

Data Analytics for Scanning

by

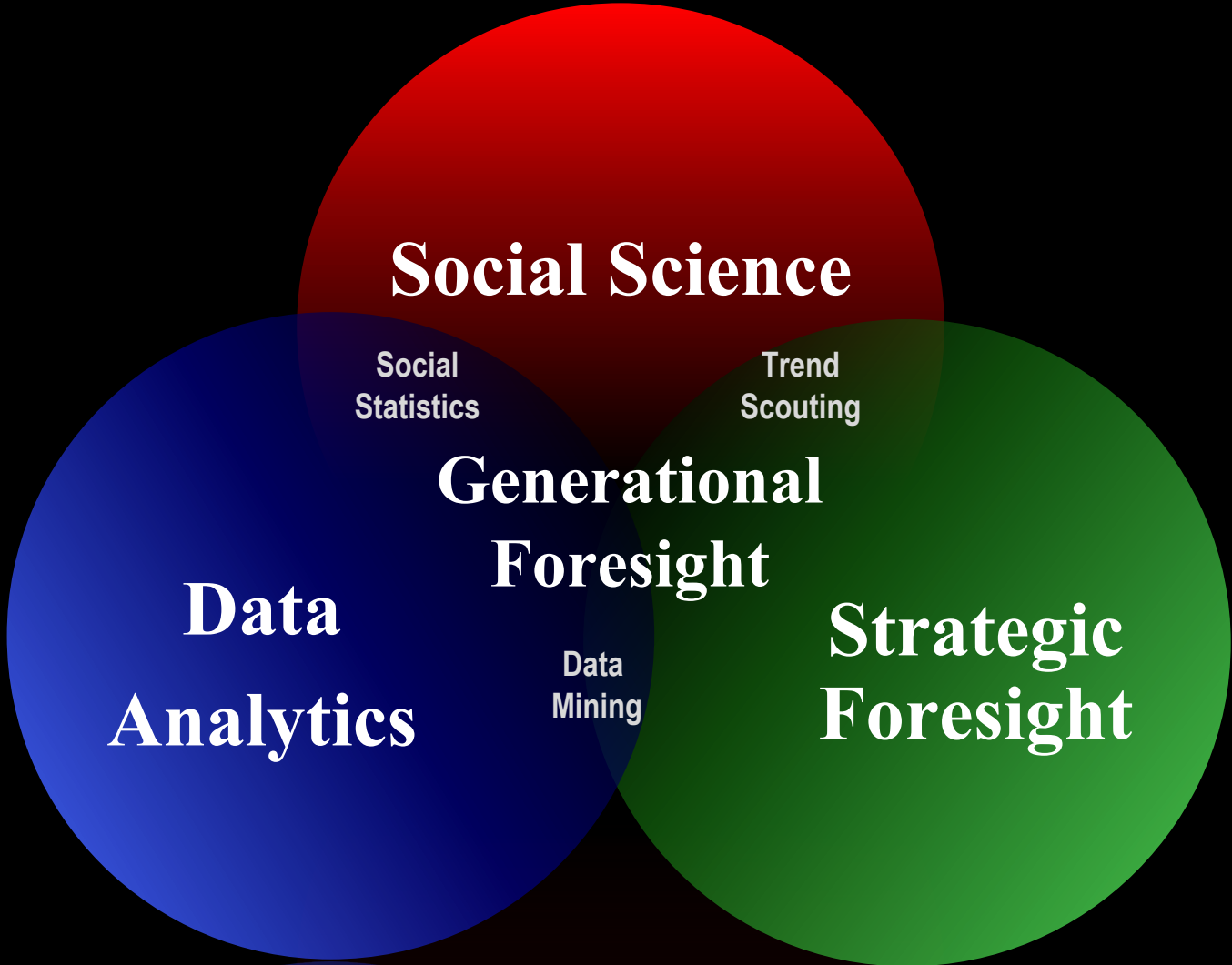
Anne Boysen

UH Annual gathering 2018



Data in Zettabytes





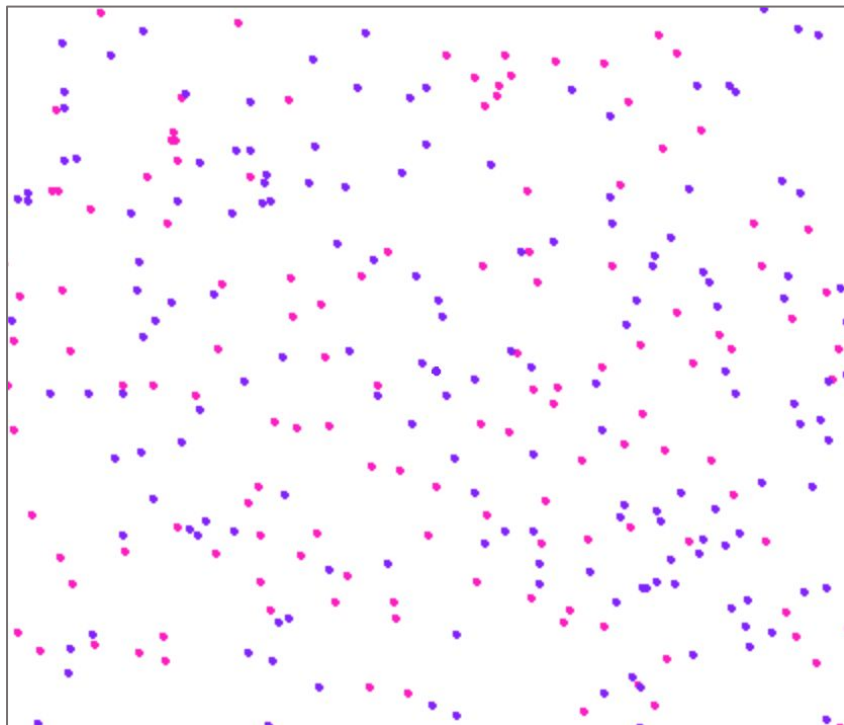
