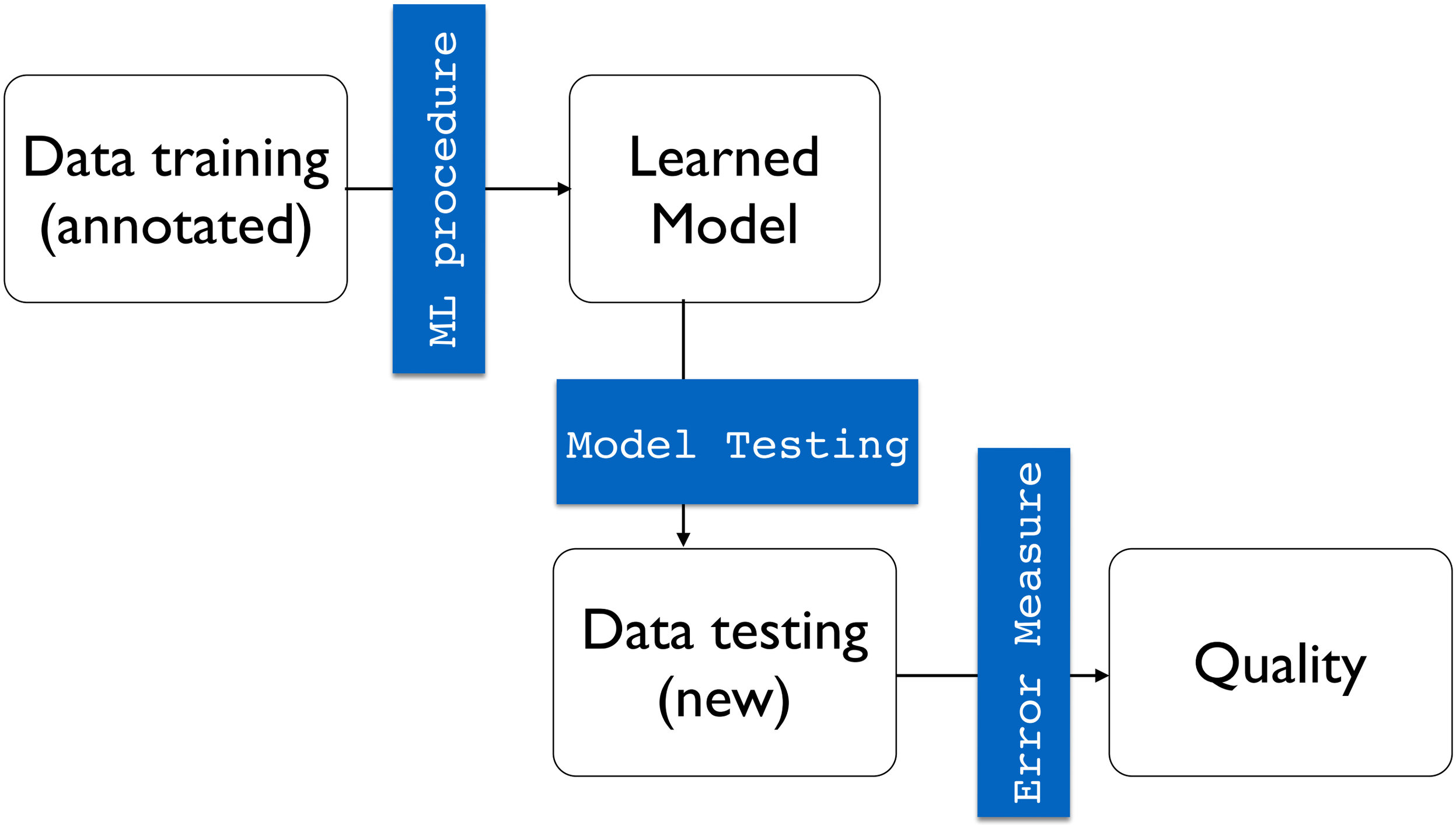
Intro to Machine Learning

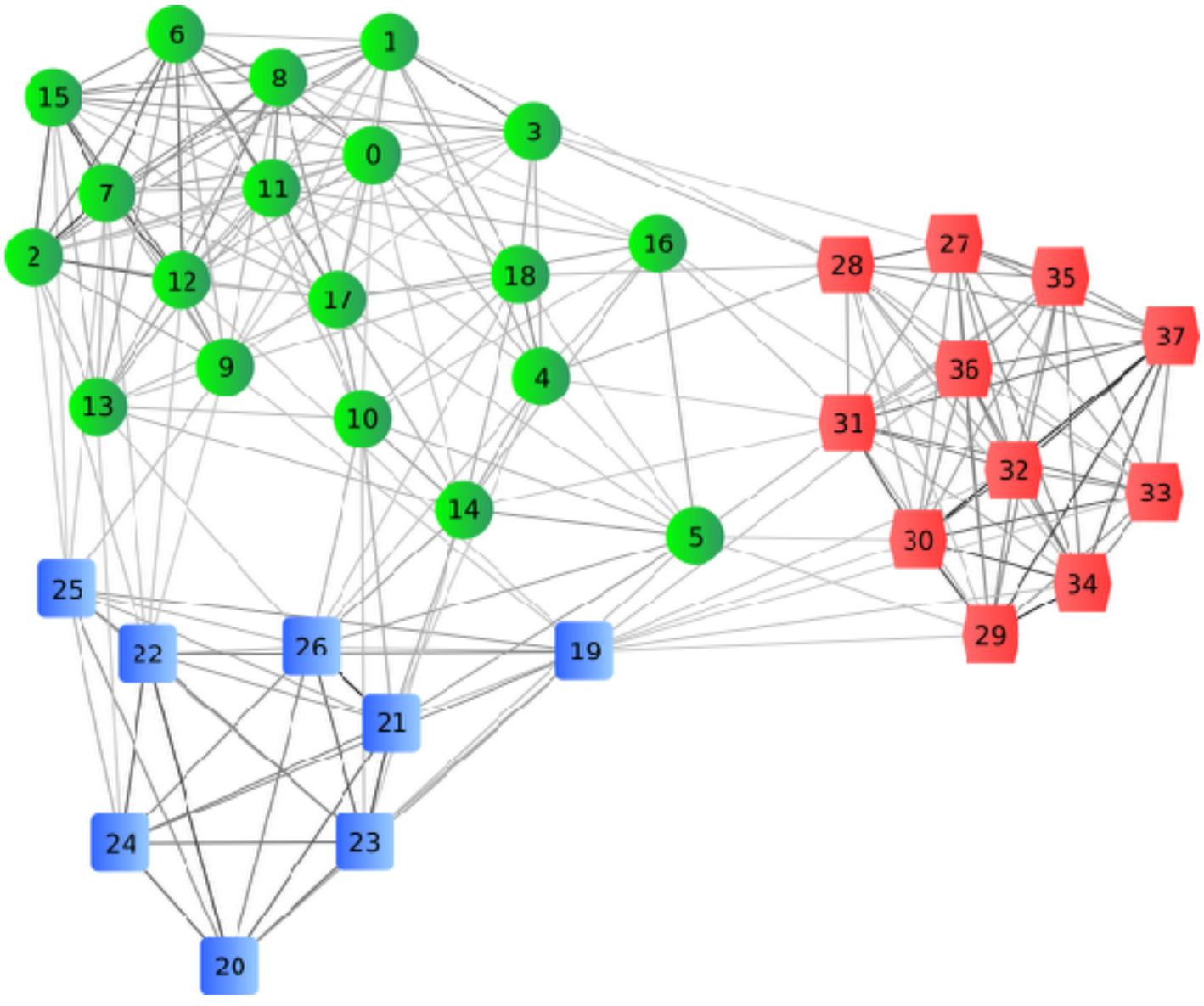
Module 1 Objectives / Intro, Evaluation

- Intro to Machine Learning - what is learning ?
- Data - Matrix type
  - algebraic notations
- Heuristics and Quantitative rules
- Error measurement
  - training VS testing error, Cross Validation
  - overfitting

What is machine learning ? Supervised learning

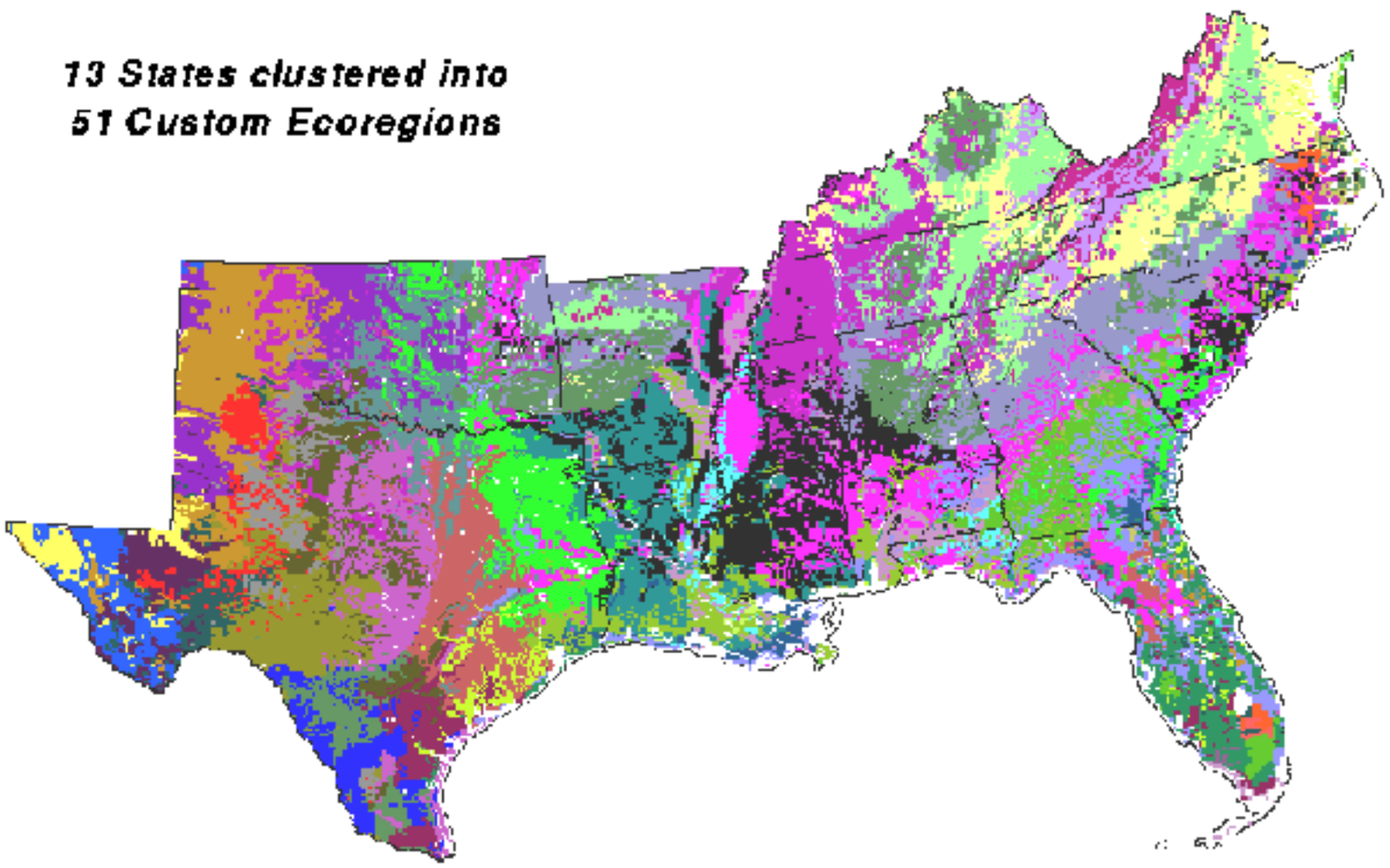


What is machine learning ? Graph learning



- data defined by links or analogies or connections
- for example social networks, or web links
- task: identify object properties from links
- taks: detect graph patterns

