

# Project Suggestions

# Enterprise Search

- Omar's pitch on the 1<sup>st</sup> class
  - People in industry (usually) don't talk about stuff they are really working on
  - But there is more talk than you might expect
    - <https://www.amazon.com/Regional-Advantage-Culture-Competition-Silicon/dp/0674753402>
- Enterprise Search: Some history of failures
  - [https://en.wikipedia.org/wiki/Google\\_Search\\_Appliance](https://en.wikipedia.org/wiki/Google_Search_Appliance)
  - Enterprise search is harder than web search because of network effects
    - [https://en.wikipedia.org/wiki/Metcalfe%27s\\_law](https://en.wikipedia.org/wiki/Metcalfe%27s_law)
    - When benefits scale with edges (links between pages)
      - and costs scale with nodes (web pages)
      - then rich get richer (easier to find good stuff in larger graphs than smaller graphs)
      - (There are more links to good stuff in larger graphs)

# Conference Automation

- Routing of submissions to reviewers
- <https://blog.allenai.org/conference-peer-review-with-the-semantic-scholar-api-24ab9fce2324>
- Current status:
  - Many venues use routing tools:
    - <https://openreview.net/>
    - <https://easychair.org/>
    - <https://www.softconf.com/>
    - <https://cmt3.research.microsoft.com/>
  - Not clear these tools are safe and effective (better than manual routing)
  - Tools are too hard for program committees (see blog above)

# Recommender Systems

- I regularly receive spam suggestions
  - Rarely credible
  - At best, recommendations are
    - recent,
    - buzz-word compliant
    - but not credible
  - Credible (Seminal) >> Recent

Web of Science



Greetings! Your work has been cited.

[View citing publications](#)

2 publications have cited your work since Jan 10th 2023.

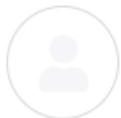
[A review on microwave band pass filters: Materials and design optimization techniques for wireless communication systems](#)

Krishna, V. Neeraj; Padmasine, K. G.  
Materials Science In Semiconductor Processing

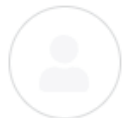
Microwave band pass filters play an important function in radio frequency for suppressing out of band emissions that are necessary



**Kenneth Ward Church**  
222 Publications • 18,486  
Citations • Computer  
Science



**K. Church**  
21 Publications • 940  
Citations • Materials  
Science



**K. Church**  
49 Publications • 616  
Citations • Materials  
Science



SEMANTIC SCHOLAR

Hi Kenneth, we found 10 new papers for you in the past day.

From Your Feed

## Your Papers

### Phoneme-Level BERT for Enhanced Prosody of Text-to-Speech with Grapheme Predictions

Yinghao Aaron Li, Cong Han, ... N. Mesgarani

**TLDR** Subjective evaluations show that the phoneme-level BERT encoder has significantly improved the mean opinion scores (MOS) of rated naturalness of synthesized speech compared with the state-of-the-art (SOTA) StyleTTS baseline on out-of-distribution (OOD) texts.

Save Not Relevant

### From English to More Languages: Parameter-Efficient Model Reprogramming for Cross-Lingual Speech Recognition

Chao Yang, ...Tara N. Sainath, ... Trevor Strohman

In this work, we propose a new parameter-efficient learning framework based on neural model reprogramming for cross-lingual speech recognition, which can...

