CENTER FOR URBAN SCIENCE+PROGRESS

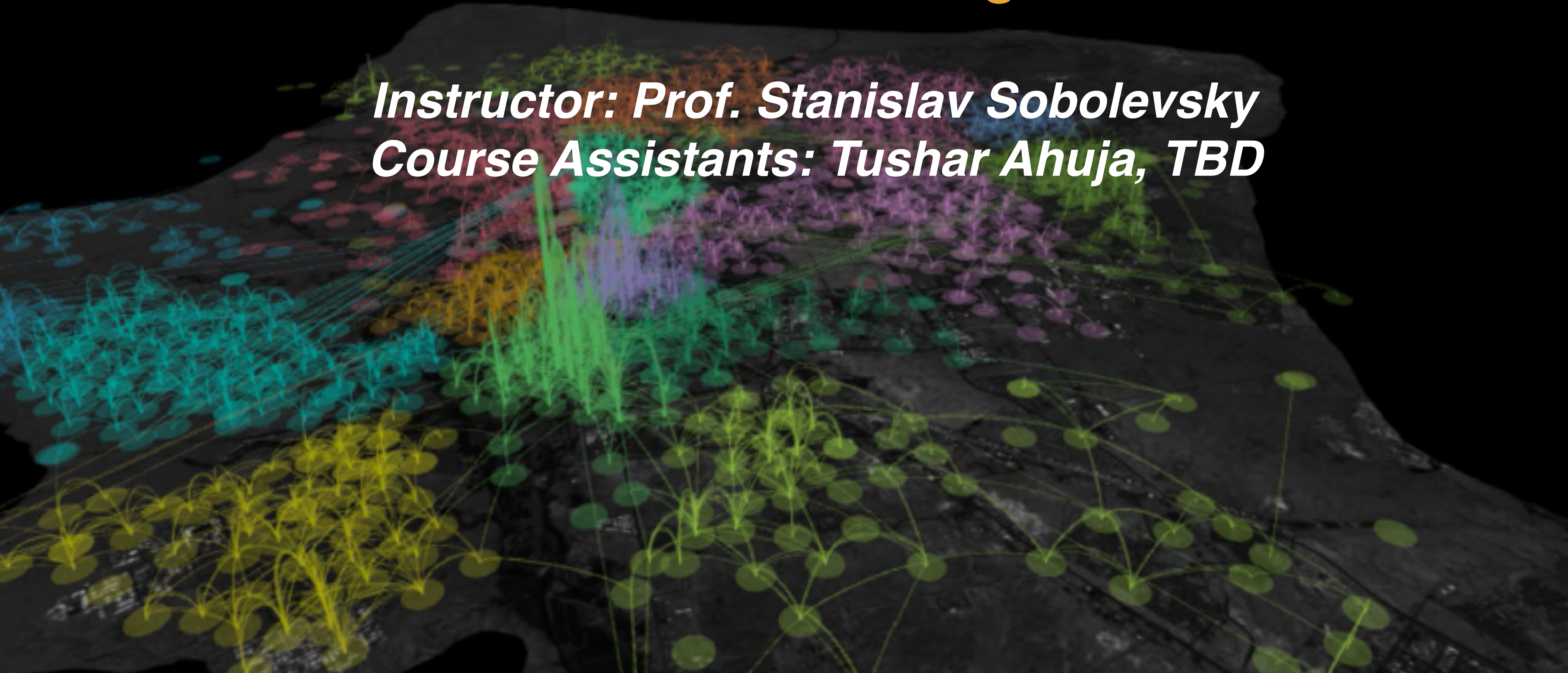
APPLIED DATA SCIENCE

5004.001/2, fall 2017

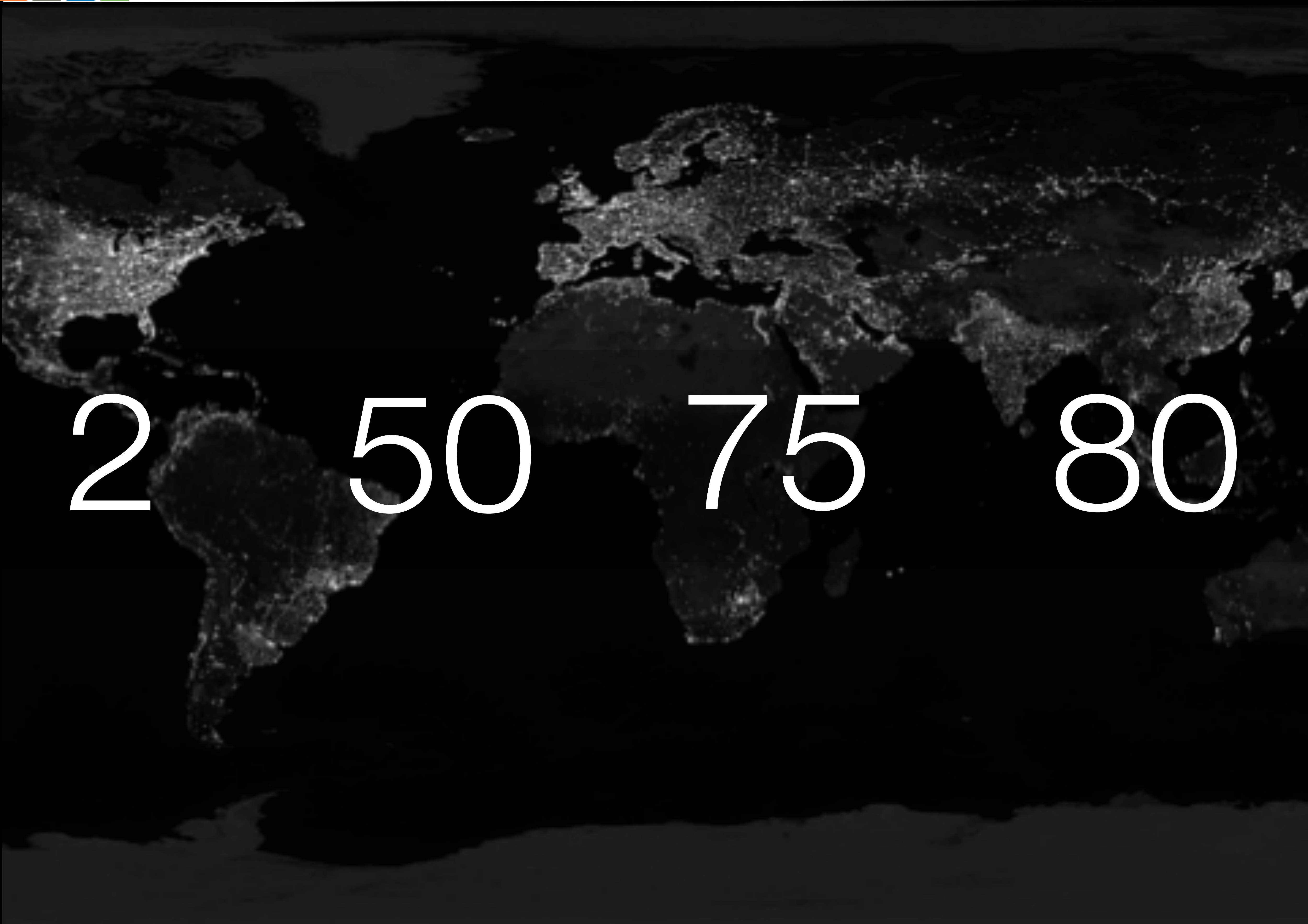
Session1: Introduction to Urban Data Science and Machine Learning

Instructor: Prof. Stanislav Sobolevsky

Course Assistants: Tushar Ahuja, TBD



…1800-3%, 1900-14%, 1950 – 30%, 2008 – 50%, 2014 – 54%, 2050-66%

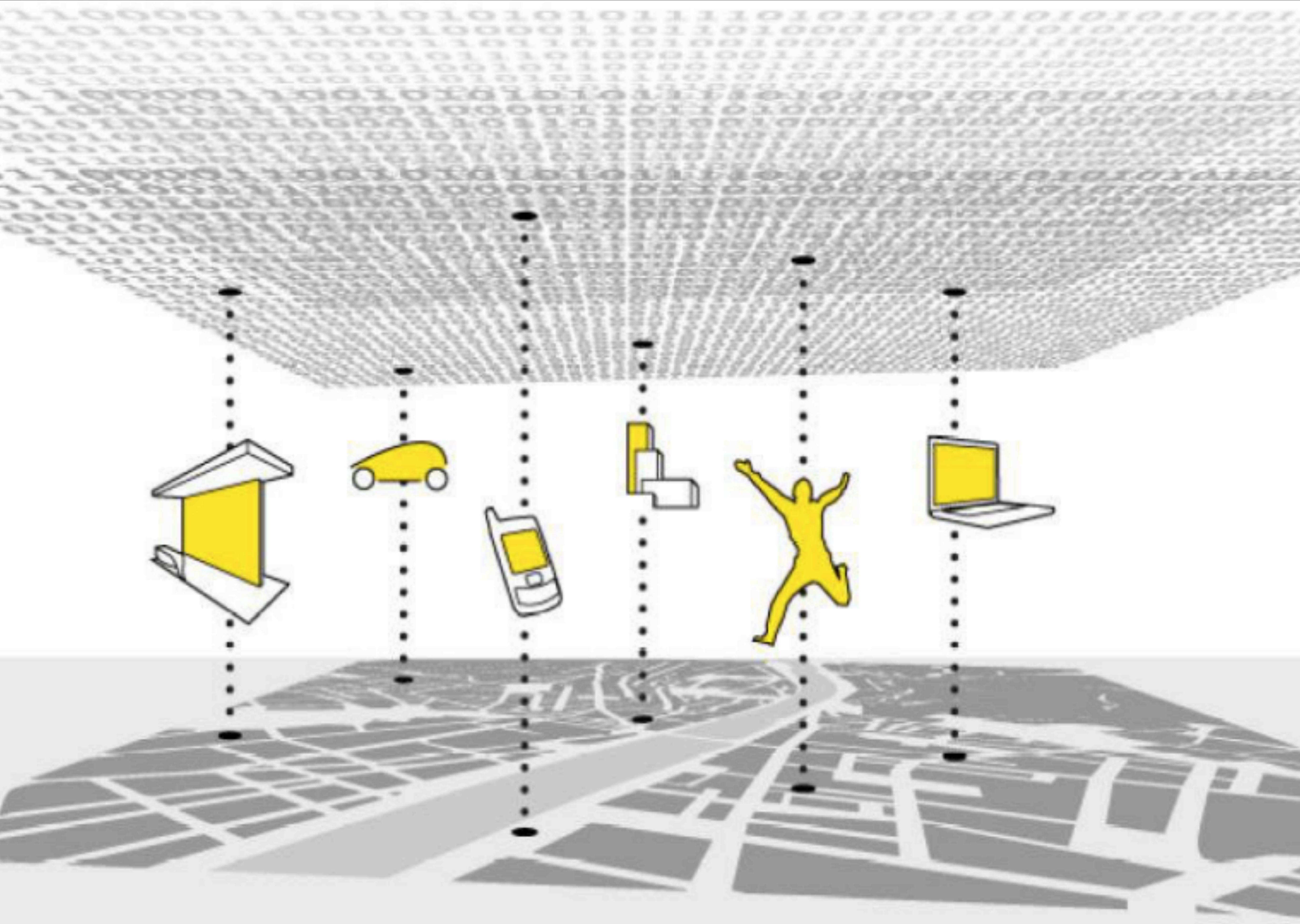

2
50
75
80

...2003 =

= 5,000,000,000GB =

=48...36...24 hours today

Google CEO, Eric Schmidt



Credits to: MIT SENSEable city lab

"... society can only be understood through a study of the messages and communication facilities which belong to it; and that in the future development of these messages and communication facilities, messages between man and machines, between machines and man, and between machine and machine, are destined to play an ever increasing part."

Norbert Wiener (1954)

Urban datasets

- Public transportation check-ins/outs
- Taxi ridership
- 311 complains; 911 calls
- Pedestrian and traffic counts
- Energy usage
- Waste collection
- Environmental: weather, air quality, noise levels
- Real-estate sales
- Social media: Twitter, Instagram, Foursquare, Flickr, Yelp
- Airbnb
- WiFi usage
- Cell phone data
- Credit card transactions
- Health records

NYTE
new york talk exchange

Globe Encounters

Globe Encounters visualizes the volumes of Internet data flowing between New York and cities around the world based on data collected over the past 24 hours. The size of the glow on a particular city location corresponds to the amount of IP traffic flowing between that place and New York City. A larger glow implies a greater IP flow.

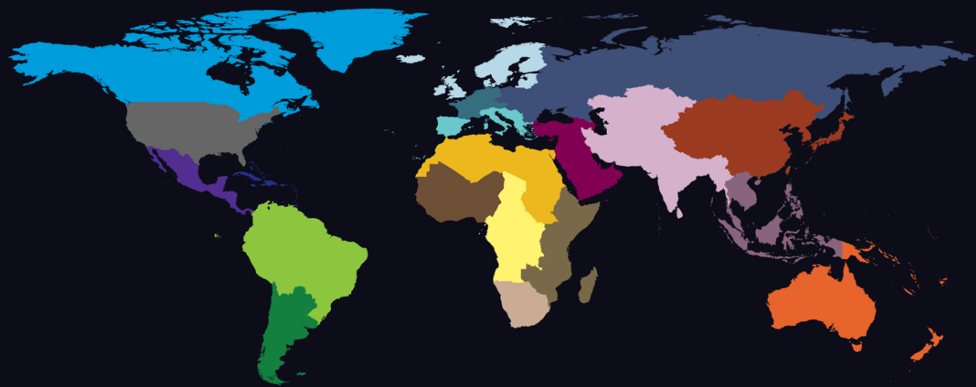
Data is continuously updated.

IP traffic | total
outgoing from new york

new york time | night | morning | afternoon | evening

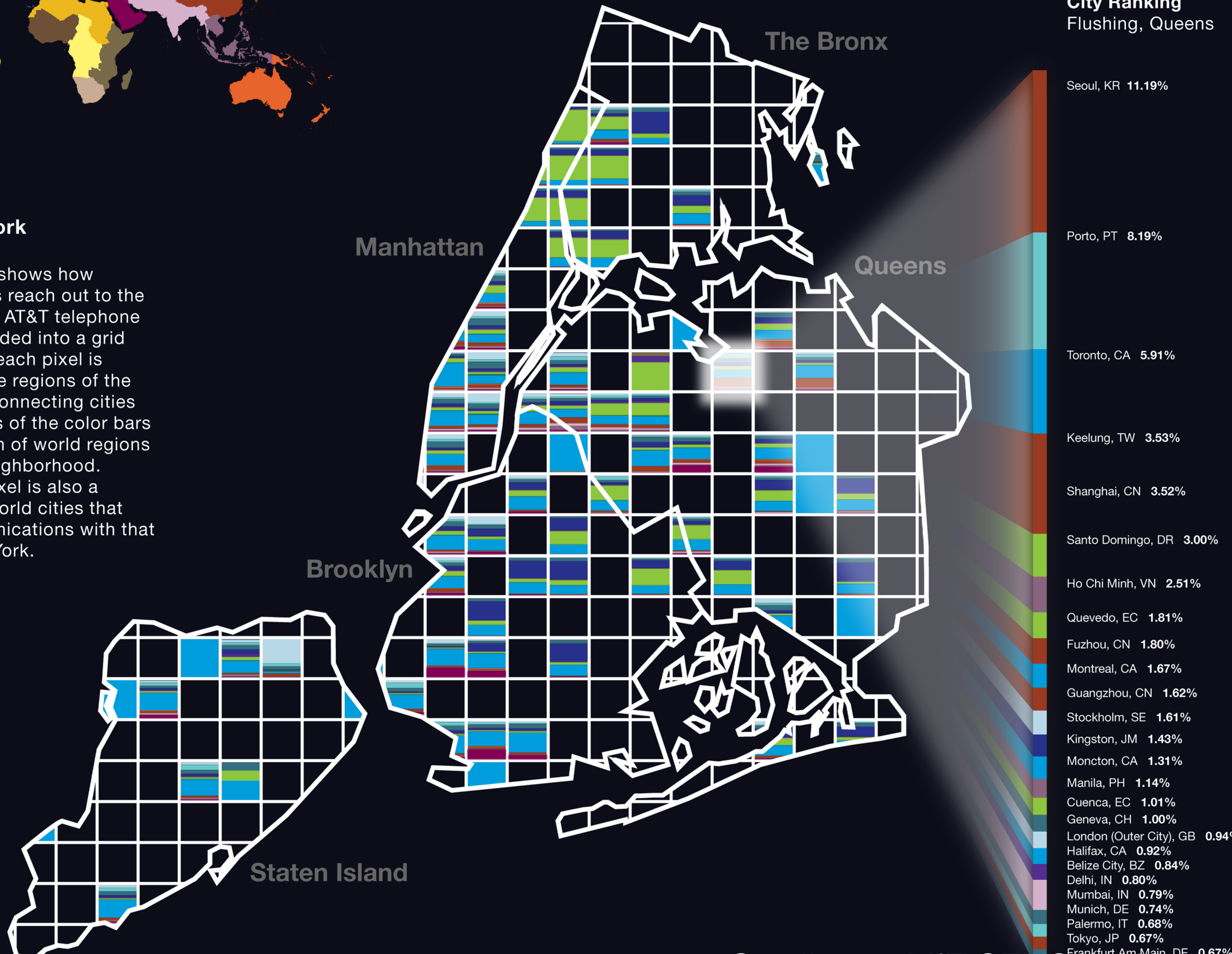
Credits to: MIT SENSEable city lab

NYTE
new york talk exchange

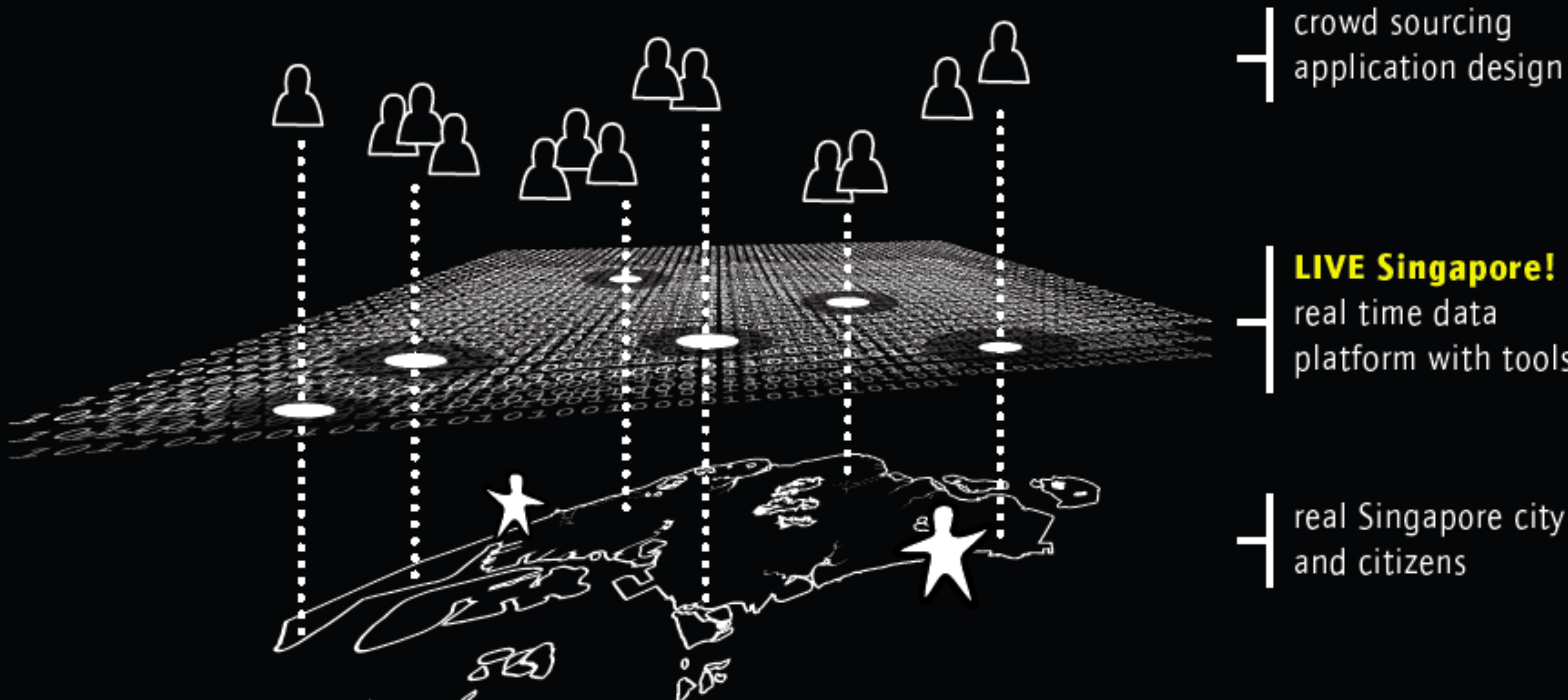


World Within New York

World Within New York shows how different neighborhoods reach out to the rest of the world via the AT&T telephone network. The city is divided into a grid of square pixels where each pixel is colored according to the regions of the world wherein the top connecting cities are located. The heights of the color bars represent the proportion of world regions in contact with each neighborhood. Encoded within each pixel is also a list of the top ranking world cities that account for the communications with that particular area of New York.



Credits to: MIT SENSEable city lab



LIVE Singapore is no single one application but an **enabling platform** for applications

QUANTIFIED COMMUNITIES

HUDSON YARDS

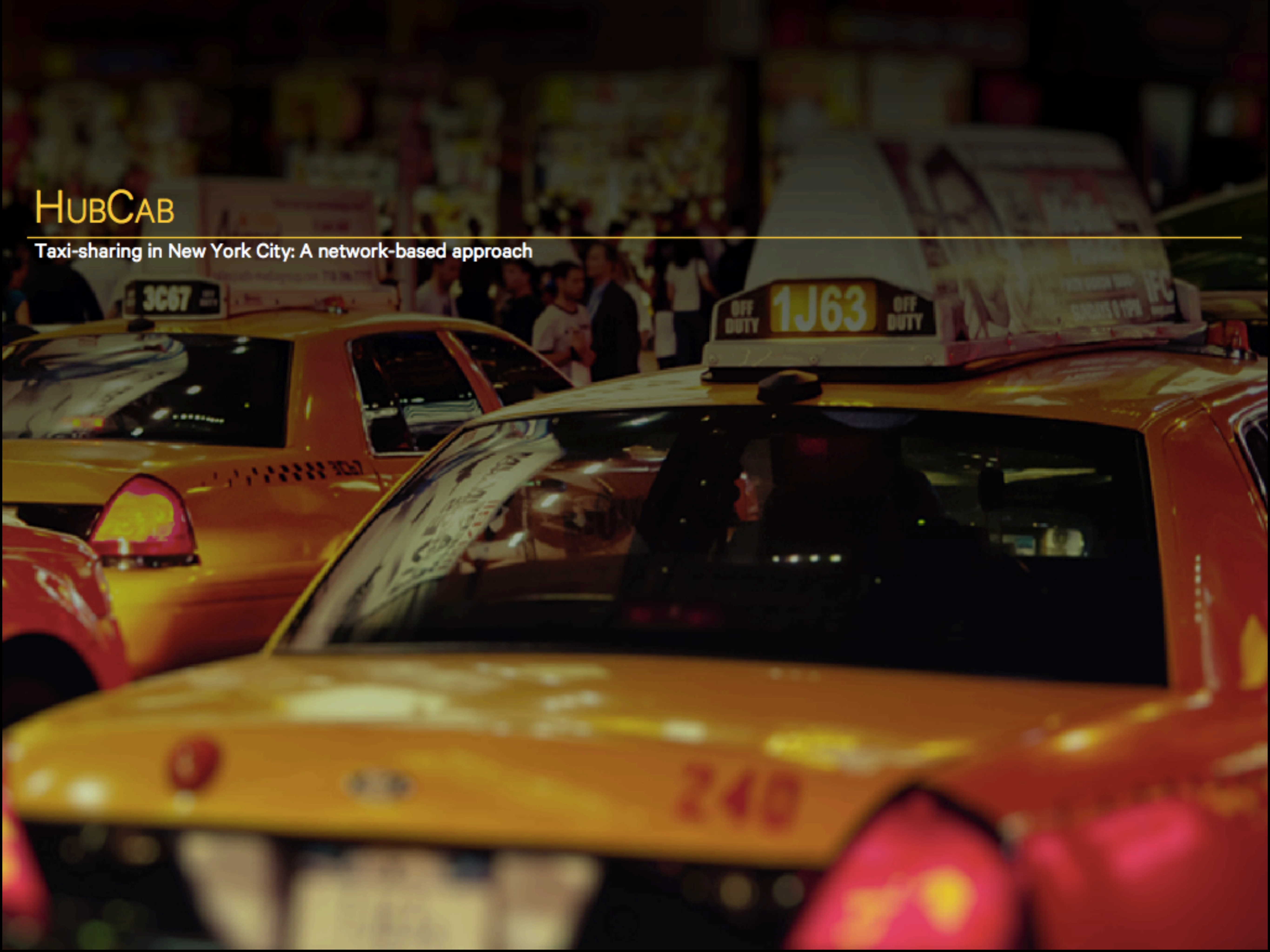
Applications of urban data science

- Urban planning
  land use classification, zoning, borders, transport infrastructure
- Urban policy
  impact of decisions, foresee and support
- Urban operations
  transportation optimization
- Urban modeling
  traffic, energy, disasters/attacks, epidemiology
- Urban applications
  UBER, Yelp, e.g. Urban Lens