What is machine learning ? Clustering

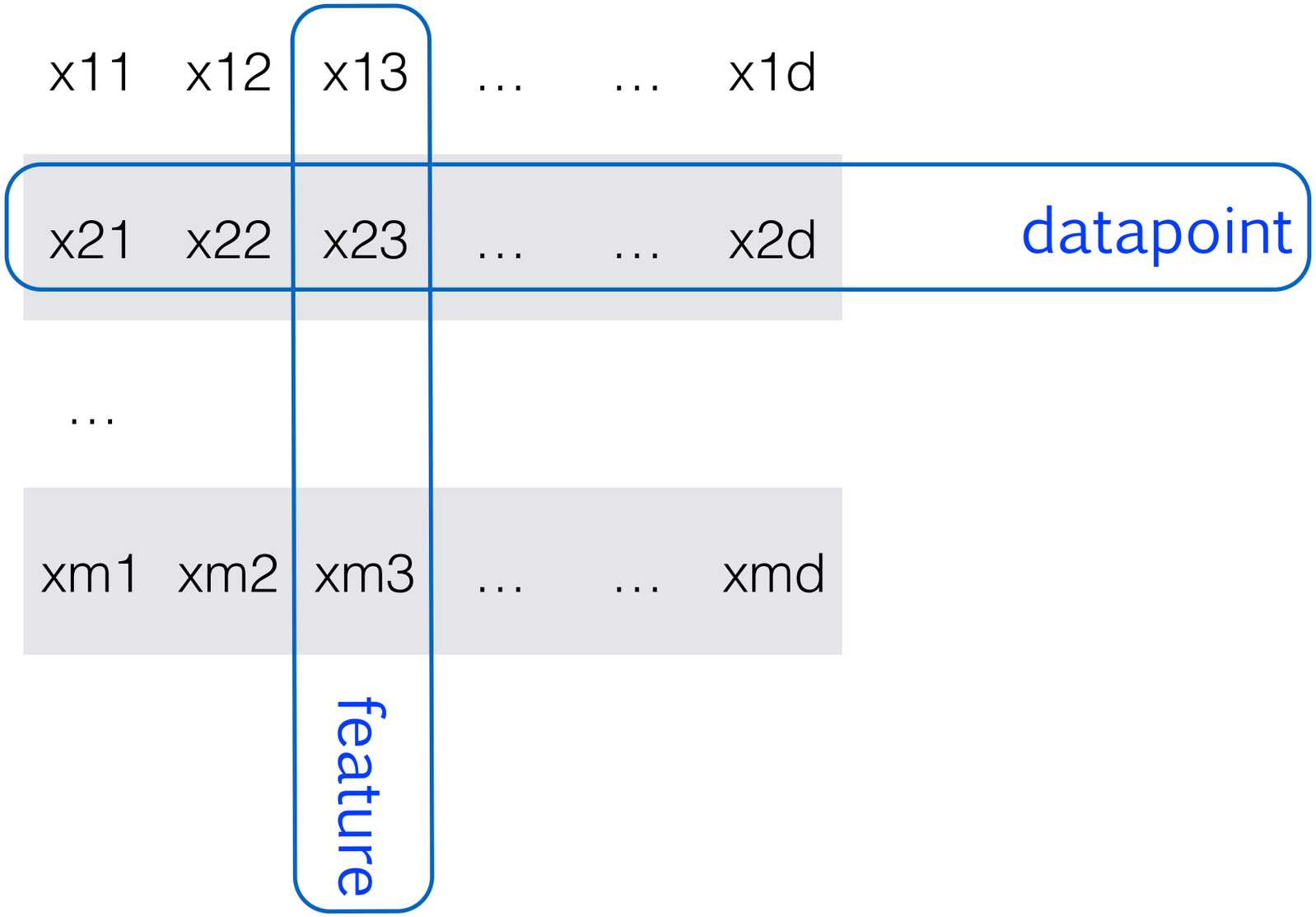


- data given without labels
- task: group similar data points

What is machine learning ? Time series analysis

- data that evolves with time
- like stocks or patient records
- task: predict future behavior
- task: detect anomalies

Matrix data



- m datapoints/objects X=(x1,x2,…,xd)
- d features/columns f1, f2, …, fd

Heuristic rules / decisional

- If fever>100, patient has flu
- If email contains words “free” or “porn”, it is spam
- If a web page contains ngram “Michael Jackson”, it is relevant to the user
- If age<22 and sex=F and highschool_diploma=Yes, then eligible for application
- If income_per_capita<$1000, region prone to civil war
- If romantic=Yes and comedy=Yes and Orlando_Bloom=Yes, then movie success among females aged 20-40
- If Nasdaq_Computer_Index=Gain and Apple announces new Ipad, then AAPL_Stock=Buy

Heuristic rules / quantitative

- if 3*exam_grade+2*HW_grade>55, then student can pass
- if blood_pressure/log(age)>3, recommend medicine
- if rent+food+bills<1/2 salary, loan for 1/2 salary possible

