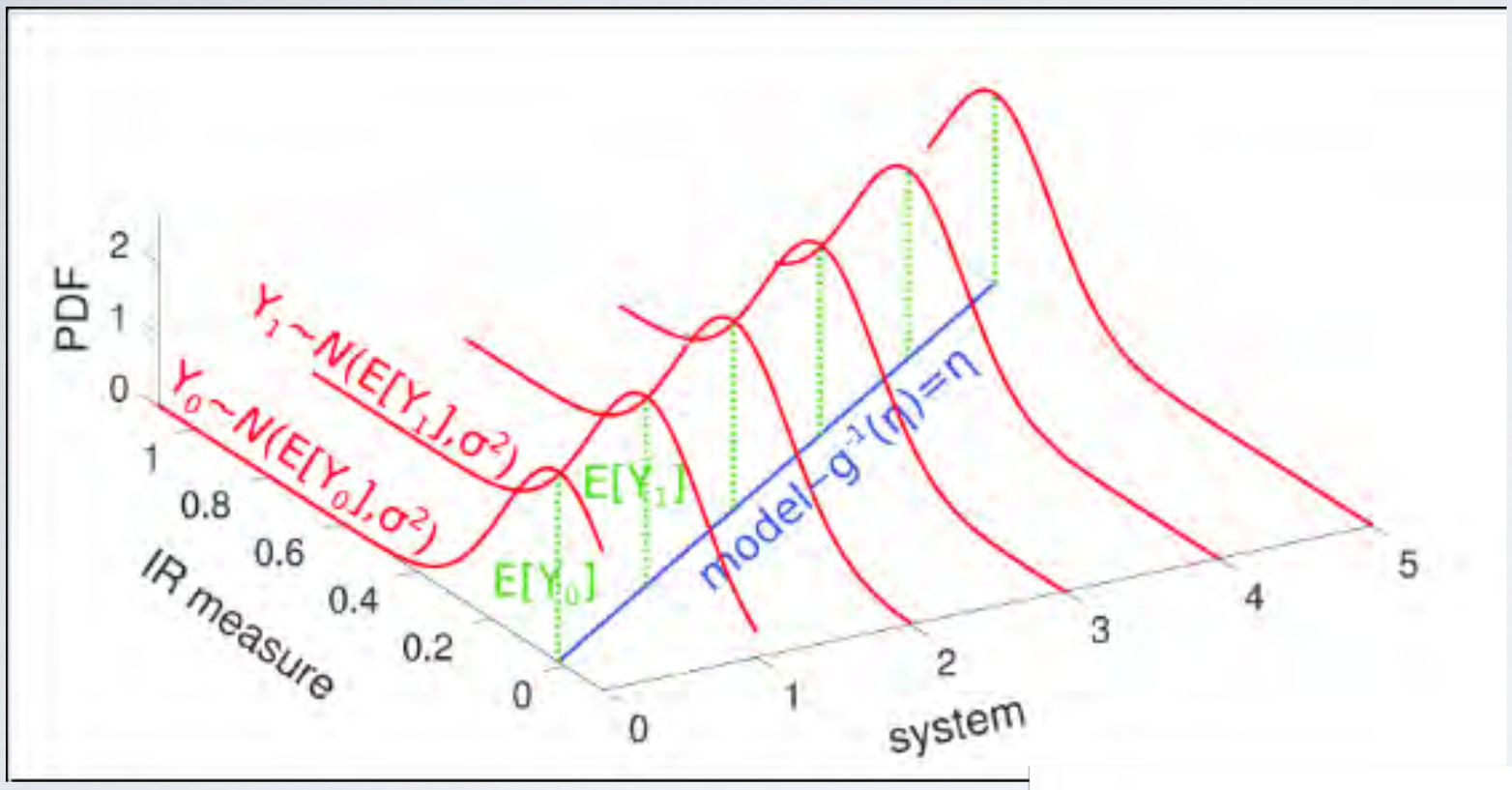
ANOVA Assumptions?

- **Independence**: topics and systems can be considered (reasonably) independent
- **Normality**: typical IR measures are bounded in [0, 1] so they cannot be normal
  - the normal distribution is unbounded
- **Homoskedasticity**: variance changes across systems

- ANOVA is considered robust to violations of normality and also homoskedasticy when the sample sizes are equal and large (our typical case)

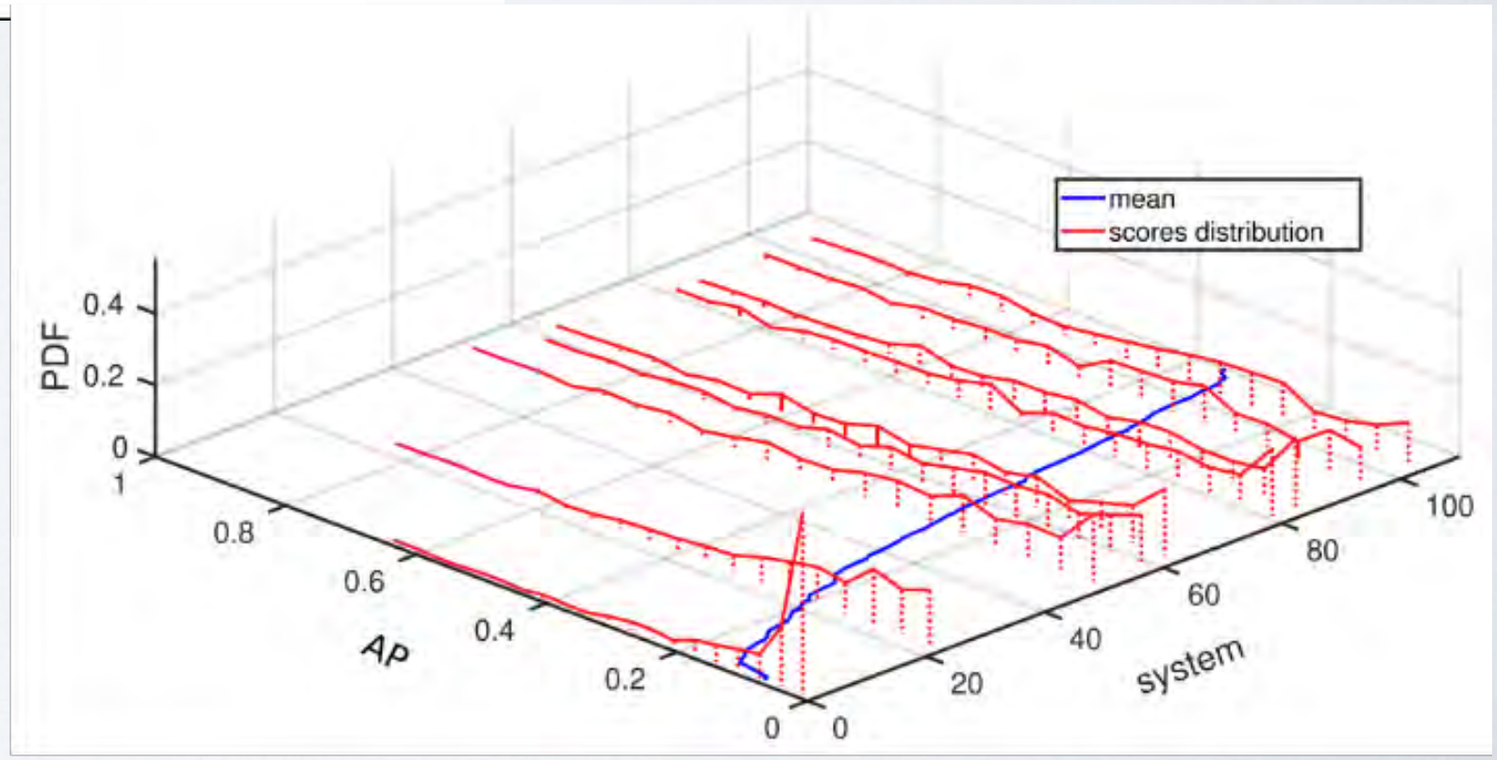
Eisenhart, C. (1947). The Assumptions Underlying the Analysis of Variance. *Biometrika*, 3(1):1–21.
Ito, P. K. (1980). Robustness of ANOVA and MANOVA test procedures. In Krishnaiah, P. R., editor, *Handbook of Statistics – Analysis of Variance*, volume 1, pages 199–236. Elsevier, The Netherlands.
Tague-Sutcliffe, J. M. and Blustein, J. (1994). A Statistical Analysis of the TREC-3 Data. In Harman, D. K., editor, *The Third Text REtrieval Conference (TREC-3)*, pages 385–398. National Institute of Standards and Technology (NIST), Special Publication 500-225, Washington, USA.
Carterette, B. A. (2012). Multiple Testing in Statistical Analysis of Systems-Based Information Retrieval Experiments. *ACM Transactions on Information Systems (TOIS)*, 30(1):4:1–4:34.

ANOVA Assumptions?

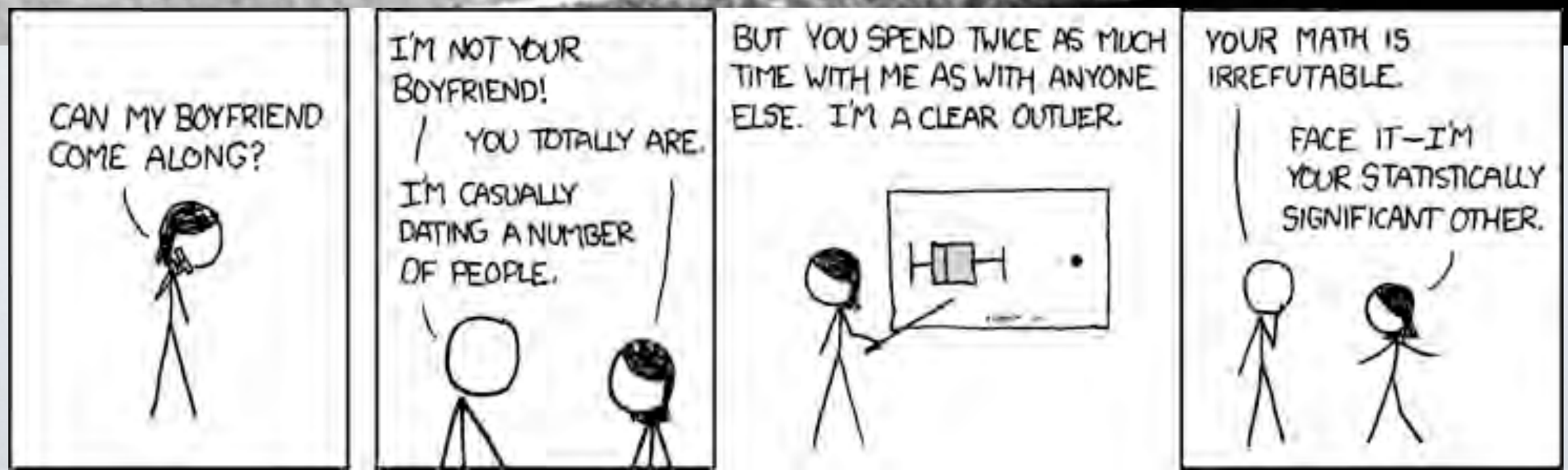


ANOVA assumptions

Real IR data



Questions?


CAN MY BOYFRIEND COME ALONG?
I'M NOT YOUR BOYFRIEND!
YOU TOTALLY ARE.
I'M CASUALLY DATING A NUMBER OF PEOPLE.
BUT YOU SPEND TWICE AS MUCH TIME WITH ME AS WITH ANYONE ELSE. I'M A CLEAR OUTLIER.
YOUR MATH IS IRREFUTABLE.
FACE IT—I'M YOUR STATISTICALLY SIGNIFICANT OTHER.

Foundations of Measurement

The Problem

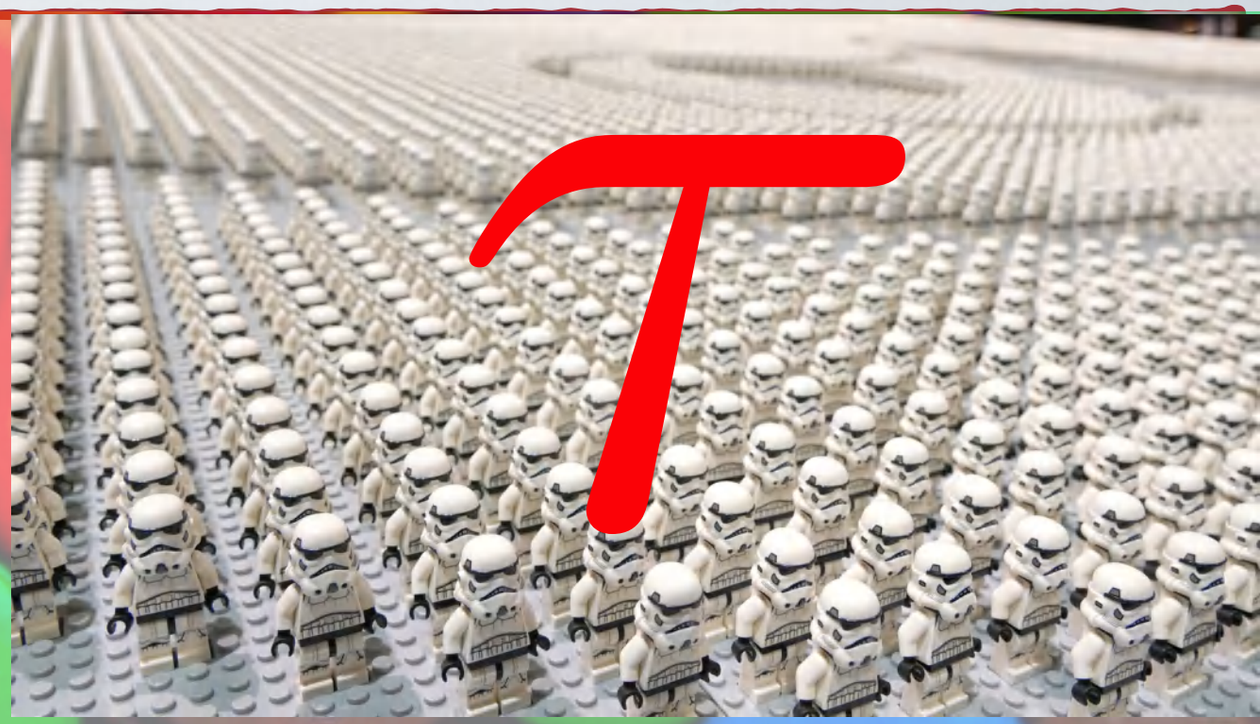
Deeply Rooted…

Experimentation

A Science of Measures...

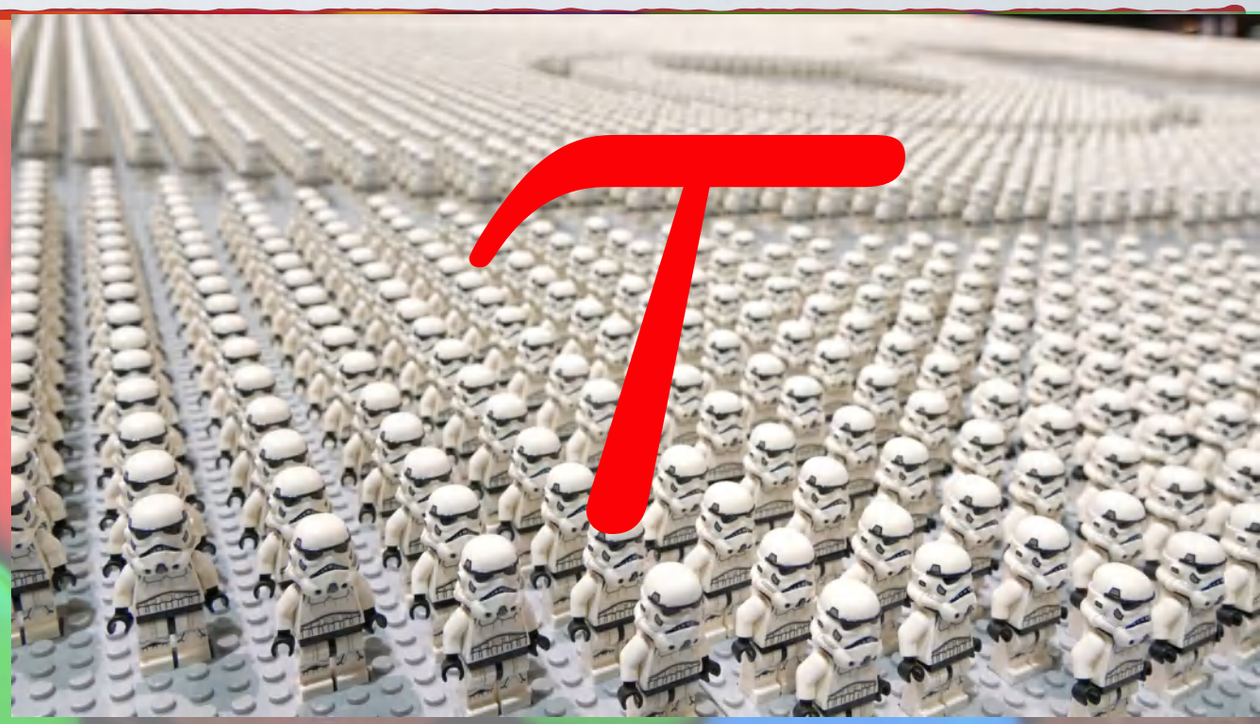
A Science of Measures...

