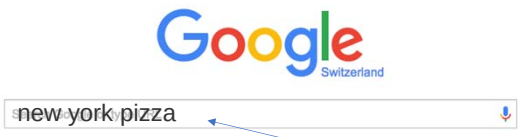
Big Data and Internet Search

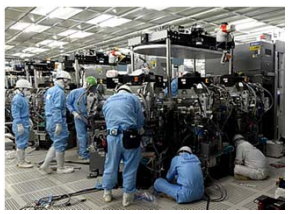
James Larus

Professor and Dean IC

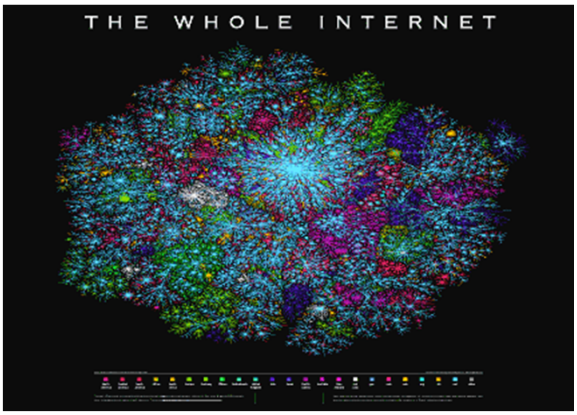
December 13, 2016

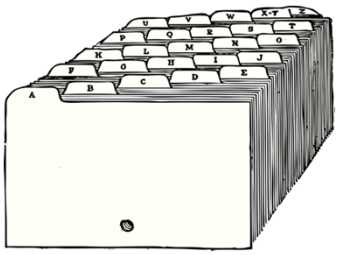
How Does a Search Engine Work?





Easy Part



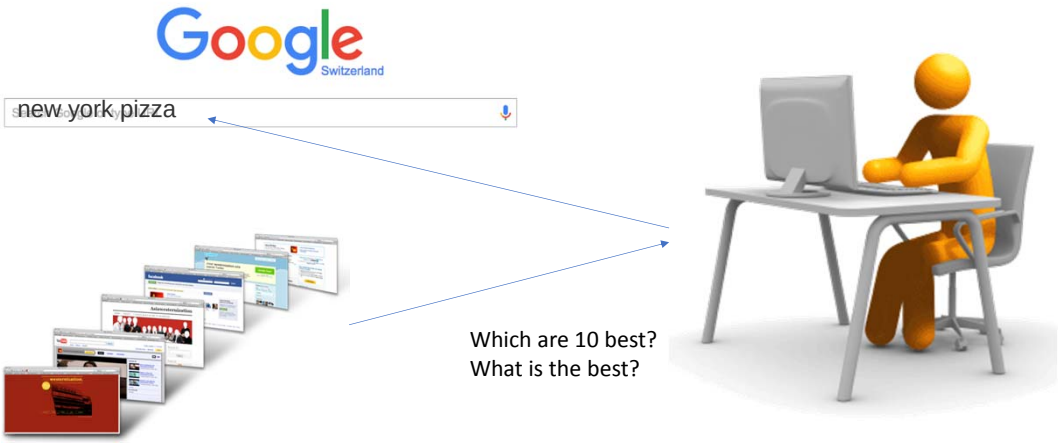


Index every non-trivial word on every page

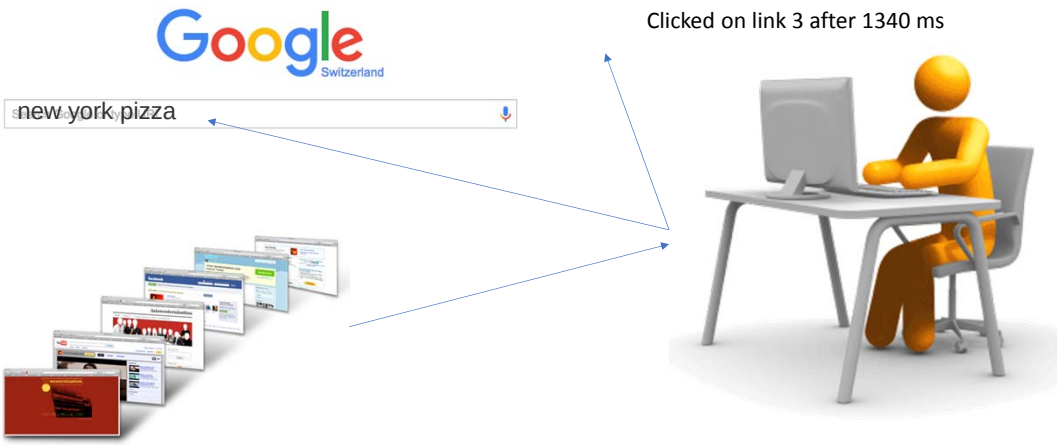
Just Takes Computers

1632 for ~50 billion web pages
(x6 for availability and throughput)

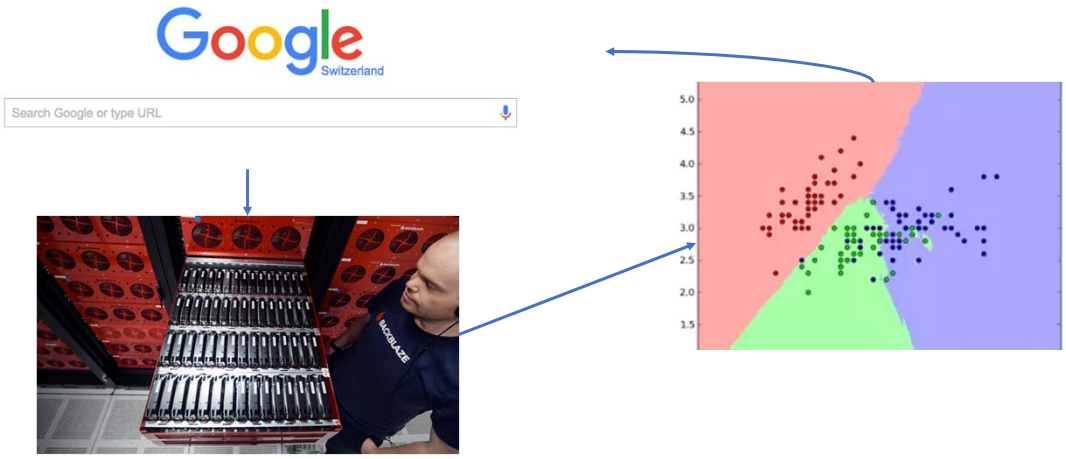
Hard Part



Currently Machine-Learned Algorithm



Clickstream is BIG Data



Search Engine Relevance (NDCG)

Machine Learning

EPFL Master Program in Data Science

Martin Jaggi
EPFL

13th Dec 2016

What is Machine Learning?

What is Machine Learning?

