ETH-RAT

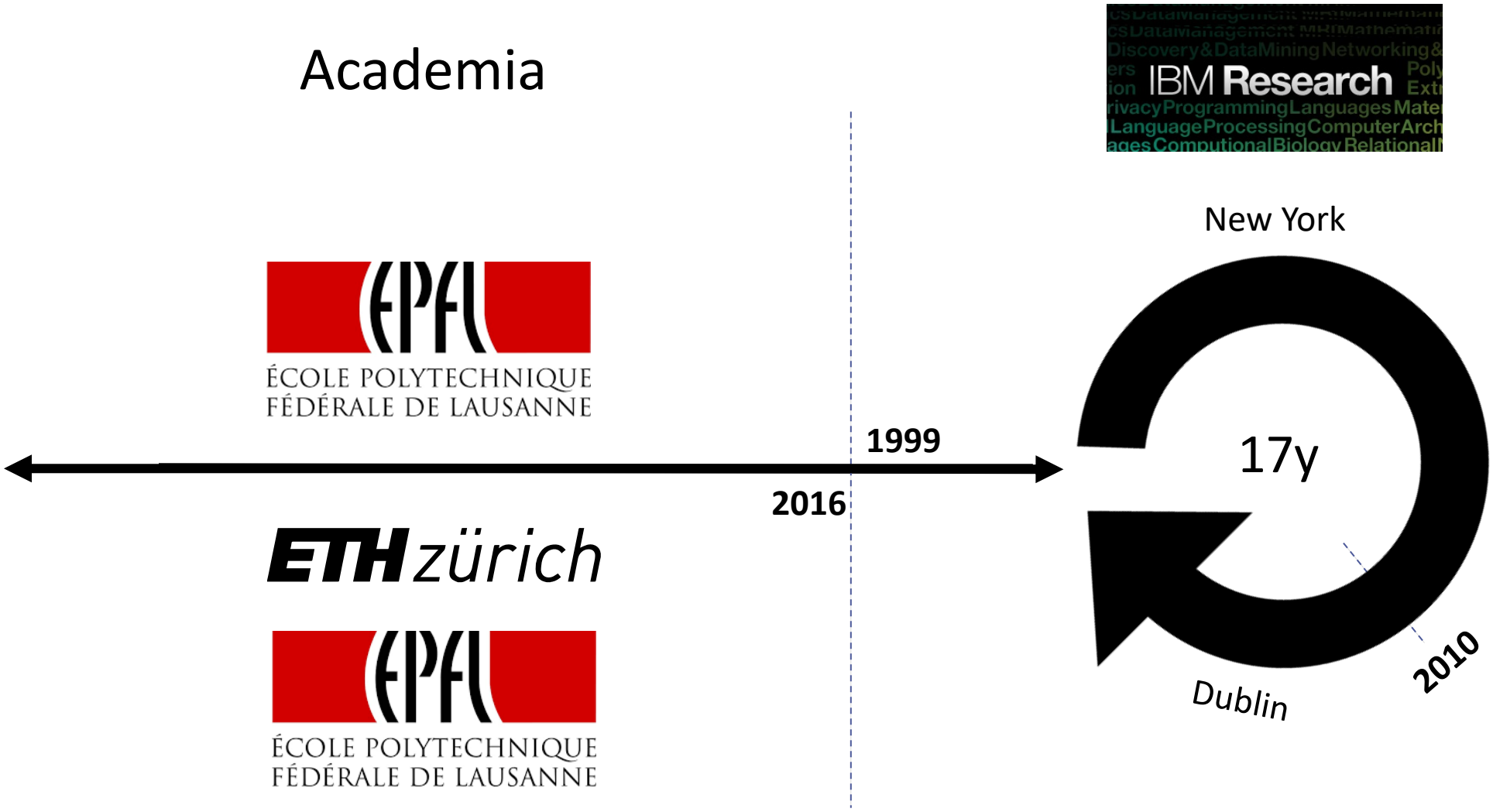
SDSC

DATA SCIENCE IN THE REAL WORLD
CHALLENGES & OPPORTUNITIES

Olivier Verscheure, PhD
Swiss Data Science Center
EPFL & ETH Zurich

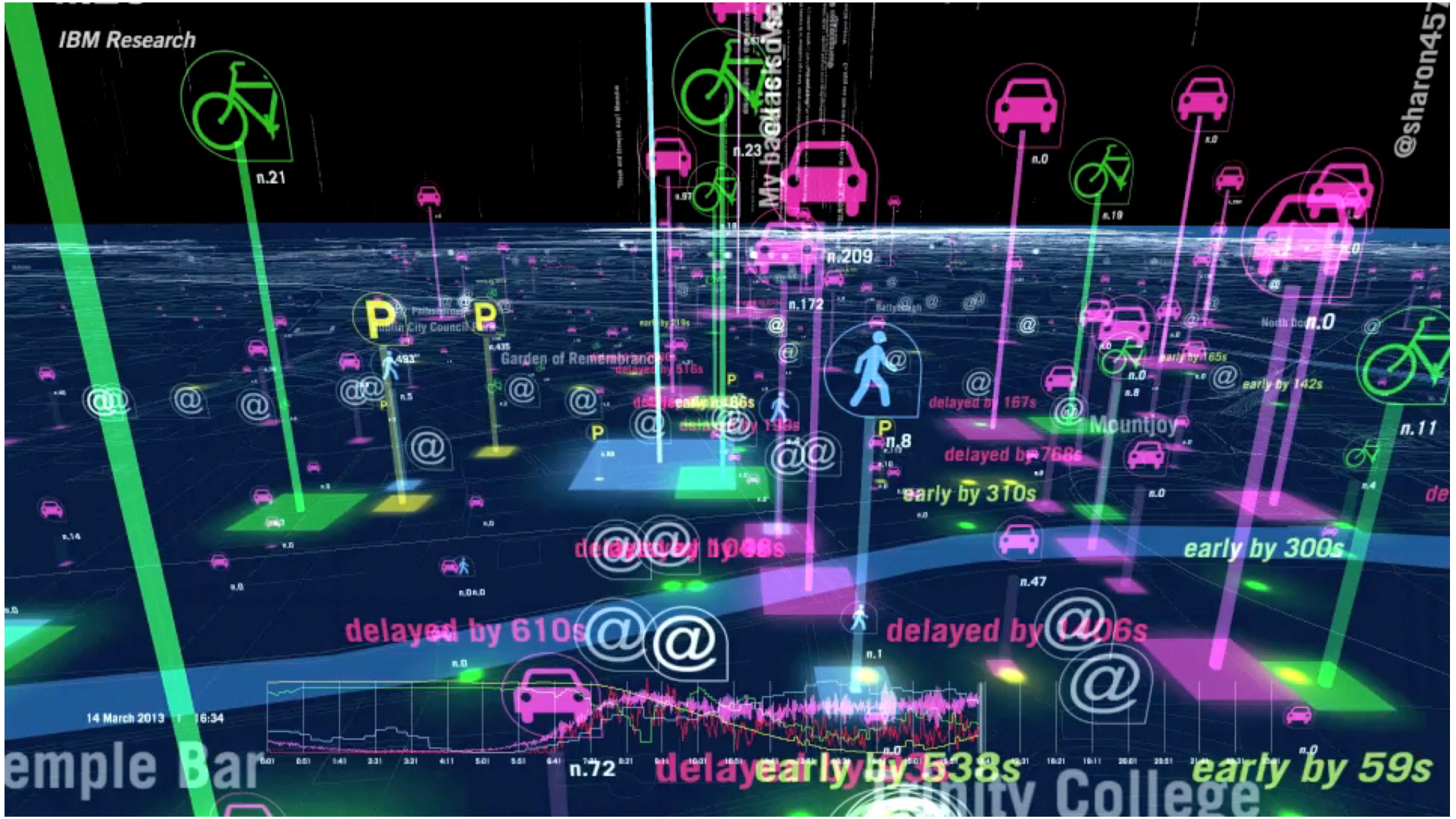
About me

Academia



What Do You See?