Matrix data / training VS testing

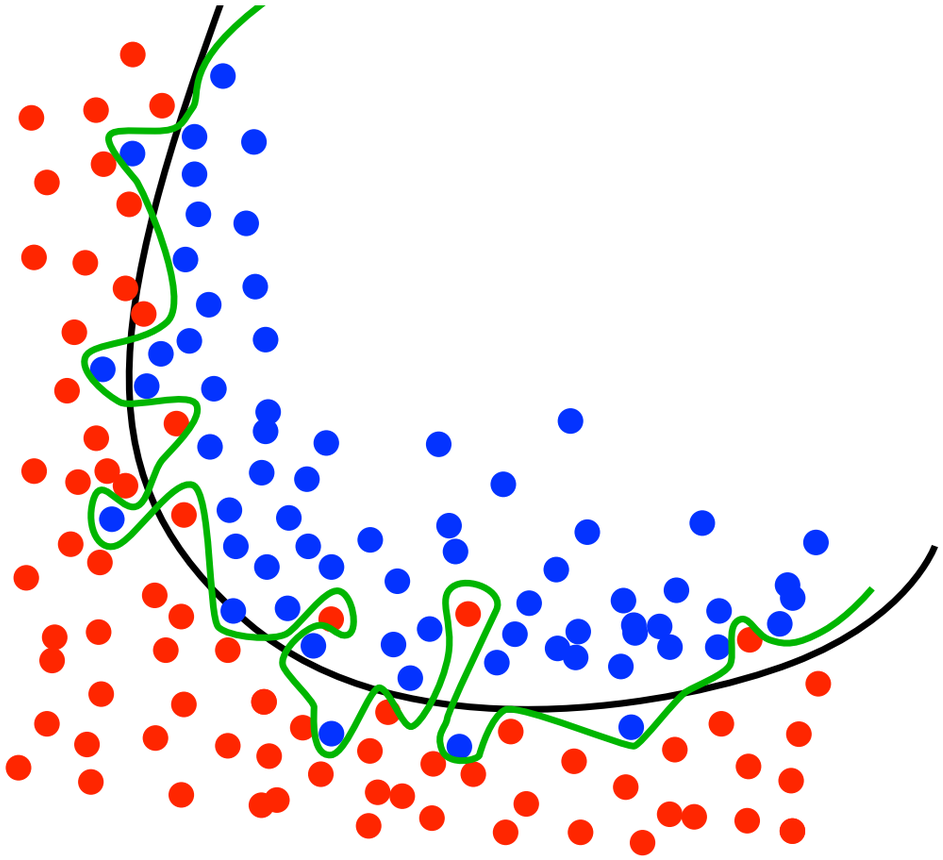
	AUT	BEL	BUL	CYP	CZE	DEN	EST	FIN	FRA	GER	GRE	HUN	IRL	ITA	LAT	LTU	LUX	MLT	NED	POL	POR	ROM	SVK	SLO	ESP	SWE	GBR
T-01	64.4	135.0	44.7	7.0	124.1	51.3	14.9	66.6	363.5	837.4	98.2	66.8	42.8	446.6	6.6	11.6	8.4	2.1	174.9	360.8	64.8	90.7	26.9	16.1	304.9	49.8	578.2
T-02	7.1	7.8	10.3	0.7	21.0	5.0	1.9	4.5	33.9	47.0	9.3	7.6	13.1	29.6	1.8	3.3	0.3	0.3	16.7	38.3	17.4	25.7	4.2	2.1	37.3	5.6	48.5
T-03	5.3	11.0	4.4	0.7	8.0	7.0	0.8	6.9	72.3	66.3	9.1	9.7	8.6	40.5	1.5	5.0	0.4	0.3	17.6	31.1	8.1	16.8	3.7	1.3	29.8	7.7	39.6
T-04	11.8	14.1	9.0	1.0	19	14	1.6	1.0	1,80.1	71.8	12.8	20.9	17.4	36.1	3	4.1	0.6	0.5	26.5	26.1	12.9	57.0	3.0	12.4	24.4	29.6	35.1
T-05	91.2	1,45.4	33.7	9.1	39.1	30.6	8	86.4	10,05.8	9,35.3	1,15.2	51.6	43.1	5,26.7	1.9	7.9	0.3	4.2	1,35.4	2,75.0	36.1	4	17.5	9.5	5,01.1	77.7	9,22.1
T-06	4.7	4.3	1.4	1.6	8.5	2.2	0.6	2.1	1,35.4	4,74.0	21.0	20.1	9.5	45.0	0.8	1.1	0.4	0.8	33.7	8.4	1.9	0	1.7	1.5	27.2	14.2	1,14.3
T-07	64.4	1,23.0	44.7	7.0	1,24.1	51.3	14.9	56.6	3,63.5	8,37.4	9.2.2	56.8	42.8	4,46.6	6.5	11.6	8.4	2.1	1,74.8	3,06.8	64.8	90.7	26.9	15.1	3,04.9	49.8	5,78.2
T-08	7.8.2	1,12.6	4.8.0	8.2	7.7.9	9.6.8	12.0	5.7.8	8,7.3.8	6,17.4	1,04.5	84.6	1,84.3	3,72.1	19.2	40.6	3.8	3.8	1,81.7	3,48.8	83.4	2,02.8	32.2	23.2	4,47.6	55.7	4,48.8
T-09	22.6	13.6	8.4.8	2.6	29.1	13.7	6.3	24.4	1,41.0	1,39.8	32.8	20.4	17.8	1,95.3	7.6	15.4	3	1.2	98.4	1,22.1	64.7	74.0	21.1	6.8	1,29.6	21.8	2,30.6
T-10	30.2	1,04.6	76.4	9.6	1,03.3	54.6	12.4	53.0	6,41.0	8,23.1	1,11.5	1,00.6	43.0	5,92.1	2.3	30.7	4.7	4.1	1,63.9	3,81.3	1,06.2	2,14.4	53.9	23.2	4,51.2	91.5	6,05.9
T-11	30.5	1.1	11.2	0.8	12.5	10.9	8.9	29.7	6.1.8	1,19.8	4.3	8.3	1.6	33.8	8.7	5.9	0.4	0.1	9.5	23.2	4.7	5.8	11.0	1.5	49.6	31.6	25.1
T-12	50.1	49.7	31.4	4.4.8	37.3	35.4	35.2	44.8	49.1	54.6	34.8	28.0	38.5	58.1	29.7	38.4	65.9	52.5	42.3	31.4	57.2	14.9	22.2	45.6	45.4	45.6	46.1
T-13	29.2	6.4.1	13.1	5.3	36.8	17.1	6.0	22.0	1,97.9	2,65.0	43.6	13.2	16.2	1,28.1	4.7	6.5	6.2	1.9	94.7	44.5	33.2	21.2	7.4	6.3	1,57.5	36.2	1,82.7
T-14	15.2	8.2.4	3.7.4	4.5	5.8.8	36.4	6.8	8.2.8	4.8.4	5.2.6	5.3.3	3.7.1	2.3.2	3.0.8	6.3	9.4	6.1	2.1	10.2.4	13.4.1	4.6.3	4.9.6	2.8.5	1.3.7	2.4.1.8	1.3.7.8	3.4.5.2
T-15	9.00	17.26	3.47	0.01	9.80	4.52	1.44	4.86	45.41	102.00	2.34	14.48	4.20	80.61	1.91	2.92	1.58	0.00	51.20	15.87	4.30	18.00	6.00	1.10	27.01	0.38	98.47
T-16	3.00	3.10	7.40	0.00	19.40	5.50	0.00	0.20	13.10	82.40	8.81	2.90	0.00	17.40	0.00	0.20	3.10	0.00	7.50	58.40	3.70	7.60	3.90	0.00	12.30	2.20	43.80