User Models



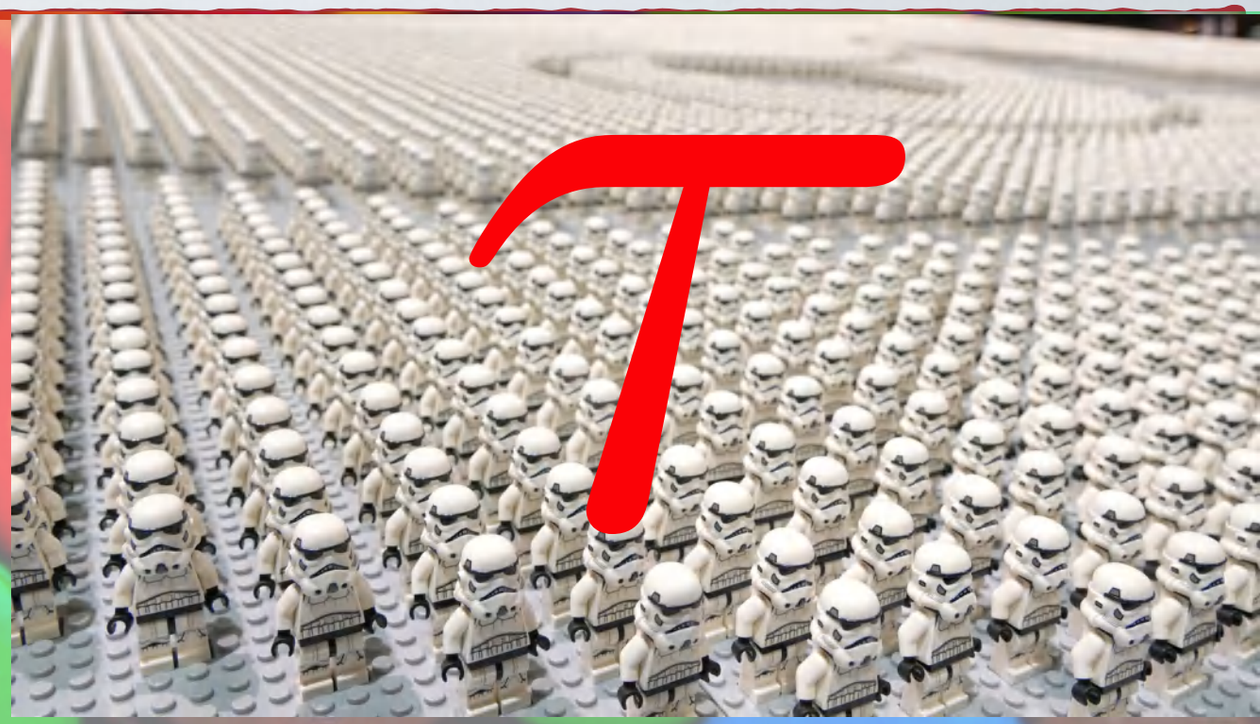
A Science of Measures...

User Models



Top-heaviness

A Science of Measures...

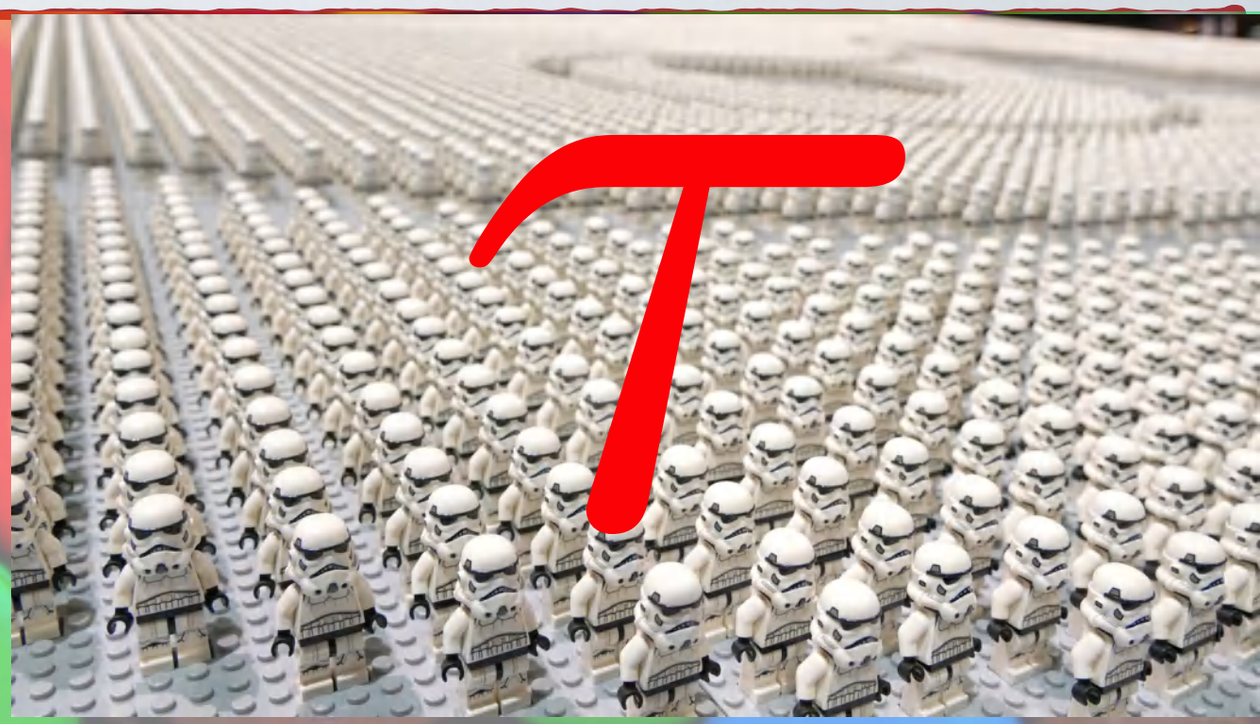


User Models

Top-heaviness Incomplete Information

A Science of Measures...

User Models



Top-heaviness

Incomplete Information

Sensitivity

… But, Wait, Assumptions?

- The **operations** you are **allowed** to perform with the values of a measure depend on the notion of **measurement scale**
  - mean
  - variance
  - correlation
  - statistical tests
  - ….

… But, Wait, Assumptions?

- The **operations** you are **allowed** to perform with the values of a measure depend on the notion of **measurement scale**
  - mean
  - variance
  - correlation
  - statistical tests
  - ….

- ✦ Do IR measures **comply with** those assumptions?
- ✦ How much are (statistical) analyses impacted by **departures** from those assumptions?
- ✦ What is the **validity** of our experiments?

… But, Wait, Assumptions?

Measurement Scales

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (invariantive)
NOMINAL	Determination of equality	*Permutation group* $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
ORDINAL	Determination of greater or less	*Isotonic group* $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
INTERVAL	Determination of equality of intervals or differences	*General linear group* $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
RATIO	Determination of equality of ratios	*Similarity group* $x' = ax$	Coefficient of variation

- An **interval scale** is an **equi-spaced scale** where a **difference** of one unit has the **same meaning** all over the range

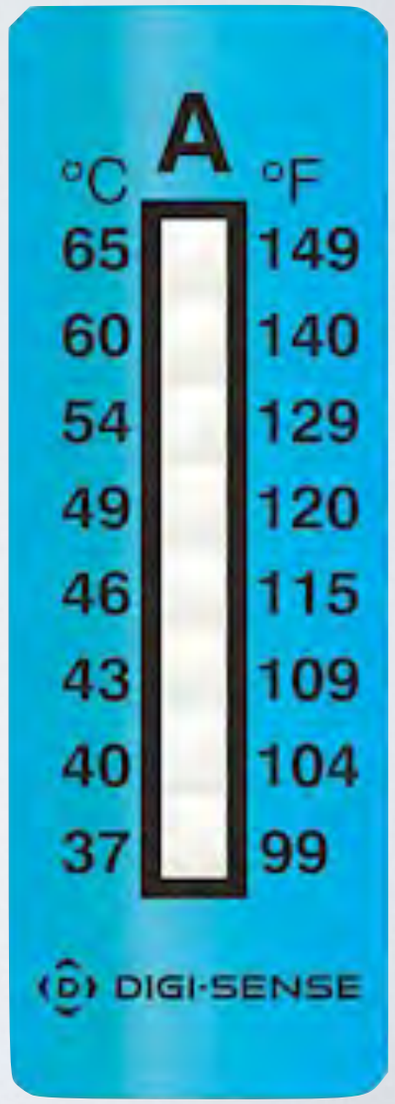
Stevens, S. S. (1946). On the Theory of Scales of Measurement. Science, *New Series*, 103(2684):677–680.

Temperature: An Interval Scale

- 20 °C is not twice as hot as 10 °C, i.e. multiplication and division are not allowed
  - Division is **not invariant** wrt the transformation

$$\frac{20\,^\circ\mathrm{C}}{10\,^\circ\mathrm{C}} = 2 \quad \text{but} \quad \frac{68\,^\circ\mathrm{F}}{50\,^\circ\mathrm{F}} = 1.36$$

$$F = \frac{9}{5}C + 32$$



- The increase between 10 °C and 20 °C is the same as the increase between 20 °C and 30 °C, i.e. addition and subtractions are allowed
  - Subtraction is **invariant** wrt the transformation

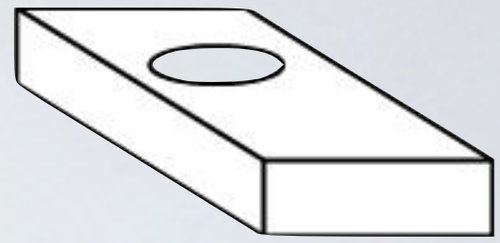
$$\begin{cases} 30\,^\circ\mathrm{C} - 20\,^\circ\mathrm{C} = 20\,^\circ\mathrm{C} - 10\,^\circ\mathrm{C} = 10\,^\circ\mathrm{C} \\ 86\,^\circ\mathrm{F} - 68\,^\circ\mathrm{F} = 68\,^\circ\mathrm{F} - 50\,^\circ\mathrm{F} = 18\,^\circ\mathrm{F} \end{cases}$$

- The ratio of intervals is **invariant** wrt to the transformation

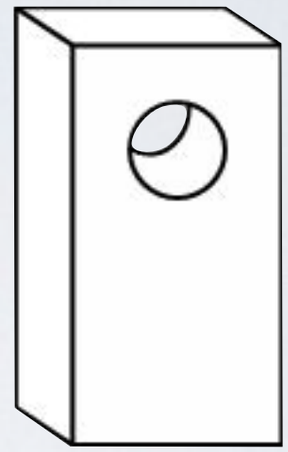
$$\frac{20\,^\circ\mathrm{C} - 10\,^\circ\mathrm{C}}{30\,^\circ\mathrm{C} - 20\,^\circ\mathrm{C}} = 1 \quad \text{and} \quad \frac{68\,^\circ\mathrm{F} - 50\,^\circ\mathrm{F}}{86\,^\circ\mathrm{F} - 68\,^\circ\mathrm{F}} = 1$$

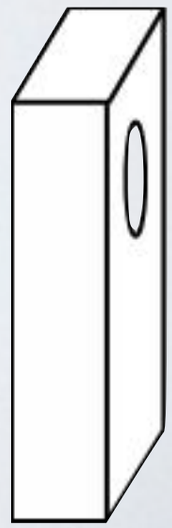
Meaningfulness

- Statistical operations on measurements of a given scale are not appropriate or inappropriate **per se** but only relative to the **kinds of statements** made about them.



- The criterion of appropriateness for a statement about a statistical operation is that the statement be empirically **meaningful** in the sense that its truth or falsity must be **invariant** under **permissible transformations** of the underlying scale





  - Meaningfulness is a distinct concept from the one of truth of a statement and it is somehow close to the notion of *invariance* in geometry

Adams, E. W., Fagot, R. F., and Robinson, R. E. (1965). A theory of appropriate statistics. *Psychometrika*, 30:99–127.

Temperature: Meaningfulness

$$T_P^C = [\ 2\ \ 2\ \ 4\ \ 8\ \ 36\ ]$$

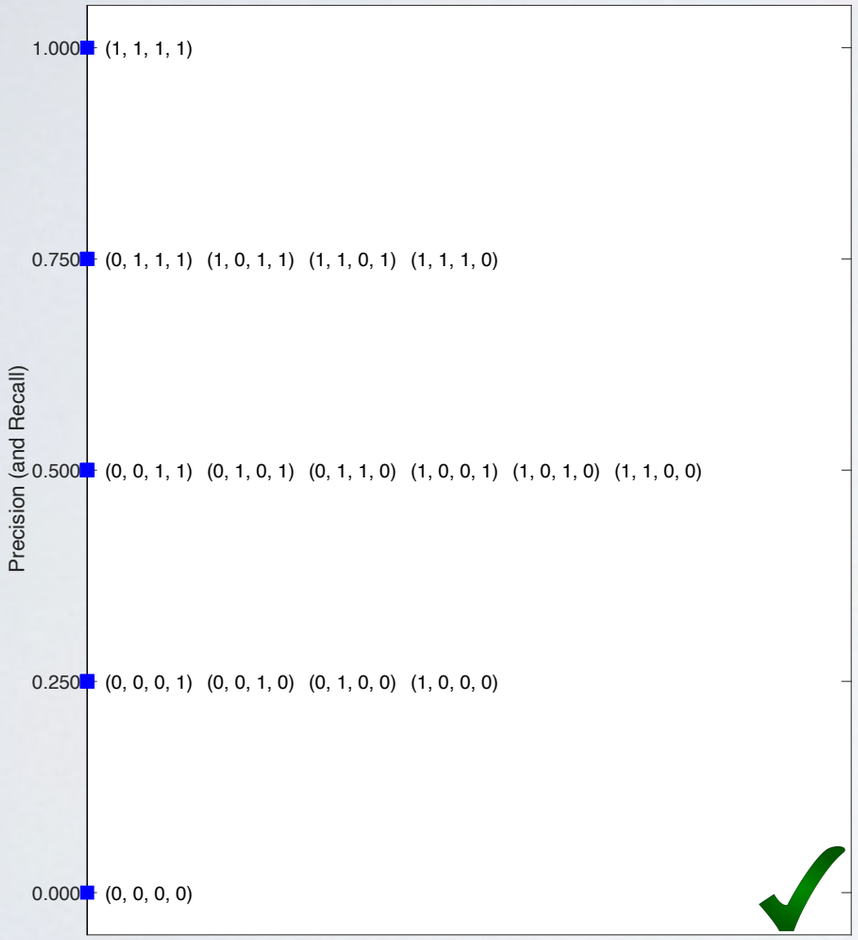
$$T_P^F = [\ 35.6\ \ 35.6\ \ 39.2\ \ 46.4\ \ 96.8\ ]$$

$$T_R^C = [\ 1\ \ 2\ \ 4\ \ 15\ \ 34\ ]$$

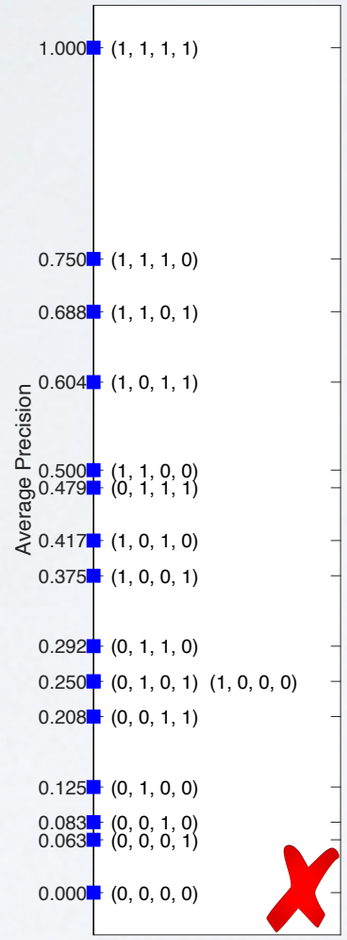
$$T_R^F = [\ 33.8\ \ 35.6\ \ 39.2\ \ 59.0\ \ 93.2\ ]$$

- "The median temperature in Paris is the same as in Rome" is **meaningful**, since 4 = 4 in Celsius degrees and 39.2 = 39.2 in Fahrenheit degrees
  - interval scales are also ordinal and quantiles are an allowable operation on ordinal scales
- "The mean temperature in Paris is less than in Rome" is **meaningful** as well, since 10.4 < 11.2 in Celsius degrees and 50.72 < 52.16 in Fahrenheit degrees
  - addition and subtraction are allowable operations on an interval scale and, as a consequence, mean is invariant to affine transformations
- "The geometric mean of temperature in Paris is greater than in Rome" is **not meaningful**, since 5.40 > 5.27 in Celsius degrees and 46.74 < 48.17 in Fahrenheit degrees
  - geometric mean involves the multiplication and division of values, which is not a permitted operation on an interval scale

How Do IR Measures Look Like?



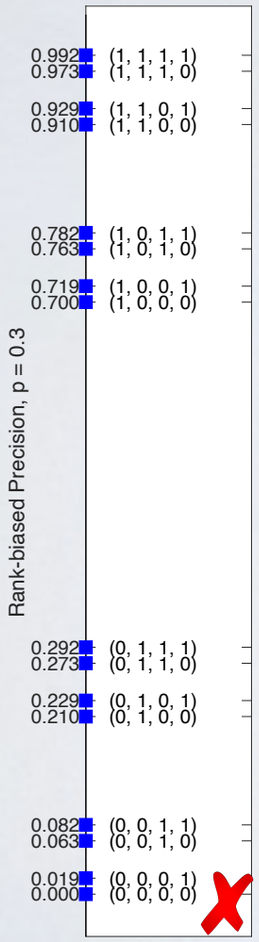
(a) Precision (and Recall).



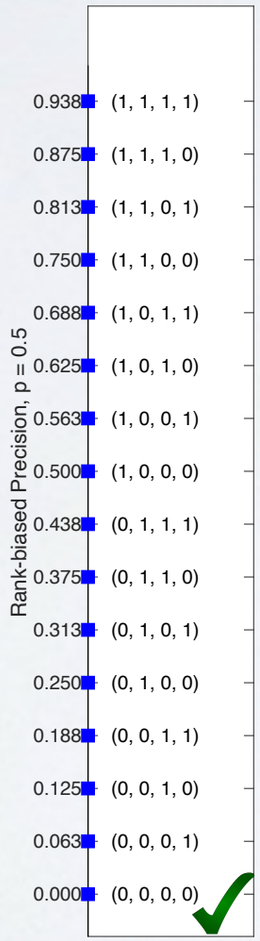
(b) AP.

(c) RR.

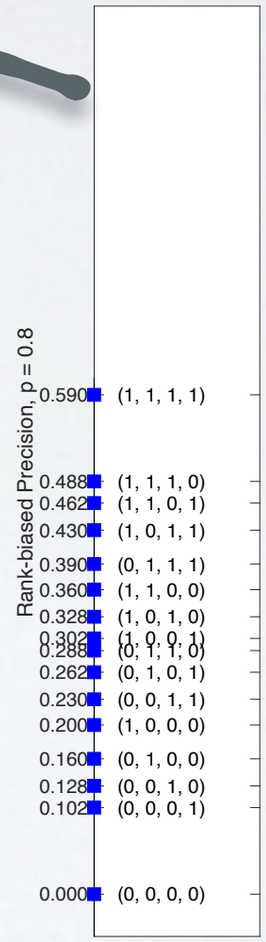
How Do IR Measures Look Like?



