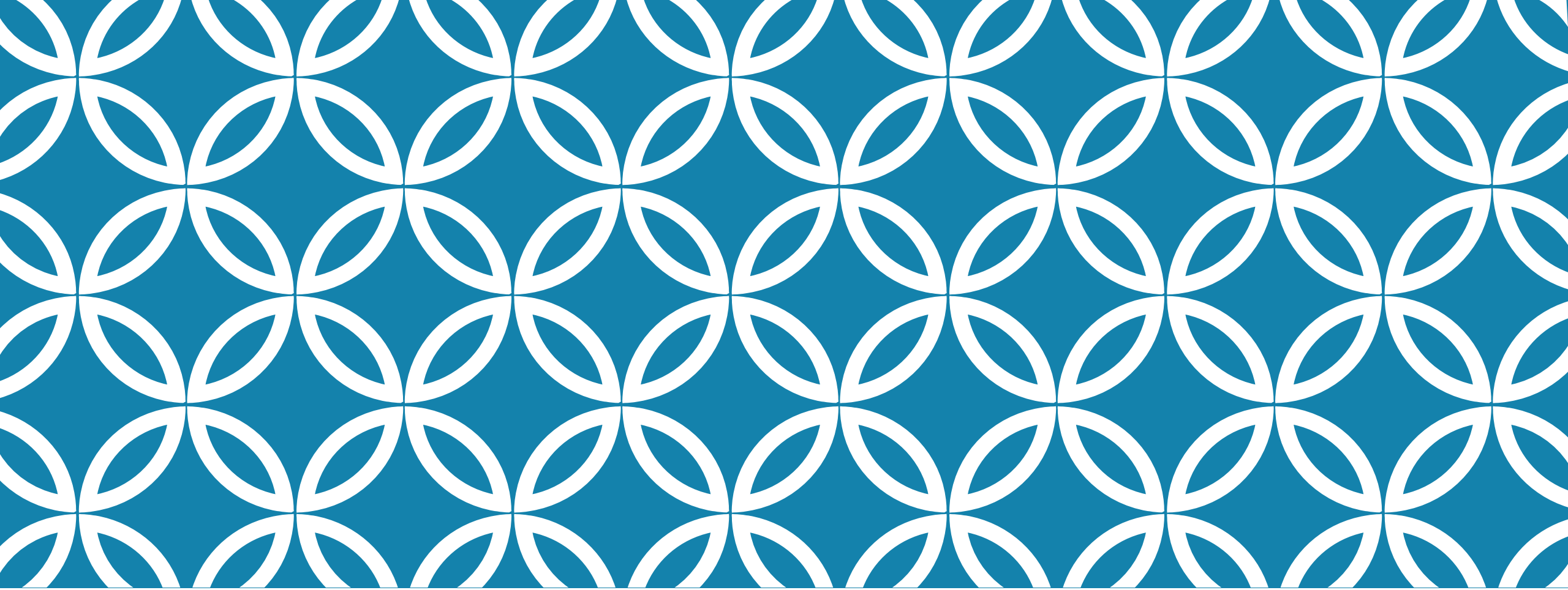


FINAL PROJECT

Predicting Biomedical Innovation

MaryJo Zaborowski

DS 04222019
August 2019

INSPIRATION

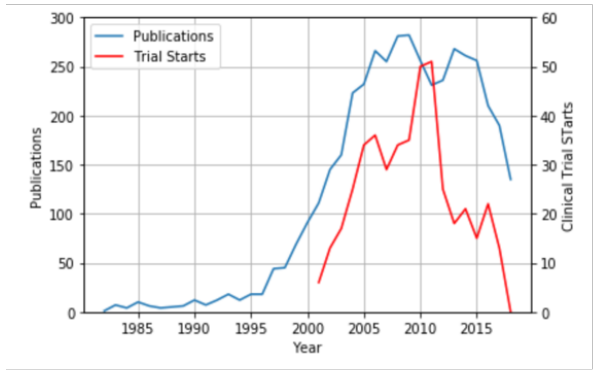
We noticed a pattern in scientific literature.

When biomedical innovation happens, there is an almost-exponential increase in publications, followed by many clinical trial starts.

