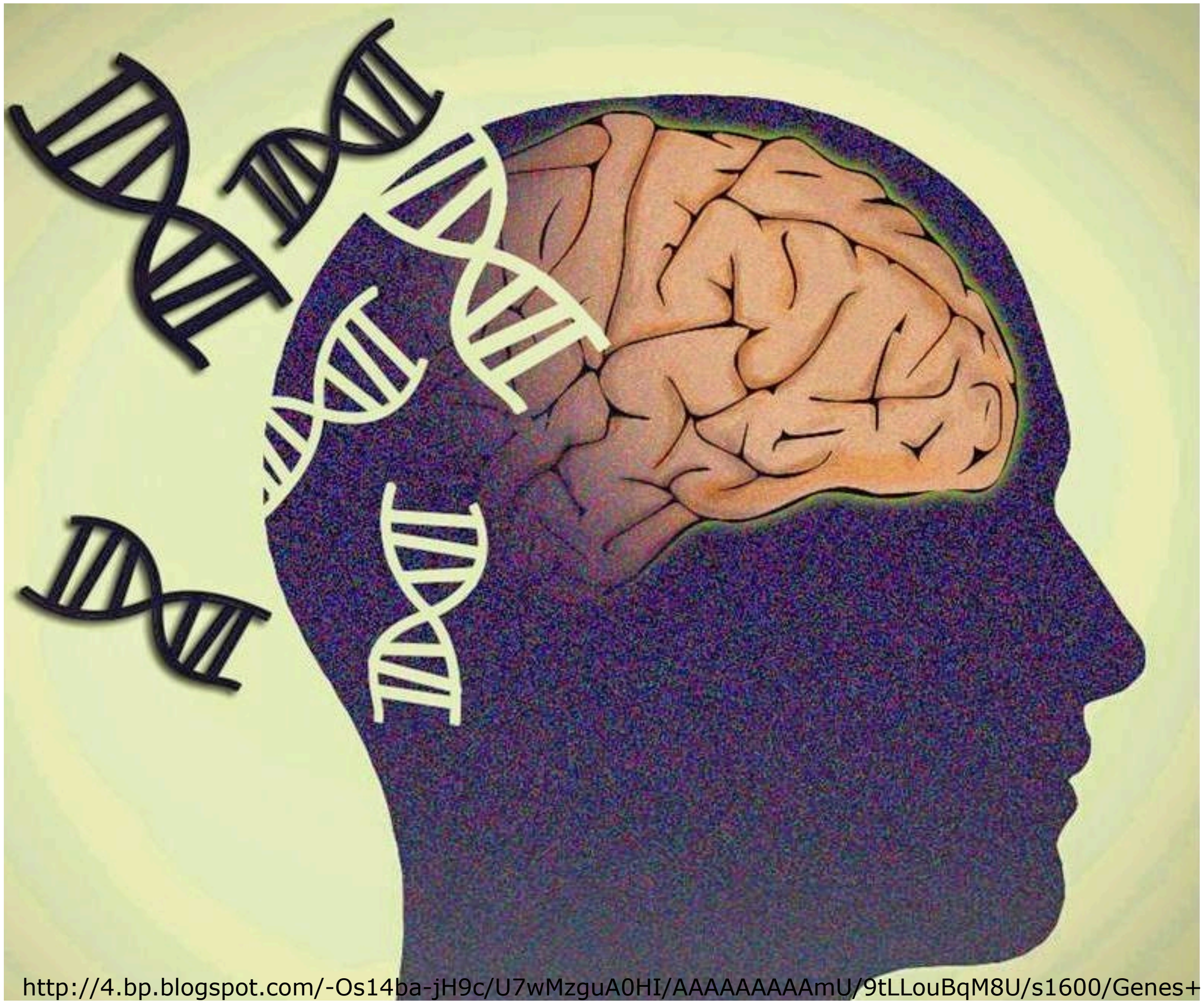
Machine Learning for early disease diagnostics

Himani Agrawal, PhD

Galvanize

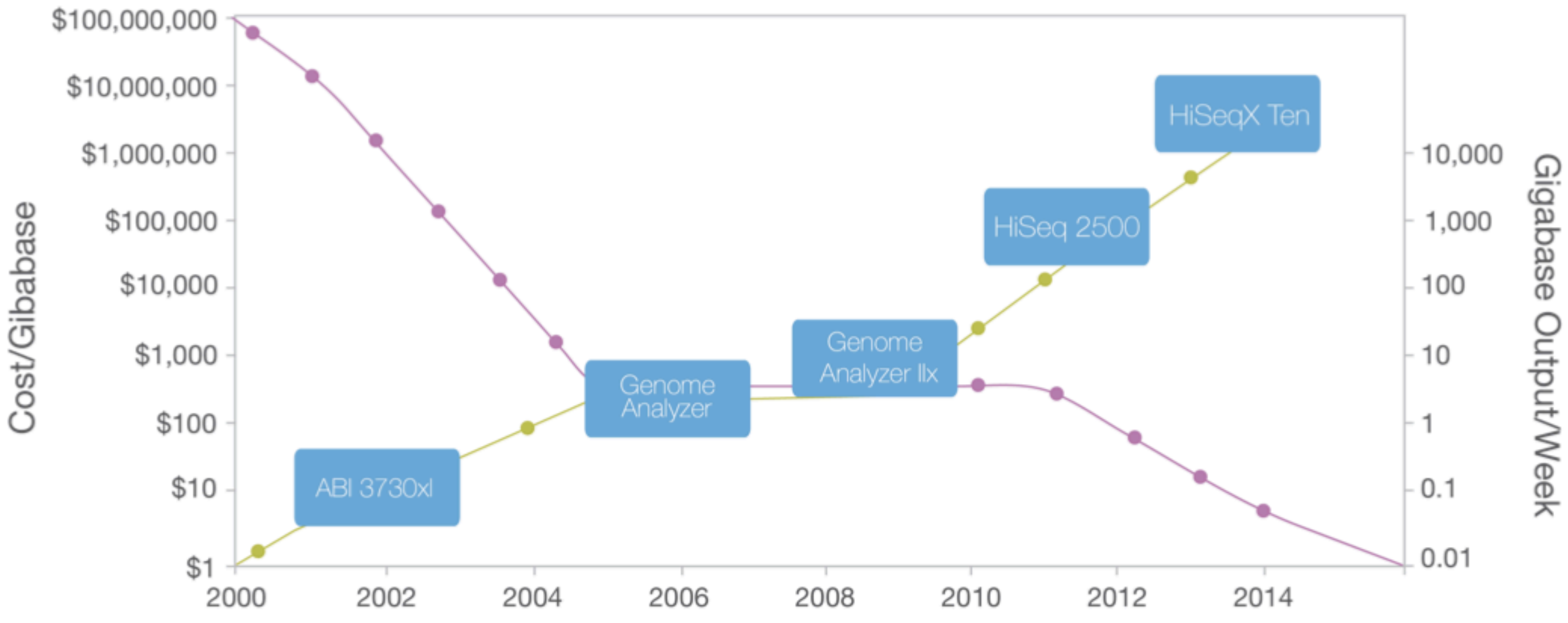
Simpatica Medicine



http://4.bp.blogspot.com/-Os14ba-jH9c/U7wMzguA0HI/AAAAAAAAAmU/9tLLouBqM8U/s1600/Genes+brain.jpg