T-17	38.9	69.2	9.8	0.8	36.8	38.5	0.8	7.7	10,97.8	2,48.1	6.8.9	7.7.0	2.3.3	16,83.0	5.3	6.0	5.8	3.0	1,46.2	8.5.0	1,24.6	27.0	26.0	6.5.0	3,40.2	17.8	3,31.3
T-18	22.7	28.9	15.7	2.3	31.6	31.7	4.2	29.7	4,17.8	2,61.2	42.0	32.5	57.3	1,68.1	6.4	16.2	0.9	0.1	40.9	1,55.7	22.8	32.7	12.0	7.2	2,18.5	28.7	1,90.9
T-19	3.5	5.8	2.3	3.2	3.8	3.6	3.3	6.4	6.4	6.1	2.5	2.4	3.9	3.0	1.5	2.0	3.4	2.1	1.8	2.3	2.5	1.6	3.2	3.3	3.1	5.4	1.0
T-20	6.9	7.7	3.3	5.3	5.4	6.2	4.5	15.5	6.8	6.3	4.5	3.2	5.9	2.3	1.7	1.3	13.5	1.3	6.4	1.5	1.8	1.1	2.0	2.4	2.3	3.7	2.5
T-21	0.46	3.43	0.19	0.00	0.43	1.01	0.08	0.19	0.09	1.82	0.47	0.23	0.45	1.00	0.04	0.21	0.00	0.00	1.51	0.28	0.22	0.27	0.27	0.17	0.44	0.27	2.76
T-22	2.9	3.8	4.9	1.0.0	7.6	8.3	10.0	3.9	8	6.2	9.5	6.0	9.6	7.9	2.9	1.7	6.7	10.0	9.0	9.8	6.5	6.3	3.0	3.5	6.0	4	7.4
T-23	13.3	17.8	7	1.3	4.1	11.1	8	12.8	78.8	73.2	10.3	5.2	16.4	32.5	1.1	1.0	2.8	1.5	22.7	7.2	9.6	2.0	2	1.3	5.1.8	23.4	96.1
T-24	80.4	33.4	6.5	1.9.2	47.1	1,03.4	5.6	70.8	5,24.8	9,07.9	9.4.5	27.4	4,26.7	5,61.2	1.0.3	5.1	8.5	1.3.7	2,61.3	3.5.5	1,01.4	17.1	7.1	7.6	4,08.6	9.0.2	9,36.0
T-25	13.0	10.3	7	0.0.0	5.3	7.8	7	9.7	8.6.0	1,07.0	8.0	4.6	19.7	39.8	7	1.0	7.4	2.2	42.9	6.8	13.8	3.6	5	1.3	47.3	12.8	9.7.7
T-26	0.13	0.19	0.10	0.02	0.57	0.12	0.06	0.10	0.32	0.31	0.10	0.22	0.17	0.35	0.13	0.14	0.00	0.00	0.28	0.35	0.17	0.00	0.22	0.20	0.15	0.17	0.27
T-27	63.9	46.4	46.3	7.3.9	28.0	73.7	43.5	46.8	54.3	69.1	43.8	45.9	74.0	54.2	31.0	37.6	7.3.6	5.1.1	82.4	24.5	44.5	38.2	26.9	42.3	51.7	48.2	38.4
T-28	4.6	1.7	4	0	4	8	0	4.7	12.9	1.7	6	4	8.7	4.7	0	1	0	0	7.9	3.1	1.5	7	9.6	1.6	8.5	3.1	6.2
T-29	32.1	62.6	1,00.4	3,71.1	1,28.9	86.3	1,25.4	19.7	16.2	1,14.0	2,24.7	57.6	2,42.1	1,47.3	38.2	1.2.8	1,70.7	2,19.6	1,57.5	1,50.1	1,20.7	74.4	7.9.8	85.1	1,24.8	4.1	1,17.0
T-30	34.7	33.0	19.7	2.8.3	22.0	37.1	20.3	33.6	3.1.2	3.1.9	2.4.0	17.5	44.5	30.3	1.8.0	1.5.8	7.1.4	2.1.3	32.1	14.4	19.8	9.1	18.8	23.4	27.4	12.2	31.5
T-31	0.0	0.0	2.0.3	4.8	0.6	7.0	0.6	0.2	3.0.5	0.1	1.8.3	1.3	0.5	7.5.4	1.0	0.1	0.1	0.1	0.2	6.9.5	11.1.8	1.6	0.1	0.7	9.2.4	1.3	0.2
T-32	2.1.7	2.0.1	3.4.1	1.3.3	3.4.3	2.1.6	2.4.1	2.8.3	1.5.4	2.5.3	0.0	2.9.6	2.5.7	2.2.6	1.8.3	3.0.1	9.7	2.0.2	2.0.8	2.9.5	2.0.9	3.6.1	3.2.5	2.9.7	2.1.1	2.5.4	1.8.4
T-33	13.4	11.7	3.4	0.8	12.7	7.2	1.3	5.1	9.5.1	2.3.1	10.7	7.0	1.6	9.5.0	2.8	6.2	0.5	0.2	10.5	2.4.9	3.9	9.8	3.7	2.0	6.5.7	1.6.7	3.7.2
T-34	9.3	3.7.0	6.3	0.0	8.1	7.6	0.9	1.2.8	5.6.3	1.2.2.7	2.1.3	8.4	3.1	1.0.0.6	0.0	0.2	0.0	0.0	8.4.7	1.8.5	1.3.6	1.4.9	6.2	0.6	6.0.3	1.9.8	8.6.0
T-35	1.0	3.2	0.5	0.1	1.4	0.3	7.3	2.5	7.6	2.2.0	0.3	1.4	0.7	6.1	0.2	0.7	0.1	0.0	1.9	1.6	2.3	2.2	0.4	0.1	3.1	1.2	8.0