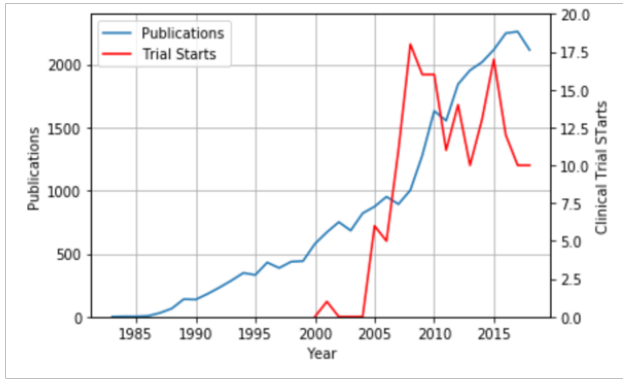
If innovation can be predicted from publication velocity...

- Investors can predict stock price increases
- Wise biomedical companies can gain early-mover advantage
- Patients can benefit from accelerated innovation

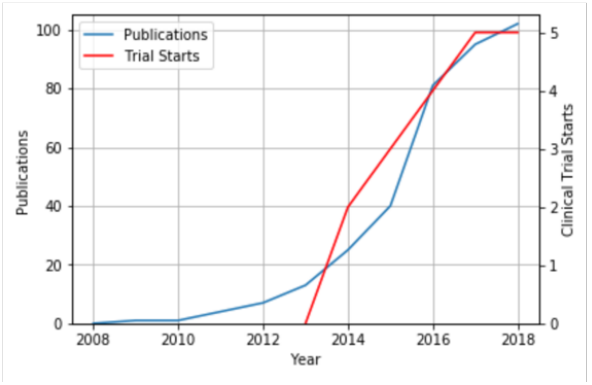
SSRI for Depression



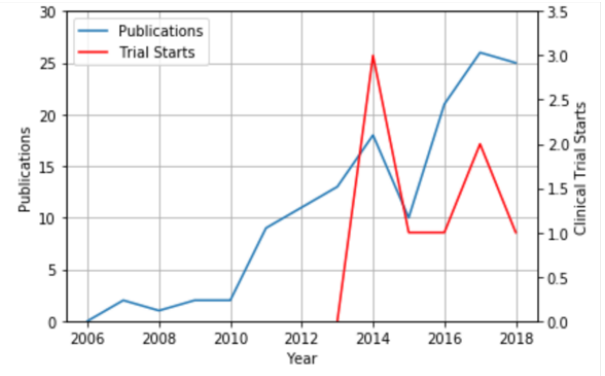
Amyloid for Alzheimer's



CAR-T for Cancer



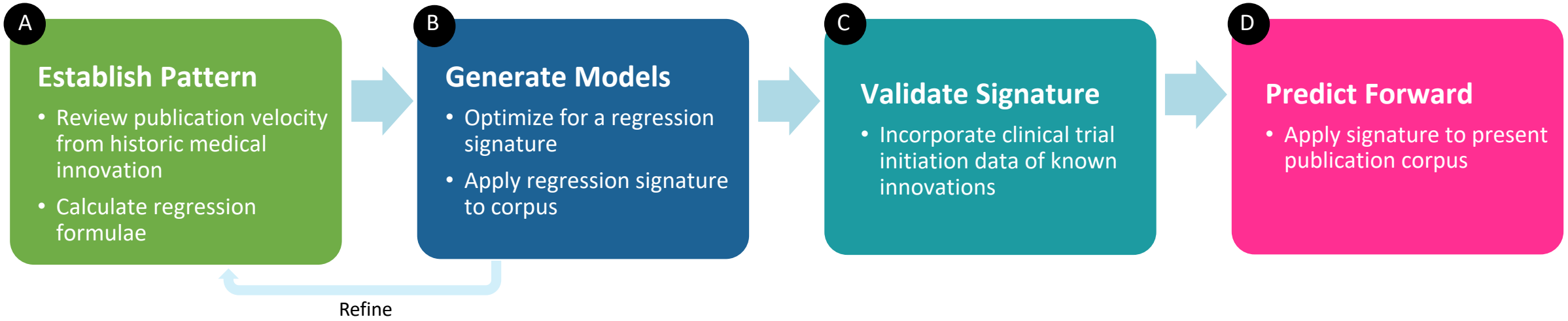
JAK for Psoriasis



PROJECT

Predict biomedical innovation from publication literature

Objective: Develop a trend signature through regression approaches to automate the identification of investment-worthy scientific breakthroughs