EXAMPLES OF URBAN DATA SCIENCE PROJECTS

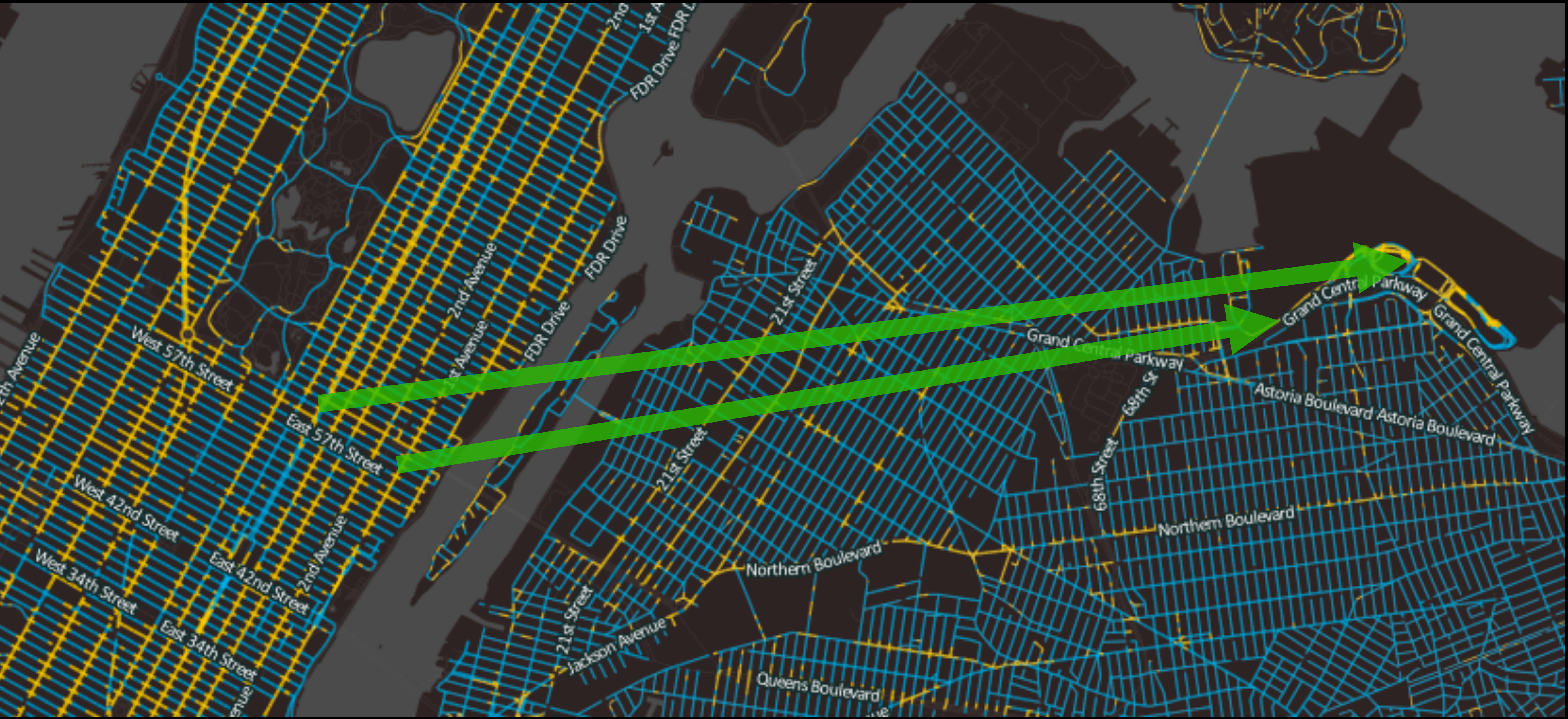
HubCab

Taxi-sharing in New York City: A network-based approach



Combine 2 trips

Regular cab



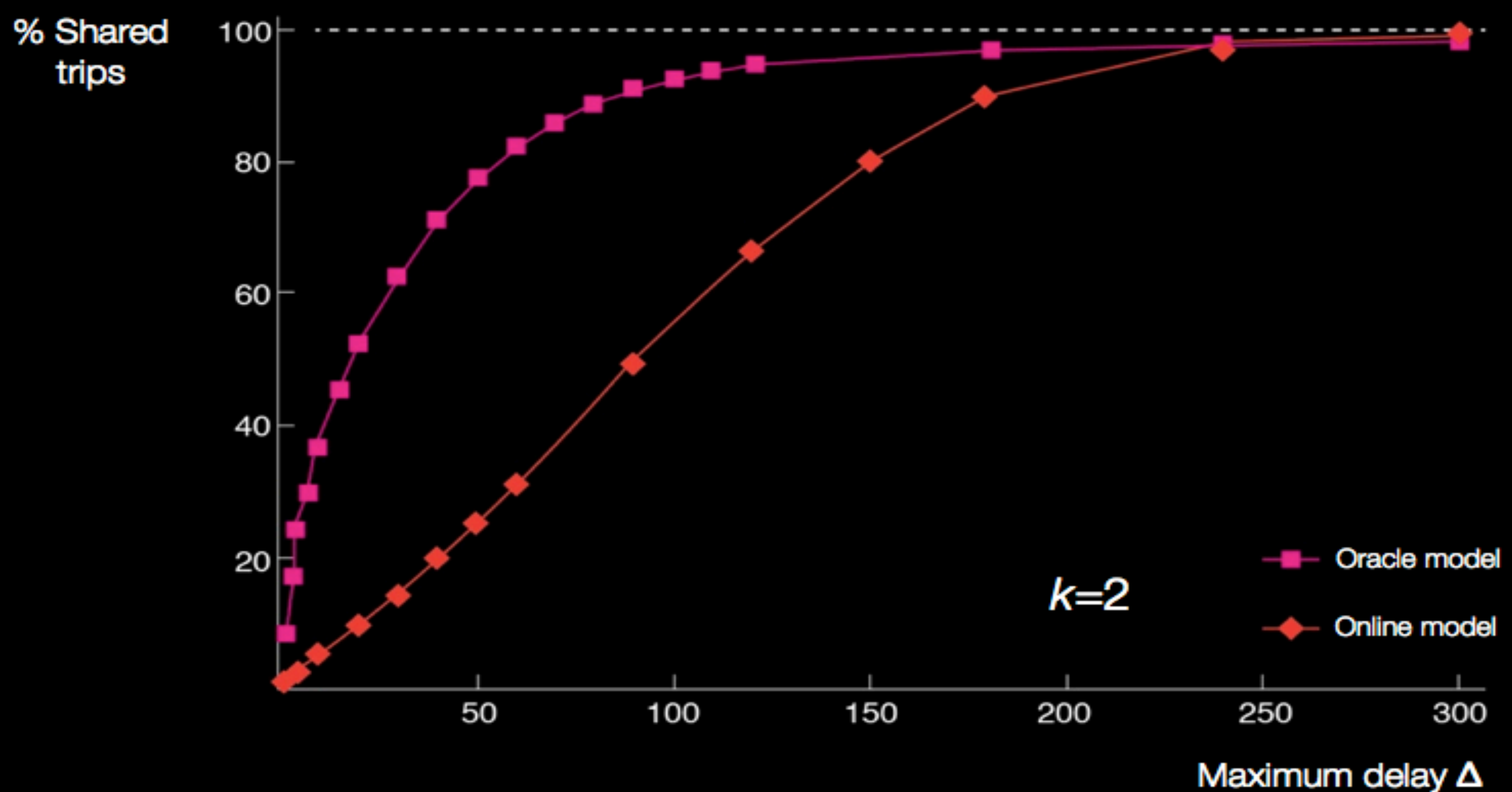
Combine k trips

“Taxi Limousine”

Transportation optimization

Santi, P., Resta, G., Szell, M., Sobolevsky, S., Strogatz, S. H., & Ratti, C. (2014). Quantifying the benefits of vehicle pooling with shareability networks. *Proceedings of the National Academy of Sciences*, *111*(37), 13290-13294.

The majority of trips is shareable!



The majority of trips is sharable with minimal passenger inconvenience

Santi, P., Resta, G., Szell, M., Sobolevsky, S., Strogatz, S. H., & Ratti, C. (2014). Quantifying the benefits of vehicle pooling with shareability networks. *Proceedings of the National Academy of Sciences*, *111*(37), 13290-13294.

CityDrive

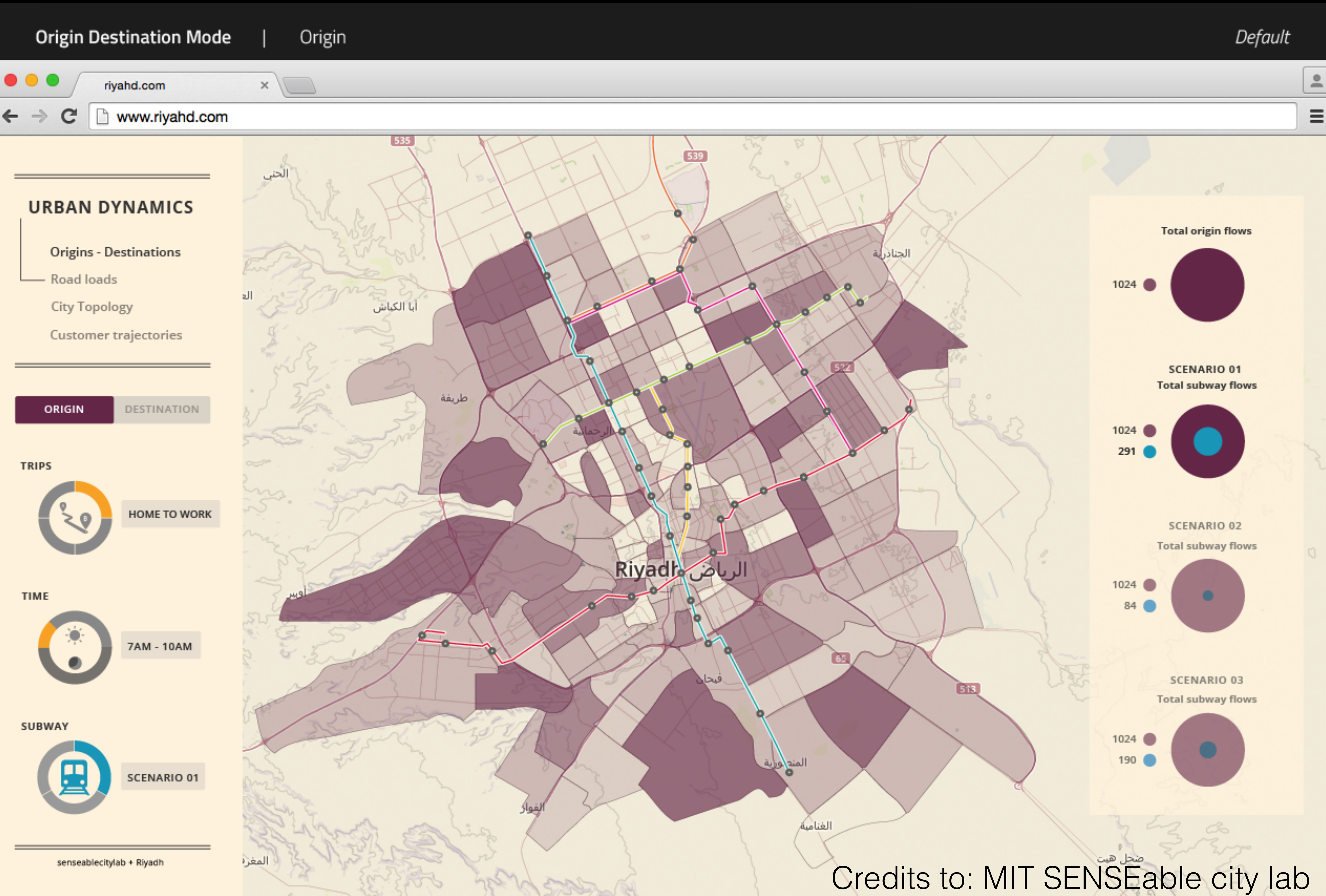
A Network-based Approach for Automated Intersection Management

Smart traffic controls

Tachet, R., Santi, P., Sobolevsky, S., Reyes-Castro, L. I., Frazzoli, E., Helbing, D., & Ratti, C. (2017). Revisiting street intersections using slot-based systems. *PloS one*, *11*(3), e0149607.

Video: credits to MIT SENSEable city lab

SMARTER RYADH – SUBWAY OPTIMIZATION

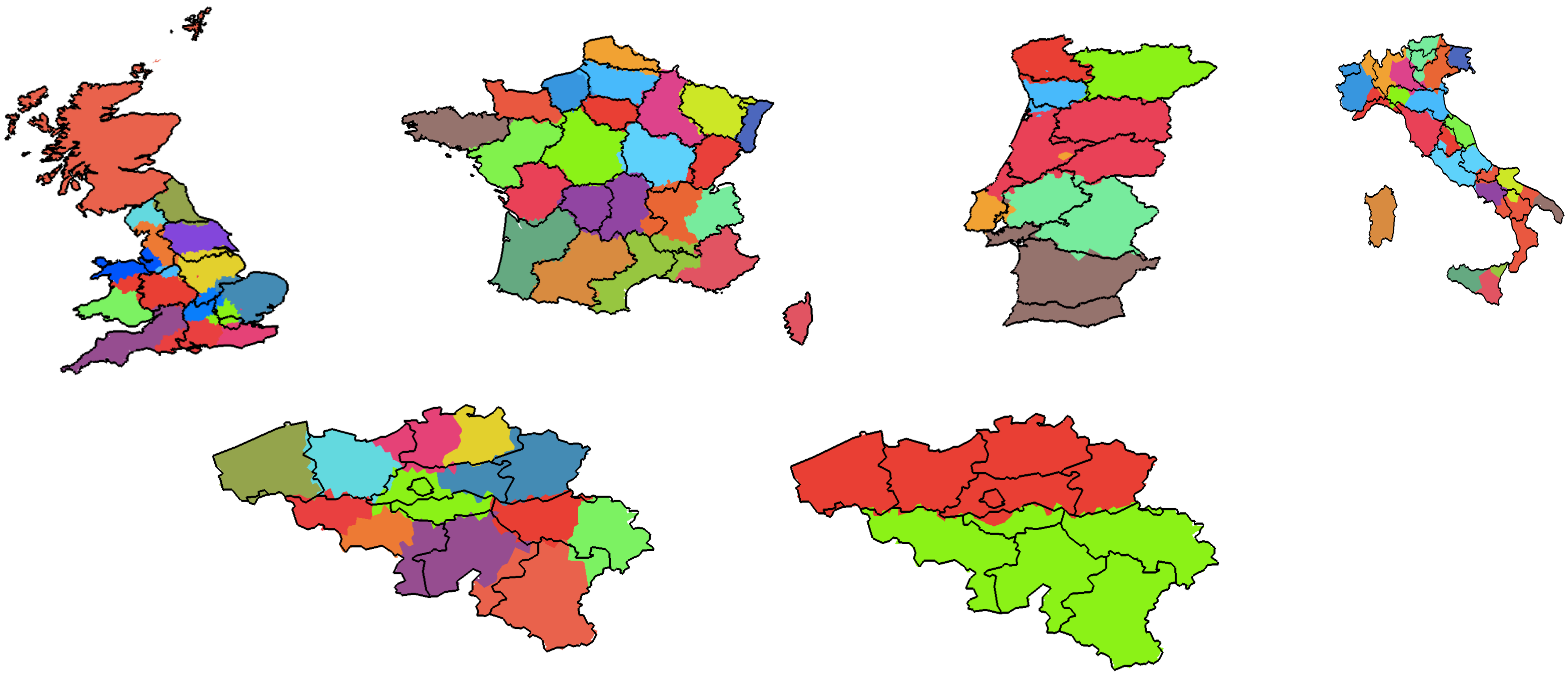


Credits to: MIT SENSEable city lab

Regional delineation

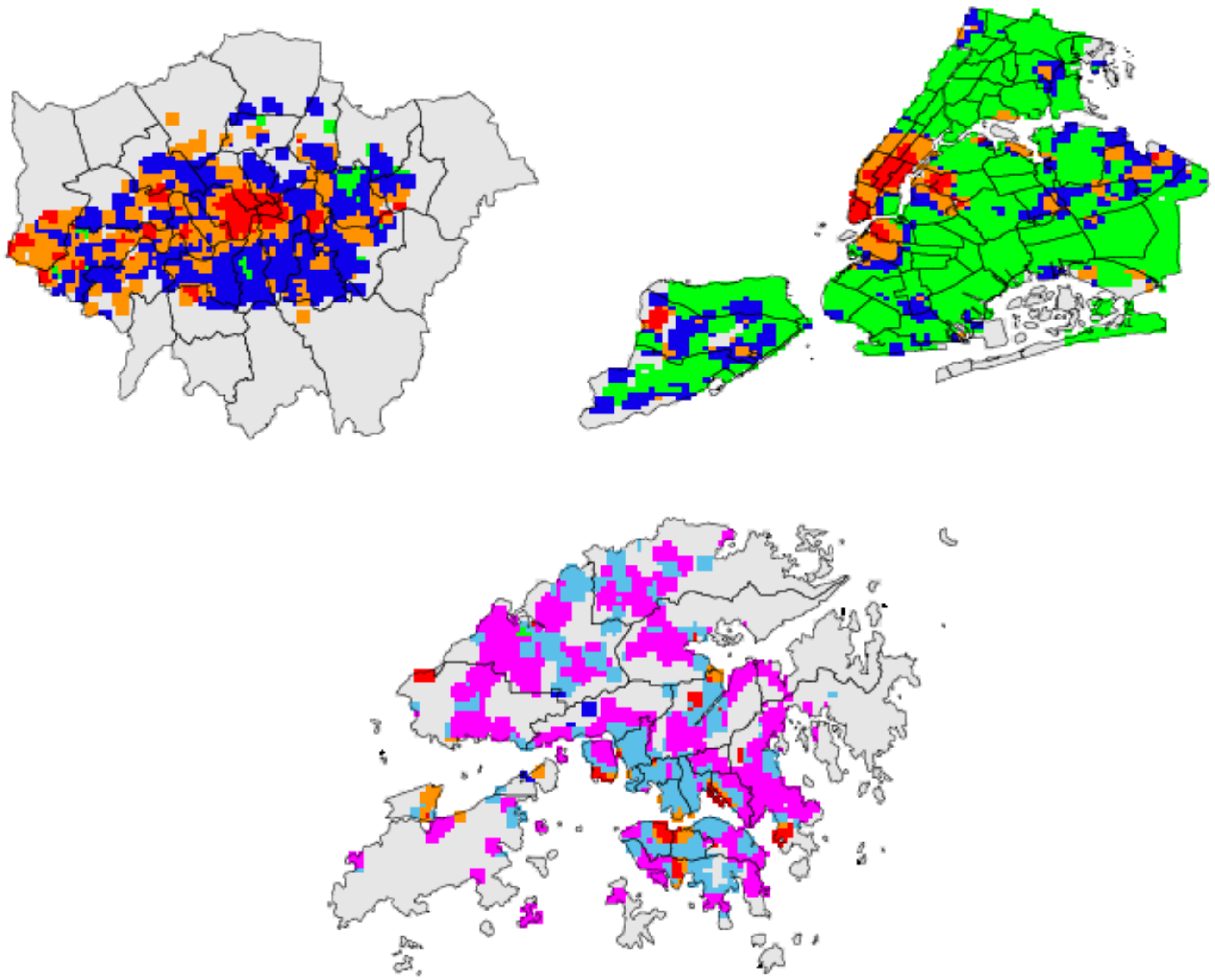
Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Martino, M., ... & Strogatz, S. H. (2010). Redrawing the map of Great Britain from a network of human interactions. *PloS one*, *5*(12), e14248.

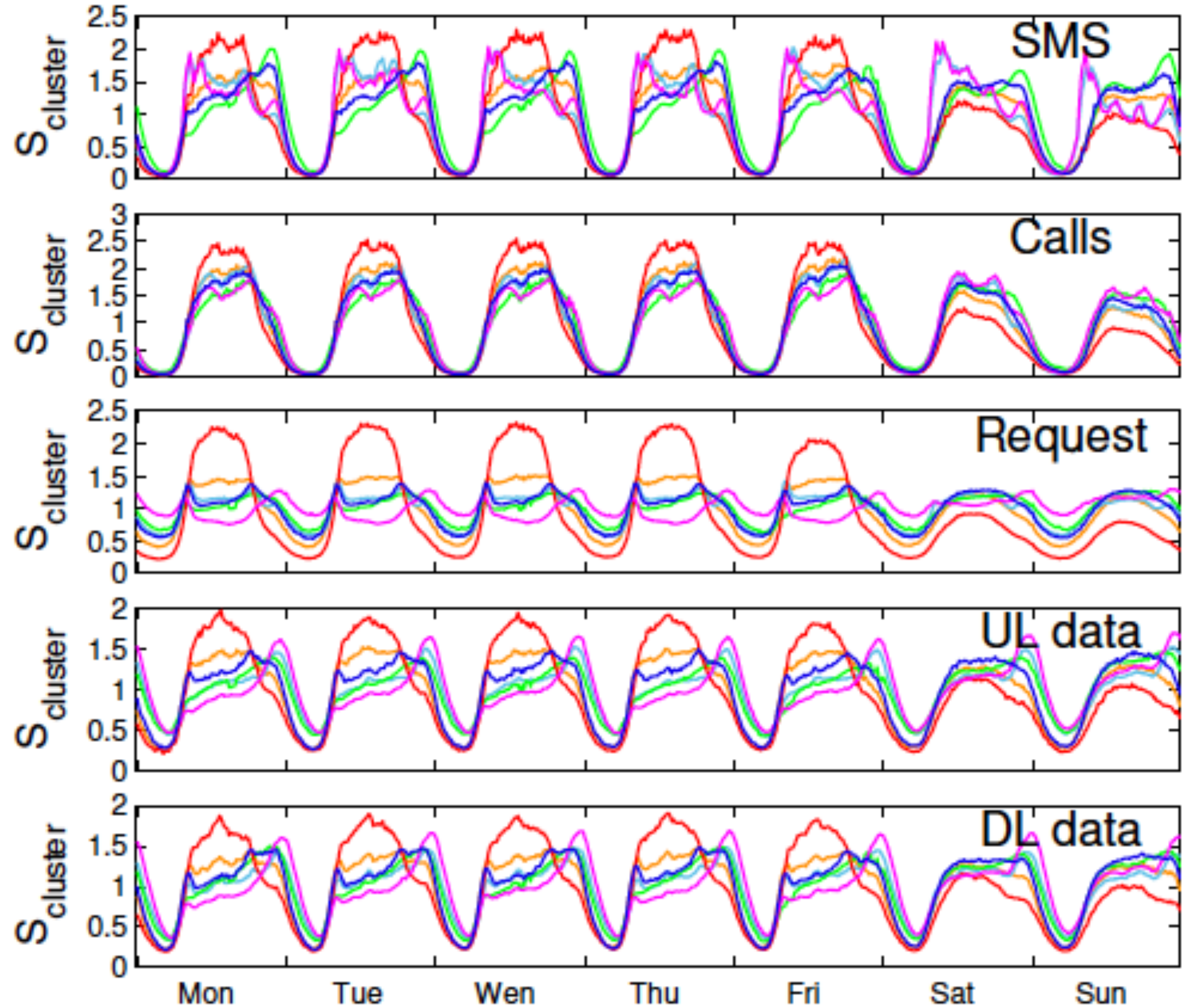
Regional delineation



Sobolevsky S., Szell M., Campari R., Couronne T., Smoreda Z., Ratti C. (2013) Delineating geographical regions with networks of human interactions in an extensive set of countries. PLoS ONE 8 (12), e81707

LAND USE CLASSIFICATION



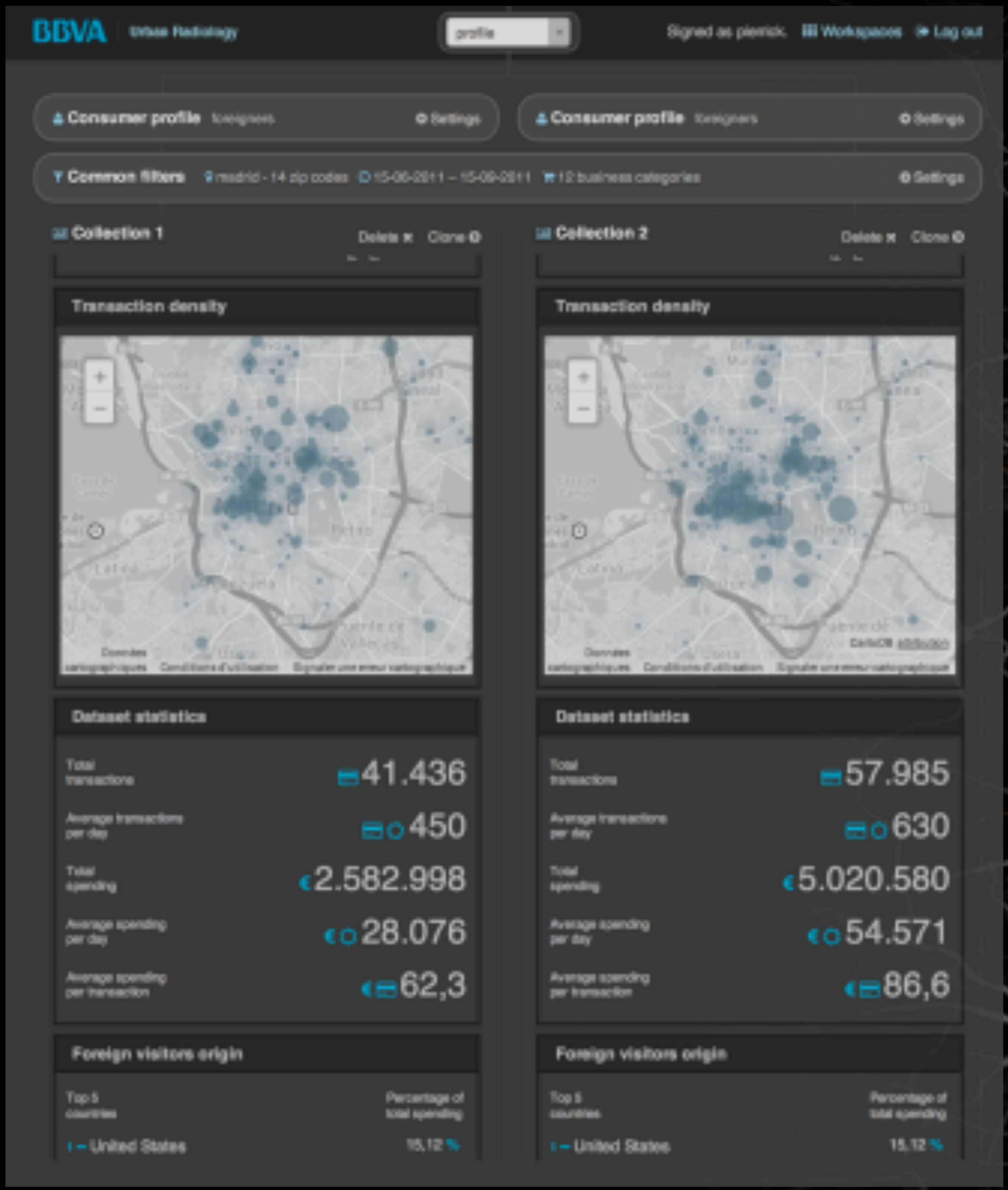


Pei, T., Sobolevsky, S., Ratti, C., Shaw, S. L., Li, T., & Zhou, C. (2014). A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*, *28*(9), 1988-2007.

Grauwin, S., Sobolevsky, S., Moritz, S., Gódor, I., & Ratti, C. (2015). Towards a Comparative Science of Cities: Using Mobile Traffic Records in New York, London, and Hong Kong. In *Computational approaches for urban environments* (pp. 363-387). Springer International Publishing.

URBAN RADIOLOGY

Translating credit card data into insights for tourism stakeholders and measuring the impact of touristic events on the urban dynamics and economy



Running prototype (2013)

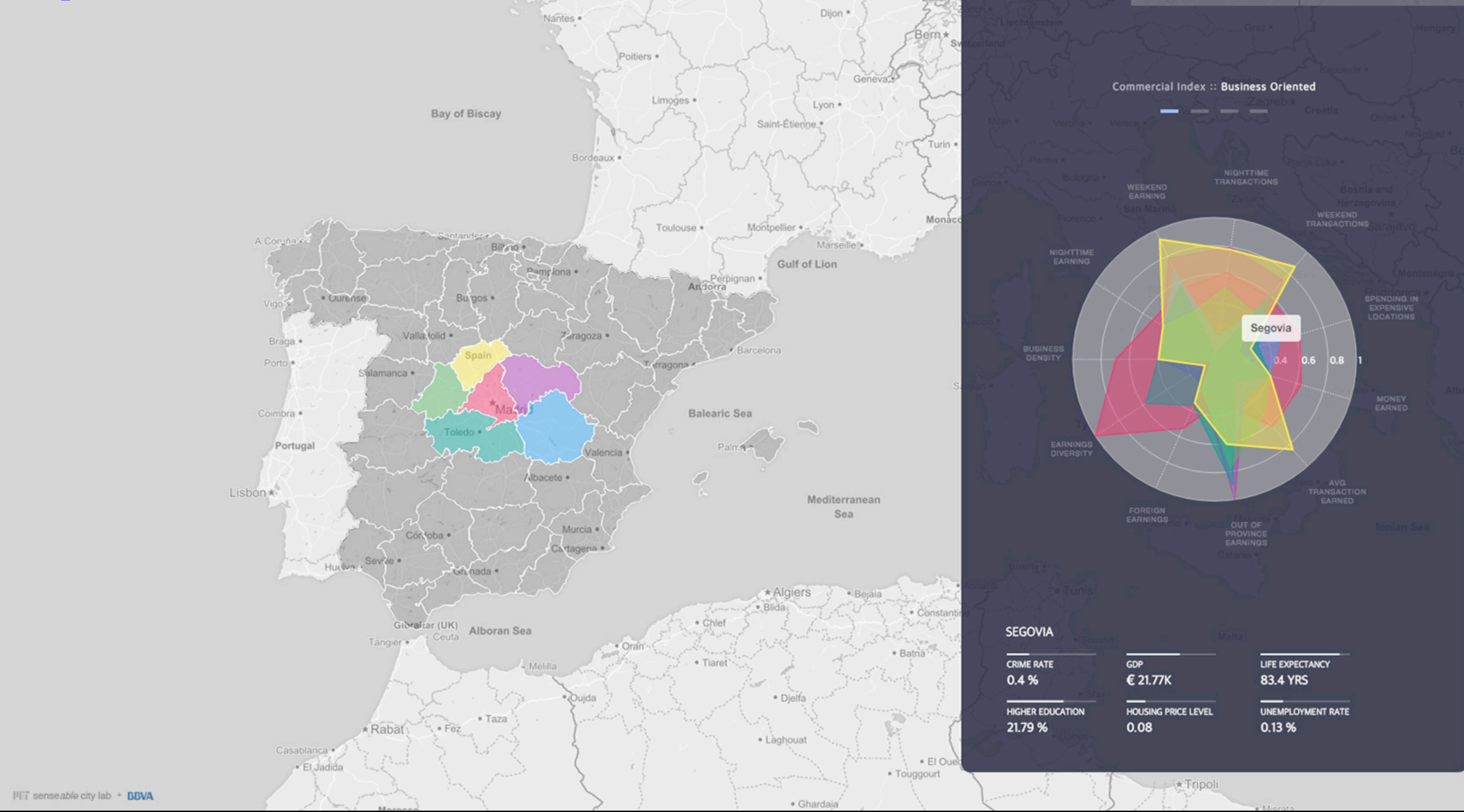
Now being developed as a product

Credits to: MIT SENSEable city lab

URBAN LENS

modeling the dynamics of socio-economic performance of regions and urban neighborhoods through bank's financial data, allowing to quantify the impact of urban interventions

http://senseable.mit.edu/urban-lens



FUTURE CITIES CATAPULT:URBAN IMPACT ASSESMENT

Midtown in Motion

LinkNYC

RideSharing

BikeSharing

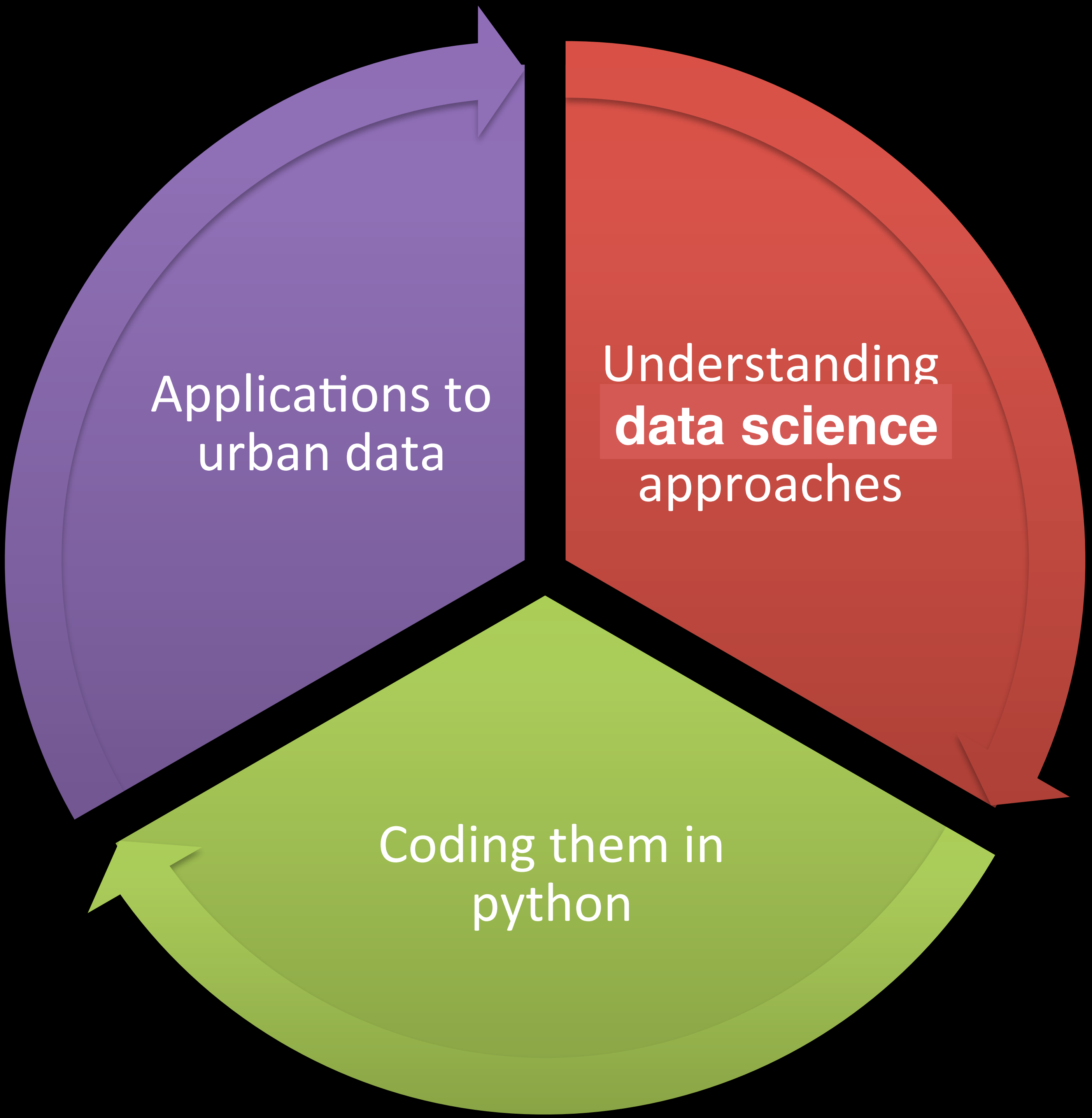
➢3 years since launch CitiBike just covered direct deployment costs

