Machine learning for smart apps

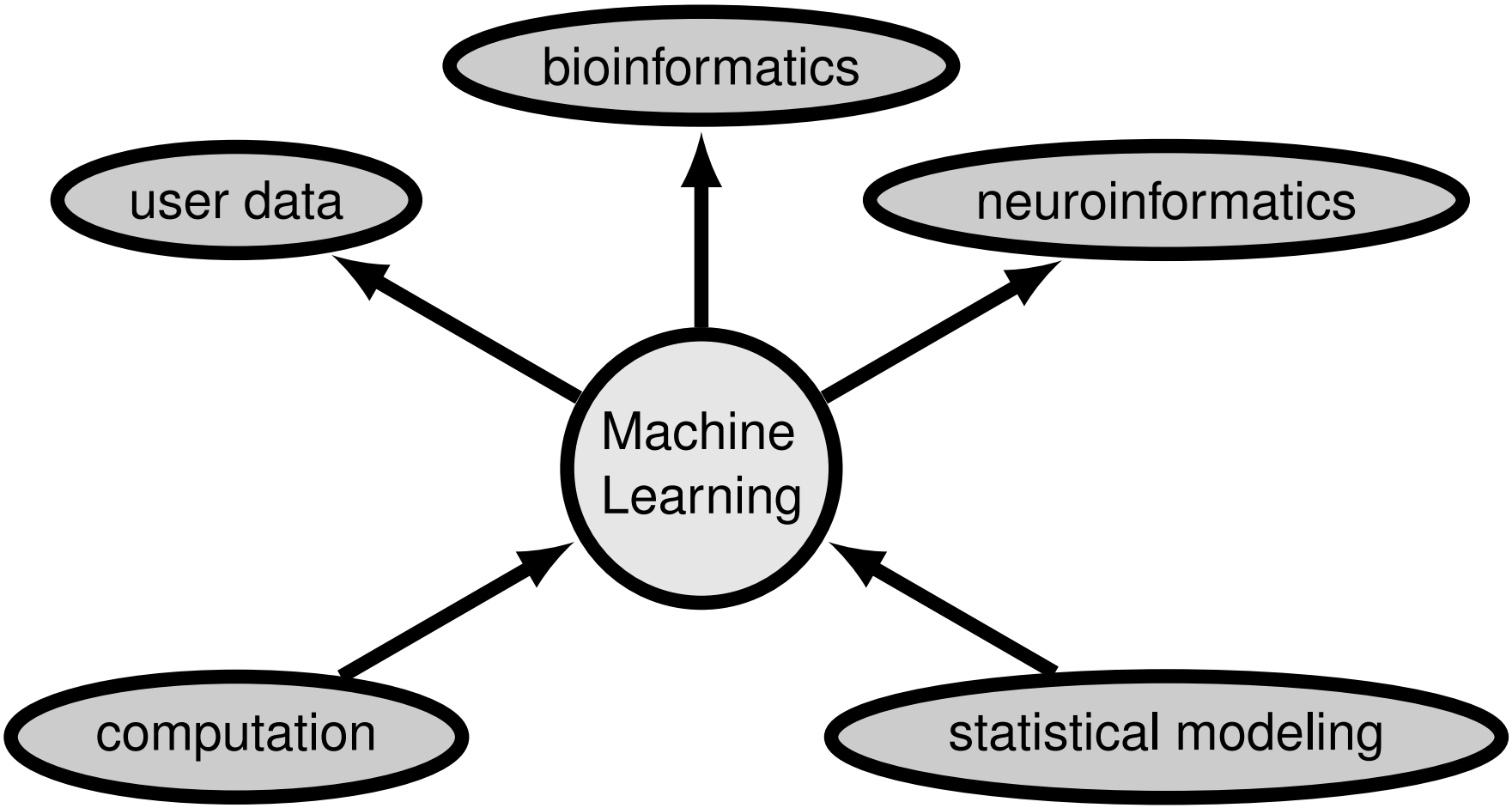
Ole Winther

Department for Applied Mathematics and Computer Science
Technical University of Denmark (DTU)

May 19, 2014

When I talk about mathematics. . .