1



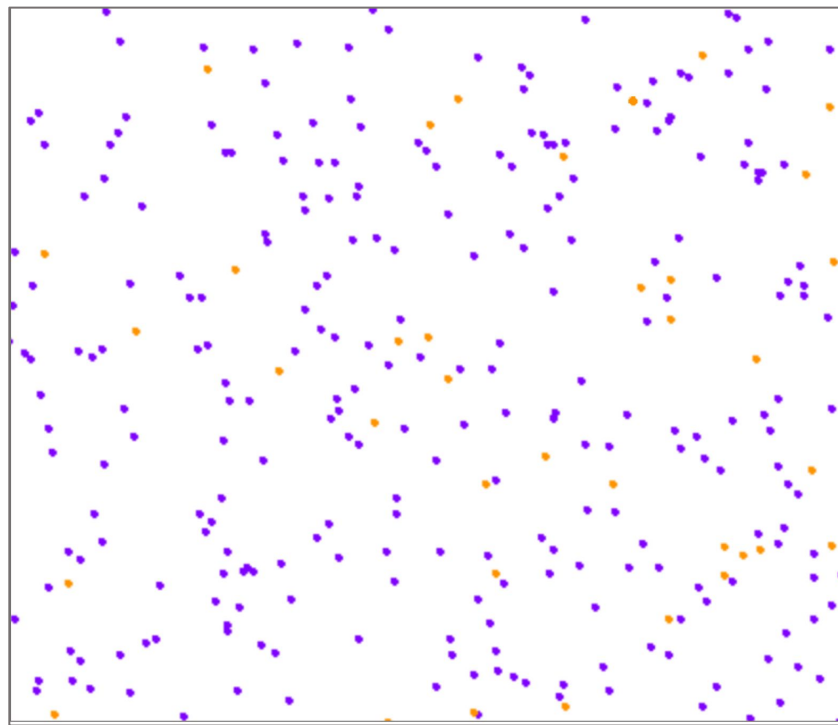
2



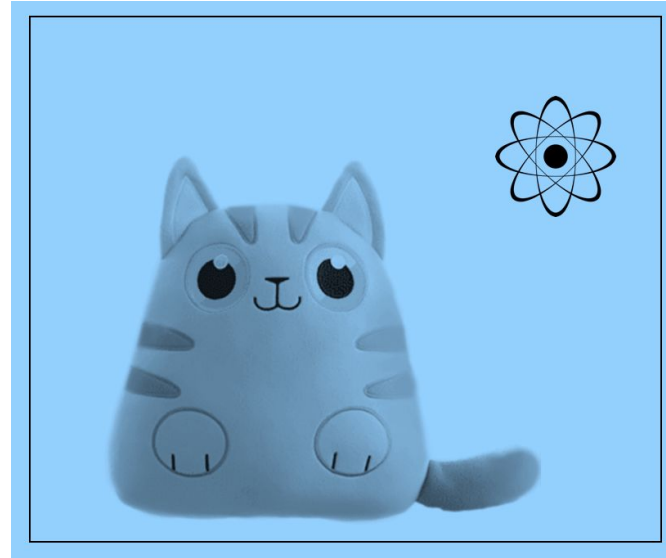
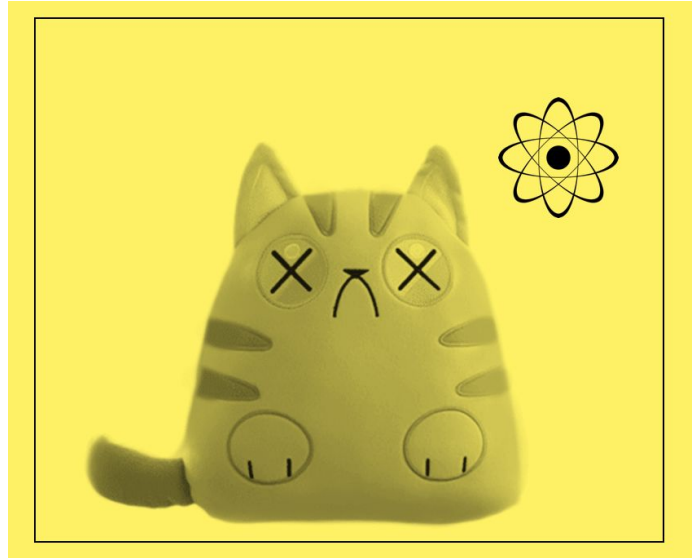
Type 1 Error