software that can

learn from data

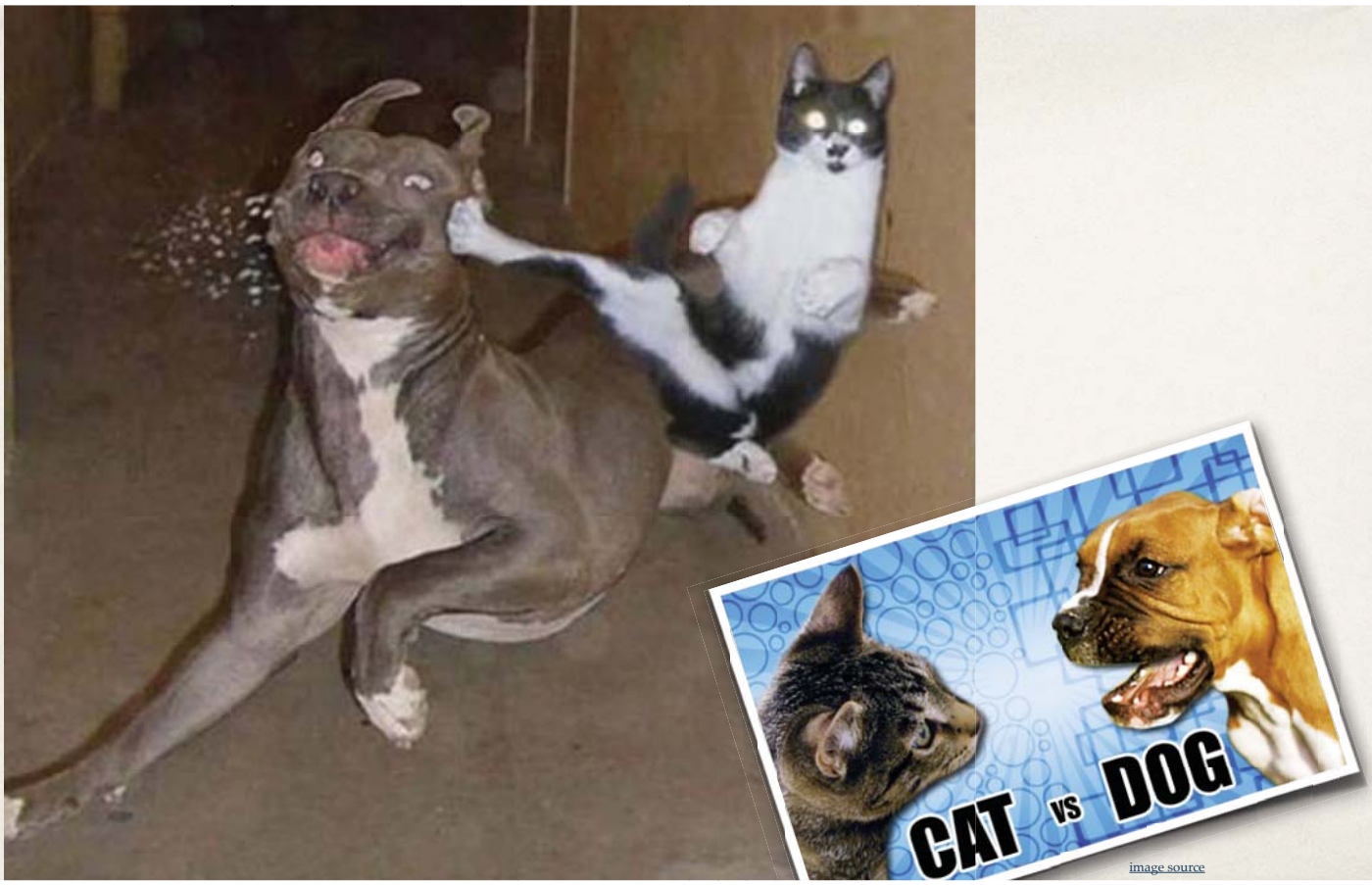
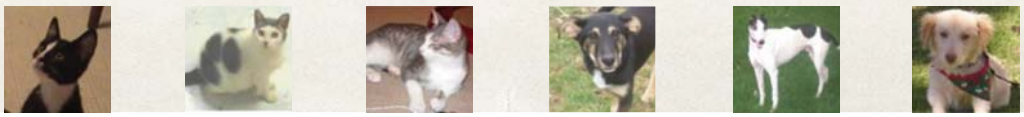