A Fantastic Source of Data

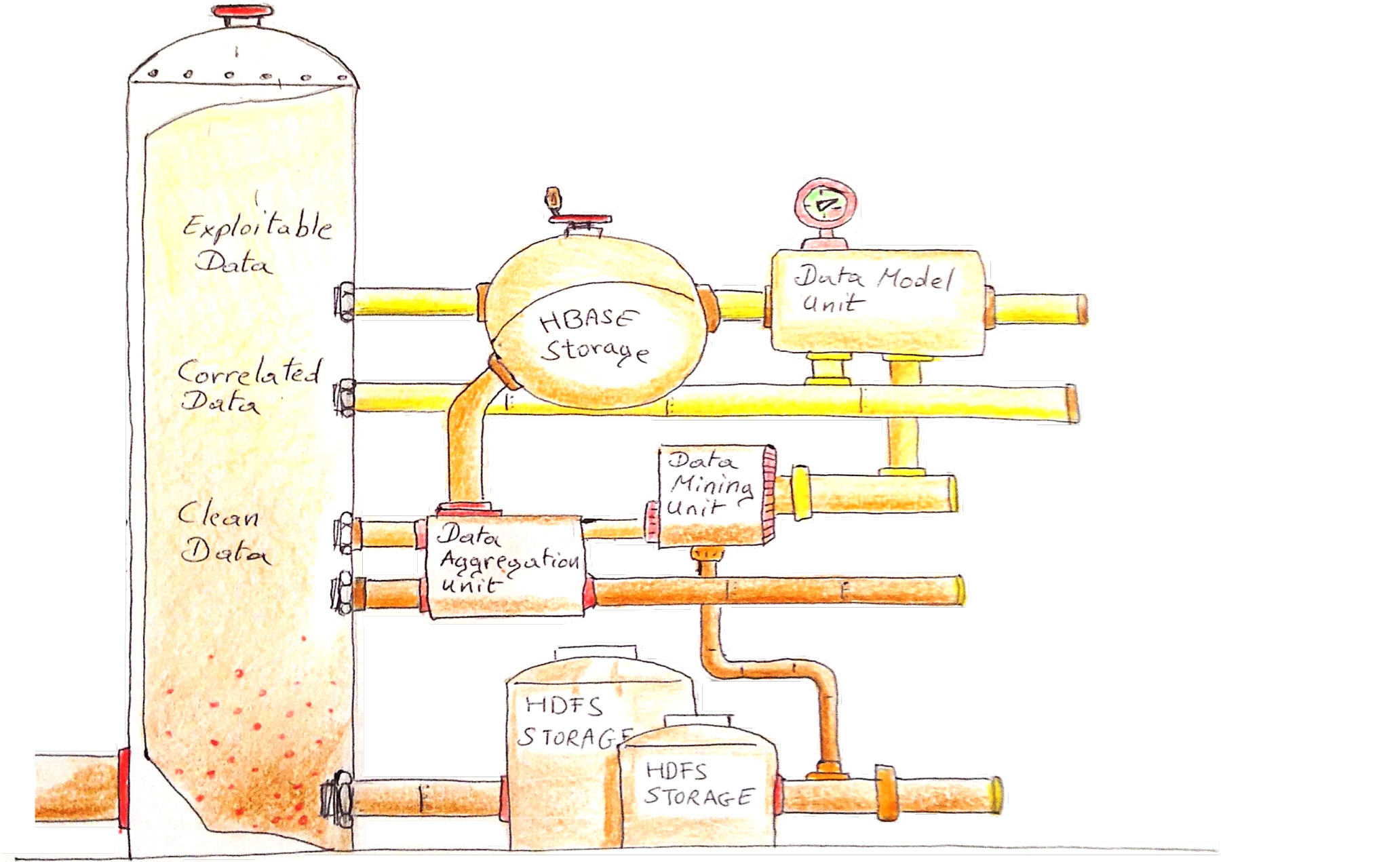


SDSC

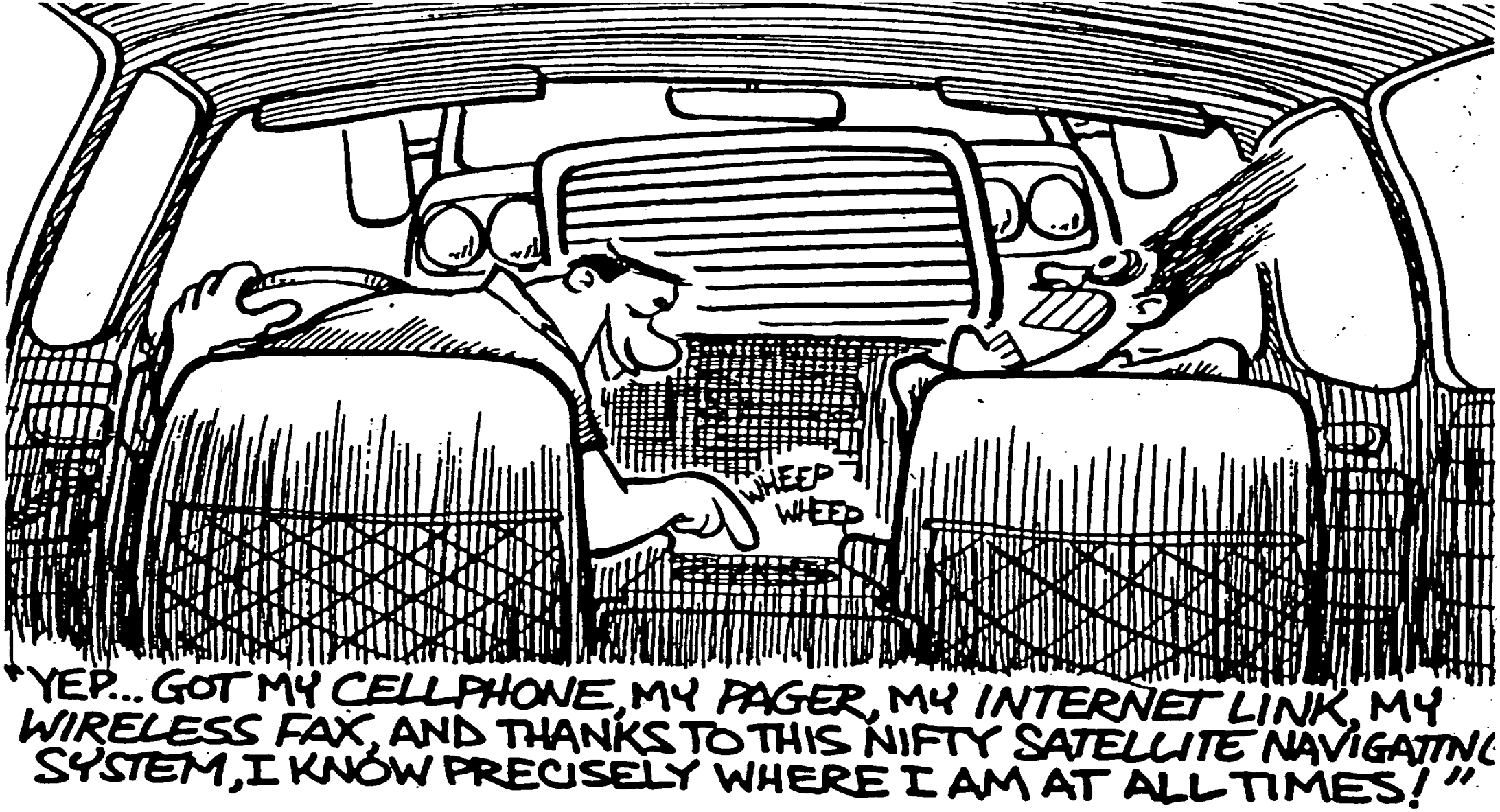
Data is the New Oil

The Economist, May 2017