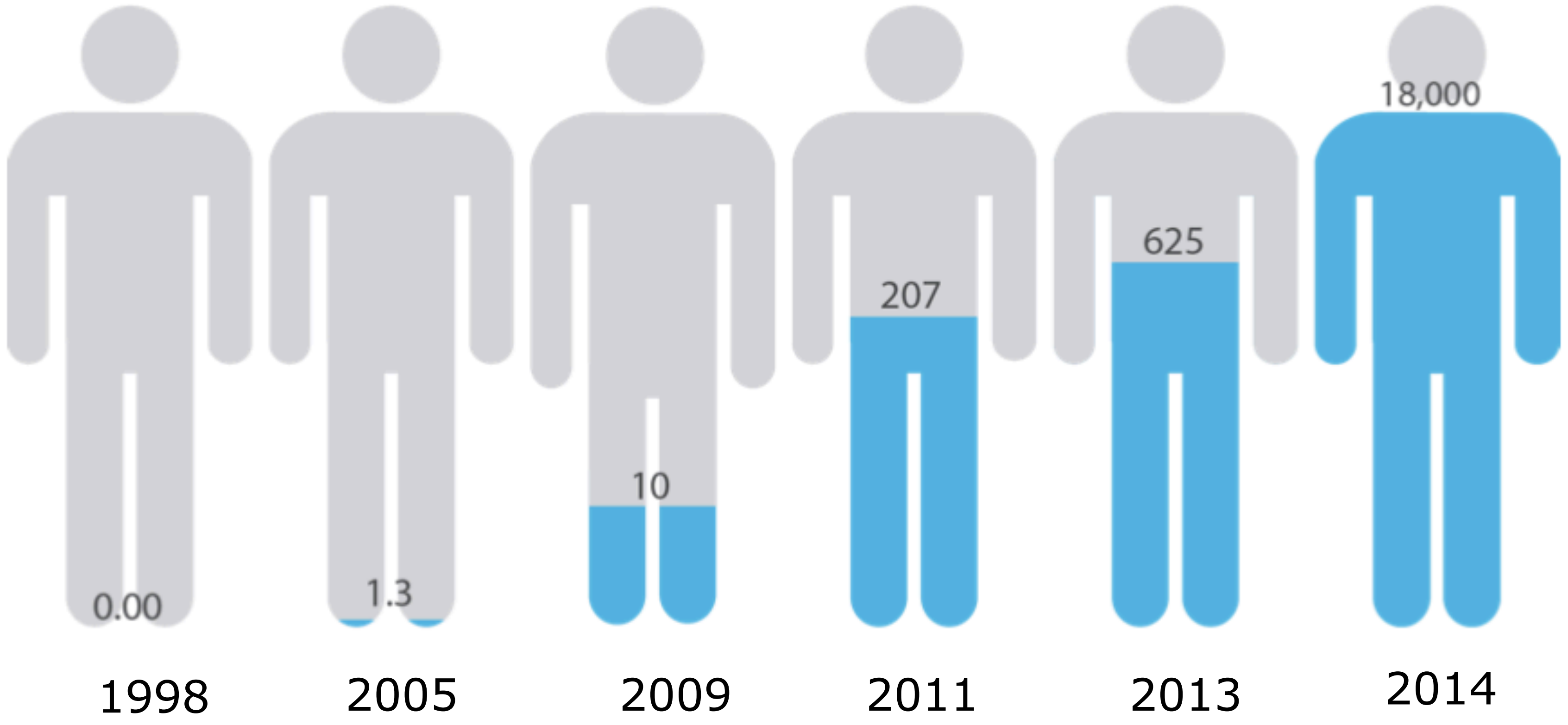
Parkinson’s disease:
- disease of the nervous system
- affects more than 10 million people
- only 5% people diagnosed before age 50
- reduced dopamine