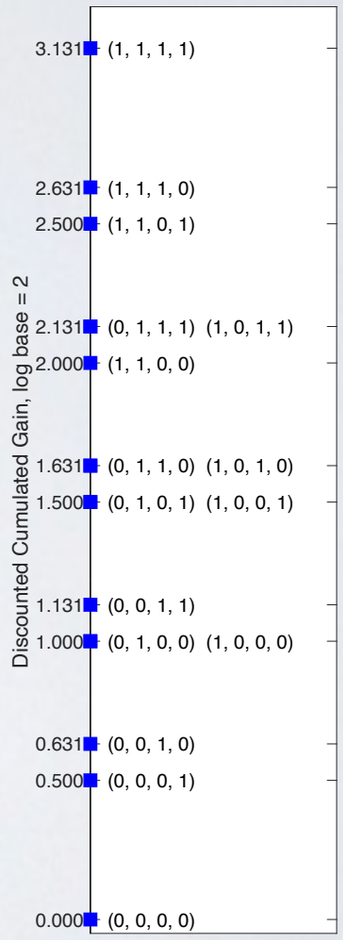
(d) RBP, $p = 0.3$.



(e) RBP, $p = 0.5$.



(f) RBP, $p = 0.8$.



(g) DCG, log base

Temperature: More on Meaningfulness

- We may be tempted to compare the results of the arithmetic mean with those of the geometric mean to gain “more **insights**”
- We might observe that the arithmetic mean in Paris is less than in Rome – 10.4 < 11.2 in Celsius degrees – but the opposite is true when we consider the geometric mean – 5.40 > 5.27 in Celsius degrees
  - We might thus highlight that this due to the fact that the first (and lowest) value 2 in Paris is double than 1 in Rome and that the **geometric mean rewards** gains at **lowest values**
  - On the other hand, the **arithmetic mean rewards** gains at **higher values** and thus 8 in Paris is (almost) half than 15 in Rome and it contributes less
- However, if we consider exactly the same temperatures just **on the Fahrenheit scale, we would reach opposite conclusions**

MAP vs GMAP: Robust Retrieval

Overview of the TREC 2004 Robust Retrieval Track

Ellen M. Voorhees
National Institute of Standards and Technology
Gaithersburg, MD 20899

Abstract

The robust retrieval track explores methods for improving the consistency of retrieval technology by focusing on poorly performing topics. The retrieval task in the track is a traditional ad hoc retrieval task where the evaluation methodology emphasizes a system's least effective topics. The most promising approach to improving poorly performing topics is exploiting text collections other than the target collection such as the web.

The 2004 edition of the track used 250 topics and required systems to rank the topics by predicted difficulty. The 250 topics within the test set allowed the stability of evaluation measures that emphasize poorly performing topics to be investigated. A new measure, a variant of the traditional MAP measure that uses a geometric mean rather than an arithmetic mean to average individual topic results, shows promise of giving appropriate emphasis to poorly performing topics while being more stable at equal topic set sizes.

The ability to return at least passable results for any topic is an important feature of an operational retrieval system. While system effectiveness is generally reported as average effectiveness, an individual user does not see the average performance of the system, but only the effectiveness of the system on his or her requests. A user whose request retrieves nothing of interest is unlikely to be consoled by the fact that the system responds better to other people's requests.

The TREC robust retrieval track was started in TREC 2003 to investigate methods for improving the consistency of retrieval technology. The first year of the track had two main technical results:

1. The track provided ample evidence that optimizing average effectiveness using the standard Cranfield methodology and standard evaluation measures further improves the effectiveness of the already-effective topics, sometimes at the expense of the poor performers.

2. The track results demonstrated that measuring poor performance is intrinsically difficult because there is so little signal in the sea of noise for a poorly performing topic. Two new measures devised to emphasize poor performers did so, but because there is so little information the measures are unstable. Having confidence in the conclusion that one system is better than another using these measures requires larger differences in scores than are generally observed in practice when using 50 topics.

The retrieval task in the track is a traditional ad hoc task. In addition to calculating scores using `trec_eval`, each run is also evaluated using the two measures introduced in the TREC 2003 track that focus more specifically on the least-well-performing topics. The TREC 2004 track differed from the initial track in two important ways. First, the test set of topics consisted of 249 topics, up from 100 topics. Second, systems were required to rank the *topics* by predicted difficulty, with the goal of eventually being able to use such predictions to do topic-specific processing.

This paper presents an overview of the results of the track. The first section describes the data used in the track, and the following section gives the retrieval results. Section 3 investigates how accurately systems can predict which topics are difficult. Since one of the main results of the TREC 2003 edition of the track was that the poor performance is hard to measure with 50 topics, section 4 examines the stability of the evaluation measures for larger topic set sizes. The final section looks at the future of the track.

1 The Robust Retrieval Task

As mentioned, the task within the robust retrieval track is a traditional ad hoc task. Since the TREC 2003 track had shown that 50 topics was not sufficient for a stable evaluation of poorly performing topics, the TREC 2004 track used

Robertson, S. E. (2006). On GMAP: and Other Transformations. CIKM 2006, pages 78–83.

Measurement Issues in IR?

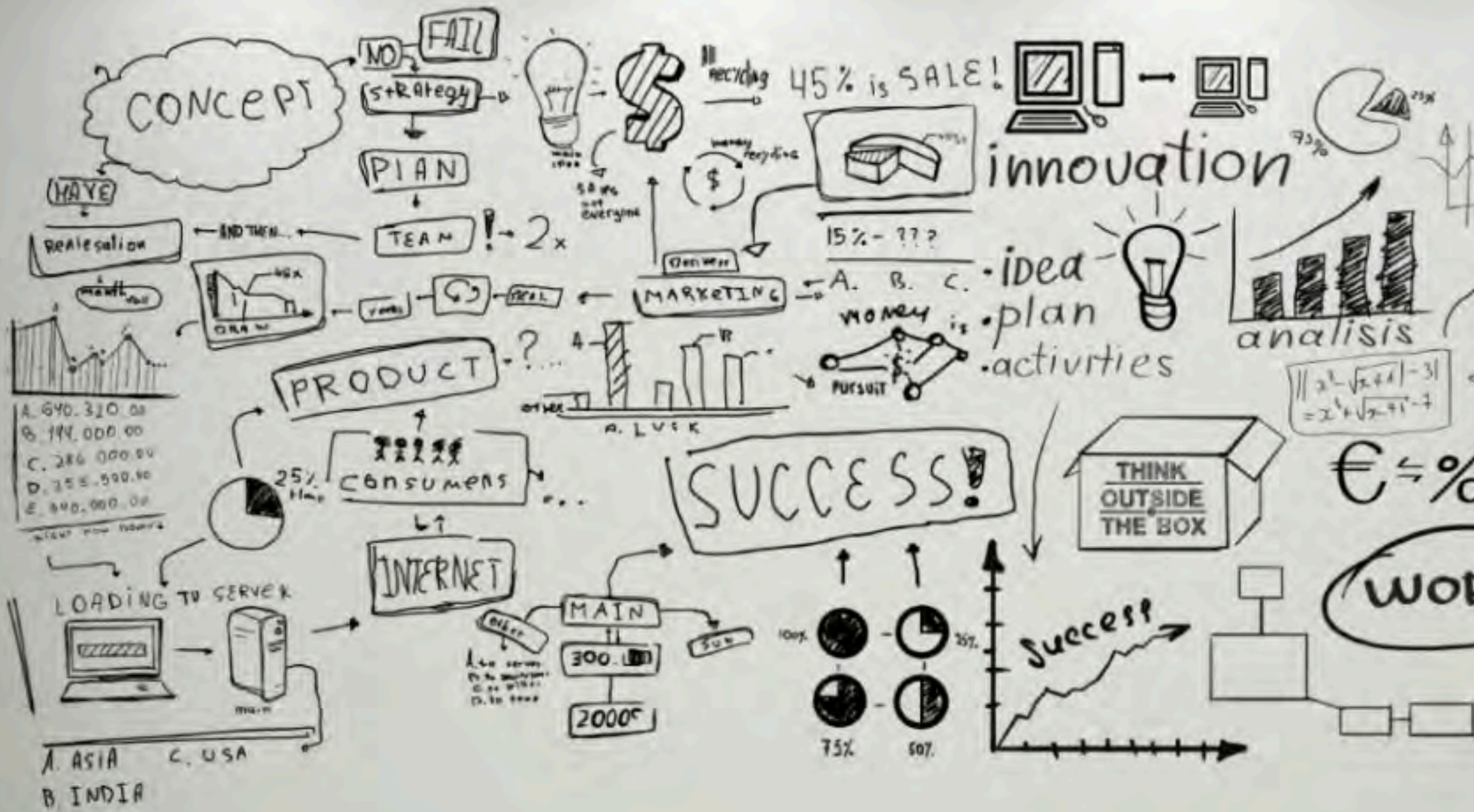
- Comparing system performance
- Topic difficulty and robust retrieval
- Score transformation and standardization techniques
- …

- Statistical significance testing
  - Sign Test: ordinal scale
  - Wilcoxon Rank Sum Test: ordinal scale
  - Wilcoxon Signed Rank Test: interval scale
  - Student's t Test: interval scale
  - ANOVA: interval scale
  - Kruskal-Wallis Test: ordinal scale
  - Friedman Test: ordinal scale

Fuhr, N. (2017). Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. *SIGIR Forum*, 51(3):32–41.

Sakai, T. (2020). On Fuhr's Guideline for IR Evaluation. *SIGIR Forum*, 54(1):p14:1–p14:8.

The Plan


CONCEPT
NO
FAIL
STRATEGY
PLAN
TEAM
2x
HAVE
REALESATION
AND THEN...
MARKETING
recycling
45% is SALE!
innovation
15% - ???
A. B. C.
idea
plan
activities
money is
PURSUIT
analisis
PRODUCT
?...
A. LUCK
25%
CONSUMERS
INTERNET
LOADING TO SERVER
A. ASIA
B. INDIA
C. USA
SUCCESS!
THINK OUTSIDE THE BOX
MAIN
300.
2000
Success!
75%
50%
€ ÷ %

Representational Theory of Measurement

- There exists an **empirical relation** $\succ$ which orders entities on the basis of their attributes
  - E.g., you can compare two rods and determine which is longer or whether they are equal
  - The empirical relation may support **concatenation** $\circ$
    - e.g., the concatenation of a rod with another one is longer than both of them
- There exists an **homomorphism** $\mathbf{M}$, the **measurement scale**, which maps entities into numbers and the empirical relation into a numerical relation which preserves the ordering (and concatenation)

$$(E, \succ, \circ) \xrightarrow{\mathbf{M}} (\mathbb{R}, >, +)$$

$$e_1 \succ e_2 \Rightarrow \mathbf{M}(e_1) > \mathbf{M}(e_2)$$

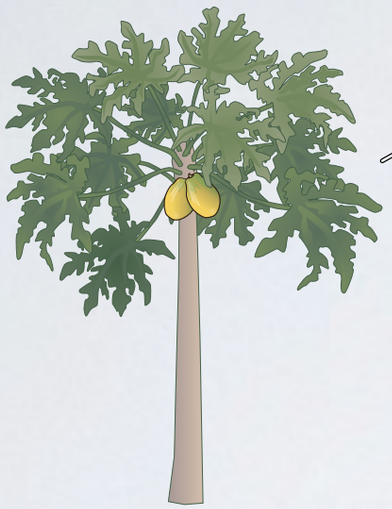
$$\mathbf{M}(e_1 \circ e_2) = \mathbf{M}(e_1) + \mathbf{M}(e_2)$$

Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A. (1971). *Foundations of Measurement. Additive and Polynomial Representations*, vol. 1. Academic Press, USA.

Fenton, N. E. and Bieman, J. (2014). *Software Metrics: A Rigorous & Practical Approach*. Chapman and Hall/CRC, USA, 3rd edition.

From Theory to Trees…

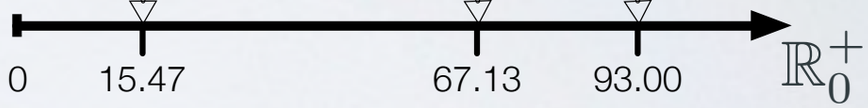
Real World $\mathrm{M}(\cdot)$ **Number System**



Papaya Tree

Banana Tree

Ananas Tree



Papaya Tree *taller than* Banana Tree

Papaya Tree *taller than* Ananas Tree

Banana Tree *taller than* Anans Tree

Papaya Tree *much taller than* Ananas Tree

Banana Tree *much taller than* Ananas Tree

$\mathrm{M}(\text{Papaya Tree}) > \mathrm{M}(\text{Banana Tree})$

$\mathrm{M}(\text{Papaya Tree}) > \mathrm{M}(\text{Ananas Tree})$

$\mathrm{M}(\text{Banana Tree}) > \mathrm{M}(\text{Ananas Tree})$

$\mathrm{M}(\text{Papaya Tree}) > \mathrm{M}(\text{Ananas Tree}) + 50$

$\mathrm{M}(\text{Banana Tree}) > \mathrm{M}(\text{Ananas Tree}) + 50$

Empirical Relations $\mathrm{M}_R(\cdot)$ **Mathematical Relations**

Measurement Issues in IR

"In the physical sciences there is usually an empirical ordering of the quantities we wish to measure [...] Such a situation does not hold for information retrieval. There is no empirical ordering for retrieval effectiveness and therefore any measure of retrieval effectiveness will by necessity be artificial"

C. J. "Keith" van Rijsbergen

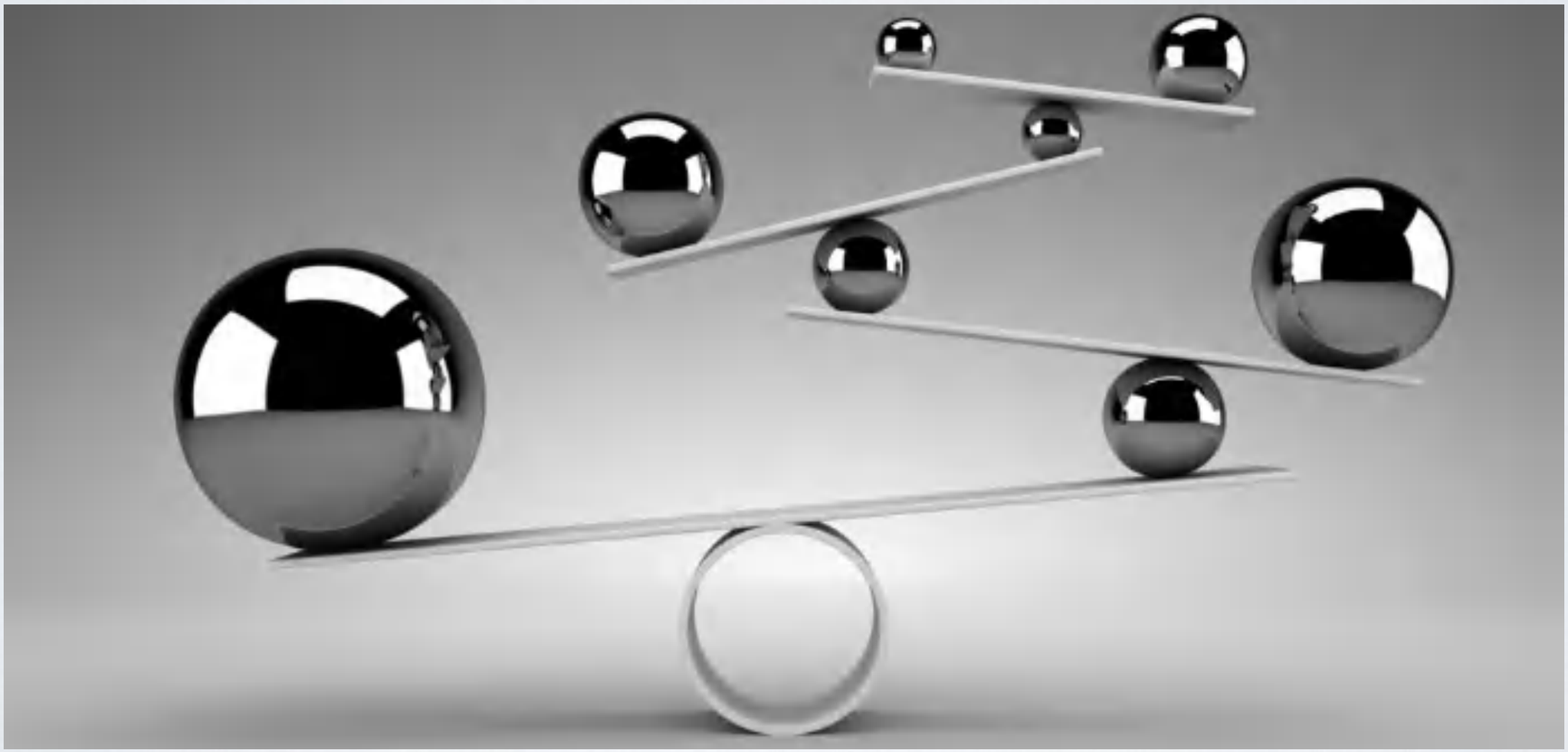
van Rijsbergen, C. J. (1981). Retrieval effectiveness. In Spärck Jones, K., editor, Information Retrieval Experiment, pages 32-43. Butterworths, United Kingdom.

Measurement Issues: Ordering?

HR
NR
NR
NR
...

≺ ? ≻

NR
R
NR
R
...



Ferrante, M., Ferro, N., and Maistro, M. (2015). Towards a Formal Framework for Utility-oriented Measurements of Retrieval Effectiveness. *ICTIR 2015*, pages 21–30.

More Measurement Issues: Intervals?

To determine whether a measure is interval-based

- We need to have a notion of interval among runs
- We need to have a notion of length of an interval among runs

Which runs fall in-between?

$\ell = 1, 2, 3, \ldots, ?$

Even More Measurement Issues: Transformations?

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (invariantive)
NOMINAL	Determination of equality	*Permutation group* $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
ORDINAL	Determination of greater or less	*Isotonic group* $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
INTERVAL	Determination of equality of intervals or differences	*General linear group* $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
RATIO	Determination of equality of ratios	*Similarity group* $x' = ax$	Coefficient of variation

Measures: Don’t Give Up!

“Measure what is measurable and make measurable what is not”

Galileo Galilei (1564-1642)

A General Theory of IR Measures

- Approach
  - Define what **orders** and **intervals** among runs are
  - Using orders and intervals of runs, introduce a proper **structure** among runs which allows us to define an **interval scale measure by construction**
  - Try to **transform** IR measures to the interval scale one and determine whether they are interval scales not
- <TL;DR>
  - Set-based measure are interval scales
  - Rank-based measures are interval scales only under very strict conditions, hardly met in practice
  - When you go multi-graded… be though!

Ferrante, M., Ferro, N., and Pontarollo, S. (2019). A General Theory of IR Evaluation Measures. *IEEE Transactions on Knowledge and Data Engineering (TKDE).* 31(3):409–422.