Like Oil, Data Must be Refined



Big Data, Bad Data



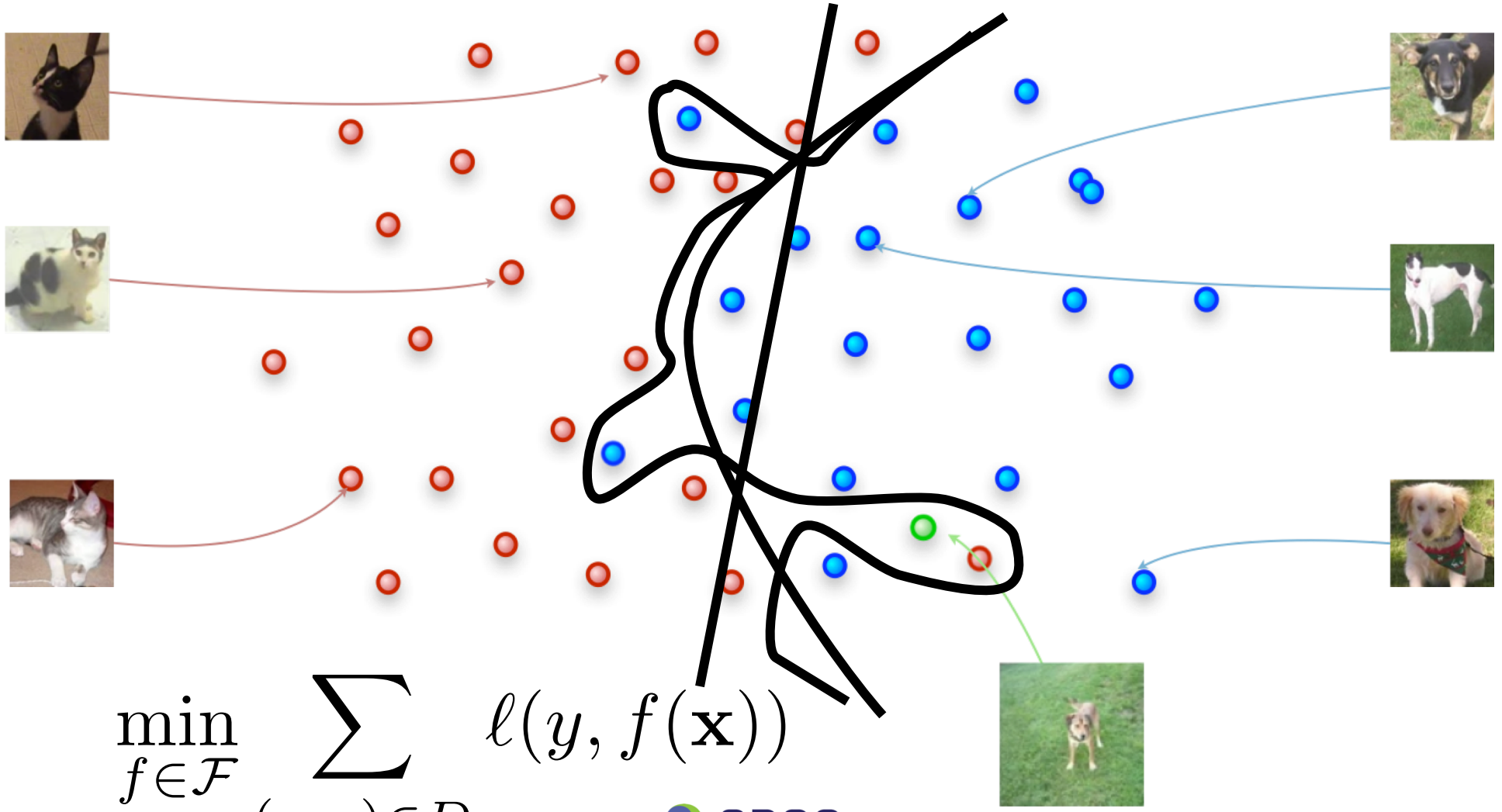
Data Science, a Fragmented Ecosystem

separates two classes of points
in multidimensional space?