The plummeting cost of genome sequencing, rise of genome data



https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

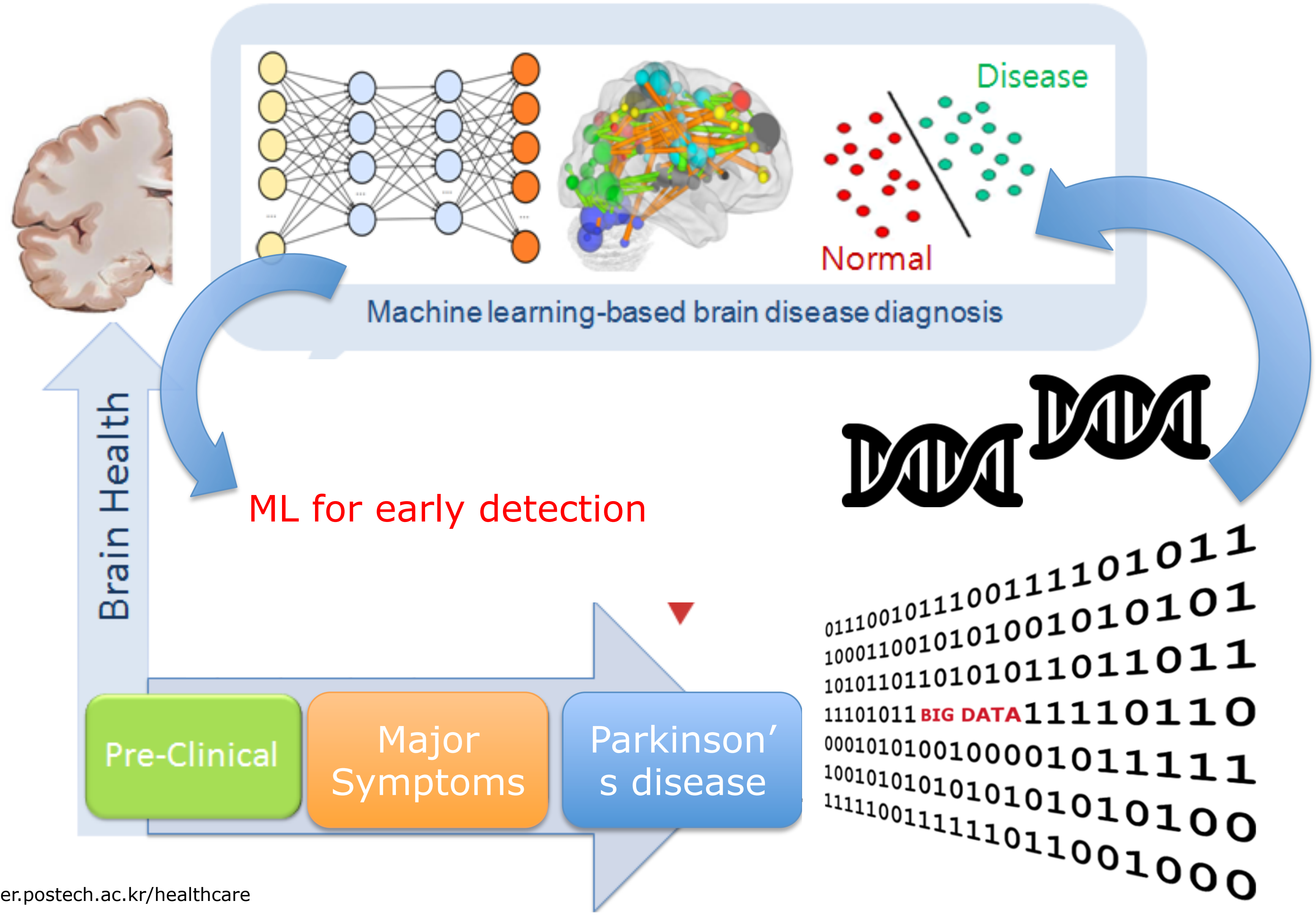
Human genomes sequenced annually



https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

RNA Sequencing

Machine learning for disease detection



ML for early detection

Brain Health

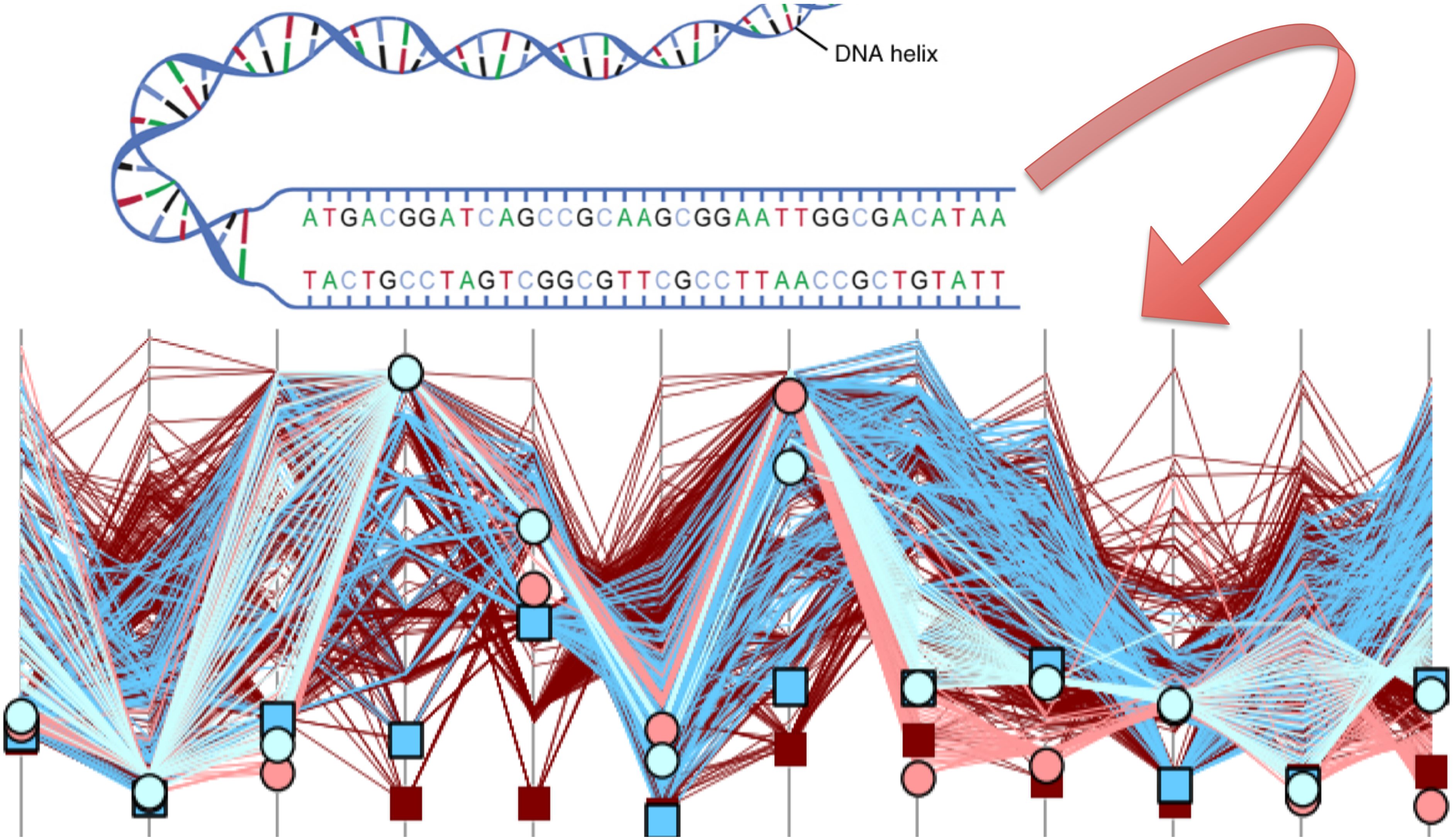
Pre-Clinical

Major Symptoms

Parkinson's disease

http://mlcenter.postech.ac.kr/healthcare

High dimensional genomics data

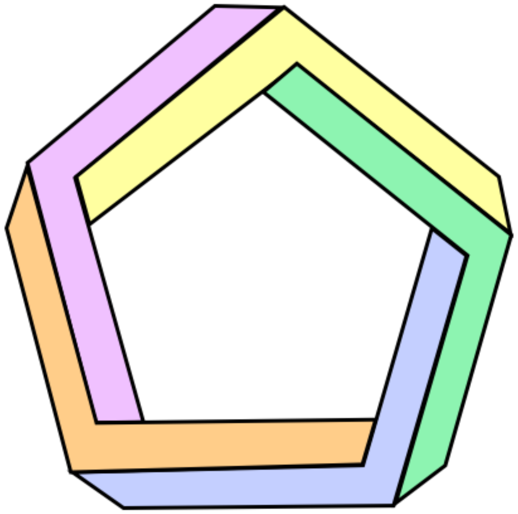

DNA helix
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAA
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATT

https://github.com/matthewjwoodruff/parallel.py

High dimensional genomics data

Needs Visualization

Needs veracity

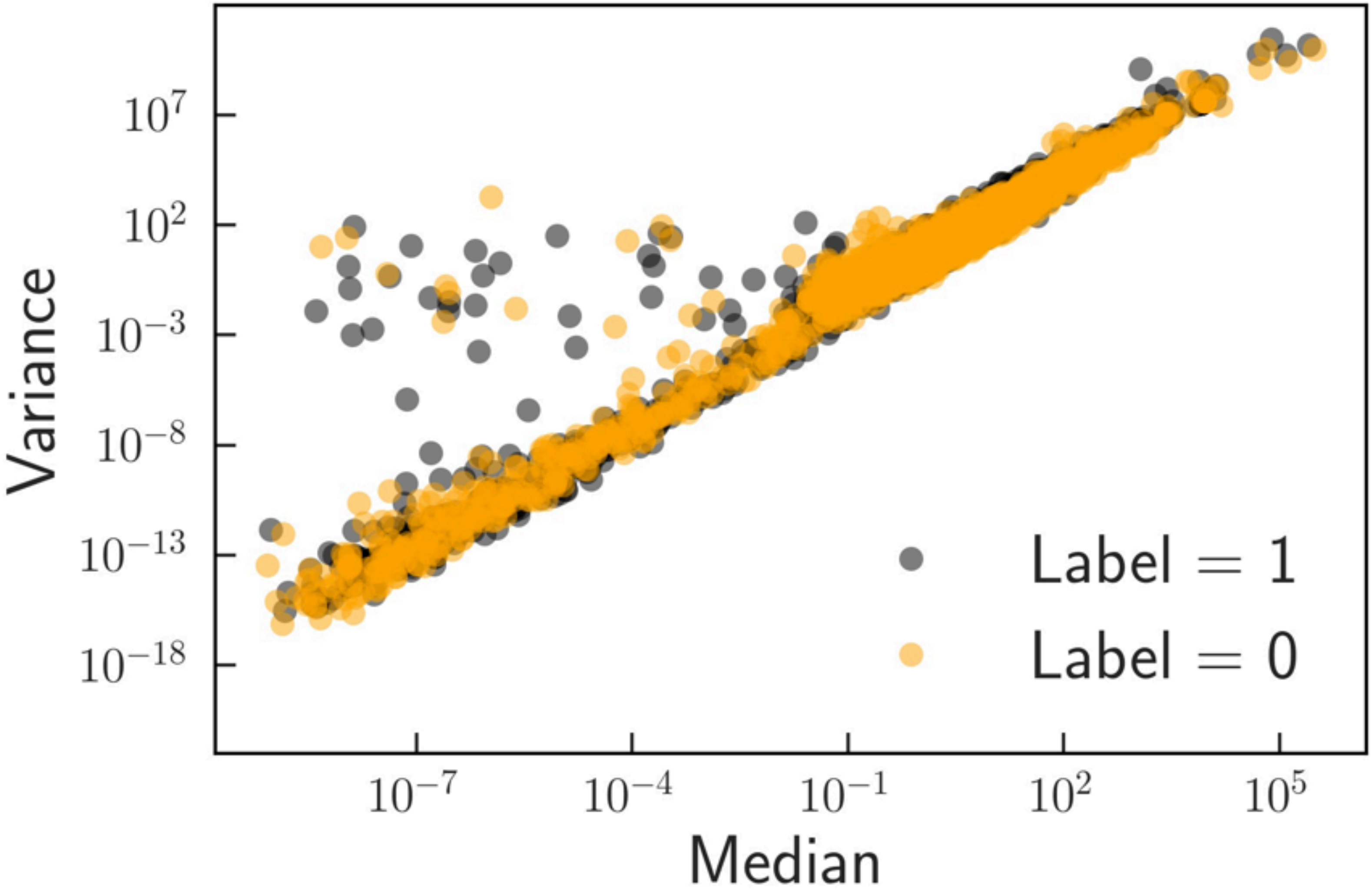


High variability

High volume

High variety

High variability: genomics data



High variety, high volume: medical data

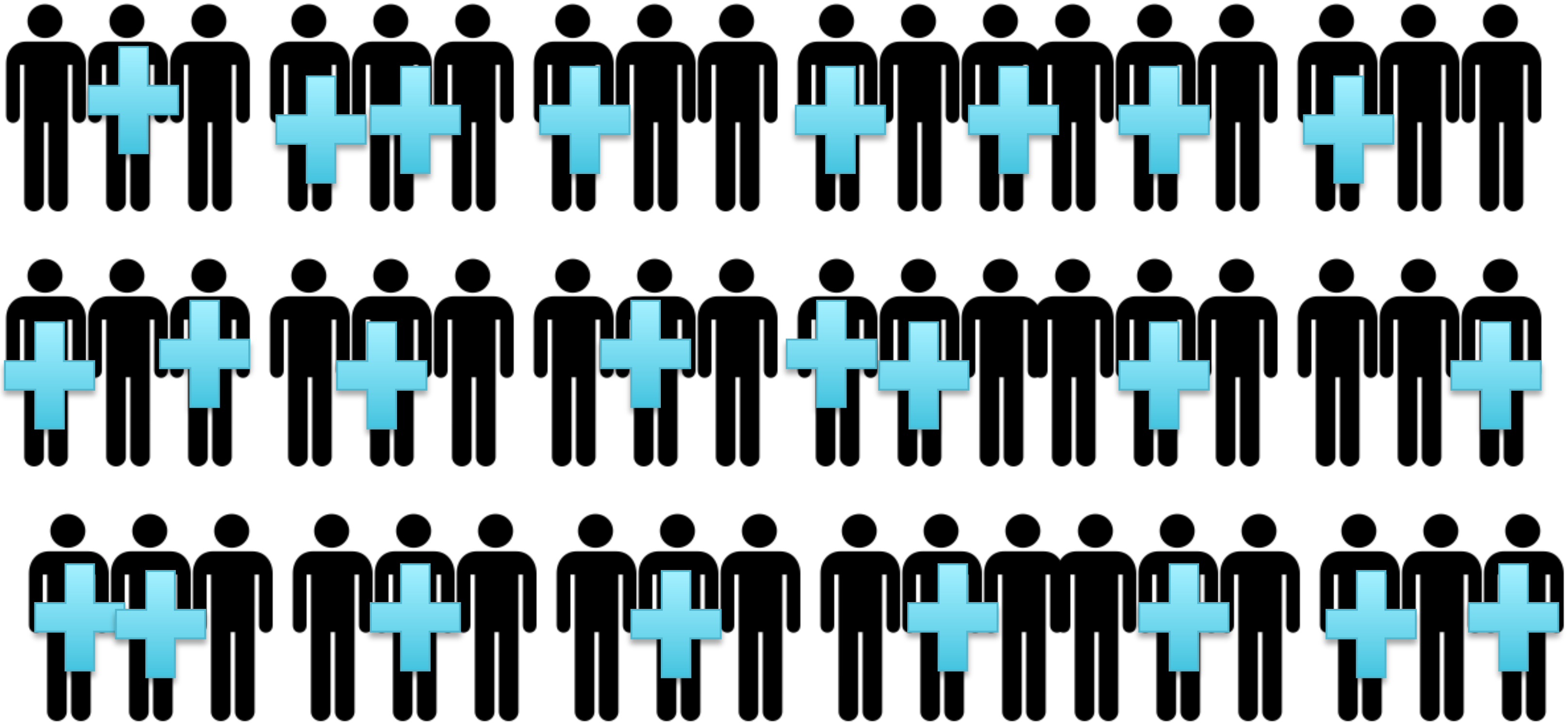
- Different formats: text, numeric, paper, digital, pictures, videos, multimedia
- Electronic Health Records
- Medical Images
- Fitness tracking