Set-based Measures: Findings

- Traditional set-based measures are all interval scales
- Precision

$$\mathrm{P}\left(\hat{r}\right) = \frac{1}{N}\sum_{i=1}^{N} \hat{r}_i = \frac{1}{N}\mathrm{SBTO}\left(\hat{r}\right)$$

- Recall

$$\mathrm{R}\left(\hat{r}\right) = \frac{1}{RB}\sum_{i=1}^{N} \hat{r}_i = \frac{1}{RB}\mathrm{SBTO}\left(\hat{r}\right)$$

- F-measure

$$\mathrm{F}\left(\hat{r}\right) = 2\frac{\mathrm{P}\left(\hat{r}\right)\cdot\mathrm{R}\left(\hat{r}\right)}{\mathrm{P}\left(\hat{r}\right)+\mathrm{P}\left(\hat{r}\right)} = \frac{2}{N+RB}\mathrm{SBTO}\left(\hat{r}\right)$$

Rank-based Measures: Findings

- RBP with $p \leq \frac{1}{2}$ is ordinal

- Only RBP with $p = \frac{1}{2}$ is interval

$$\mathrm{RBP}_{\frac{1}{2}} = \frac{1}{2} \sum_{i=1}^{N} \frac{1}{2^{i-1}} \hat{r}[i] = \frac{1}{2^N} \sum_{i=1}^{N} 2^{N-i} \hat{r}[i] = \frac{1}{2^N} \mathrm{RBTO}\left(\hat{r}\right)$$

- RBP with $p > \frac{1}{2}$ and AP are not even ordinal

Experiments

We Understood Everything, right? Significance Tests

Binary Relevance – T08, 8,256 system pairs compared		
Measure Pair	**Significantly Different Pairs**	
	Kruskal-Wallis Test	**ANOVA**
Precision	1,566 (18.97%)	2,785 (33.73%)
Recall	1,748 (21.17%)	3,259 (39.47%)
F-measure	1,721 (20.85%)	3,081 (37.32%)
SBTO	1,566 (18.97%)	2,785 (33.73%)

Binary Relevance – T08, 8,256 system pairs compared		
Measure Pair	**Significantly Different Pairs**	
	Kruskal-Wallis Test	**ANOVA**
RBP $p = 1/2$	1,677 (20.31%)	2,861 (34.65%)
RBP $p = 0.2$	1,675 (20.29%)	2,198 (26.62%)
RBP $p = 0.8$	1,783 (21.60%)	3,476 (42.10%)
AP	1,824 (22.09%)	3,320 (40.21%)
RBTO	1,677 (20.31%)	2,861 (34.65%)

Ferrante, M., Ferro, N., and Losiouk, E. (2020). How do interval scales help us with better understanding IR evaluation measures? *Information Retrieval Journal*, 23(3):289–317.

We Understood Everything, right? Significance Tests

Binary Relevance – T08, 8,256 system pairs compared		
Measure Pair	**Significantly Different Pairs**	
	Kruskal-Wallis Test	**ANOVA**
Precision	1,566 (18.97%)	2,785 (33.73%)
Recall	1,748 (21.17%)	3,259 (39.47%)
F-measure	1,721 (20.85%)	3,081 (37.32%)
SBTO	1,566 (18.97%)	2,785 (33.73%)

Binary Relevance – T08, 8,256 system pairs compared		
Measure Pair	**Significantly Different Pairs**	
	Kruskal-Wallis Test	**ANOVA**
RBP $p = 1/2$	1,677 (20.31%)	2,861 (34.65%)
RBP $p = 0.2$	1,675 (20.29%)	2,198 (26.62%)
RBP $p = 0.8$	1,783 (21.60%)	3,476 (42.10%)
AP	1,824 (22.09%)	3,320 (40.21%)
RBTO	1,677 (20.31%)	2,861 (34.65%)

Ferrante, M., Ferro, N., and Losiouk, E. (2020). How do interval scales help us with better understanding IR evaluation measures? *Information Retrieval Journal*, 23(3):289–317.

We Understood Everything, right? Significance Tests

Binary Relevance – T08, 8,256 system pairs compared		
Measure Pair	**Significantly Different Pairs**	
	Kruskal-Wallis Test	**ANOVA**
Precision	1,566 (18.97%)	2,785 (33.73%)
Recall	1,748 (21.17%)	3,259 (39.47%)
F-measure	1,721 (20.85%)	3,081 (37.32%)
SBTO	1,566 (18.97%)	2,785 (33.73%)

Binary Relevance – T08, 8,256 system pairs compared		
Measure Pair	**Significantly Different Pairs**	
	Kruskal-Wallis Test	**ANOVA**
RBP $p = 1/2$	1,677 (20.31%)	2,861 (34.65%)
RBP $p = 0.2$	1,675 (20.29%)	2,198 (26.62%)
RBP $p = 0.8$	1,783 (21.60%)	3,476 (42.10%)
AP	1,824 (22.09%)	3,320 (40.21%)
RBTO	1,677 (20.31%)	2,861 (34.65%)

Ferrante, M., Ferro, N., and Losiouk, E. (2020). How do interval scales help us with better understanding IR evaluation measures? *Information Retrieval Journal*, 23(3):289–317.

We Understood Everything, right? Significance Tests

Binary Relevance – T08, 8,256 system pairs compared		
Measure Pair	**Significantly Different Pairs**	
	Kruskal-Wallis Test	**ANOVA**
Precision	1,566 (18.97%)	2,785 (33.73%)
Recall	1,748 (21.17%)	3,259 (39.47%)
F-measure	1,721 (20.85%)	3,081 (37.32%)
SBTO	1,566 (18.97%)	2,785 (33.73%)

Binary Relevance – T08, 8,256 system pairs compared		
Measure Pair	**Significantly Different Pairs**	
	Kruskal-Wallis Test	**ANOVA**
RBP $p = 1/2$	1,677 (20.31%)	2,861 (34.65%)
RBP $p = 0.2$	1,675 (20.29%)	2,198 (26.62%)
RBP $p = 0.8$	1,783 (21.60%)	3,476 (42.10%)
AP	1,824 (22.09%)	3,320 (40.21%)
RBTO	1,677 (20.31%)	2,861 (34.65%)

Ferrante, M., Ferro, N., and Losiouk, E. (2020). How do interval scales help us with better understanding IR evaluation measures? *Information Retrieval Journal*, 23(3):289–317.

We Understood Everything, right? Kendall's Tau

Binary Relevance – T08		
Measure Pair		**Overall**
Precision vs SBTO		0.9998
Recall vs SBTO		0.8591
F-measure vs SBTO		0.9670
Precision vs Recall		0.8588

Binary Relevance – T08		
Measure Pair		**Overall**
RBP $p = 1/2$ vs RBTO		1.0000
RBP $p = 0.2$ vs RBTO		0.9225
RBP $p = 0.8$ vs RBTO		0.9043
AP vs RBTO		0.7439

We Understood Everything, right? Kendall's Tau

Binary Relevance – T08		
Measure Pair	**Topic-by-Topic**	**Overall**
Precision vs SBTO	1.0000	0.9998
Recall vs SBTO	1.0000	0.8591
F-measure vs SBTO	1.0000	0.9670
Precision vs Recall	1.0000	0.8588

Binary Relevance – T08		
Measure Pair	**Topic-by-Topic**	**Overall**
RBP $p = 1/2$ vs RBTO	1.0000	1.0000
RBP $p = 0.2$ vs RBTO	0.9985	0.9225
RBP $p = 0.8$ vs RBTO	0.8553	0.9043
AP vs RBTO	0.6099	0.7439

We Understood Everything, right? Kendall's Tau

Binary Relevance – T08		
Measure Pair	**Topic-by-Topic**	**Overall**
Precision vs SBTO	1.0000	0.9998
Recall vs SBTO	1.0000	0.8591
F-measure vs SBTO	1.0000	0.9670
Precision vs Recall	1.0000	0.8588

Binary Relevance – T08		
Measure Pair	**Topic-by-Topic**	**Overall**
RBP $p = 1/2$ vs RBTO	1.0000	1.0000
RBP $p = 0.2$ vs RBTO	0.9985	0.9225
RBP $p = 0.8$ vs RBTO	0.8553	0.9043
AP vs RBTO	0.6099	0.7439

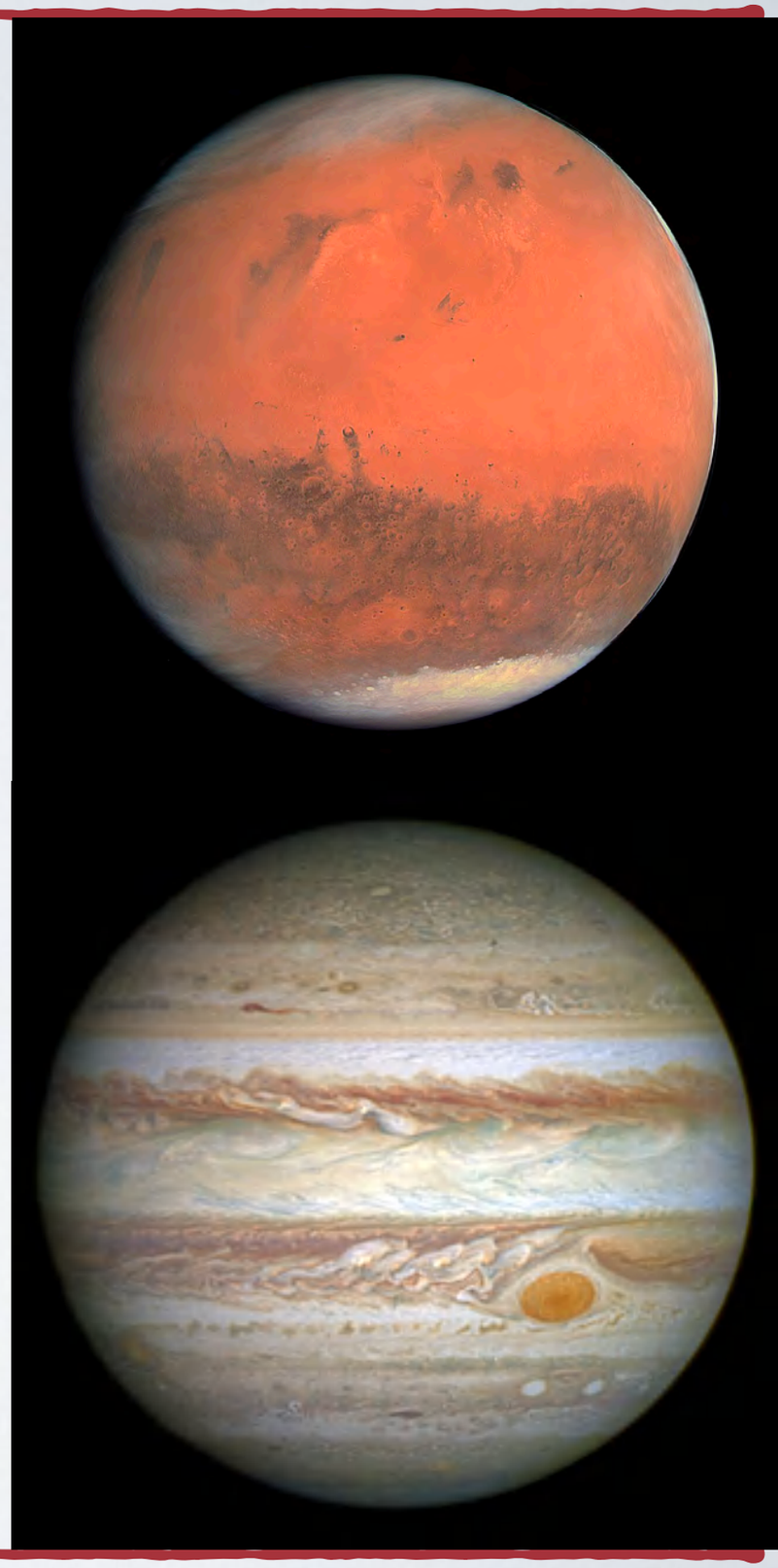
Worser and Worser Measurement Issues

- We started with a concern on whether IR measures should be **interval scales** or not
  - Whatever your stance on this, it matters since it **changes the number of SSD pairs**
- Independently from being interval scales or not, we found that **length** of the run and, especially, **recall base change the scale** you use from topic to topic
  - When you are aggregating across topics, e.g. **averaging**, you are mixing numbers from different scales
  - We can control the run length but not the recall base

Ferrante, M., Ferro, N., and Fuhr, N. (2021). Towards Meaningful Statements in IR Evaluation. Mapping Evaluation Measures to Interval Scales. *IEEE Access*, 9: 136182-136216

Precision and Recall: An Intuition

$$P(k) = \frac{1}{k} \sum_{n=1}^{k} r_n \qquad R(k) = \frac{1}{RB} \sum_{n=1}^{k} r_n$$

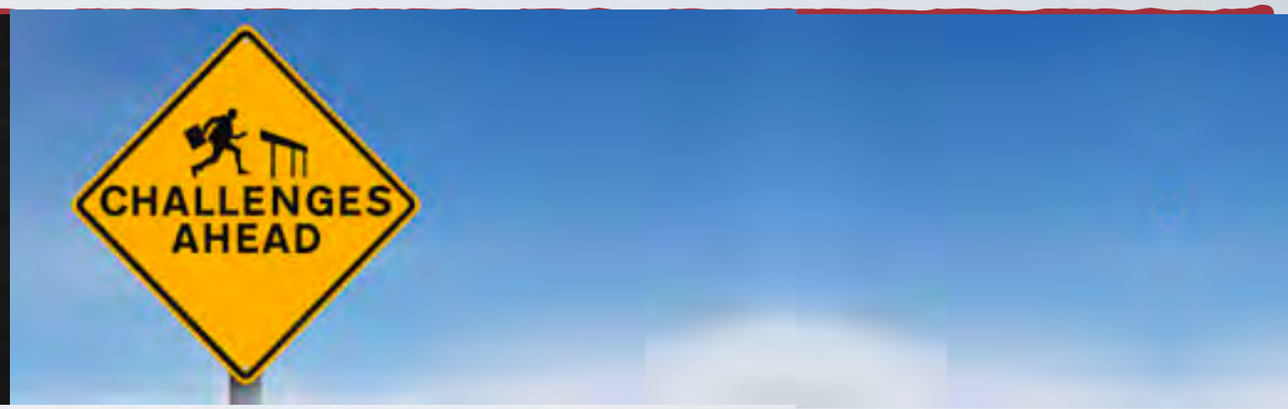


- Topics are like planets
- Precision is (somehow) like mass
- Recall is (somehow) like weight
- One the same planet, mass and weight order bodies in the same way
  However…
- You could average mass across planets but not weight

Lessons Learnt

- It is possible to develop a theory of IR measures grounded in the representational theory of measurement
- We determined the scale properties of several state-of-the-art IR measures
  - issues with intervalness but even more with recall base and run length
- Experimental results agree with the expected properties of the measures
  - you need to deep dive to really understand the behavior
  - sizeable impact on both correlation analysis and statistical significance testing

Lessons Learnt



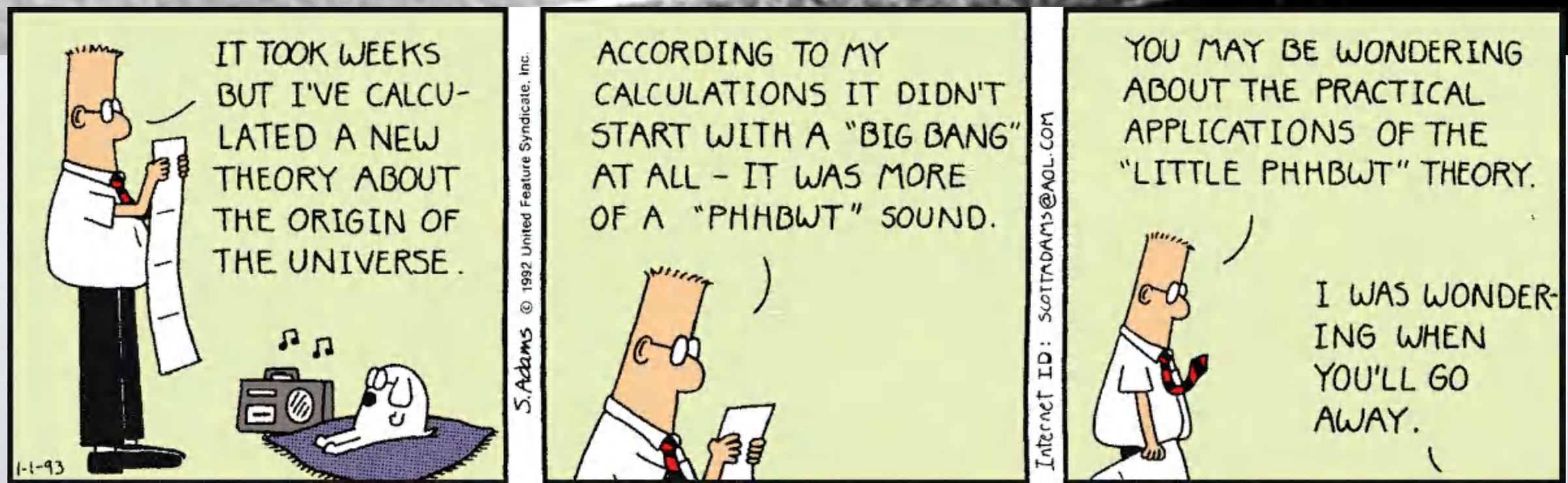
- It is possible to develop a theory of IR measures grounded in the representational theory of measurement
- We determined the scale properties of several state-of-the-art IR measures
  - issues with intervalness but even more with recall base and run length
- Experimental results agree with the expected properties of the measures
  - you need to deep dive to really understand the behavior
  - sizeable impact on both correlation analysis and statistical significance testing

- ✦ We need to **rethink** how we **use** our analytical tools and how we **explain** their outcomes
- ✦ What is the **validity** of our experiments is still an open question
- ✦ **Meaningfulness** should be a central concern in IR

No Shortcuts

No Shortcuts

Questions?


IT TOOK WEEKS BUT I'VE CALCU-LATED A NEW THEORY ABOUT THE ORIGIN OF THE UNIVERSE.
1-1-93
S. Adams © 1992 United Feature Syndicate, Inc.
ACCORDING TO MY CALCULATIONS IT DIDN'T START WITH A "BIG BANG" AT ALL - IT WAS MORE OF A "PHHBWT" SOUND.
Internet ID: SCOTTADAMS@AOL.COM
YOU MAY BE WONDERING ABOUT THE PRACTICAL APPLICATIONS OF THE "LITTLE PHHBWT" THEORY.
I WAS WONDER-ING WHEN YOU'LL GO AWAY.

stions?


