



U.S. National Library of Medicine
National Center for Biotechnology Information

PubTator

Automated concept annotation for biomedical full text articles

Chih-Hsuan Wei* Alexis Allot* Robert Leaman Zhiyong Lu (PI)

<https://www.ncbi.nlm.nih.gov/research/pubtator/>

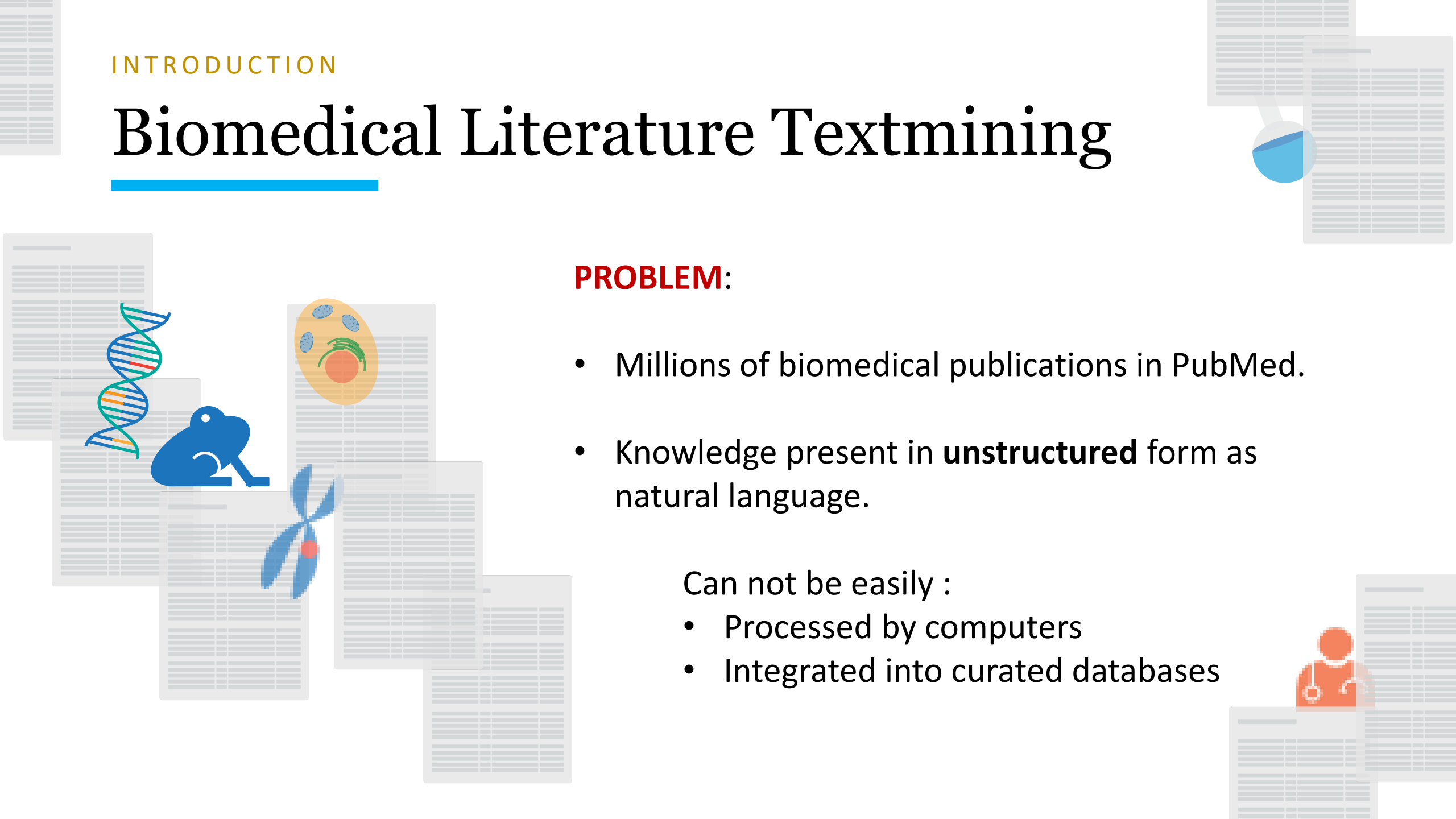
Biomedical Literature Textmining

PROBLEM:

- Millions of biomedical publications in PubMed.
- Knowledge present in **unstructured** form as natural language.