from mobile apps

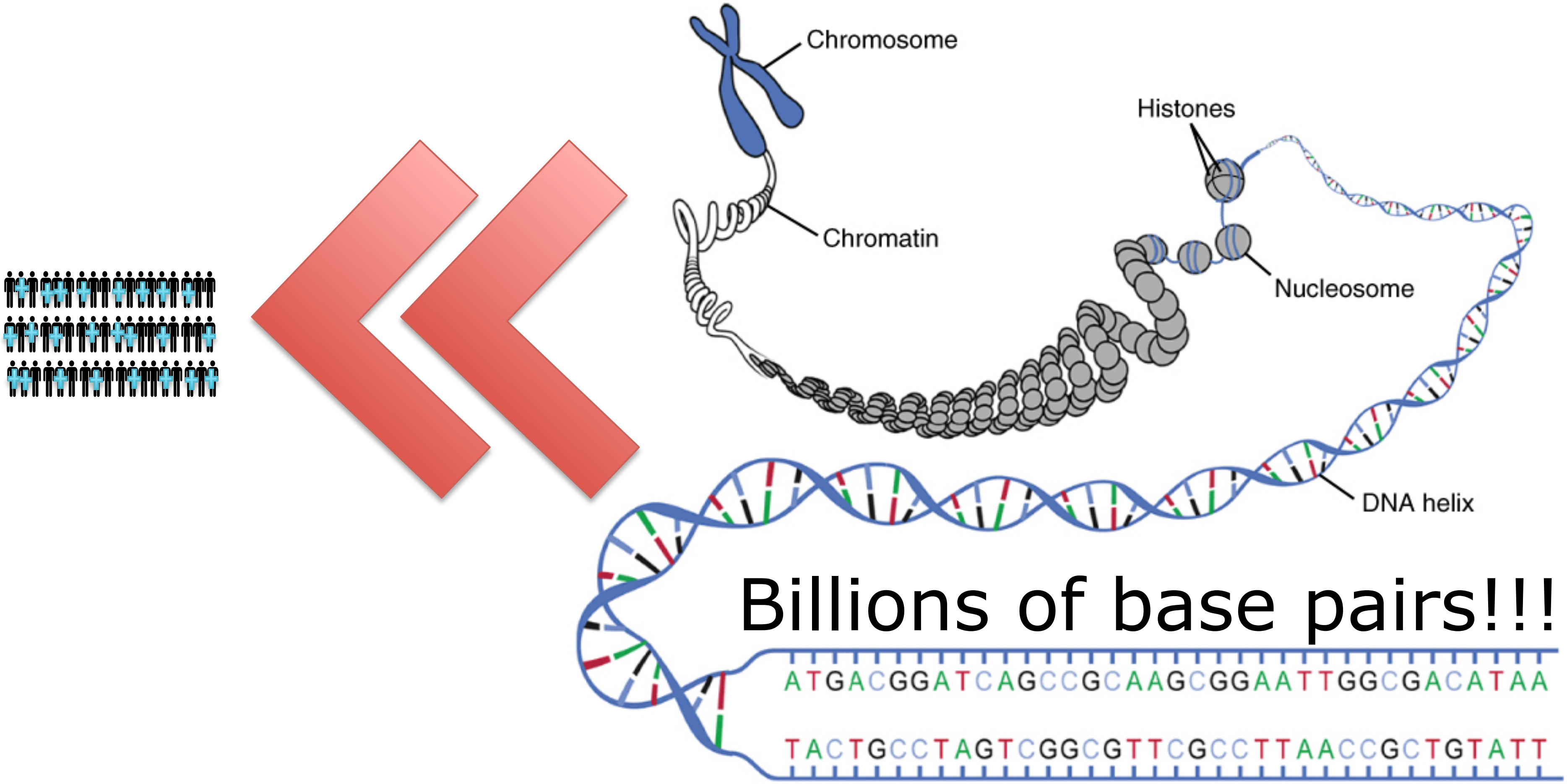
- ….etc.

Dataset characteristics

- 53 patients, 60000 features (gene expression level),
- ✚ implies presence of Parkinson