Training

Testing

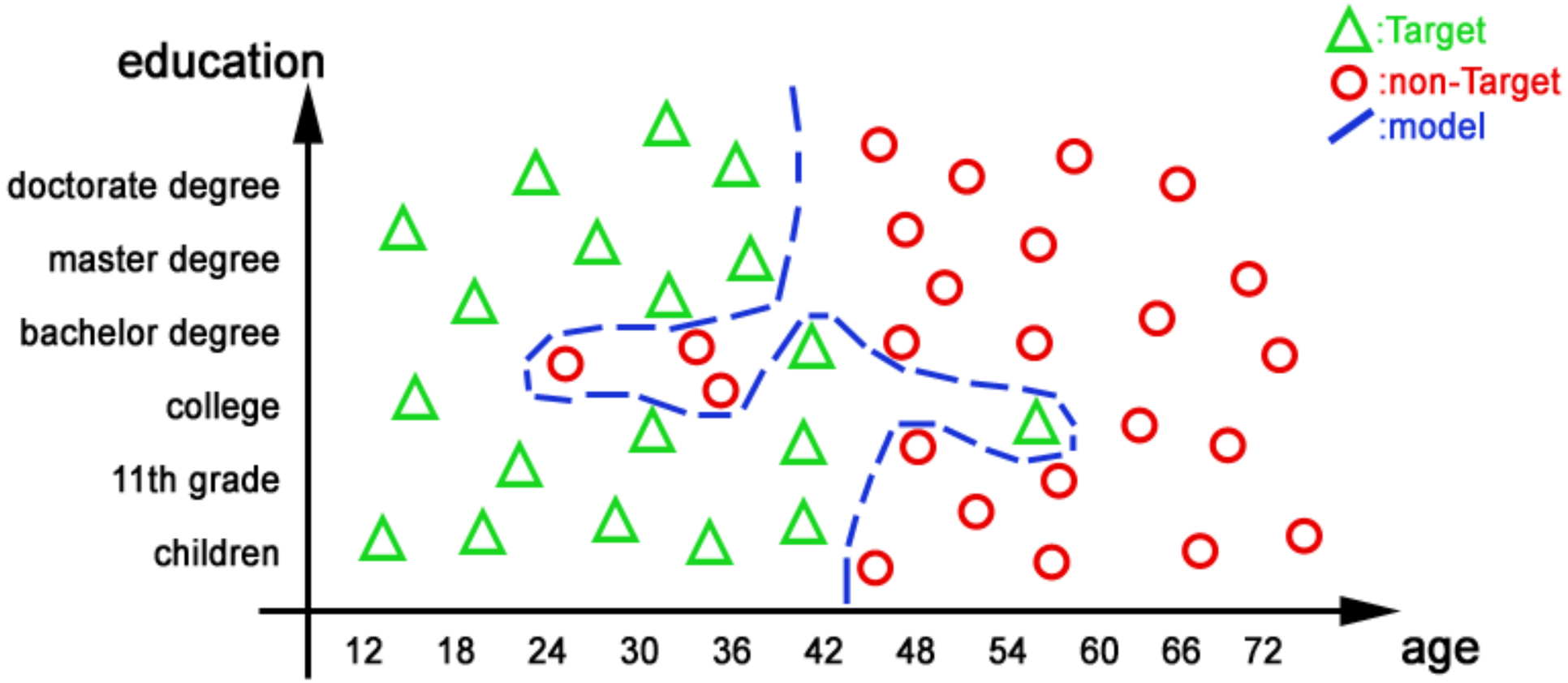
- testing set has to be independent of training set
  - or else testing result is inconclusive
  - and not reliable
- usually the data is partitioned before running any ML algorithm