IT TOOK WEEKS BUT I'VE CALCULATED A NEW THEORY ABOUT THE ORIGIN OF THE UNIVERSE.
ACCORDING TO MY CALCULATIONS IT DIDN'T START WITH A BIG BANG AT ALL - IT WAS MORE OF A "PHHBWT" SOUND
YOU MAY BE WONDERING ABOUT THE PRACTICAL APPLICATIONS OF THE "LITTLE PHHBWT" THEORY
I WAS WONDERING WHEN YOU'D GO AWAY.
SCIENCE © 1992 United Feature Syndicate, Inc.
INTERNET IDEA SCAVENGER

Reproducibility

Reproducibility: Why?

What we find reported in papers…

Reproducibility: Why?

What we find reported in papers…

… what happens to us

https://twitter.com/DevilleSy/status/958761021421903872

Reproducibility: How?

Everybody likes reproducibility…

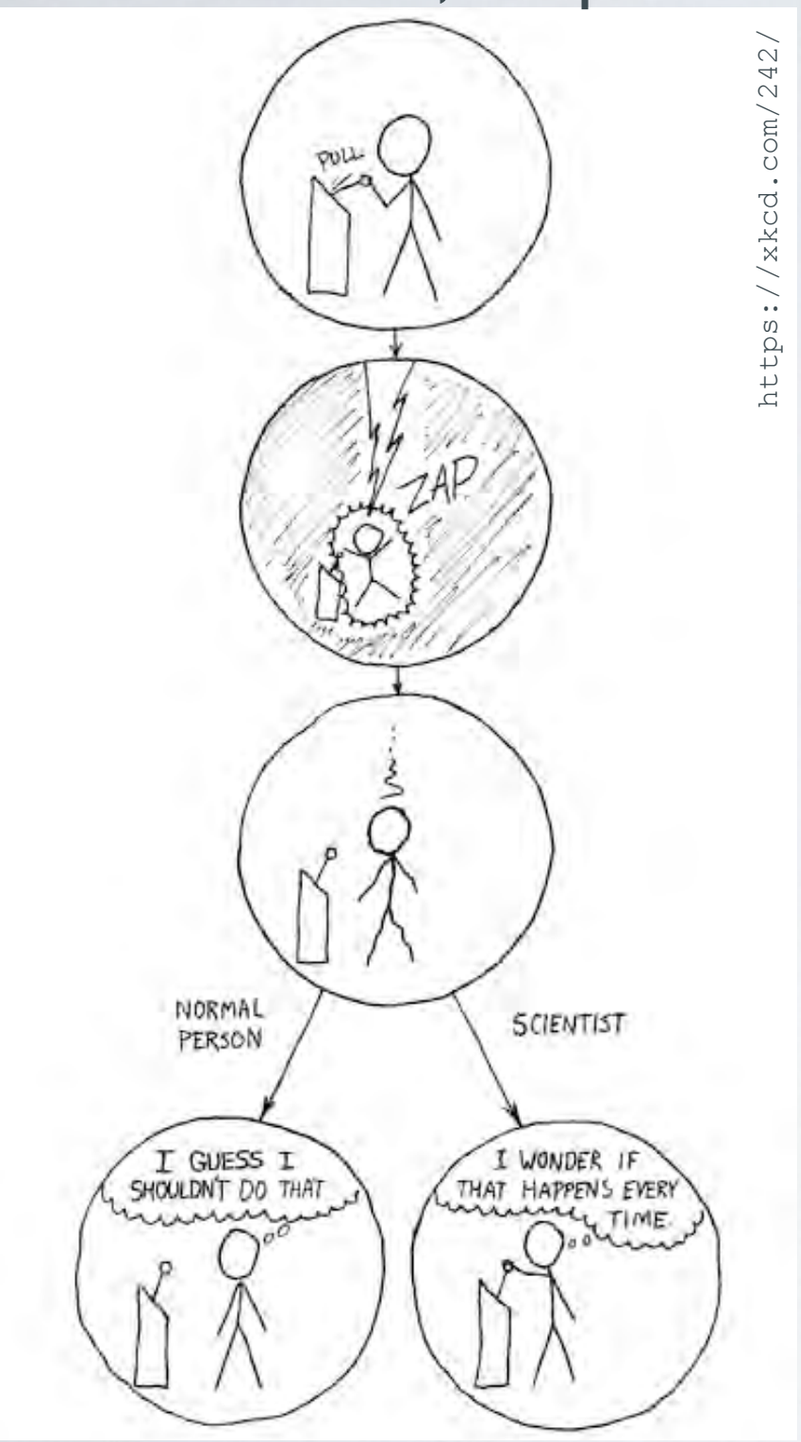
Reproducibility: How?

Everybody likes reproducibility…

…as soon as someone else does it

Reproducibility: What?

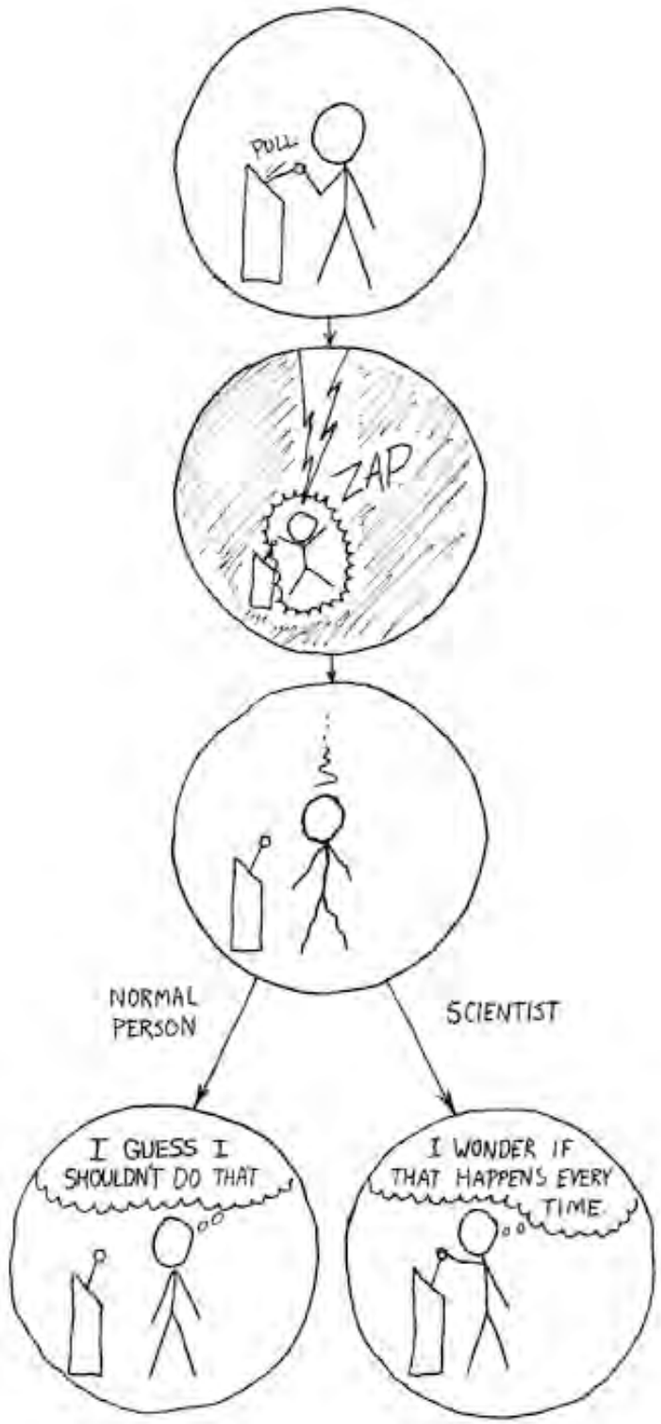
We know, reproducibly is at the core of science…



https://xkcd.com/242/

Reproducibility: What?

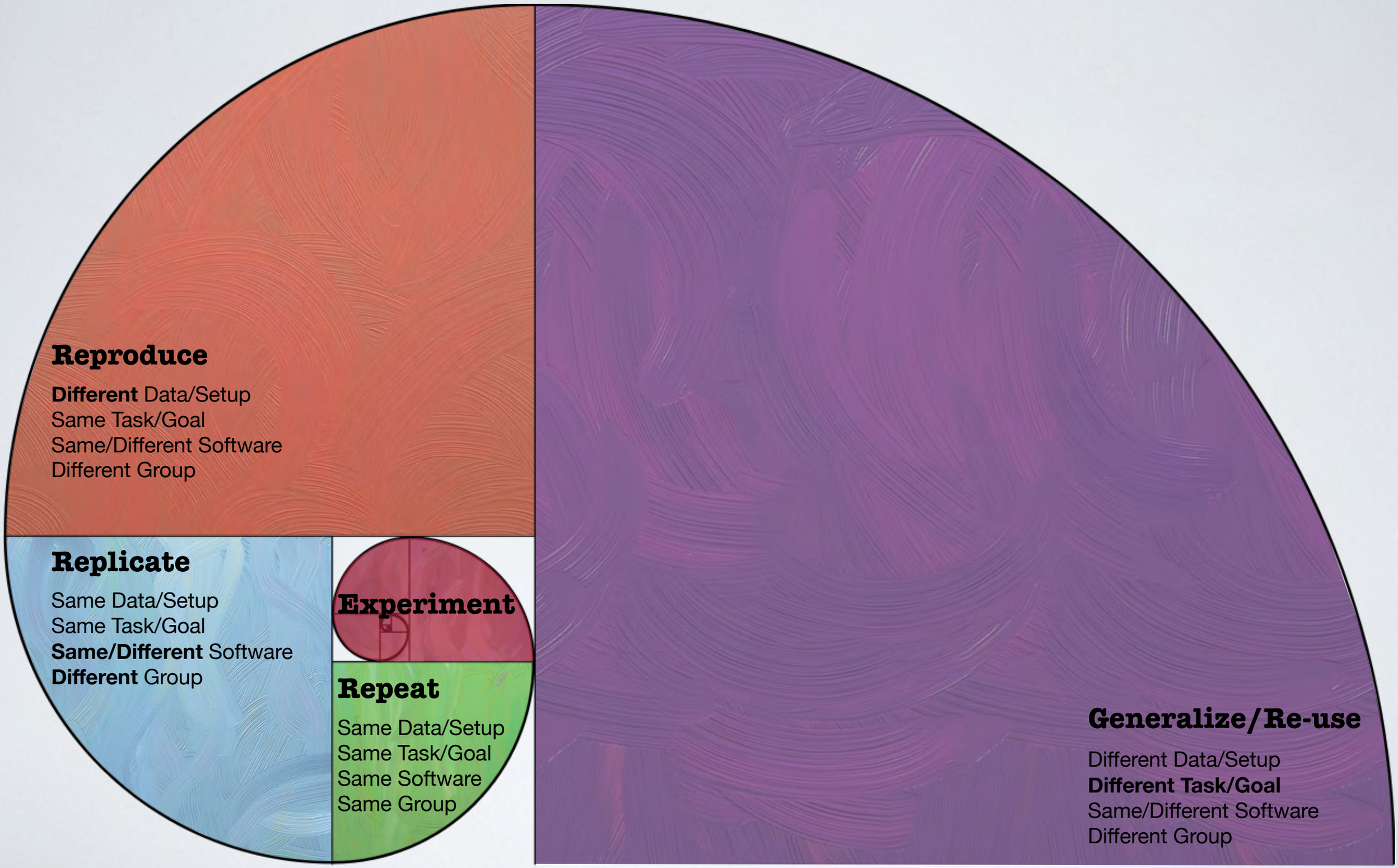
We know, reproducibly is at the core of science…



https://xkcd.com/242/

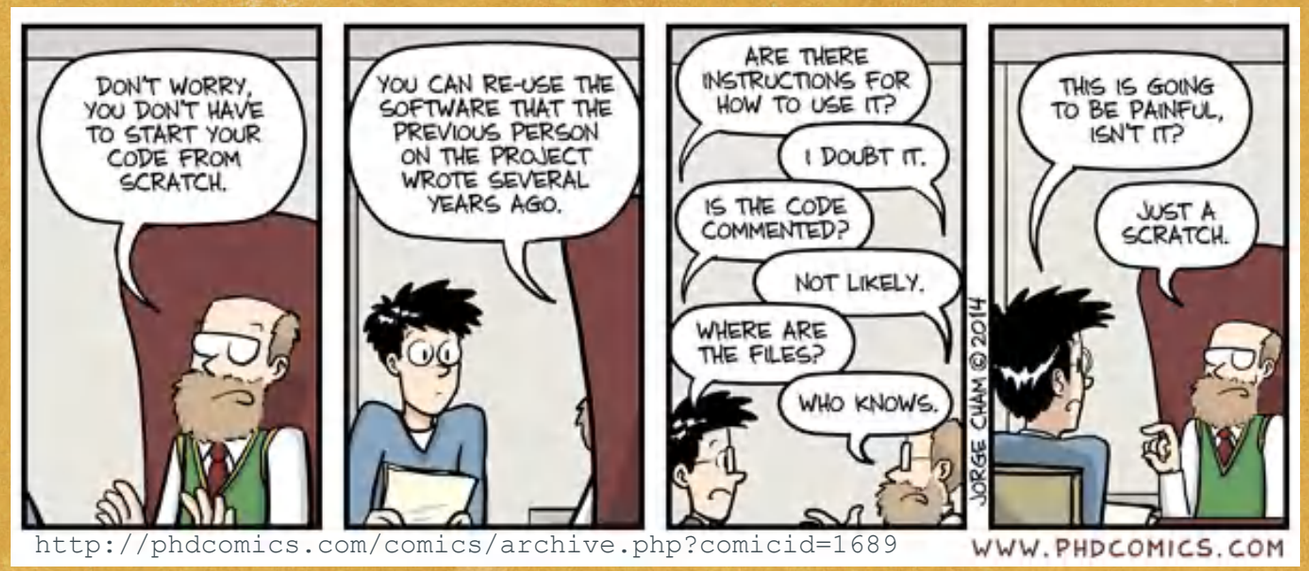
…but reproducing research is not new research

The “Reproducibility” Nautilus

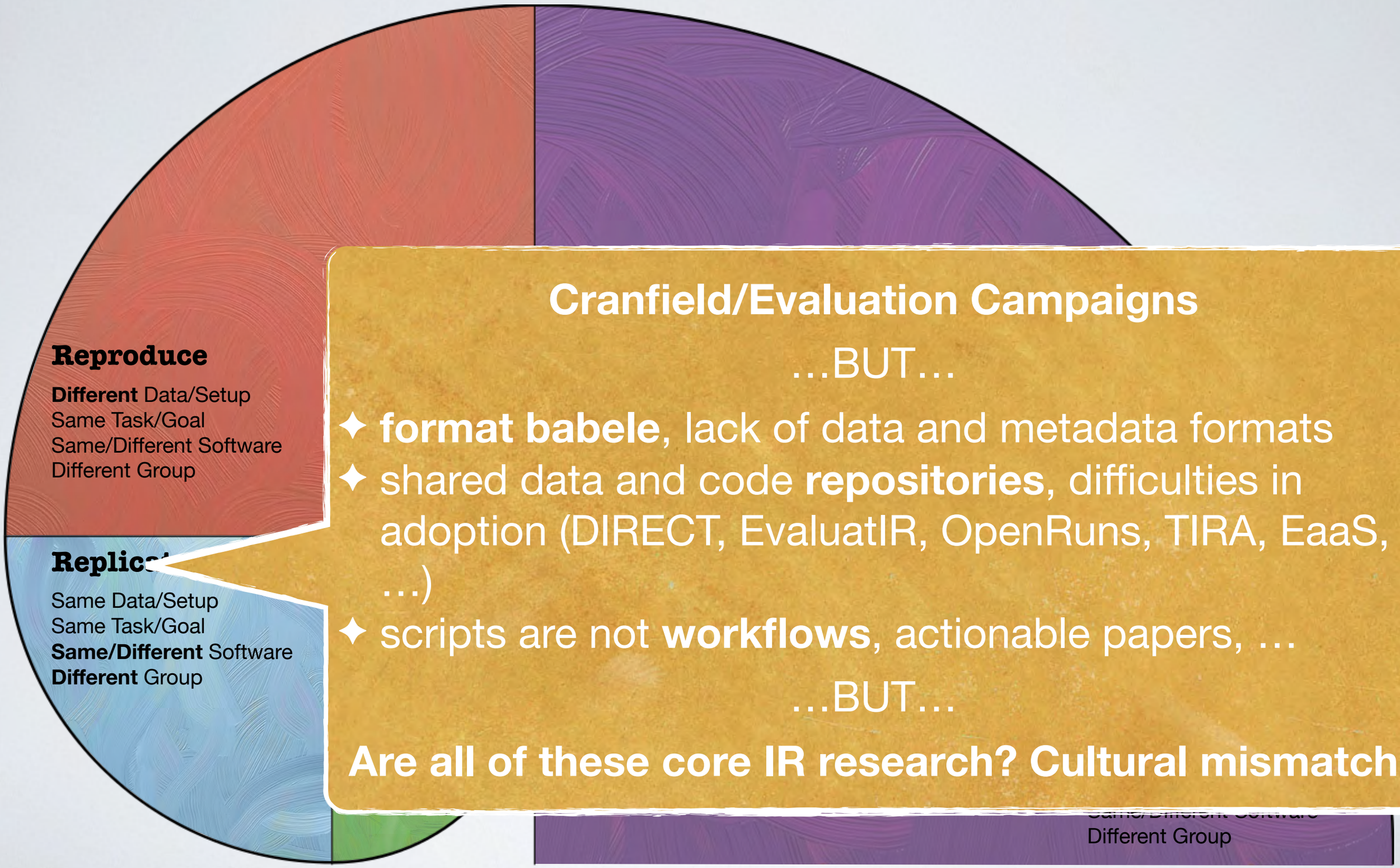


The “Reproducibility” Nautilus

Is it really that easy?



