# PyPharma NLP Workshop 2019
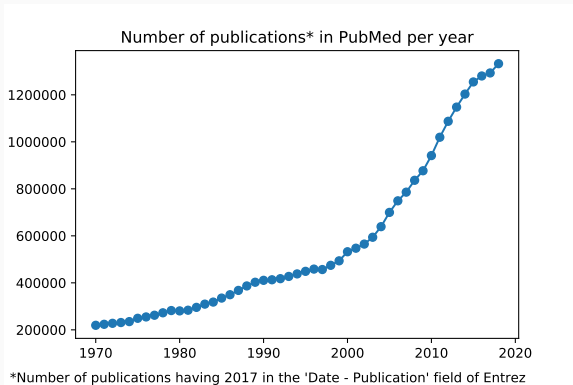
Diego Saldana

November 2019

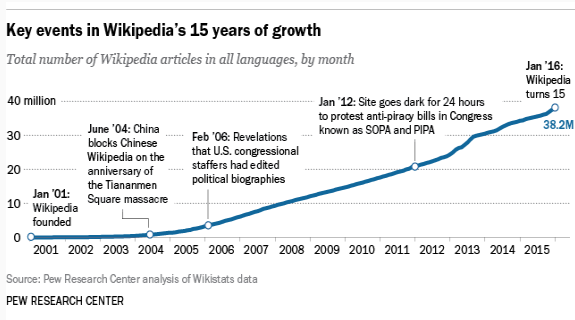# Introduction to Biomedical NLP

# Why Biomedical NLP?

- Most of the information out there is in the form of natural language: scientific papers, clinical notes, social media, textbooks, lectures, websites.



Number of publications* in PubMed per year

*Number of publications having 2017 in the 'Date - Publication' field of Entrez

- Most of the information out there is in the form of natural language: scientific papers, clinical notes, social media, textbooks, lectures, websites.

**Key events in Wikipedia's 15 years of growth**

*Total number of Wikipedia articles in all languages, by month*

Jan '16: Wikipedia turns 15

38.2M

Jan '12: Site goes dark for 24 hours to protest anti-piracy bills in Congress known as SOPA and PIPA

June '04: China blocks Chinese Wikipedia on the anniversary of the Tiananmen Square massacre

Feb '06: Revelations that U.S. congressional staffers had edited political biographies

Jan '01: Wikipedia founded

40 million
30
20
10
0

2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015

Source: Pew Research Center analysis of Wikistats data

**PEW RESEARCH CENTER**

## Why Biomedical NLP?

- Most of the information out there is in the form of natural language: scientific papers, clinical notes, social media, textbooks, lectures, websites.
- This information is potentially very useful but cannot readily be used programmatically and be stored in databases, searched, or analyzed.
- As a result this valuable information is "locked into a vault" until a human reads it, structures it and puts it into some database.
- And even when that happens, the scope in which the data can be used is usually limited and chosen by the extractors.

## Agenda

- Biomedical NLP 101: Bags of words (30 mins)
- Deep Learning for Biomedical NLP (30 mins)

  - Language Modelling (30 mins)
  - Text Classification (30 mins)
  - Named Entity Recognition (30 mins)
  - Question Answering (30 mins)
  - Integrating NLP into survival models (30 mins)