Statistical machine learning



Infinite is larger than big

Bill Gates Wired interview

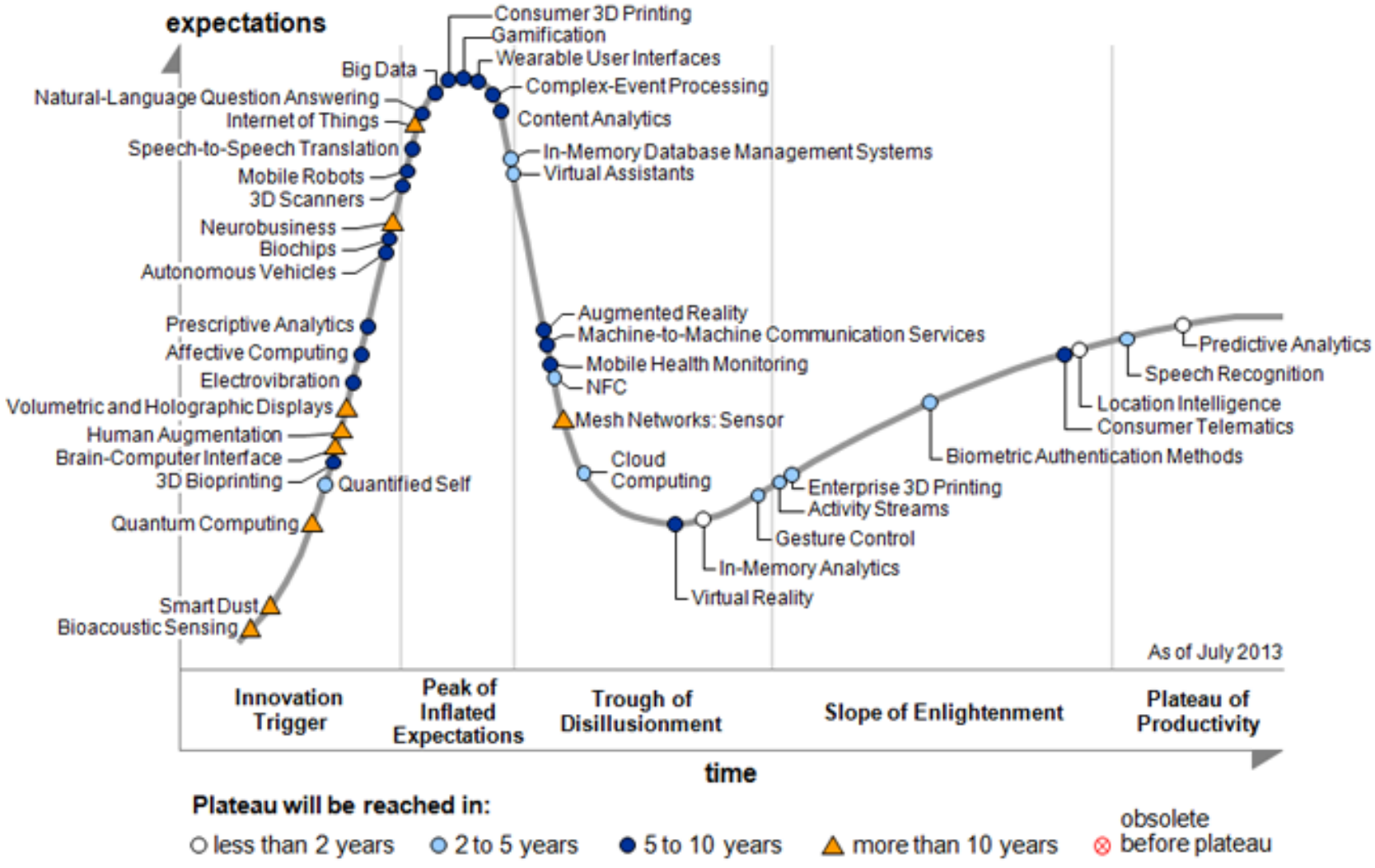
Wired: What will we be writing about in *Wired* 20 years from now?