Overfitting



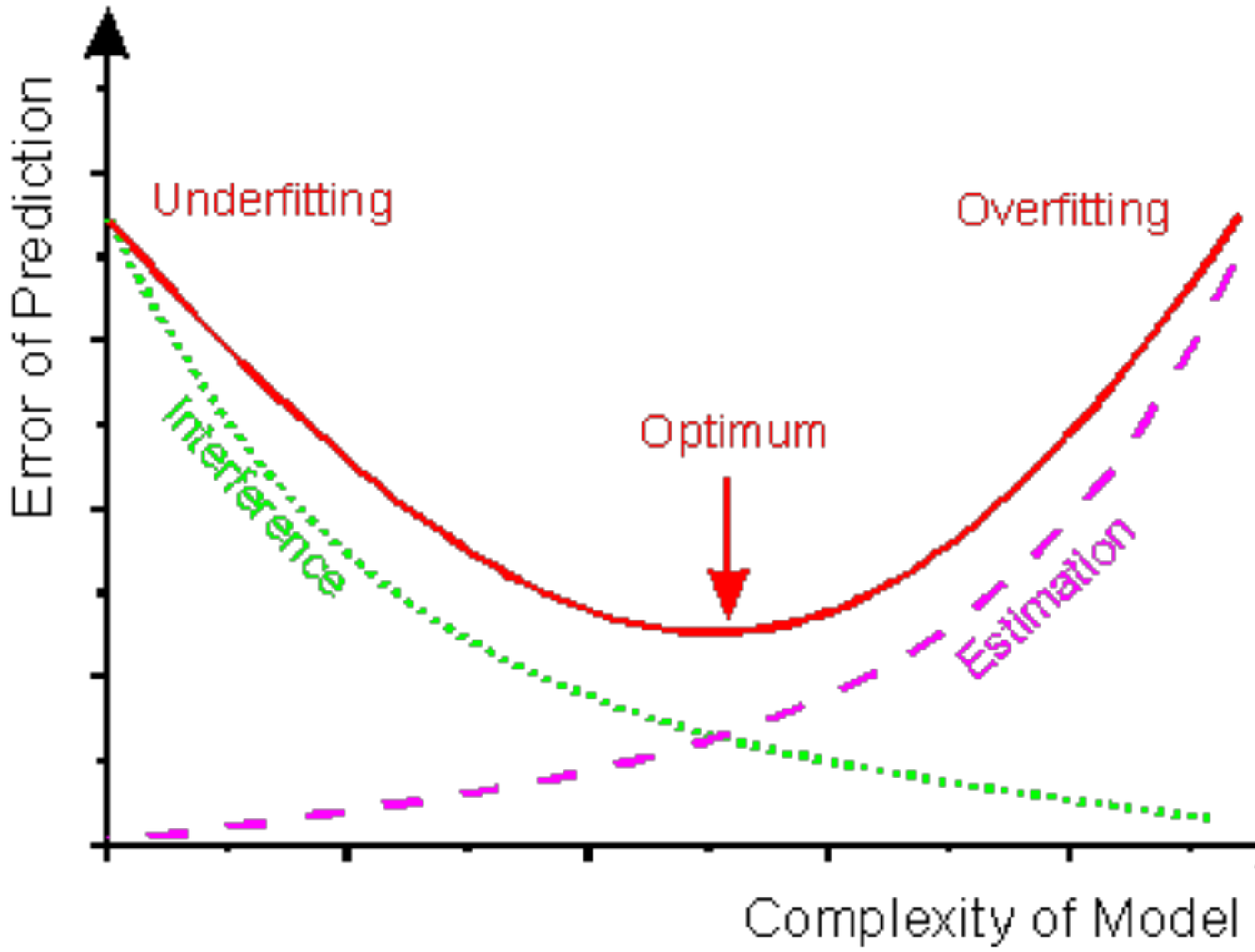
- might be capable to create a model that essentially memorizes all training dataset
  - for example a decision tree deep enough
- that is not useful : the purpose of the learning model is to applicable to new data (testing)

Overfitting



- might be capable to create a model that essentially memorizes all training dataset
  - for example a decision tree deep enough
- that is not useful : the purpose of the learning model is to applicable to new data (testing)

Overfitting



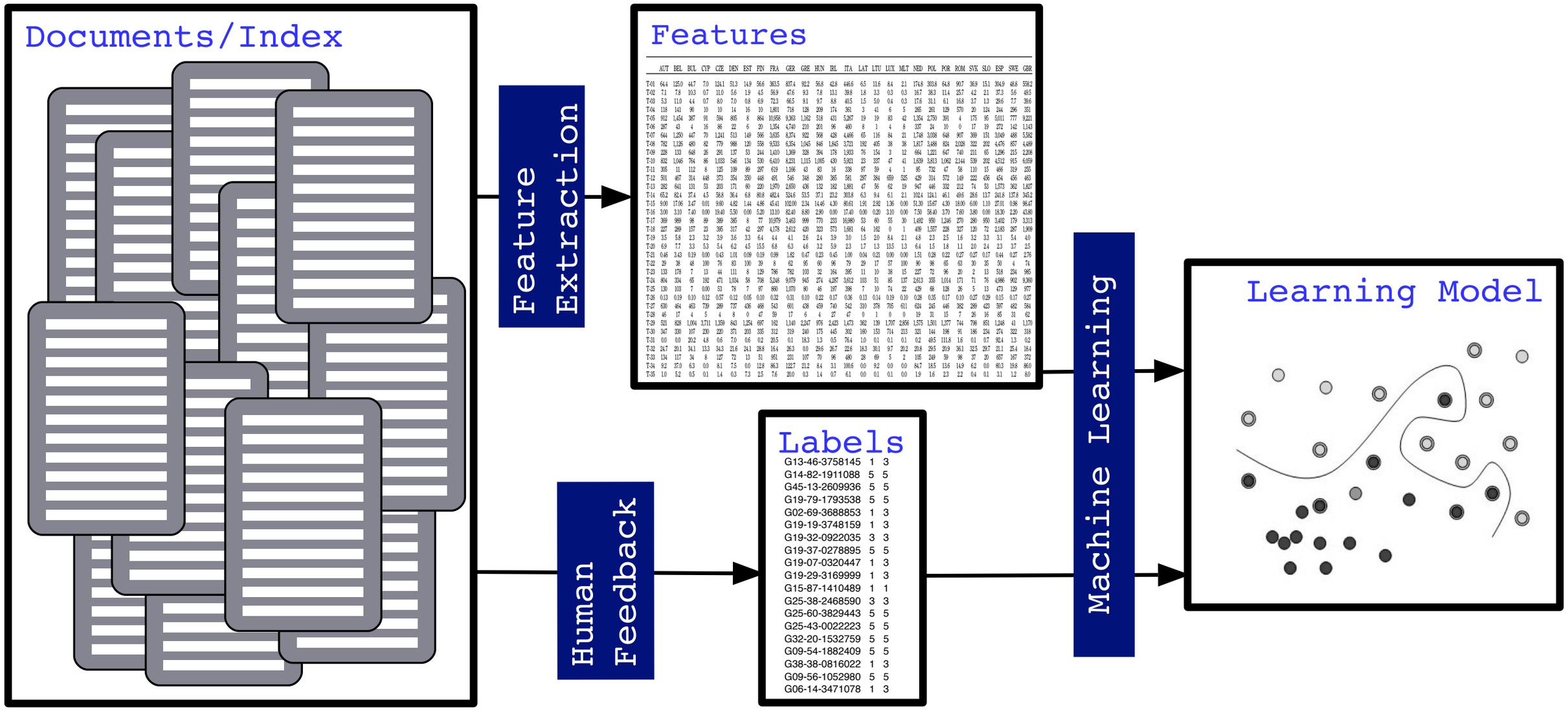
- as we keep training (insisting on ability to classify training set), the performance on the training set (green) becomes unrealistically small
  - model becomes more complex
- but at the same time ability to predict/classify new data (pink) worsens