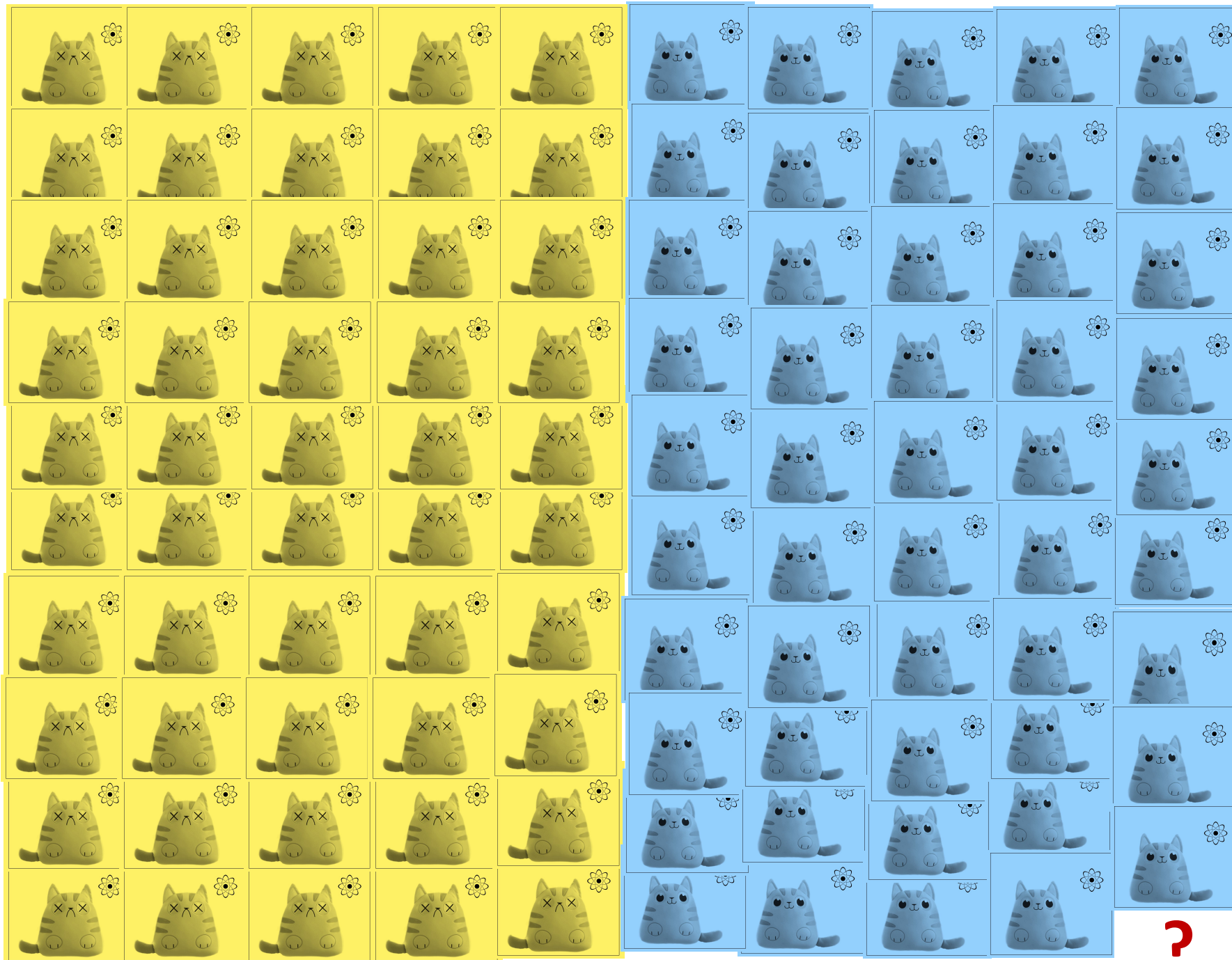


Type 2 Error



Let's play Schrödinger's cat!

