A New Paradigm for Experimental Pattern Recognition Research

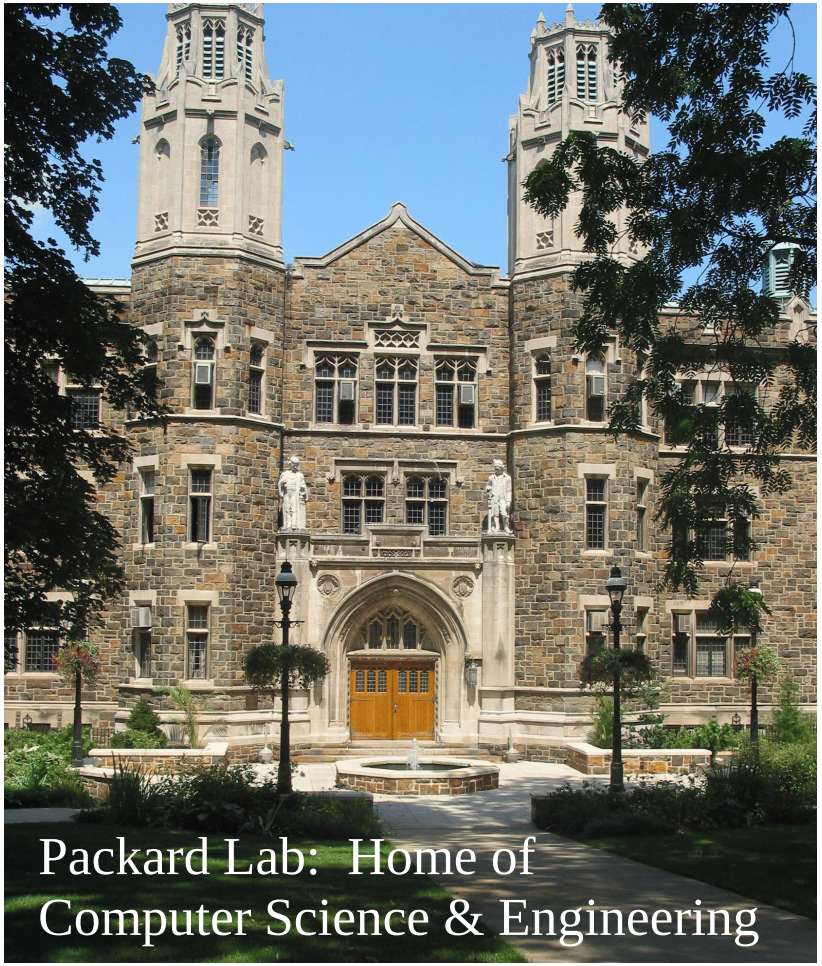
(including Document Analysis and Exploitation*)

Daniel Lopresti
Computer Science & Engineering
Lehigh University
Bethlehem, PA, USA

Bart Lamiroy
Nancy Université
Campus Scientifique, BP 239, 54506
Vandoeuvre Cedex, France

* Supported by a Congressional appropriation administered through DARPA IPTO via Raytheon BBN Technologies.

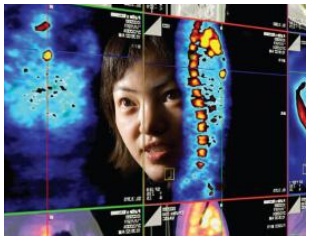
Lehigh University


Packard Lab: Home of Computer Science & Engineering

- Private research university (1865)
- Four colleges: Engineering, Arts & Sciences, Business, Education
- Faculty = 441 full-time
- Students = 2,064 grad, 4,577 undergrad
- Three campuses spread over 1,600 acres (mountain side, wooded)
- Located about 1.5 hours from NYC and Philadelphia, 3 hours from Washington
- Engineering College ranked in top 20% of Ph.D.-granting schools in U.S.
- University ranked in top 15% of U.S. national universities

Lehigh University

CSE Department





17 faculty, 60 PhD, 20 MSE …

- Bioinformatics
- Biomedical Image Analysis
- Computer Game Design
- Data Mining
- Database Systems
- Document Analysis
- Intelligent Agents
- Mobile Robotics
- Networking
- Parallel Processing
- Programming Languages
- Computer Security
- Semantic Web
- Social Networking
- Web Search and Web Systems

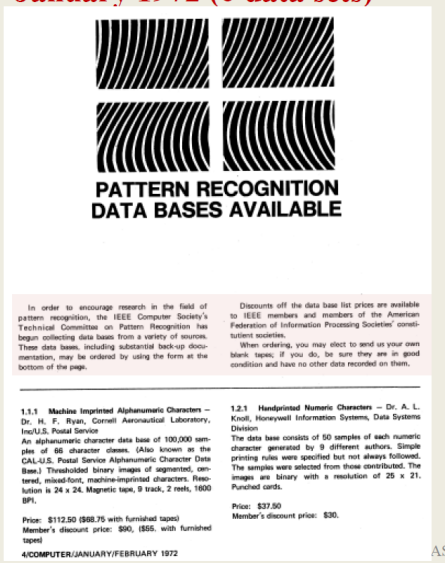
Germ of an Idea (1)

George Nagy graduat
Physics (fencing and
solving Euler's Secon
awarded the PhD at C
Rosenblatt build Tobe
network for speech re
character recognition
claims credit for IBM
reverse sabbatical at t
trains from cats' medi
Department of Compu
where he dabbled in c
1985 he has been Prof
Troy, NY. Nagy's credits in document analysis incl
Casey, "self-corrective" character recognition with
years later with Henry Baird), character recognitior

Wonderful keynote on datasets by George Nagy at DAS 2010 …

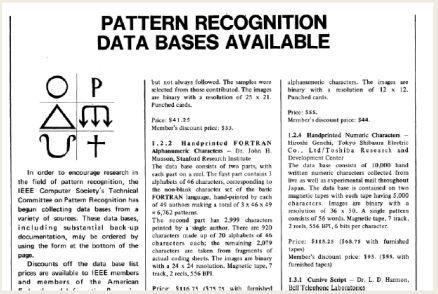
Data Sets advertised in IEEE *Computer*

January 1972 (6 data sets)

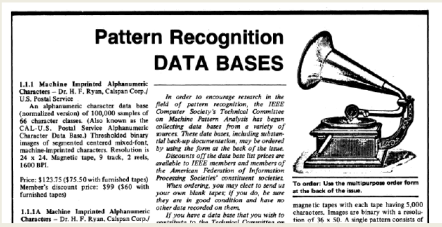


PATTERN RECOGNITION
DATA BASES AVAILABLE

January 1973



PATTERN RECOGNITION
DATA BASES AVAILABLE



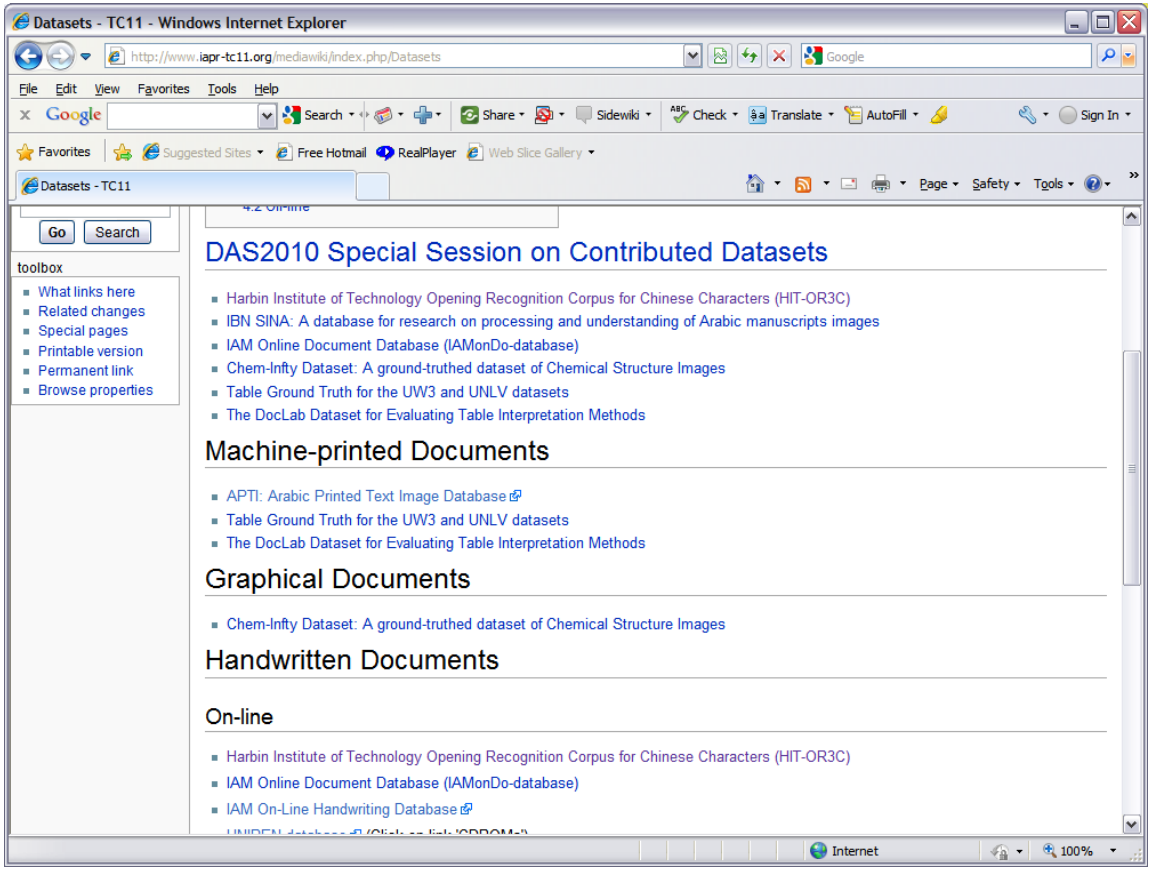
Pattern Recognition
DATA BASES

April 1976 (10 data sets)

* Slides available on the DAS 2010 website.

Germ of an Idea (2)

IAPR TC-11 ("Reading Systems") website:



http://www.iapr-tc11.org

Status Quo

Despite tremendous advances in computer and communications technologies, the way we conduct research and disseminate results is essentially unchanged over the past 50 years.*

Identify problem → conceive solution → find data →
test method → publish results (→ field system)

Modest steps forward include more stringent peer review processes (comparisons to prior work), attempts to create and share common datasets, and "competitions" at international conferences.

* See the excellent keynote talk delivered by George Nagy at DAS 2010:
http://cubs.buffalo.edu/DAS2010/GN_testing_DAS_10.pdf

Broad Aims for Research

Research goals for work in machine perception*, broadly-stated:

- Develop algorithms that are robust and approach human levels of performance for specific tasks of interest.
- Invent new methods that are better than known techniques.
- Generate experimental results that are well-documented, understandable in context, and reproducible by others.
- Build on past knowledge to yield new knowledge which moves us toward solutions for problems of vital importance.