Machine Learning in a Nutshell

$$f : \quad \longrightarrow \text{Cat}$$

Machine Learning (ML) in a Nutshell



$$\min_{f \in \mathcal{F}} \sum_{(\mathbf{x},y) \in D} \ell(y, f(\mathbf{x}))$$

SDSC

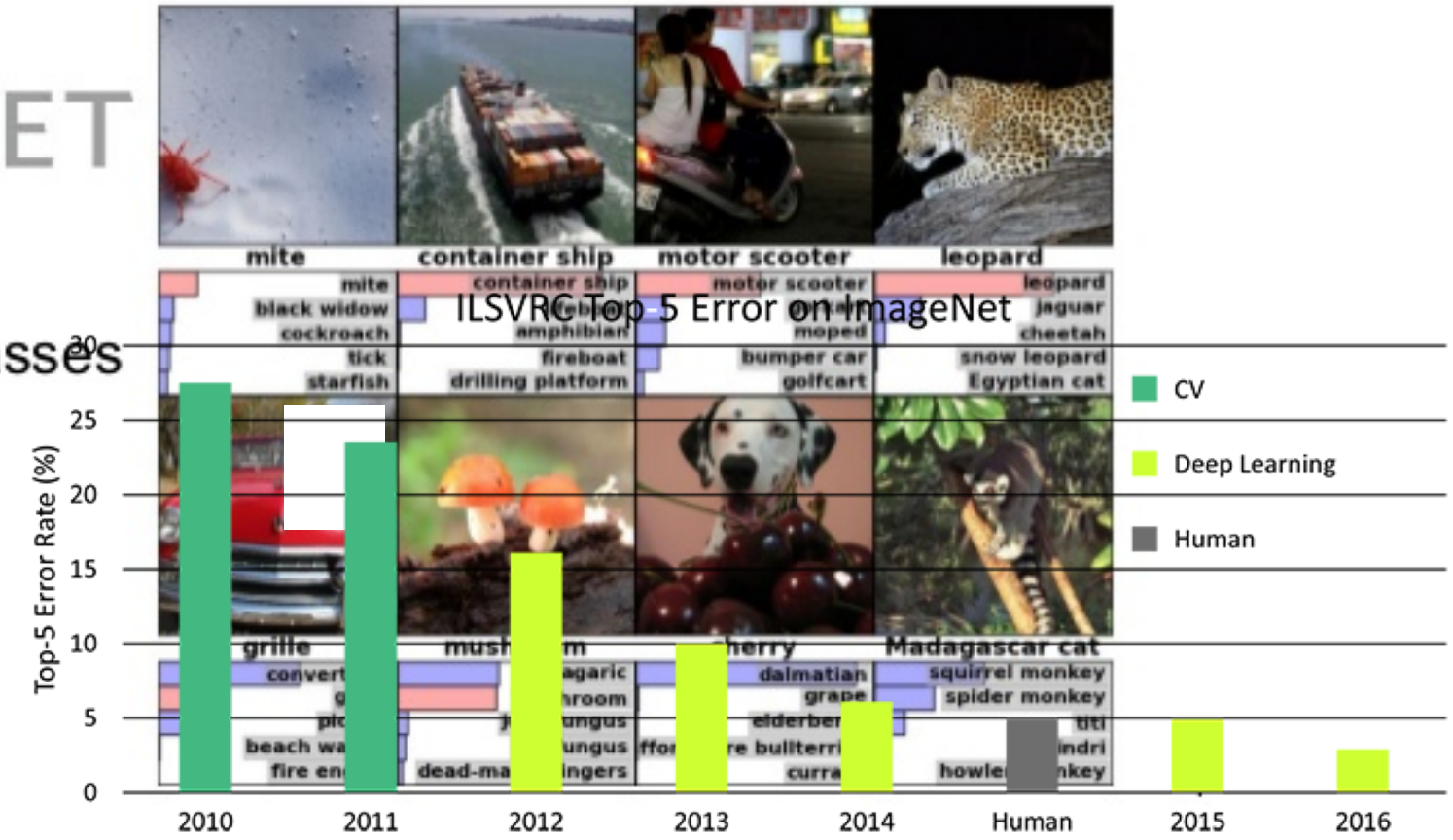
Some Key Players

Recent advances in ML: Deep Learning

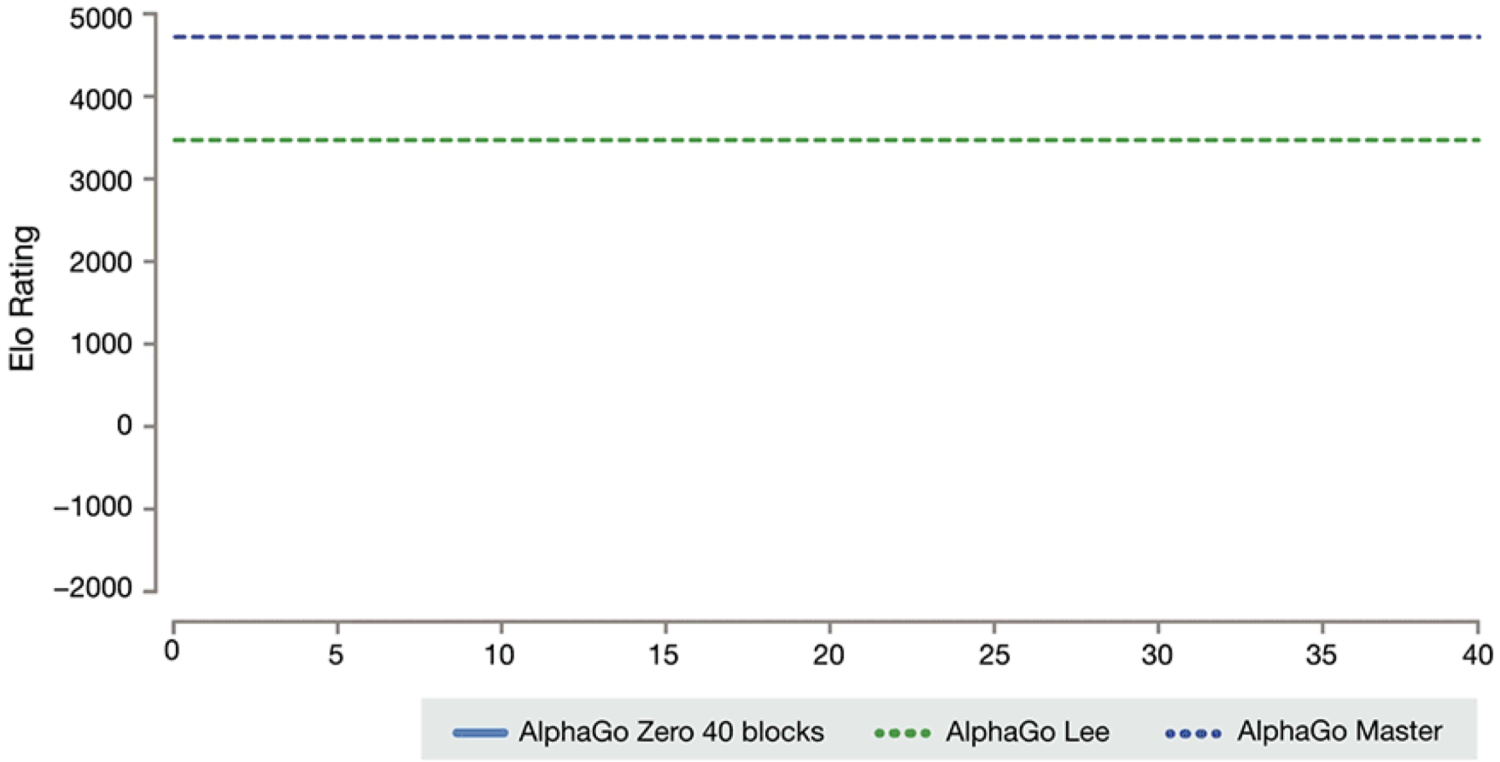
ImageNet Challenge

IMAGENET

- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.



Recent advances in ML: Deep Reinforcement Learning



SDSC

This Success Relies On…

1. Large dataset of labelled data
2. Good quality data
3. Enough computing power
4. Clear and measurable objectives

An Unexpected Outcome

Title: Universal adversarial perturbations
Authors: Moosavi-Dezfooli, Seyed-Mohsen; Fawzi, Alhussein; Fawzi, Omar; Frossard, Pascal
Publication: eprint arXiv:1610.08401
Publication Date: 10/2016

This is not a sock

- It's an Indian elephant!
- At least after adding a universal noise to the image
- Deep learning models do not mimic brain activity