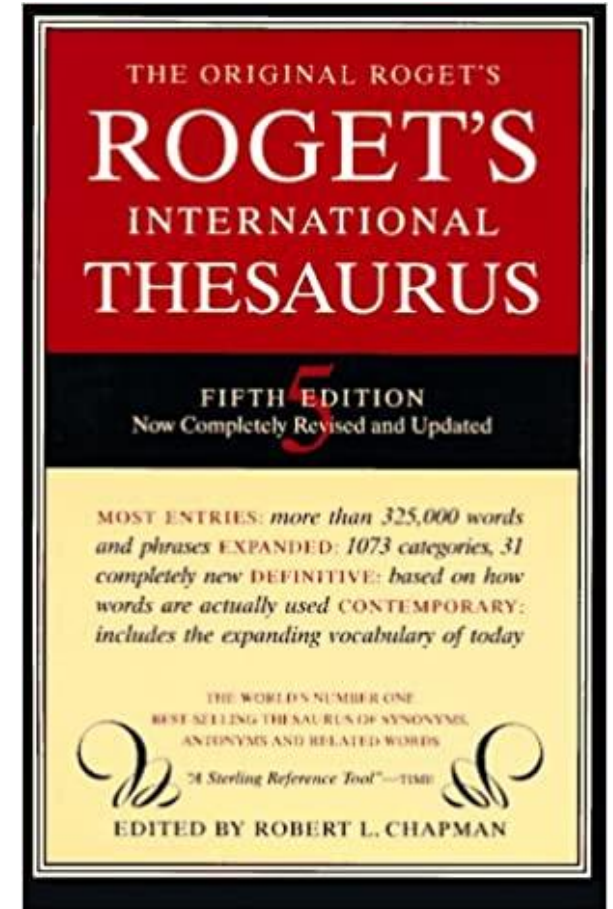
Tara N. Sainath authored 10 papers you cited Save Not Relevant

Not useful/credible  
(18k over 50 years >> 2/week)

Semantic Scholar

# Thesaurus → Deep Nets (ChatGPT)

- Good writers use resources
  - Dictionaries, Thesauruses, etc.
- Active vocab << Passive vocab
  - Hard to find the right word,
  - But easier when prompted
  - (Especially in 2<sup>nd</sup> language)
- So too, there may be a role for deep nets
  - <https://www.npr.org/2023/01/24/1151160196/how-to-stop-worrying-and-love-or-at-least-live-with-chatgpt>
- Deep nets are better on fluency than truth
  - <https://www.cambridge.org/core/journals/natural-language-engineering/article/gpt3-whats-it-good-for/0E05CFE68A7AC8BF794C8ECBE28AA990>
  - “Unreliable doesn’t mean useless”



# Nanny Cameras for Young and Old

- In China, massive migration from rural areas to cities
  - Cities are not designed for families
  - Workers move to cities, and leave children with their parents in rural areas
  - Children miss their parents (might see them just once a year at New Year)
  - Workers can do 30 minutes of homework per day before calling home
  - Can we summarize a day of video down to 30 minutes?
- In much of the world, seniors are living alone, far from their children
  - Children would like to know that their parents are ok
  - If parents fall, or need help, could nanny cam take appropriate action
    - No need for wake-up words
    - My father couldn't remember ``Alexa''
    - If he fell, he wouldn't think to ask for help (because that wasn't a thing in his day)

# Where is my phone? My kid? My father?

- Where did I leave my keys?
- Where are my eyeglasses?
  - You shouldn't need your eyeglasses to find your eyeglasses...
- Where is that book I was reading a few years ago?

# Where did my day go?

- Can you help me be more productive?
- Time and motion study of my life
- <https://en.wikipedia.org/wiki/MyLifeBits>



# Workflow

- Example scenario: Access to medical expertise
  - Better hospitals in urban areas than rural areas in China (and everywhere)
  - Medical expertise tends to be concentrated near top medical schools
  - Biopsies from rural areas are digitized and sent to experts
  - No need for expertise to be physically near patients
- Traditionally, Amazon Mechanical Turk is used for low-skill tasks
  - Is there a Human-in-the-Loop opportunity for high-skill task?
- If so, can we log workflows to collect data for machine learning?
- Can we create exchanges like Ad exchange, Futures exchange (FTX)
  - [https://en.wikipedia.org/wiki/Ad\\_exchange](https://en.wikipedia.org/wiki/Ad_exchange)