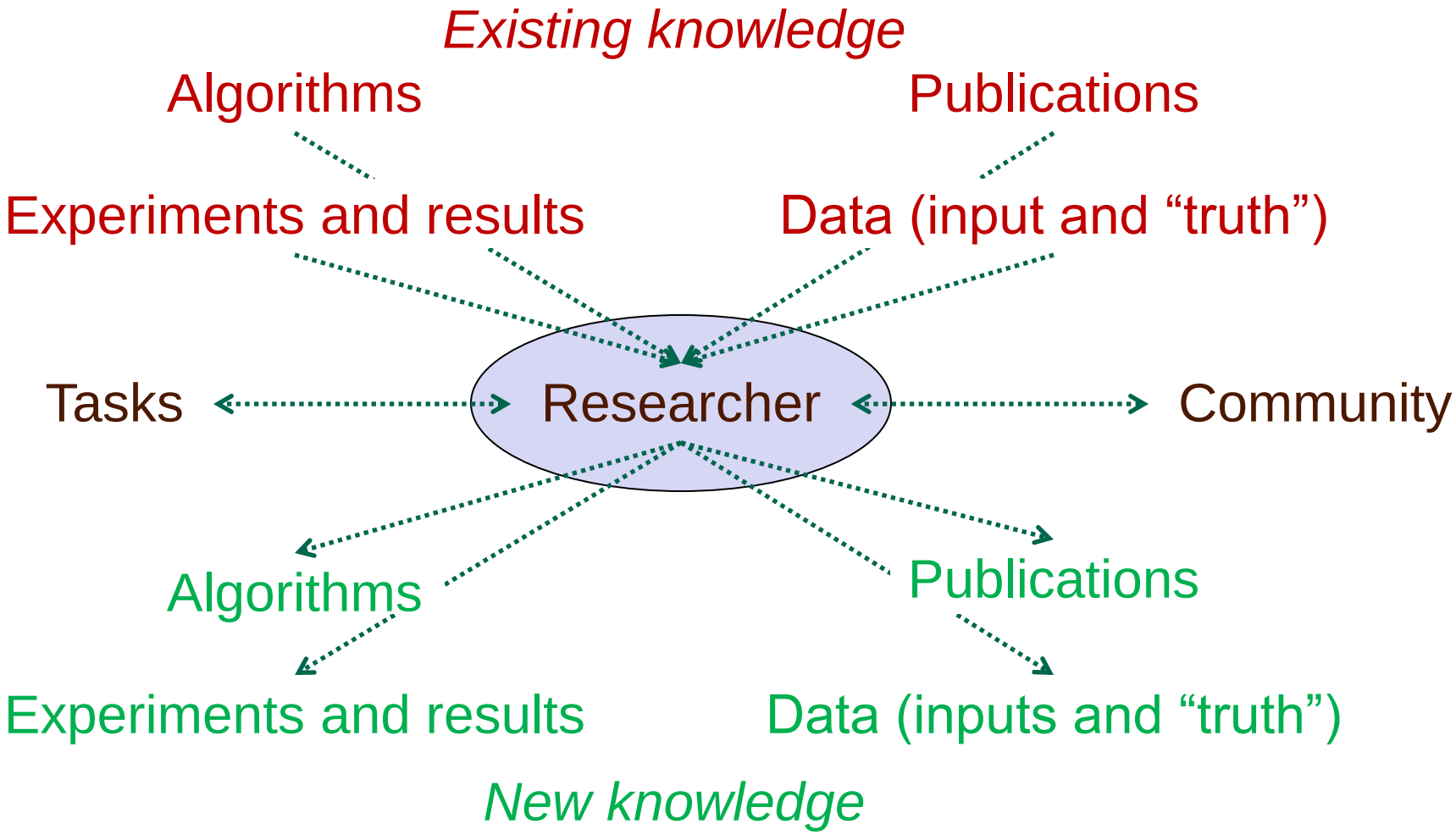
* *Machine perception* refers to any pattern recognition task that attempts to mimic human behavior, including computer vision, document analysis, biomedical image processing, speech recognition, natural language understanding, etc.

Moving Forward

A fundamentally new paradigm for experimental research:

- Not a minor "tweak" – an enormous, ground-breaking change.
- Enabled by infrastructure that largely exists – most pieces are already present, they just need to be put into place.
- Could dramatically accelerate the rate of scientific discovery.
- Power derived from community participation, but vision is so compelling that some version of this future is inevitable.
- Critical mass of early adopters (thought-leaders) will play a key role in guiding the vision to fruition. Despite open-access model, significant proprietary benefits are also possible.
- Large potential for "meta" research (developing the paradigm).

Current Paradigm



Problems with Status Quo (1)

Stated goal:

- Develop algorithms that are robust and approach human levels of performance for specific tasks of interest.

Observations:

- We want algorithms to be general, but too often they are tested on small, overused datasets far removed from the real world.
- Nearly all experimental results reported in the literature are biased by algorithm developer's intimate knowledge of the data.
- Current practices lack convincing evidence of generality.
- What does "human levels of performance" mean? Even experts can disagree on all but the most trivial of cases.

Problems with Status Quo (2)

Stated goal:

- Invent new methods that are better than known techniques.

Observations:

- How do we know when we have succeeded?
- Need to compare against previously published results creates over-reliance on standard datasets (which is counter-productive).
- Attempts to re-implement a published algorithm are problematic (incomplete descriptions, inherent conflict of interest).
- Competitions can be useful, but are infrequent.
- Most papers do not even bother to make such comparisons.

Problems with Status Quo (3)

Stated goal:

- Generate experimental results that are well-documented, understandable in context, and reproducible by others.

Observations:

- Are published descriptions sufficient to reproduce experiments? How often is this even attempted?
- Explicit and/or implicit bias in selecting and using data (e.g., discarding hard cases) makes context difficult to recover.
- "Publish or perish" mindset leads to overstated claims, poor understanding of generalizability of results.

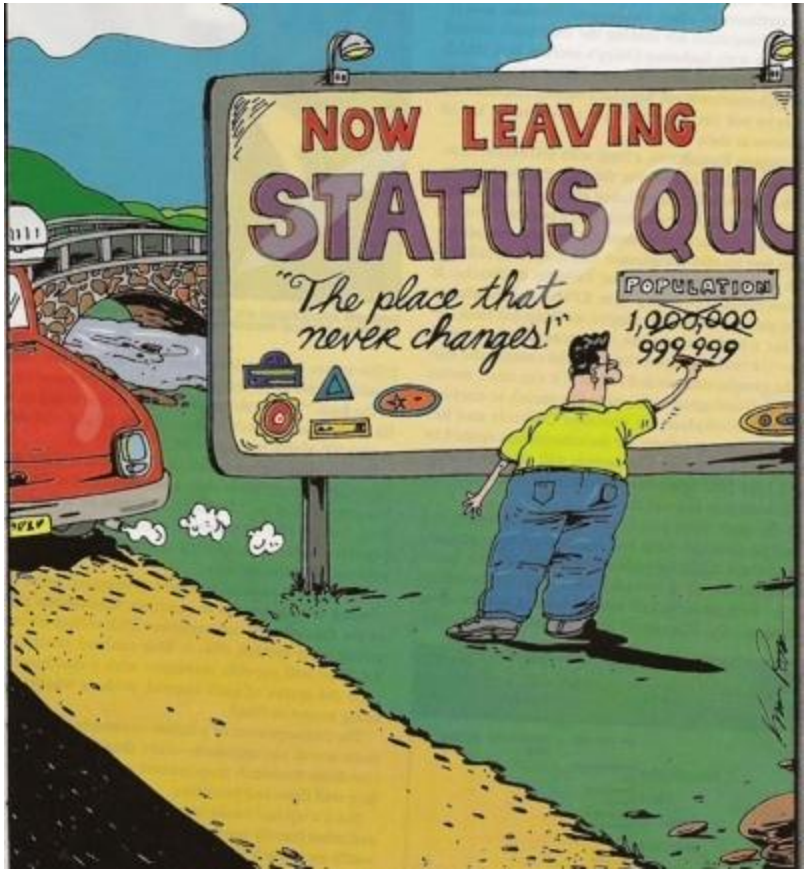
Problems with Status Quo (4)

Stated goal:

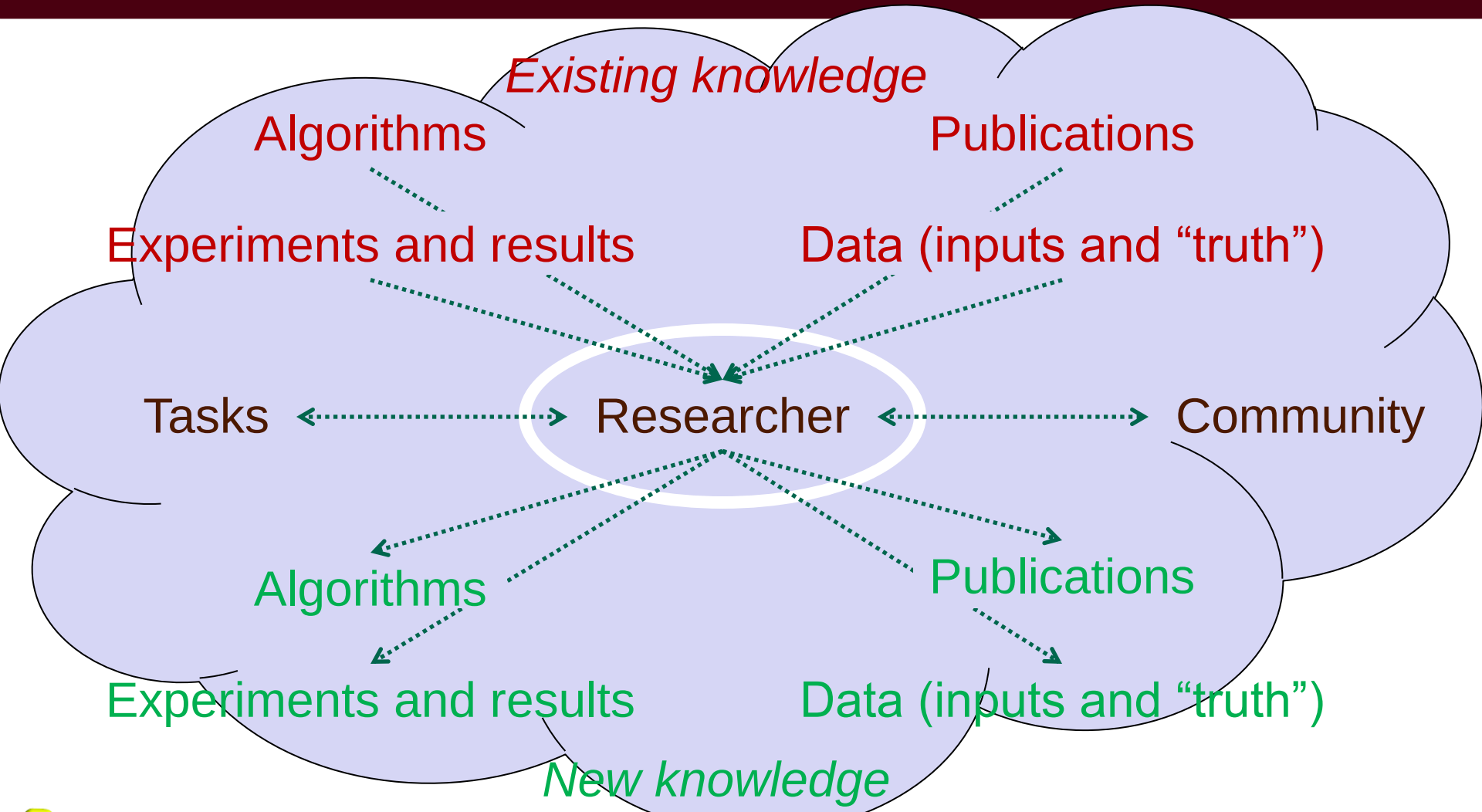
- Build on past knowledge to yield new knowledge which advances the field toward solving problems of vital importance.