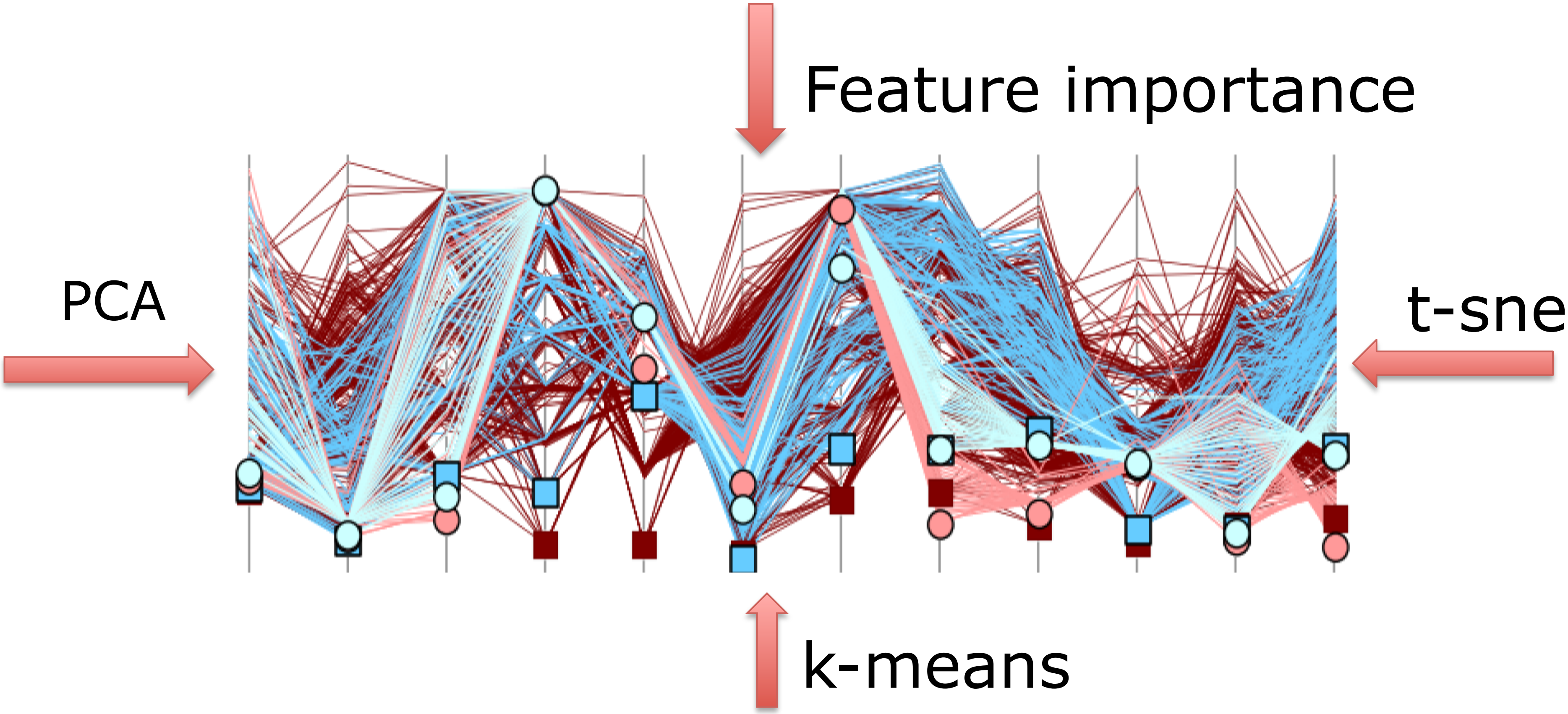
High dimensional analytics
53 << 60,000


Chromosome
Histones
Chromatin
Nucleosome
DNA helix
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAA
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATT

Billions of base pairs!!!

… https://i.stack.imgur.com/ibE8u.jpg

Dimensionality reduction: Smush the data !!!



Feature importance

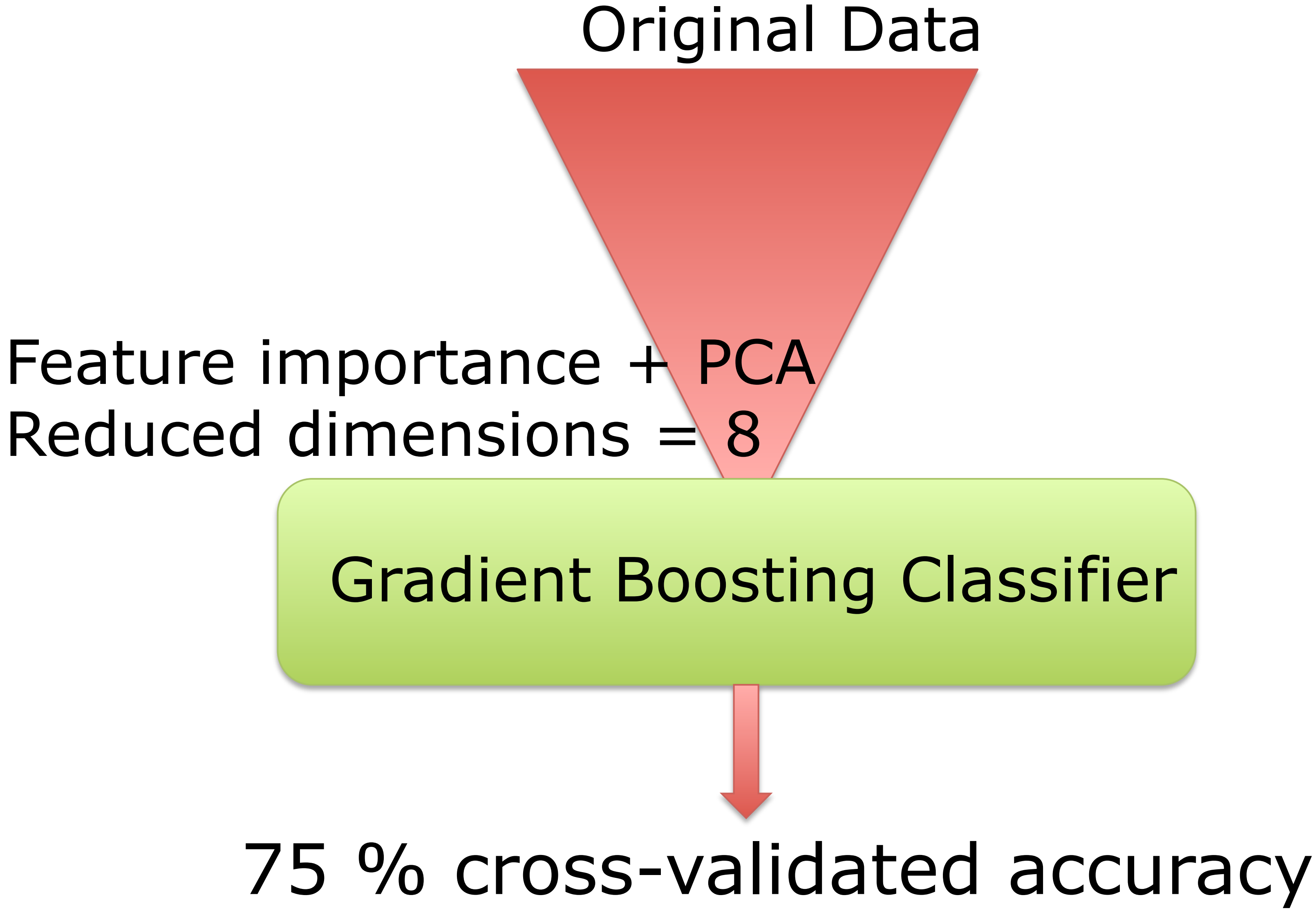
PCA

t-sne

k-means

- Principal Component Analysis (PCA)
- Feature importance using Random Forest
- K-means clustering