A Disturbing Outcome

Turning a **STOP** sign into a **45 mph speed limit**

Structured vs Unstructured Data

	Tissue Cate	Cell ID	Total Cells	Tissue Cate	Cell Density	Cell X Posit	Cell Y Posit	Process Re	Distance fr	Distance fr	Category R	Distance fr	Nucleus An	Nucleus An	Nucleus Co	Nucleus Mi	Nucleus Ma	Nucleus Ax	Nucleus PC	Nucleus PD1 (FITC) Mean (Normalized Counts, Total Weighting)	Nucleus PC	Nucleus PC	Nucleus PC	Nucleus CE	Ni
86-1A_image 1.im3	No tissue	1				751	4	#N/A	#N/A		55	#N/A	101	0.03%	0.66	7.11	17.01	2.39	0.558		0.812	1.138	0.108	82.011	0.787	
86-1A_image 1.im3	Tumor	2				878	4	#N/A	#N/A		1	#N/A	57	0.01%	0.66	6.01	13	2.16	0.975		1.236	1.765	0.159	70.433	0	
86-1A_image 1.im3	Tumor	3				910	4	#N/A	#N/A		1	#N/A	61	0.01%	0.59	7.04	13	1.85	0.126		0.42	0.638	0.094	25.642	0	
86-1A_image 1.im3	Tumor	4				1002	5	#N/A	#N/A		1	#N/A	96	0.01%	0.65	10.02	14.02	1.4	0.281		0.48	0.711	0.083	46.074	0	
86-1A_image 1.im3	Stroma	5				1214	5	#N/A	#N/A		11	#N/A	82	0.03%	0.68	8.09	14.05	1.75	0.272		0.567	0.829	0.101	46.499	0	
86-1A_image 1.im3	Stroma	6				1306	5	#N/A	#N/A		13	#N/A	67	0.03%	0.73	9.05	10.97	1.21	0.26		0.484	0.68	0.095	32.437	0	
86-1A_image 1.im3	No tissue	7				175	6	#N/A	#N/A		54	#N/A	66	0.02%	0.68	9.03	11.01	1.22	0.989		1.267	1.596	0.141	83.595	0.468	
86-1A_image 1.im3	Stroma	8				789	6	#N/A	#N/A	#N/A		2	107	0.04%	0.76	12	12.01	1	0.267		0.641	1.522	0.196	68.541	0	
86-1A_image 1.im3	Tumor	9				1056	5	#N/A	#N/A		1	#N/A	115	0.01%	0.75	9	15	1.67	0.327		0.541	0.79	0.083	62.226	0	
86-1A_image 1.im3	Stroma	10				484	7	#N/A	#N/A		9	#N/A	86	0.03%	0.68	8.65	13.16	1.52	0.728		1.15	1.884	0.216	98.887	0	
86-1A_image 1.im3	Tumor	11				980	7	#N/A	#N/A		1	#N/A	139	0.02%	0.73	12.05	14.99	1.24	0.208		0.438	0.719	0.083	60.924	0	
86-1A_image 1.im3	Stroma	12				373	7	#N/A	#N/A		9	#N/A	65	0.02%	0.71	8.16	11.29	1.38	0.282		0.44	0.611	0.073	28.606	0	
86-1A_image 1.im3	No tissue	13				1252	8	#N/A	#N/A	#N/A		1	214	0.06%	0.43	13.1	27.29	2.08	0.248		0.672	1.053	0.135	143.823	0	
86-1A_image 1.im3	No tissue	14				295	8	#N/A	#N/A		54	#N/A	66	0.02%	0.72	8.08	10.9	1.35	1.087		1.464	1.882	0.199	96.636	0.847	
86-1A_image 1.im3	No tissue	15				774	8	#N/A	#N/A			2	131	0.03%	0.66	12.01	14.03	1.17	0.564		1.267	2.651	0.477	158.173	0	
86-1A_image 1.im3	Tumor	16				1118	10	#N/A	#N/A	#N/A		3	129	0.02%	0.45	8.74	20.1	2.3	0.321		0.599	1.036	0.116	77.253	0	
86-1A_image 1.im3	Tumor	17				836	11	#N/A	#N/A		1	#N/A	161	0.02%	0.44	11.77	20.61	1.75	0.304		0.529	0.809	0.094	85.12	0	
86-1A_image 1.im3	Tumor	18				1017	9	#N/A	#N/A		1	#N/A	241	0.03%	0.6	17.99	18.03	1	0.287		0.535	0.812	0.089	129.007	0	
86-1A_image 1.im3	Tumor	19				1035	10	#N/A	#N/A		1	#N/A	200	0.03%	0.54	13.78	19.36	1.41	0.33		0.956	2.096	0.315	191.28	0	
86-1A_image 1.im3	No tissue	20				1241	10	#N/A	#N/A		57	11	68	0.02%	0.48	9.41	12.51	1.33	0.42		0.73	1.257	0.141	49.66	0	
86-1A_image 1.im3	Tumor	21				805	12	#N/A	#N/A	#N/A		6	157	0.02%	0.33	9.02	26.08	2.89	0.239		0.501	0.944	0.127	78.634	0	
86-1A_image 1.im3	Tumor	22				940	14	#N/A	#N/A		1	#N/A	286	0.04%	0.37	14.29	31.73	2.22	0.217		0.49	0.854	0.091	140.025	0	
86-1A_image 1.im3	No tissue	23				1299	13	#N/A	#N/A	#N/A		1	105	0.03%	0.75	11.03	12.03	1.09	0.275		0.581	0.85	0.124	61.031	0	
86-1A_image 1.im3	Tumor	24				910	14	#N/A	#N/A		1	#N/A	130	0.02%	0.63	11.63	13.74	1.18	0.308		0.529	0.862	0.098	68.76	0	
86-1A_image 1.im3	Stroma	25				783	14	#N/A	#N/A		10	7	70	0.03%	0.55	8.45	12.38	1.46	0.383		0.754	1.569	0.262	52.812	0	
86-1A_image 1.im3	Tumor	26				972	17	#N/A	#N/A		1	#N/A	169	0.02%	0.57	11.59	20.18	1.74	0.291		0.525	0.793	0.1	88.741	0	
86-1A_image 1.im3	Tumor	27				1006	15	#N/A	#N/A		1	#N/A	119	0.01%	0.51	12.34	13.07	1.06	0.33		0.506	0.759	0.084	60.251	0	
86-1A_image 1.im3	Tumor	28				1073	13	#N/A	#N/A		1	7	271	0.03%	0.53	14.58	25.85	1.77	0.219		0.479	0.777	0.095	129.686	0	
86-1A_image 1.im3	Stroma	29				514	18	#N/A	#N/A		9	#N/A	179	0.07%	0.55	10.69	21.81	2.04	0.253		0.627	0.947	0.119	112.241	0.126	
86-1A_image 1.im3	Tumor	30				820	15	#N/A	#N/A		1	#N/A	178	0.02%	0.66	15.01	16.62	1.07	0.309		0.519	0.842	0.095	92.361	0	
86-1A_image 1.im3	Tumor	31				884	18	#N/A	#N/A		1	#N/A	489	0.06%	0.45	23.3	30.32	1.3	0.315		0.598	1.38	0.146	292.35	0	
86-1A_image 1.im3	Stroma	32				494	18	#N/A	#N/A		9	#N/A	107	0.04%	0.73	11.01	13	1.18	0.425		0.653	0.893	0.101	69.832	0.028	
86-1A_image 1.im3	Tumor	33				861	18	#N/A	#N/A		1	#N/A	218	0.03%	0.5	15.64	21.05	1.35	0.303		0.643	2.072	0.193	140.107	0	
86-1A_image 1.im3	Stroma	34				484	18	#N/A	#N/A		9	#N/A	105	0.04%	0.6	10.21	15.17	1.49	0.466		0.775	1.5	0.17	81.383	0.685	
86-1A_image 1.im3	Tumor	35				755	23	#N/A	#N/A	#N/A		4.1	117	0.01%	0.73	11.63	13.05	1.12	0.306		0.539	0.755	0.073	63.073	0	
86-1A_image 1.im3	No tissue	36				1380	22	#N/A	#N/A		57	#N/A	162	0.04%	0.34	9.37	28.07	2.99	0.312		0.645	1.072	0.159	104.44	0	
86-1A_image 1.im3	No tissue	37				1241	23	#N/A	#N/A		57	10	205	0.09%	0.56	14.01	20.03	1.43	0.737		1.696	3.679	0.547	347.687	0	
86-1A_image 1.im3	Stroma	38				426	24	#N/A	#N/A		9	#N/A	93	0.04%	0.81	10.04	11.03	1.1	0.473		0.766	1.202	0.136	71.217	0	
86-1A_image 1.im3	Tumor	39				935	25	#N/A	#N/A		1	#N/A	124	0.02%	0.65	12.38	14.3	1.16	0.271		0.477	0.727	0.087	59.176	0	
86-1A_image 1.im3	Tumor	40				992	25	#N/A	#N/A		1	#N/A	260	0.04%	0.56	15.03	26.01	1.73	0.321		0.611	1.525	0.129	171.192	0	
86-1A_image 1.im3	No tissue	41				1256	25	#N/A	#N/A		57	7.2	97	0.02%	0.69	10.01	13	1.3	0.717		1.036	1.511	0.145	102.466	0	
86-1A_image 1.im3	Stroma	42				360	26	#N/A	#N/A		9	19.6	122	0.05%	0.67	10.4	15.32	1.47	0.41		0.603	0.959	0.107	73.557	0.034	
86-1A_image 1.im3	Stroma	43				792	26	#N/A	#N/A		10	8	123	0.05%	0.53	10.2	16.55	1.62	0.249		0.55	1.102	0.115	67.647	0	
86-1A_image 1.im3	Tumor	44				832	26	#N/A	#N/A		1	#N/A	128	0.02%	0.53	11.89	15.82	1.33	0.392		0.79	3.708	0.511	101.163	0	
86-1A_image 1.im3	Tumor	45				1170	25	#N/A	#N/A	#N/A		2	84	0.01%	0.81	9.03	11.01	1.22	0.165		0.311	0.488	0.073	26.088	0	
86-1A_image 1.im3	Tumor	46				733	27	#N/A	#N/A		1	19.4	221	0.03%	0.66	14.91	20.36	1.37	0.288		0.542	0.842	0.113	119.89	0	
86-1A_image 1.im3	Tumor	47				867	28	#N/A	#N/A		1	#N/A	59	0.01%	0.46	7.03	13.38	1.9	0.367		0.571	0.837	0.088	33.716	0	
86-1A_image 1.im3	Tumor	48				1012	27	#N/A	#N/A		1	#N/A	192	0.02%	0.45	15.22	20.2	1.33	0.305		0.561	0.901	0.094	96.268	0	
86-1A_image 1.im3	Tumor	49				685	33	#N/A	#N/A	#N/A		2.8	57	0.01%	0.91	8.02	8.02	1	0.23		0.336	0.443	0.056	19.143	0	