Gates: You'll still be talking about the fear of robots. That's a good one to chew on for a long time.

Wired: Which robots?

Gates: The article-writing robots. Seriously, what's unique about human intelligence will be a topic of interest for way more than 20 years. But the biggest thing in that time period will be the completion of pervasive computing: vision, speech, handwriting, goggles, every surface, infinite machine learning, infinite storage, infinite reliability, at essentially no cost.

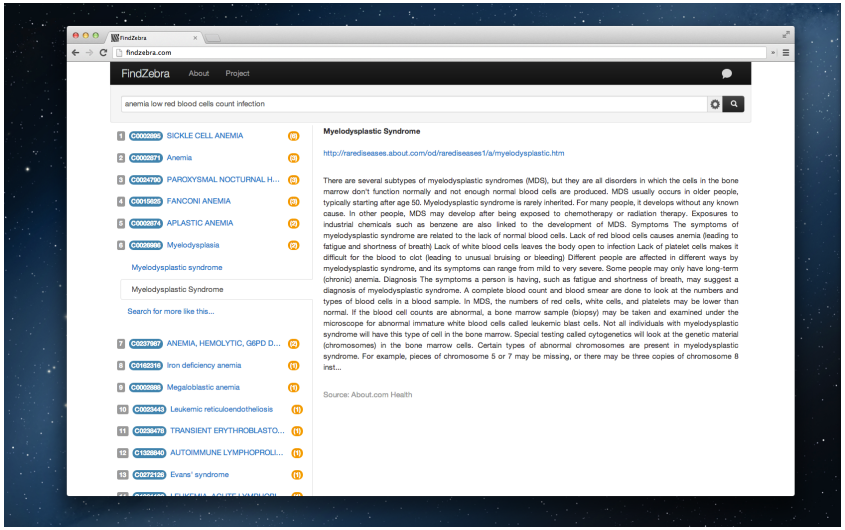
The hype curve



http://www.gartner.com/newsroom/id/2575515

Two machine learning cases

- Collaborative filtering — the Netflix Prize and one-class CF
- Specialised search — `findzebra.com`



Collaborative filtering

- Collaborative filtering from Wikipedia:
- . . . Applications of collaborative filtering typically involve very large data sets. Collaborative filtering (CF) methods have been applied to many different kinds of data . . . in electronic commerce and web 2.0 applications where the focus is on user data, etc.
- The method of making automatic predictions (filtering) about the interests of a user by collecting taste information from many users (collaborating). The underlying assumption of CF approach is that those who agreed in the past tend to agree again in the future. . . .
- Some companies using collaborative filtering: Amazon, . . . , eBay, . . . , Netflix, . . .

Netflix prize

- Improve Netflix Cinematch system by 10% to win prize.
- Data details
  - $M = 17.770$ movies
  - $N = 480.189$ users
  - training.txt – 10^8 quadrules

$$(\text{user}, \text{movie}, \text{rating}, \text{time-stamp})$$

  - rating: ★ to ★★★★★
  - qualifying.txt – 2.817.131

$$(\text{user}, \text{movie}, ?, \text{time-stamp})$$

- Competition - at most once a day:
  - submit (continuous) predictions and
  - Netflix returns a RMSE.
- Data sparse:

$$\frac{10^8}{MN} = 0.015 \ .$$

18,000 movies

480,000 users

x	1	1	x	...	x
x	x	x	5	...	x
x	x	3	x	...	x
x	4	3	x	...	2
...	x	x	x	...	x
x	5	x	1	...	x
x	x	3	3	...	x
x	1	x	x	...	2

- $\mathbf{v}_i$: "taste" vector of user i, $\text{length}(\mathbf{v}_i) = K$.
- $\mathbf{u}_j$: "profile" vector movie j.
- Rating model:

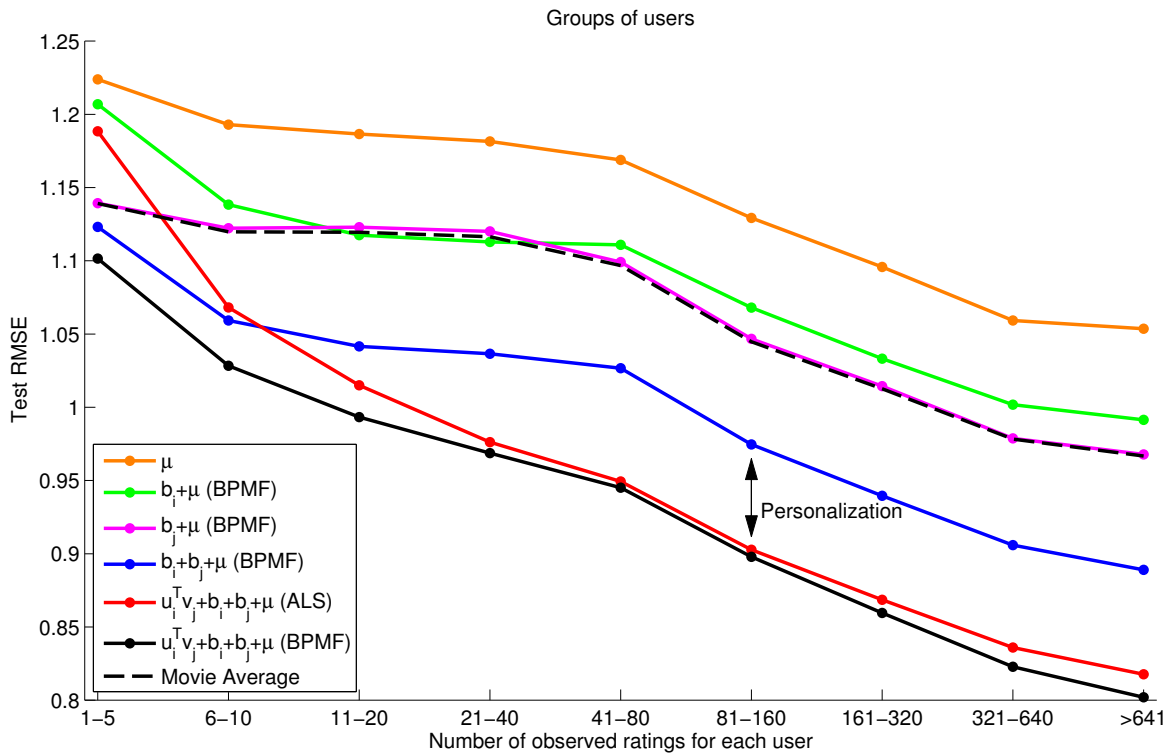
$$r_{ij} = \mathbf{u}_i \cdot \mathbf{v}_j + \epsilon_{ij}$$

- Learn **U** and **V** from rating matrix. Computation!

- Delineate personalisation from biases:

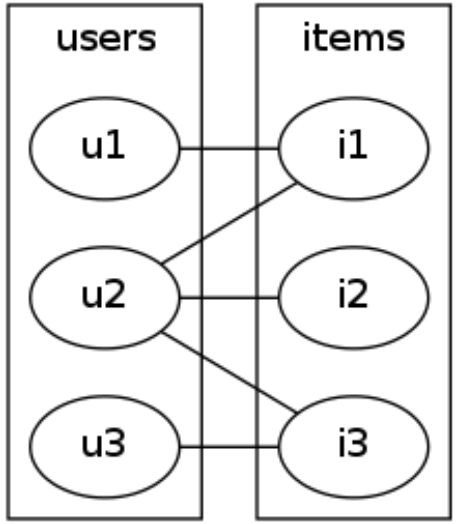
$$r_{ij} = \mathbf{u}_i \cdot \mathbf{v}_j + b_i + b_j + \mu + \epsilon_{ij}$$

- Likelihood calculation $\propto$ training data - 10^8 ratings.
- Inference over $K(M+N) \sim 10^8$ parameters:
  - Least square with regularisation (ALS)
  - Bayesian - Gibbs sampling inference (BMF)



- Bayesian averaging works!