Can not be easily :

- Processed by computers
- Integrated into curated databases



Biomedical Literature Textmining



SOLUTION:

Automated **text mining** allows to easily access and **extract knowledge within the biomedical literature.**

1. For downstream text mining applications

- gene prioritization
- genetic disease analysis
- literature-based knowledge discovery

2. For faster biocuration

Ex: curating a database (such as UniProt)

INTRODUCTION

Text Mining Tools

Our team has developed many dedicated tools.

GNormPlus

Chih-Hsuan Wei, Hung-Yu Kao, Zhiyong Lu. (2015) **GNormPlus: An integrative approach for tagging genes, gene families, and protein domains.** Biomed Res Int.



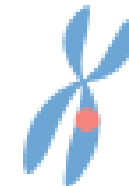
SR4GN

Chih-Hsuan Wei, Hung-Yu Kao, Zhiyong Lu. (2012) **SR4GN: a species recognition software tool for gene normalization.** PLoS One.



tmVar 2.0

Chih-Hsuan Wei, Lon Phan, Juliana Feltz, Rama Maiti, Tim Hefferon, Zhiyong Lu. (2017) **tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine.** Bioinformatics.

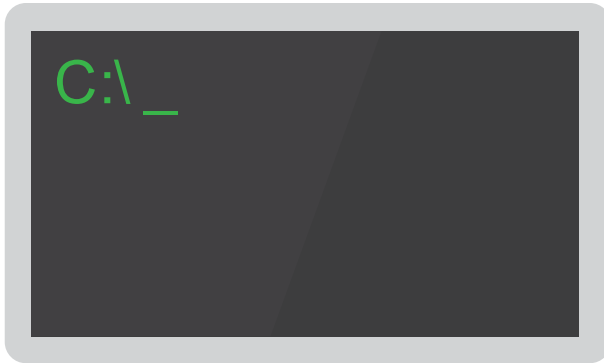


TaggerOne

Robert Leaman, Zhiyong Lu. (2016) **TaggerOne: joint named entity recognition and normalization with semi-Markov Model.** Bioinformatics.



Biomedical Literature Textmining



PROBLEM: Text-mining command line tools require bioinformatics expertise to execute them.



SOLUTION: Web-based tools can simplify distributing results from text mining systems to a wide range of users:

- **No installation or maintenance**
- **No infrastructure requirement**

What is PubTator ?

PubTator is a Web-based system providing **automatic annotations of biomedical concepts** such as genes and mutations in PubMed abstracts and PMC full-text articles.

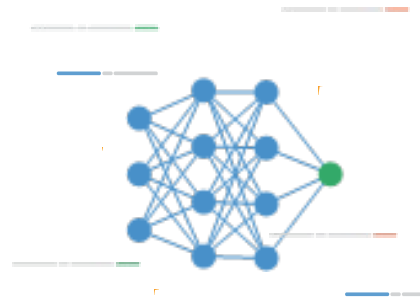
FULL-TEXT ARTICLES

PubTator includes the **full-text articles** in the PMC Open Access subset (nearly 3 million) in addition to the 30+ million abstracts in PubMed.



DEEP LEARNING

Cutting-edge machine learning and **deep learning techniques** are applied to concept disambiguation for improved accuracy.



ALWAYS UP-TO-DATE

PubTator adds **new articles every day** to always keep in sync with PubMed and PMC.



PubTator supports six concept types: **genes/proteins**, **genetic variants**, **diseases**, **chemicals** and **cell lines**

PREPROCESSING

Preprocessing Pipeline

Entities Extraction



PMC

Gene
GNormPlus

Disease
TaggerOne

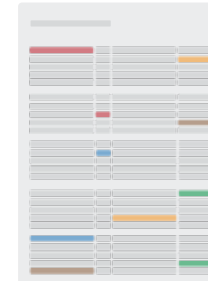
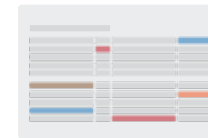
Chemical
TaggerOne

CellLine
TaggerOne

Mutation
tmVar 2.0

Species
SR4GN

Disambiguation



Daily Updates



Boc XML

Boc JSON

Pubtator

Multi-formats
generation



MongoDB

3M full-texts
30M abstracts

WEBSITE

Features



Semantic Search

Leverages PubTator annotations to find all publications mentioning an entity, regardless of which entity name the author uses