# BOTUS

<https://www.npr.org/transcripts/522897876>

- Today on the show PLANET MONEY builds a robot,
  - a bot to trade stocks with real money.
- The official Twitter handle, yeah. BOTUS, bot of the United States.
- Task
  - Input: Trump Tweets
  - Output: Trades
- Technologies
  - Crawl tweets
    - Filter for Trump tweets
  - Sentiment analysis
    - Tweet → Buy / Sell
  - Named Entity Recognition (NER)
    - Tweet → Stock
- Spoiler Alert:
  - They lost \$\$

# HuggingFace Tutorials

- Sentiment Analysis
  - <https://huggingface.co/blog/sentiment-analysis-python>
- Named Entity Recognition (NER)
  - <https://huggingface.co/dslim/bert-base-NER>
  - <https://huggingface.co/course/chapter7/2>
- General Fine-Tuning (GFT)
  - <https://github.com/kwchurch/gft>

# PubTator

<https://www.ncbi.nlm.nih.gov/research/pubtator/?view=docsum&query=PMC6982432>

PubTator

PMC6982432

Q

NIH>> NLM

MENTIONS

group ▼ sort ▼

type freq

Search...

GENE

ESR1 100

ER 14

HER2 8

KIT 4

PR 3

more

DISEASE

MBC 53

TUMOR 19

INVASIVE DUCTAL CARCINOMA 6

DEATH 2

STRUCK TUBES 1

more

CHEMICAL

PBC 5

MUTATION

Y537S 12

D538G 7

PMID32021303 • PMC6982432

2020

**Prevalence of ESR1 Mutation in Chinese ER-Positive Breast Cancer**

Zhu W, Ren C ... Liao N • Onco Targets Ther. 2020 Jan 21

BiocXML

Background

ESR1 mutation and its possible relation to endocrine therapy resistance in ER-positive breast cancers have been studied with respect to genetic sequencing data from Western patients but rarely from Chinese patients. This study aimed to investigate the prevalence of ESR1 mutation in Chinese primary and metastatic ER-positive breast cancer.

Methods

Tumor samples from 297 primary breast cancer (PBC) patients and blood samples from 43 metastatic breast cancer (MBC) patients were obtained to perform whole exon sequencing of the ESR1 gene through next-generation sequencing (NGS). Clinicopathological features of MBC patients were listed and grouped to explore potential factors in ESR1 mutations.

Results

☒ BIOCONCEPTS

☒ GENE

☒ DISEASE

☒ CHEMICAL

☒ MUTATION

☒ SPECIES

☒ CELLLINE

NAVIGATION

[TITLE](#)

[INTRODUCTION](#)

[MATERIALS AND METHODS](#)

[RESULTS](#)

[DISCUSSION](#)

[CONCLUSIONS](#)

[FUNDING](#)

[DISCLOSURE](#)

# Medical Resources

- PubTator: Abstracts (papers) with annotations (entities)
  - <https://www.ncbi.nlm.nih.gov/research/pubtator/?view=docsum&query=PMC6982432>
- Pubmed: Abstracts (papers) with annotations (MeSH)
  - <https://www.ncbi.nlm.nih.gov/mesh/>
- MeSH: Medical Subject Headings (Ontology / Knowledge Graph)
  - <https://www.ncbi.nlm.nih.gov/mesh/>

# MeSH

<https://meshb.nlm.nih.gov/search>

### Medical Subject Headings 2023

The files are updated each week day Monday-Friday by 8AM EST

FullWord ▾

Exact Match

All Fragments

Any Fragment

Sort by: Relevance ▾

Results per Page: 20 ▾

☐ All Terms

☒ Main Heading (Descriptor) Terms

☐ Qualifier Terms

☐ Supplementary Concept Record Terms

☐ MeSH Unique ID

☐ Search in all Supplementary Concept Record Fields

☐ Heading Mapped To

☐ Indexing Information

☐ Pharmacological Action

☐ Search Related Registry and CAS Registry/EC Number/UNII Code/NCBI Taxonomy ID Number (RN)