One-class collaborative filtering



- Modeling likes, buys or views
- Corresponds to links in bipartite graph

- Model1: Simple: popularity model works quite well:

$$p(\text{link}(i,j)|\pi_i,\psi_j) = \pi_i\,\psi_j$$

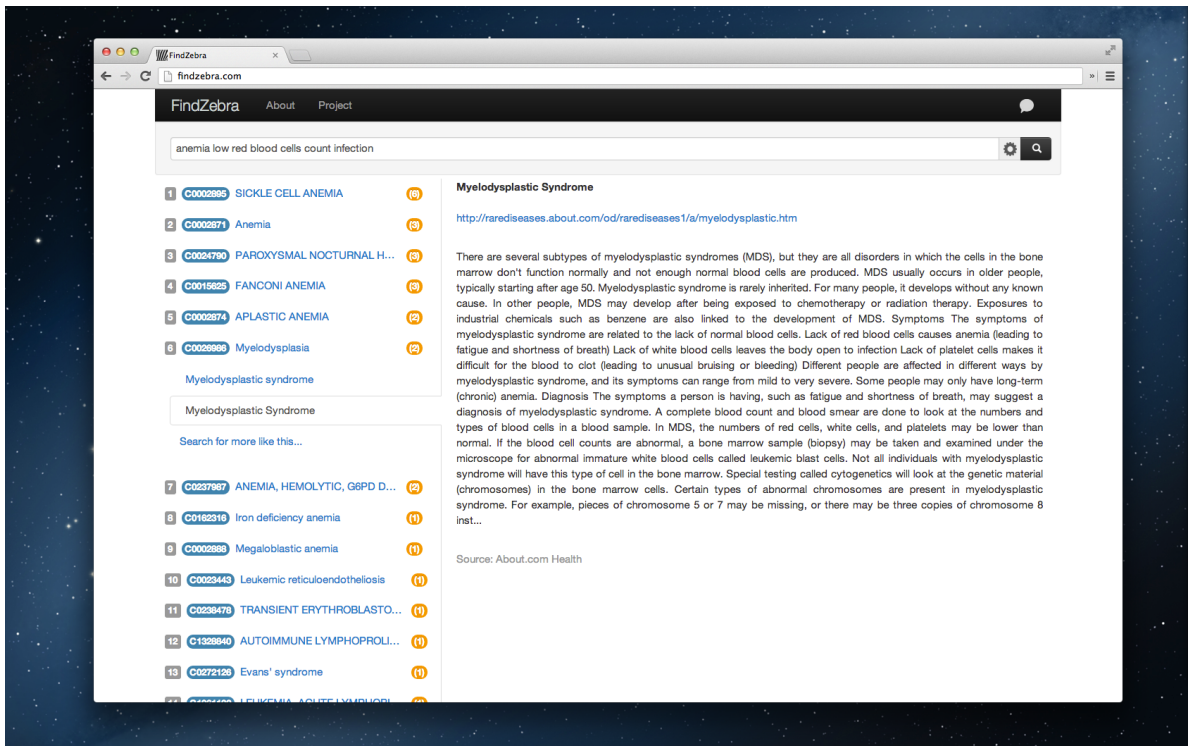
- π_i probability of user i likes something
- ψ_j probability that item j is liked.
- Model 2: Personalised preference function: $\sigma(\mathbf{u}_i^T\mathbf{v}_j) \in [0,1]$

$$p(\text{link}(i,j)|\pi_i,\psi_j,\mathbf{u}_i,\mathbf{v}_j) = \pi_i\,\psi_j\,\sigma(\mathbf{u}_i^T\mathbf{v}_j)$$

- $\sigma(\ldots)$ is logistic function.

FindZebra -
The search engine for difficult medical cases

- Links
  - `www.ijmijournal.com/article/S1386-5056(13)00016-6/abstract`
  - `arxiv.org/abs/1303.3229`,
  - `findzebra.com`



Ellen's case story

For 25 years, Ellen struggled to find a diagnosis for the multitude of debilitating symptoms that seemed to increase year after year.