Structured

Unstructured

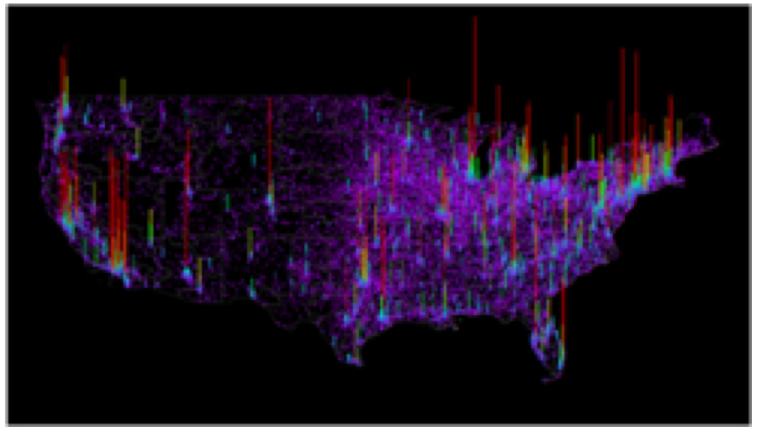
Semi-structured

SDSC

The Structured World

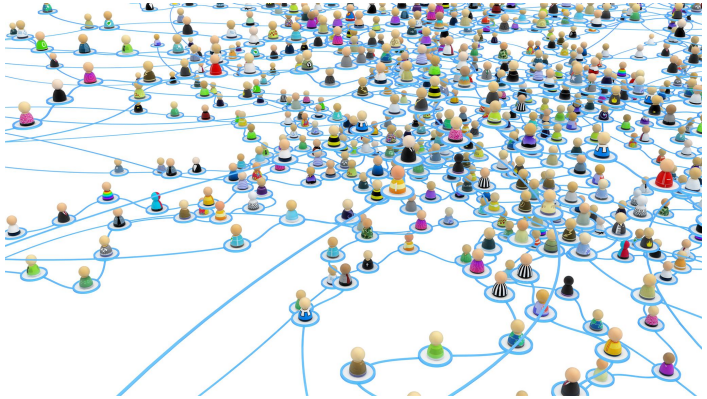
Health care

Predictive maintenance



Energy networks

Financial time series



Social networks

SDSC

[csr.lanl.gov/sim/]

A Sobering View of Data Science

Obstacles to a Wider Adoption in a Structured World

1. ~~Large dataset of labelled data~~ -> **Labelling is expensive**

2. ~~Good quality data~~ -> **Data is usually missing/Increased uncertainty**

3. ~~Clear and measurable objectives~~ -> **Knowledge discovery/causality**

4. **Lack of interpretability/lack of trust**

Upcoming Challenges for Structured Data

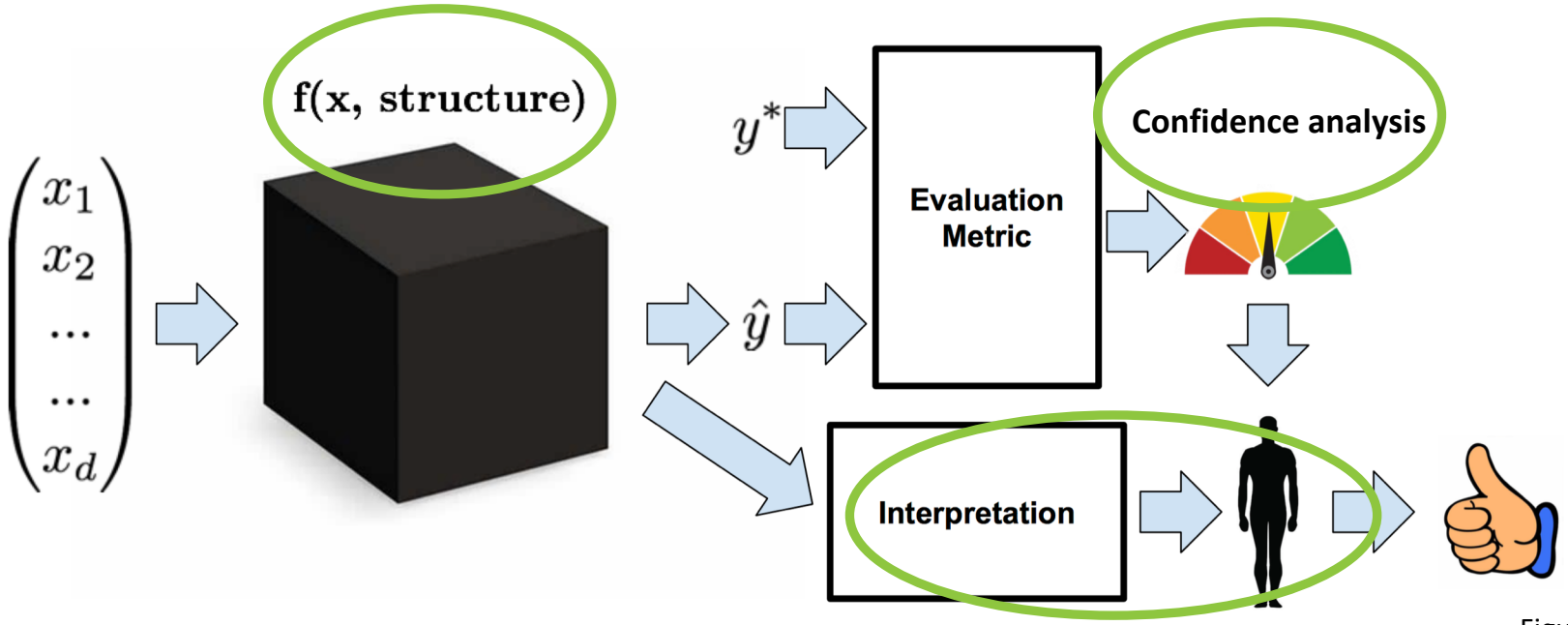


Figure adapted from Z. Lipton

1. Incorporating structure knowledge in the model for data curation
2. Dealing with uncertainties
3. Promoting causality/interpretability
4. Focusing on unsupervised/semi-supervised learning

Interpretable Machine Learning – Use Case

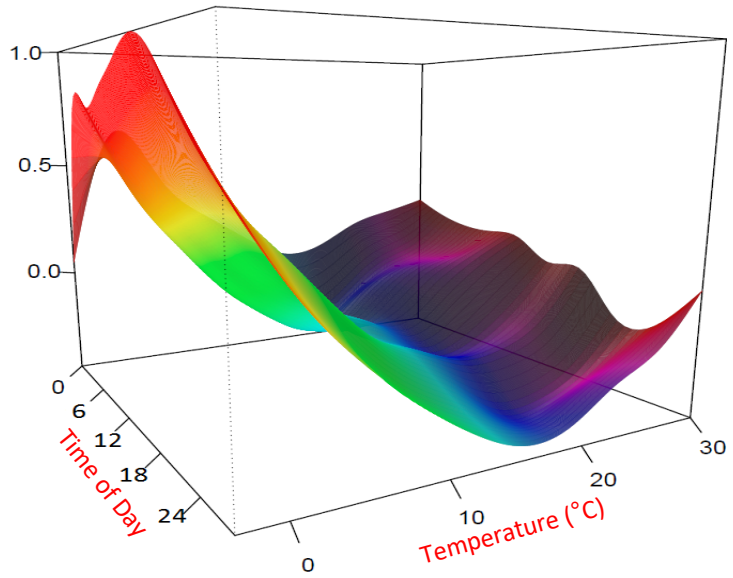
- Forecasting demand in electricity (France)