☐ Related Registry Search

☐ CAS Registry/EC Number/UNII Code/NCBI Taxonomy ID Number (RN)

☐ Search in all Free Text Fields

☐ Annotation

☐ ScopeNote

☐ SCR Note

<https://meshb.nlm.nih.gov/record/ui?ui=D001943>

## Breast Neoplasms MeSH Descriptor Data 2023

Details	Qualifiers	MeSH Tree Structures	Concepts
<b>MeSH Heading</b>	Breast Neoplasms		
<b>Tree Number(s)</b>	C04.588.180 C17.800.090.500		
<b>Unique ID</b>	D001943		
<b>RDF Unique Identifier</b>	<a href="http://id.nlm.nih.gov/mesh/D001943">http://id.nlm.nih.gov/mesh/D001943</a>		
<b>Annotation</b>	human only; BREAST NEOPLASMS, MALE is also available; for animal, index MAMMARY NEOPLASMS, ANIMAL or MAMMARY NEOPLASMS, EXPERIMENTAL; coordinate IM with histological type of neoplasm (IM)		
<b>Scope Note</b>	Tumors or cancer of the human BREAST.		
<b>Entry Version</b>	BREAST NEOPL		
<b>Entry Term(s)</b>	Breast Cancer Breast Carcinoma Breast Tumors Cancer of Breast Cancer of the Breast Human Mammary Carcinoma Malignant Neoplasm of Breast Malignant Tumor of Breast Mammary Cancer Mammary Carcinoma, Human Mammary Neoplasm, Human Mammary Neoplasms, Human Neoplasms, Breast Tumors, Breast		
<b>See Also</b>	<a href="#">Breast Cancer Lymphedema</a>		
<b>Date Established</b>	1966/01/01		
<b>Date of Entry</b>	1999/01/01		
<b>Revision Date</b>	2018/06/14		

# Many start-up companies in this space

- Use cases
  - Personalized medicine:
    - Many rare diseases
    - Literature is large (and growing quickly)
    - Doctors can't keep up
    - Opportunity to help doctors find papers relevant to patient
  - Bloomberg-terminal for drug companies
  - Electronic Health Records
    - Map records to codes (for insurance purposes)
      - Maximize bills (for doctors)
      - Minimize bills (for insurance companies)

# Wikipedia search

- Can we improve Wikipedia search?
- What would be an alternative to 10-blue links for Wikipedia?
- [https://en.wikipedia.org/wiki/Wikipedia:Database\\_download](https://en.wikipedia.org/wiki/Wikipedia:Database_download)



# Q&A systems

- Question-answering system exist from the early days.
- Quora, StackOverflow, etc.
- ChatGPT is a new incarnation
- What to do when there are potentially alternative answers?
- How to organize non-factoid answers

# Labeling

- Ground truth is a key component of ML projects
- Current tools like Mechanical Turk are somewhat primitive
- Are there new alternatives for content moderation tasks?
- Further reading
  - <https://www.wired.com/2014/10/content-moderation/>
  - <https://time.com/6247678/openai-chatgpt-kenya-workers/>

Exclusive: OpenAI Used Kenyan Workers on  
Less Than \$2 Per Hour to Make ChatGPT Less  
Toxic

ADRIAN CHEN BACKCHANNEL OCT 23, 2014 6:38 AM

## The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed

Inside the soul-crushing world of content moderation, where low-wage workers soak up the worst of humanity and keep it off your screens.

# Reddit

- Lots of interesting content on Reddit
- Not well organized and not easy to find
- Perception that Google's search is not as good as it used to be
- How would you organize Reddit data for better access?
- <https://dkb.io/post/google-search-is-dying>
- <https://paperswithcode.com/dataset/reddit>



Why People Are Adding REDDIT To Their Google Searches