Observations:

- Effort is wasted developing methods that do not improve on existing techniques, or are ill-suited for the task at hand.
- Much time spent "reinventing the wheel."
- Like trying to build a pyramid out of shifting sand without first forming it into blocks.
- Impossible to fix without a paradigm shift across the community.


NOW LEAVING
STATUS QUO
"The place that never changes!"
POPULATION
1,000,000
999,999

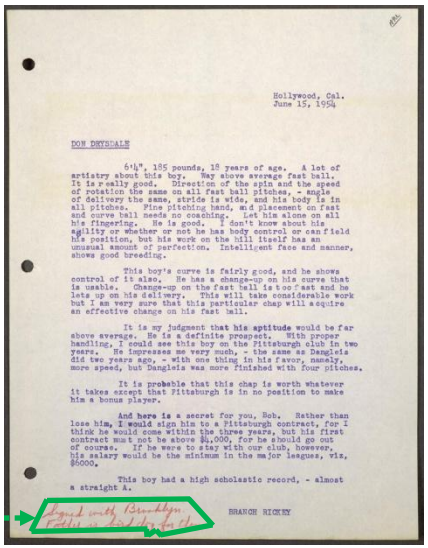
Vision: A Communal Infrastructure



Looking Ahead: A Scenario

Sometime in 2020, Jane, a young researcher just getting started in the field turns her attention to a specific task: given a page image, identify regions that contain handwritten notations.*

- Jane wants a fully general method that can accept any scanned page as input.
- She plans to develop an algorithm that will place a bounding polygon as tight as possible around each handwritten region.
- Each polygon should enclose a logical "unit."



Hollywood, Cal.
June 15, 1954

DON DRYSDALE

6'4", 185 pounds, 19 years of age. A lot of artistry about this boy. Way above average fast ball. It is really good. Direction of the spin and the speed of rotation the same on all fast ball pitches, - angle of delivery the same, stride is wide, and his body is in all pitches. Fine pitching hand, and placement on fast and curve ball needs no coaching. Let him alone on all his fingering. He is good. I don't know about his agility or whether or not he has body control or can field his position, but his work on the hill itself has an unusual amount of perfection. Intelligent face and manner, shows good breeding.

This boy's curve is fairly good, and he shows control of it also. He has a change-up on his curve that is usable. Change-up on the fast ball is too fast and he lets up on his delivery. This will take considerable work but I am very sure that this particular chap will acquire an effective change on his fast ball.

It is my judgment that his aptitude would be far above average. He is a definite prospect. With proper handling, I could see this boy on the Pittsburgh club in two years. He impresses me very much, - the same as Dangleis did two years ago, - with one thing in his favor, namely, more speed, but Dangleis was more finished with four pitches.

It is probable that this chap is worth whatever it takes except that Pittsburgh is in no position to make him a bonus player.

And here is a secret for you, Bob. Rather than lose him, I would sign him to a Pittsburgh contract, for I think he would come within the three years, but his first contract must not be above $6,000, for he should go out of course. If he were to stay with our club, however, his salary would be the minimum in the major leagues, viz, $6000.

This boy had a high scholastic record, - almost a straight A.

Signed with Brooklyn

BRANCH RICKEY

* If you prefer, replace this task by any one in machine perception you like.

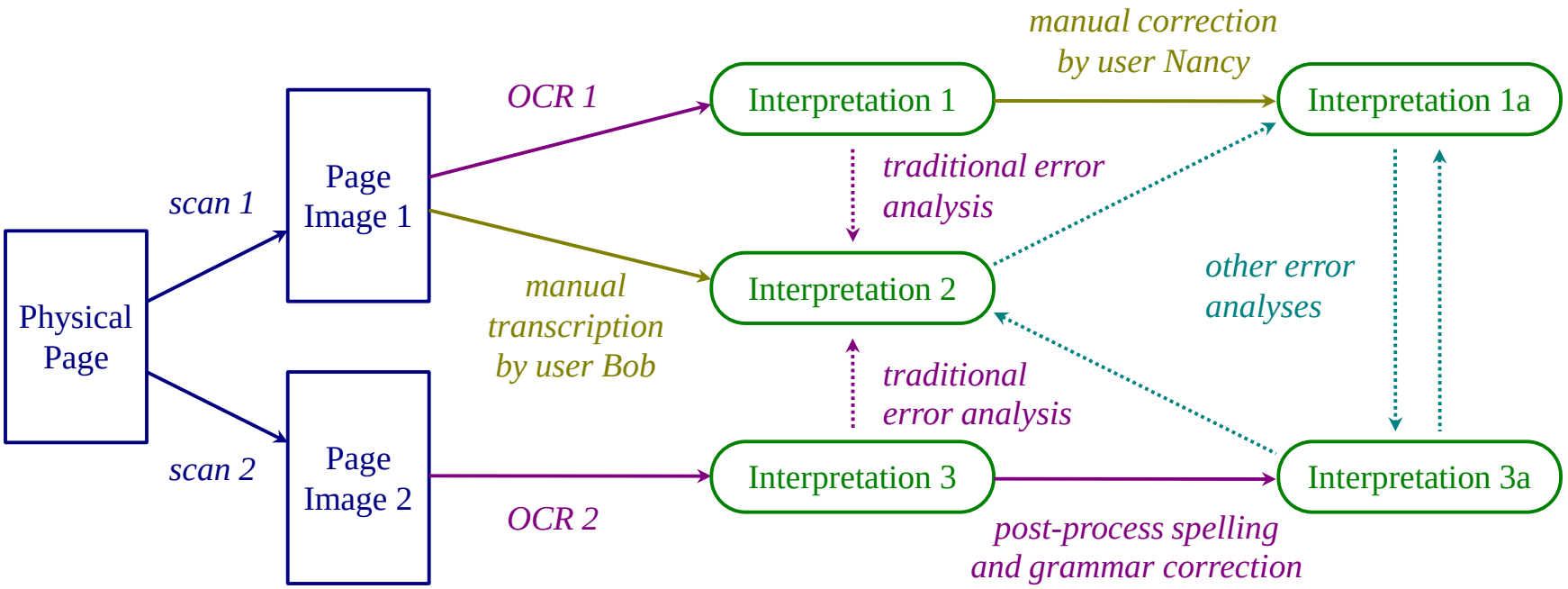
Truth vs. Interpretation

What should Jane be aiming for?

- In this version of the future, there is no such thing as "ground truth" – that term is no longer used.
- Rather, we talk about the *intent* of the author (which can be hard to determine, although sometimes we have access to it) …
- … or the *interpretation* arrived at by a reader of the document (which could be a human or an algorithm).
- There are no "right" or "wrong" answers – interpretations may naturally differ – although for some applications, we expect that users who are fluent in the language and the domain of the document will agree nearly all of the time.

Alternate Interpretations

Entities on a page are located and interpreted by humans interacting with the system, as well as by algorithms which have been invoked to process the page. This raises the possibility of alternate interpretations and relativistic error analyses:

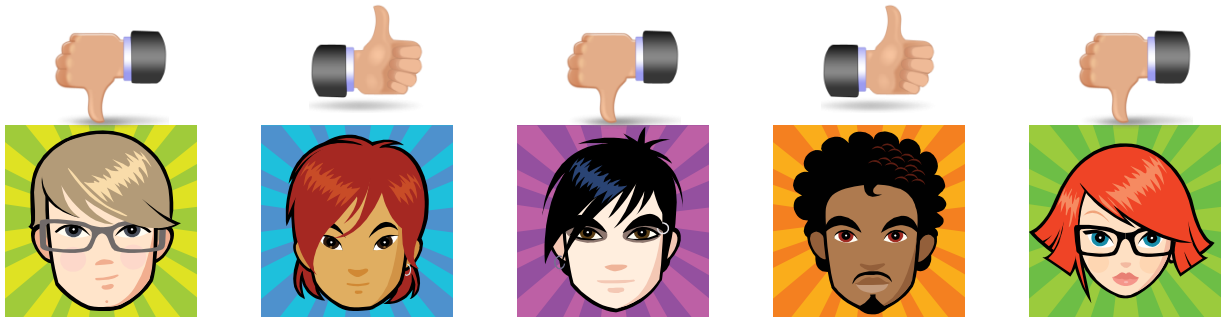


The Role of Reputation

With multiple competing interpretations, how does Jane proceed to develop a new method that mimics a careful, fluent human expert?

- Some people are more careful, fluent, and/or expert than others.
- The concept of online reputation as originally derived for use in social networking (Web 2.0) determines whose interpretations Jane will trust when developing her new method.
- This use of reputation is one key feature of the new paradigm.

Users who have contributed interpretations



Kinds of Reputation

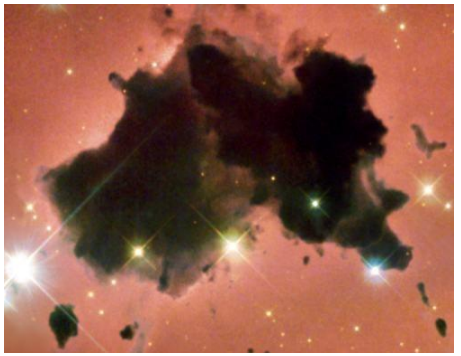
We have always had a notion of reputation, even in the early days. What is different in 2020 is this has been formalized to support experimental research.

Various entities and possible bases for their reputations:

- Algorithm (how well does it address a stated task?)
- Implementation (is it buggy, or robust to a variety of inputs?)
- Dataset (is it representative for a particular problem?)
- Interpretation (is it trustworthy?)
- Publication (is it highly cited?)
- Researcher (amalgamation of contributions to above)

The DARE Server

After Jane has defined her task, she turns to a web server called DARE (for "Document Analysis Research Engine"), to request a set of sample pages for use in developing her new algorithm.



- DARE server lives in the "cloud" – it is not a single machine.
- A community-managed resource which has become a de facto standard (just as UW-1 dataset was once considered a standard).
- Collects datasets, interpretations, algorithms, citations, etc.
- Manages reputations and provides a query interface and API.
- Not just a storage system – DARE supports our new paradigm for experimental document analysis research.

Building Datasets via Queries

Jane queries the server to give her 1,000 random pages from the various collections it knows about. Through the query interface, she specifies:

- Each page should 300 dpi color TIF. Jane may allow conversion from other formats (via built-in utilities), or she may prohibit this.
- Each page should be "predominately" printed material: text, line art, photographs, etc. This might mean ≥ 60% of the page area.
- A reasonable number of pages should contain at least one annotation. For Jane's application, this might be 10% to 25%.
- The handwritten annotations should be delimited in a way consistent with the intended output from Jane's algorithm.

The File Format Issue

In the past, there were attempts to unify the research community around common file formats. Since there was no compelling benefit to the individual researcher, these were routinely ignored.