image source

Classification

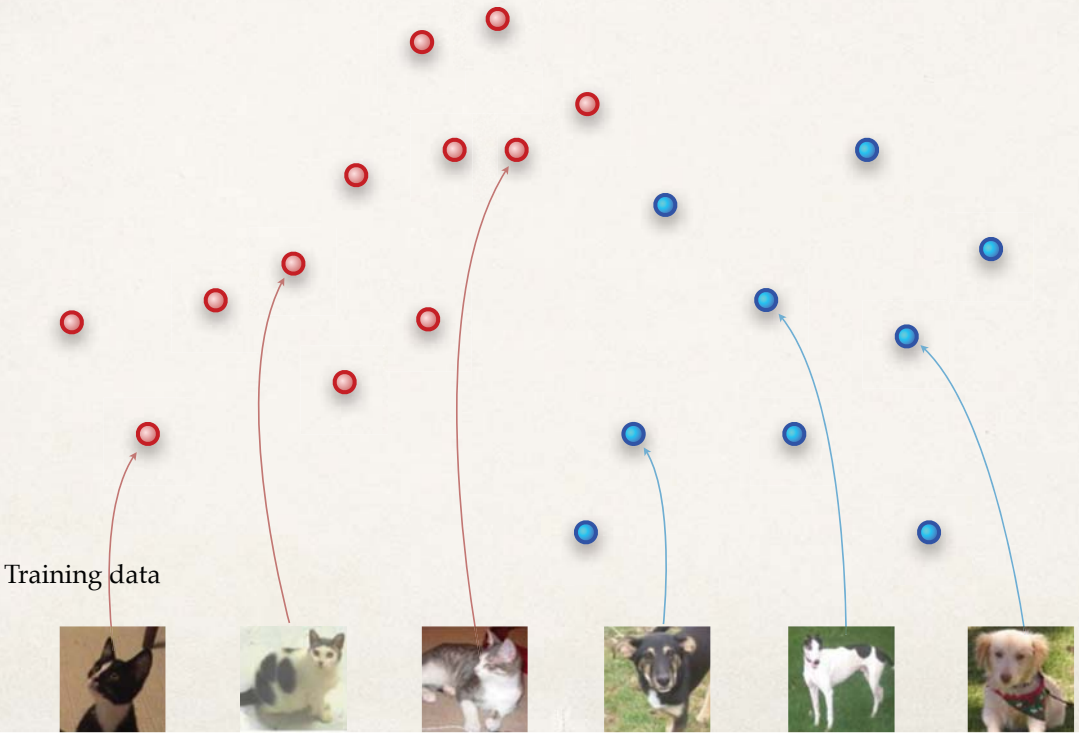
$\boldsymbol{x}_i \in \mathbb{R}^d$

Training data



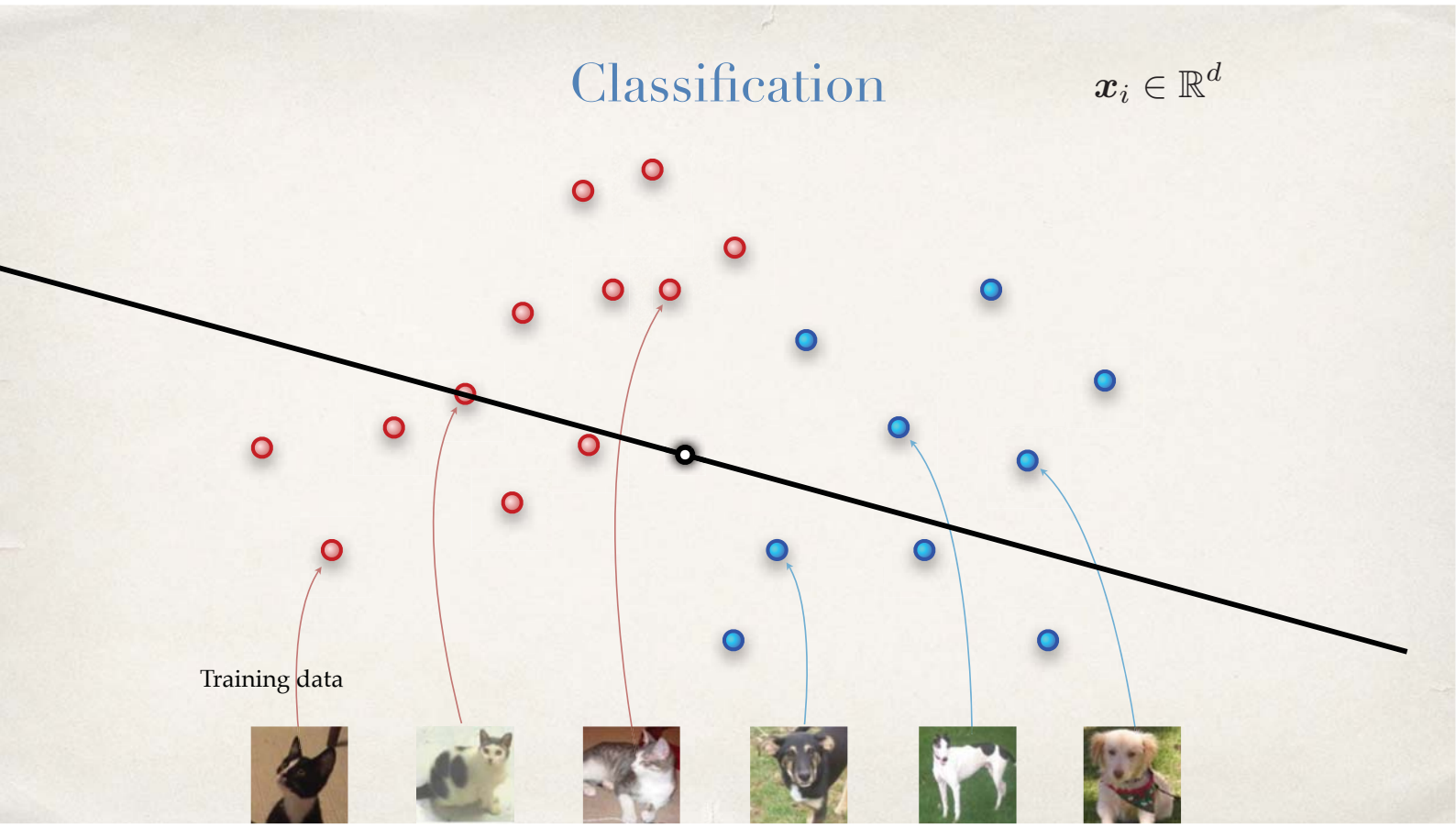
Classification

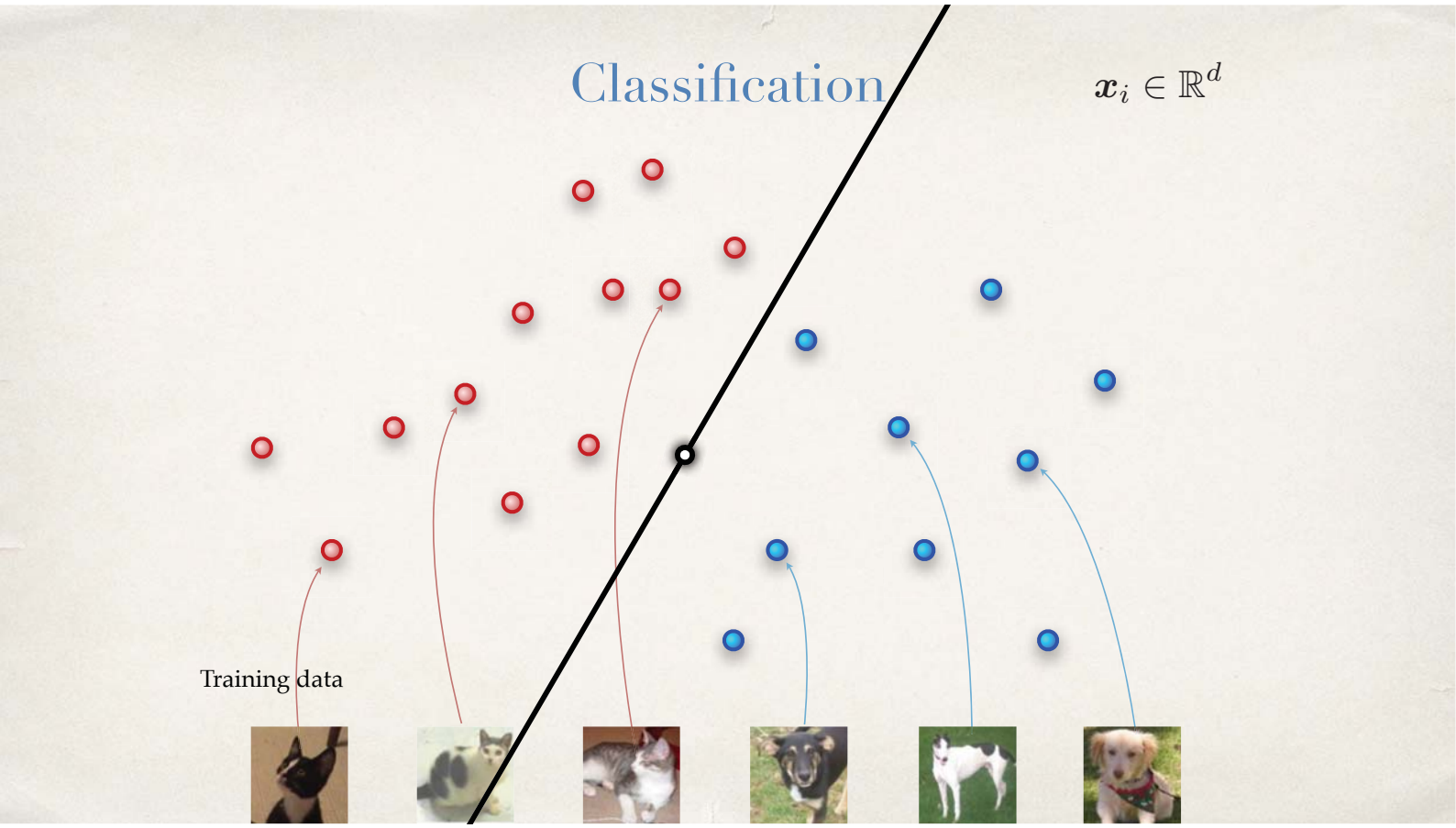
$\boldsymbol{x}_i \in \mathbb{R}^d$

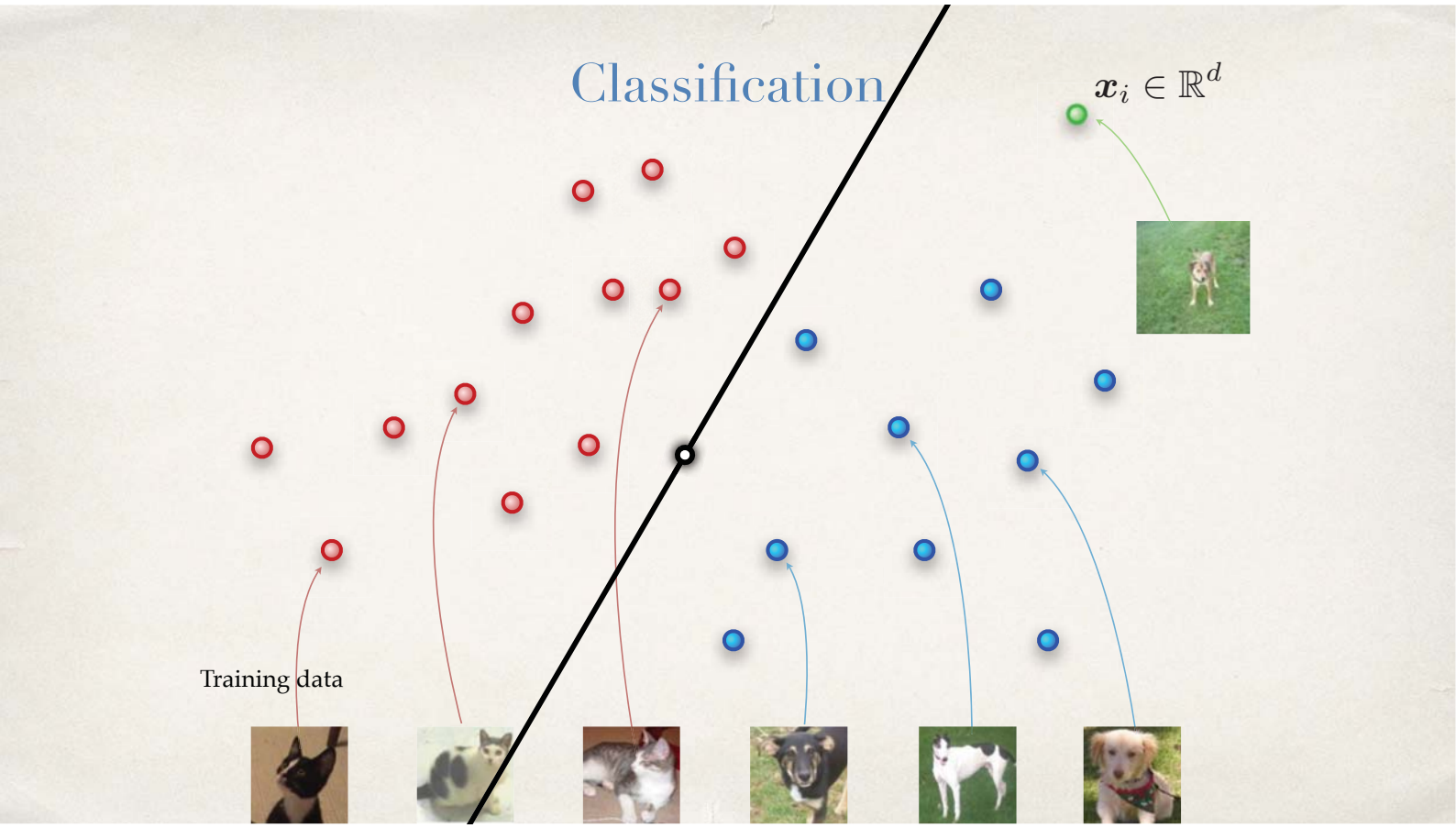


Classification

$\boldsymbol{x}_i \in \mathbb{R}^d$




Classification
$\boldsymbol{x}_i \in \mathbb{R}^d$
Training data


Classification
$\boldsymbol{x}_i \in \mathbb{R}^d$
Training data

towards…

understanding intelligence

?