The “Reproducibility” Nautilus



Cranfield/Evaluation Campaigns

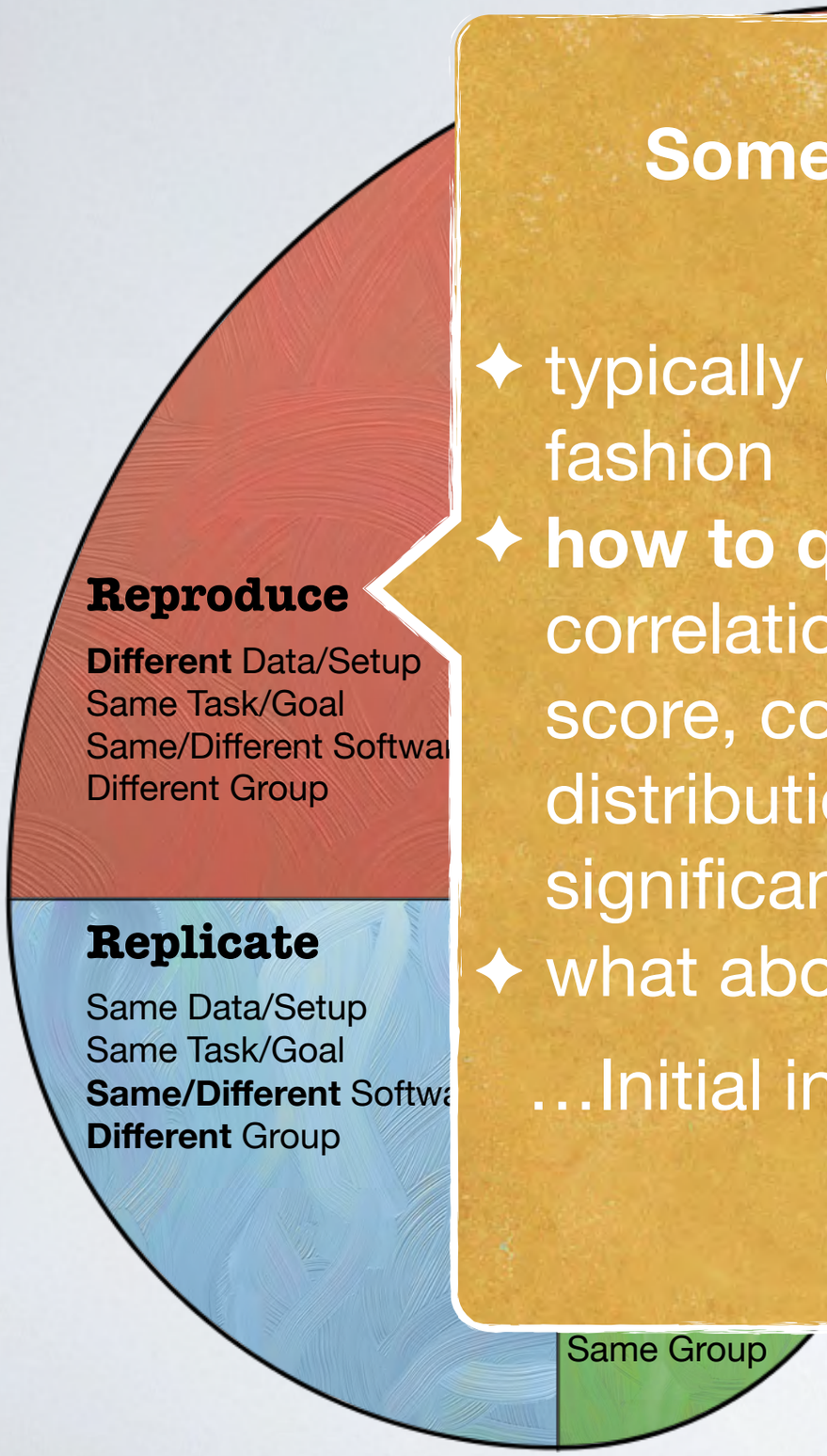
…BUT…

- **format babele**, lack of data and metadata formats
- shared data and code **repositories**, difficulties in adoption (DIRECT, EvaluatIR, OpenRuns, TIRA, EaaS, …)
- scripts are not **workflows**, actionable papers, …

…BUT…

Are all of these core IR research? Cultural mismatch

The “Reproducibility” Nautilus



Somehow standard approach in IR evaluation

…BUT…

- typically done **with-in group**, in a repeatability-like fashion
- **how to quantify when “reproduced”** - same ranked list, correlation among ranked lists, same effectiveness score, confidence bounds on effectiveness score, close distributions of effectiveness score, similar statistical significance (p-values, effect sizes, …), …
- what about the **user-side** 😱?

…Initial investigation in CENTRE@CLEF/NTCIR/TREC…

http://www.centre-eval.org/

The “Reproducibility” Nautilus

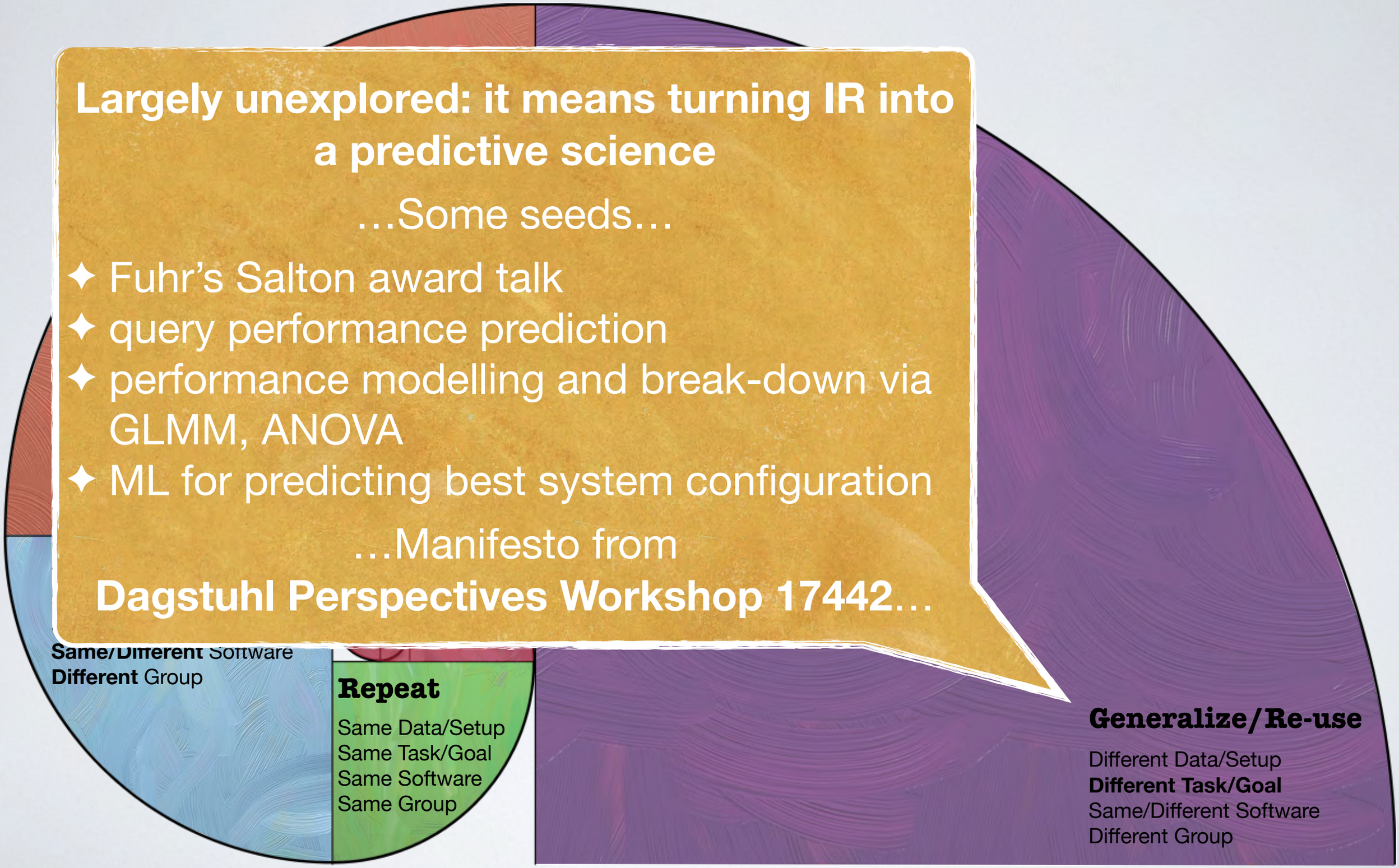
Largely unexplored: it means turning IR into a predictive science

…Some seeds…

- ✦ Fuhr’s Salton award talk
- ✦ query performance prediction
- ✦ performance modelling and break-down via GLMM, ANOVA
- ✦ ML for predicting best system configuration

…Manifesto from

Dagstuhl Perspectives Workshop 17442…

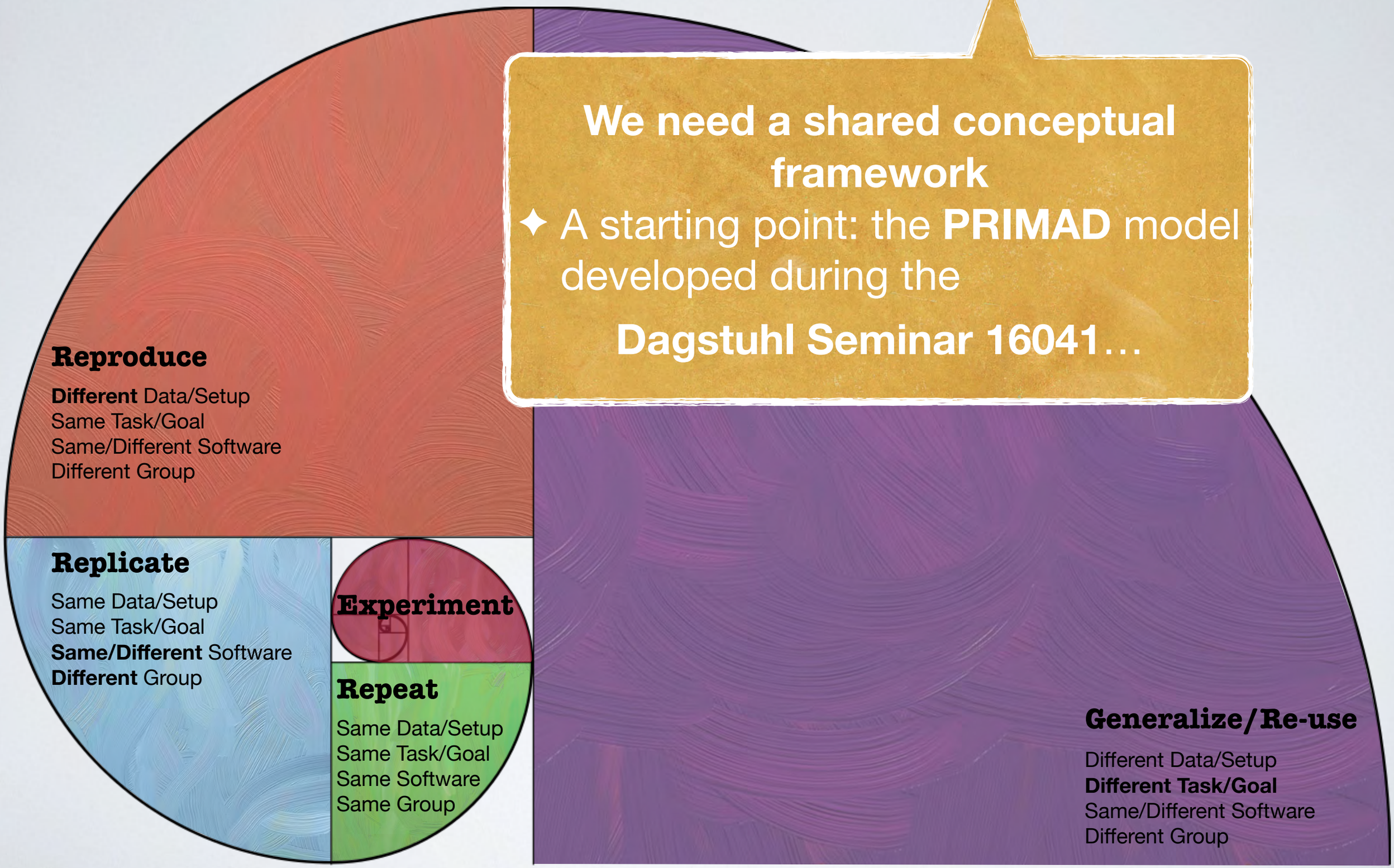


The “Reproducibility” Nautilus

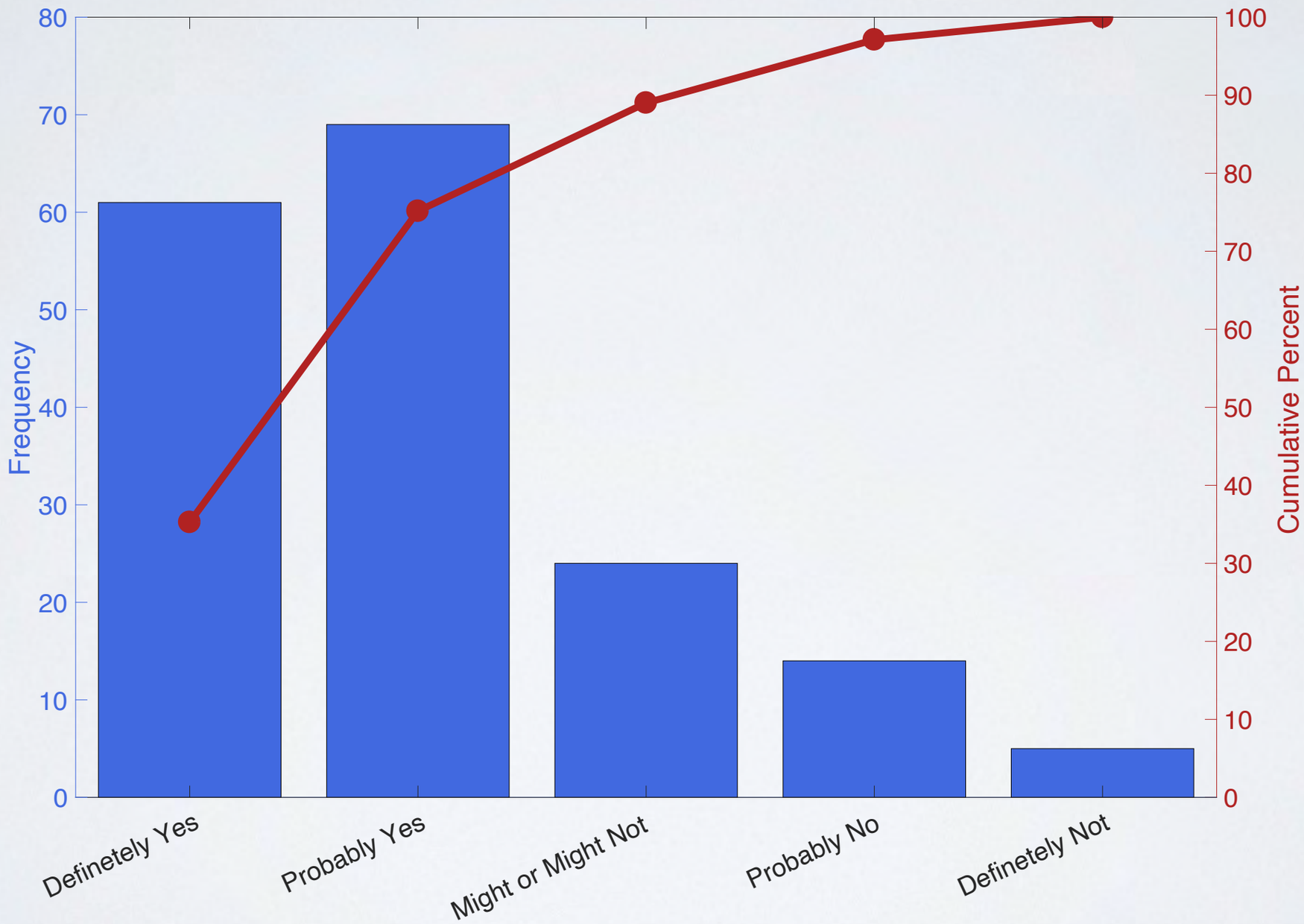
We need a shared conceptual framework

- A starting point: the **PRIMAD** model developed during the

Dagstuhl Seminar 16041…



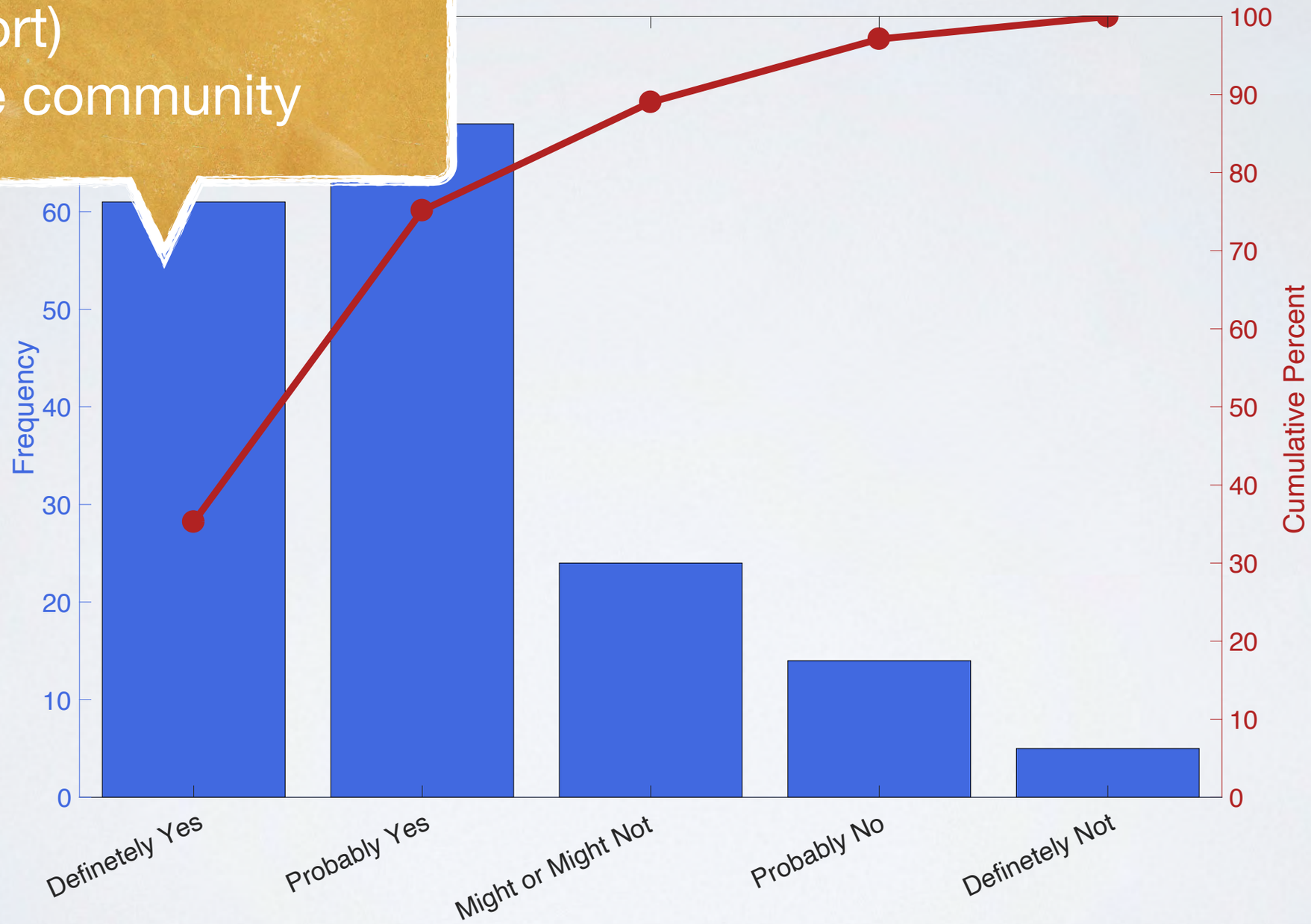
What about Introducing Badges?