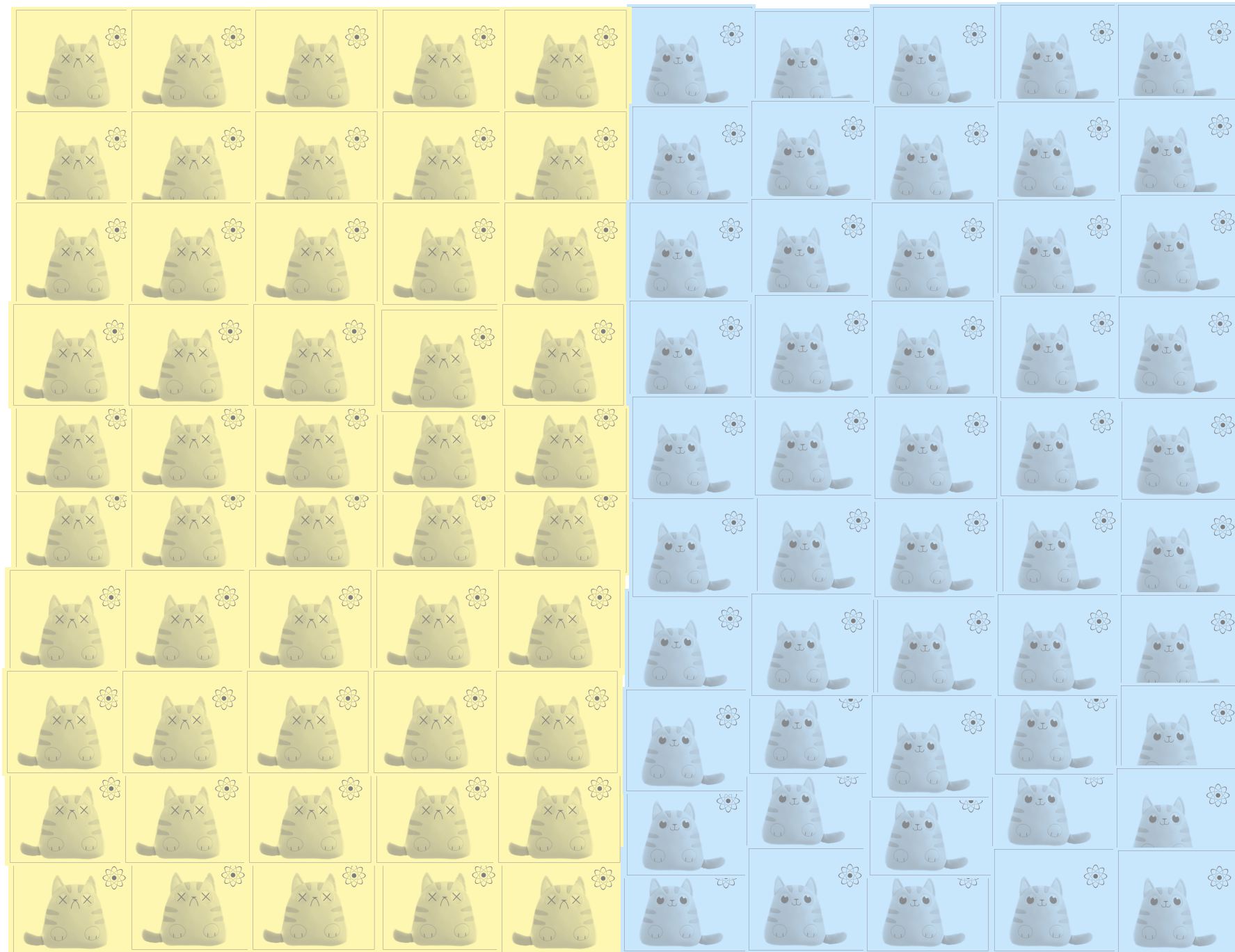




Probability
the last cat is
alive?



Probability 50
out of 100 will
be alive?



**Disease or no
disease?**



Does failure lead to success?



🕒 April 18, 2017 👤 Tomas Ondrejka

📁 Career, Entrepreneurship, Famous Resumes

Elon Musk's Resume of Failures Proves That Your Failures Aren't Big Enough

1.6k
shares



Descriptive

creates a summary of historical data.

Predictive

Identifies the likelihood of future outcomes based on data mining, algorithms and machine learning techniques

Prescriptive

finds the best course of action for a given situation

Structured Data

Surveys

Association rules

Forecasting

Decision Trees

Logistic Regression

Simulation

Clusters

Stemming

Neural Networks

Linear Programming

Text Analytics

Sentiment Analysis

NLP

CNN

Unstructured Data

Judging a survey

relevance – validity - reliability

- What is being asked?
- Who are asked?
- How are they being asked?
- How is it interpreted?



Norwegian media barometer

UPDATED

April 20, 2017

NEXT UPDATE

Currently not determined

Relevance

Does it answer what you *really* need to know?

67%

watching television on average

Share that has used different mass media an average day (9-79 years)

	percent			
	1991	2000	2015	2016
Printed	84	77	42	39
Television	81	82	67	67
Radio	71	57	59	59
audio media	43	50	38	37
weekly magazine	21	17	7	5
Books	24	20	23	25
Journal	18	14	8	8
comic Blade	11	9	3	3
Series / Movie / Video ¹	10	10	21	26
Internet	..	27	87	89