➢End user cost-benefit ratio from **2.75 to 6.92**

Direct/ Indirect	**Impacts (social, environmental, economic)**	**Stakeholder**	**Positive/Negative**	**Metrics and results in 2015**
Direct (Intended)	Efficiency gains in mobility (social/economic)	End user	Positive	Travel time savings of **8.19** hours and **$496.21** savings in time-cost per user annually
	Public health (social)	End user, city	Positive	Calories burnt: **12,437** calories per user annually
	Reducing emissions (environmental)	City	Positive	Gas savings of **3.16** liters per user translated into emission reduction for New York City
	Public safety (social)	End user, City	Negative	0.11% chance of getting injured, 0.00036% chance of getting killed, health-related loss: **$80.78** per year per user during 3 years after the initial deployment
Indirect (Unintended)	Property values (economic)	City, real-estate owners/renters	Positive for the sellers and the city, negative for customers	Growth in residential real-estate prices (**8.4%** within 200ft from the deployment)
	Local commercial activity, specifically food-related businesses (economic)	City, local businesses	Positive	Growth of sales in food-related businesses (**2.6%** in Jersey City deployment areas being translated into $36,720 in sales taxes per station)

Direct/Indirect	Impacts (social, environmental, economic)	Stakeholder	Positive/ negative	Metrics and results
Direct (Intended)	More affordable and faster transportation for the end users (social/economic)	End user	Positive	Travel cost and time savings ($1.14 million (with 95% confidence interval between $0.5 and $2.5 million) over 9 months for the entire NYC and travel time savings of 80,000 hours). Given the total cost-time of the UberPOOL rides providing those savings estimated at $21.3 million, this is translated into **1.053** benefit/cost ratio.
	Reduced road congestion (social)	City	Positive	Number of trips (473500 trip reduction over 9 months for the entire NYC)
	Reduced gas consumption and associated emission cuts (environmental)	City	Positive	Gas savings and associated emission cuts (124,412 liters equivalent to 292 metric tons of carbon emissions cut over 9 months for the entire NYC)
Indirect (Unintended)	Jobs (social)	Taxi drivers	Negative	Trips taken from taxi drivers: 5 full time taxi driver jobs lost over 9 months after the deployment due to the UberPOOL alone; Uber itself causes 400 jobs lost over the same period
	Jobs (social)	Uber drivers	Positive	Jobs created for Uber drivers: equivalent of 75 full time jobs

Goals


Applications to urban data
Understanding data science approaches
Coding them in python

Grading

Activity	*% of the total grade*	*Deadline*
Homework	40%	12 days each
Midterm	20%	October, 31/November, 1
Project	40%	December, 10 Proposals - October, 9, 2017 12pm (noon)

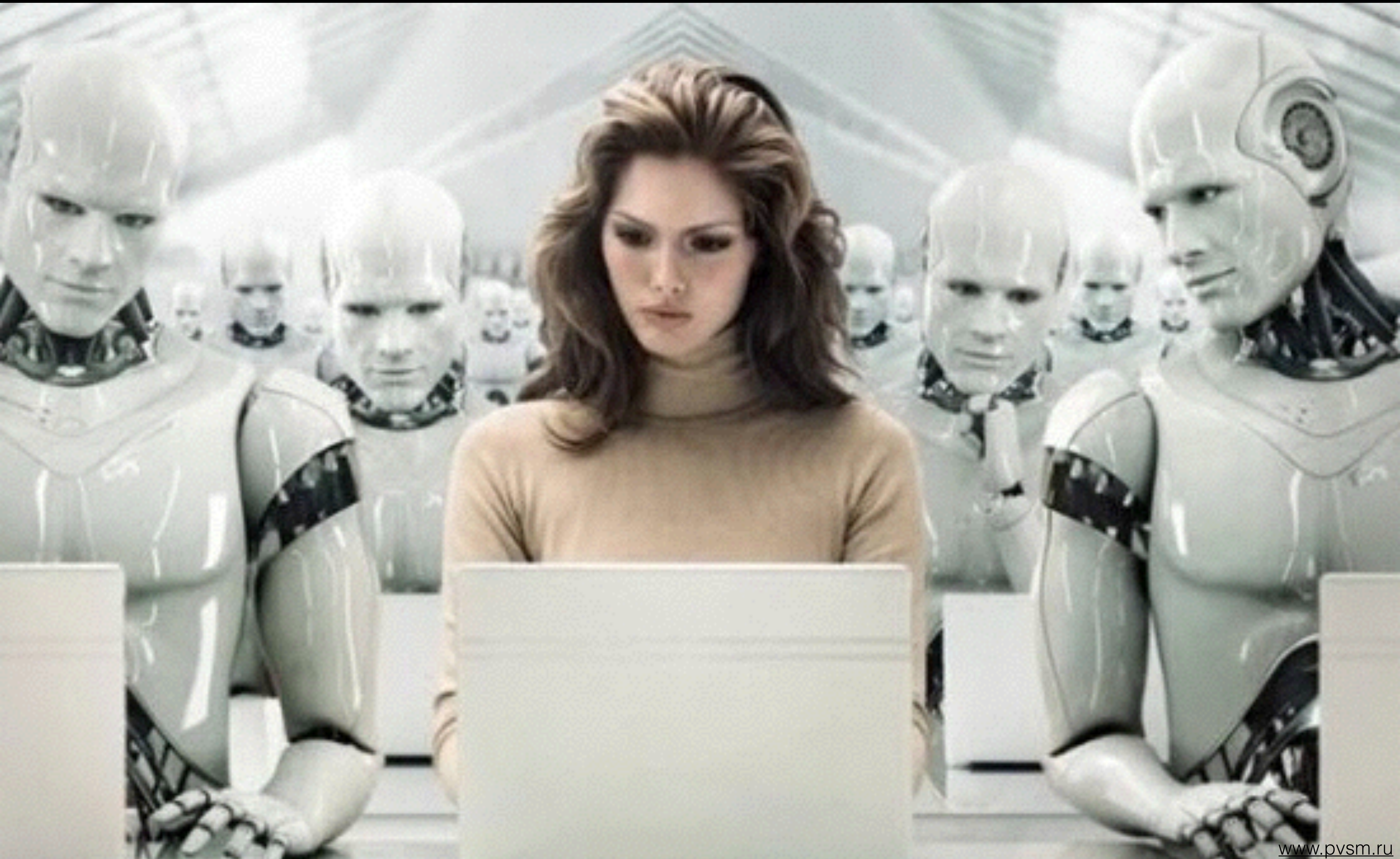
Date	Session	Topics	Assignment
9/12-13	Session 1	Introduction to Urban Data Science. Basic Machine Learning concepts	
9/26-27	Session 2	Single-attribute linear regression and its applications	Homework 1
9/29	Session 3	Lab practicum session – handling urban data in python, performing basic regression analysis and visualizations	
10/3-4	Session 4	Multivariate linear regression. Multicollinearity and overfitting	Homework 2
10/10-11	Session 5	Regression diagnostics and hypothesis testing. Confidence intervals	
10/13	Session 6	Presentations and discussion of ADS project ideas	Homework 3
10/17-18	Session 7	Dealing with multicollinearity and overfitting. Dimensionality reduction through Principle Component Analysis	
10/24-25	Session 8	Unsupervised learning: clustering techniques. K-means, k-medians. Classification through Logistic regression. Max-likelihood estimate (theory only)	Homework 4
10/31-11/1	Session 9	Clustering and classification lab session. Midterm quiz	***Midterm quiz***
11/07-08	Session 10	Introduction to Bayesian Inference. Linear regression revisited	Homework 5
11/14-15	Session 11	Introduction to Network Analysis-I	Homework 6
11/21-22	no classes	Thanksgiving break	
11/28-29	Session 12	Introduction to Network Analysis-II	
12/04-05	Session 13	Spatial regression	
12/11-12	Session 14	Project final presentations	

METHODOLOGY - MACHINE LEARNING

Why machine learning?

Opportunity	Computerization&Digitalization Big data
Challenge	Increasing complexity, size, dynamics, need for scalability
Promise	Multiple application revolutionizing computer science and the world

Can machines learn like humans?



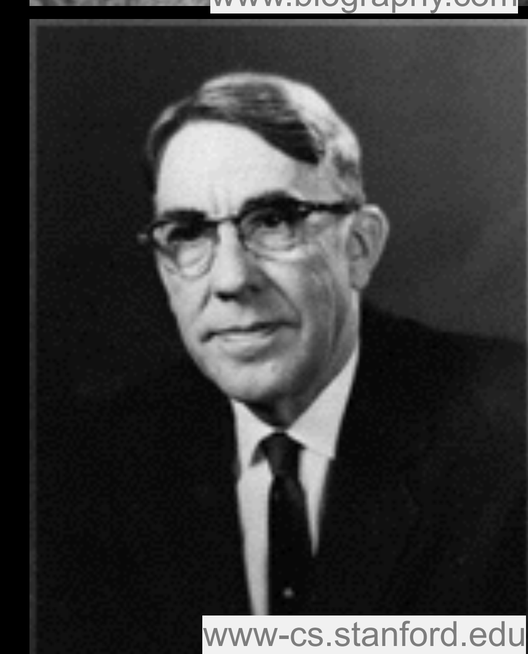
Machine learning vs programming

Ada Lovelace (1815-1852)
1842 - first program

"The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform. It can follow analysis; but it has no power of anticipating any analytical relations or truths."

1950 Alan Turing
"Computing Machinery and Intelligence":

I propose to consider the question, "Can machines think?
We may hope that machines will eventually compete with men in all purely intellectual fields



Arthur Samuel (1959):

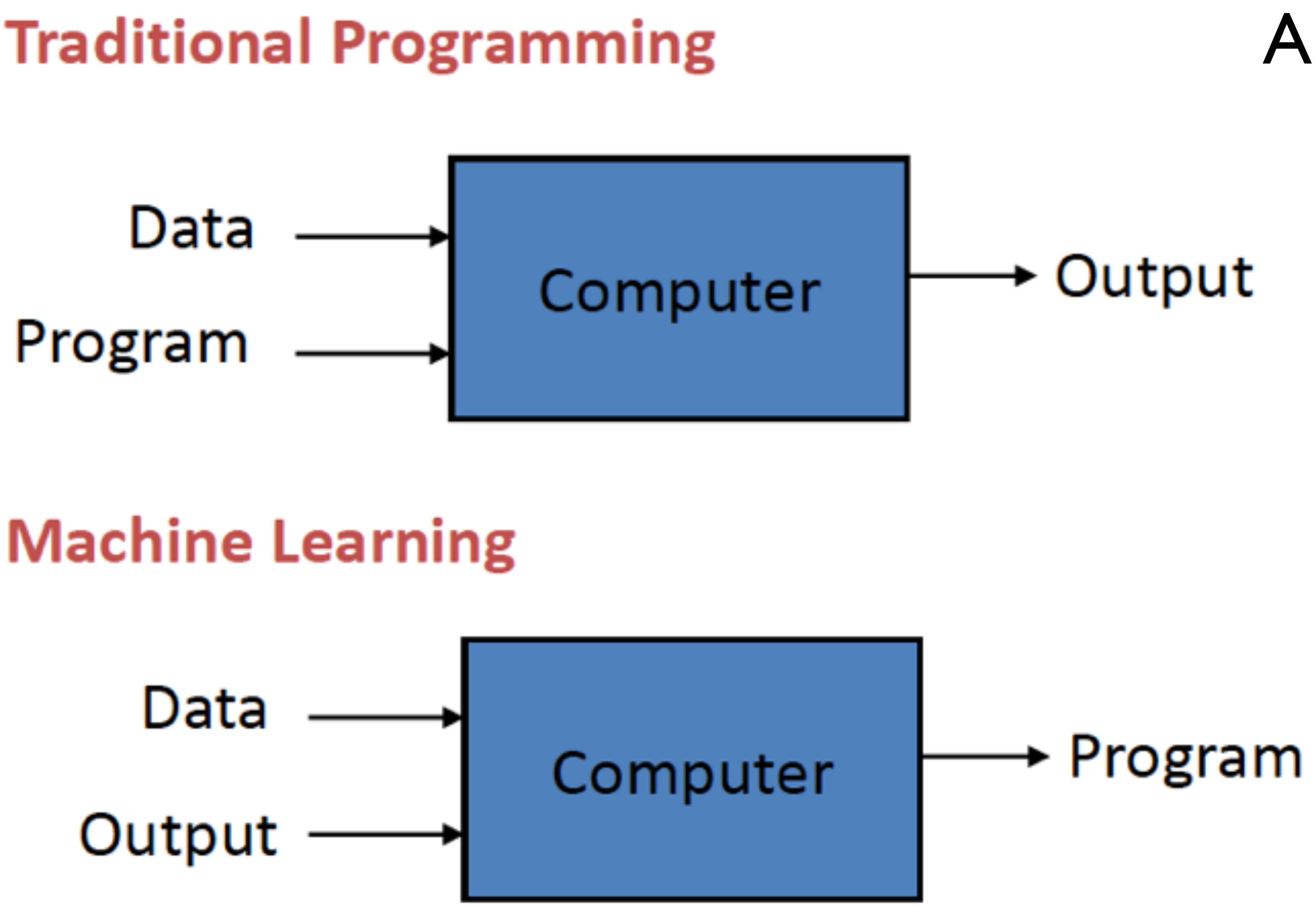
field of study that gives computer the ability to learn without being explicitly programmed (checkers program – learned to play better than him)

Arthur Samuel’s checker game experiment

1949-1960

IBM's Poughkeepsie Laboratory, working on 701 computer

What is machine learning?



Automating automation

No, more like gardening

- **Seeds** = Algorithms
- **Nutrients** = Data
- **Gardener** = You
- **Plants** = Programs

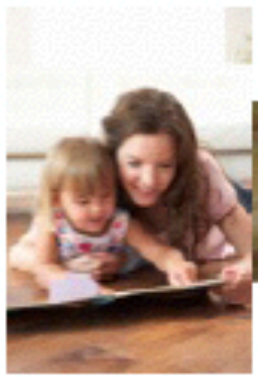
http://www.computerhope.com

Learning like humans

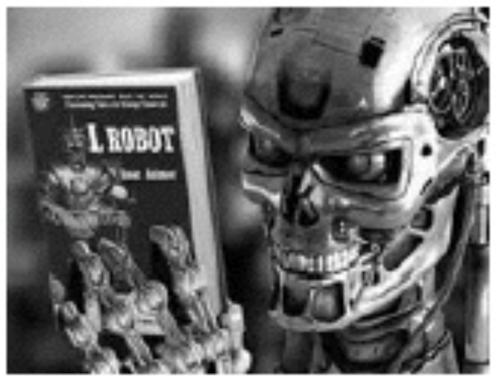
Learning as a natural function of humans and leaving species. How to make machines do the same?

Human vs Machine Learning

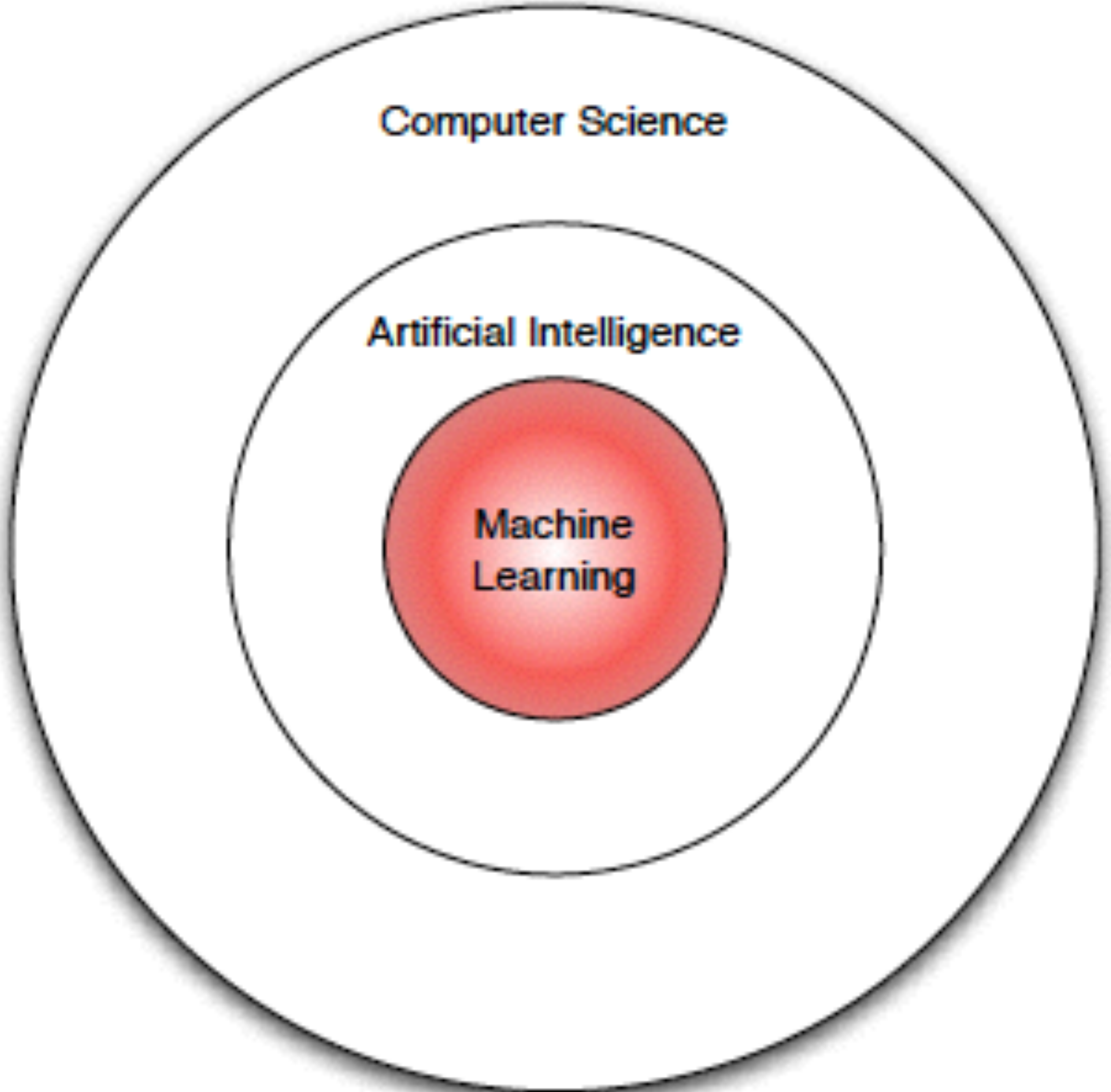
Dr. Theo Damoulas's slides, University of Warwick



How do humans learn…?



- **Supervisory** role?
- **Unsupervised** [grouping, similarity, patterns]
- Internal **reward** system [dopamine]
- Neural structure [Hebbian learning]
- Classical Conditioning (Pavlov's dog)



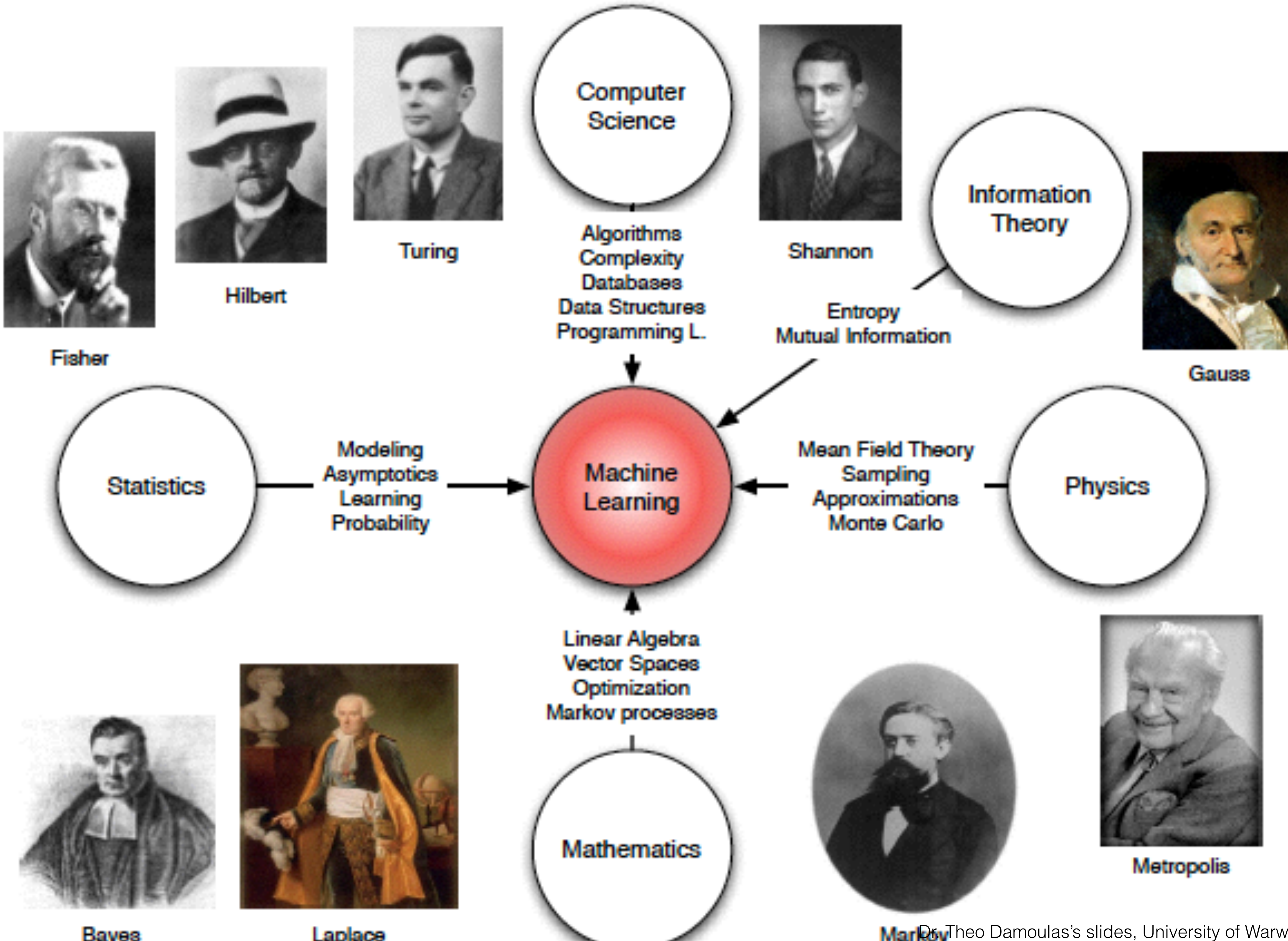
Dr. Theo Damoulas's slides, University of Warwick

Artificial intelligence

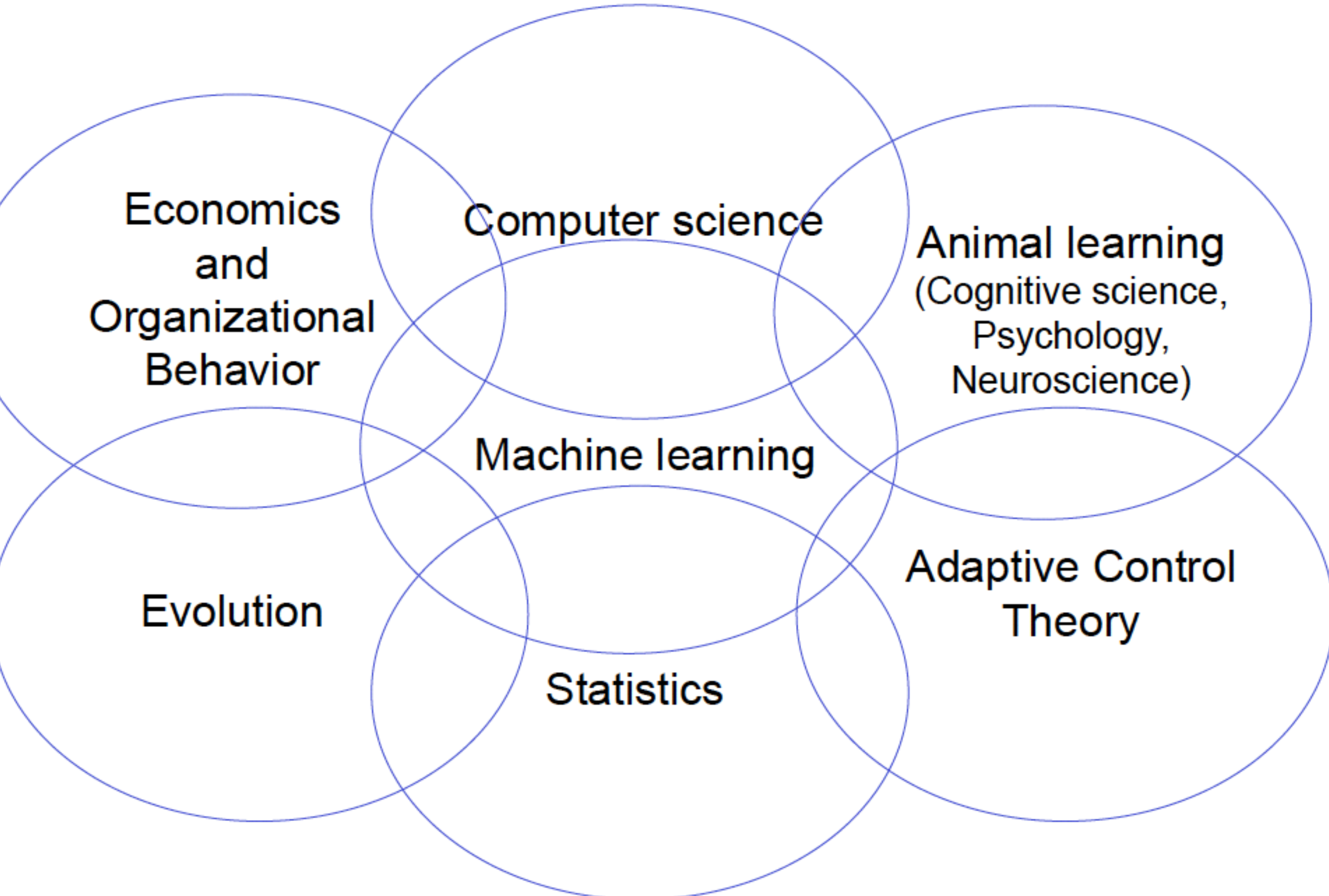
Artificial intelligence is the branch of computer science concerned with making computers behave like humans. The term was coined in 1956 by John McCarthy at the Massachusetts Institute of Technology.
http://www.webopedia.com

The theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.