if-then-else

≠

intelligence

towards…

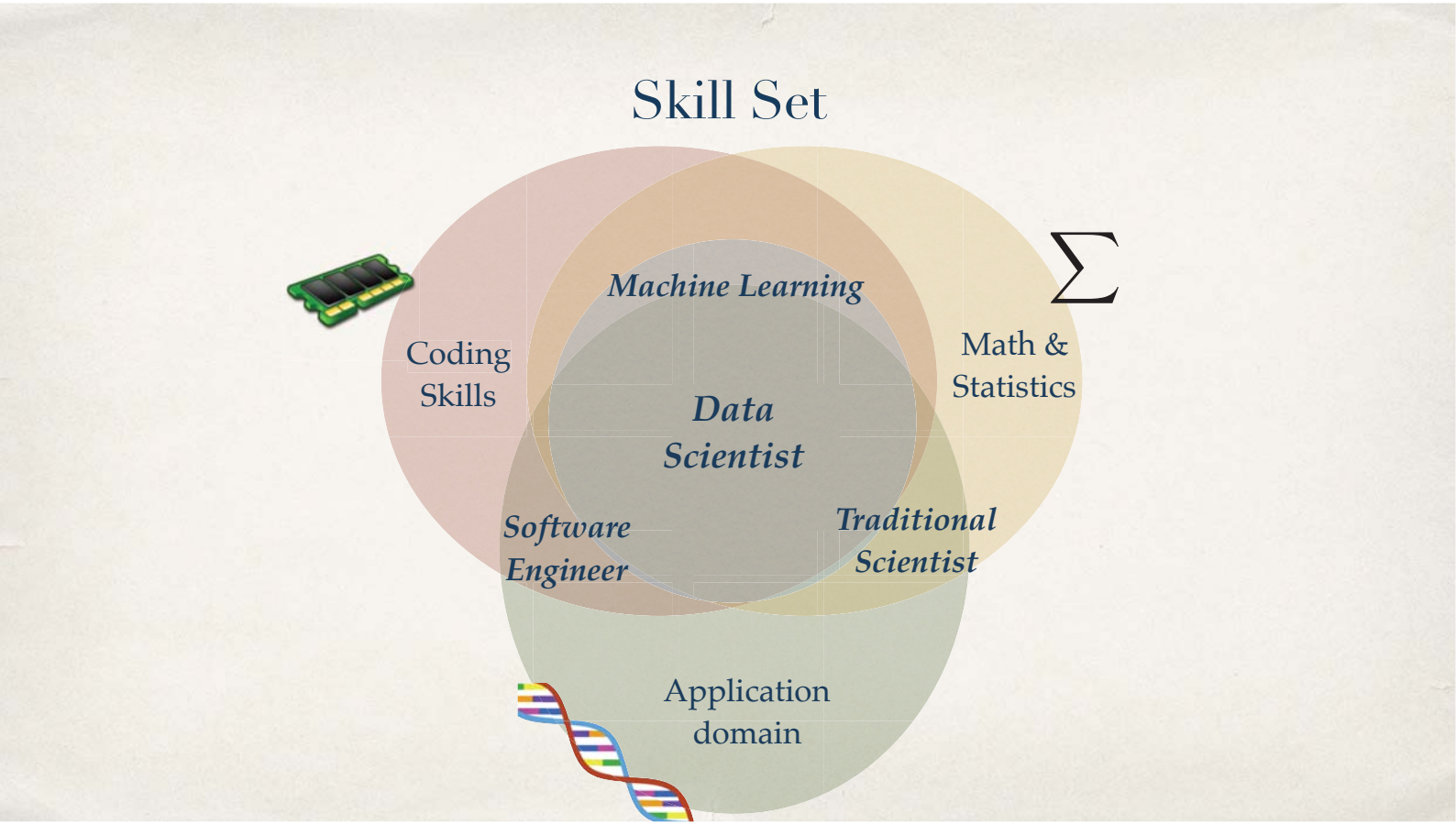
understanding intelligence

?

Machine Learning

vs

Neuroscience


Skill Set
Machine Learning
Coding Skills
Math & Statistics
Data Scientist
Software Engineer
Traditional Scientist
Application domain

why ML?

Applications

Industry Applications

Applications in Other Sciences

- increasingly data driven
  - science of X ➔ *digital* science of X

Image Data

- Astronomy
- Face recognition
- 2D + 3D medical imaging
- OCR
- self-driving cars

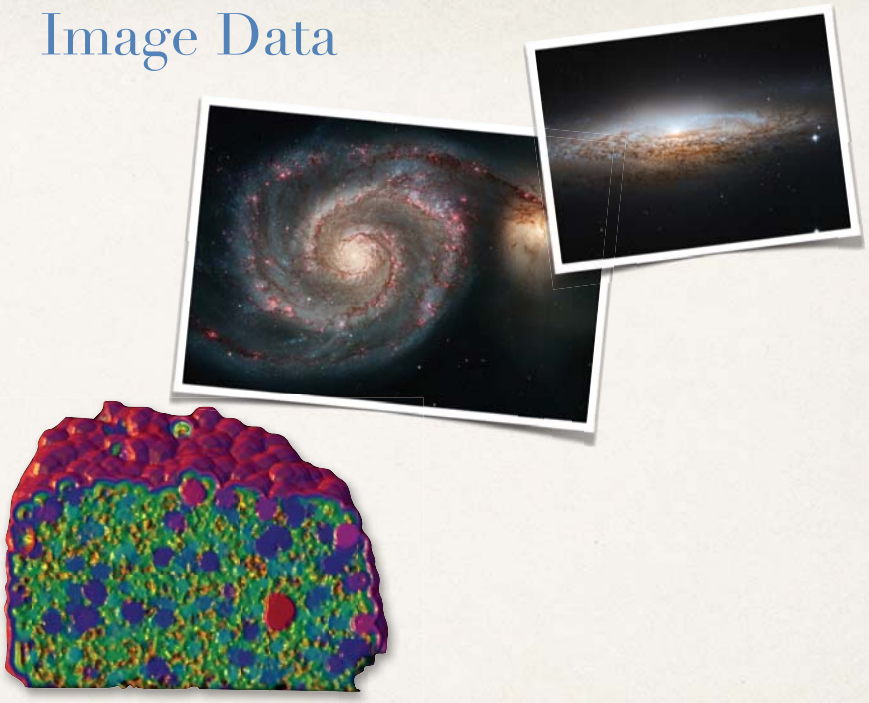


Image Data

- Astronomy
- Face recognition
- 2D + 3D medical imaging
- OCR
- self-driving cars

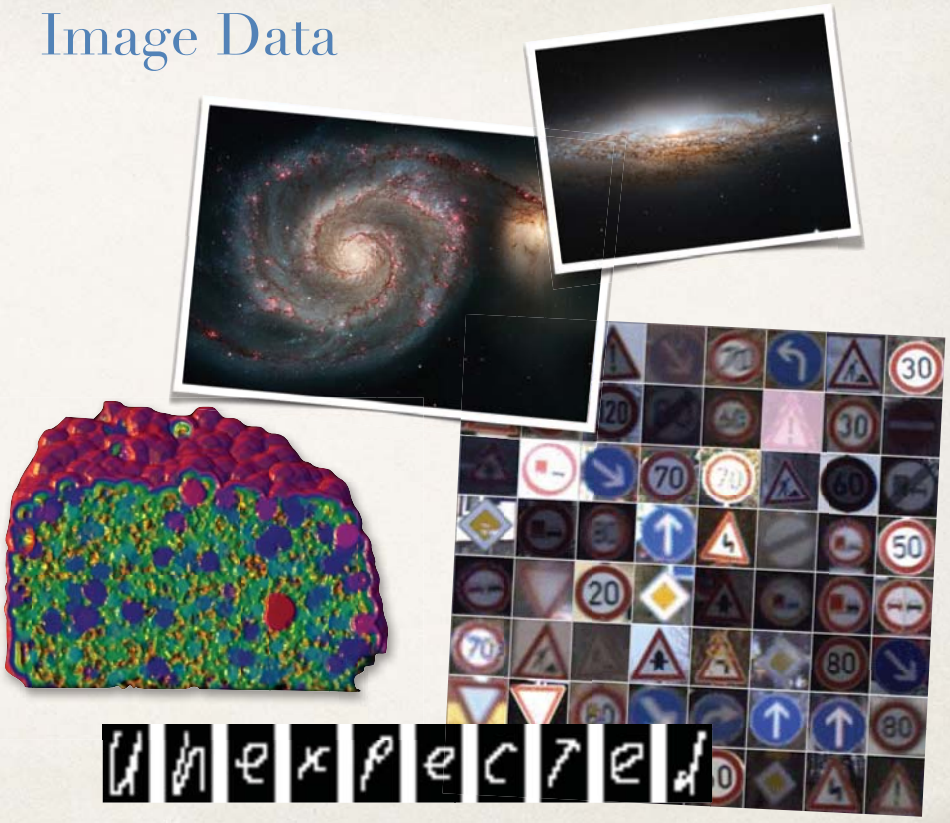
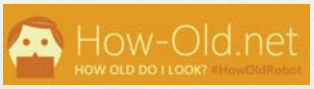