- The status of the DARE web server as the de facto standard has enticed most researchers to use compatible file formats.
- DARE is the preferred platform to support their research.
- There is no coercion – it is just easier for users to go along since so much data is now delivered from the server and no longer lives locally on their own machines anyway.
- Some find this analogous to the earlier rise of other standard file formats such as TIF, JPG, PDF, PS, RTF, etc.

Logging and Transparency

In processing Jane's request, the DARE server returns the set of 1,000 random pages along with their associated interpretations.

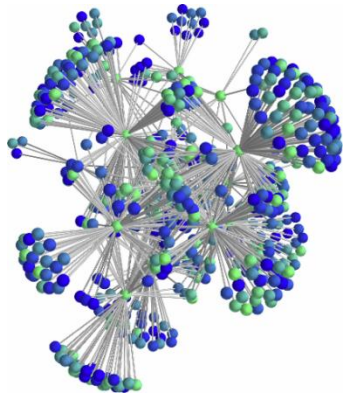
- It also makes a permanent record of her query and provides a URL that will return exactly the same set of pages each time it is run.
- Any user with the URL can access the dataset and see Jane's parameter settings.
- All accesses are logged.

http://www.dareserver.org/query?ID=124881371744

Dataset created by Jane Simmons on July 26, 2020.
Parameter settings:
- 1,000 random page images
- 300 dpi color TIF format (no conversions)
- At least 60% foreground (printed) material
- 10% to 25% containing at least one annotation
- XMLpoly interpretation format
- Highest available reputation for data and interpreter

Click here to download the dataset.
Click here to access the log for this dataset.

Resource Discovery



Based on Jane's browsing behavior on the server, and her eventual query to create a dataset, DARE is smart enough to glean what she may be doing. Other people have made similar queries in the past.

- DARE suggests a number of previously published papers addressing similar problems. Jane already knows about some of these, but others are new to her. She notes this for reference.
- The server also reports that several researchers have registered algorithms that Jane may choose to test her own against later. (For security and reliability, the DARE server runs contributed code using a virtual machine architecture.)

The Copyright Question

In the field's early days, the sharing of datasets was hampered by copyright issues. A group could create data for its own use, but was prevented from sharing it with others. By 2020, this is no longer a problem:

- Vast quantities of image data are available on the WWW. Digital libraries, both commercial and public, contain billions of pages.
- OCR and manual transcriptions have made this data searchable via simple text queries, but more sophisticated methods are needed to deal with "information overload" (hence Jane's work).
- Some digital libraries require a subscription. The DARE server knows which libraries Jane has access to and manages the authentication process for her seamlessly.

Other Kinds of Datasets

The DARE server is not limited to delivering randomly-generated datasets. It supports other contributions from the community including:

- Legacy datasets like UW-1 referenced by their own URL's: http://www.dareserver.org/dataset?ID=UW-1
- Components available for synthesizing page images using pre-defined models for layout, typography, and noise effects.
- A "black box" testing protocol for data that is stored securely on the DARE server but that cannot be re-distributed.

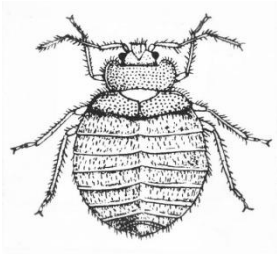
The Creative Process is Unchanged

With dataset in hand, Jane starts to work on her algorithm. This step is no different from what researchers do today.

- Jane may study the pages in the dataset provided by the server to better understand the cases her method should handle.
- If her algorithm employs machine learning, she could use some of the pages as a training set and others as a "test" set.
- Since it is understood that the design of her algorithm may be biased by knowledge of the specific pages in the dataset, and since her goal is a more general method, Jane will only use this dataset for development purposes and never when it comes time to publish her results.

Buggy Data

The DARE server manages large quantities of data contributed by the community with no direct oversight. Hence, it is understood that bugs will arise occasionally.



- While working with the data, Jane notices a few problems. One of the page images was delivered to her upside down (rotated by 180 degrees). In another case, the TIF file for a page was unreadable, at least by the version of the library she is using.
- Being a responsible member of the research community (and wanting her online reputation to reflect this fact), Jane logs onto the DARE server and, with a few mouse clicks, reports both of these problems – it just takes a minute.
- Jane's reports will be checked by other community members.

Differing Interpretations

In a few other cases, Jane disagrees with the interpretation provided for the page in question.

- In her opinion, on one page the bounding polygons are drawn improperly and, on another, an annotation has been missed.
- Rather than just make local changes to her own private copies of the data (as would have happened in the past), Jane records her interpretations on the DARE server. Her opinions are added to the existing ones, enriching the entire collection.
- To facilitate this, the server provides a wiki interface with editing and markup tools that run in any web browser.

Unbiased Testing

After Jane is done fine-tuning her algorithm, she prepares to write a paper about it. This will involve testing the claim that her method works for arbitrary inputs, not just for the data she has been using for the past six months. The DARE server gives her two options for performing random, unbiased testing of her method. These are:

- Option #1 (local): Jane can "wrap" her code in a driver provided by the DARE server that runs the algorithm on a series of page images, calling the server each time a new page is required.
- Option #2 (remote): Jane can choose to upload her code to the DARE server, thereby contributing it to the community. The server runs her algorithm on a series of previously unseen page images according to her specifications.

Unbiased Testing (Local)

Jane wraps her code in a driver provided by the DARE server …

- The code continues to run on Jane's machine, with the server delivering a series of random page images it has not seen before but that satisfy certain properties she has specified in advance.
- The algorithm performs its computations and responds to the server within a few seconds – too fast for a human to intercede.
- Testing takes place over a period of time, with random delays between when the next page is requested and when it is delivered. It would be too tedious for a human to try to "game" the system.
- Hence, everyone can be certain that the results reported are in fact the unbiased output from Jane's program.

Unbiased Testing (Remote)

Instead, Jane can choose to upload her code to the DARE server …

- In this case, the server will run her algorithm on a series of previously unseen test pages according to her specifications.
- Since the testing is under the complete control of the DARE server, there is no risk of someone trying to "game" the system.
- A virtual machine architecture is used for security and reliability. Malicious code is contained and cannot compromise the server.
- The DARE server will also maintain Jane's code in the system and use it in future comparisons when other researchers test their own new algorithms on the same task.

Local vs. Remote Testing Models

The two schemes, local vs. remote, provide tremendous flexibility:

- The first option allows Jane to keep her code private, if she so wishes, since the code runs entirely on her own machine. Perhaps she is still developing her method and would rather not publicize it until she can improve it, or maybe she works for a company or a government agency, or she is planning to file a patent application.
- The second option, of course, allows Jane to contribute her algorithm to the community (and, again, raise her online reputation), just as anyone can add new data or interpretations to the DARE server.

Certified Testing Results

At the end of the evaluation, Jane is provided with:

- A set of summary results showing how well her algorithm matched human performance on the task.
- Another set of summary results showing how well her algorithm fared in comparison to other methods tested on the same pages.
- A *certificate* (i.e., a unique URL) that guarantees the integrity of the results and which can be cited in the paper she is writing. Anyone who enters the certificate into a web browser can see the results delivered directly from the (trusted) DARE server.
- Now there can be no doubt that what Jane reports in her paper is true and scientifically reproducible.

Certified Testing Results

http://www.dareserver.org/query?ID=08272348481991

Certified evaluation run by Jane Simmons on December 3, 2020.
Parameter settings:
- 2,000 random page images guaranteed previously unseen
- 300 dpi color TIF format (no conversions)
- At least 60% foreground (printed) material
- 10% to 25% containing at least one annotation
- XMLpoly interpretation format
- Highest available reputation for data and interpreter

Performance relative to human interpretation:
- 92% recall / 89% precision
- No result returned for 2% of inputs (click here to access these pages).

Performance of best previously registered algorithm on same pages:
- 87% recall / 82% precision
- +5% net improvement in recall / +7% net improvement in precision

Click here to download the dataset.
Click here to access the log for this dataset.

Guaranteed unbiased testing

Comparison to human expert

Comparison to earlier method

Promoting Robustness

It was perhaps a bit optimistic of Jane to believe that her method would handle all possible inputs and, in fact, she learns that her code crashed on 2% of the test pages.

- The server allows Jane to download these pages to see what is wrong (she failed to dimension a certain array to be big enough).
- If Jane is requesting a certificate, the DARE server will guarantee that her code never sees the same page twice.
- If she is not requesting a certificate, there is no restriction and the server will deliver the same page as often as she wishes.
- Past researchers could remove troublesome inputs from their test sets, but the DARE server now prohibits this.

Task-Based Evaluation

In the past, it was well known that simple summary measures like recognition accuracy and precision / recall may not correlate with performance in a target application. Nothing could be done about this, however, since the extra effort required to implement a full end-to-end system was too much work for a single researcher.

- Now, in 2020, Jane can choose to evaluate her algorithm as part of a pre-programmed pipeline that consumes the annotations it identifies, performs handwriting recognition, and tests the results in prototypical information extraction and retrieval applications.
- As before, relative improvement over existing methods is key.
- A variety of pipelines are fielded on the server for testing.

Other Notable Benefits

The new paradigm brings with it a range of significant benefits:

- Experimental results are now verifiable and reproducible.
- Beginning researchers no longer waste their time pursuing methods that are inferior to known techniques (since the DARE server will immediately inform you if another registered algorithm did a better job on the test set you were given).
- The natural (often innocent) tendency to bias an algorithm based on knowing the details of the test set has been eliminated.
- The overuse of small "standard" datasets so prevalent in the early days of the field is now no longer a problem.

Questions Never Before Answerable

Jane's queries are just one example of what DARE makes possible. A system such as this will allow users to ask questions like:

- What is the best known algorithm for a certain task? (This is the method that comes closest to matching human performance.)
- What is the domain of expertise for a particular algorithmic technique? (These are the categories of inputs for which the method performs significantly better than it does elsewhere.)
- What classes are most vexing to document image analysis? (These are the inputs for which known methods perform poorly.)
- What is the "hardest" open problem in a given field of inquiry? (This is the task for which the current gap between machine and human performance is the largest.)

Attempts to Defeat the System

The DARE server is not foolproof – it provides many features to encourage and support good science, but it cannot completely eliminate the possibility of abuse by a malicious individual.