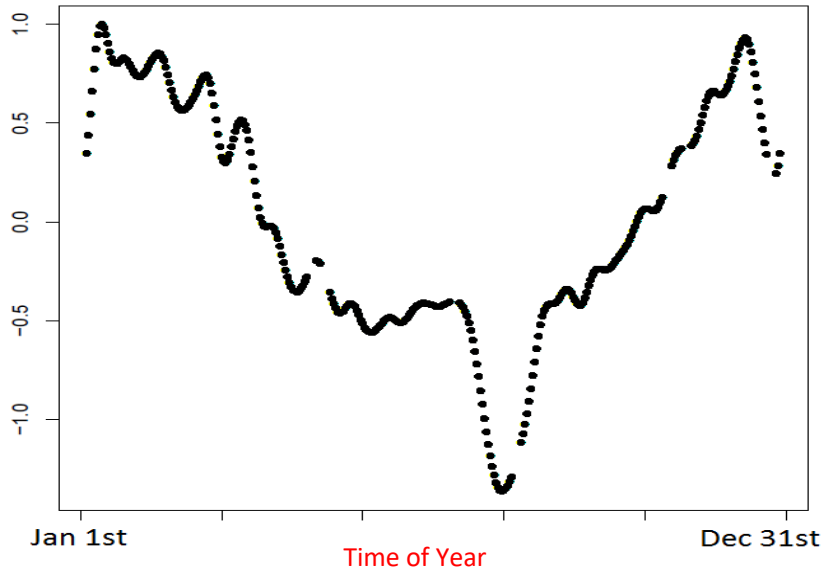
Trend Lag load Day-type specific daily pattern

$$y_k = \beta^{\text{Intercept}} + f^{\text{Trend}}(k) + f^{\text{LagLoad}}(y_{k-48}) + \sum_{l=1}^{6} \mathbf{1}(x_k^{\text{DayType}} = l)(\beta_l^{\text{DayType}} + f_l^{\text{TimeOfDay}}(x_k))$$
$$+ f^{\text{CloudCover}}(x_k) + f^{\text{Temperature/TimeOfDay}}(x_k) + f^{\text{LagTemperature}}(x_{k-48})$$
$$+ f^{\text{TimeOfYear}}(x_k) + x_k^{\text{LoadDecrease}} f^{\text{LoadDecrease}}(x_k) + \epsilon_k.$$

Lag temperature (accounting for thermal inertia)

Transfer functions learned from data:





© IBM Research, EDF R&D

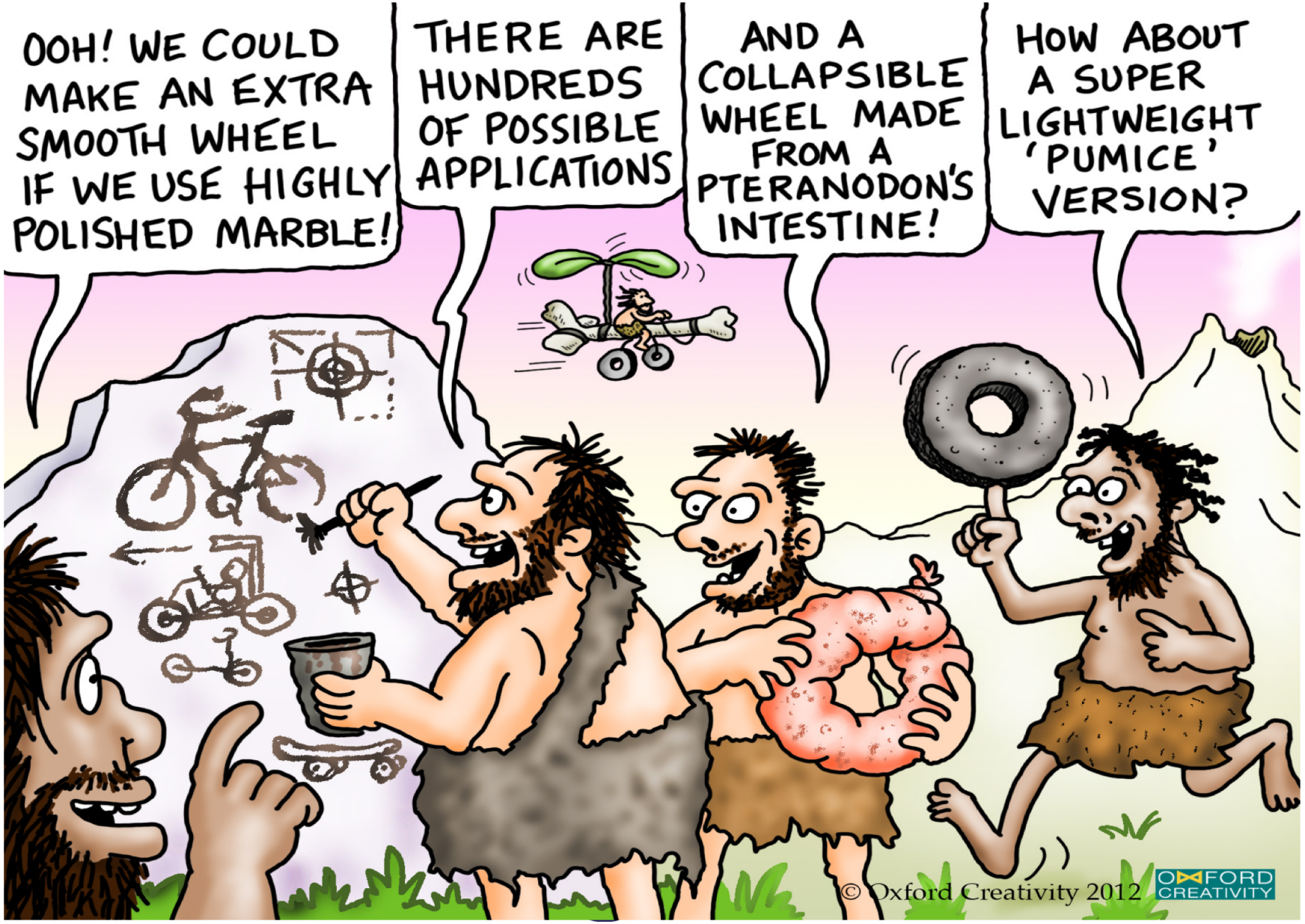
Other Roadblocks in Data Science Ventures

credit: oxford creativity, https://www.triz.co.uk/

Facilitate communication to foster innovation

credit: oxford creativity, https://www.triz.co.uk/

Foster multidisciplinary collaborations



credit: oxford creativity, https://www.triz.co.uk/

Swiss Data Science Center (SDSC)

Multi-disciplinary team of data scientists, computer and data scientists, and domain experts

How can I best match the right drug with the right dosage to the right patient at the right time?

ETH zürich + ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

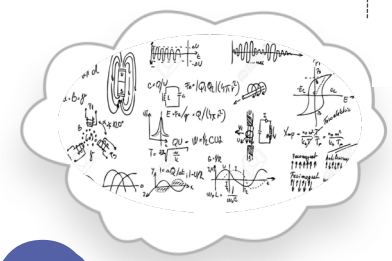
Domain experts

What is the hyperplane that best separates two classes of points in multidimensional space?

