



Μελέτη και Κατασκευή Μοντέλων για Πρόβλεψη του Ρυθμού Εγκατάλειψης σε Πρόγραμμα Αποταμιεύσεων

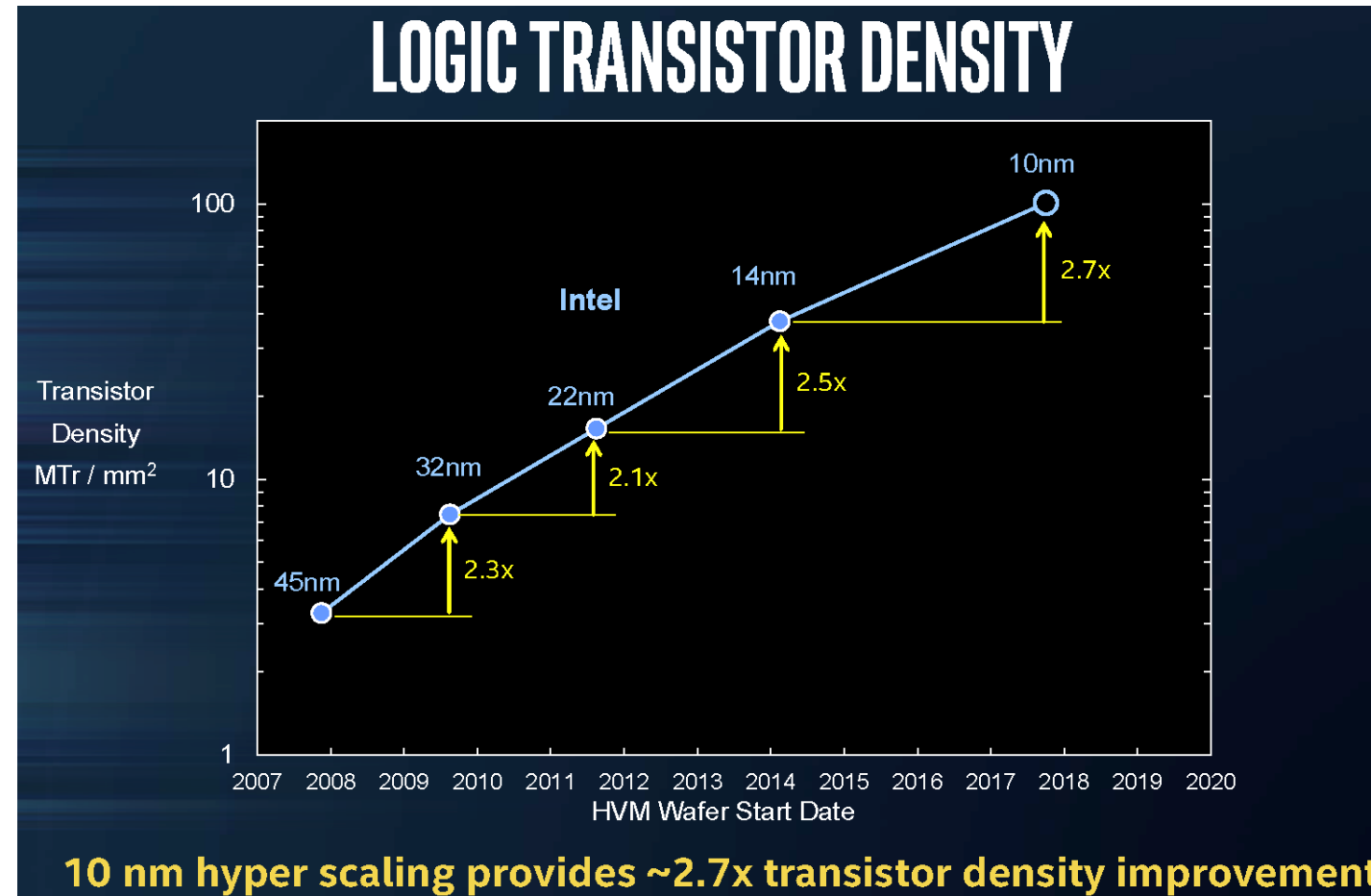
Διπλωματική Εργασία
Γεωργακόπουλος Παναγιώτης
0310 9757