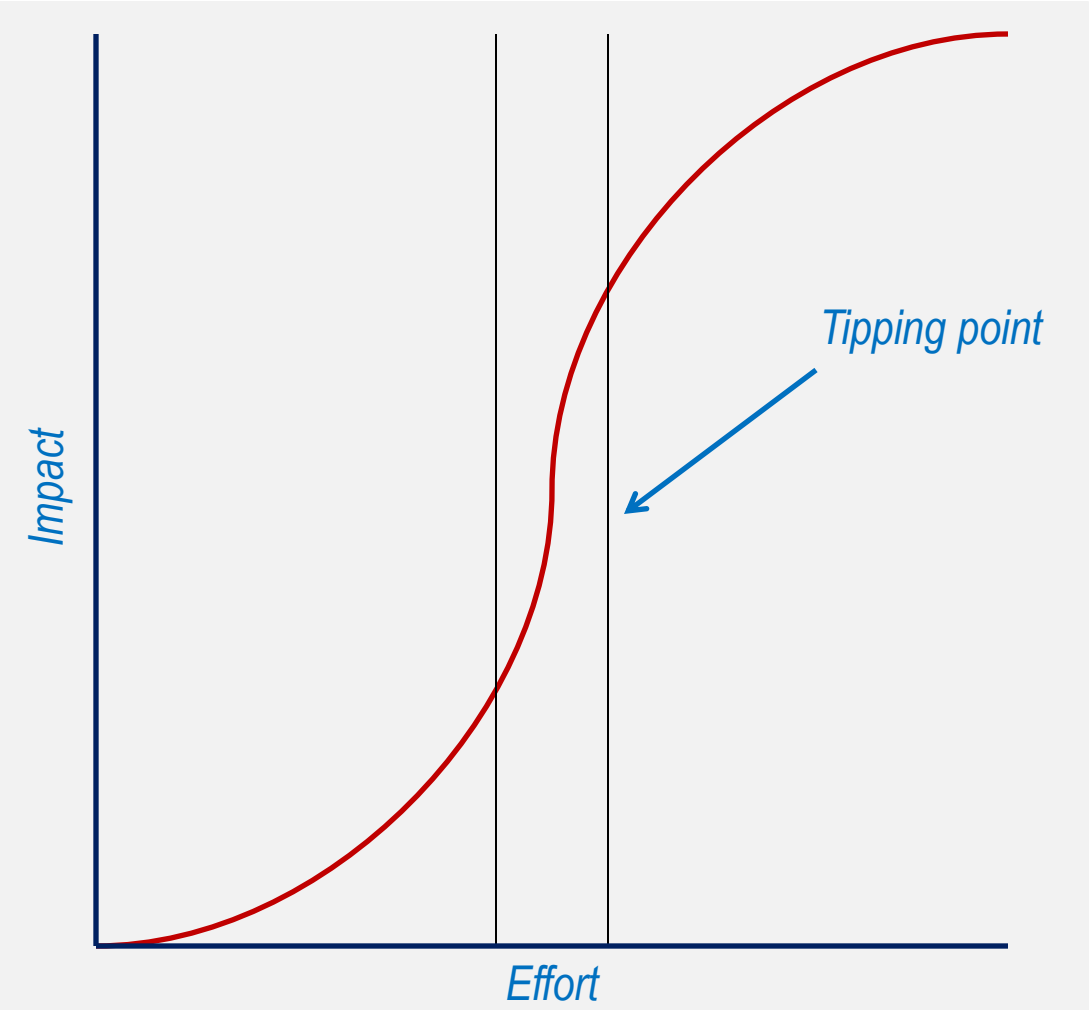
- For example, in order to guarantee that a given page is not seen twice, it is necessary to register both the user and the algorithm.
- Just as with any other web service, it is possible to create multiple accounts in an attempt to circumvent this rule.
- However, due to the community nature of DARE, all records are open and visible to every user of the system, increasing the risk of being caught to the degree that legitimate researchers would never be willing to take that chance.

Key Pieces Already Present

Much of this should sound familiar because it is already happening elsewhere in other contexts (including Web 2.0):

- amazon.com: recommendations based on searches and past purchases, independent reviewers who in turn are rated by users.
- Netflix: movie suggestions based on ratings by >1M customers.
- Wikipedia: community authorship of a valuable online resource.
- YouTube and Flickr: user-provided and annotated multimedia.
- NLM model for interactive scientific publication.
- Work by NLM / Lehigh team on multi-observer segmentation.
- Recent research on data provenance (e.g., VisTrails system).
- IAPR TC-11 and TC-10 efforts to collect standard datasets.

Where is the Tipping Point?



$64,000 question: is required effort modest or large?

We believe it should be possible to achieve key aspects of what we have described here with a few year's work.

Possible Paths Forward

Currently ongoing:

- Lehigh DAE project.
- IAPR TC-11 and TC-10 dataset efforts.

Worth exploring:

- Grants, partnerships, evangelizing within the community.

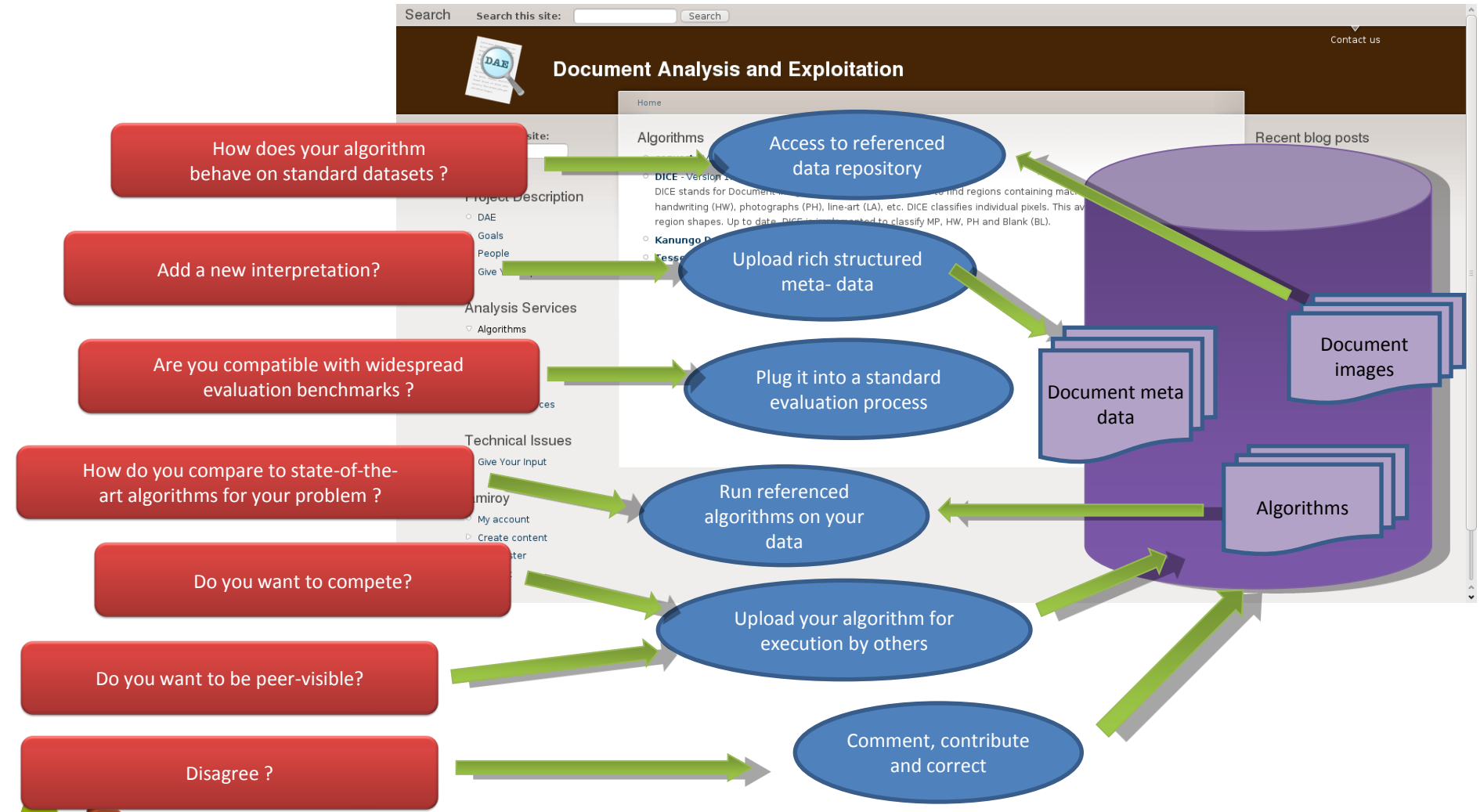
From the international research community:

- Ideas?
- Interest??
- Involvement???

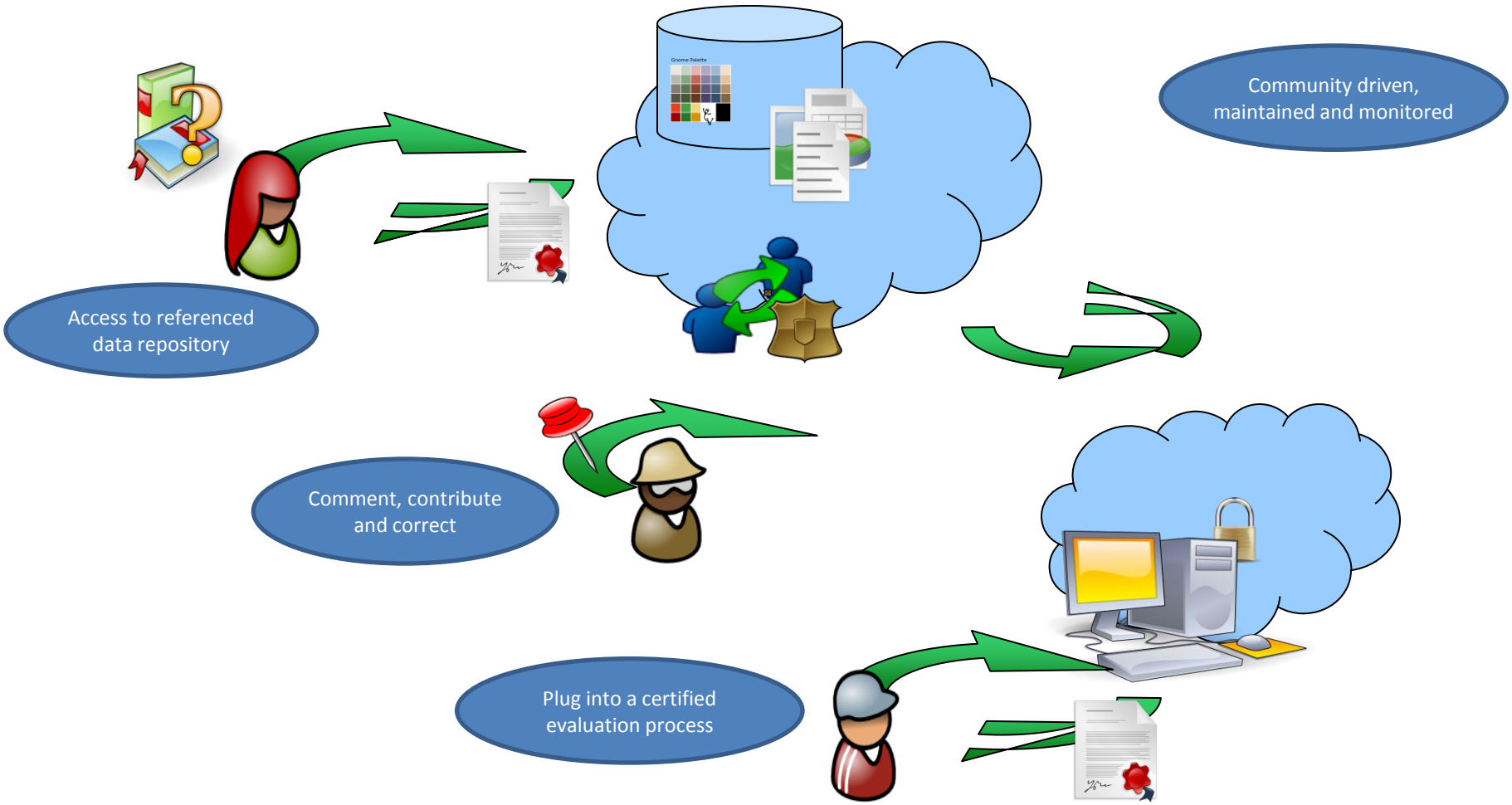
Why this Future is Inevitable

- Not everything discussed is under our control (e.g., copyright is an enormous legal issue), so why are we so certain about this?
- Small hurdles mean scenario will likely play out differently, but enough is doable that it is entirely feasible to make this happen.
- Focus on one key question: is the vision compelling and will it advance science? If the answer is "yes," then it must happen.
- From a competitiveness standpoint, even though what we have described is a communal resource, some individuals / institutions / countries will be better positioned to exploit it at first than others.
- The tipping point may arrive sooner than we expect.