Ferro, N. and Kelly, D. (2018). SIGIR Initiative to Implement ACM Artifact Review and Badging. *SIGIR Forum*, 52(1):4–10.

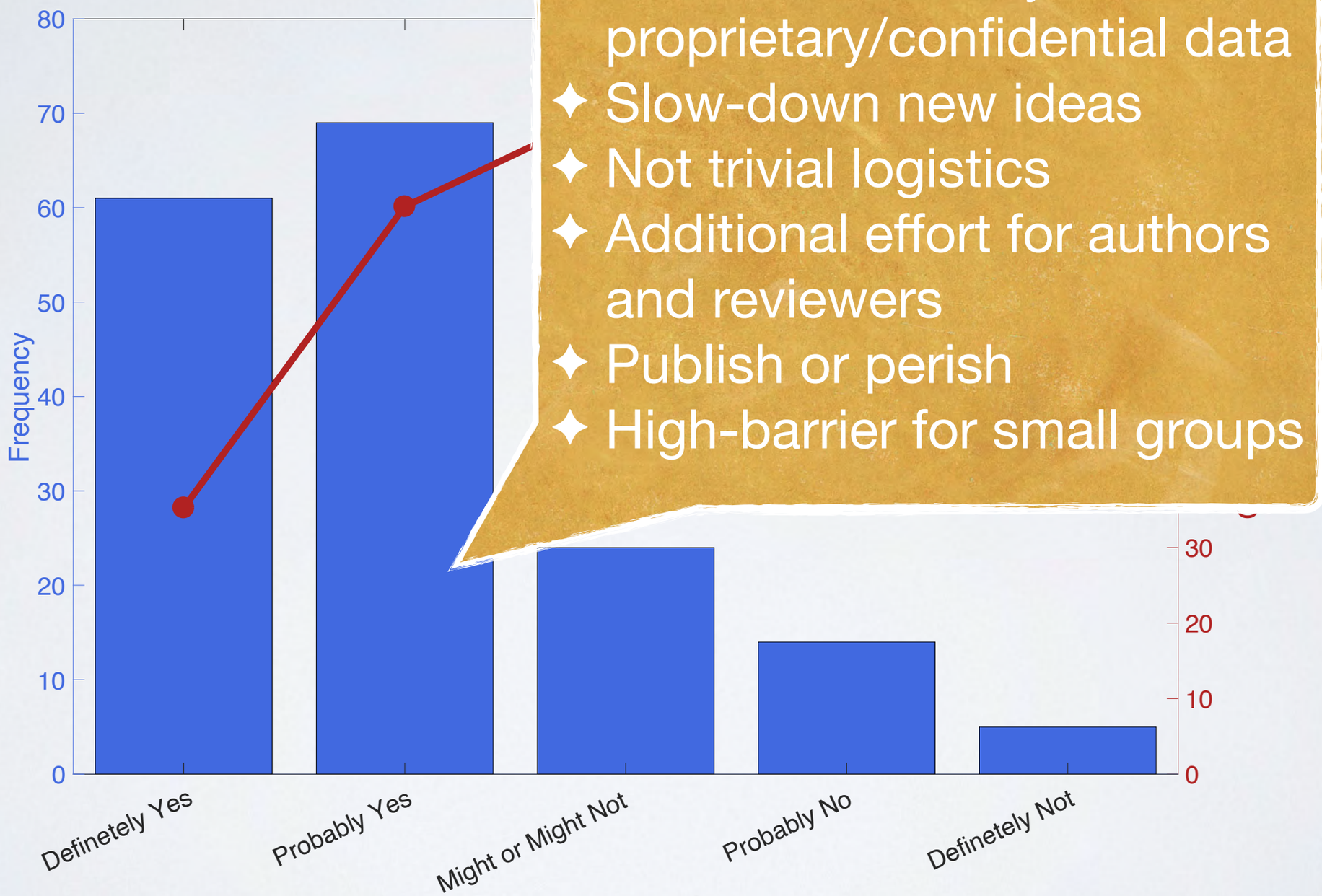
What about Introducing Badges?

- Better science
- Better baselines (and less effort)
- Improve community



Ferro, N. and Kelly, D. (2018). SIGIR Initiative to Implement ACM Artifact Review and Badging. *SIGIR Forum*, 52(1):4–10.

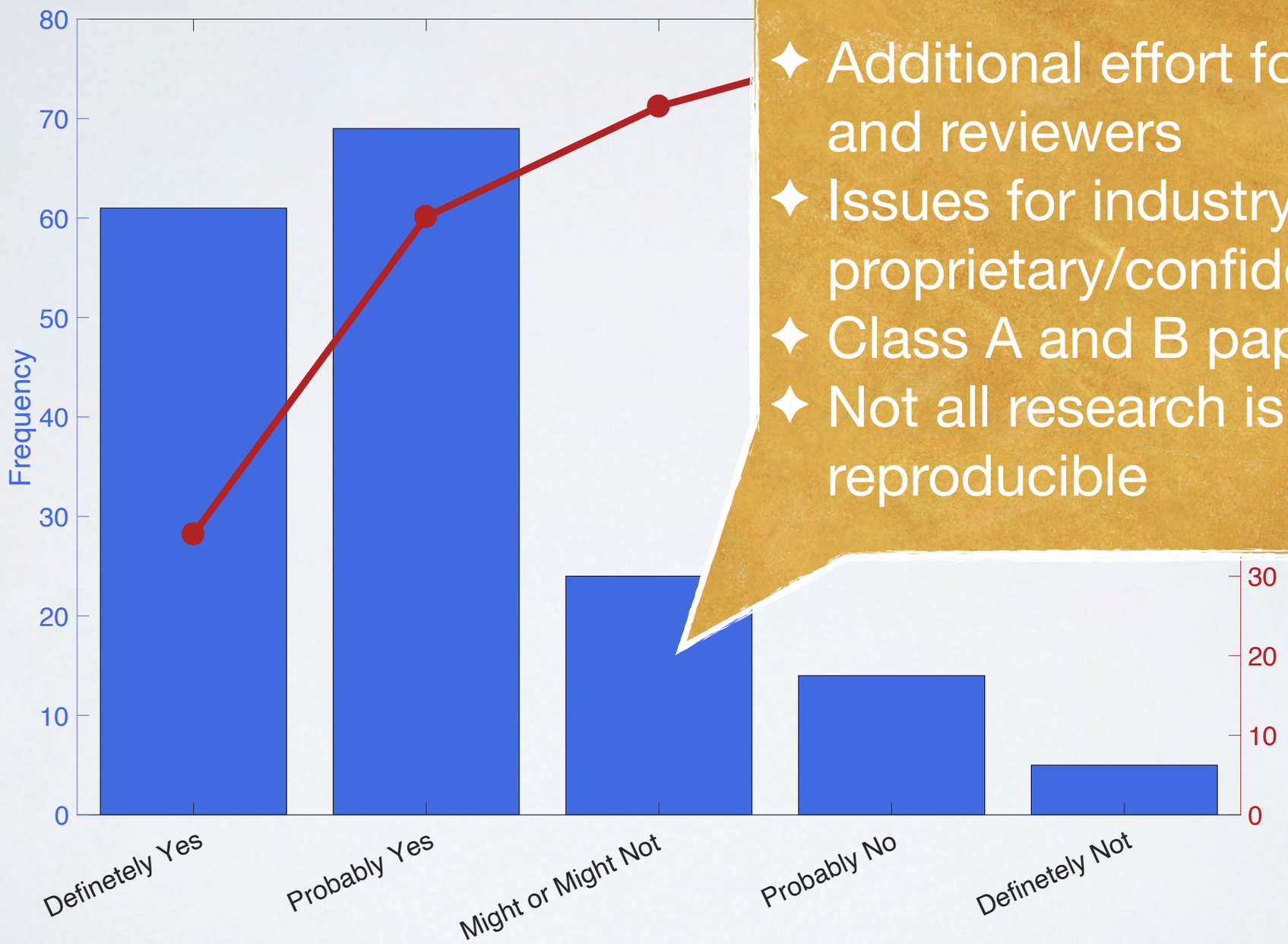
What about Introducing Badges?



- Better science
- Issues for industry, proprietary/confidential data
- Slow-down new ideas
- Not trivial logistics
- Additional effort for authors and reviewers
- Publish or perish
- High-barrier for small groups

Ferro, N. and Kelly, D. (2018). SIGIR Initiative to Implement ACM Artifact Review and Badging. *SIGIR Forum*, 52(1):4–10.

What about Introducing Badges?

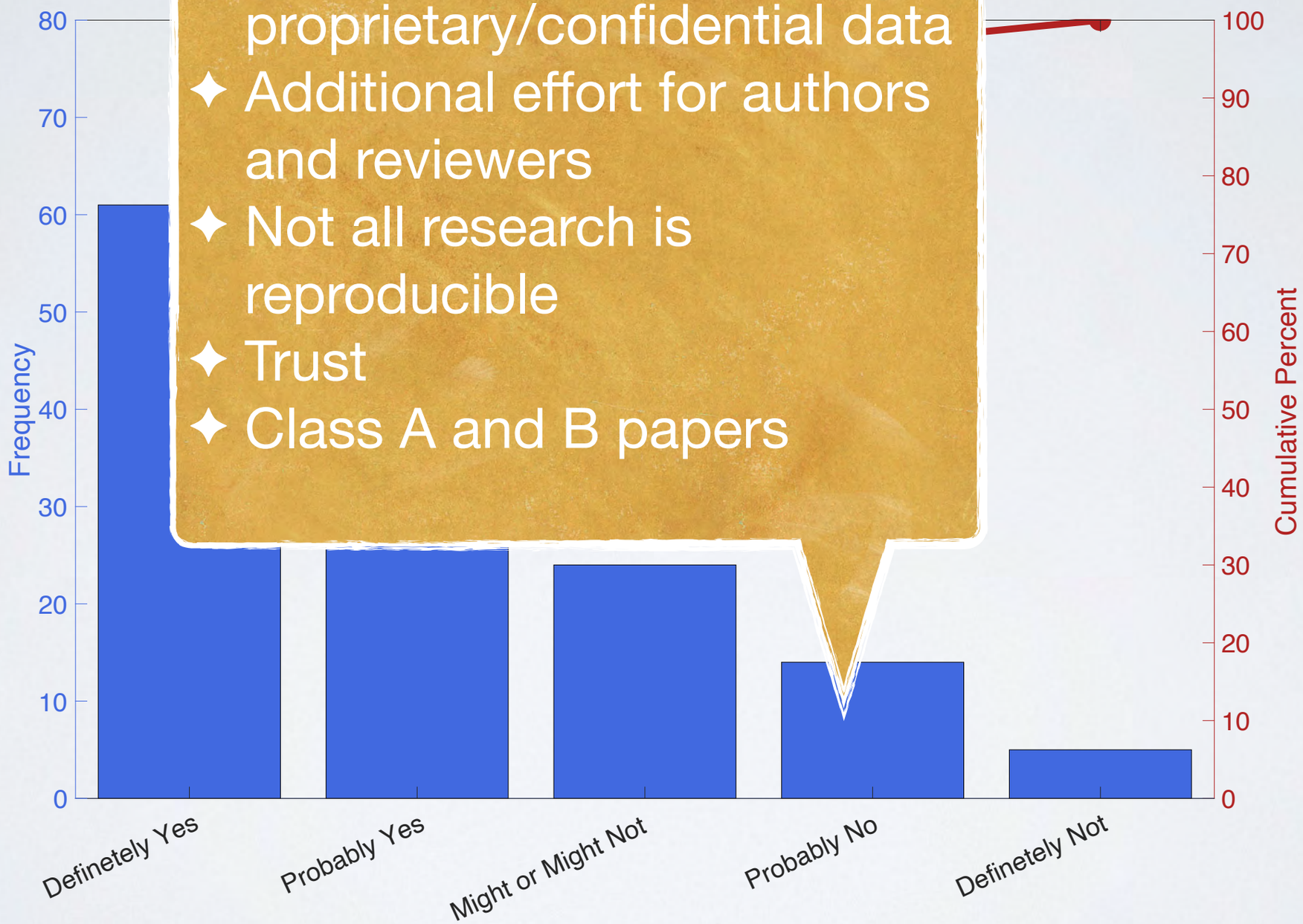


- ✦ Additional effort for authors and reviewers
- ✦ Issues for industry, proprietary/confidential data
- ✦ Class A and B papers
- ✦ Not all research is reproducible

Ferro, N. and Kelly, D. (2018). SIGIR Initiative to Implement ACM Artifact Review and Badging. *SIGIR Forum*, 52(1):4–10.

What about Introducing Badges?

- Issues for industry, proprietary/confidential data
- Additional effort for authors and reviewers
- Not all research is reproducible
- Trust
- Class A and B papers

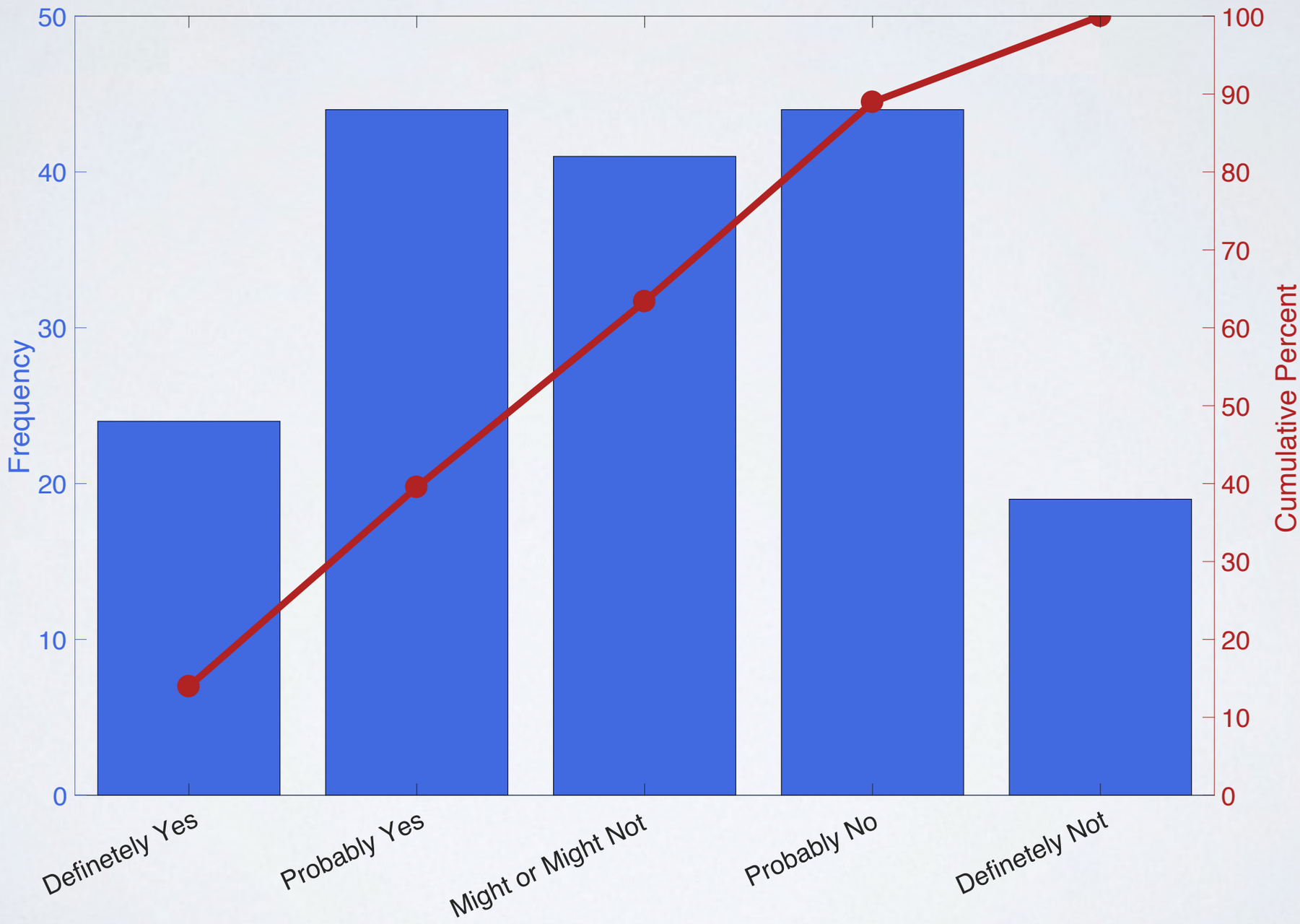


Ferro, N. and Kelly, D. (2018). SIGIR Initiative to Implement ACM Artifact Review and Badging. *SIGIR Forum*, 52(1):4–10.

What about Introducing Badges?