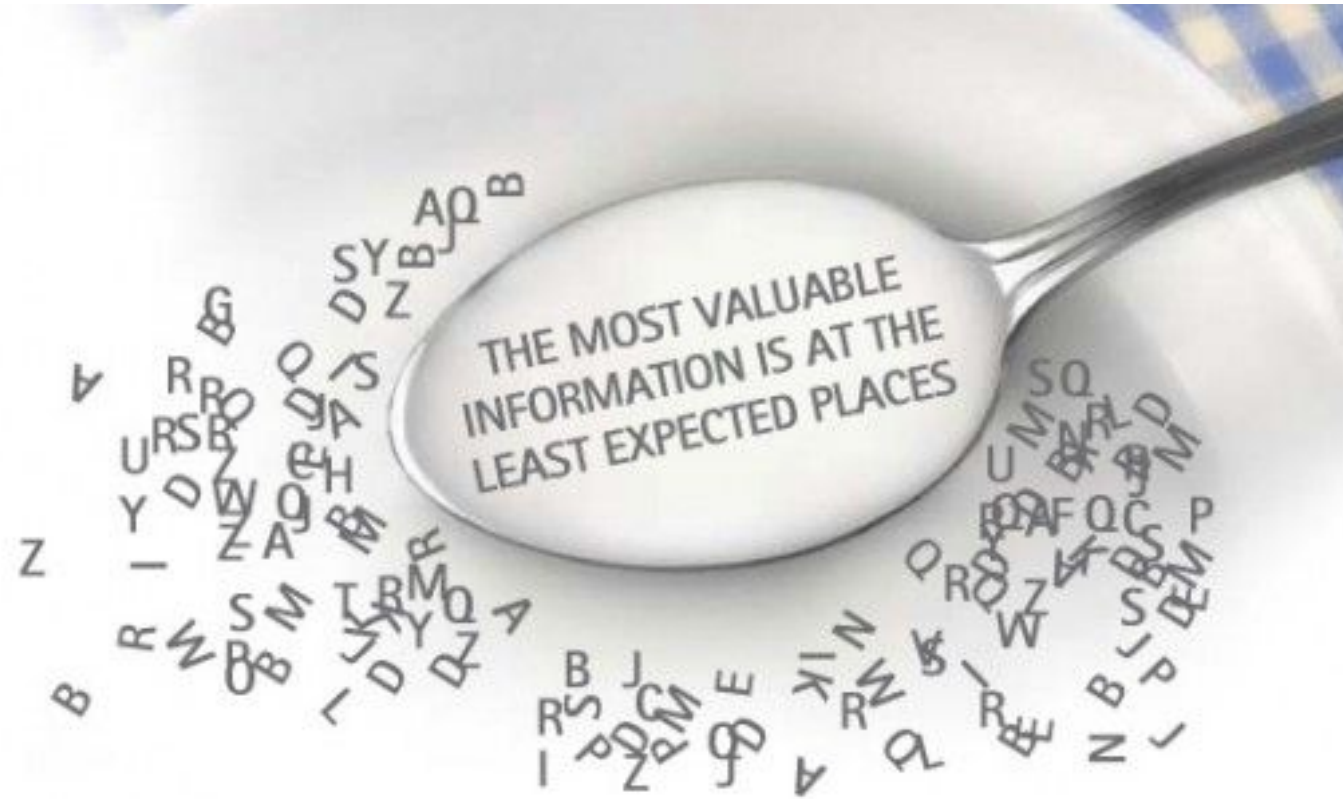
¹ Also includes streaming services via the internet



Do pay for a professional census-representative sample



Text mining marries qualitative and quantitative methods



Decision trees

Node Id:	1	
Statistic	Train	Validation
0:	36.67%	36.21
1:	25.67%	25.91
2:	23.67%	23.26
3:	14.00%	14.62
Count:	300	300

Replacement: Repla...

0 Or Missing

Node Id:	8	
Statistic	Train	Validation
0:	45.53%	39.66
1:	26.02%	22.41
2:	18.70%	25.00
3:	9.76%	12.93
Count:	123	11

Replacement: Oper...

Node Id:	9	
Statistic	Train	Validation
0:	25.86%	29.82
1:	25.86%	30.70
2:	30.17%	22.81
3:	18.10%	16.67
Count:	116	11

Replacement: Benefit

Node Id:	10	
Statistic	Train	Validation
0:	39.34%	40.85
1:	24.59%	23.94
2:	21.31%	21.13
3:	14.75%	14.08
Count:	61	7

0

1 Or Missing

2

3

0 Or Missing

1

2

3

Node Id:	11	
Statistic	Train	Validation
0:	47.83%	39.29
1:	27.54%	27.38
2:	21.74%	23.81
3:	2.90%	9.52
Count:	69	9

Node Id:	12	
Statistic	Train	Validation
0:	40.91%	54.55
1:	45.45%	9.09
2:	0.00%	27.27
3:	13.64%	9.09
Count:	22	1

Node Id:	13	
Statistic	Train	Validation
0:	40.91%	35.71
1:	4.55%	0.00
2:	36.36%	35.71
3:	18.18%	28.57
Count:	22	1

Node Id:	14	
Statistic	Train	Validation
0:	50.00%	28.57
1:	20.00%	28.57
2:	0.00%	14.29
3:	30.00%	28.57
Count:	10	

Node Id:	15	
Statistic	Train	Validation
0:	21.74%	28.57
1:	34.78%	32.14
2:	28.26%	26.79
3:	15.22%	12.50
Count:	46	5

Node Id:	16	
Statistic	Train	Validation
0:	14.29%	13.64
1:	32.14%	22.73
2:	28.57%	27.27
3:	25.00%	36.36
Count:	28	2