Machine learning model



Original Data

Feature importance + PCA
Reduced dimensions = 8

Gradient Boosting Classifier

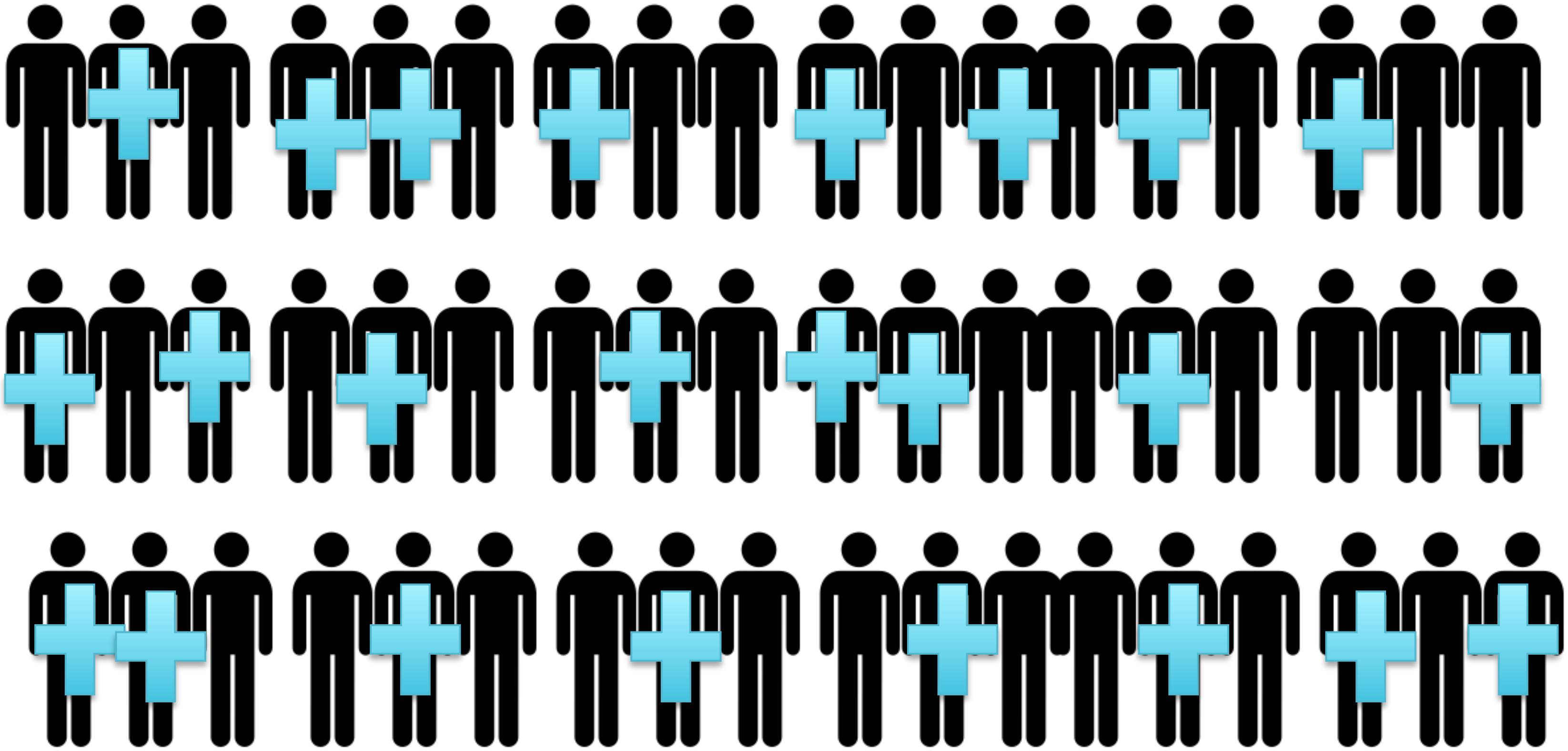
75 % cross-validated accuracy

Importance

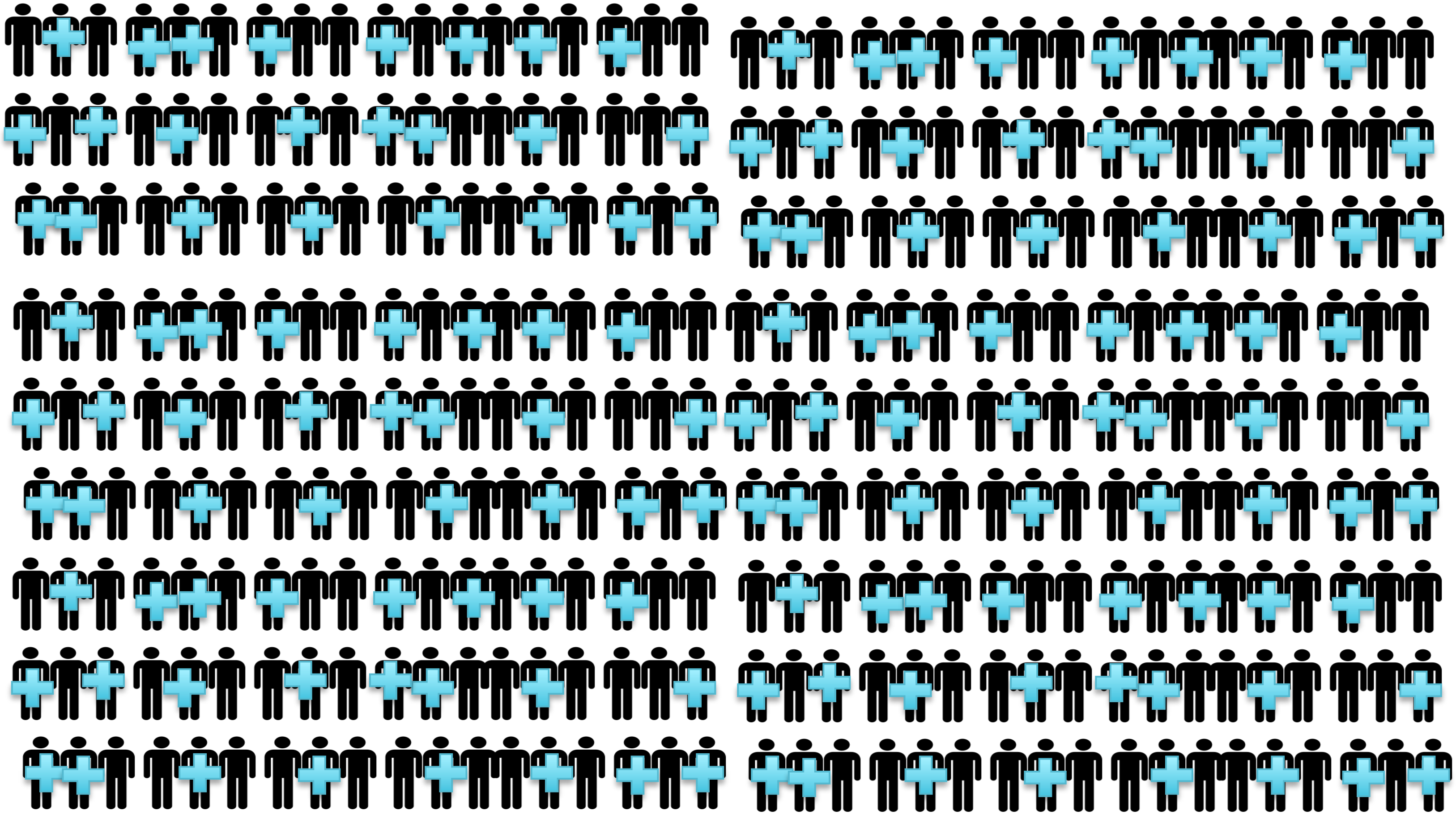
- There is no definitive diagnostic of Parkinson’s disease
- Doctor looks into medical history, neurological exam, major symptoms
- Response to Parkinson’s medications for further evidence
- Imaging system for dopamine levels
- But no accurate diagnosis, particularly in early stages --- ML can help!