DAE Server (dae.cse.lehigh.edu/)



DAE Server Technical Overview



Hardware & Software

Hardware:

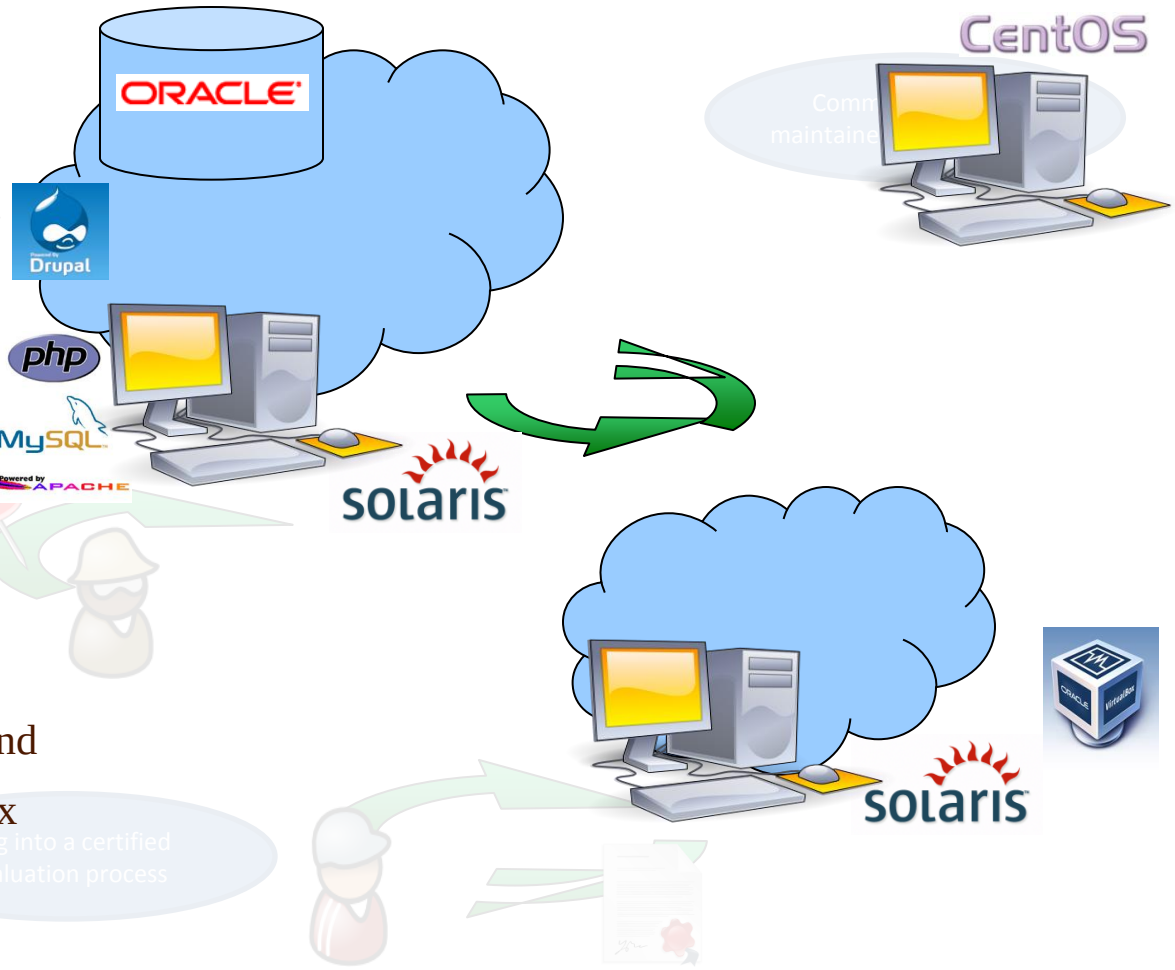
- Sun Fire X4540
  - 48TB raw storage,
  - dual six-core AMD Opteron

Architecture:

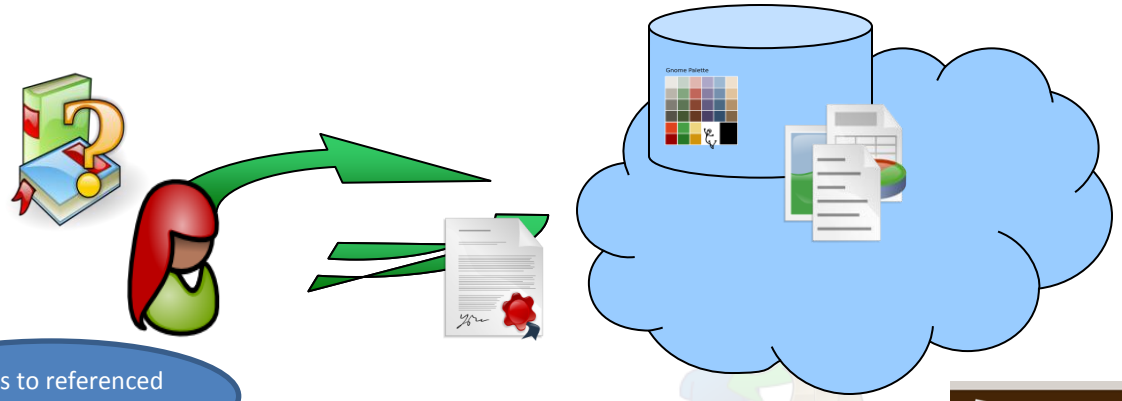
- Solaris Zones
- VirtualBox
- Beowulf HPC

Software:

- Drupal powered LAMP front-end
- Custom open-source PhP – Ajax
- Oracle back-end



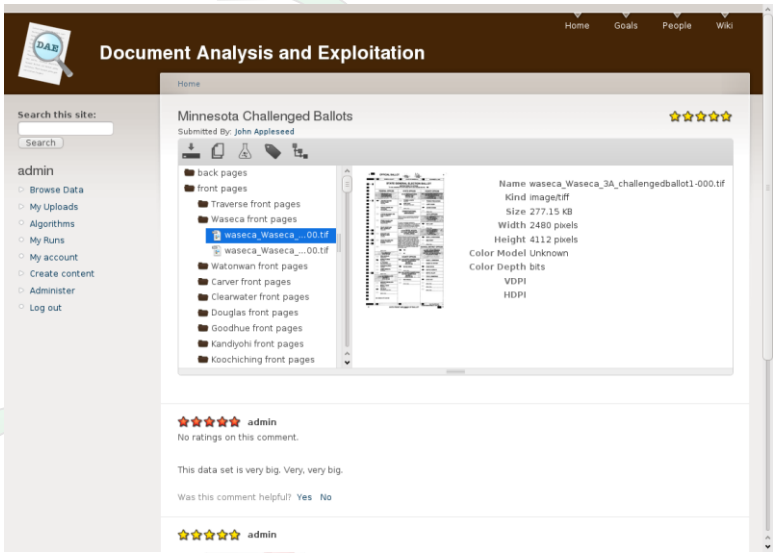
Intelligent Data Repository



Community driven, maintained and monitored

Currently Implemented:

- Web-based browsing and downloading
- XML annotation format for uploading
- Full raw SQL interface
- Search by tag name