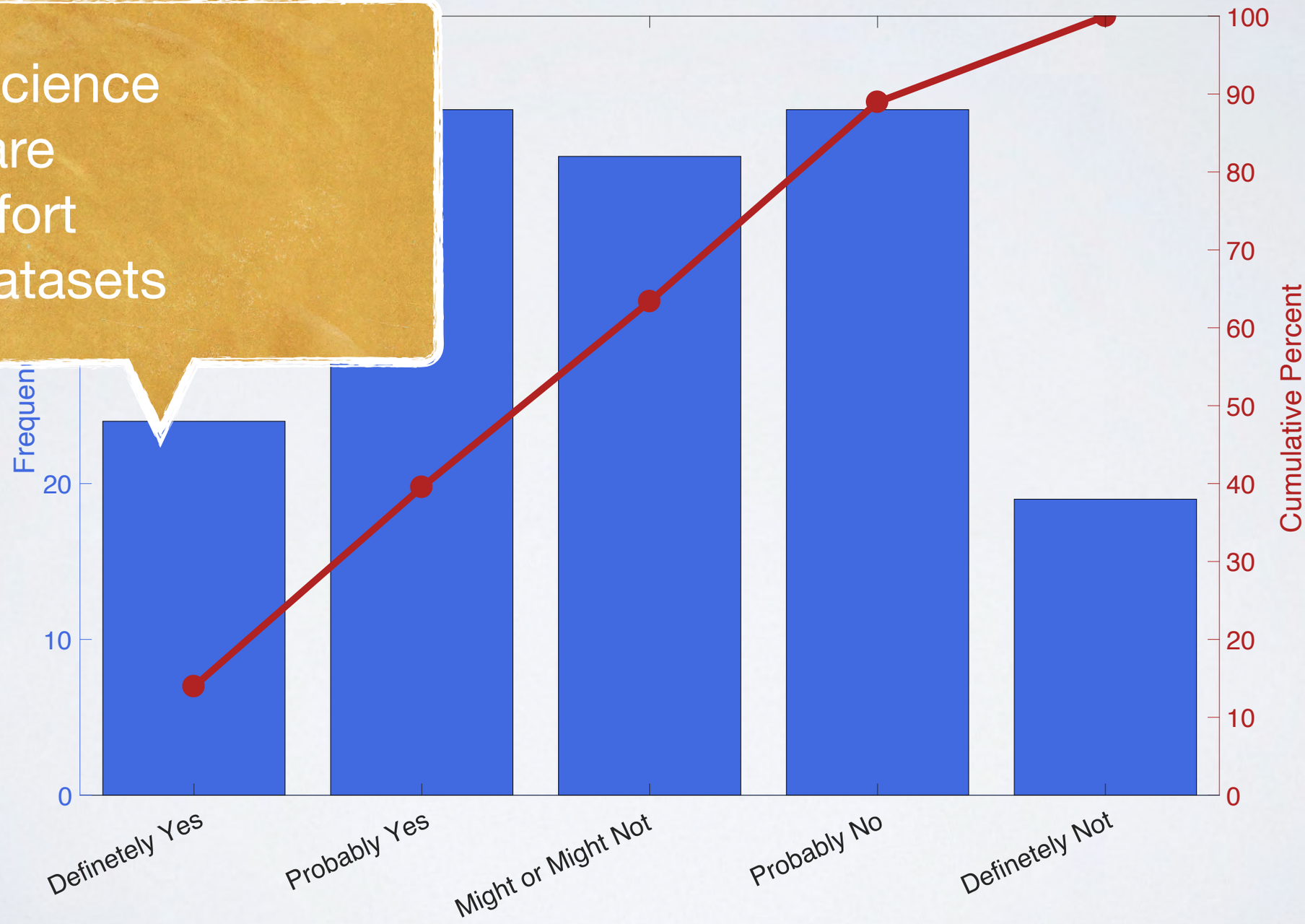
Ferro, N. and Kelly, D. (2018). SIGIR Initiative to Implement ACM Artifact Review and Badging. *SIGIR Forum*, 52(1):4–10.

Would Badges Change Your Research?



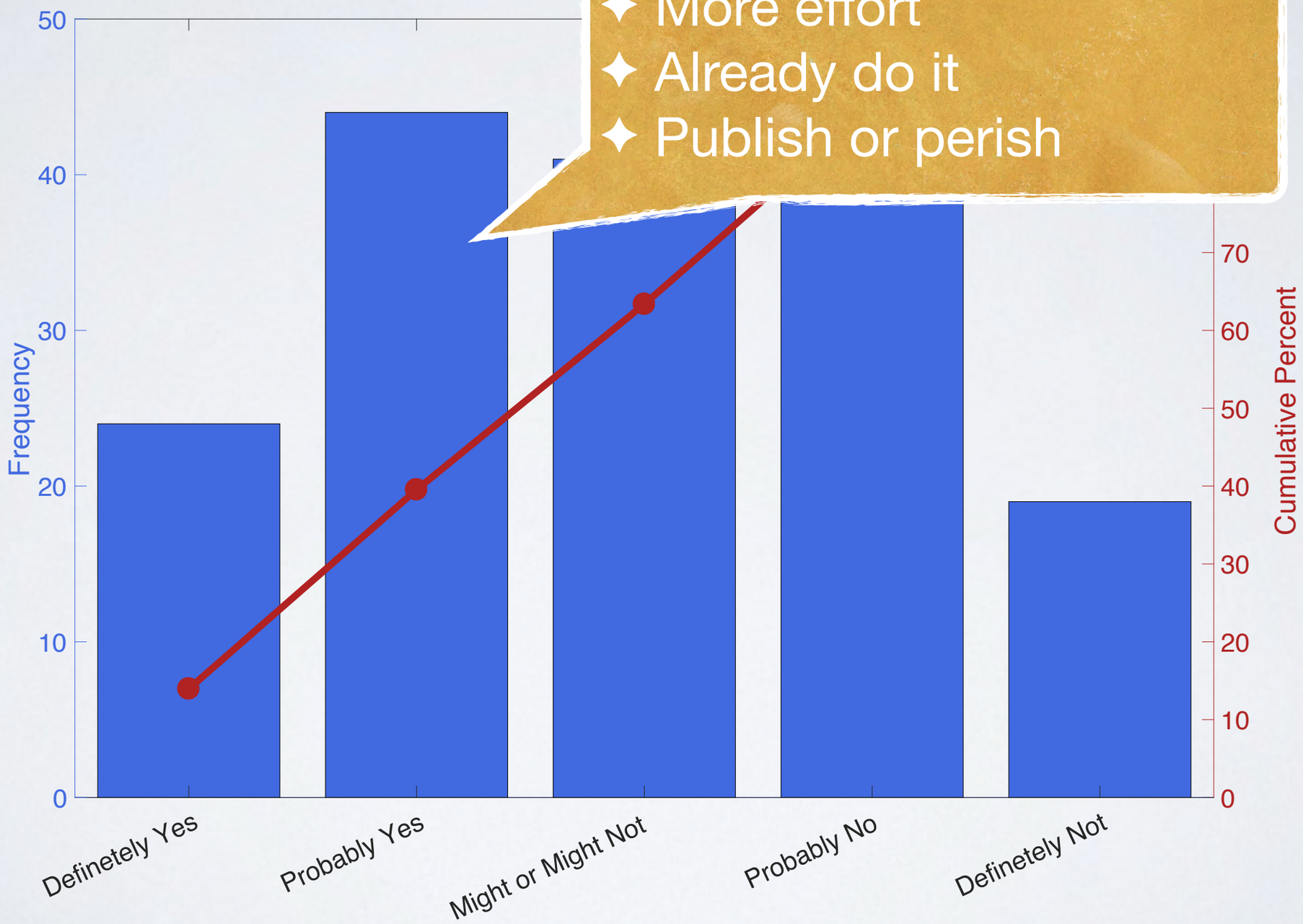
Would Badges Change Your Research?

- ✦ Better science
- ✦ More care
- ✦ More effort
- ✦ Open datasets



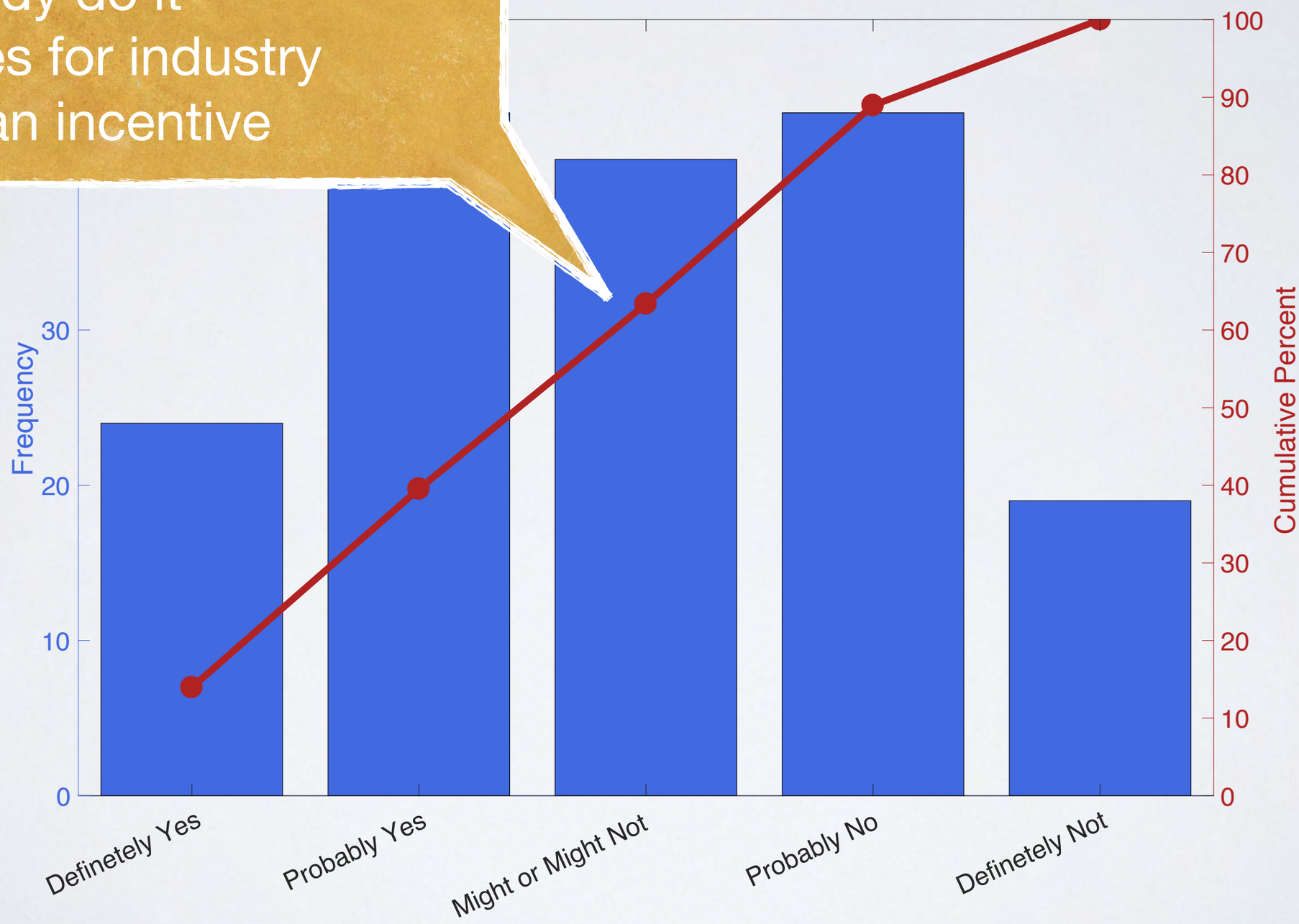
Would Badges Change Your Research?

- Better science
- More care
- More effort
- Already do it
- Publish or perish



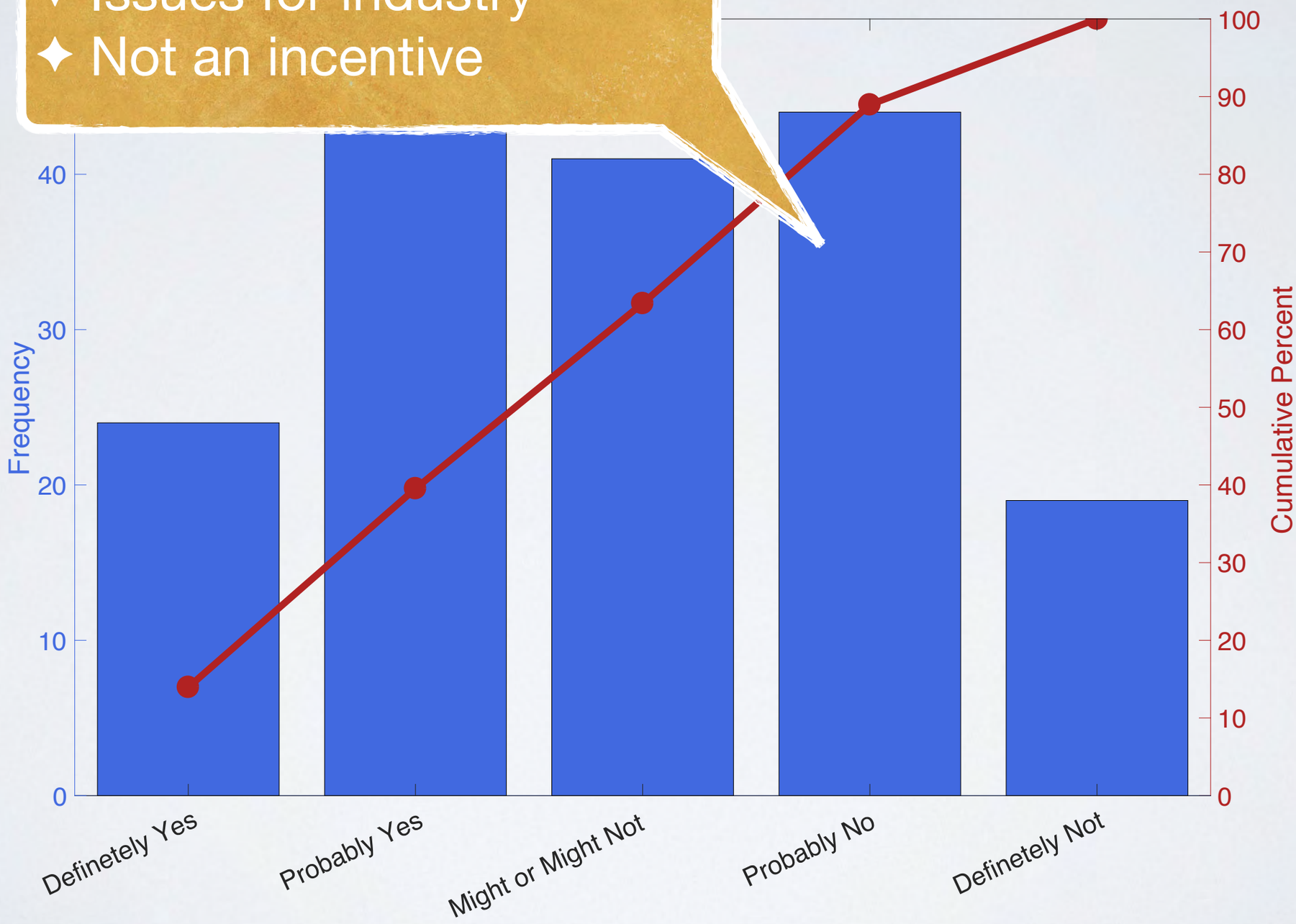
Would Badges Change Your Research?

- More care
- More effort
- Already do it
- Issues for industry
- Not an incentive

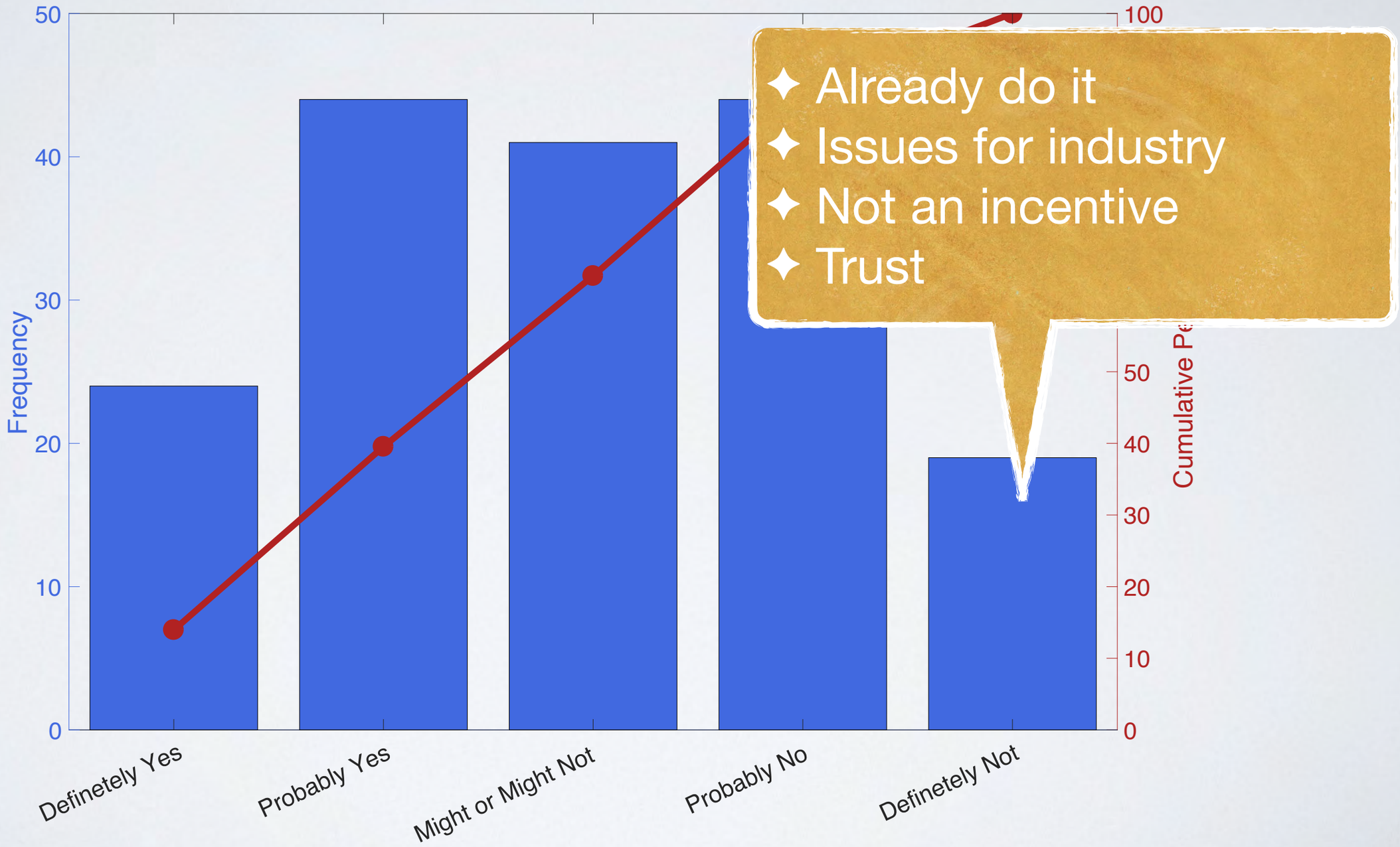


Would Badges Change Your Research?

- ✦ Already do it
- ✦ Issues for industry
- ✦ Not an incentive



Would Badges Change Your Research?



Reproducibility: Some Needs

- Shift in culture
  - more work needed to put reproducibility in action
  - what about the pressure to publish?
  - acknowledgment in careers
- Systematic but focused approach
  - how to choose what to reproduce?
- Quantitative assessment
  - when do we consider something as “reproduced”?
- Infrastructures (evaluation campaigns?)
  - lightweight tools and protocols… but they need adoption!

Questions?

http://phdcomics.com/comics/archive.php?comicid=139