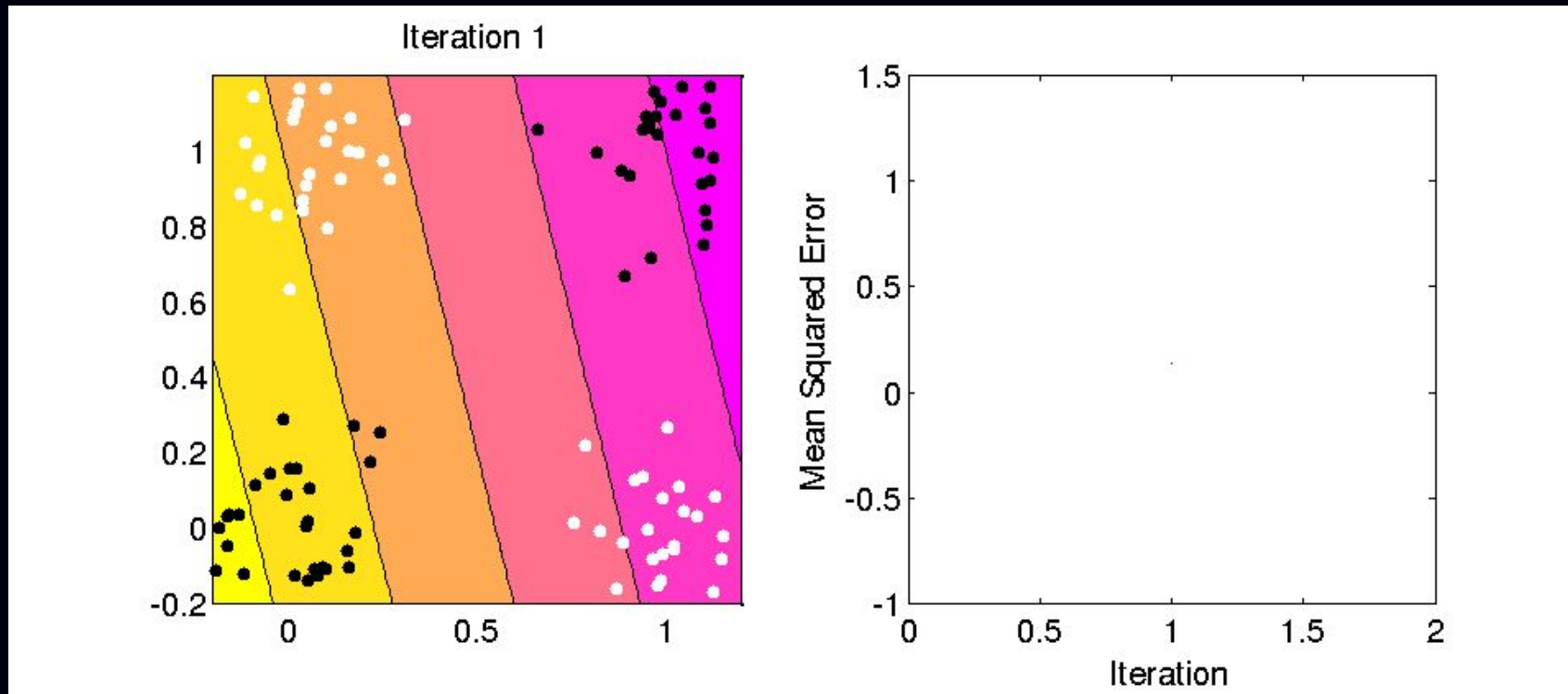
Node Id:	17	
Statistic	Train	Validation
0:	47.83%	38.46
1:	17.39%	30.77
2:	17.39%	7.69
3:	17.39%	23.08
Count:	23	1

Node Id:	18	
Statistic	Train	Validation
0:	26.32%	43.48
1:	5.26%	34.78
2:	52.63%	17.39
3:	15.79%	4.35
Count:	19	2

Predictive Weak Signals Scanning

1. Determine outliers	Novelty and detection algorithms, discriminant analysis, ethnographic approaches, desk research or brainstorm
2. Identify relationships to other variables	Use association rules to see how outliers are tied to more frequently occurring data.
3. Identify emerging clusters and segments	Use co- occurrences to identify clusters or segments. Cluster Analysis (K-means, Hierarchical clusters)
4. Explore target segments	Capture behavioral and attitudinal data via surveys, web-scraping etc. to examine direction of outlier trend