Data scientists

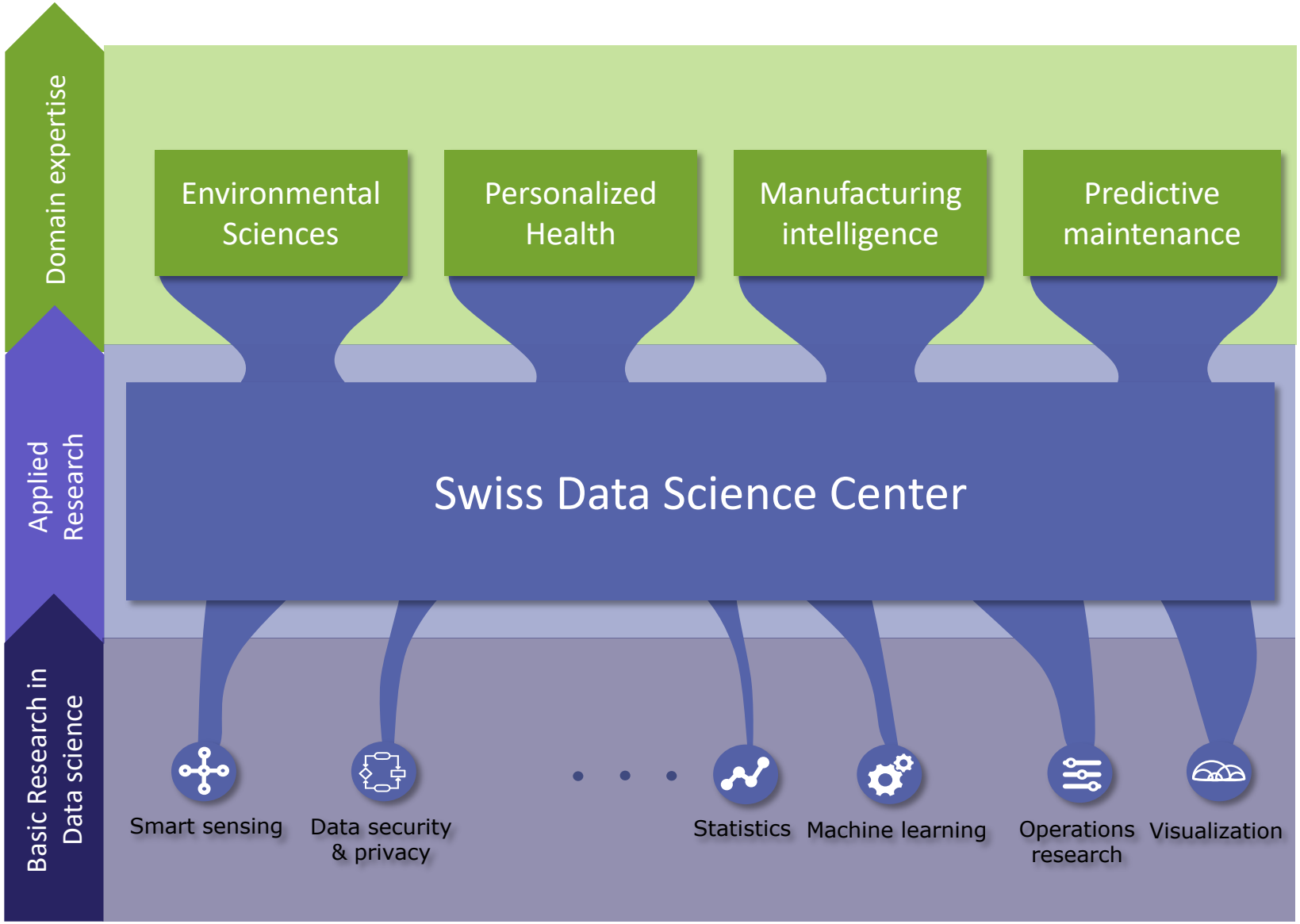
Data providers

How is my data protected?
How private is it?
How exactly is it used?

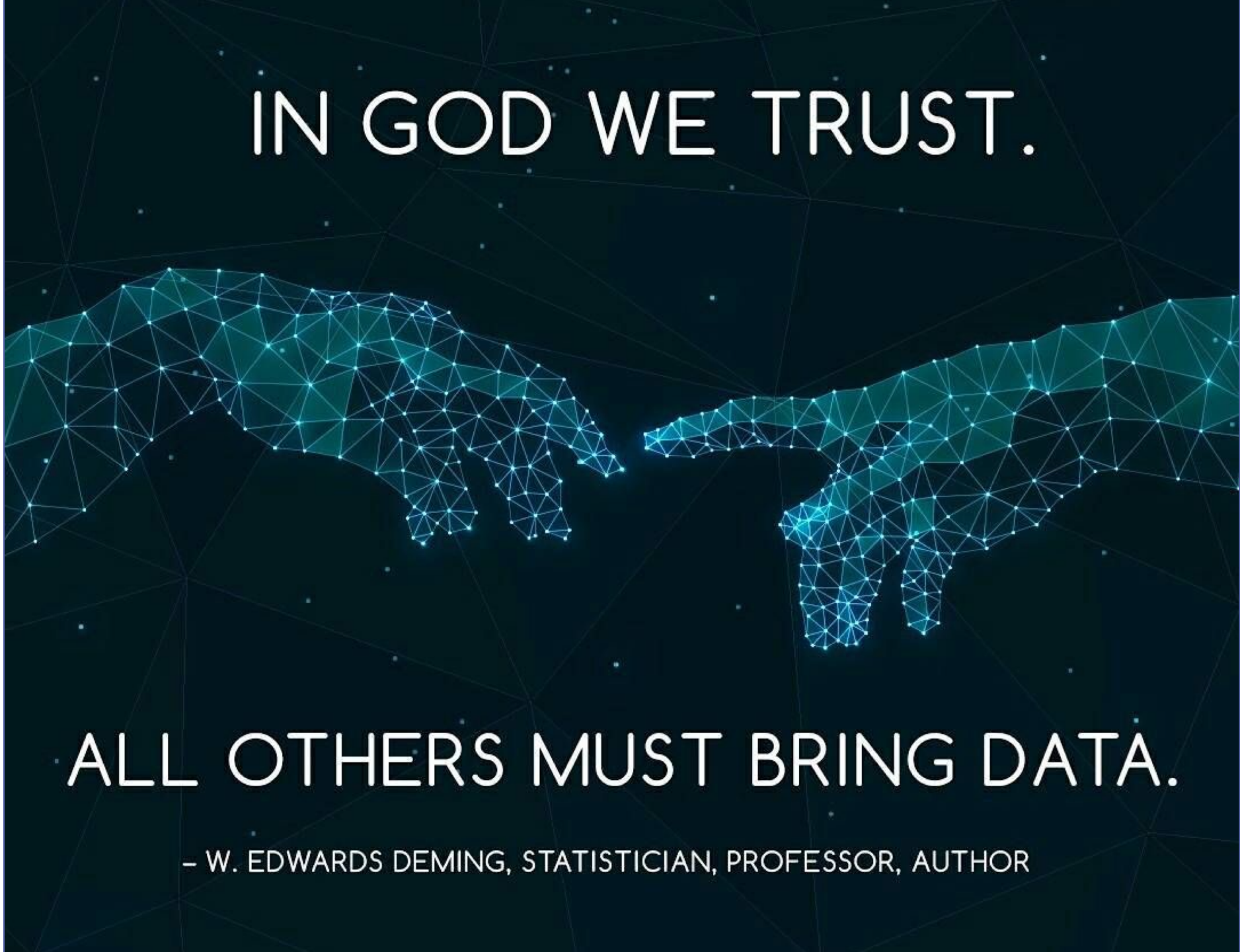


SDSC

Key Actor in a Complex Ecosystem



SDSC


IN GOD WE TRUST.
ALL OTHERS MUST BRING DATA.
– W. EDWARDS DEMING, STATISTICIAN, PROFESSOR, AUTHOR