Image Data

- Astronomy
- Face recognition
- 2D + 3D medical imaging
- OCR
- self-driving cars

how-old.net

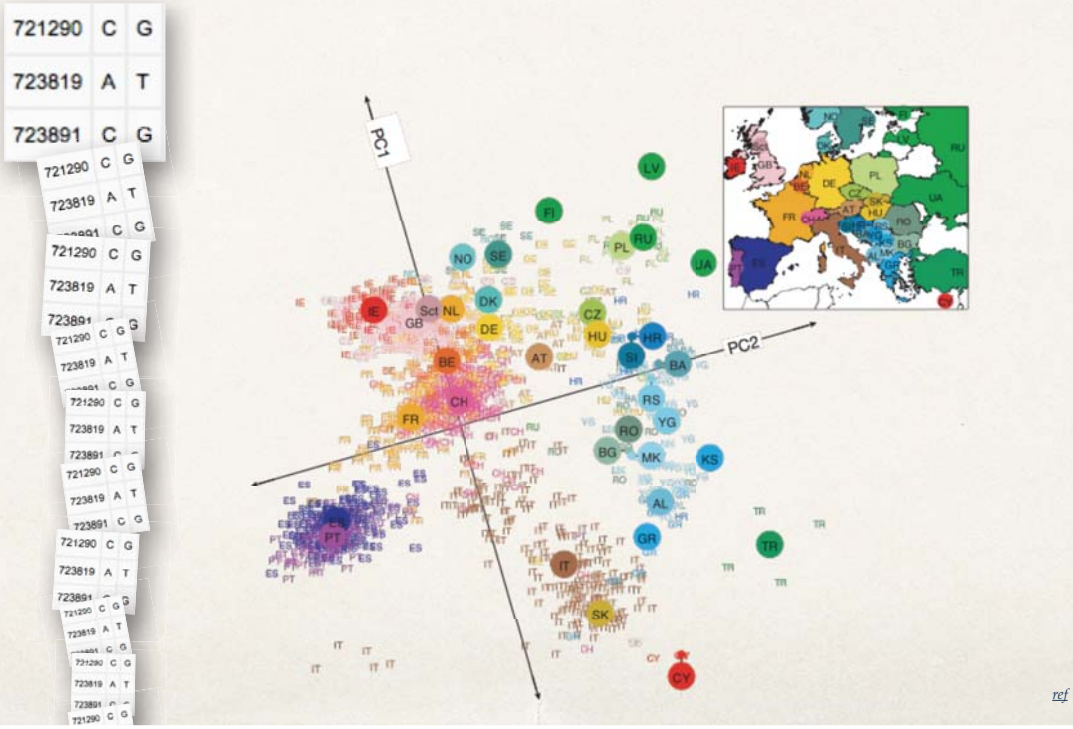


Text Data

- Spam Detection
- User Content
- Medical Data

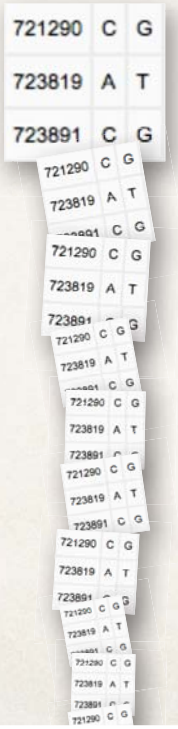
negative	*neutral*	But i wanna wear my Concords tomorrow though but i don't feel like it
positive	*neutral*	Gonna watch Grey's Anatomy all day today and tomorrow(:
negative	*neutral*	@CoachVac heey do you know anything about UVA's fallll fest loll they invited me
neutral	*neutral*	@DustyEf when that sun is high in that Texas sky, I'll be buckin it to county fair. A
neutral	*positive*	Up 20 points in my money league with Vernon Davis and L. Fitz still to go tomorro
neutral	*positive*	DEEJAYING this FRIDAY in THE FIRST CHOP it's CHRIS actual SMITH with a smash
		The Rick Santorum signing that was scheduled for tomorrow at the Books A Million
		@dreami9 lol yep looks like it! Was after El Clasico on Sunday. I didn't like her lol
		Back in Stoke on Trent for the 2nd time today!
		First Girls Varsity Basketball Game tomorrow at 6:00 pm Then Football Senior nigh
		#UFC lightweights @Young__Assassin VS @jamievarner set for TUF 16 Finale on t
		@OOOOO_WEEEE slide thru sometime this weekend ill have somethin yu can sip o
		@DannyB618 Sure absolutely-- I meant out of the Bachmann, Perry, Santorum, H
		@RichardGordon48 re Levein discussion on Wed. Can't keep changing boss, but he
		Today In History November 02, 1958 Elvis gave a party at his hotel before going o
		Hustle cause you got to then kick back n party everyday like its Fri
		I can't sleep. Way too exited about Vancouver tomorrow! I'm like a kid at Christma

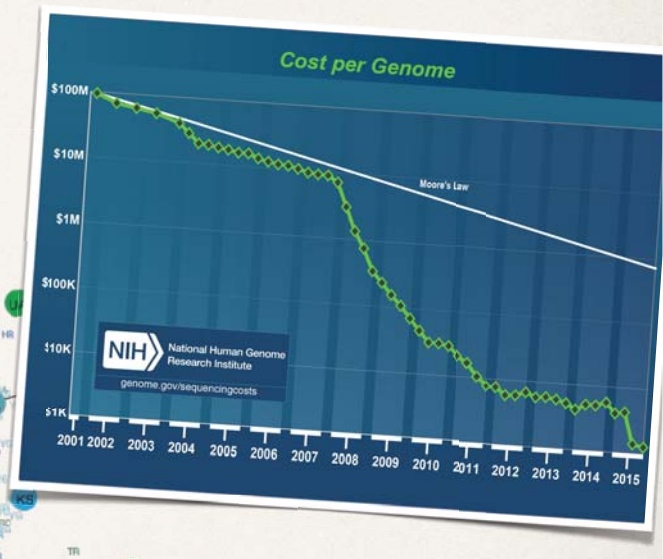
Medical: Genetic Data



ref

Medical: Genetic Data

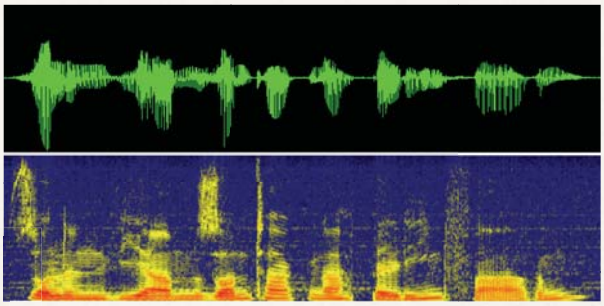




ref

Audio Data

- Hearing aids
- Voice Recognition
- Automatic Translation



Numerical / Sensor Data

- Cern
- SKA, Telescopes
- Fitness Trackers
- Weather Forecast
- Robotics
- Kinect

Numerical / Sensor Data

- Cern
- SKA, Telescopes
- Fitness Trackers
- Weather Forecast
- Robotics
- Kinect

Numerical / Sensor Data

- Cern
- SKA, Telescopes
- Fitness Trackers
- Weather Forecast
- Robotics
- Kinect

Numerical / Sensor Data

- Cern
- SKA, Telescopes
- Fitness Trackers
- Weather Forecast
- Robotics
- Kinect

Numerical / Sensor Data

- Cern
- SKA, Telescopes
- Fitness Trackers
- Weather Forecast
- Robotics
- Kinect

Numerical / Sensor Data

- Cern
- SKA, Telescopes
- Fitness Trackers
- Weather Forecast
- Robotics
- Kinect