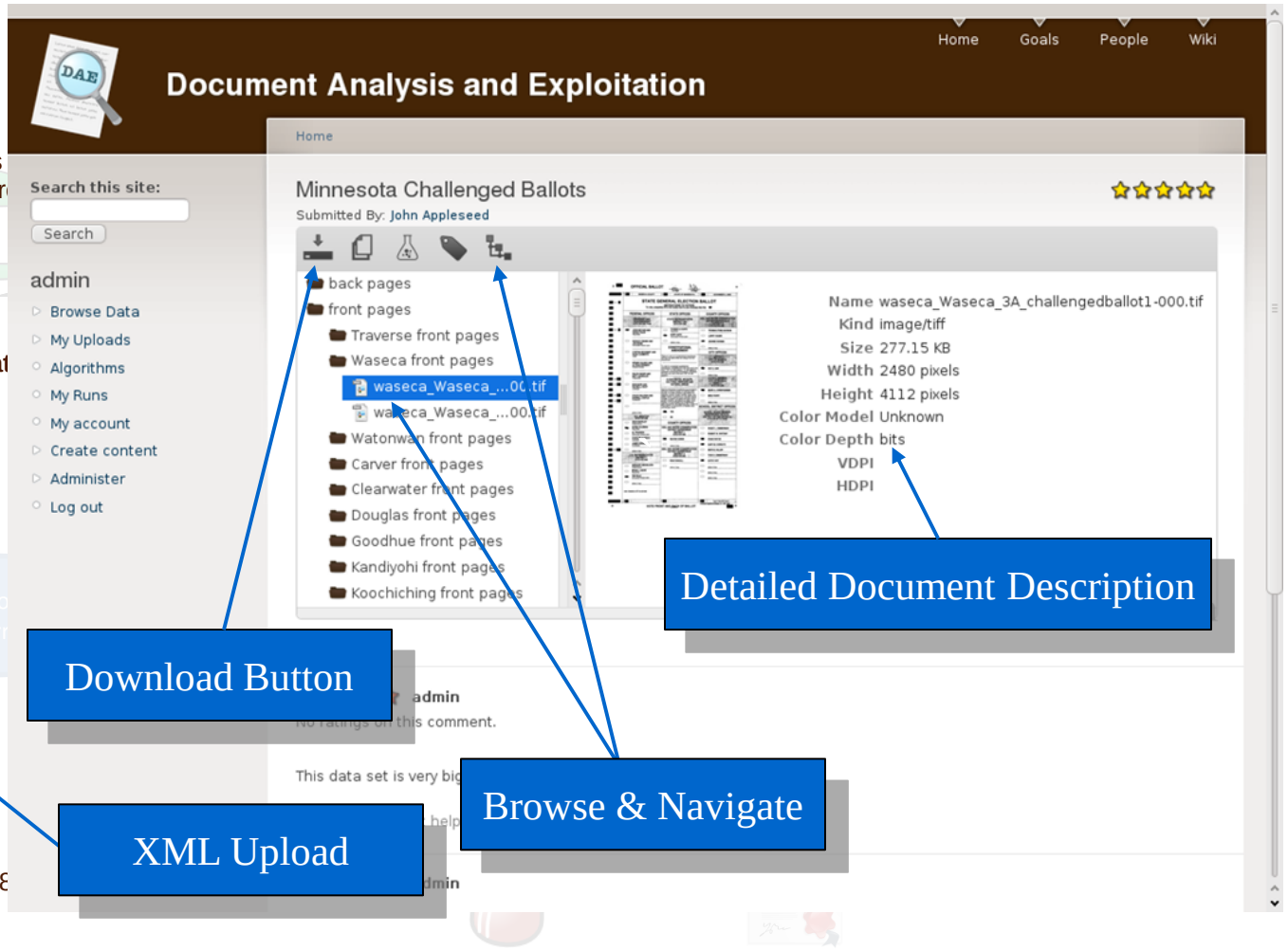
Comment, contribute

Plug into a standard evaluation process

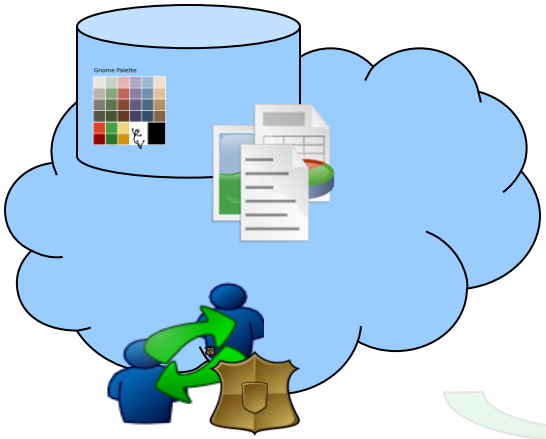
Intelligent Data Repository

```
<dataset>
  <id>3_G</id>
  <name>UNLV 3_G Sub-Dataset
        for testing</name>
  <description>
    For specific information about the contents
    consult the file(s) README.V3 on the appr
  </description>
</dataset>

<page_image>
 <id>ISRI-OCRtk/3/3_G/0148_479.3G</id>
 <path>ISRI-OCRtk/3/3_G/0148_479.3G</pat
 <vdpi>300</vdpi>
 <hdpi>300</hdpi>
 <skew></skew>
 <type_list>
     <type_list_item>
       <id>TIFF file, big-endian</id>
     </type_list_item>
     <type_list_item>
       <id>image/tiff</id>
     </type_list_item>
 </type_list>
 <contributor_list>
  <contributor_list_item>
     <id>isri-info@isri.unlv.edu</id>
  </CONTRIBUTOR_LIST_item>
 </contributor_list>
 <copyright>UNLV Copyright</copyright>
 <description>UNLV ISRI-OCRtk/3/3_G/0148
 <image_dataset_list>
   <image_dataset_list_item>
```

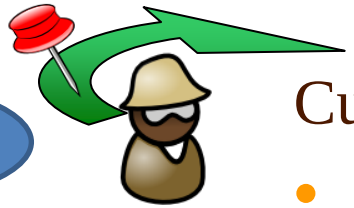


Web 2.0 Interaction



Community driven, maintained and monitored

Access to referenced data repository

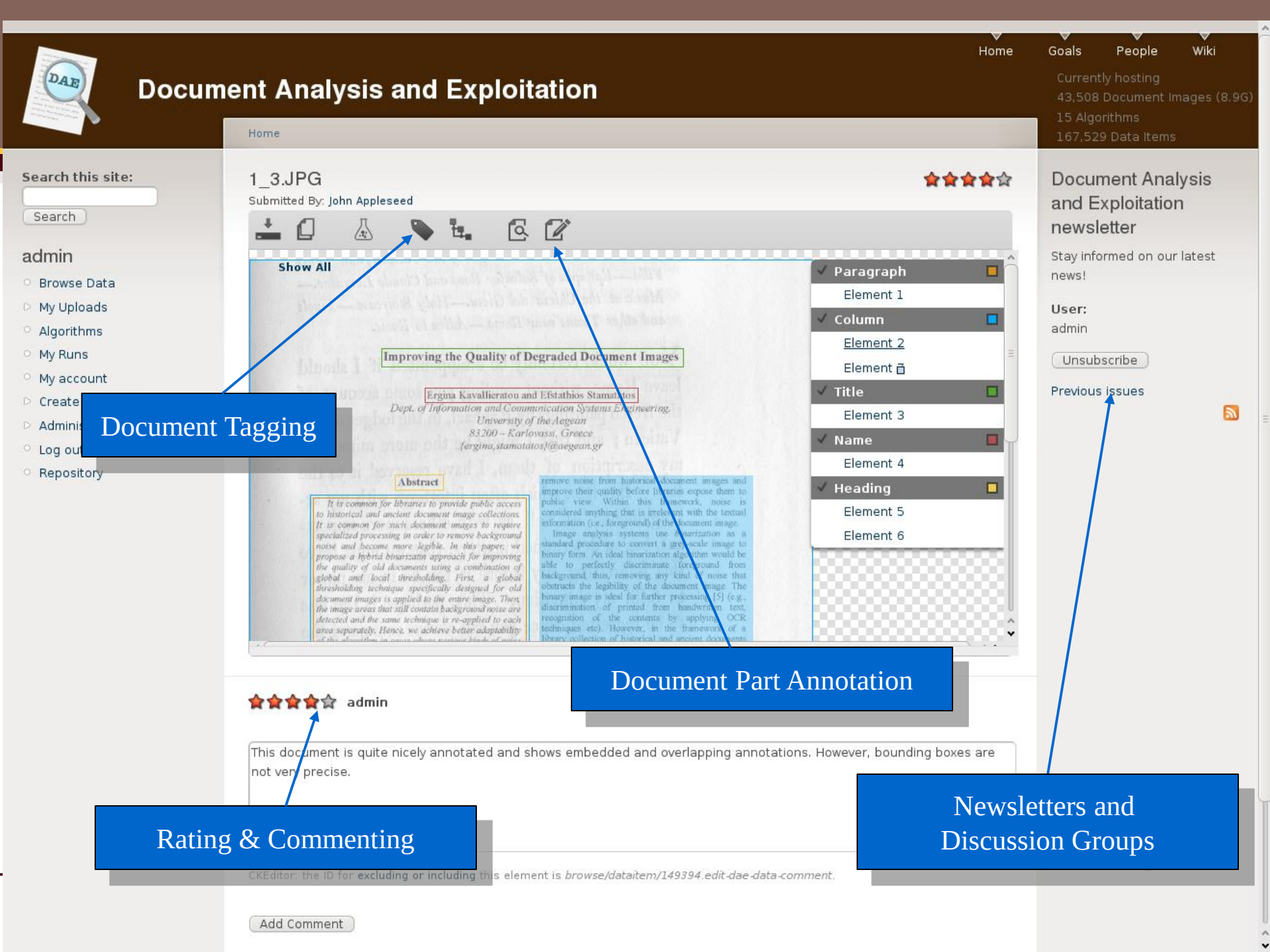


Comment, contribute and correct

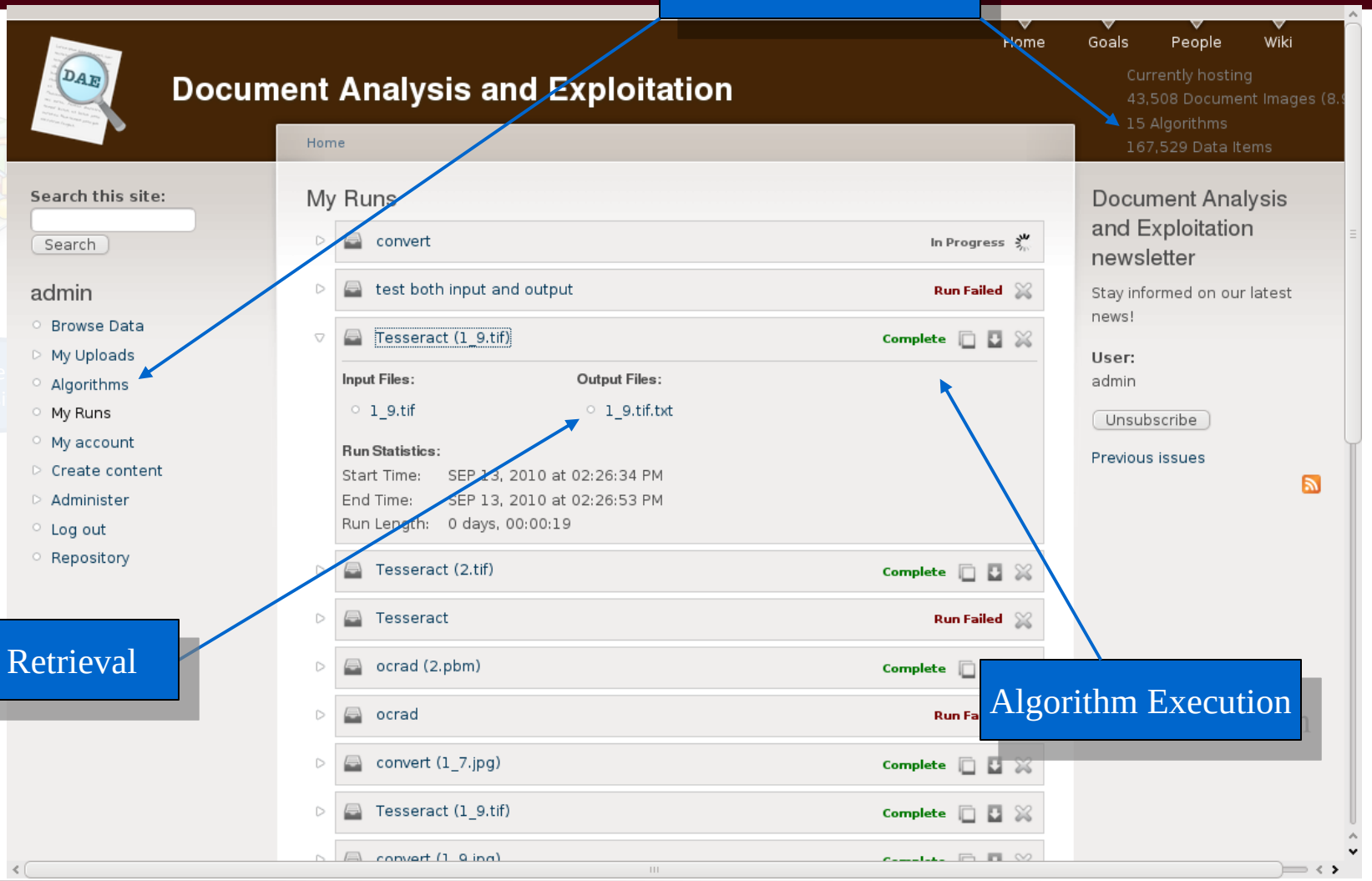
Plug into a standard evaluation process

Currently Implemented:

- Individual annotation and tagging of document sub-parts
- Commenting and rating
- Discussion forums


Document Analysis and Exploitation
Home
Goals
People
Wiki
Currently hosting
43,508 Document Images (8.9G)
15 Algorithms
167,529 Data Items
Home
Search this site:
Search
admin
Browse Data
My Uploads
Algorithms
My Runs
My account
Create
Admini
Log ou
Repository
1_3.JPG
Submitted By: John Appleseed
Show All
Improving the Quality of Degraded Document Images
Ergina Kavallieratou and Efstathios Stamatatos
Dept. of Information and Communication Systems Engineering,
University of the Aegean
83200 – Karlovassi, Greece
{ergina,stamatatos}@aegean.gr
Abstract
Paragraph
Element 1
Column
Element 2
Element
Title
Element 3
Name
Element 4
Heading
Element 5
Element 6
admin
This document is quite nicely annotated and shows embedded and overlapping annotations. However, bounding boxes are not very precise.
CKEditor: the ID for excluding or including this element is browse/dataitem/149394.edit-dae-data-comment.
Add Comment
Document Analysis and Exploitation newsletter
Stay informed on our latest news!
User:
admin
Unsubscribe
Previous issues
Document Tagging
Document Part Annotation
Rating & Commenting
Newsletters and Discussion Groups