What fields make up Machine Learning?



Dr. Theo Damoulas's slides, University of Warwick



Prof. T. Mitchell's slides/Carnegie-Mellon

Definition: Machine Learning!

- T. Mitchell: Well posed machine learning
  - Improving performance via experience
  - Formally, A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, it its performance at tasks in T as measured by P, improves with experience.
- H. Simon
  - Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the task or tasks drawn from the same population more efficiently and more effectively the next time.

The ability to perform a task in a situation which has never been encountered before (Learning = Generalization)

Examples of tasks and performance measures

Task	Performance metric	Experience
Play checkers	Percentage of wins against given opponent	Games previously played with outcomes
Recognize hand written digits	Percentage of correct recognitions	Set of digit writing with labels
Control a self-driving car	Average speed in given conditions provided that safety standards are met	Previous driving record and its evaluation
Predict stock prices	Average prediction accuracy	History of stock prices

What is machine learning?

Machine learning addresses the question of how to build computers that improve automatically through experience. It is one of today's most rapidly growing technical fields, lying at the intersection of computer science and statistics, and at the core of artificial intelligence and data science. Recent progress in machine learning has been driven both by the development of new learning algorithms and theory and by the ongoing explosion in the availability of online data and low-cost computation. The adoption of data-intensive machine-learning methods can be found throughout science, technology and commerce, leading to more evidence-based decision-making across many walks of life, including health care, manufacturing, education, financial modeling, policing, and marketing.

M. I. Jordan, T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. Science. 2015

Machine learning as optimization problem

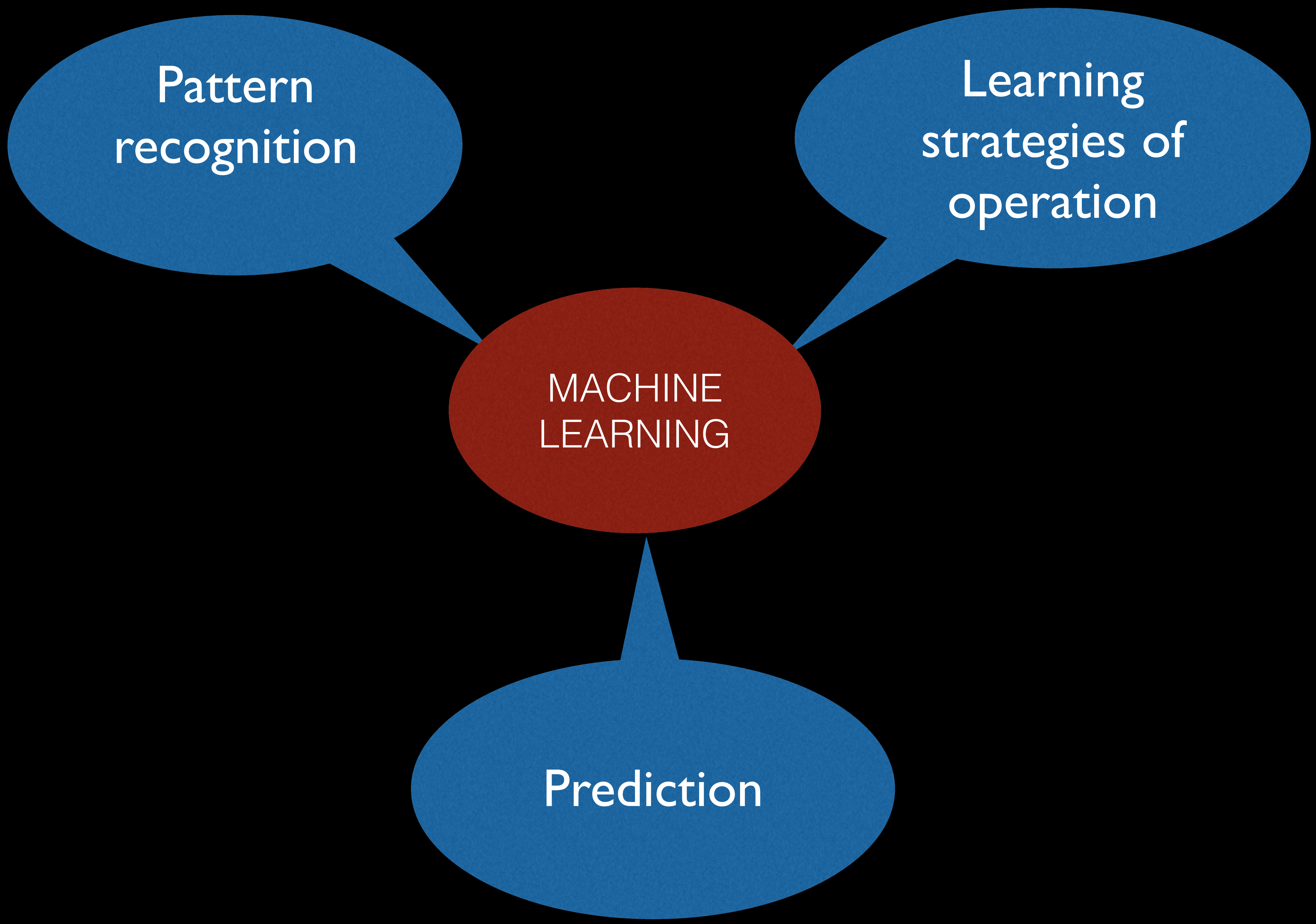
1. Select performance metric and dataset to evaluate it

2. Pick up a machine learning model depending on the unknown parameters to learn as well as the dataset to train it

3. Look for the set of model parameters that optimize the given performance metric

4. Evaluate different models and finally pick up the best one

Objectives of learning



Some quotes

”Machine learning is the hot new thing” (John Hennessy, President, Stanford)

“A breakthrough in machine learning would be worth ten Microsofts” (Bill Gates, Microsoft)

“Machine learning is the next Internet” (Tony Tether, Former Director, DARPA)

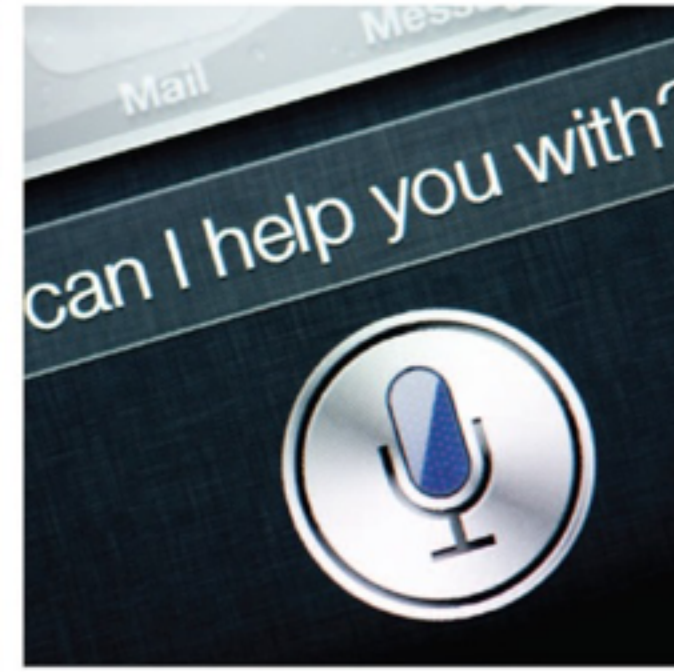
“Web rankings today are mostly a matter of machine learning” (Prabhakar Raghavan, Dir. Research, Yahoo)

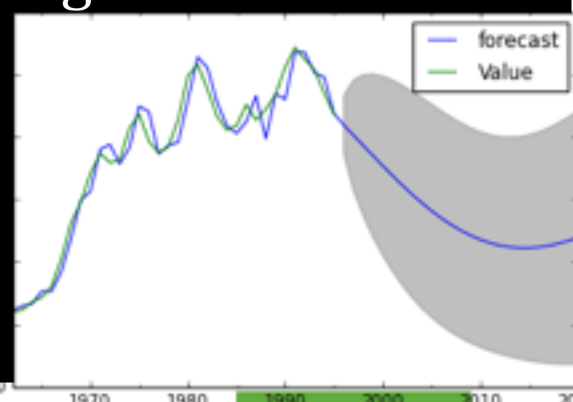
“Machine learning is going to result in a real revolution” (Greg Papadopoulos, CTO, Sun)

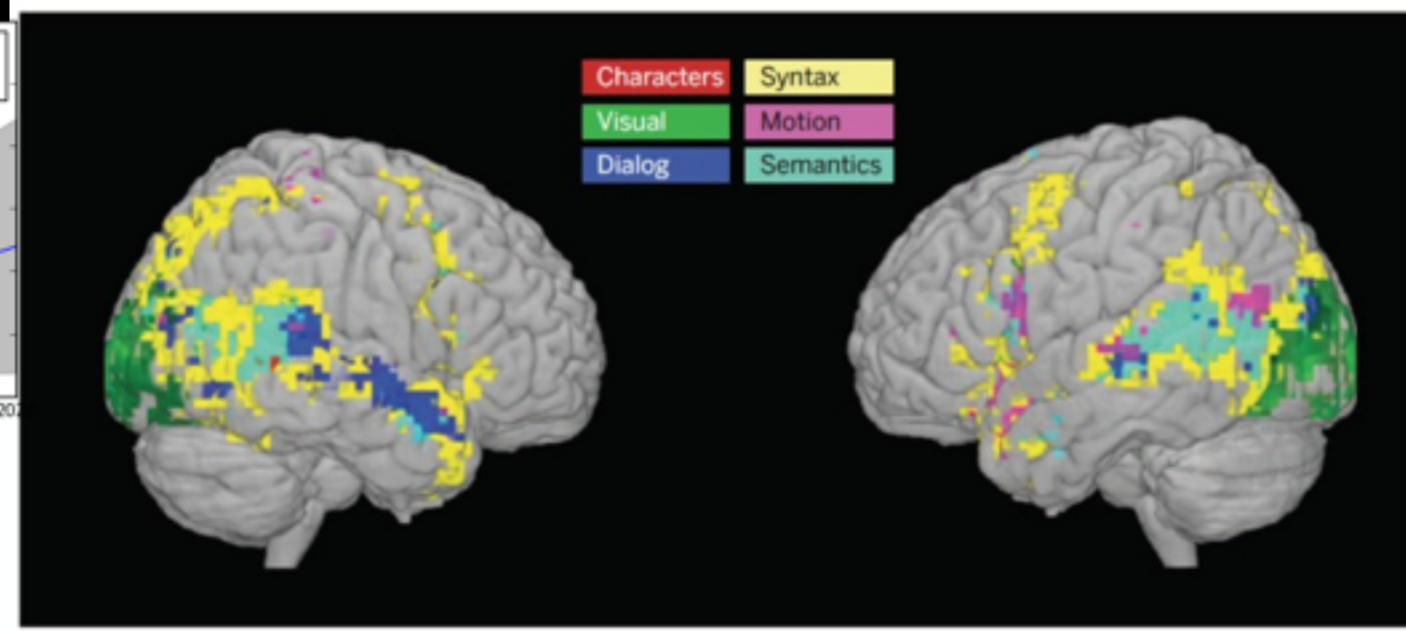
APPLICATION EXAMPLES

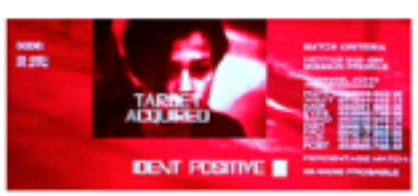
What can machine learning do?

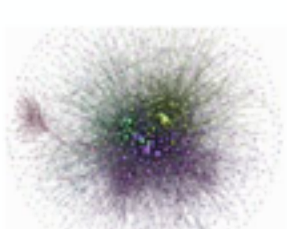
Control
Voice recognition
Image recognition
Recommendation systems
Natural language processing
Prediction

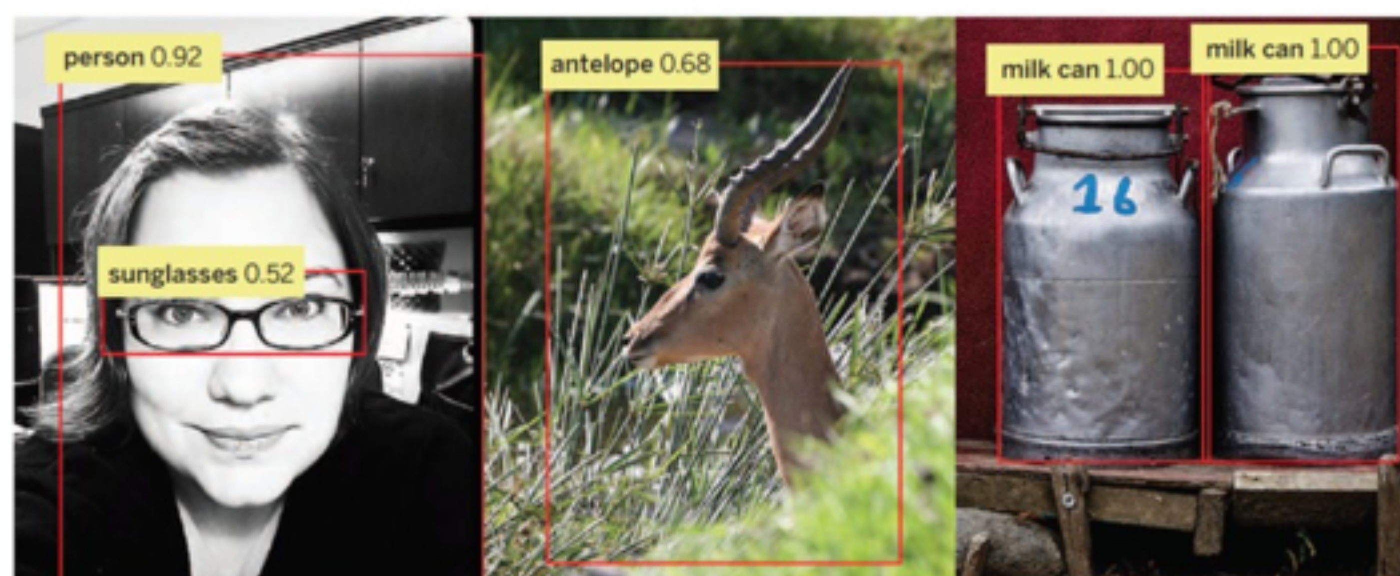


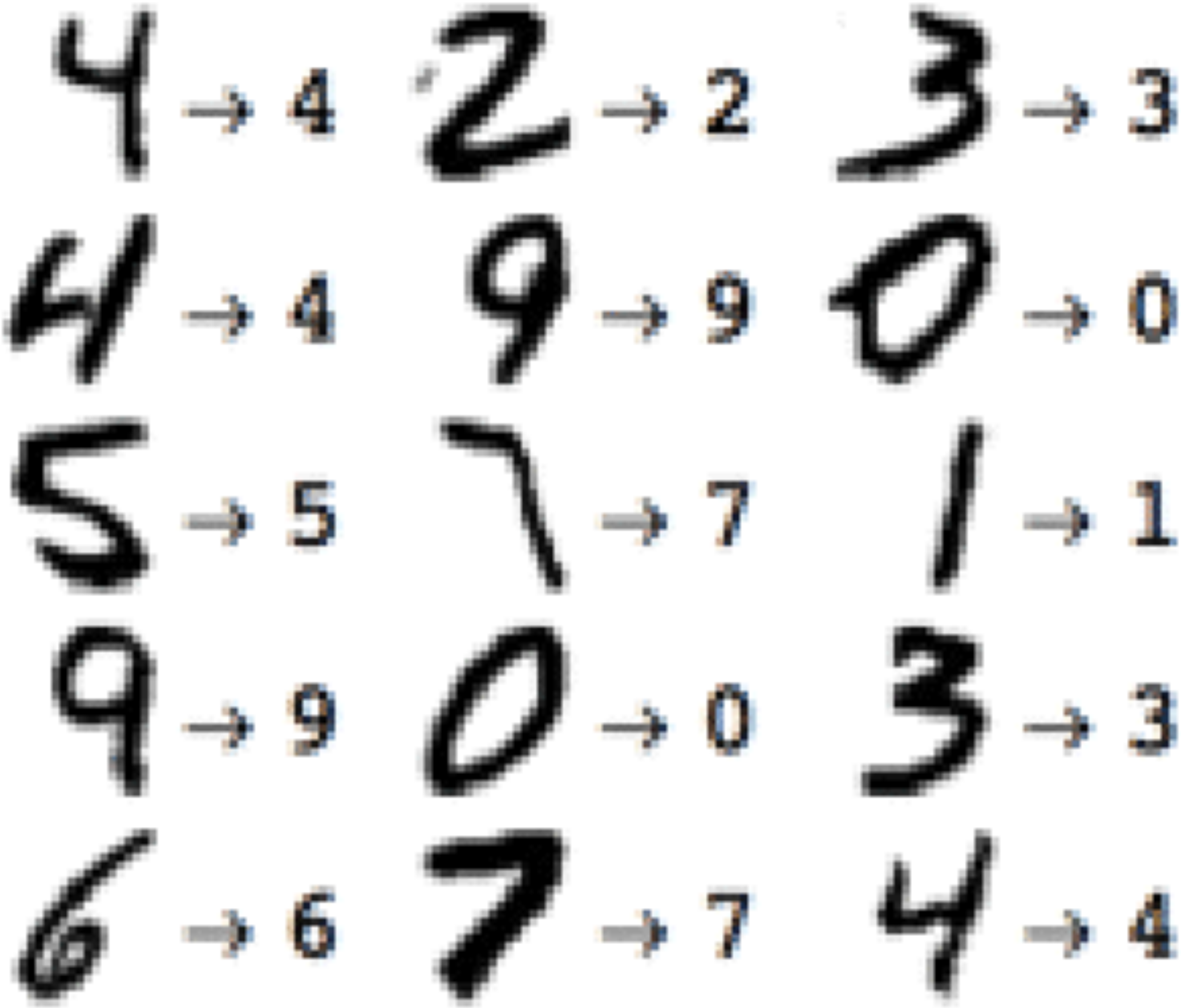




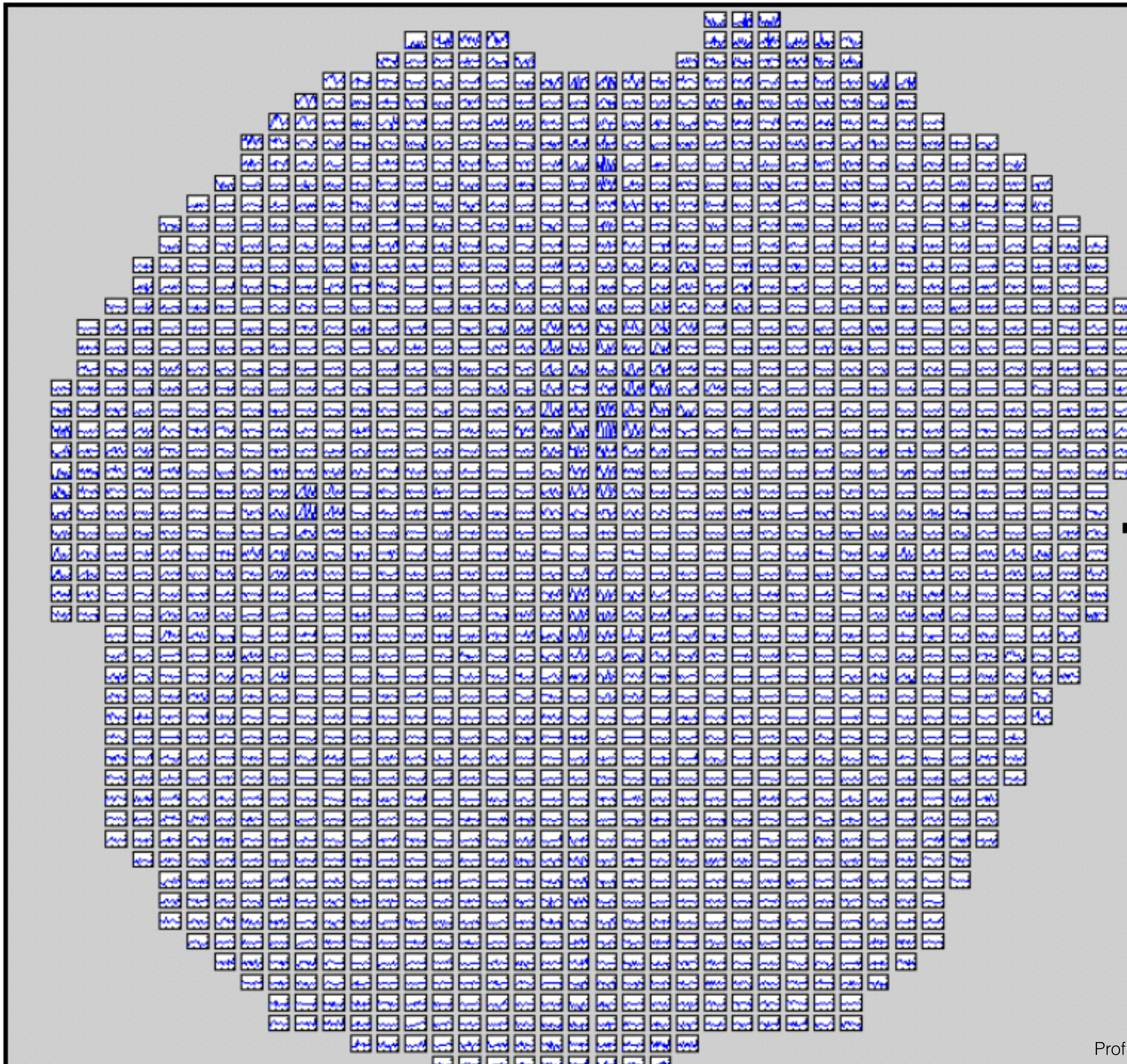








wolfram.com



Reading a noun (vs verb)

[Rustandi et al., 2005]

Prof. T. Mitchell's slides/Carnegie-Mellon

→ Company home page

vs

Personal home page

vs

University home page

vs

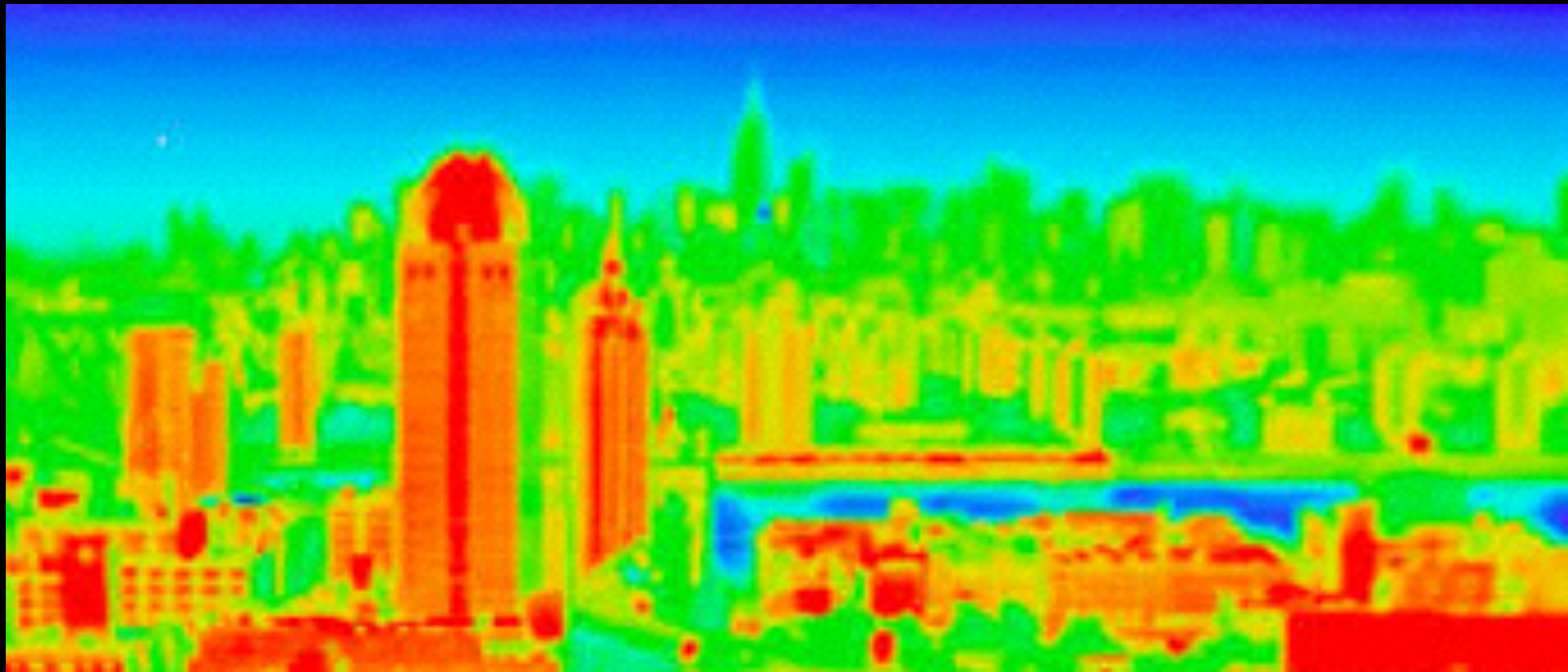
. . .