Games & Simulations

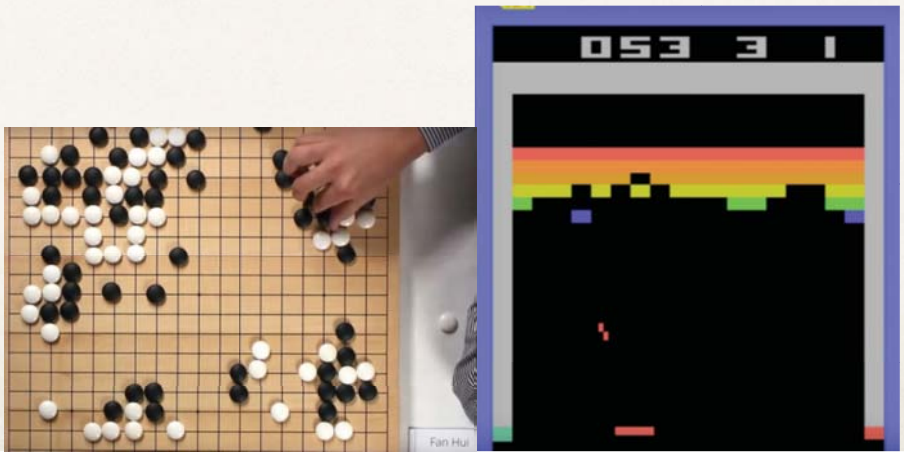
- Immediate Feedback

Games & Simulations

- Immediate Feedback
- Chess, Go

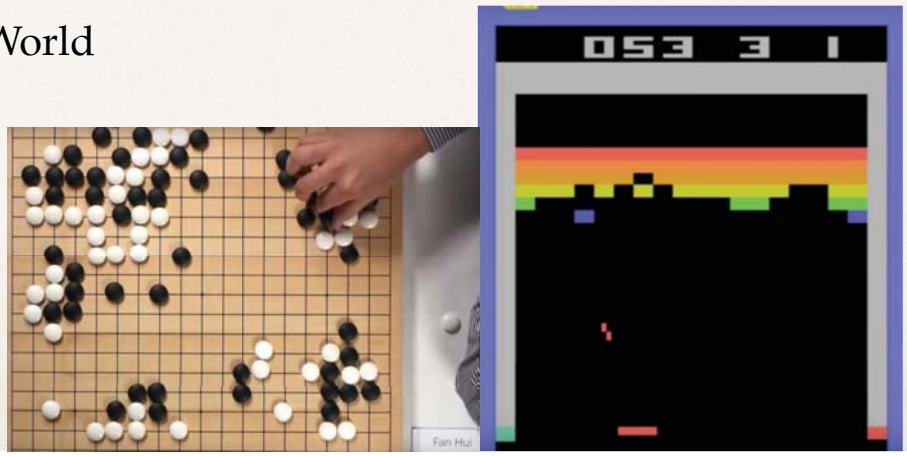
Games & Simulations

- Immediate Feedback
- Chess, Go



Games & Simulations

- Immediate Feedback
- Chess, Go
- Physical World



Internet Data

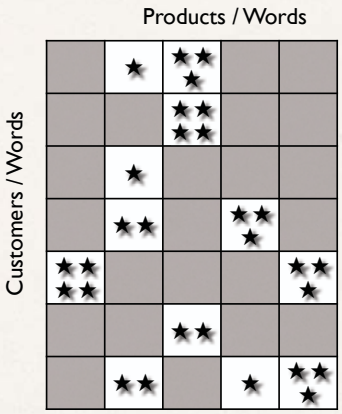
WIKIPEDIA

NETFLIX

facebook

amazon.com

Google



$\approx UV^T$

Prediction Markets

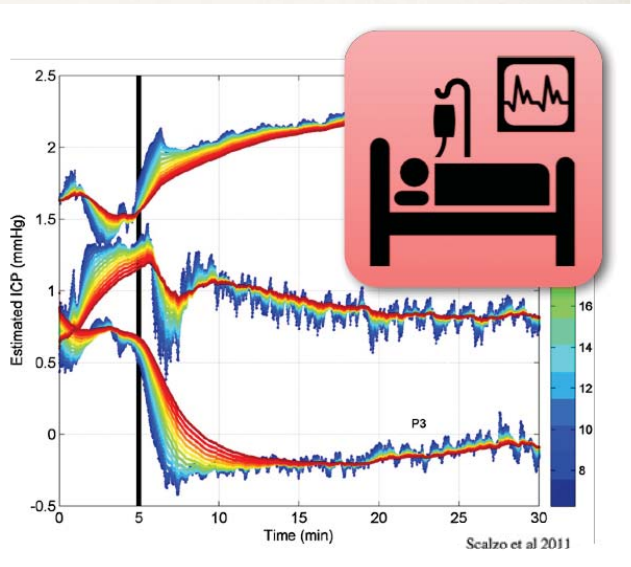
ML Applications by Master Students

Generating Maps from Satellite Images

MSc Thesis Project
Pascal Kaiser

Medical Applications in Intensive Care

Forecasting of intracranial hypertension

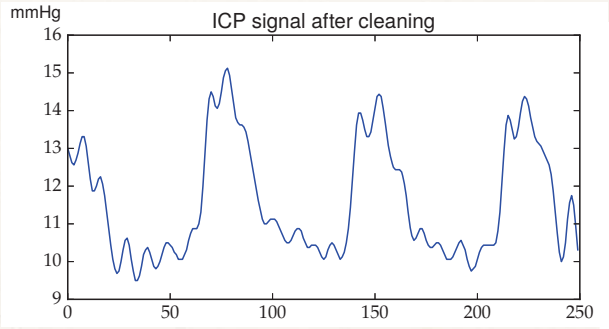


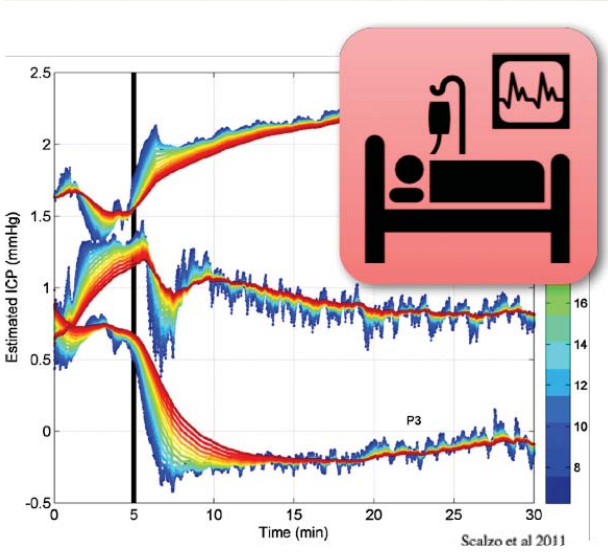
43 GB of raw CSV text data
67 days of ICP signal
sampling rate 125 Hz

MSc Thesis Projects
Matthias Hüser & Adrian Kündig

Medical Applications in Intensive Care

Forecasting of intracranial hypertension





43 GB of raw CSV text data
67 days of ICP signal
sampling rate 125 Hz

MSc Thesis Projects
Matthias Hüser & Adrian Kündig

Text Understanding

sentiment of tweets

SemEval 2016:
winner out of 34
teams, >20 countries

negative	*neutral*	But i wanna wear my Concords tomorrow though but i don't feel like it
positive	*neutral*	Gonna watch Grey's Anatomy all day today and tomorrow(:
negative	*neutral*	@CoachVac heey do you know anything about UVA's fallll fest loll they invited me
neutral	*neutral*	@DustyEf when that sun is high in that Texas sky, I'll be buckin it to county fair. A
neutral	*positive*	Up 20 points in my money league with Vernon Davis and L. Fitz still to go tomorro
neutral	*positive*	DEEJAYING this FRIDAY in THE FIRST CHOP it's CHRIS actual SMITH with a smash
negative	*negative*	The Rick Santorum signing that was scheduled for tomorrow at the Books A Million
positive	*neutral*	@dreami9 lol yep looks like it! Was after El Clasico on Sunday. I didn't like her lol
neutral	*neutral*	Back in Stoke on Trent for the 2nd time today!
neutral	*neutral*	First Girls Varsity Basketball Game tomorrow at 6:00 pm Then Football Senior nigh
neutral	*neutral*	#UFC lightweights @Young__Assassin VS @jamievarner set for TUF 16 Finale on t
neutral	*neutral*	@OOOOO_WEEEE slide thru sometime this weekend ill have somethin yu can sip c
negative	*negative*	@DannyB618 Sure absolutely-- I meant out of the Bachmann, Perry, Santorum, H
negative	*negative*	@RichardGordon48 re Levein discussion on Wed. Can't keep changing boss, but he
neutral	*neutral*	Today In History November 02, 1958 Elvis gave a party at his hotel before going o
neutral	*positive*	Hustle cause you got to then kick back n party everyday like its Fri
positive	*positive*	I can't sleep. Way too exited about Vancouver tomorrow! I'm like a kid at Christma