Cross Validation Setup

	AUT	BEL	BUL	CYP	CZE	DEN	EST	FIN	FRA	GER	GRE	HUN	IRL	ITA	LAT	LTU	LUX	MLT	NED	POL	POR	ROM	SVK	SLO	ESP	SWE	GBR	
T-01	64.4	125.0	44.7	7.0	124.1	51.3	14.9	58.6	383.5	637.4	92.2	58.8	42.8	448.6	6.5	11.6	8.4	2.1	174.8	303.8	84.8	90.7	36.9	15.1	304.9	48.8	558.2	Fold 1
T-02	7.1	7.8	10.2	0.7	11.0	5.6	1.9	4.5	56.9	47.6	9.3	7.9	13.1	29.6	1.6	3.3	0.3	0.3	16.7	38.3	11.4	25.7	4.2	2.1	37.3	5.6	48.5	
T-03	5.3	11.0	4.4	0.7	8.0	7.0	0.8	6.9	72.3	68.5	9.1	9.7	8.6	40.5	1.5	5.0	0.4	0.3	17.6	31.1	6.1	16.8	3.7	1.3	29.8	7.7	39.6	
T-04	119	141	90	10	10	14	16	10	1,801	718	128	209	174	361	3	41	6	5	205	251	129	570	20	124	244	296	351	
T-05	912	1,464	387	91	894	805	8	864	10,958	9,983	1,162	518	431	5,207	19	19	82	42	1,364	2,760	824	4	175	95	5,011	777	9,221	Fold 2
T-06	267	43	4	16	80	22	6	20	1,354	4,740	210	201	96	400	8	1	4	8	337	24	10	0	17	19	272	142	1,143	
T-07	844	1,280	447	70	1,241	513	149	588	3,635	8,374	922	588	428	4,486	65	116	84	21	1,748	3,038	848	907	369	151	3,049	488	5,582	
T-08	782	1,126	490	82	779	988	120	728	9,133	6,354	1,045	846	1,845	3,721	192	405	38	38	1,817	3,488	824	2,028	322	202	4,476	957	4,489	
T-09	228	133	648	26	291	137	53	244	1,410	1,389	328	394	178	1,933	76	104	3	12	664	1,221	847	740	211	65	1,298	215	2,208	Fold 3
T-10	852	1,046	784	86	1,023	546	134	520	6,410	8,231	1,115	1,005	430	5,921	23	207	47	41	1,039	3,613	1,062	2,144	538	202	4,512	915	6,059	
T-11	305	11	112	8	125	109	89	297	619	1,166	43	83	16	358	97	50	4	1	55	732	47	68	110	15	456	319	266	
T-12	501	467	314	448	373	354	350	448	481	546	348	280	385	581	297	284	679	725	428	314	372	149	222	456	434	455	463	
T-13	382	641	131	63	908	171	60	220	1,970	2,650	495	132	182	1,881	47	56	68	19	947	446	392	812	74	63	1,573	362	1,827	Fold 4
T-14	65.2	82.4	37.4	4.5	58.8	30.4	6.8	80.8	482.4	524.5	53.5	37.1	23.2	303.8	6.3	9.4	6.1	2.1	102.4	124.1	46.1	49.6	25.6	13.7	241.8	137.8	345.2	
T-15	9.00	17.06	3.47	0.01	9.60	4.82	1.44	4.86	45.41	102.00	2.34	14.46	4.30	80.61	1.91	2.98	1.36	0.00	51.50	15.67	4.30	18.00	6.00	1.10	27.01	0.98	98.47	
T-16	3.00	3.10	7.40	0.00	10.40	5.50	0.00	5.30	13.10	82.40	8.80	2.90	0.00	17.40	0.00	0.20	3.10	0.00	7.50	58.40	3.70	7.80	3.80	0.00	18.30	2.20	43.80	
T-17	360	989	98	89	389	385	8	77	10,979	3,463	990	770	238	16,980	53	60	56	30	1,482	950	1,246	870	880	950	3,402	179	3,313	
T-18	227	280	157	23	305	317	42	287	4,178	2,512	420	323	573	1,681	64	162	0	1	408	1,557	228	327	120	72	2,183	287	1,909	
T-19	3.5	5.8	3.3	3.2	3.9	3.6	3.3	6.4	4.4	4.1	2.6	2.4	3.9	3.0	1.5	2.0	8.4	2.1	4.8	2.3	2.5	1.6	3.2	3.3	3.1	5.4	4.0	
T-20	6.8	7.7	3.3	5.3	5.4	6.2	4.5	15.5	5.8	6.3	4.6	3.2	5.9	2.3	1.7	1.3	13.5	1.3	6.4	1.5	1.8	1.1	2.0	2.4	2.3	3.7	2.5	
T-21	0.46	3.03	0.19	0.00	0.43	1.01	0.09	0.19	0.99	1.82	0.47	0.23	0.45	1.00	0.04	0.21	0.00	0.00	1.51	0.28	0.22	0.37	0.27	0.17	0.14	0.27	2.76	
T-22	29	38	48	100	78	83	100	39	8	62	95	60	96	79	29	17	57	100	50	86	55	63	30	35	50	4	74	
T-23	133	178	7	13	44	111	8	128	786	782	103	32	164	395	11	10	38	15	227	72	96	20	2	13	518	234	695	
T-24	804	334	65	192	471	1,034	53	708	5,248	9,079	945	274	4,287	3,612	103	51	85	127	2,813	355	1,014	171	71	78	4,586	902	9,360	
T-25	130	108	7	0.00	53	78	7	97	860	1,070	80	46	197	398	7	10	74	22	428	68	128	26	5	13	673	129	977	Fold K-1
T-26	0.13	0.19	0.10	0.12	0.57	0.12	0.05	0.10	0.32	0.31	0.10	0.22	0.17	0.36	0.13	0.14	0.19	0.10	0.28	0.35	0.17	0.10	0.27	0.29	0.15	0.17	0.27	
T-27	630	464	463	739	289	737	436	468	543	601	438	459	740	542	310	378	705	611	634	245	445	382	289	423	597	482	584	
T-28	46	17	4	5	4	8	0	47	59	17	8	4	27	47	0	1	0	0	19	31	15	7	28	18	85	31	62	
T-29	521	628	1,004	3,711	1,359	843	1,254	697	162	1,140	2,247	971	2,423	1,473	362	139	1,707	2,856	1,575	1,501	1,377	744	798	851	1,248	41	1,170	Fold K
T-30	347	330	107	230	220	371	200	335	312	319	240	175	445	302	160	153	71.4	213	321	144	198	91	188	234	274	322	318	
T-31	0.0	0.0	20.2	4.6	0.6	7.0	0.6	0.2	20.5	0.1	18.3	1.3	0.5	76.4	1.0	0.1	0.1	0.1	0.3	49.5	111.8	1.5	0.1	0.7	92.4	1.3	0.3	
T-32	24.7	20.1	34.1	13.3	34.3	21.8	24.1	28.8	16.4	28.3	0.0	23.6	26.7	22.6	18.3	20.1	9.7	20.2	20.8	29.5	20.9	38.1	32.5	23.7	21.1	26.4	18.4	
T-33	124	117	34	8	127	72	13	51	851	231	107	70	91	491	28	39	3	2	702	249	39	98	27	20	607	107	272	
T-34	9.2	37.0	6.3	0.0	8.1	7.5	0.0	12.8	88.3	122.7	21.2	8.4	3.1	100.6	0.0	9.2	0.0	0.0	84.7	18.5	13.6	14.9	8.2	0.0	60.3	19.8	80.0	
T-35	1.0	5.2	0.5	0.1	1.4	0.3	7.3	2.5	7.6	20.0	0.3	1.4	0.7	6.1	0.0	0.1	0.1	0.0	1.9	1.6	2.3	2.2	0.4	0.1	3.1	1.2	8.0	