Prof. T. Mitchell's slides/Carnegie-Mellon

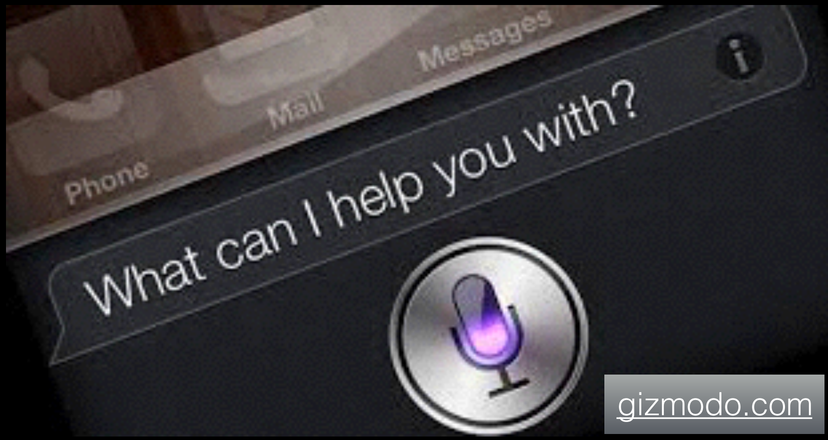
What can machine learning do: Image recognition

- Handwritten digits (zip code reading US mail)
- Reading checks
- Face recognition (security, search)
- Classification of land types by satellite images
- Astronomic images – patters in galactics
- Pattern detection in the infrared image of the city

http://www.fastcoexist.com/3037771/a-secret-urban-observatory-is-snapping-9000-images-a-day-of-new-york-city



What can machine learning do: Voice recognition



- Cocktail party – separate person's voice from a set of different speakers
- street noise
- speech recognition (Siri)

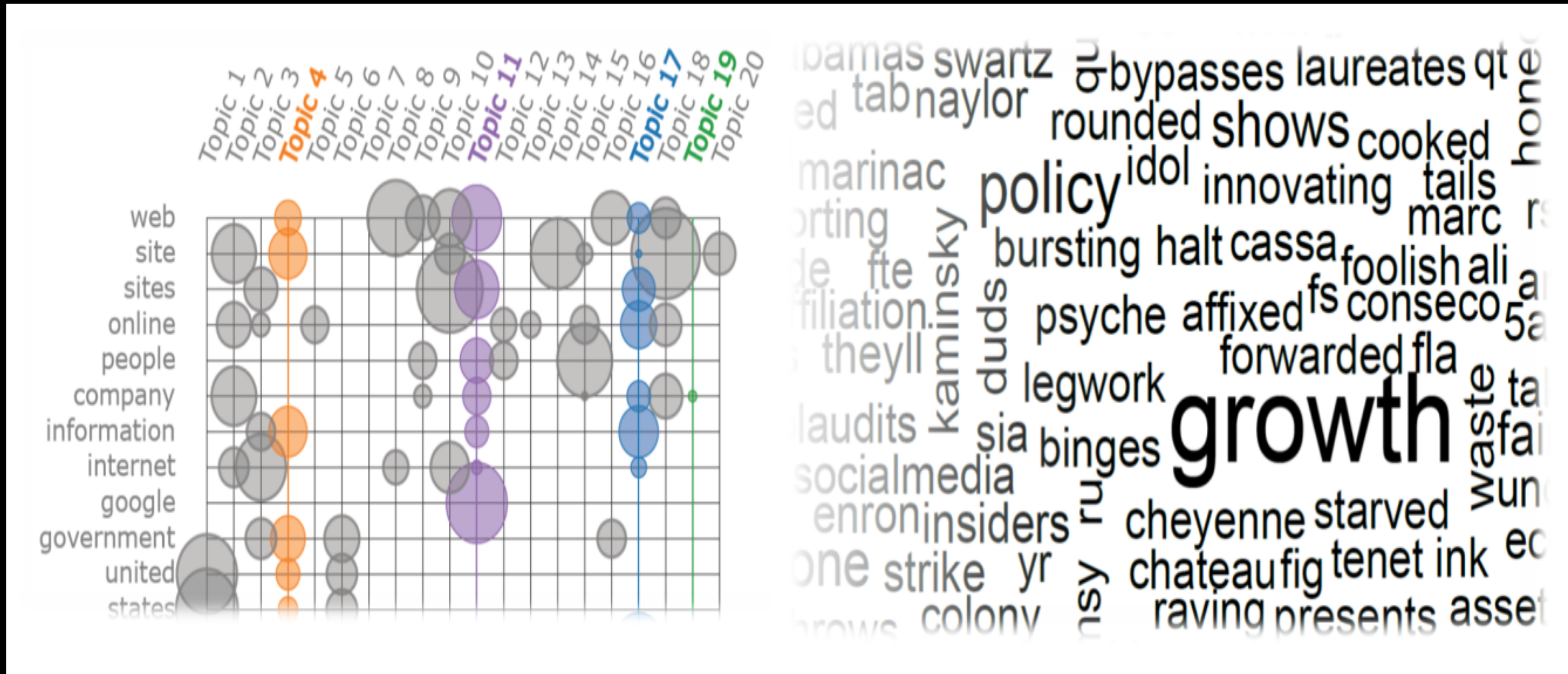
What can machine learning do: Robotics

- Program to fly a helicopter
- Self-driving vehicles
- 3d landscape recognition from the camera - robotics
- Parking airplane (saving Delta $0.5M/day gas)
- Robotics – teach robots move in space, or even off-road driving

What can machine learning do: Recommendation systems

What can machine learning do: Natural language processing

- Sentiment analysis
- Document classification
- Webpage ranking
- Automated translation



What can machine learning do: Prediction

- Weather forecast
- Trading
- Disease development
- Traffic models

What can machine learning do: and much more

- Reconstruct 3d model of the world from 2d pictures
- Social network analysis
- Organizing computer clusters…

What can machine learning do: Data Mining

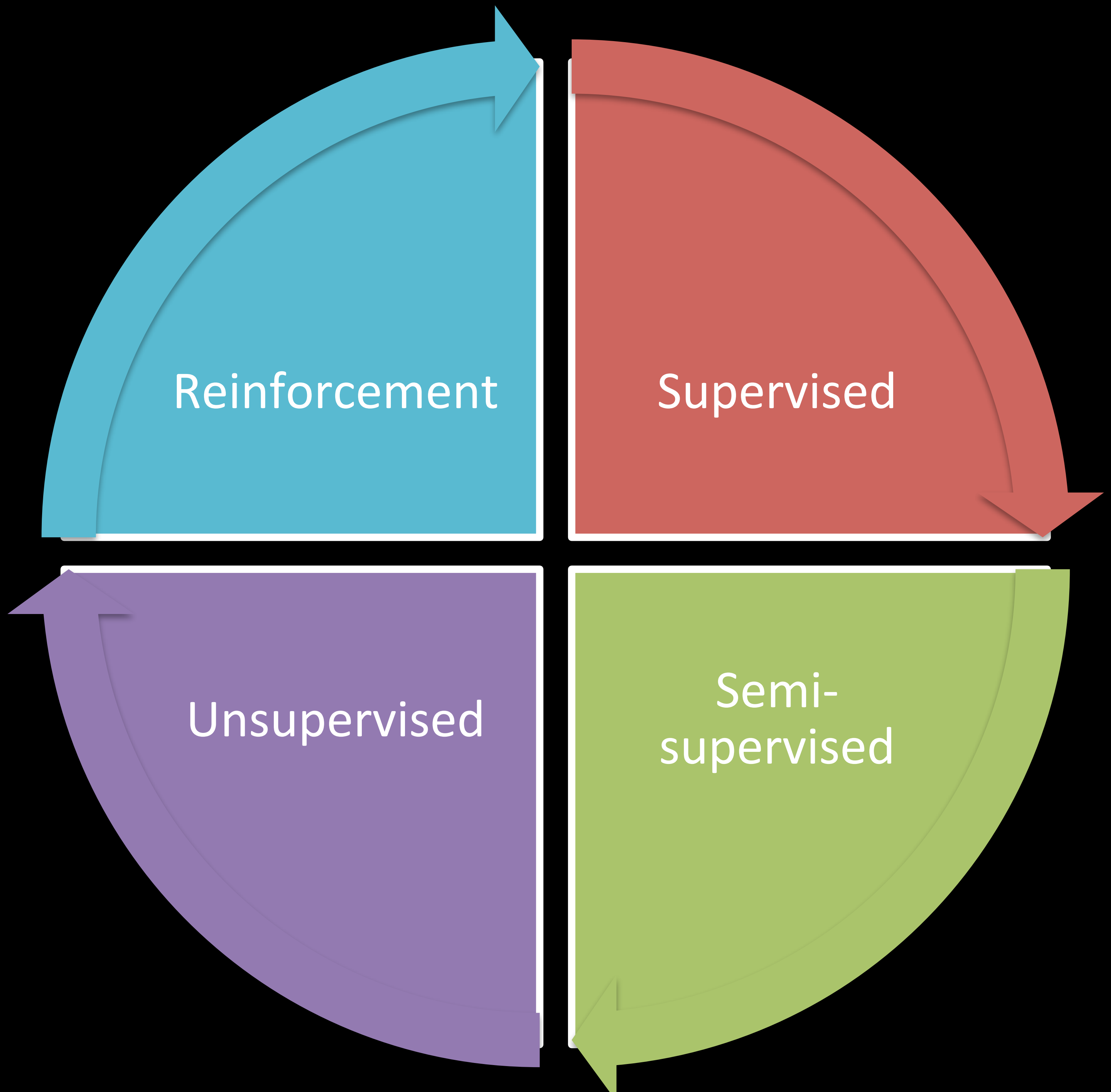
- Medical records -> medical knowledge, detect trends, practices
- Behavioral analysis
- Fraud detection – credit cards

Urban applications

- Inferring urban dynamics
- Energy usage prediction
  http://www.sciencedirect.com/science/article/pii/S1877705815027988
- Traffic prediction
- Economic patterns detection and prediction
- Public health pattern detection and prediction
- Land use classification
- Event detection from urban activity
- Computer vision: pedestrian/traffic counts, security (face recognition), traffic accident detection, remote sensing (air content)
- Street noise decomposition
- Detecting trends from social media (natural language processing)

TYPES OF MACHINE LEARNING APPROACHES

Types of learning



Supervised learning

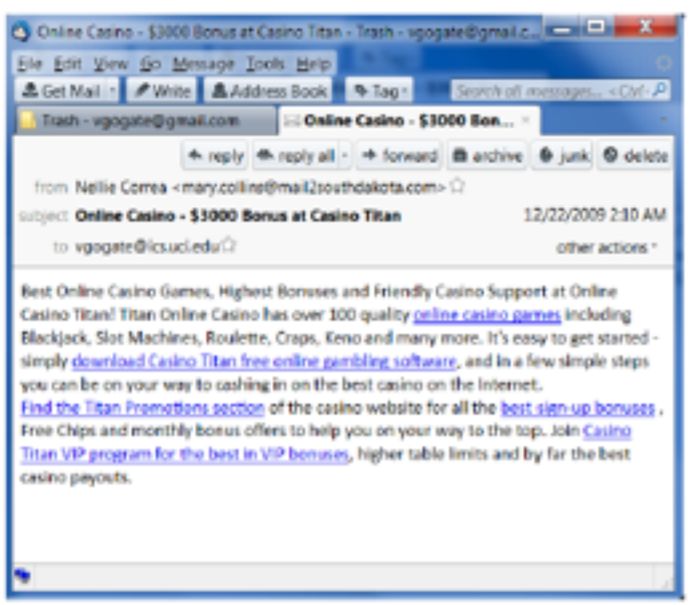
Data/input	Labels/output	Dependence
x_1	y_1	$y = f(x)$
x_2	y_2	
$\ldots$	$\ldots$	
x_N	y_N	

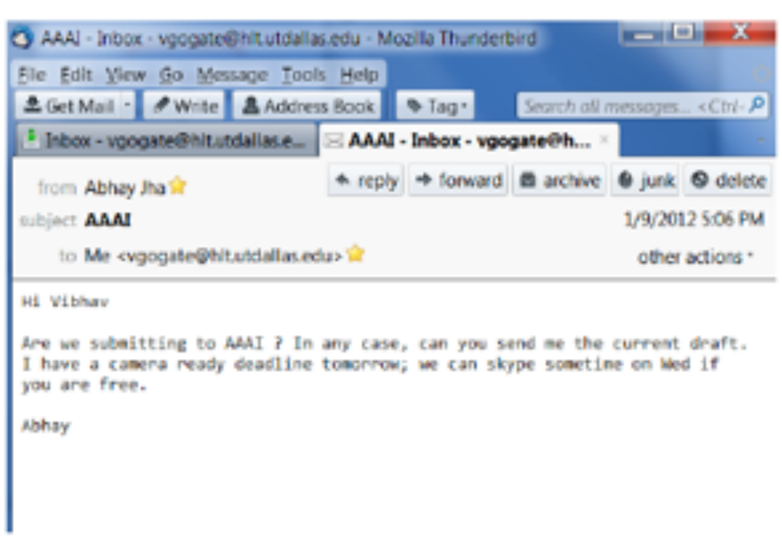
Supervised learning

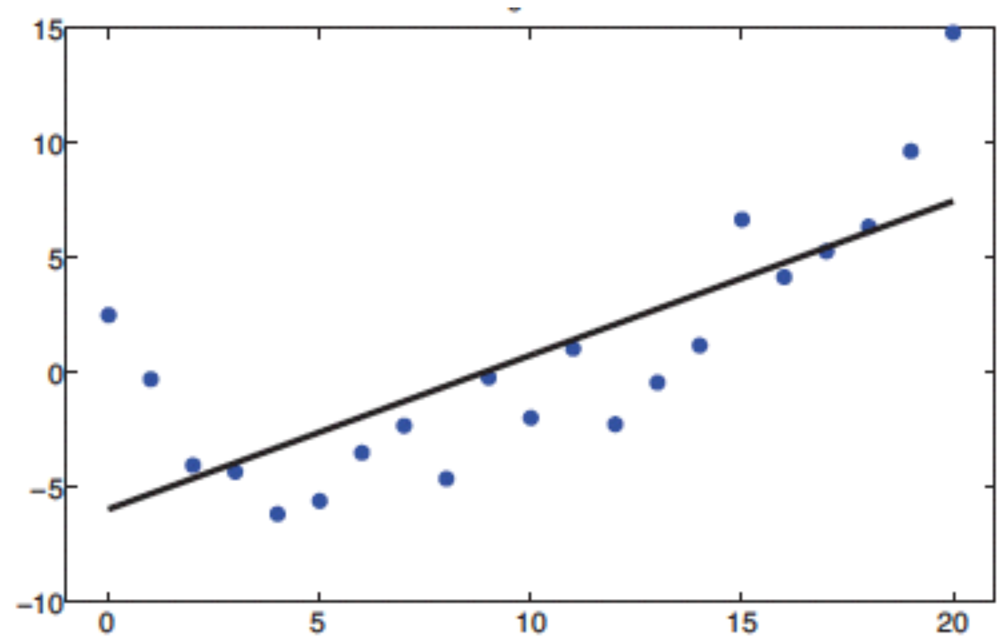
$$y = f(x)$$

Discrete y - classification

Continous y - regression







Murphy, Kevin P. *Machine learning: a probabilistic perspective*. MIT press, 2012.
These materials are included under the fair use exemption and are restricted from further use

Unsupervised learning

Analysis of the unlabeled data under assumptions about structural properties

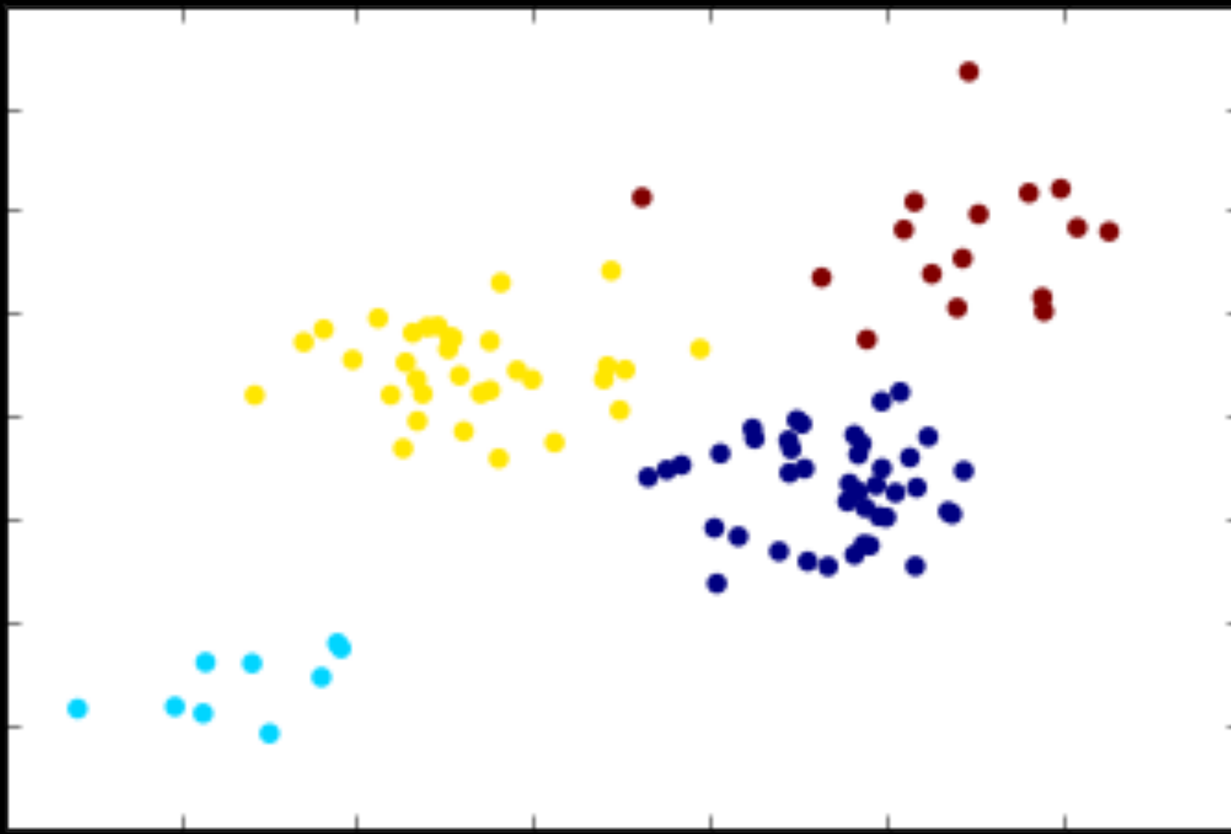
Data → Patterns

x_1

x_2

$\dots$

x_N



Semi-supervised learning

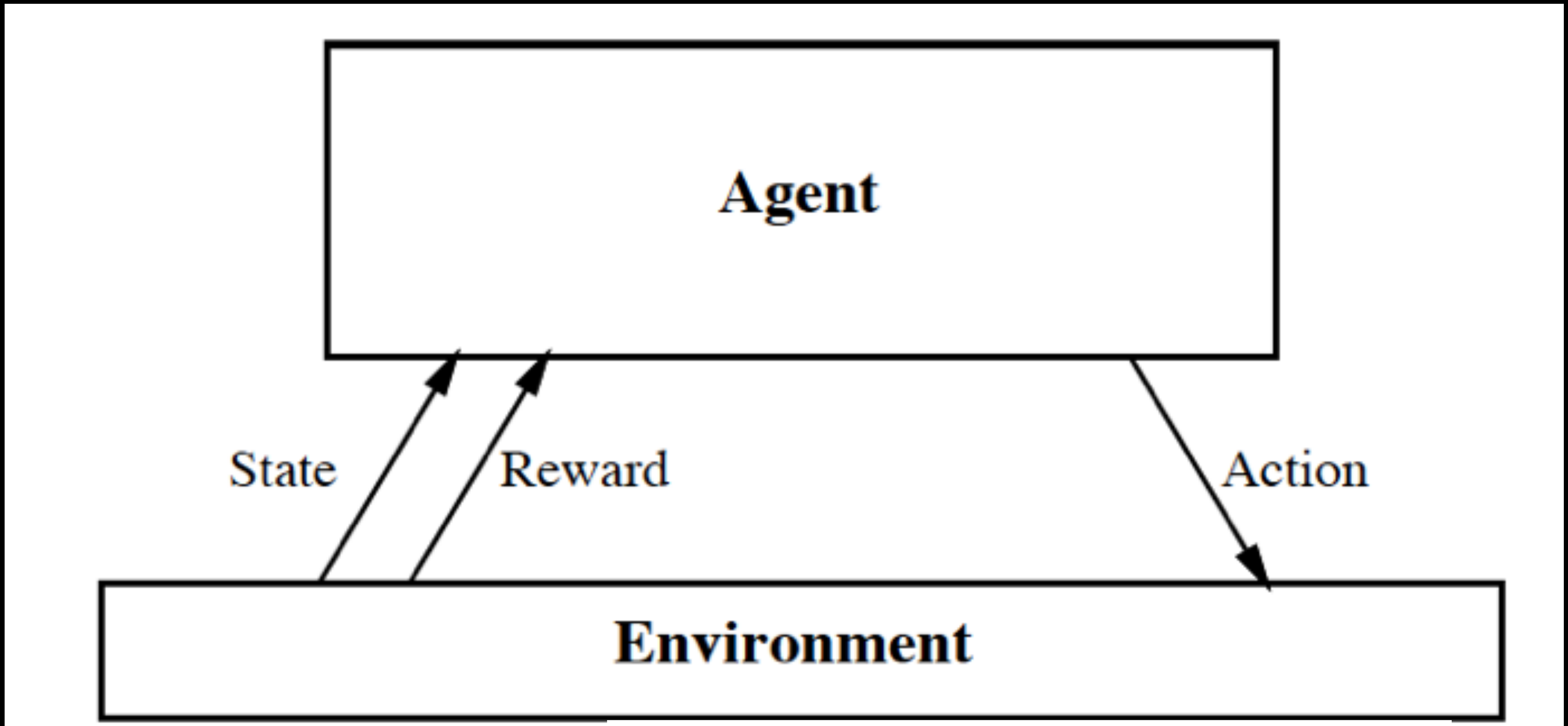
Data	Partial labels	Dependence
x_1	-	$y = f(x)$
x_2	y_2	
$\ldots$	$\ldots$	
x_N	-	

- Unsupervised step (clustering, learning latent factors etc)
- Supervised step

Reinforcement learning

sequential decision-making with delayed reward

- Robotics - learn control of a moving robot (helicopter, self-driving car, vacuum cleaner etc)
- Learn strategy to play a game
- Manufacturing or marketing strategy
- Cell phone network routing
- Web-page indexing
- Multi-agent systems



lecture slides for textbook *Machine Learning*, T. Mitchell,

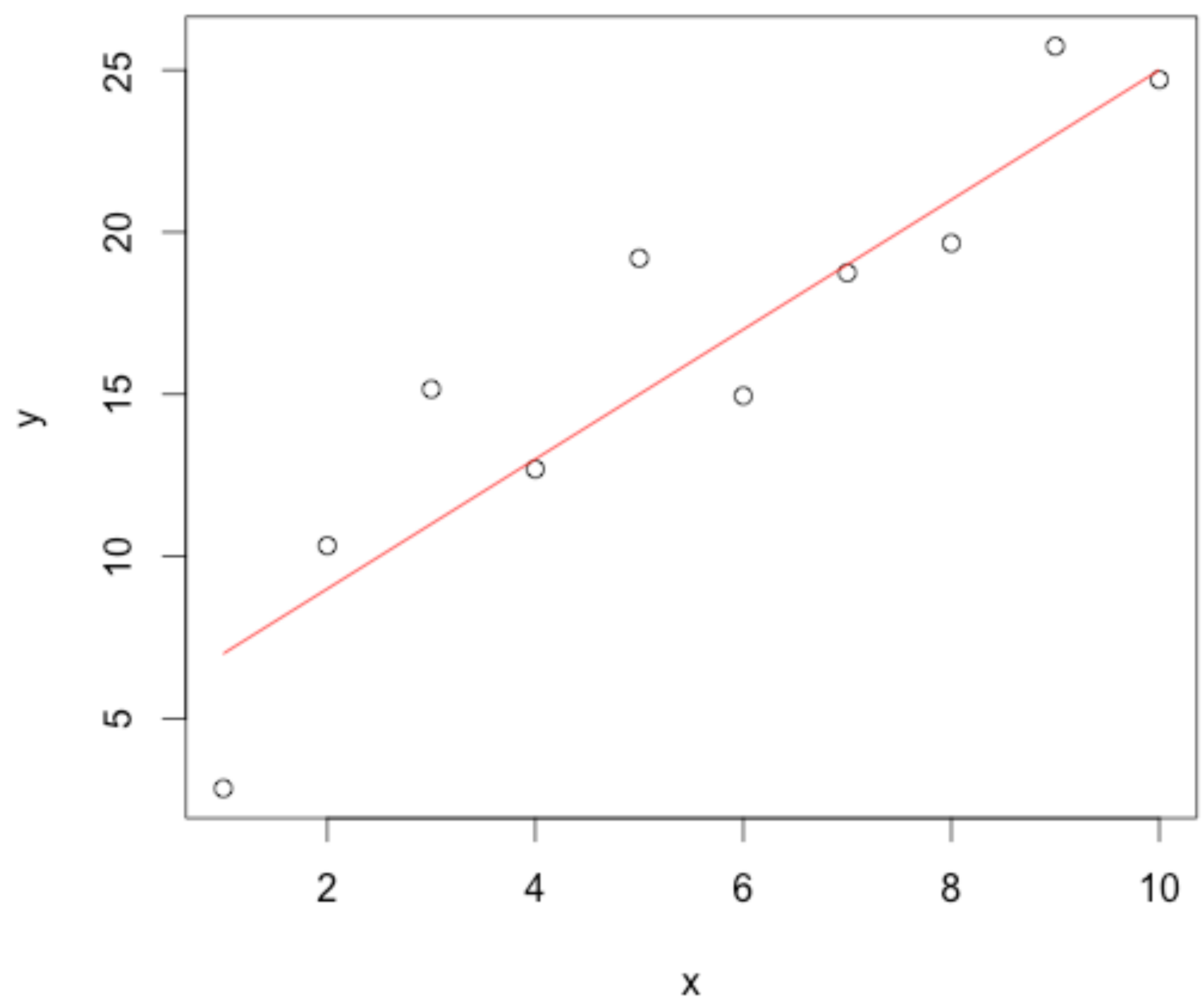
SUPERVISED LEARNING APPROACHES AND EXAMPLES

Linear regression

$$y = w_1 x + w_0 + \varepsilon$$

$$y = w^T x + \varepsilon$$

$$p(y|x, w, \sigma) = \mathcal{N}(y|w^T x, \sigma^2)$$



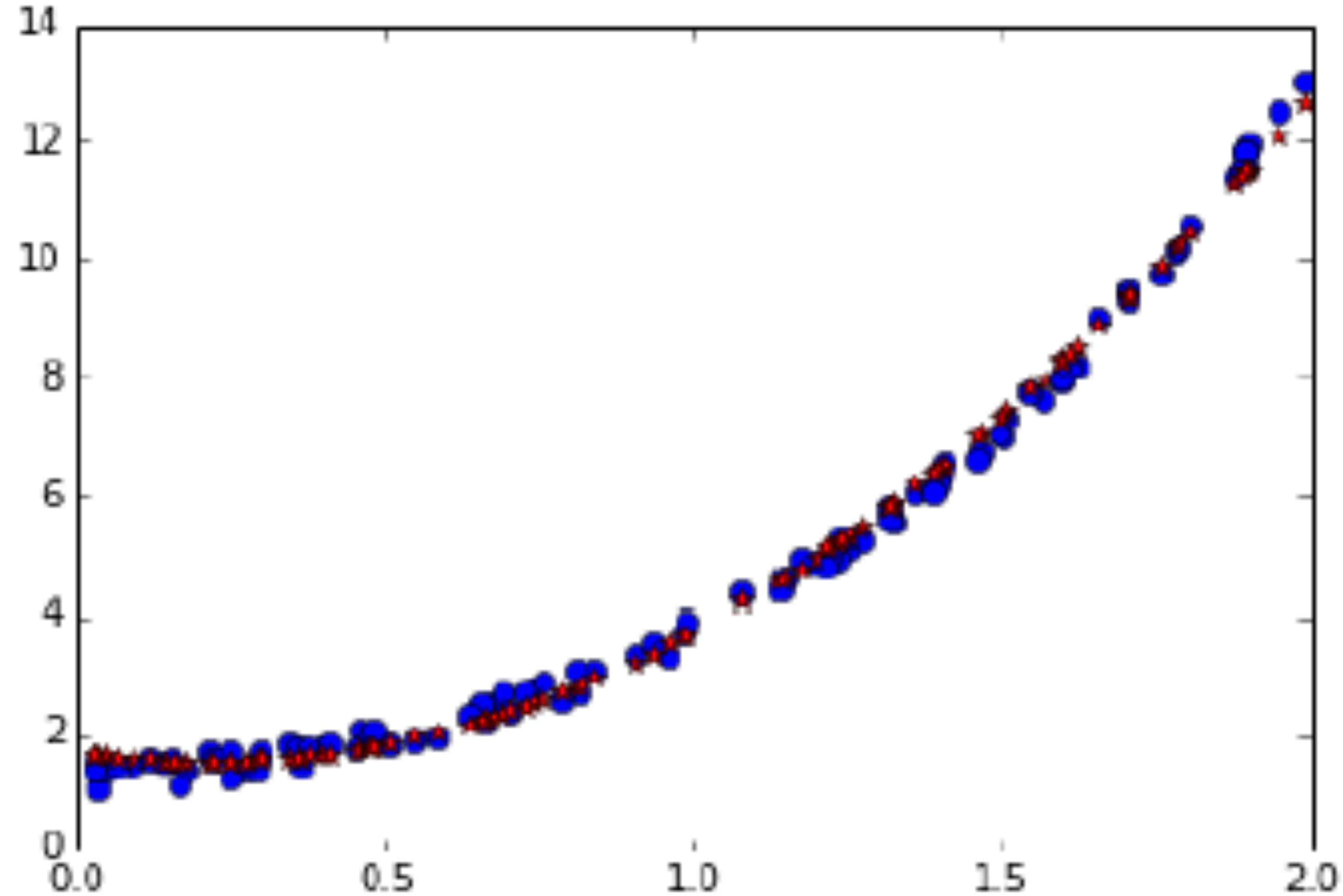
Polynomial Models

$$y = w_m * x^m + w_{m-1}x^{m-1} + \ldots + w_1 x + w_0 + \epsilon$$

$$y \sim 1, x, x^2, \ldots, x^m$$

$$y \sim w_{2,0}x_1^2 + w_{1,1}x_1x_2 + w_{2,0}x_2^2 + w_{1,0}x_1 + w_{0,1}x_2 + w_{0,0}$$