MSc Thesis Projects
Jan Deriu & Maurice Gonzenbach

Audio: Bird Songs & Ecology

another kind of tweets

winner of
BirdCLEF 2016

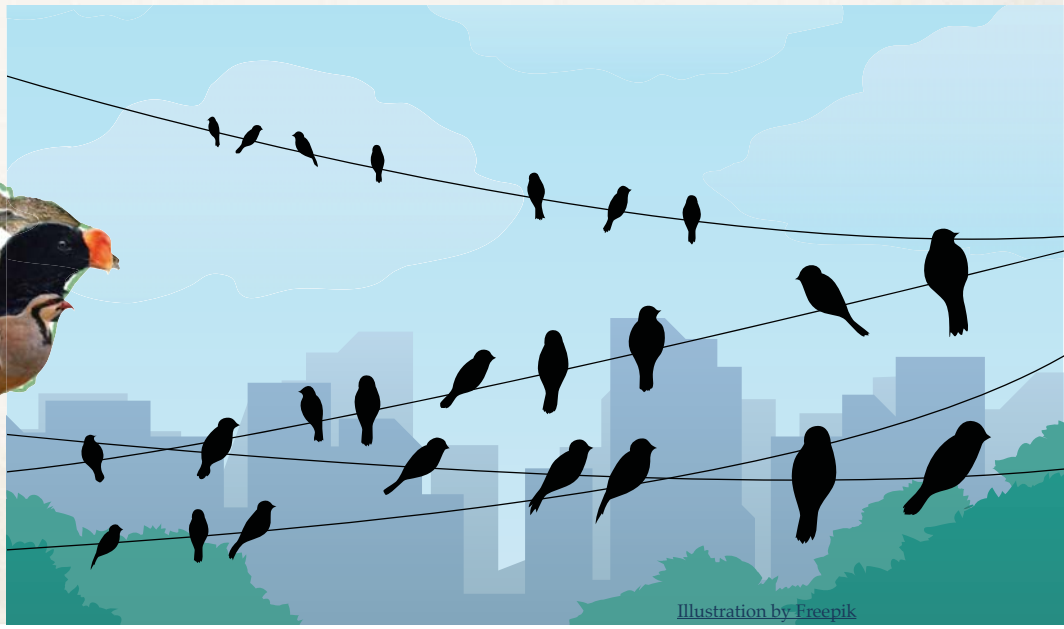


Illustration by Freepik

MSc Thesis Project
Elias Sprengel

thanks

MS in Data Science
- starting Fall 2017 -

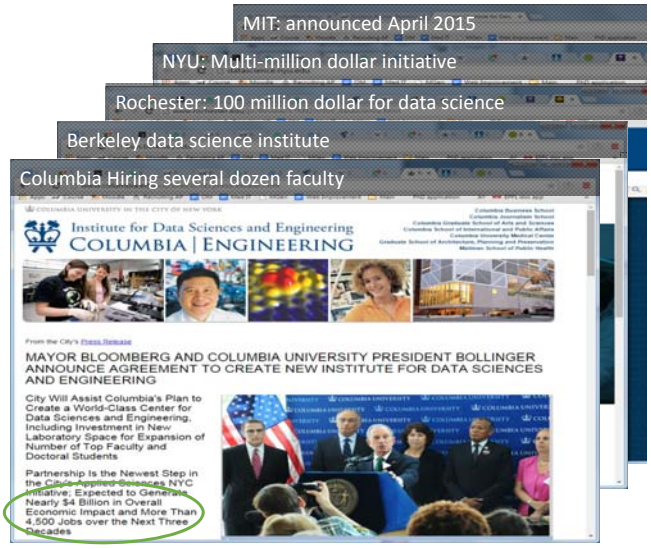
Séance information aux étudiants bachelor en IC

Dec 13, 2016

Data Science job market

Data science

Elsewhere



Data Science Overview - **Carnegie Mellon University**
Master's of Science in Data Science (**U of Minnesota**)
Data Science - School of Informatics and Computing (**Indiana U**)
Data Science Institute - **Columbia University**
Institute for Data Science : **University of Rochester**
Center for Data Science - **New York University**
University of Warwick: BSc in Data Science
Data Science Institute - **University of Virginia**
Data Science - **University of San Francisco**
Data Science - **University of Southern California**
Data Science Certificate - **Harvard** Extension
Data Science - **University of Glasgow**
Data Science - **Kent State University**
Data Science Center **Eindhoven** (DSC/e)
Major: Data Science | LSA Students | **University of Michigan**
Delft University of Technology: The Data Science ... - **TU Delft**
Data Science - Computer Science Department (**U Maryland**)
Data Science - **UC Irvine** Extension
Imperial and Zhejiang University launch data science ...
Information & Data Sciences | **Boston University**
Analytics and Data Science | **University of Technology, Sydney**
Data Science - **Carleton University**
MSc in Data Science at **Heriot-Watt University, Edinburgh**
Data Science - **University of St. Thomas**
Data Science MSc at **Lancaster University**
The Warren Center for Data & Network Sciences - **U Penn**
Data Science Center - **Tel Aviv University**
Data Science Center - **U Twente**
Centre for Doctoral Training in Data Science - **U Edinburgh**
Data Science Research Center – **Amsterdam**
Warwick Data Science Institute (WDSI) - **University of Warwick**
Paris-Saclay Center for Data Science
Erasmus Centre for Data Science and Business Analytics
...

Yes, Data Science programs are everywhere, but…

Complete academic programs:

- They are rare.
- Many existing programs are "vocational training".

Focus on Foundations:

- Statistics
- Information Theory, Signal Processing

Focus on Algorithms:

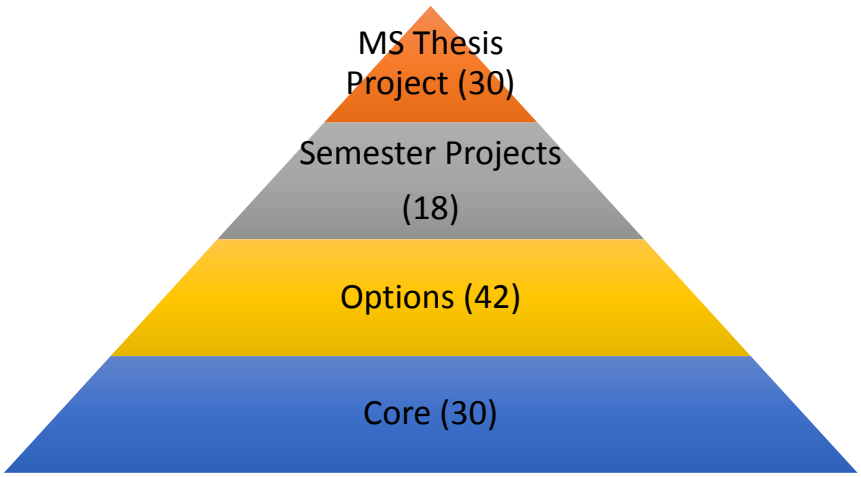
- Machine Learning

Focus on Systems Design:

- Database systems, Big Data
- Information Security and Privacy

Focus on Practical Real-World Data

Structure of Program (120 credits)



Core (30 credits: 5 out of 8 classes)

Existing Classes:

- Advanced Algorithms (CS-450)
- Machine Learning (CS-433)

Evolutions of Existing Classes:

- Systems for Data Science (DS-?)
- Applied Data Analysis (CS-401, new in Fall 2016)

New Classes lined up (more or less):

- Statistics for Data Science (MATH-?)
- Information security and privacy (COM-402)

New Classes being designed:

- Information Theory and Signal Processing
- Optimization for Machine Learning and Data Science

Optional Courses (42 credits)

- Current plan is based on ***existing*** classes mostly.
- Classes from
  - Computer Science
  - Communication Systems
  - Mathematics
  - Statistics
  - Electrical Engineering
- "Data Visualization"
- Optional Project (8 credits)

- Future optional courses to be developed specifically for Data Science.
  - IC is hiring in Machine Learning and Data Science

Semester Projects (18 credits)

- Semester projects can be carried out in IC (just like our MS students in Computer Science and SysCom)
- Or in other data science related labs on campus (SV and many more).