Running Algorithms on DAE



Currently Being Implemented

Enhanced User Experience

- Improved Browsing and Retrieval
- Complex Query Interface

Resilient Queries

- Dataset Certification

Algorithm Execution Overhaul

- Full Web-based REST-WSDL Services
- Flexible Orchestration for Pipeline Execution
- Full OS Independent Virtualization

Getting Involved

Freely Downloadable Source Code (GPL)

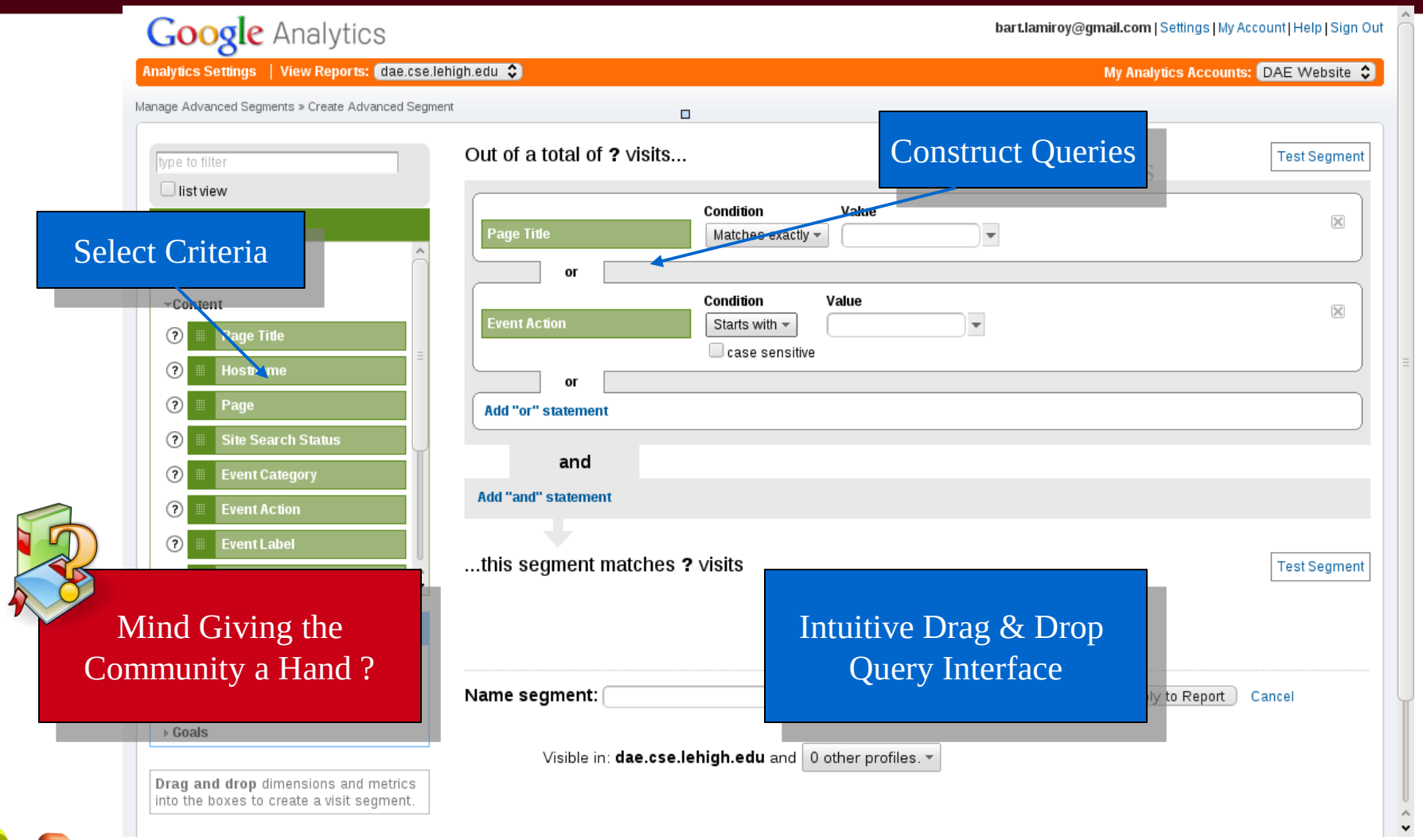
- Standalone version available (database part still a bit tricky)
- FusionForge download site coming soon (tomorrow, promised)
- Open for contributions, help and critical user feedback

Mind Giving the Community a Hand ?

Some Challenging Feature Requests:

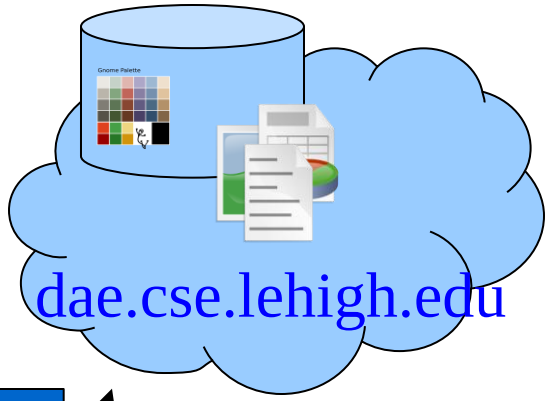
- Creating a drag-and-drop query interface
- Help making algorithm execution fully web-service enabled (and thus 100% distributed and cloud-ready)
- Import / export functions to other annotation formats

Open Challenges



Open Challenges

Making DARE Algorithm Execution Cloud-Ready

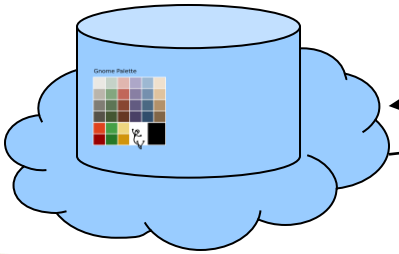


Mind Giving the Community a Hand ?

1. Query Data & Algorithm URIs

2. Connect to Algorithm URI and provide Data URIs

somewhere.org

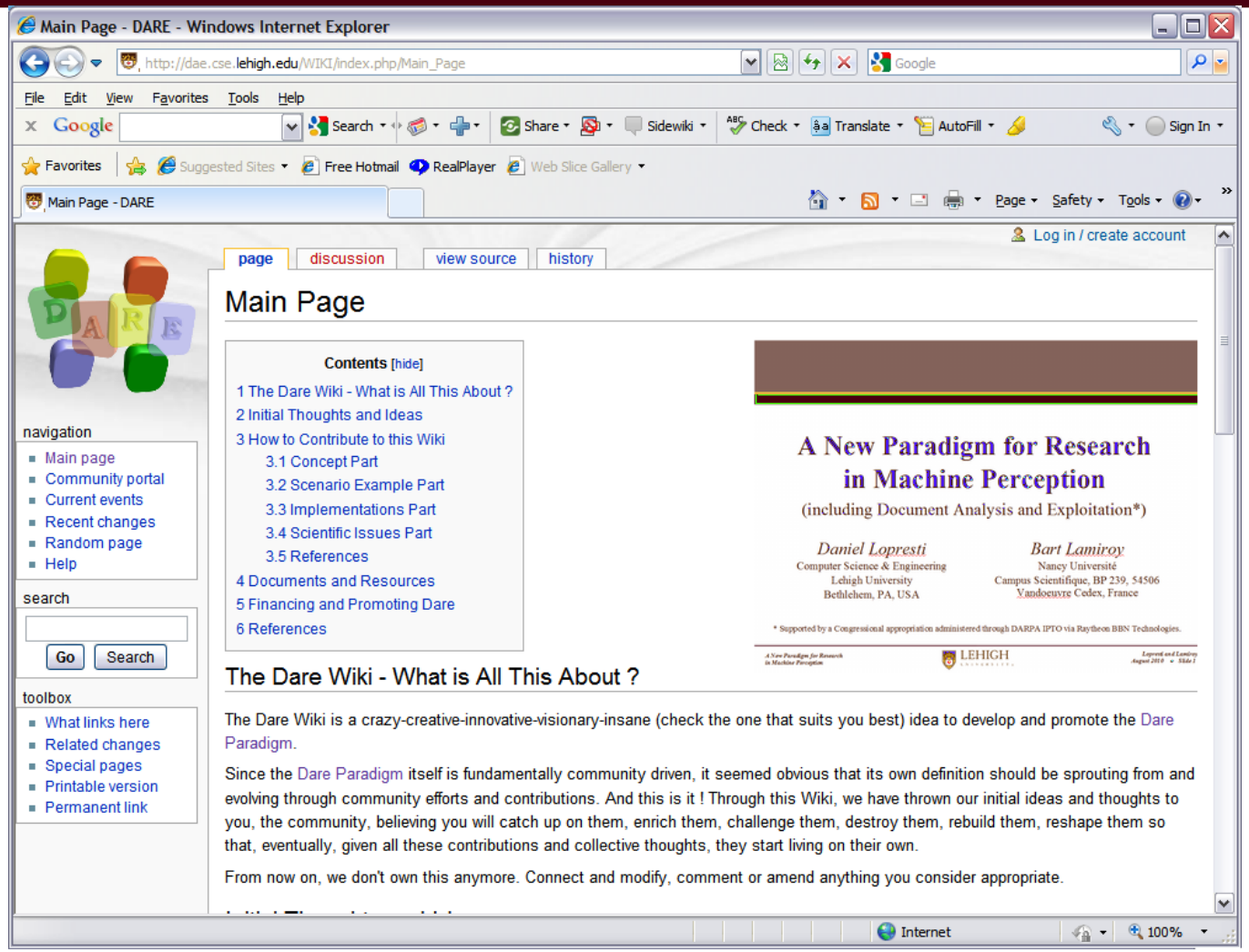


3. Data Fetching and Algorithm Execution

Further Challenges

- Guaranteeing Reproducible Queries
- Controlling Data Clutter
- Implementing a Distributed Data Model
- Maintaining Data Coherence
- Batch Data Processing and Performance
- Monitoring, Tracking and Exploiting Reputation

Add Your Thoughts to DARE Wiki



http://dae.cse.lehigh.edu/WIKI

Acknowledgements

Technical Contributors:

- Dezhao Song
- Mike Kot
- Weidong Chen
- Yingjie Li
- Yang Yu
- Xingjian Zhang
- Mike Caffrey
- Sai Lu Mon Aung

Lehigh Faculty:

- Prof. Hank Korth
- Prof. Jeff Heflin
- Prof. Brian Davison

Supporters to date:

- Prof. G. Fink (Technische Universität, Dortmund, Germany)
- Dr. D. Karatzas (Universitat Autònoma de Barcelona, Spain)
- Prof. K. Kise (Osaka Prefecture University, Japan)
- Prof. J. Lladós (Universitat Autònoma de Barcelona, Spain)
- Prof. G. Nagy (Rensselaer Polytechnic Institute, USA)
- Prof. J.M. Ogier (Université de La Rochelle, France)
- Prof. K. Tombre (INRIA, France)