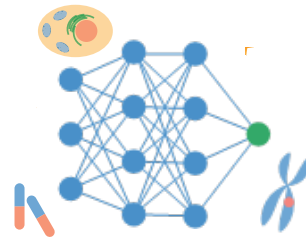
ESR1 vs **estrogen receptor**

The screenshot shows the PubTatorCentral interface. At the top, there is a search bar containing the ID '30813596' and navigation links for 'TUTORIAL', 'API', and 'FTP'. Below the search bar, there are filters for 'group by' (type) and 'sort by' (freq). The main content area displays the title 'Low-Frequency Mutational Heterogeneity of Invasive Ductal Carcinoma Subtypes: Information to Direct Precision Oncology' and the abstract text. The abstract discusses the role of low-frequency hotspot cancer-driver mutations (CDMs) in breast carcinogenesis and therapeutic response, mentioning specific CDMs like PIK3CA H1047R, KRAS G12D, and BRAF V600E. On the left side, there is a sidebar with filters for 'GENE', 'DISEASE', 'CHEMICAL', and 'MUTATION'. On the right side, there is a 'BioConcepts' panel with checkboxes for 'GENE', 'DISEASE', 'CHEMICAL', 'MUTATION', 'SPECIES', and 'CELLLINE'. Below the abstract, there is a '1. INTRODUCTION' section starting with 'Breast cancer is a heterogeneous disease, presenting with a spectrum of clinical features'.

Features

In-Document Search

A menu displays a list of bioentities in a publication, allowing users to easily **navigate** to entities of interest.



PubTatorCentral 30813596

group by type sort by freq

Search...

GENE

- PIK3CA (122)
- KRAS (88)
- HER2 (82)
- BRAF (68)
- HRAS (52)
- more

DISEASE

- IDCs (72)
- TUMOR (52)
- BREAST CANCER (47)
- BREAST CARCINOGENESIS (10)
- BREAST IDCs (4)
- more

CHEMICAL

- 5'-FLUORESCIN (5)
- POLYACRYLAMIDE (4)
- CDM (4)
- TRITON (4)
- DNTPs (4)
- more

MUTATION

- G12D (85)
- H1047R (54)
- V600E (51)
- E545K (41)
- G12V (41)
- more

Low-Frequency Mutational Heterogeneity of Invasive Ductal Carcinoma Subtypes: Information to Direct Precision Oncology

PMID30813596 PMC6429455 No AUTHORS LISTED 2019 full-text

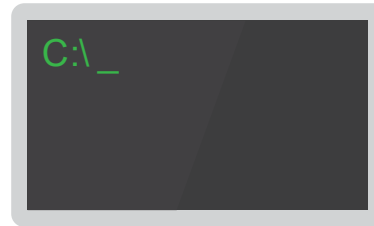
BioC XML

Information regarding the role of low-frequency hotspot cancer-driver mutations (CDMs) in breast carcinogenesis and therapeutic response is limited. Using the sensitive and quantitative Allele-specific Competitor Blocker PCR (ACB-PCR) approach, mutant fractions (MFs) of six CDMs (PIK3CA H1047R and E545K, KRAS G12D and G12V, HRAS G12D, and BRAF V600E) were quantified in invasive ductal carcinomas (IDCs; including ~20 samples per subtype). Measurable levels (i.e., $\geq 1 \times 10^{-5}$, the lowest ACB-PCR standard employed) of the PIK3CA H1047R, PIK3CA E545K, KRAS G12D, KRAS G12V, HRAS G12D, and BRAF V600E mutations were observed in 34/81 (42%), 29/81 (36%), 51/81 (63%), 9/81 (11%), 70/81 (86%), and 48/81 (59%) of IDCs, respectively. Correlation analysis using available clinicopathological information revealed that PIK3CA H1047R and BRAF V600E MFs correlate positively with maximum tumor dimension. Analysis of IDC subtypes revealed minor mutant subpopulations of critical genes in the MAP kinase pathway (KRAS, HRAS, and BRAF) were prevalent across IDC subtypes. Few triple-negative breast carcinomas (TNBCs) had appreciable levels of PIK3CA mutation, suggesting that individuals with TNBC may be less responsive to inhibitors of the PI3K/AKT/mTOR pathway. These results suggest that low-frequency hotspot CDMs contribute significantly to the intertumoral and intratumoral genetic heterogeneity of IDC, which has the potential to impact precision oncology approaches.