- split data in K folds
- execute K independent learning trials:
  - train on K-1 folds
  - test on remaining fold
  - measure testing performance
- average results across K trials

Learning / Training with text objects



- for objects like text documents or images:
  - extract features (to obtain matrix form)
  - annotate (to obtain labels)

Spambase dataset

- about 4000 emails
- 54 features numerical
- two classes: spam / no_spam

Housing dataset

- 1300 houses
- 13 features (numerical)
- label : purchase prices (quantitative)

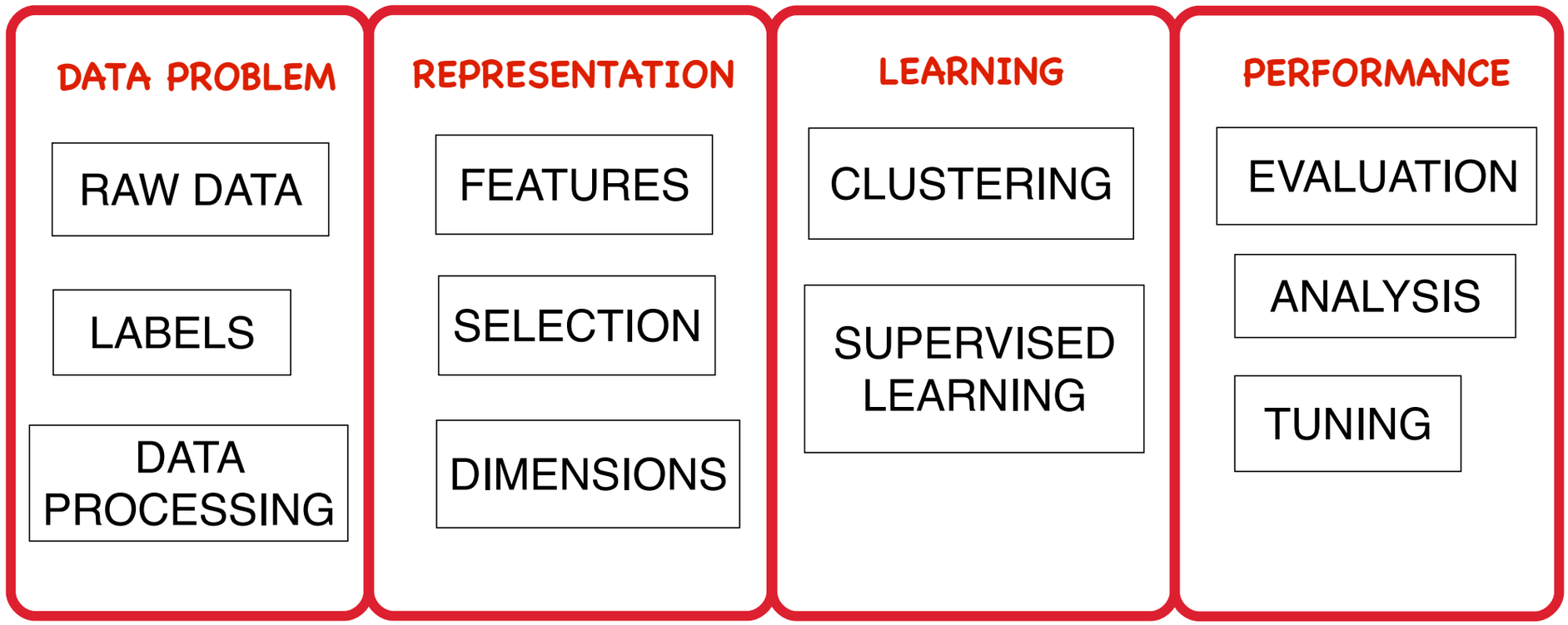
Digits dataset

- 60000 images of scanned digits
- 26x26 pixel per image, black or white
- features not extracted
- 10 classes : 0,1,2, …, 9

Documents dataset

- 20,000 news articles (text)
- features not extracted
- 20 categories: religion, music, computers, sports, etc.

course map



- main focus: learning algorithms
- main focus: hands-on practice on datasets
- secondary focus: analysis, error measurement
- secondary focus: features, representation

typical module subtaks / objectives

- THEORY
  - explain/understand fundamental mechanism
  - proof (math, intuition)
  - pseudocode
- CODE
  - run existing code
  - implement and demo your code
  - data handling: features, dimensionality, scale, missing values, normalization
  - computational issues : memory, cache, CPU, disk
- EVALUATION
  - setup
  - performance measurement, comparison
  - analysis/failure of procedure behavior
- HOWTO
  - practical advise, hacks, heuristics
  - communicate on topic well : email, forums
  - where to look online