- SHS Project (6 credits, as usual)

Masters Project (30 credits)

- MS Projects will be carried out in industry or at EPFL (as all MS projects in IC).

- **Fact:** More than half of the current MS projects in IC (Computer Science and Communication Systems) ***already*** involve Data Science anyway.

Who can access the MS in Data Science?

- It is a regular MS in IC.
- All IC Bachelor students can enter just like they can enter the MS in Computer Science and the MS in Communication Systems.

- It is open to all EPFL BS students *on dossier*.
- Generically, this will require additional *prerequisite* course work (as any switch in study major at EPFL).
- Some of this could be taken already in the 3rd year BS as optional credit.

Key Prerequisites (Draft)

- Already mandatory for both CS and SysCom:
  - Algorithms (CS 250)
  - Theory of Computation (CS 251)
- Additionally recommended for SysCom (mandatory for CS):
  - Introduction to Database Systems (CS 322)
  - Parallelism and Concurrency (CS 206)
  - Functional Programming (CS-210)
- Additionally recommended for CS (mandatory for SysCom):
  - Circuits & Systems II (EE-205)
  - Stochastic Models (COM 300)
  - Signal Processing (COM 303)

Footnotes

- **"Specializations":**
  Initially, this MS will not offer specializations.
  - It is sufficiently specialized.
- But specializations are definitely an option for the future.

- **Digital Humanities:**
  - No overlap in key Job Markets.
  - Small overlap in intellectual contents:
    - Two core courses ("Applied Data Analysis" and "Pattern Classification and Machine Learning")

2017-2018 **DATA SCIENCE** **Cycle Master**

Code	Matières	Enseignants sous réserve de modification	Sections	Semestres MA1 c	MA1 e	MA1 p	MA2 c	MA2 e	MA2 p	Crédits	Période des épreuves	Type examen
	Groupe "Core courses et options"									72		
	Groupe 1 "Core courses"									min. 30		
CS-450	Advanced Algorithms	Svensson	IN				4	2	1	7	E	écrit
DS-401	Applied Data Analysis	Catasta	SC	2	2					6	H	écrit
COM-402	Information security and privacy	Ford	IN				2		2	6	E	écrit
DS-402	Information Theory and Signal Processing	Gastpar / Telatar / Urbanke	SC	4	2					6	H	écrit
CS-433	Machine learning	Jaggi / Urbanke	IN / SC	4	2					7	H	écrit
DS-403	Optimization for Machine Learning	Jaggi	IN				2	2		4	E	écrit
MATH-???	Statistics for Data Science	Panaretos	MA	4	2					6	H	écrit
DS-404	Systems for Data Science	Koch	IN				2	2	2	6	E	écrit
	Groupe 2 "Options"	**(la somme des crédits des groupes 1 et 2 doit être de 72 crédits au minimu**										
---	Cours à option	Divers enseignants	Divers									
	Bloc "Projets et SHS" :									18		
DS-416	Projet de semestre en data science	divers enseignants	SC	◄		2			►	12	sem A ou P	
HUM-nnn	SHS : introduction au projet	divers enseignants	SHS	2		1				3	sem A	
HUM-nnn	SHS : projet	divers enseignants	SHS						3	3	sem P	
	Total des crédits du cycle master									90		

Stage d'ingénieur :

Voir les modalités dans le règlement d'application

Mineurs :

Le cursus peut être complété par un des mineurs figurant dans l'offre de l'EPFL (renseignements à la page sac.epfl.ch/mineurs),
à l'exclusion des mineurs "Computer Engineering", "Informatique", "Information security" et "Systèmes de communication" qui ne peuvent pas être choisi.
Parmi les mineurs offerts par l'EPFL, la section recommande à ses étudiants les mineurs suivants :

- Biocomputing (SIN)
- Computational Science and Engineering (SMA)
- Management de la technologie et entrepreneuriat (SMTE)
- Technologies biomédicales (SMT)
- Technologies spatiales (SEL)

Le choix des cours de tous les mineurs se fait sur conseil de la section de l'étudiant et du responsable du mineur.

2017-208 **DATA SCIENCE - Options** **Cycle Master**

Code	Matières	Enseignants sous réserve de modification	Sections	Semestres MA1 c	MA1 e	MA1 p	MA2 c	MA2 e	MA2 p	Crédits	Période des épreuves	Type examen	Cours biennaux donnés en
EE-558	A Network Tour of Data Science	Bresson/Vandergheynst	EL	2	2					4	sem A		
COM-501	Advanced cryptography	Vaudenay	SC				2	2		4	E	écrit	
COM-417	Advanced probability and applications	Lévêque	SC				3	2		6	E	écrit	
CS-435	Analytic algorithms	Vishnoi	IN	2	1					4	sem A		
CS-???	Artificial neural networks	Gerstner	IN				2	1		4	sem P		
COM-415	Audio signal processing and virtual acoustics	Faller/Kolundzija	SC	2	2					4	H	écrit	
EE-592	Automatic speech processing	Bourlard	EL	2	1					3	H	écrit	
BIO-465	Biological modeling of neural networks	Gerstner	IN				2	2		4	E	écrit	
MATH-460	Combinatorial optimization	Eisenbrand	MA	2	2					5	H	écrit	
MATH-453	Computational linear algebra	vacat	MA				2	2		5	E	oral	
CS-413	Computational Photography	Süsstrunk	SC				2		2	5	E	oral	
CS-442	Computer vision	Fua	IN				2	1		4	E	écrit	
CS-454	Convex optimization and applications	Lebret	MTE				1	2		4	sem P		
COM-401	Cryptography and security	Vaudenay	SC	4	2					7	H	écrit	
DS-405	Data Visualization	Benzi	SC	2		2				4	sem A		
CS-411	Digital education & learning analytics	Dillenbourg/Jermann	IN	2		2				4	H	oral	
CS-423	Distributed information systems	Aberer	IN				2	1		4	E	écrit	
ENG-466	Distributed intelligent systems	Martinoli	SIE	2	2	1				5	H	oral	
MATH-360	Graph Theory	Kupavskii	MA				2	2		5	E	écrit	
CS-486	Human-computer interaction	Pu	IN				2	1	1	4	sem P		
EE-451	Image analysis and pattern recognition	Thiran J.-P.	EL				2		2	4	sem P		
CS-430	Intelligent agents	Faltings	IN	3	3					6	sem A		

CS-431	Introduction to natural language processing	Chappelier/Rajman	IN				2	2		4	E	écrit	
MATH-341	Linear models	Panaretos	MA	2	2					5	H	écrit	
COM-516	Markov chains and algorithmic applications	Lévêque/Macris	SC	2	2					4	A	écrit	
COM-514	Mathematical foundations of signal processing	Kolundzija/Scholefield/Bejar/ Parhizkar	SC	3	2					6	H	écrit	
EE-556	Mathematics of data: from theory to computation	Cevher	EL	2	2					4	sem A		
COM-512	Networks out of control	Thiran P./Celis	SC				2	1		4	E	écrit	2017-2018
DS-590	Optional project in Data Science	Divers enseignants	SC			2				8	sem A ou P		
COM-503	Performance evaluation (pas donné en 2017-2018)	Le Boudec	SC				3	1	2	7	E	oral	2018-2019
MATH-447	Risk, rare events and extremes	Davison	MA	2	2					5	H	écrit	
MATH-441	Robust and nonparametric Statistics	Morgenthaler	MA	2	2					5	H	oral	
EE-553	Speech Processing	Drygajlo	EL				2	1		3	E	oral	
COM-421	Statistical Neuroscience (pas donné en 2017-2018)	Gastpar	SC				2	2		4	E	écrit	
MATH-442	Statistical Theory	Panaretos	MA	2	2					5	E	écrit	
MATH-474	Statistics for genomic data analysis	Goldstein	MA	2	2					5	H	écrit	
COM-506	Student seminar : Security protocols and applications	Oechslin/Vaudenay	SC				2			3	E	écrit	
CS-410	Technology Ventures in IC	Bugnion	IN				2		2	4	sem P		
MATH-342	Time Series	Thiebaud	MA				2	2		5	E	écrit	
CS-455	Topics in theoretical computer science (pas donné en 2017-2018)	Svensson	IN	3	1					4	sem A		
CS-444	Virtual reality	Boulic	IN				2	1		4	sem P		