DATA

Public resources of biomedical literature and biomedical clinical trials

PubMed comprises more than 29 million citations for biomedical literature from MEDLINE, life science journals, and online books.

https://www.ncbi.nlm.nih.gov/pubmed

ClinicalTrials.gov, established in 2000, is a registry of clinical trials information for both federally and privately funded trials conducted under investigational new drug applications. IT contains information for over 310,000 trials. https://clinicaltrials.gov

Reference for drug targets: A comprehensive map of molecular drug targets, Santos et al, Nat Rev Drug Discov. 2017 Jan; 16(1): 19–34.

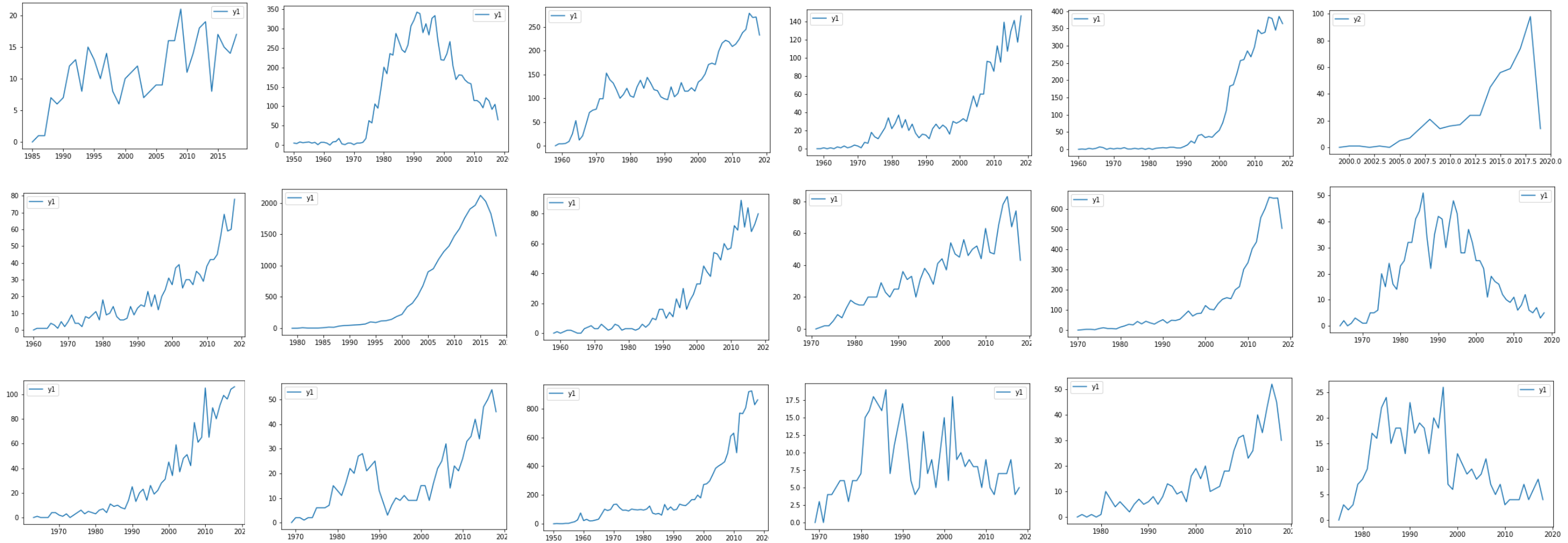
PUBLICATION VELOCITY

Over 400 Target-Indication Graphs Evaluated

Evaluated Through:

Model 1*: $y=ae^{-bx} + cx$

Model 2: sliding three-period slope



* Courtesy of Michael Carlisle

RESULTS

Model 1*: $y = ae^{-bx} + cx$

	TI_Pair_Data	exp_lin_reg_result	exp_interest_rate	lin_interest_rate	formula
0	ALK_Lung_Cancer	[3.0, 233.89433467, 27.12807883]	2.338943e+02	27.128079	3.0e^(-233.89x) + 27.13x
1	PDL1_NSCLC	[0.99999999, 105.12396945, 11.82577034]	1.051240e+02	11.825770	1.0e^(-105.12x) + 11.83x
2	Checkpoint	[1.00000001, 214.45576576, 33.42857144]	2.144558e+02	33.428571	1.0e^(-214.46x) + 33.43x
3	CART_Cancer	[1.00000001, 56.71391299, 8.69607843]	5.671391e+01	8.696078	1.0e^(-56.71x) + 8.7x
4	LiquidBiopsy	[1.0, 245.69957927, 36.28571433]	2.456996e+02	36.285714	1.0e^(-245.7x) + 36.29x
5	SSRI_Depr	[112.245059, 4.28290434e-07, -1.63016597]	4.282904e-07	-1.630166	112.25 + -1.63x
6	Serpin-C1_anticoagulant	[5.0007223, 3.98230284, 3.85346712]	3.982303e+00	3.853467	5.0e^(-3.98x) + 3.85x
7	ALPPL_radiation	[1.00000001, 53.42888388, 7.22691445]	5.342888e+01	7.226914	1.0e^(-53.43x) + 7.23x
8	ACHE_arrhythmia	[4.0, 26.94428673, 4.2070614,]	2.694429e+01	4.207061	4.0e^(-26.94x) + 4.21x
9	ACHE_nootropic	[1.0, 79.79056161, 0.93151403]	7.979056e+01	0.931514	1.0e^(-79.79x) + 0.93x
10	Abeta_Schizophrenia	[1.0, 455.01030565, 1.42809122]	4.550103e+02	1.428091	1.0e^(-455.01x) + 1.43x
11	KIT_antineoplastic	[0.99999999, 30.8946471, 4.0652239,]	3.089465e+01	4.065224	1.0e^(-30.89x) + 4.07x
12	ACHE_glaucoma	[1.00828754, 2.24536424, 0.74880017]	2.245364e+00	0.748800	1.01e^(-2.25x) + 0.75x
13	HPRT_antineoplastic	[2.0, 53.2716796, 0.57847813]	5.327168e+01	0.578478	2.0e^(-53.27x) + 0.58x
14	ACHE_dryMouth	[1.0, 289.77329865, 1.29959443]	2.897733e+02	1.299594	1.0e^(-289.77x) + 1.3x
15	C5_inflammation	[2.08575945, 1.05487533, 0.73964979]	1.054875e+00	0.739650	2.09e^(-1.05x) + 0.74x

* Courtesy of Michael Carlisle

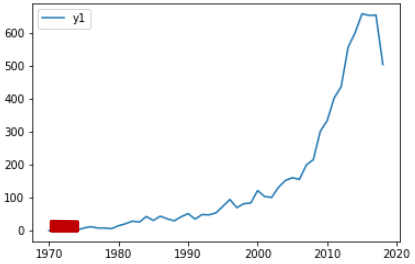
Initial estimation was to regress to an exponential function with a high growth rate: $y = Ce^{rx}$

After review, the constant was dropped and the pure exponential function was modified with a linear parameter

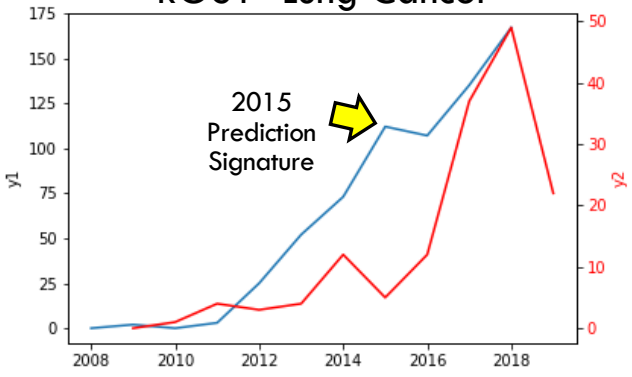
This predictive model was not yet tested