Multilayer Perceptrons / Neural Networks



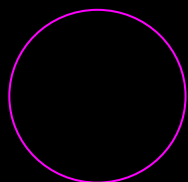
Input Layer

Input nodes

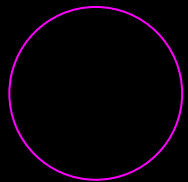
Hidden layers

Output Layer

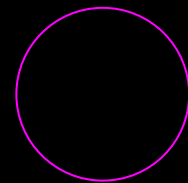
x_1



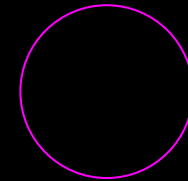
x_2



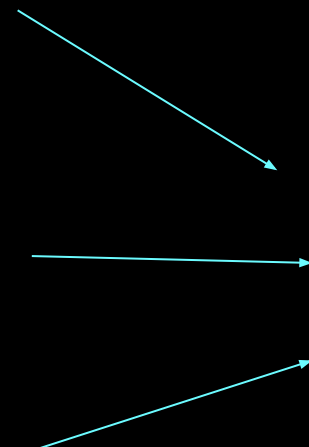
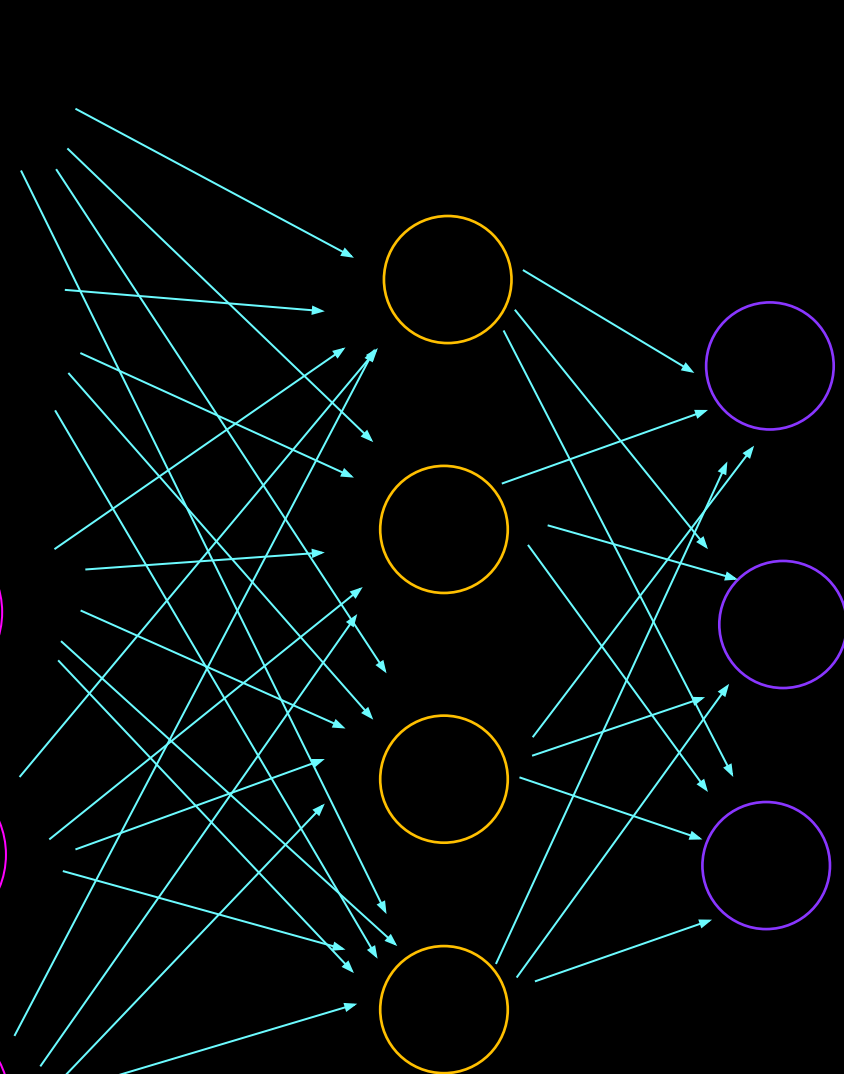
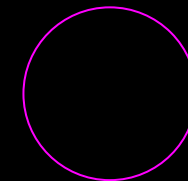
x_3



x_4



x_5



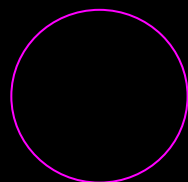
Output

Input Layer

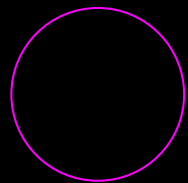
Input nodes

Output Layer

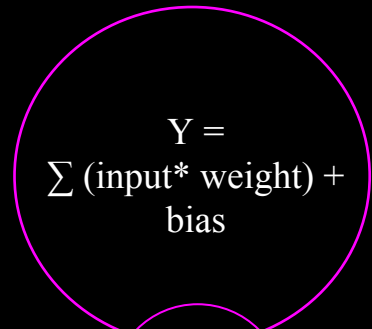
X_1



X_2

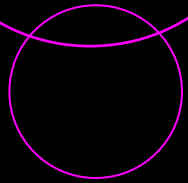


X_3

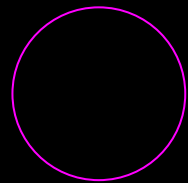


Output

X_4



X_5



Activation function



To get non-linear outcomes:

- Logistic/ sigmoid function
- Hyperbolic Tangent (tanh)
- ReLU

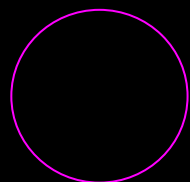
Input Layer

Input nodes

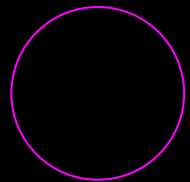
Hidden layers

Output Layer

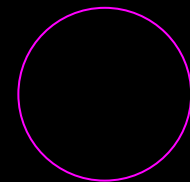
X_1



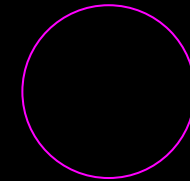
X_2



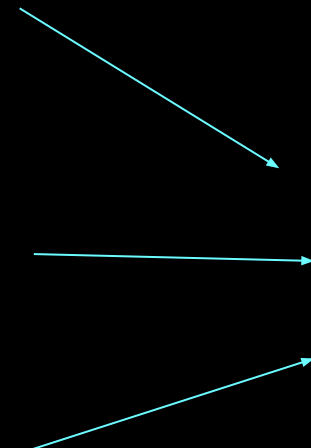
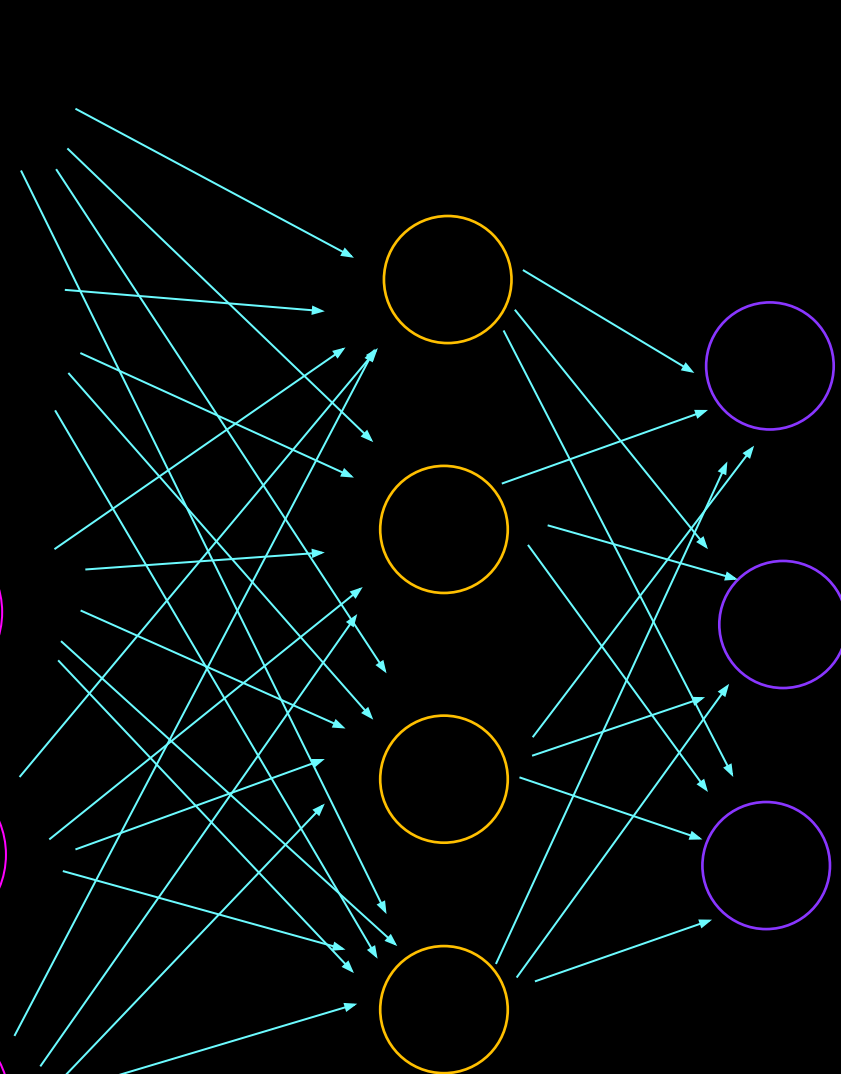
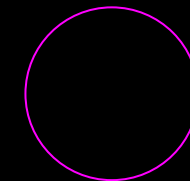
X_3



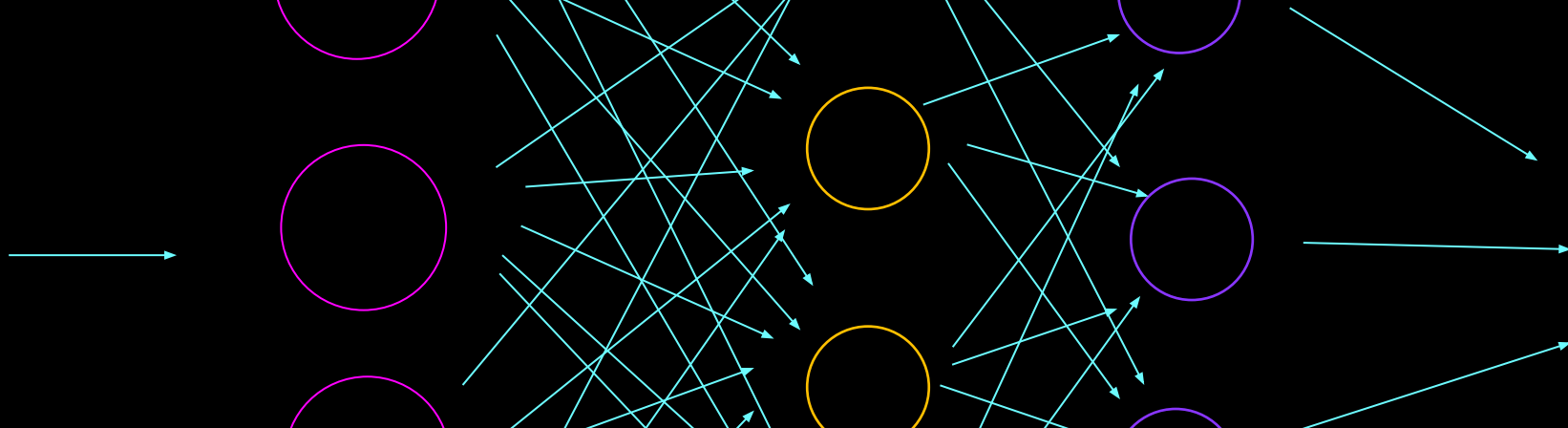
X_4



X_5



Output



*In a time of drastic change it is the (machine)
learners who inherit the future. The learned
usually find themselves equipped to live in a world
that no longer exists*

~ Eric Hoffer