$$y \sim 1, x_1, x_2, x_1^2, x_2^2$$



Generalized linear regression

$$y = f(x\beta)$$

$$P(y|x) = P(y|x\beta)$$

$$y = (x\beta + \gamma)^n$$

$$y = e^{x\beta}$$

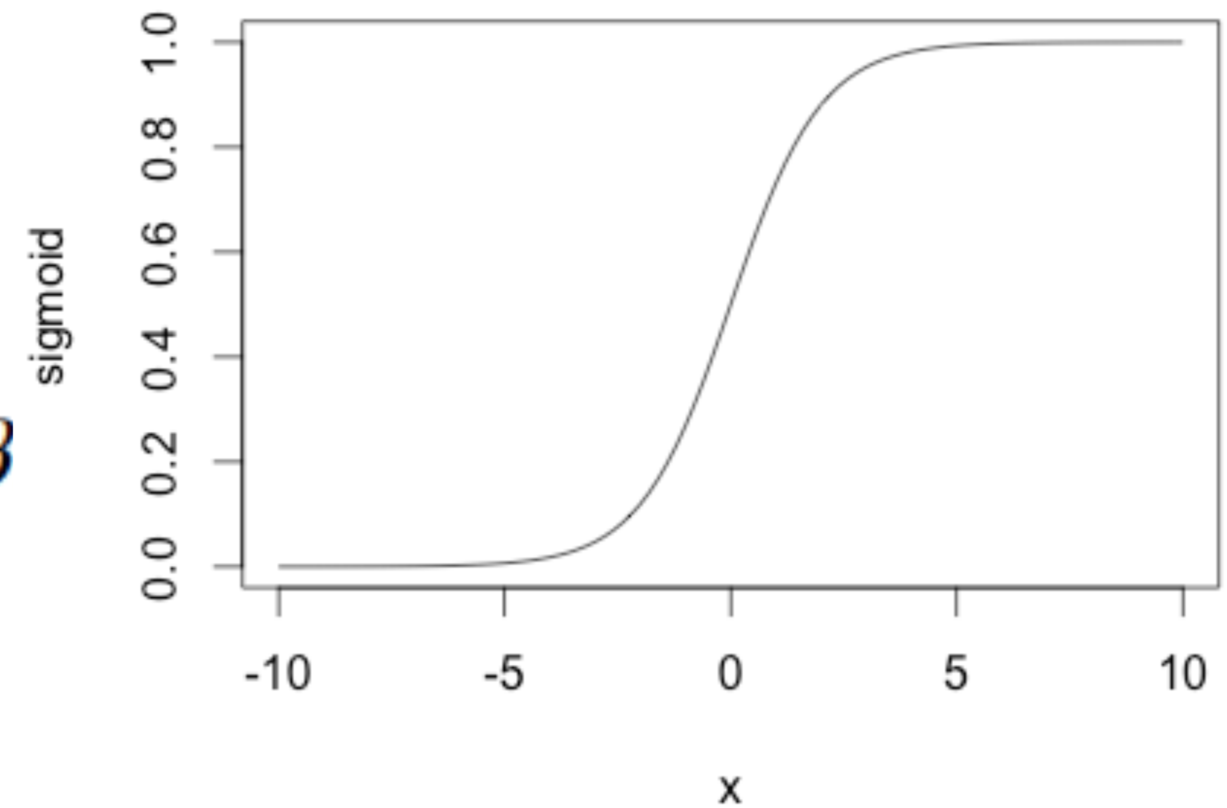
Generalized linear regression. Logistic regression

$$y = f(x\beta)$$

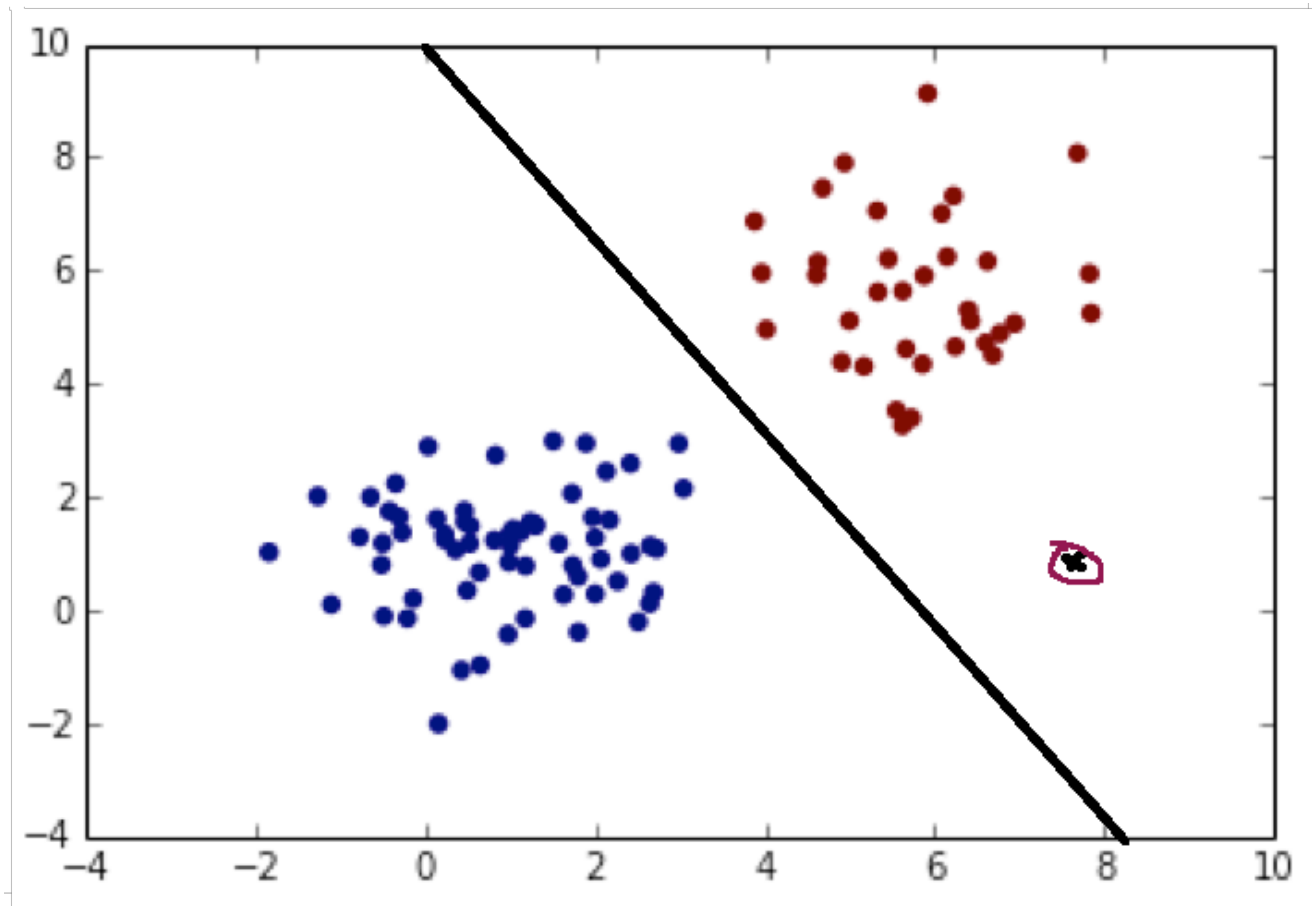
$$P(y|x, \beta) = Bern(y|\mu(x, \beta))$$

$$\mu(x, \beta) = f(x\beta) \qquad \mu(x, \beta) = x\beta$$

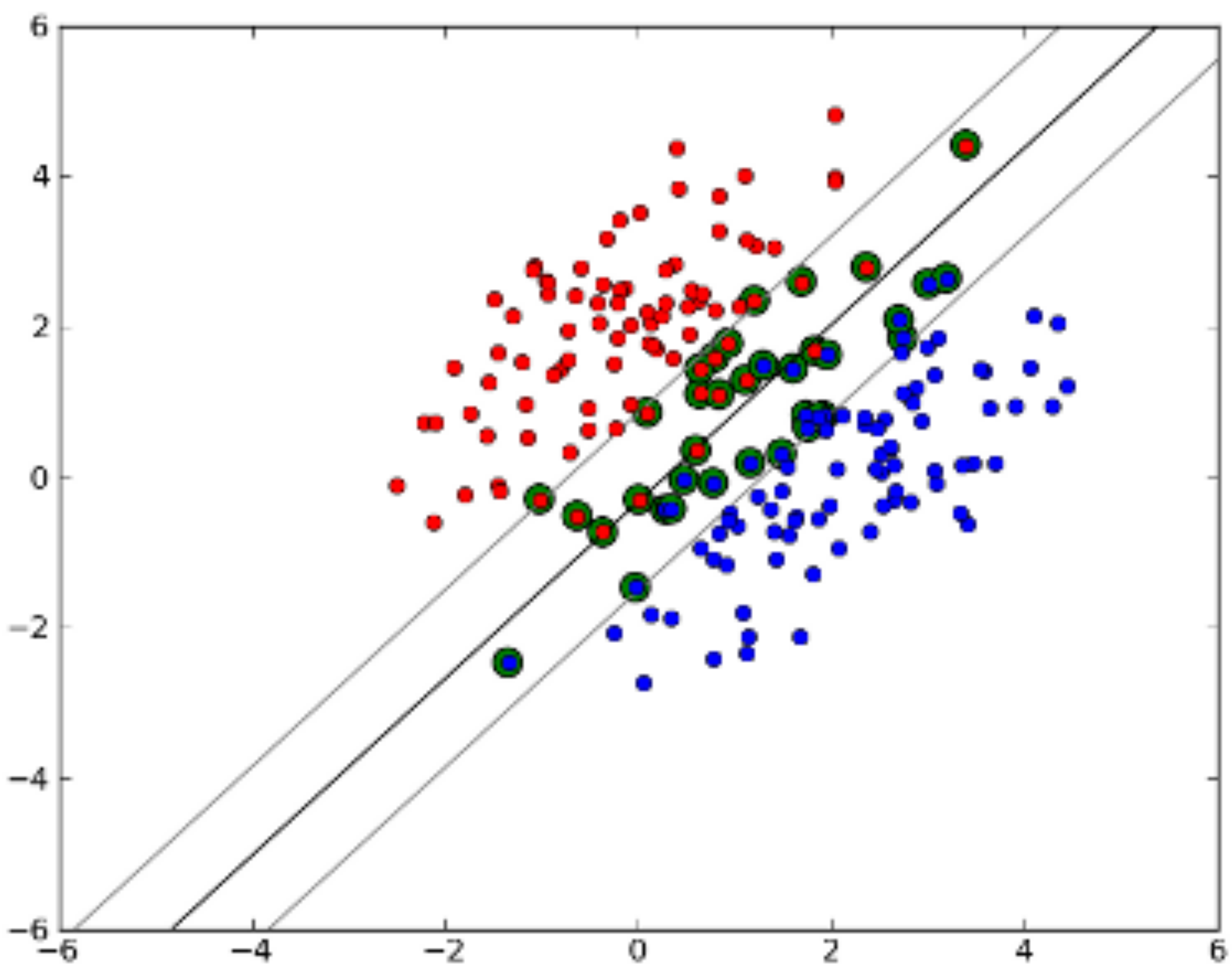
$$f(x) = \sigma(x) = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}}$$



Support Vector Machines

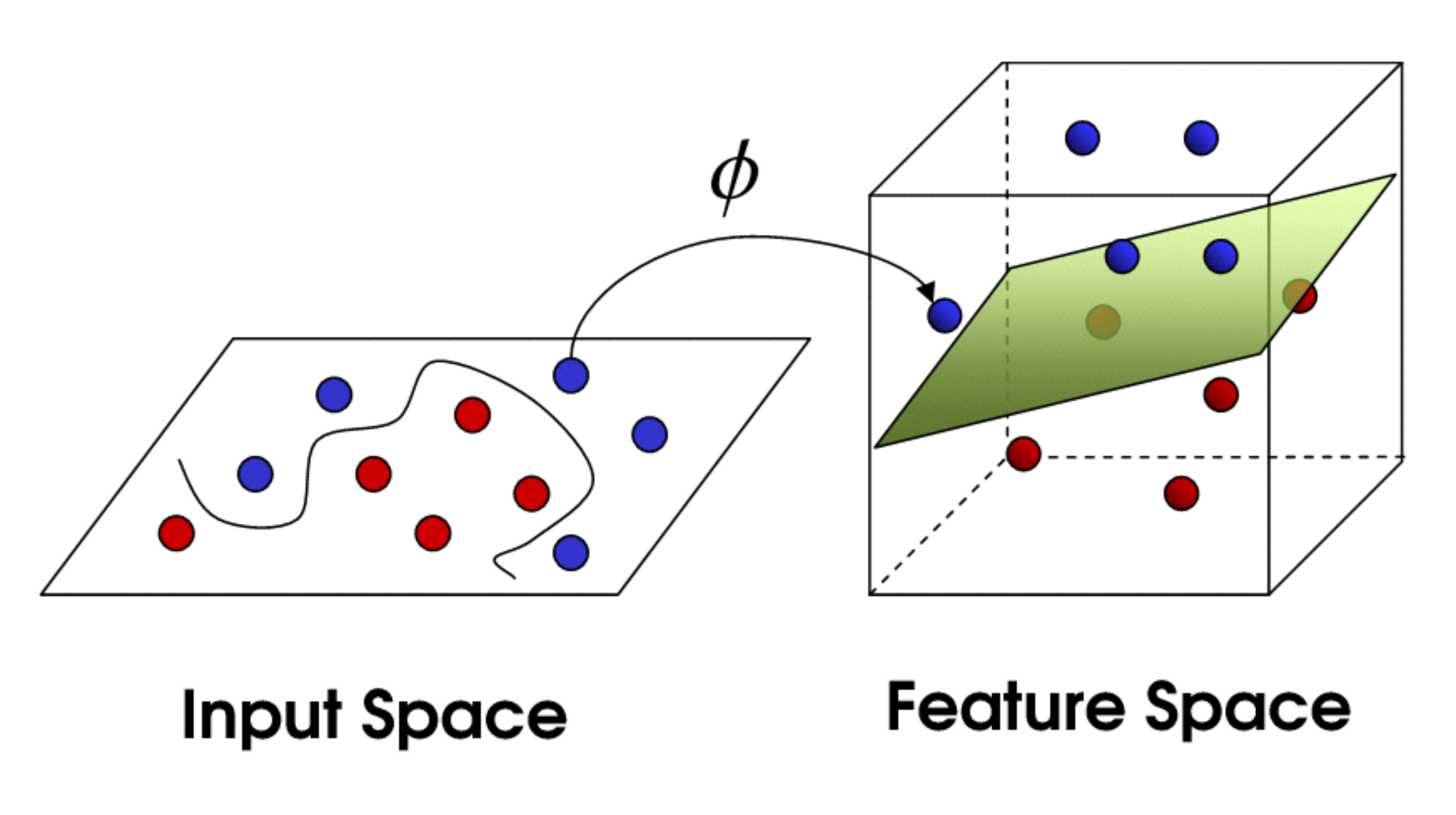


Support Vector Machines

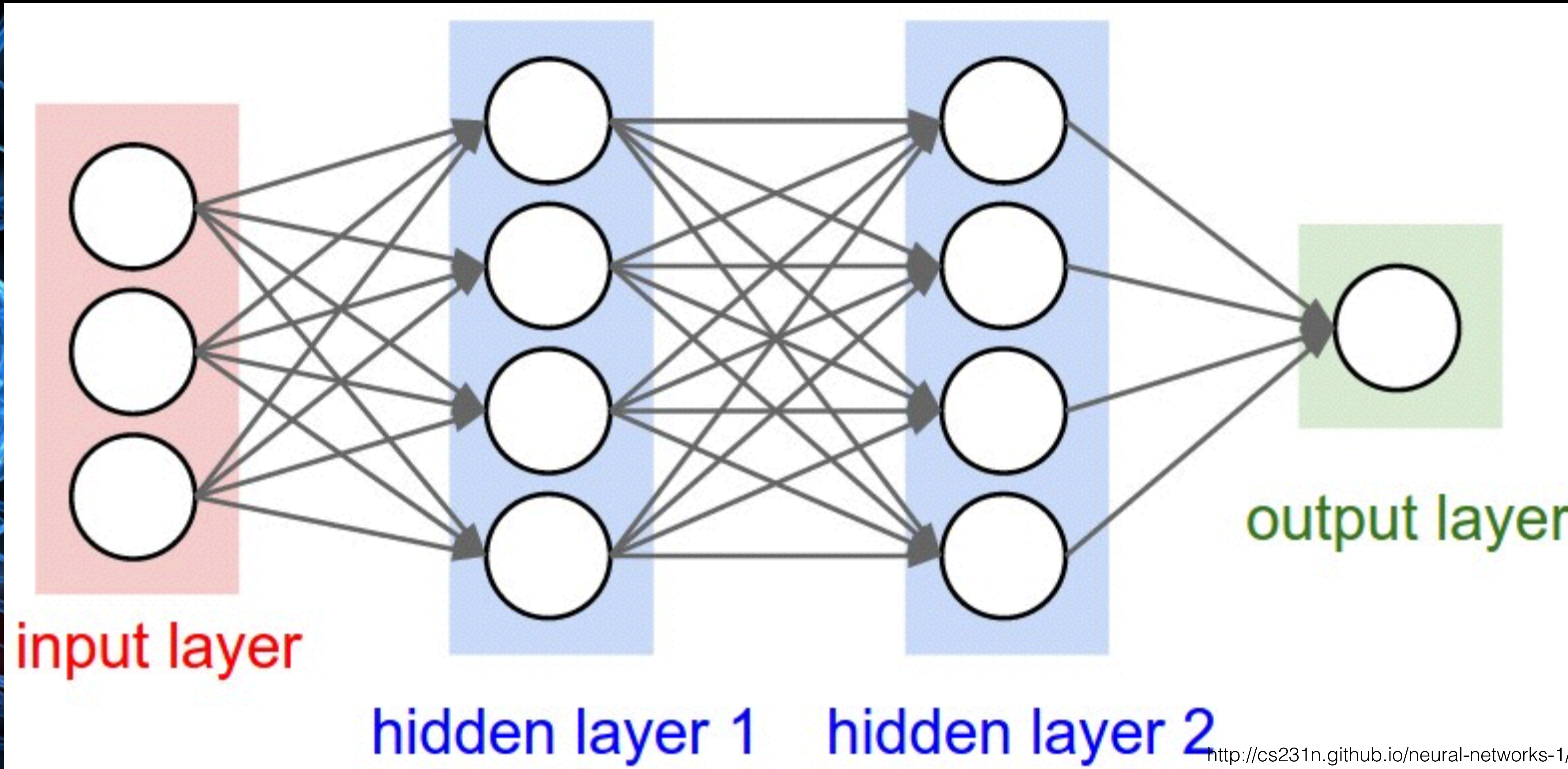


http://www.mblondel.org/journal/2010/09/19/support-vector-machines-in-python/

Support Vector Machines

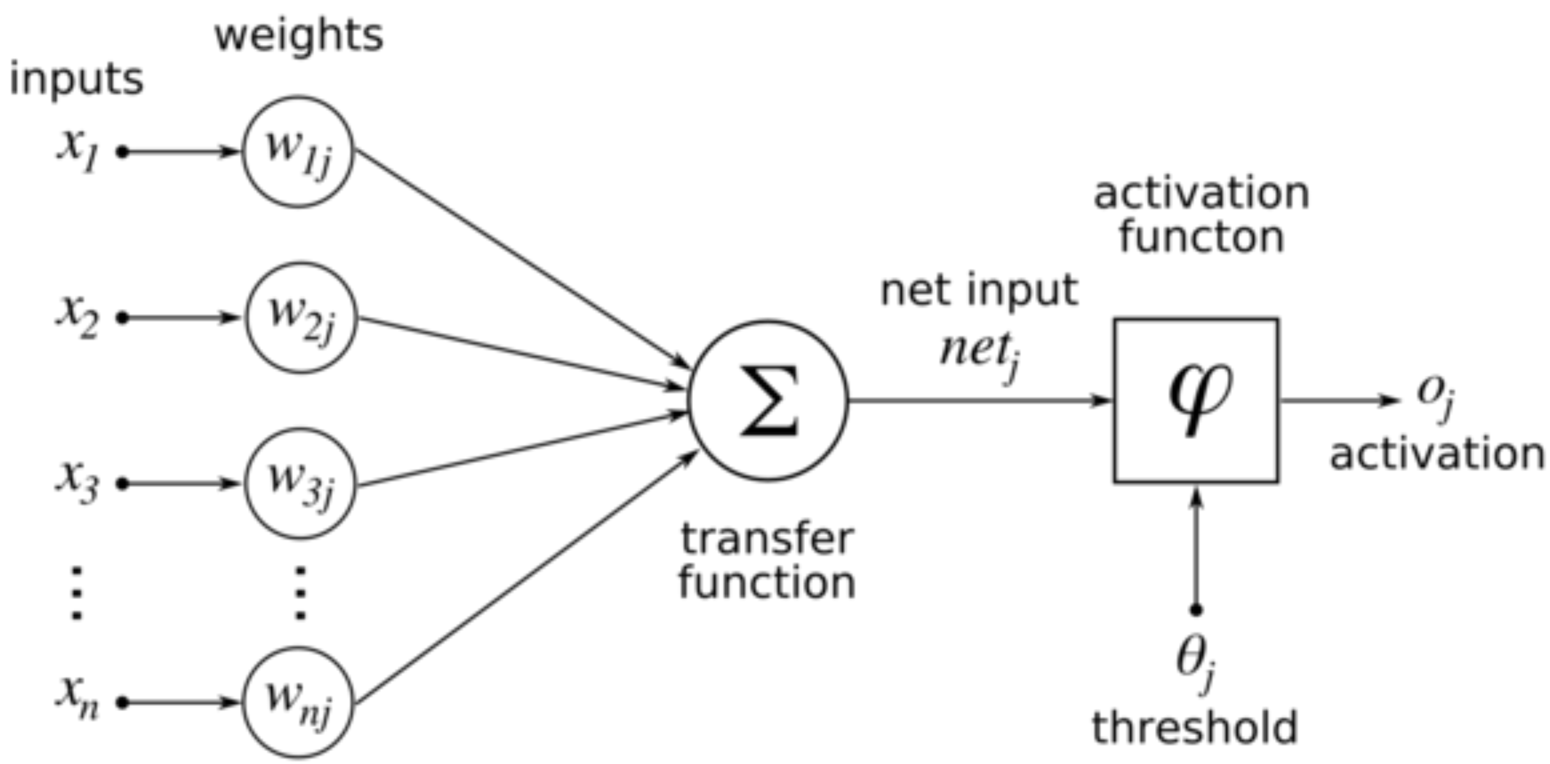


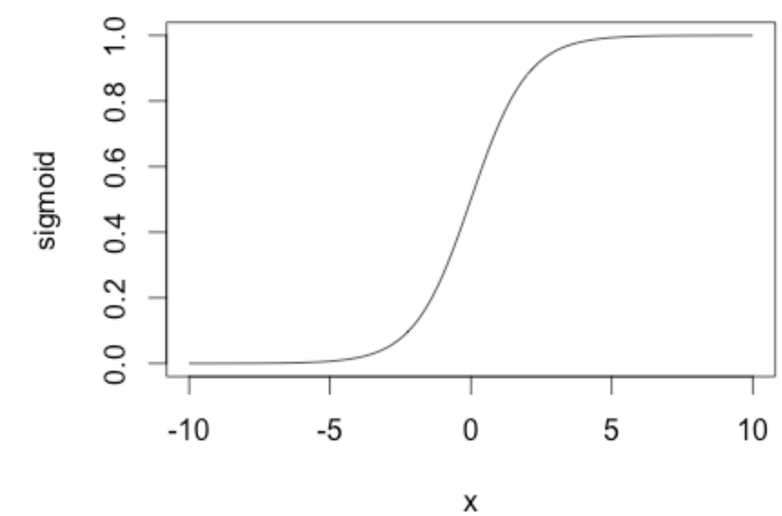
Artificial neural networks



www.extremetech.com

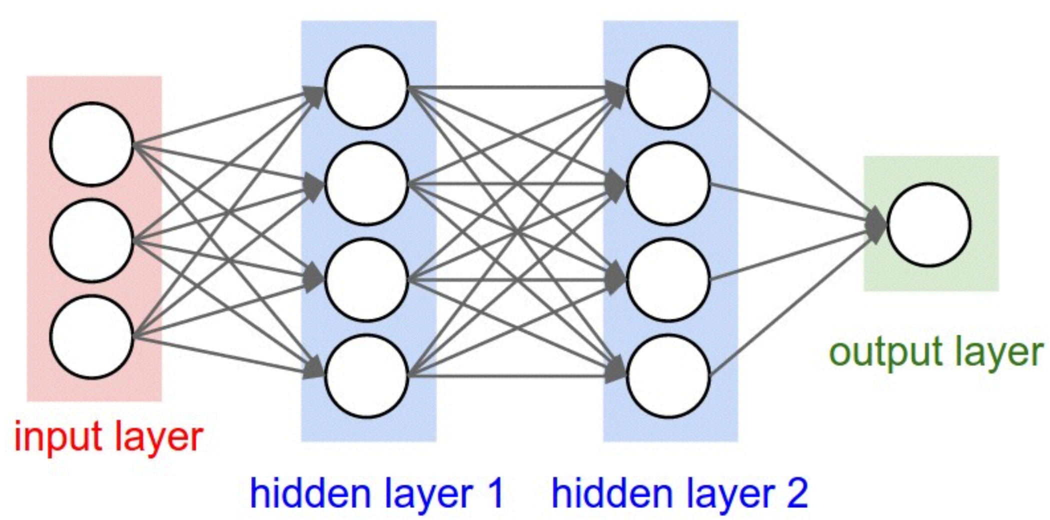
Artificial neural networks - neuron activation scheme





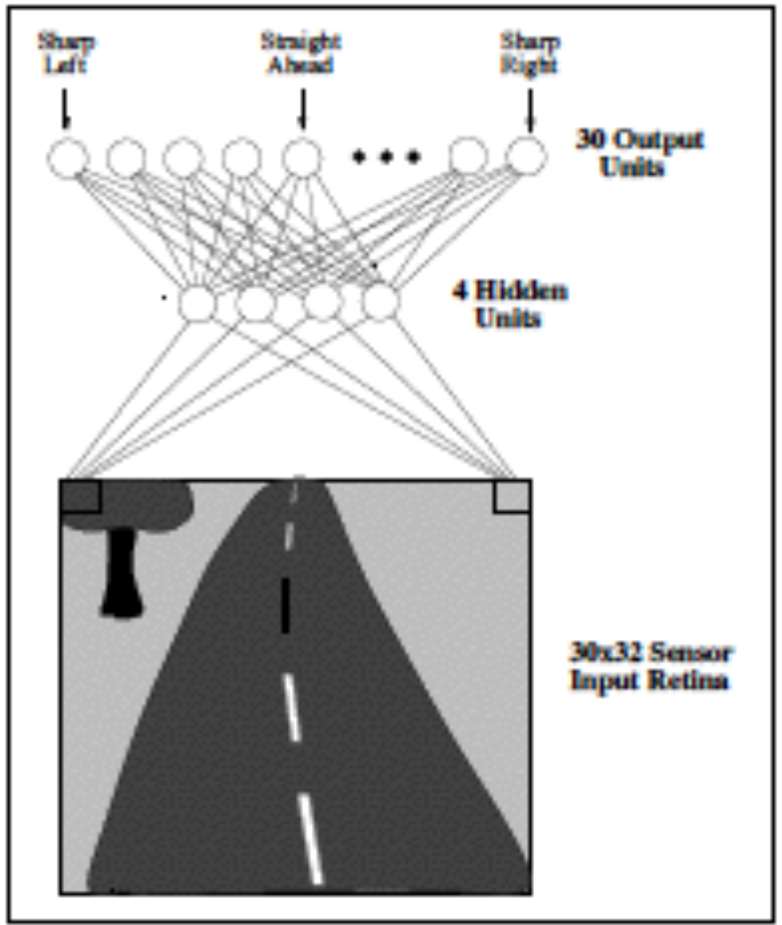
$$o_j = \varphi\left(\sum_{i=1}^{n} w_{ij}x_i + \theta_j\right)$$