RESULTS

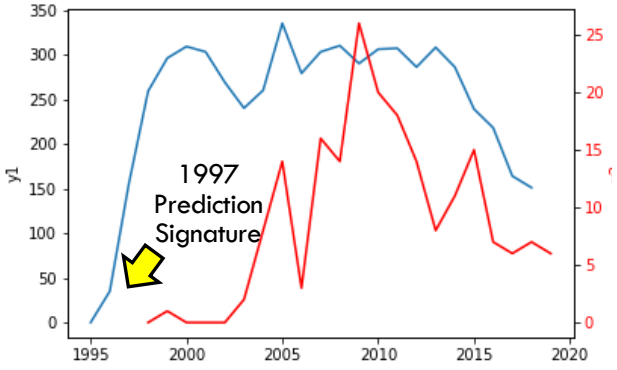
Model 2: sliding three-period slope



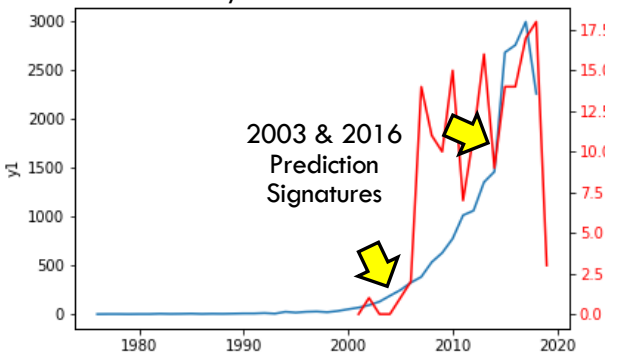
ROS1- Lung Cancer



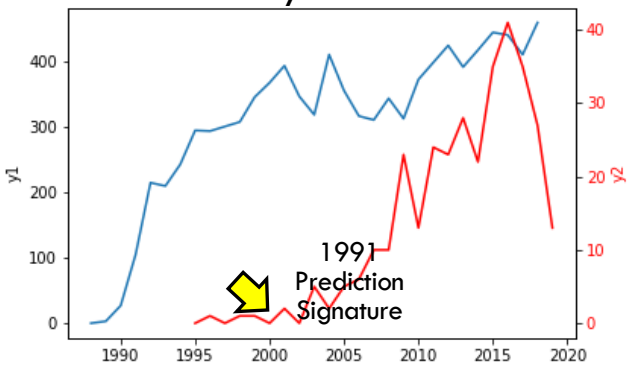
CCR5-HIV



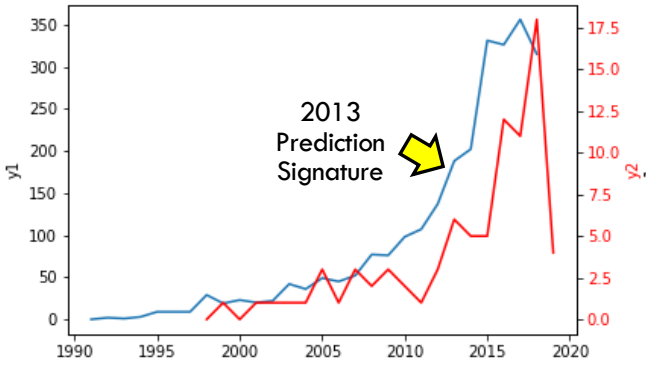
ACC/SIRT3 - NASH



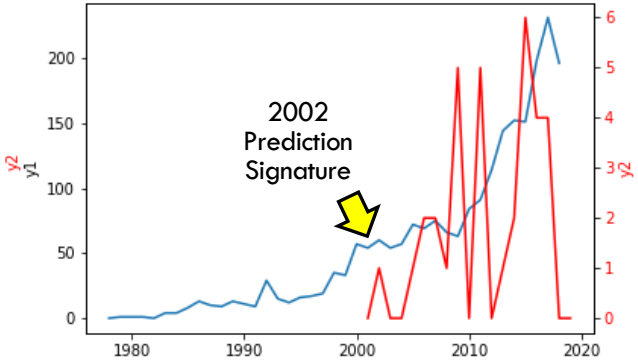
CFTR – Cystic Fibrosis



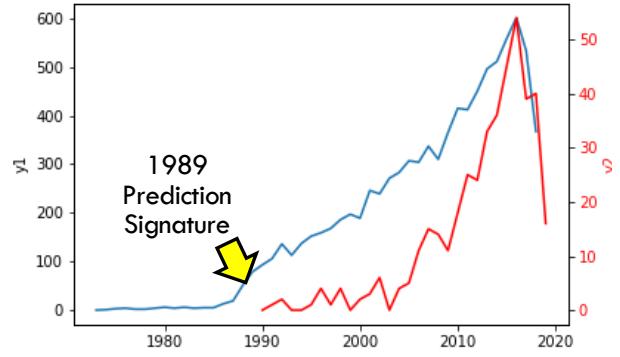
BCL2 - Cancer



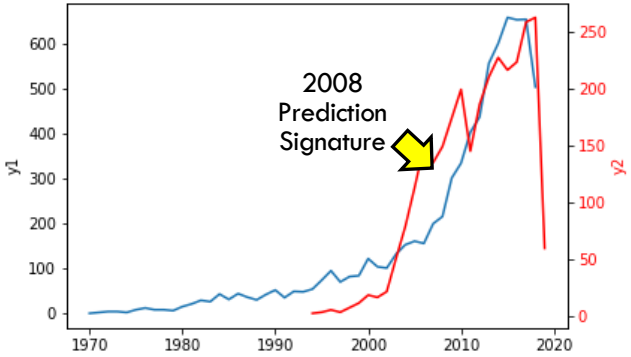
ACHE – Alzheimer's



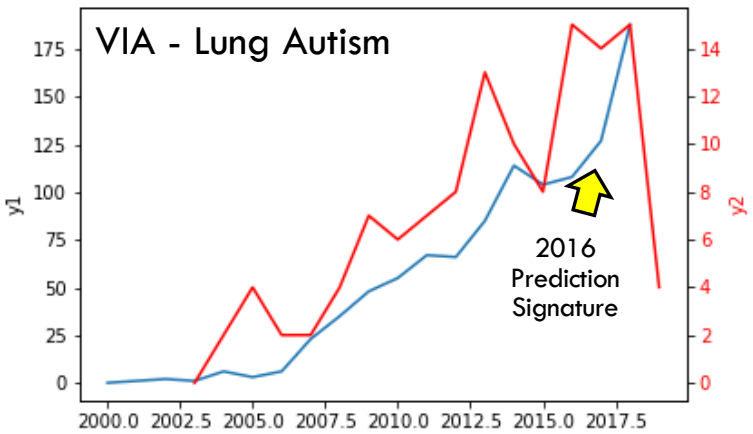
AR - Cancer

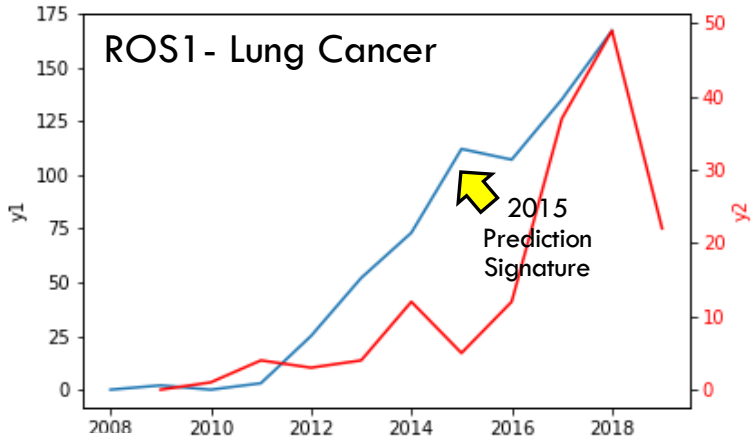


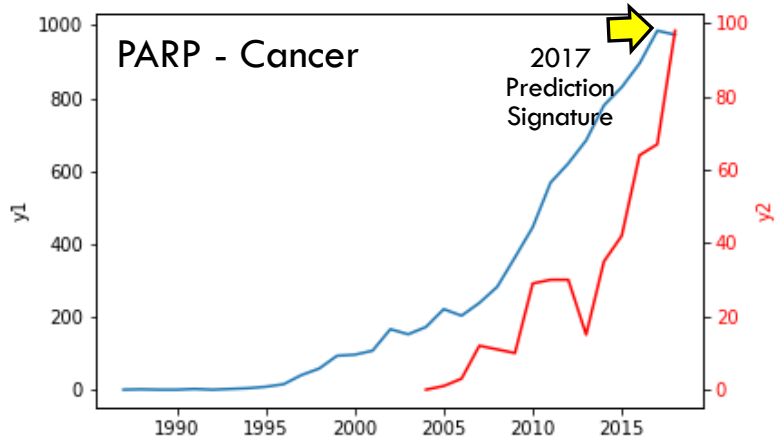
MET - Cancer

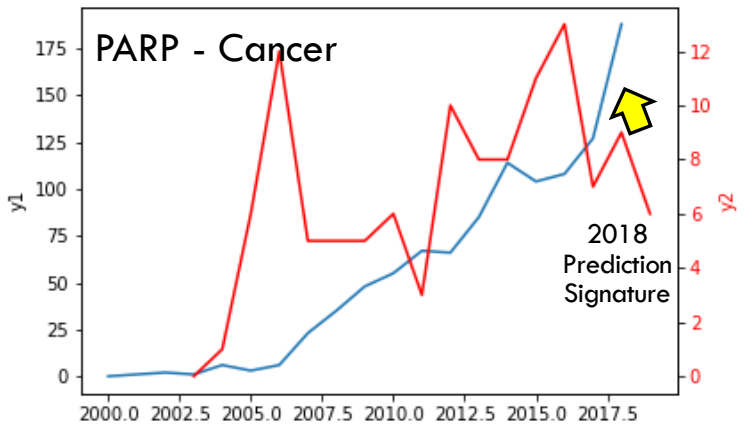


FUTURE INNOVATIONS (?)