1. INTRODUCTION

breast cancer is a heterogeneous disease, presenting with a spectrum of clinical features

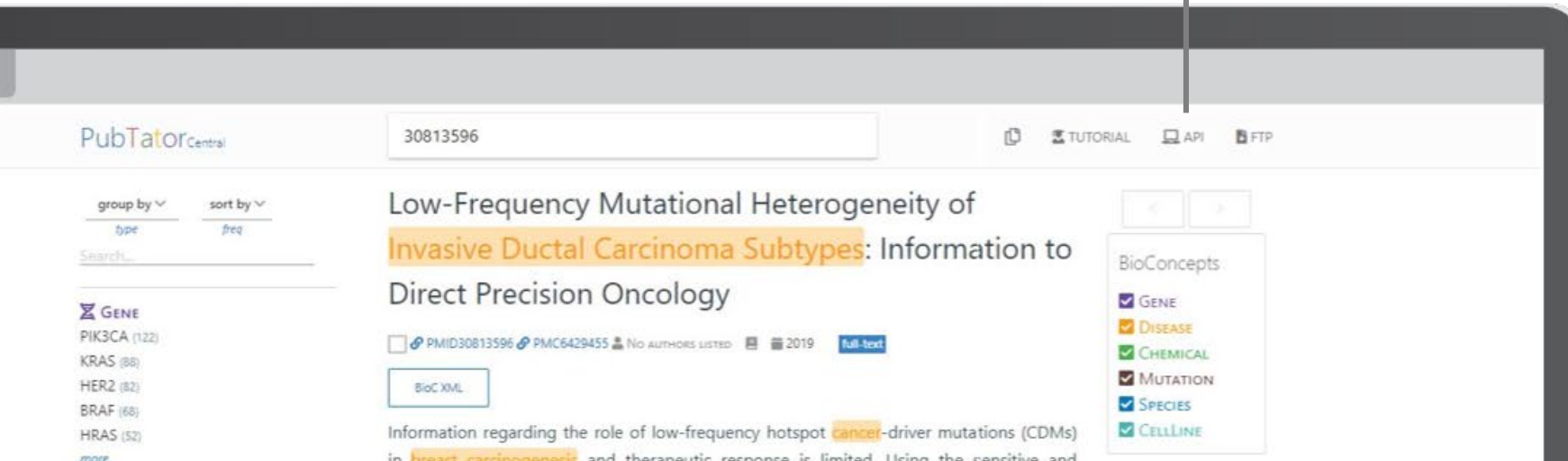
Features



Free Access

PubTator Central data is **free** and can be accessed :

- **interactively** through a web browser
- **programmatically** via **RESTful API**
- **downloaded** in bulk via **FTP**



Features

Collections

Articles can be **organized** into collections, and then be **viewed** or **downloaded** together.

Articles may be added by:

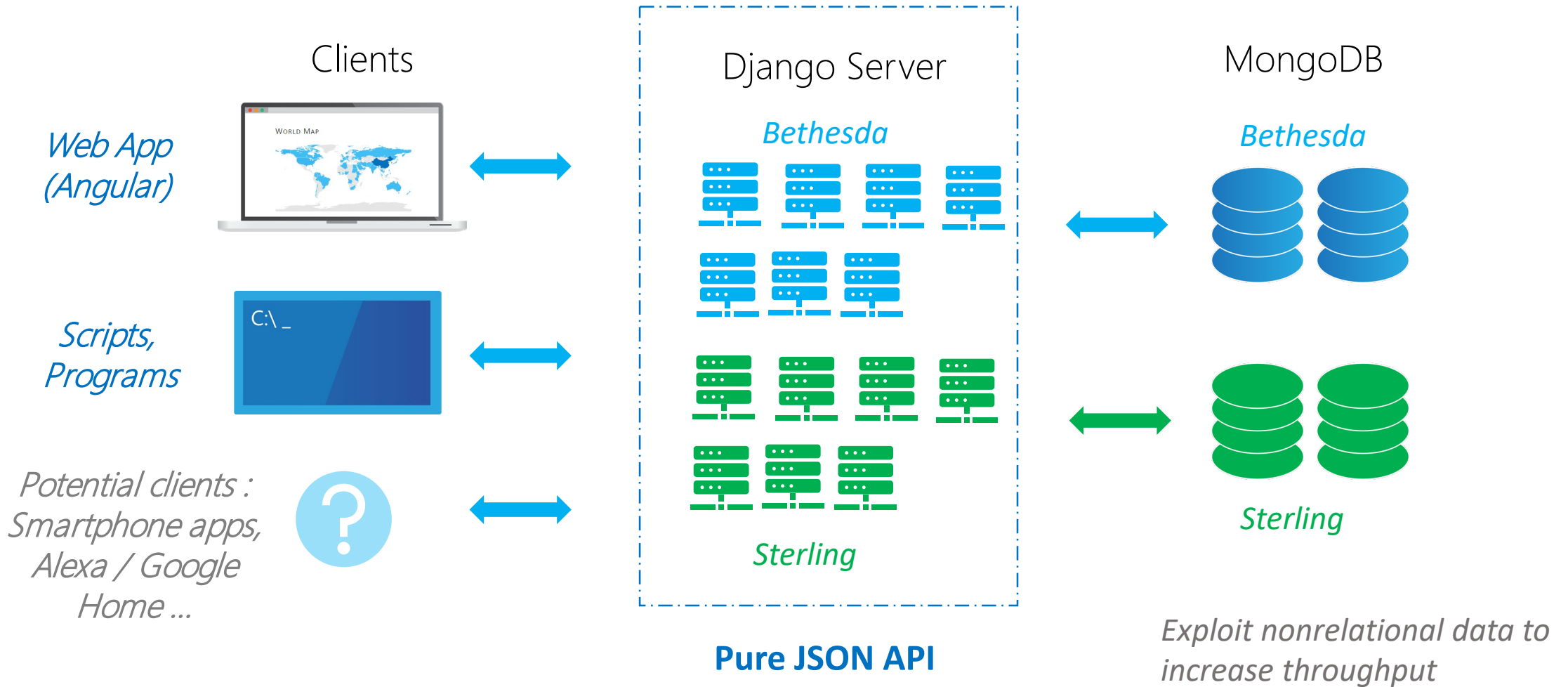
- **Selecting** specific publications
- Entering a **query**
- Entering a list of **PMIDs**



The screenshot displays the PubTator Central interface. At the top, the search bar contains the PMID 30813596. The article title is "Low-Frequency Mutational Heterogeneity of Invasive Ductal Carcinoma Subtypes: Information to Direct Precision Oncology". The article is categorized under "DISEASE" and "GENE". The abstract text is visible, mentioning "breast carcinogenesis" and "therapeutic response". The interface includes navigation options like "TUTORIAL", "API", and "FTP", and a "BioConcepts" sidebar with filters for GENE, DISEASE, CHEMICAL, MUTATION, SPECIES, and CELLLINE.

WEBSITE

(Future proof) Architecture



USAGE

Usage



744M

Total API hits

USE CASE 1 : PRIORITIZING PROTEINS ASSOCIATED WITH GENETIC MUTATIONS IN CANCERS

“The phrase “mutation cancer” was used as the search term to retrieve the commonly mutated genes tagged by PubTator. The proteins co-published with each of the identified genetic mutations were retrieved, respectively.”

Yu KH, Lee TM, Wang CS, et al. Systematic Protein Prioritization for Targeted Proteomics Studies through Literature Mining. J Proteome Res. 2018;

USE CASE 2 : UNIPROT DATABASE CURATION

“With the assistance of the PubTator text-mining tool, we tagged more than 10000 articles to assess the ratio of papers relevant for curation.”

Poux S, Arighi CN, Magrane M, et al. On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. Bioinformatics. 2017;

CONCLUSIONS

Conclusions

PubTator is a web-based system for automated concept annotations in PubMed abstracts and PMC-TM full text articles

FTP (Updated monthly)

<ftp://ftp.ncbi.nlm.nih.gov/pub/lu/PubTatorCentral/>

API (Updated daily)

<https://www.ncbi.nlm.nih.gov/research/pubtator-api/publications/export/biocxml?full=true&pmids=30375428>

Raw Text annotation service

<https://www.ncbi.nlm.nih.gov/research/pubtator/api.html>

Sample client codes in multiple languages : python, java and perl

<https://www.ncbi.nlm.nih.gov/research/pubtator/api.html>

<https://www.ncbi.nlm.nih.gov/research/pubtator/>