- Her symptoms included muscle cramps, intense headaches, rapid weight gain, fatigue, edema, intolerance to heat, excessive sweating, joint pain, tingling in her hands and feet, frequent bone fractures, acid reflux, intense anxiety and panic attacks, high blood pressure, high cholesterol, high blood sugar, sleep apnea, menstrual irregularities, peripheral vision loss and double vision.
- Source: `http://www.uptodate.com/home/ellen-uses-uptodate-find-diagnosis`
- Any suggestions? - Get back to case in demo.

Rare diseases - enter FindZebra.com

"When you hear hoofbeats behind you, don't expect to see a zebra"

- Rare diseases hard to diagnose.
- Physicians use Google and PubMed. A good idea?
- We set up evaluation and FindZebra.com (public IR + data)
- Google 18/56 and FindZebra 38/56 cases in top 20
- Conclusion: Specialized search engine works better!

Moonshots and big data

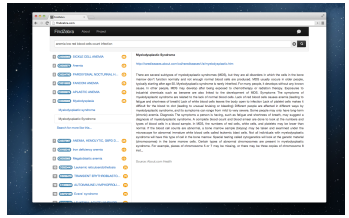
- Can information technology help change the culture of medical diagnosis?

- Larry Page, co-founder and CEO Google

$$10\% \rightarrow 10x$$

- Wired interview February 2013

- FindZebra: Small data of high quality
- 33.000 documents from specialized sources on rare diseases
- Simple document ranking algorithm - use only document-query match



Data sources

Resource	Entries
Online Mendelian Inheritance in Man (OMIM) http://www.ncbi.nlm.nih.gov/omim	20,369
Genetic and Rare Diseases Information Center (GARD) http://rarediseases.info.nih.gov/GARD	4578
Orphanet, http://www.orpha.net	2967
Wikipedia, http://www.wikipedia.org/	2239
National Organization for Rare Disorders (NORD) http://rarediseases.org	1230
Genetics Home Reference http://ghr.nlm.nih.gov	626
GeneReviews http://www.ncbi.nlm.nih.gov/books/NBK1116/	599
Madisons Foundation Rare Paediatric Disease Database http://www.madisonsfoundation.org	522
Health on the Net Foundation Rare Disease Database http://www.hon.ch	183
Swedish National Board of Health and Welfare www.socialstyrelsen.se/rarediseases	114

Ranking algorithms - how to score each document

- Google's secret, got 200 parameters including PageRank.
- We use a much simpler scoring function:
- Independence of terms:

$$\begin{aligned}\text{Score}(\text{`hypertension, adrenal mass'}) &= \text{Score}(\text{`hypertension'}) + \\ &\quad \text{Score}(\text{`adrenal'}) + \\ &\quad \text{Score}(\text{`mass'})\end{aligned}$$

- Interpolation between document and corpus frequency

$$\text{Score}_{\text{doc}}(\text{term}) = \log\left[\frac{f_{\text{doc}}(\text{term}) + \frac{\mu}{l_{\text{doc}}} f_{\text{corp}}(\text{term})}{1 + \frac{\mu}{l_{\text{doc}}}}\right]$$

Test queries - examples

- Normally developed boy age 5, progressive development of talking difficulties, seizures, ataxia, adrenal insufficiency and degeneration of visual and auditory functions: ?
- 14 year old, teenage boy, mild mental retardation, proximal muscle weakness, unable to walk (wheelchair-bound), premature ventricular complexes, ophthalmoparesis: ?
- fever, anterior mediastinal mass and central necrosis: ?