Future work

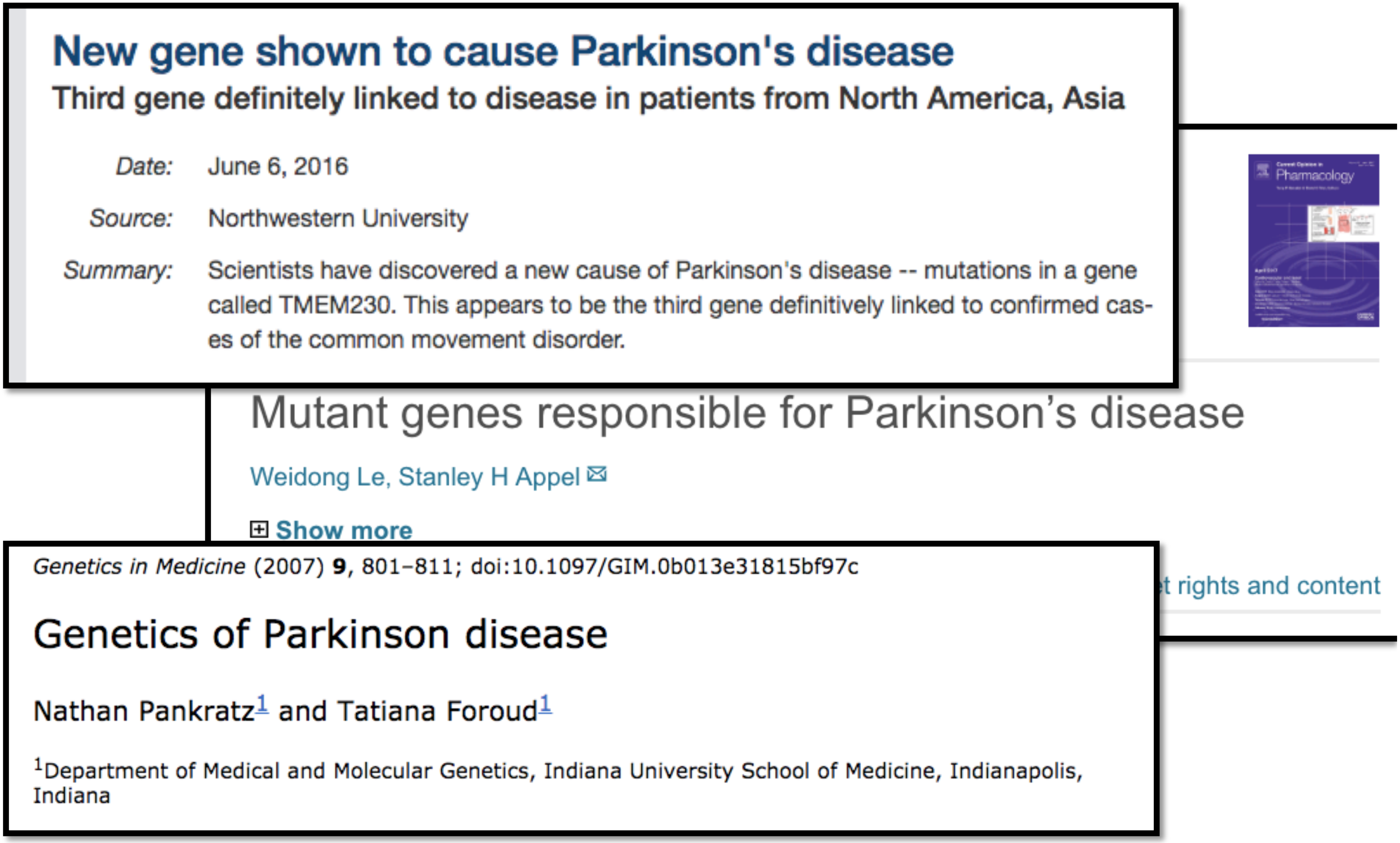
Small sample size



Higher sample size, more robust model



Incorporate published findings into ML model



New gene shown to cause Parkinson's disease

Third gene definitely linked to disease in patients from North America, Asia

Date:	June 6, 2016
Source:	Northwestern University
Summary:	Scientists have discovered a new cause of Parkinson's disease -- mutations in a gene called TMEM230. This appears to be the third gene definitively linked to confirmed cases of the common movement disorder.

Mutant genes responsible for Parkinson's disease

Weidong Le, Stanley H Appel ✉

⊞ **Show more**

t rights and content

Genetics in Medicine (2007) **9**, 801–811; doi:10.1097/GIM.0b013e31815bf97c

Genetics of Parkinson disease

Nathan Pankratz[1] and Tatiana Foroud[1]

[1] Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indiana

Thank you!!

?

Questions

VISIT ME AT

Web site

hagrawal.me.uh.edu

e-MAIL

1991himani@gmail.com

LinkedIN

https://www.linkedin.com/in/himaniagrawalgalvanize/