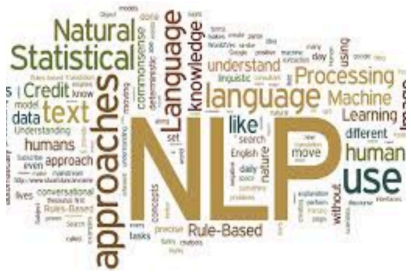


AUTOMATING FEATURE DETECTION

t-Distributed Stochastic Neighbor Embedding (T-SNE)
and Word Cloud

LEARNINGS

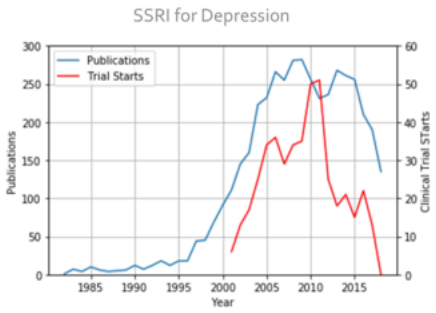
Data resources use different vocabularies. Ensuring comprehensive and consistent results from each source requires more effort with NLP approaches.

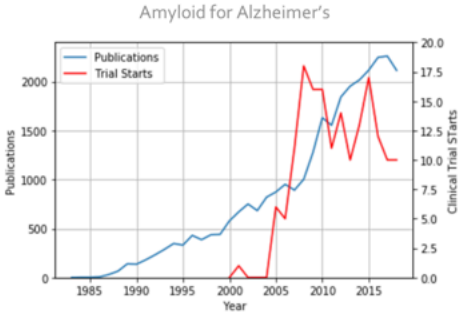


NLP approach for surfacing innovation keywords did not deliver the required specificity, requiring need for populated dictionaries of Target-Indication for this project. With BioLemmatizer, this may be corrected.

Model construction taught me there is more work to do in understanding regression and the underlying mathematics.

SUMMARY





Built multiple models to establish a predictive signature for biomedical innovation

Completed a first-pass assessment of one model

Refinements to the models are needed, as are refinements to automated feature detection and overall process automation

Although not yet in validated form, the first draft model predicts future innovation in the following areas:

- Autism from therapies that pursue the VIA target
- Lung Cancer from therapies that pursue the ROS1 target
- Lung Cancer from therapies that pursue the PARP target
- Psoriasis from therapies that pursue IL-23 and the Th-17 pathways