Test queries - examples

- Normally developed boy age 5, progressive development of talking difficulties, seizures, ataxia, adrenal insufficiency and degeneration of visual and auditory functions: Adrenoleukodystrophy autosomal neonatal form
- Ranks: FindZebra=2 and Google search = -
- 14 year old, teenage boy, mild mental retardation, proximal muscle weakness, unable to walk (wheelchair-bound), premature ventricular complexes, ophthalmoparesis: Autosomal recessive centronuclear myopathy (ARCNM)
- Ranks: FindZebra=2 and Google search = -
- fever, anterior mediastinal mass and central necrosis: Lymphoma
- Ranks: FindZebra=7 and Google search = 1

Predictive methods

- are entering in new domains all the time.
- Many niches unexplored.
- Collaborative filtering: ★ to ★★★★★ and one-class
- Medical diagnosis: Physicians make diagnostic errors
- Graber et. al. divides them into:
  - Context errors,
  - availability errors,
  - premature closure.
- A change of culture and better tools can reduce errors.
- Remember Infinite machine learning is coming. ;-)

Thank you!

Acknowledgements

- FindZebra developer team:
  - Dan Svenstrup
  - Philip Henningsen
  - Robert Kristjansson
- Team physician
  - Henrik L Jorgensen
- Former contributors:
  - Radu Dragusin
  - Paula Petcu
  - Christina Lioma
  - Birger Larsen
  - Ingemar J. Cox
  - Lars Kai Hansen
  - Peter Ingwersen
- Recommender systems:
  - Ulrich Paquet (Microsoft Research)
  - Noam Koenigstein (Microsoft Israel)
  - Blaise Thomson (Cambridge U)

www.ijmijournal.com/article/S1386-5056(13)00016-6/abstract, arxiv.org/abs/1303.3229, findzebra.com


MIT Technology Review
THE TIMES
ORF 1 RADIO ÖSTERREICH 1
NewScientist
theguardian
Smithsonian.com
The Telegraph
الجزيرة.نت ALJAZEERA.NET
NETWORKWORLD
Khaleej Times
search engine land
SCOPE
Kronen Zeitung
MNT Medical News Today
The New Zealand Herald

MIT Technology Review

A powerful new search engine designed to help diagnose rare diseases could prove a boon for both medics and the public.

— MIT Technology Review: The Rare Disease Search Engine That Outperforms Google

THE TIMES

Danish researchers have unveiled a free site called "FindZebra" in an effort to give doctors a new weapon to deal with tricky medical cases.

— The Times: Online 'doctor' can diagnose rare illnesses

The Telegraph

FindZebra is designed to look through a number of selected databases of rare diseases, meaning searches that could take hours if done manually take seconds.

— The Telegraph: New search engine means doctors don't need Google to find rare diseases

theguardian

Most of us have had occasion to consult Dr Google. Rather than waste a GP's time with your embarrassing worries, just type your symptoms into a search engine, hit return and terrify yourself with the results. [...] I'm no doctor, and for that reason I recommend FindZebra unreservedly.

— The Guardian: FindZebra diagnoses rare diseases – how will it interpret my 'symptoms'?

NewScientist

Frustrated patients and doctors can also turn to FindZebra, a recently launched search engine for rare diseases.

— New Scientist: Crowd diagnosis could spot rare diseases doctors miss

Smithsonian.com

[S]earch engines such as Google are not designed to help a physician weed out possibilities behind an obscure set of symptoms [...] To fill this gap, researchers from Denmark built a new search engine dedicated to rare diseases called FindZebra.

— Smithsonian.com: This New Search Engine Helps Doctors Diagnose Rare, Obscure Diseases

search engine land

[...] FindZebra.com may be the answer to helping physicians correctly diagnose diseases that are often misdiagnosed, can take years to identify.

— Search Engine Land: Search Engine Designed By Denmark Researchers Helps Medics Diagnose Rare Diseases

NETWORKWORLD

[D]octors are well advised to initially see common symptoms as evidence of common maladies, which is all well and good until the patient happens to be suffering from a rare disease. It is for the latter circumstance that researchers at the Technical University of Denmark are developing a specialized search engine called FindZebra.

— NetworkWorld: Medical search engine said to find 'zebras' better than Google

SCOPE

[A]s many patients with rare diseases know, using conventional Internet search engines to diagnose a condition that occurs in less than 1 in 2000 of the population can prove tricky. So a group of European researchers developed an alternative [...]

— Scope: New search engine designed to help physicians and the public in diagnosing rare diseases