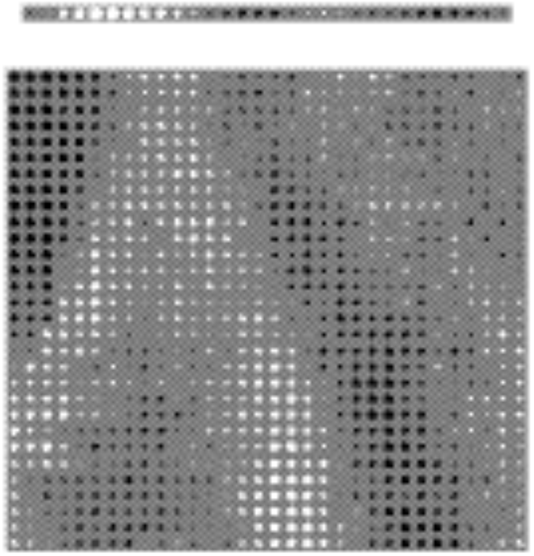
$$o_j = \begin{cases} 1, & \varphi\left(\sum_{i=1}^{n} w_{ij}x_i\right) > \theta_j \\ 0, & \varphi\left(\sum_{i=1}^{n} w_{ij}x_i\right) \le \theta_j \end{cases}$$



https://en.wikibooks.org/wiki/Artificial_Neural_Networks/Activation_Functions

Artificial neural networks - self-driving example



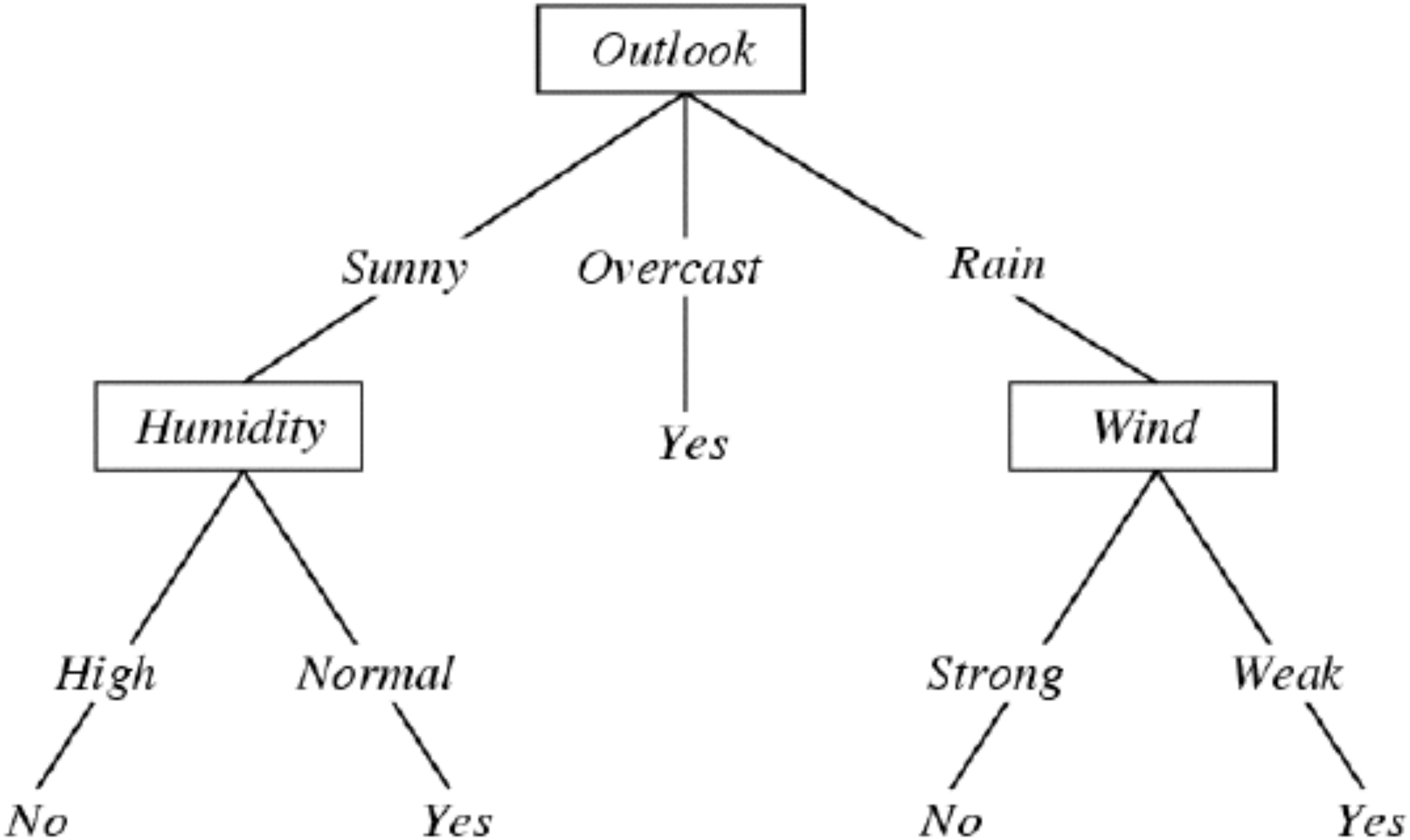


Prof. T. Mitchell's slides/Carnegie-Mellon

Decision tree classifier

A Decision tree for

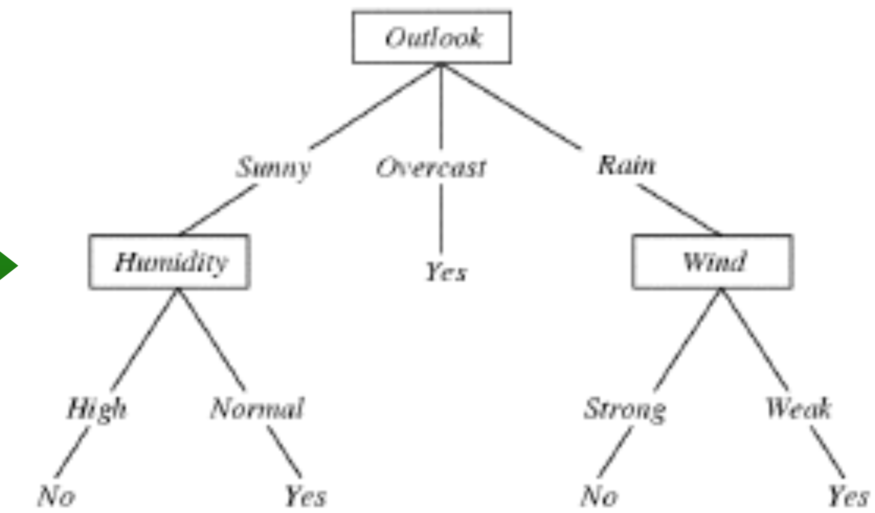
F: <Outlook, Humidity, Wind, Temp> → PlayTennis?



Prof. T. Mitchell's slides/Carnegie-Mellon

Decision tree classifier

Day	Outlook	Temperature	Humidity	Wind	PlayTenr
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

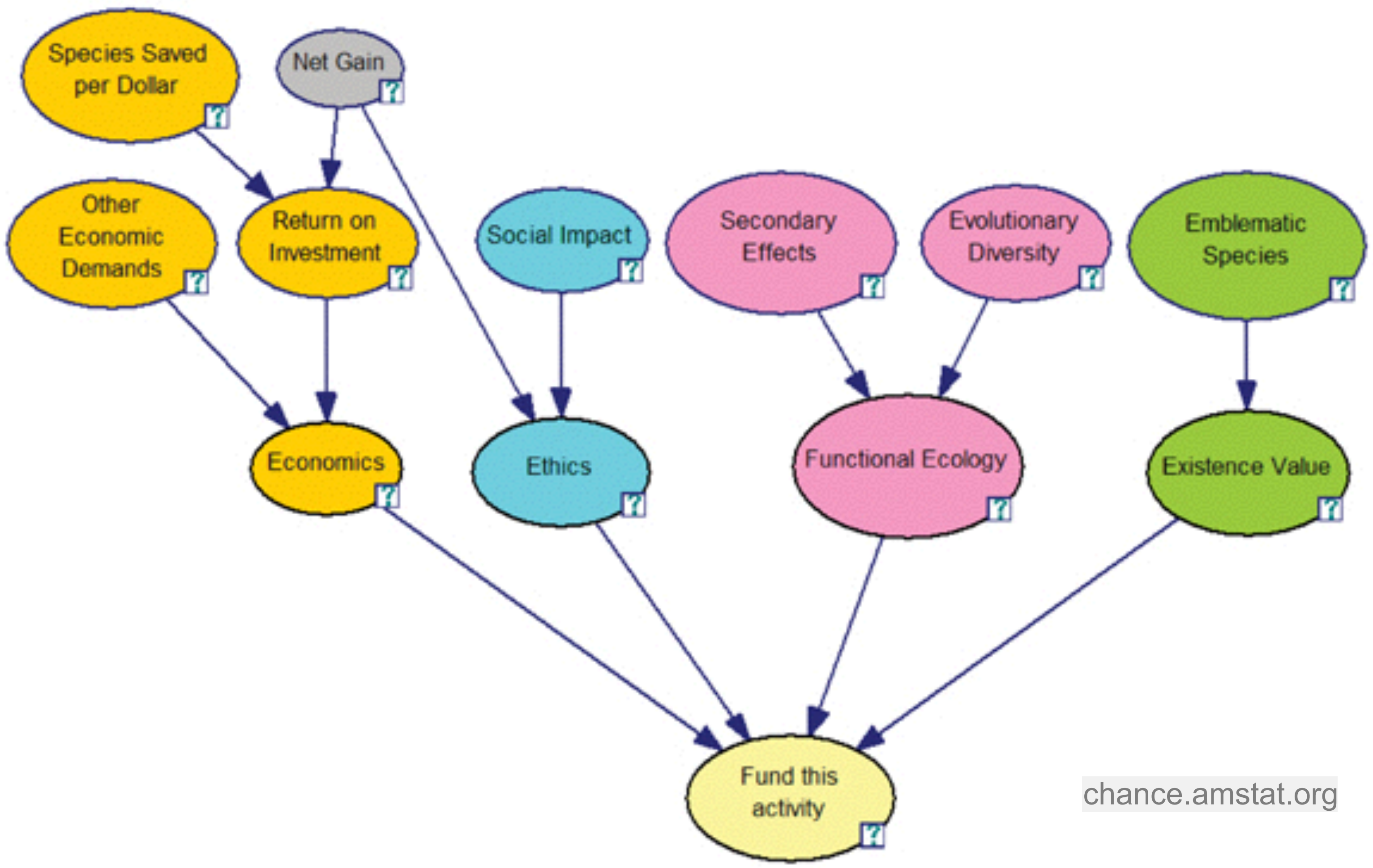


$$H(X) = -\sum_{i=1}^{n} P(X=i)\log_2 P(X=i)$$

$$H(X|Y) = \sum_{v \in values(Y)} P(Y=v) H(X|Y=v)$$

Prof. T. Mitchell's slides/Carnegie-Mellon

Bayesian networks



chance.amstat.org

Other types of supervised learning

ranking

structured prediction

UNSUPERVISED LEARNING APPROACHES AND EXAMPLES

Unsupervised learning

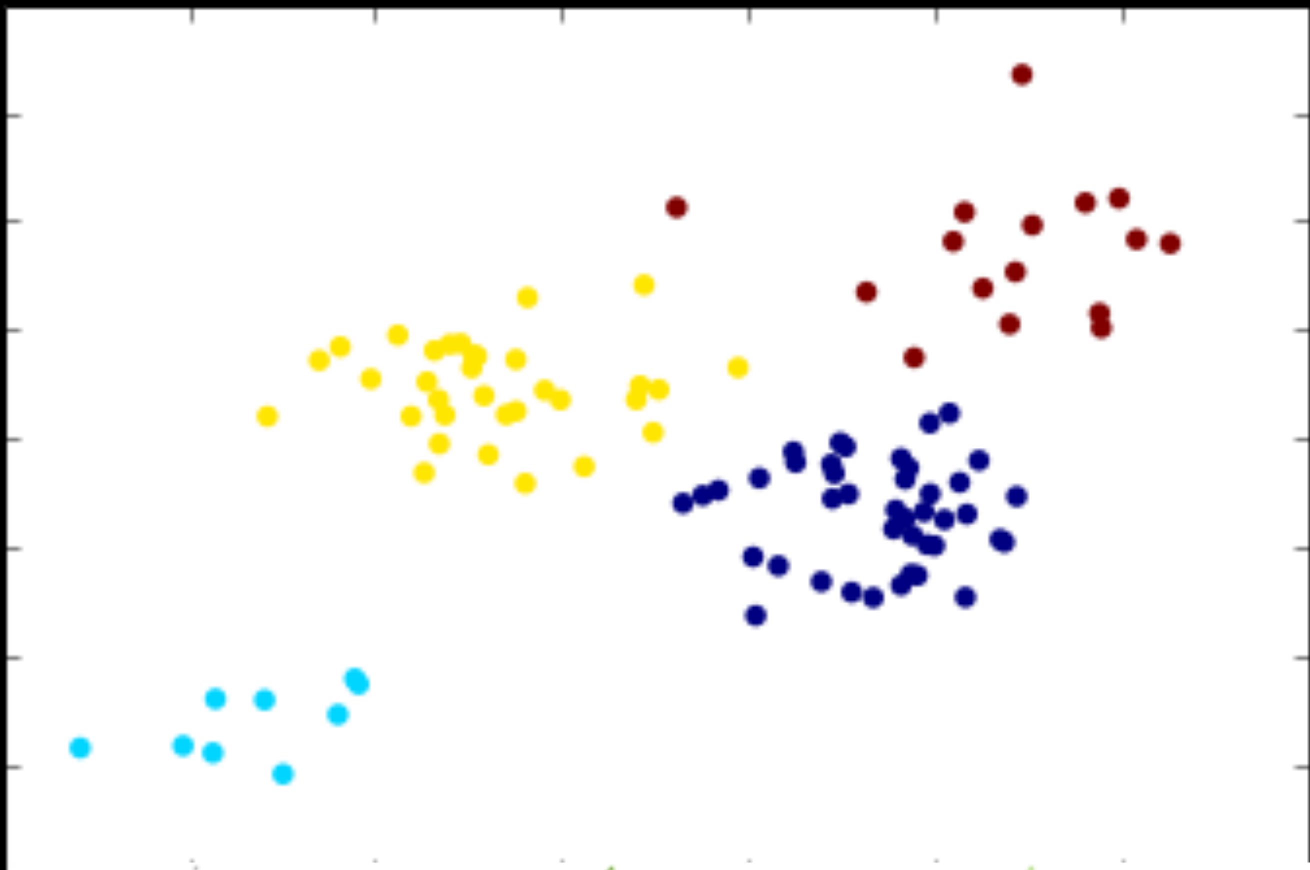
Learning latent variables and patterns

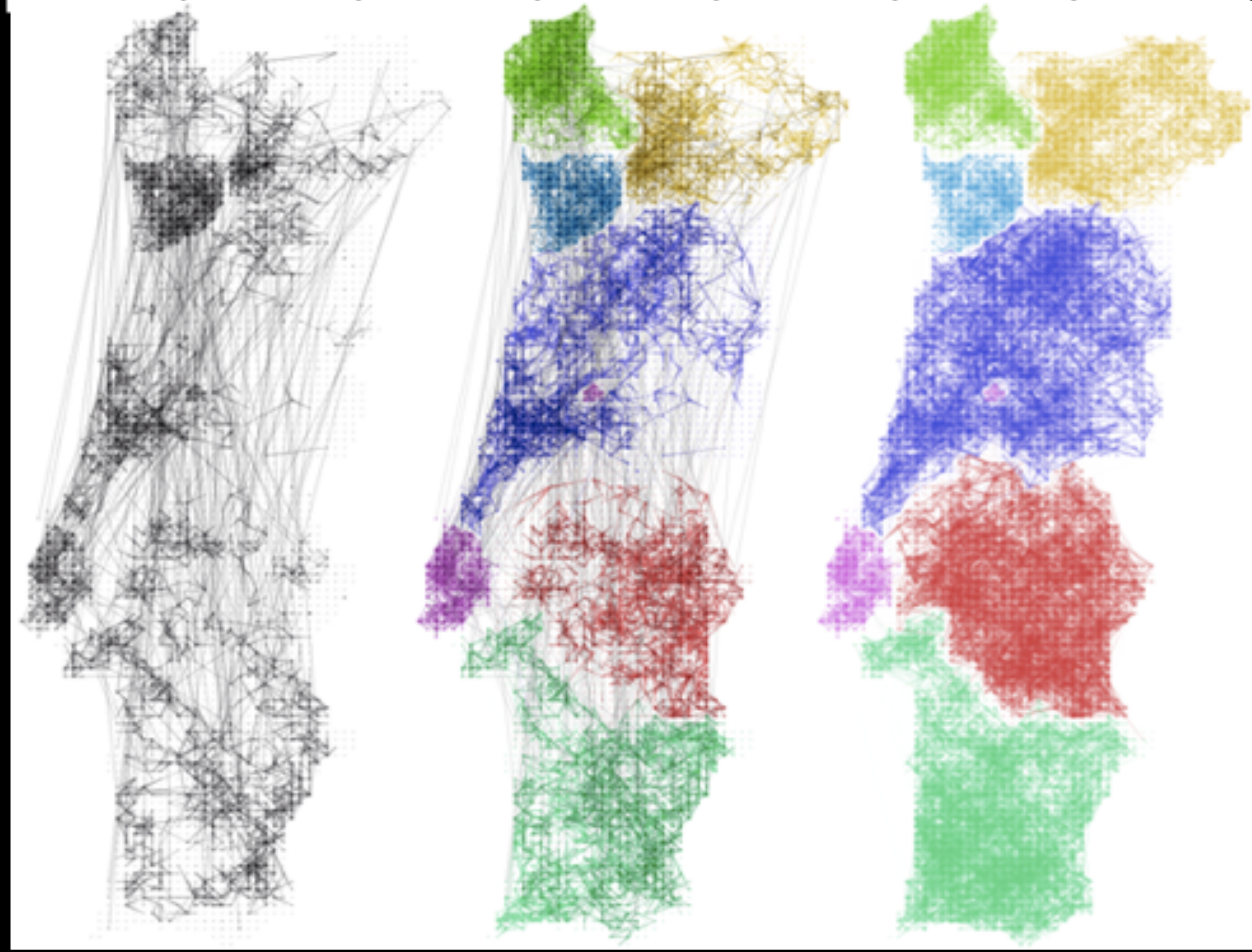
- clustering: learning cluster numbers
- principle component analysis: learning a feature transform to independent feature space
- signal extraction and time-series decomposition
- modeling

Clustering

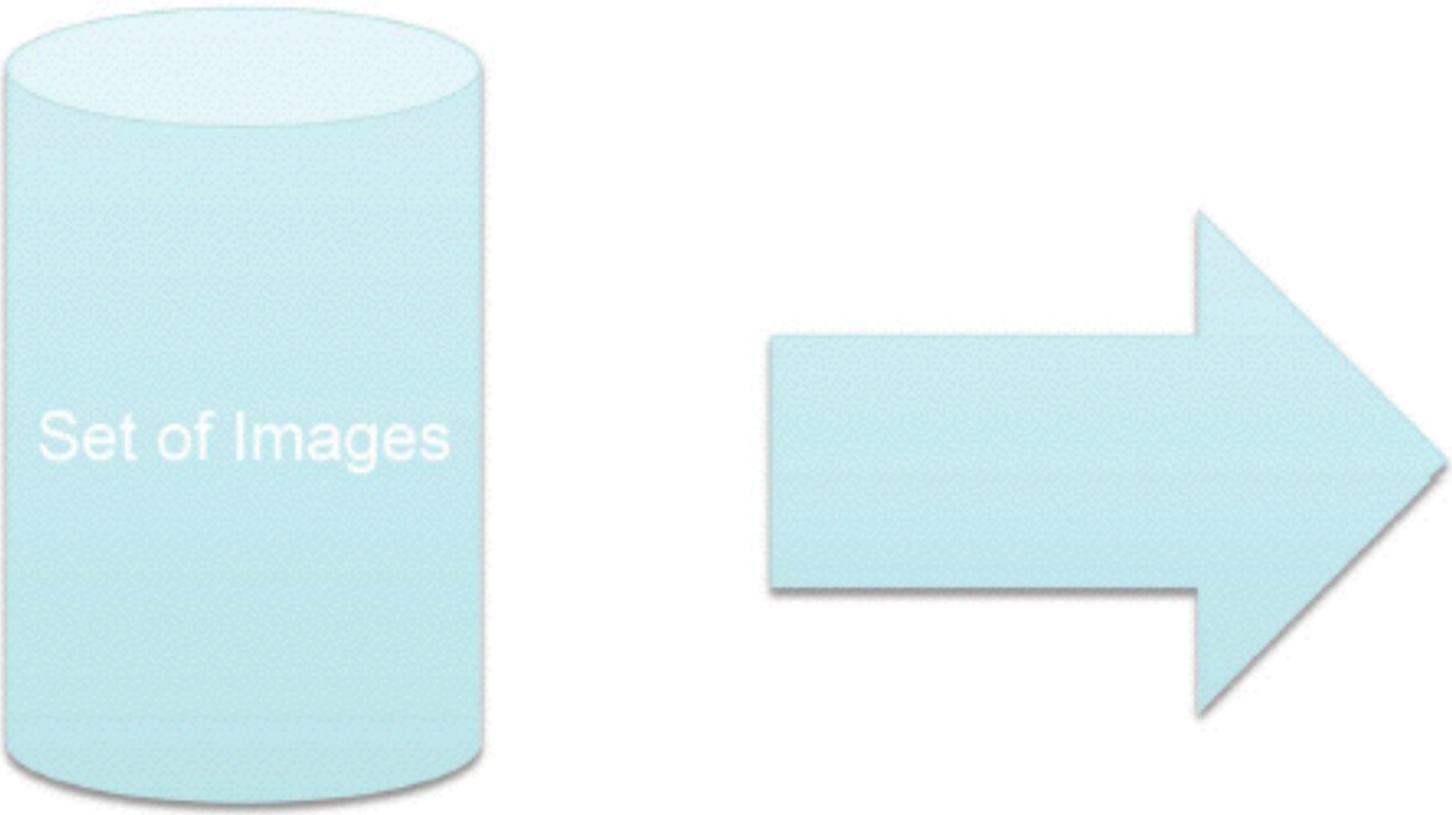
learning cluster numbers

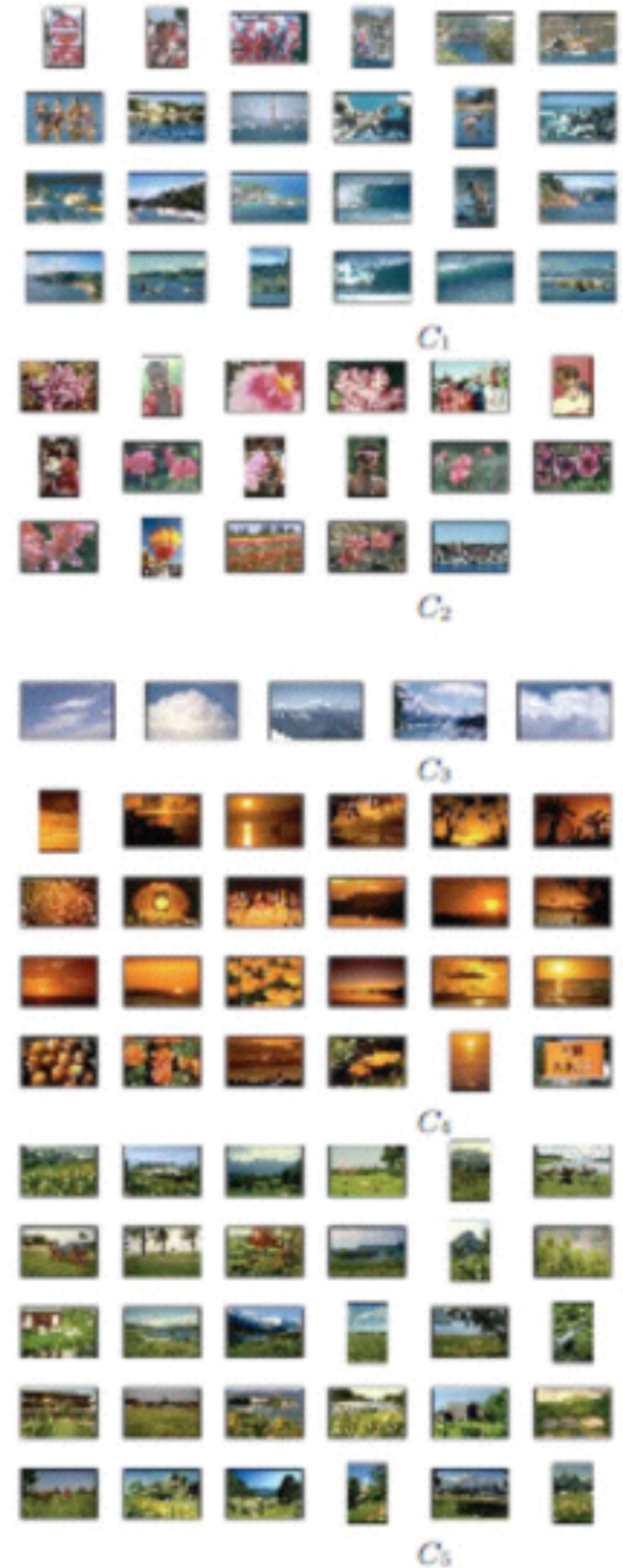
- k-means
- k-medians
- hierarchical clustering
- mixture models
- network community detection





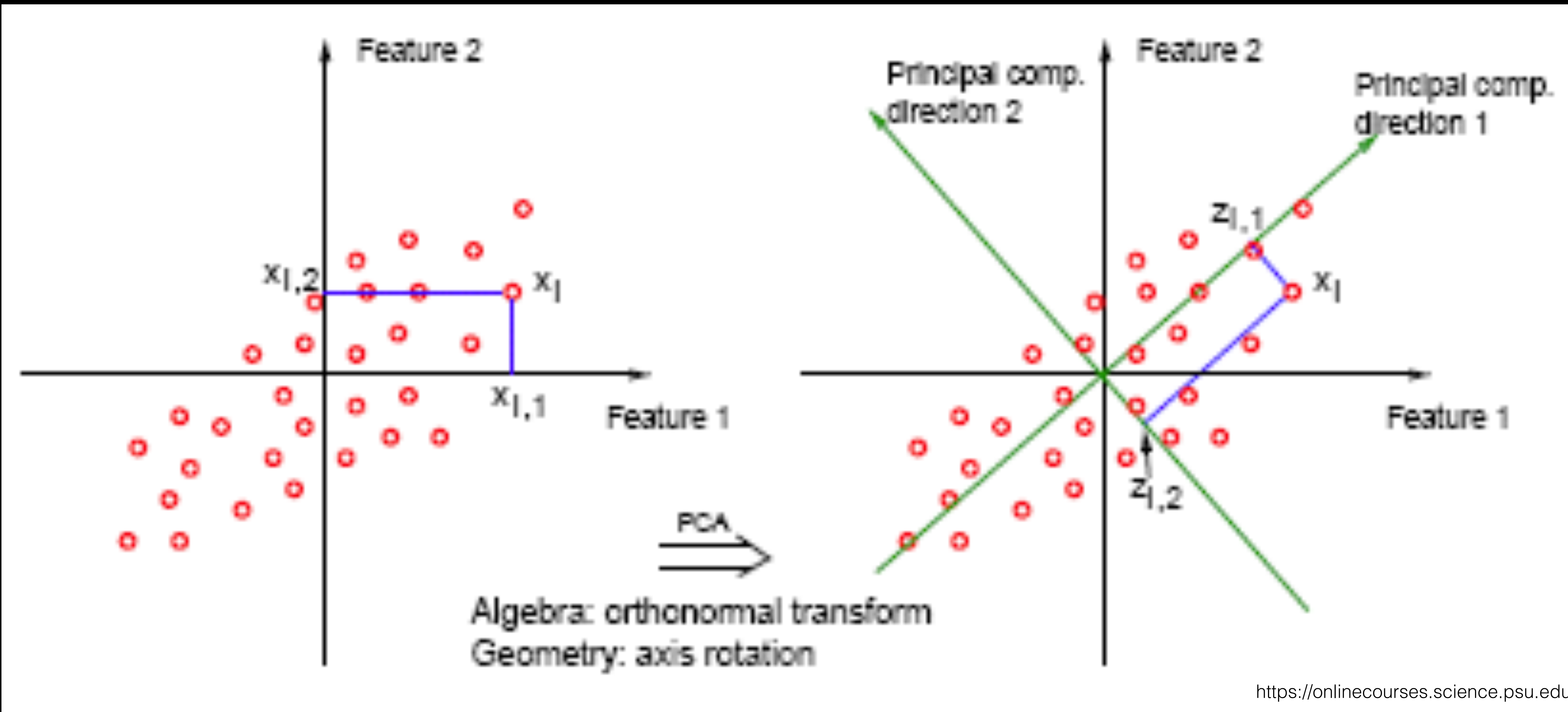
Clustering images





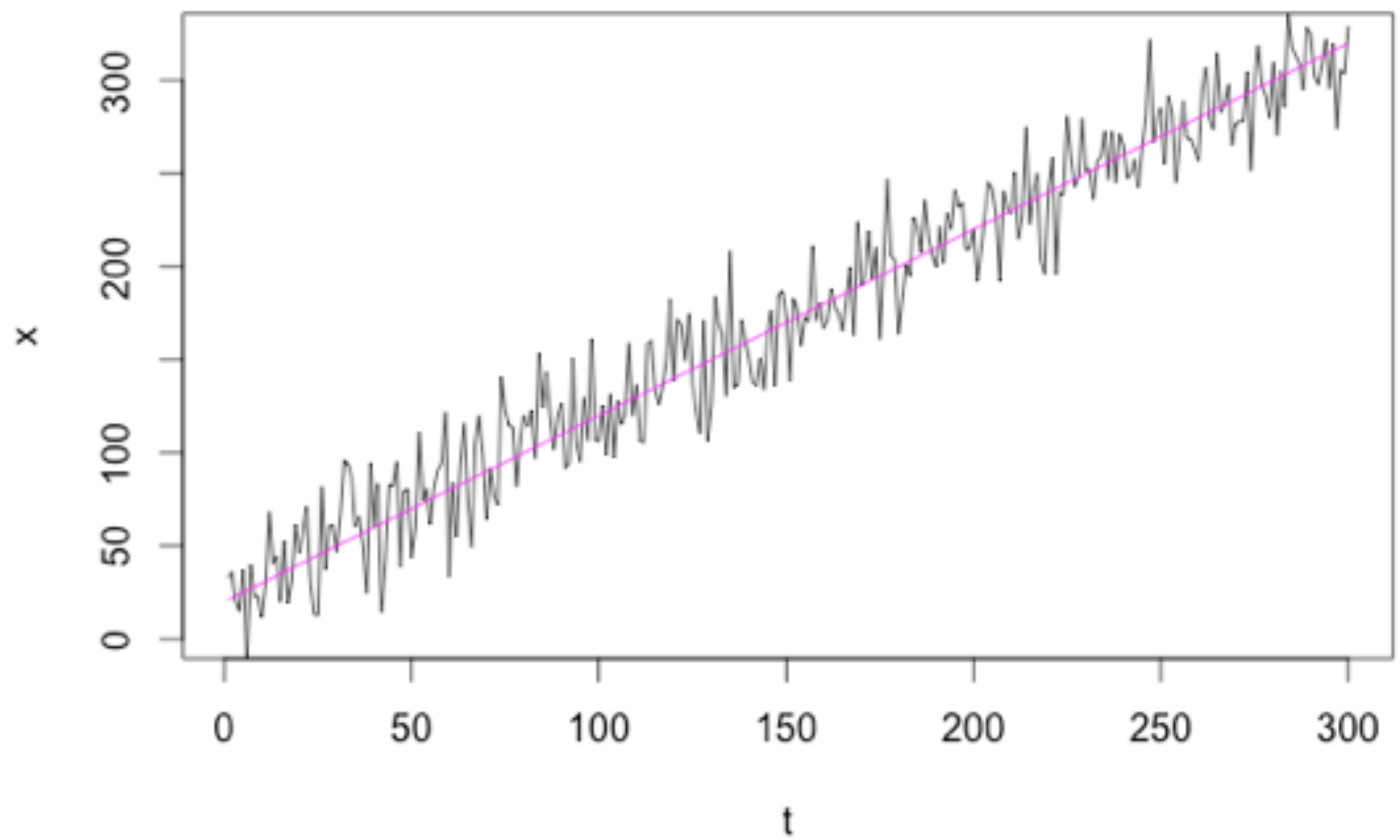
[Goldberger et al.]

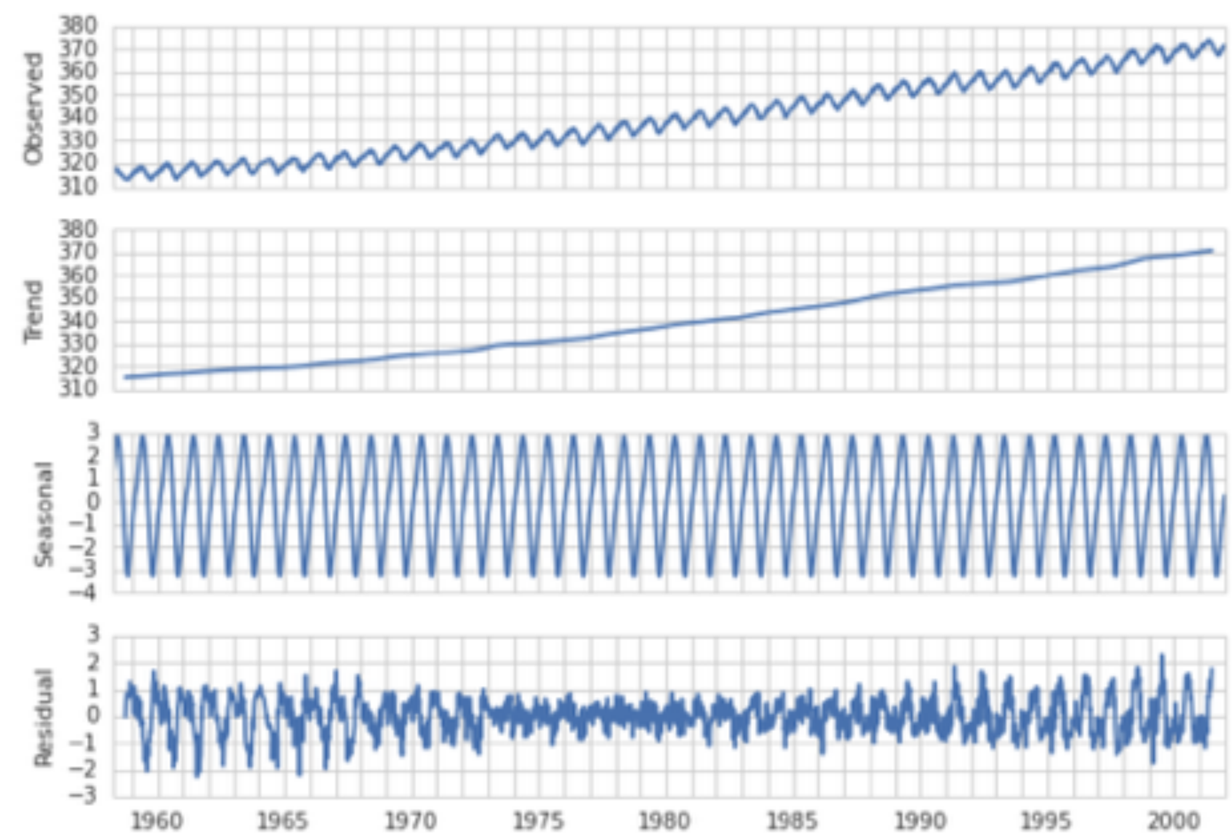
Principle component analysis



Nonlinear - manifold learning

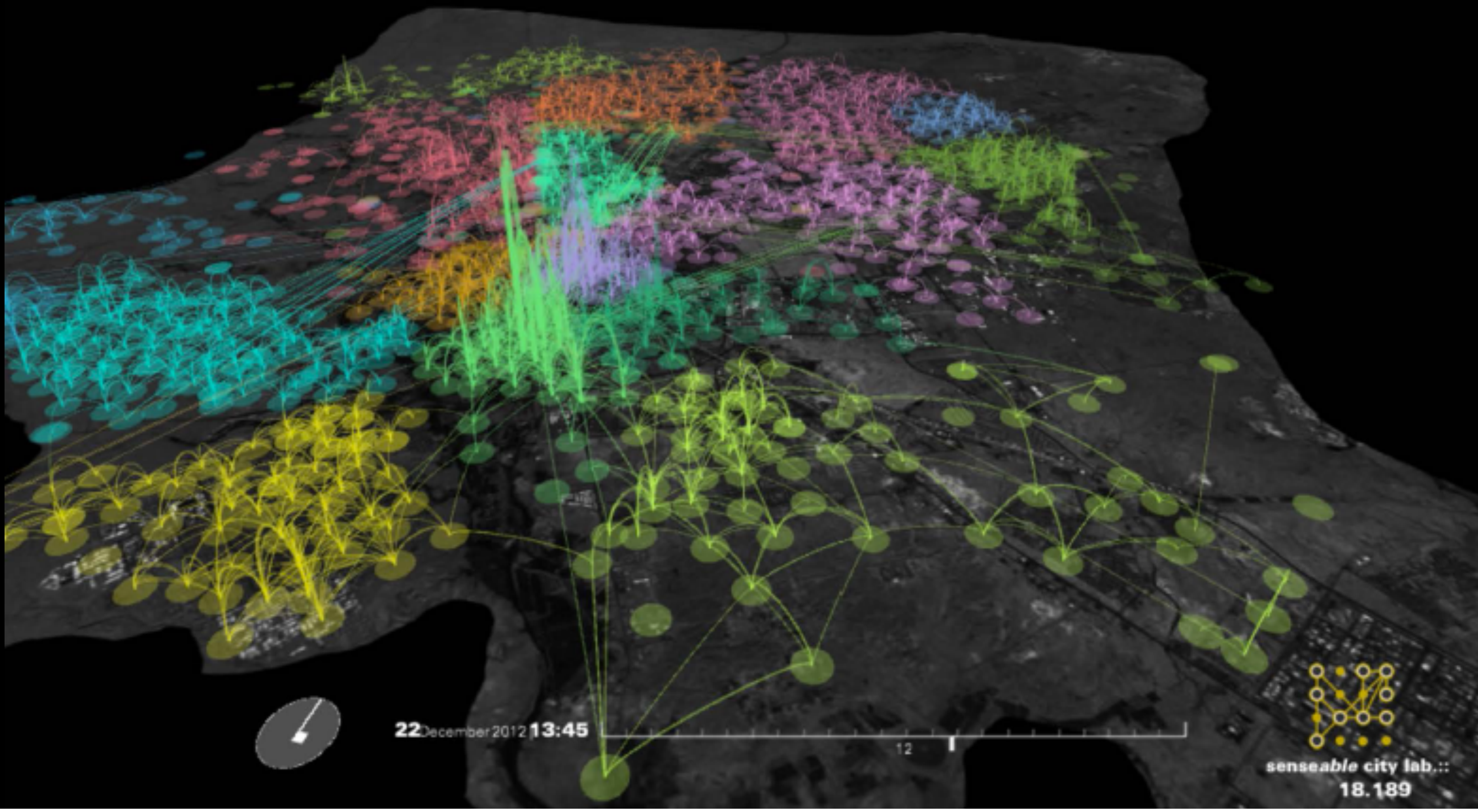
Signal extraction and time-series decomposition





Modeling

selecting a model to describe the data and fitting model parameters



$$e(a,b) = kw(a)^{\alpha} w(b)^{\beta} d(a,b)^{-\gamma}$$

Reinforcement learning

sequential decision making and control problems: distributing delayed rewards among agent actions turning it closer to supervised learning framework

Markov decision processes:
value iteration and policy iteration

$$s_0 \xrightarrow{a_0} s_1 \xrightarrow{a_1} s_2 \xrightarrow{a_2} s_3 \xrightarrow{a_3} \dots \qquad \pi : S \mapsto A$$

$$\pi^*(s) = \arg\max_{a \in A} \sum_{s' \in S} P_{sa}(s') V^*(s')$$

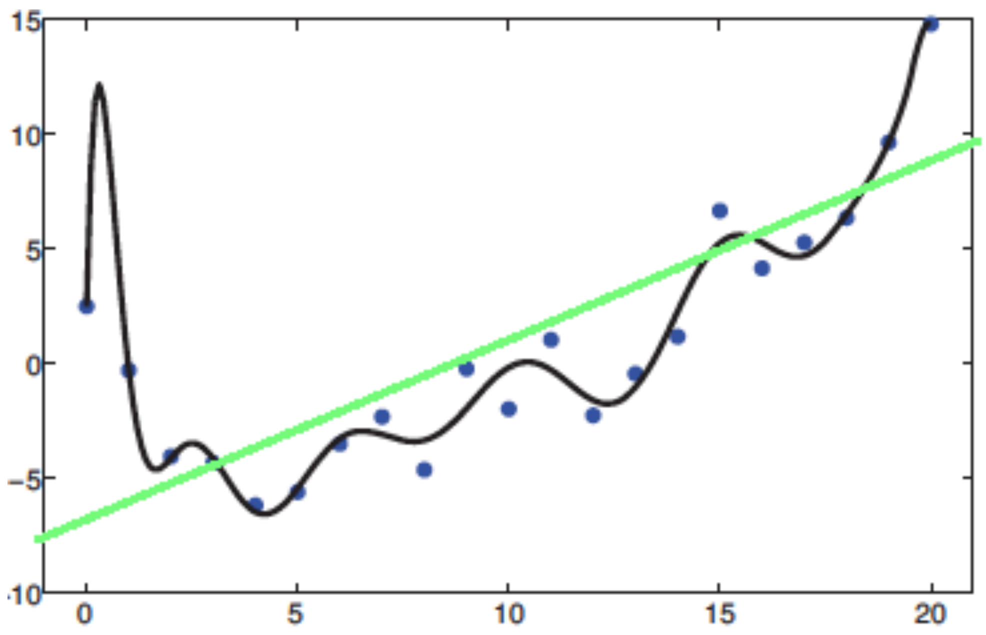
MACHINE LEARNING CHALLENGES

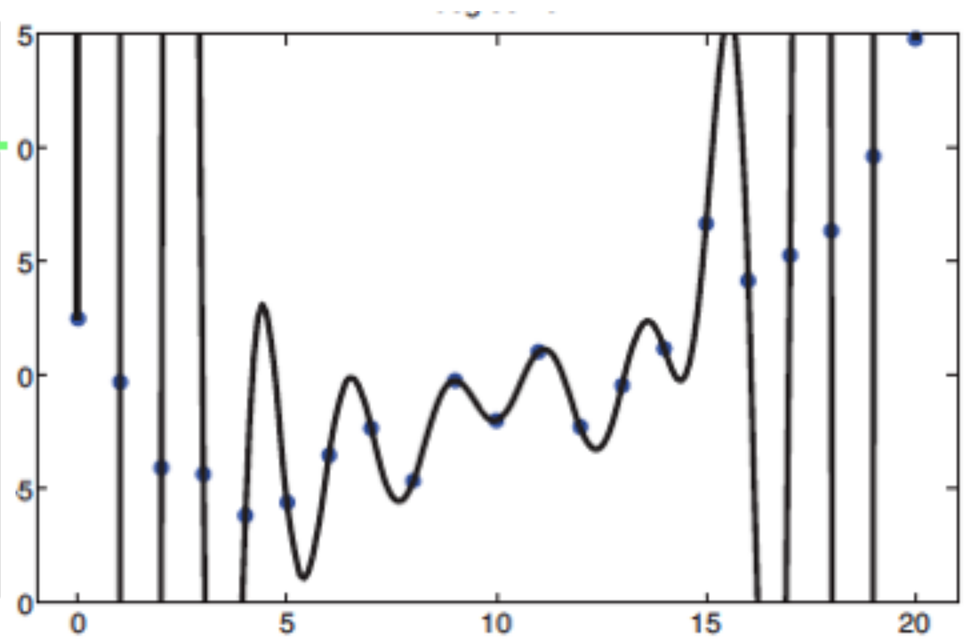
ML challenges

- getting enough training data
- observing enough relevant features
- are all the features relevant? overfitting
- what model to select? adjusting model complexity to the data available
- dealing with unobserved (latent) features

- diagnostics - find issues to fix

Overfitting





Murphy, Kevin P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

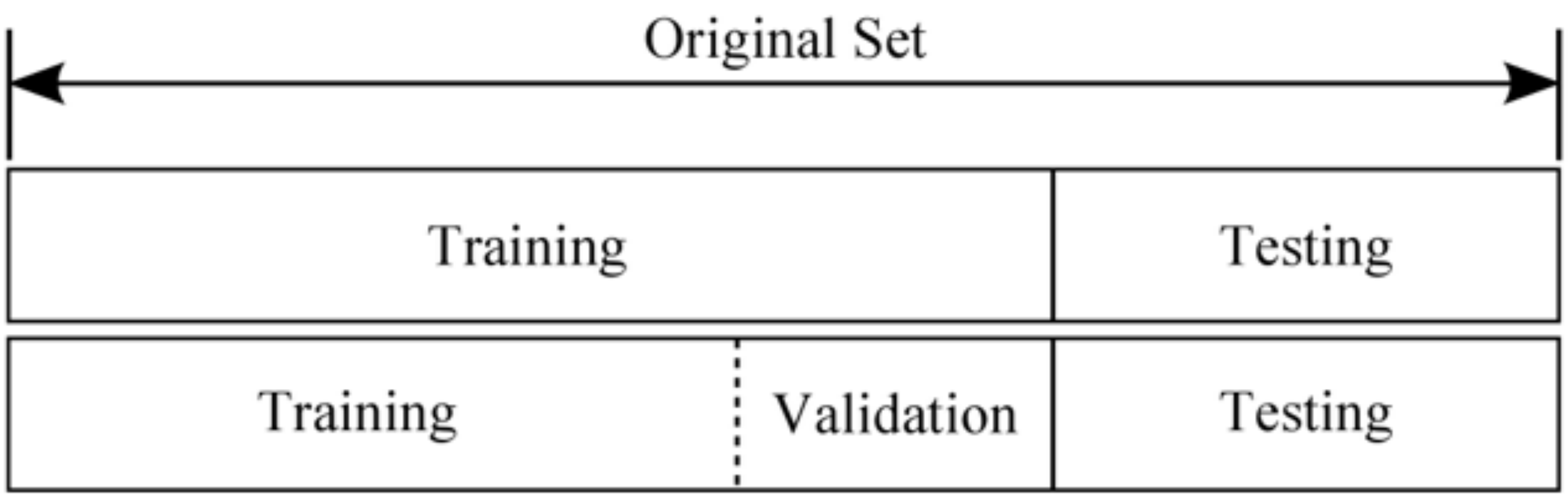
The more sophisticated model always fits better.
But it is generalizable?

Solutions:

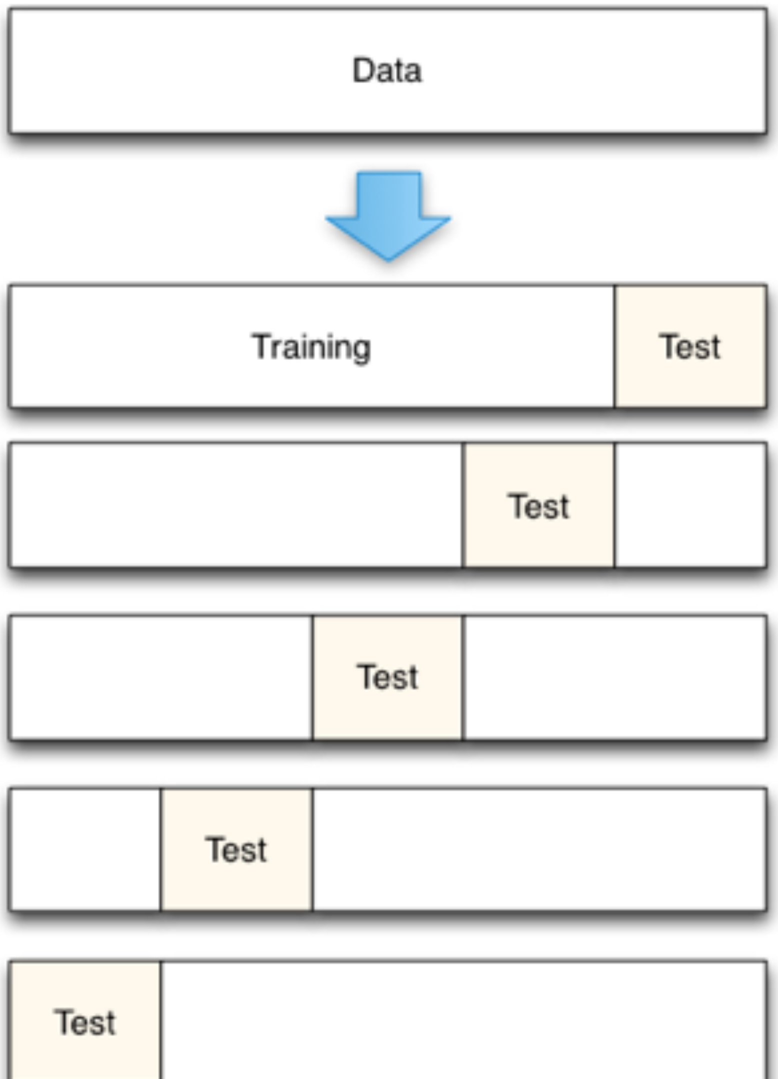
- model selection through cross-validation
- dimensionality reduction/feature selection
- regularization

Cross-validation

- train the model on one subsample
- perform feature/model selection over the different validation subsample
- test performance over the training set



www.intechopen.com



blog.kaggle.com

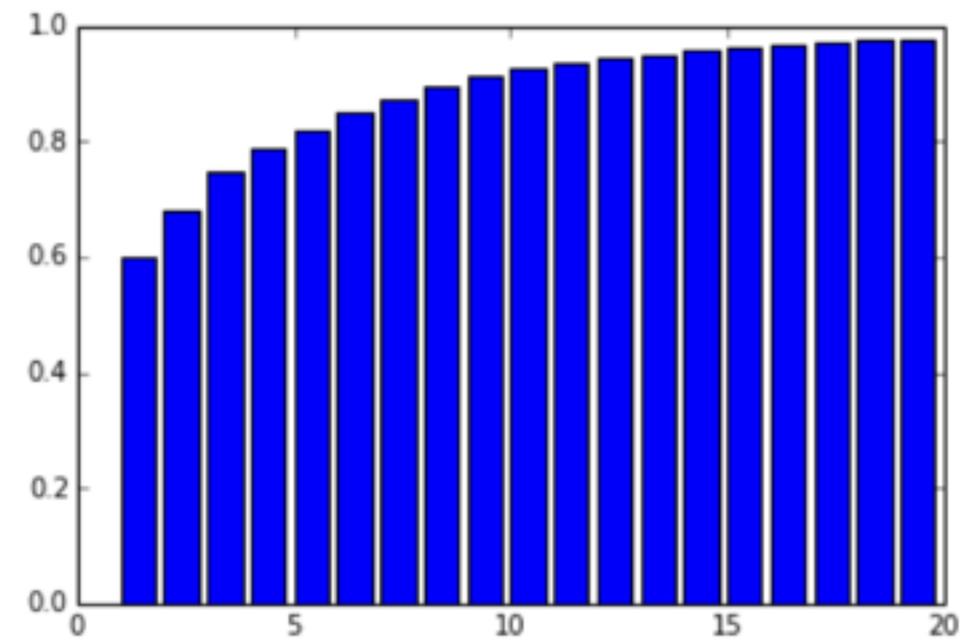
Feature selection

From the available variety select features which help the model the most (cross-validation):

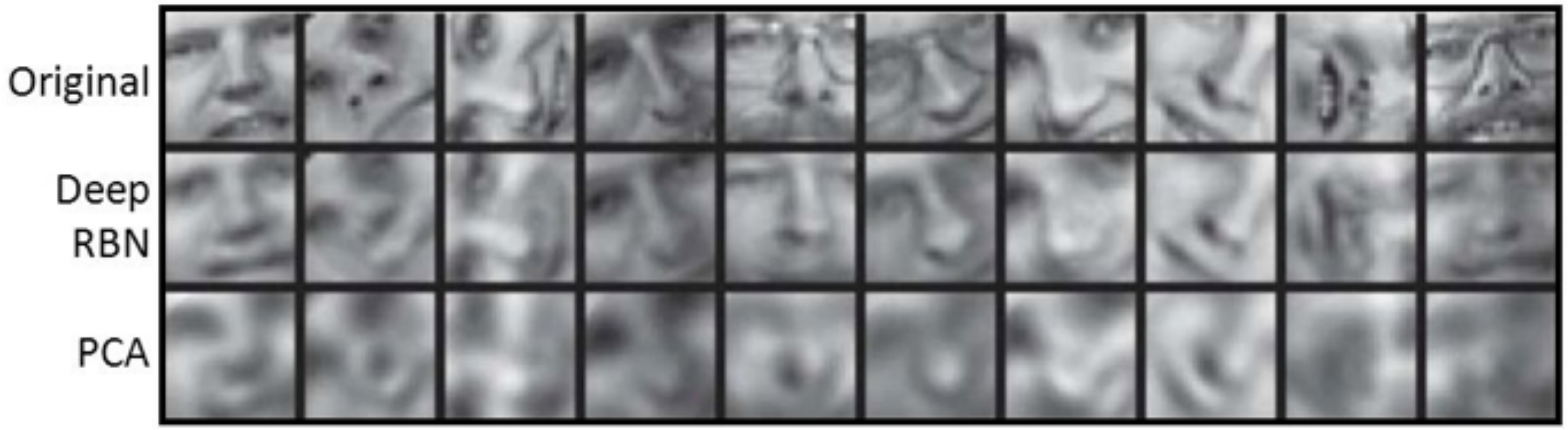
- greedy heuristics: forward/backward step-wise
- minimum-redundancy-maximum-relevance
- global optimization

Dimensionality reduction

$$U = XV = W\Sigma V^T V = W\Sigma$$



Olivetti face data, 25x25 pixel images reconstructed from 30 dimensions (625 → 30)



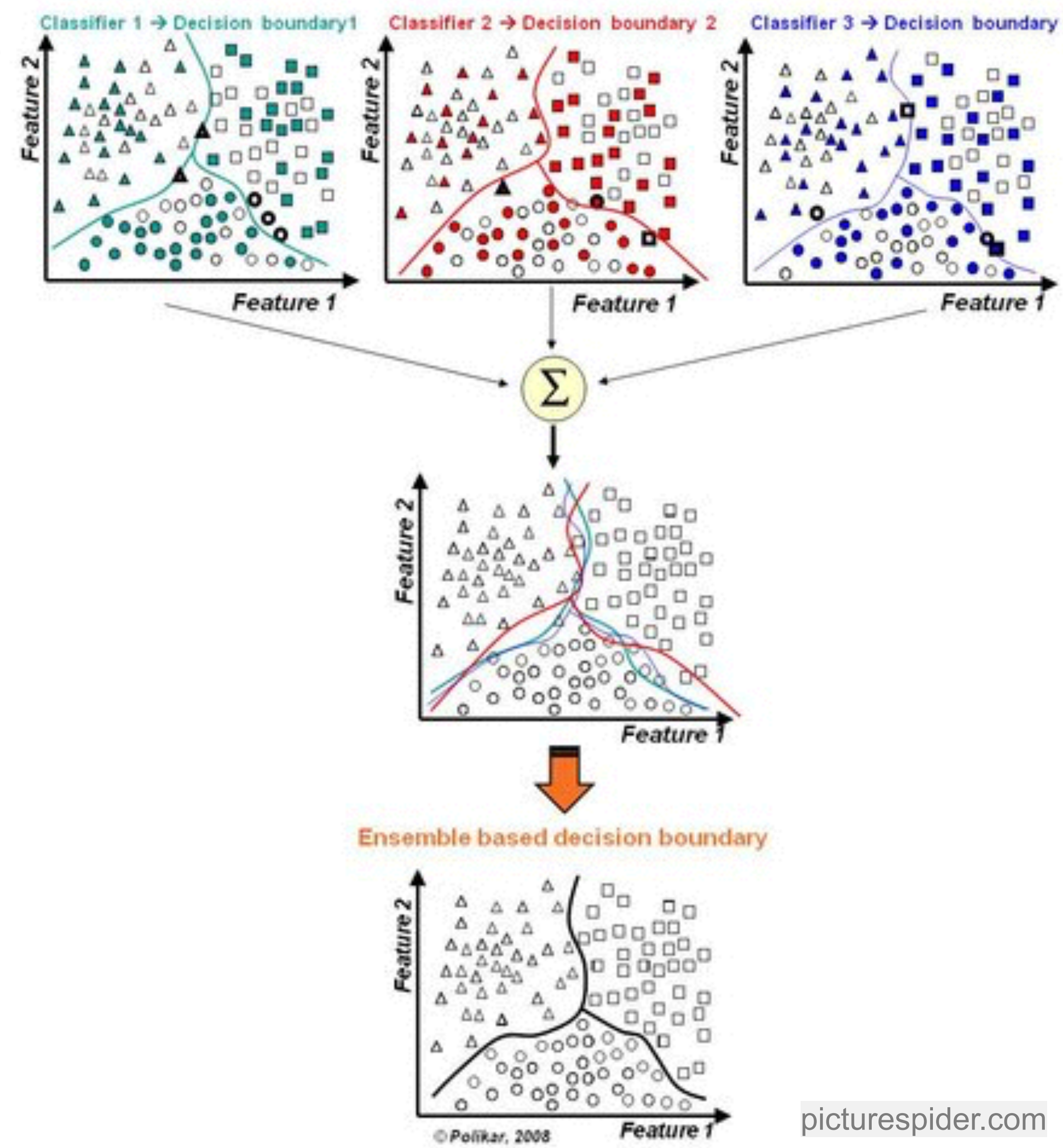
Ensemble learning

Which model to select?
Too much or too little data

- random forests
- k-NN
- bootstrap/bagging
- stacking
- Bayesian model combination

Ensemble Learning:
The Wisdom of Crowds

Lior Rokach's slides



picturespider.com

TECHNIQUES. EXPERIENCE. STRATEGY