19 Ιουλίου 2017

Moore's Law

- Moore's Law (1965) predicts **x2 improvement** in computing capabilities every ~12-18 months
- Moore's Law alive and well after 52 yrs
- *"In my 34 years in the semiconductor industry, I have witnessed the **advertised death of Moore's Law no less than four times**. As we progress from 14 nanometer technology to 10 nanometer and plan for 7 nanometer and 5 nanometer and even beyond, **our plans are proof that Moore's Law is alive and well**"*

– Intel CEO Brian Krzanich, 2016



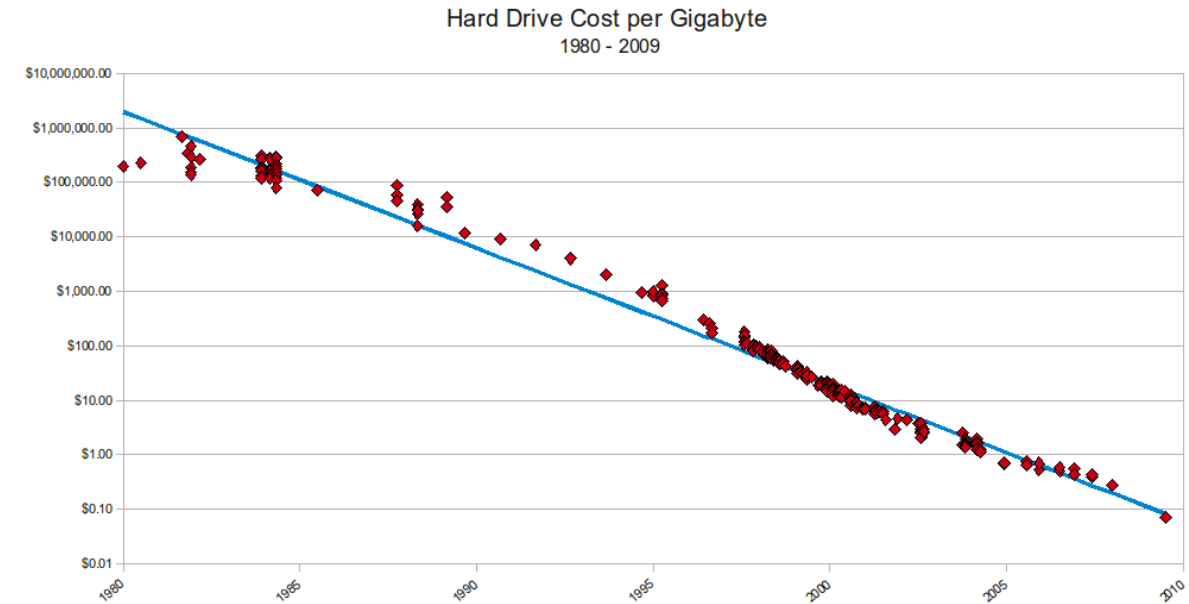
10 nm hyper scaling provides ~2.7x transistor density improvement

Source: Intel



Why Moore's Law Matters for Data Science

- Computers are **cheaper** and **faster** than ever
- Computers are **everywhere**: Industry, Homes, Mobile, Cars, Embedded, Wearables
- Exponential growth in computing capabilities is followed by exponential growth in data available for collection and storage
- Cost of 1 Gigabyte of storage now **below \$0.05**, from **\$200,000 - \$400,000** back at 1980
- **Easy** and **cheap** to keep all data possible



Collect tremendous amount of data. Then what?

- We have established that collecting data is **easy** and **cheap**.
- The straightforward business decision is to collect and store everything
- *Next step:*
Leverage the information to gain advantage.
- Need to **understand, analyze** and **model** data, to be able to extract **valuable, actionable insights** about your **business**



Who can benefit

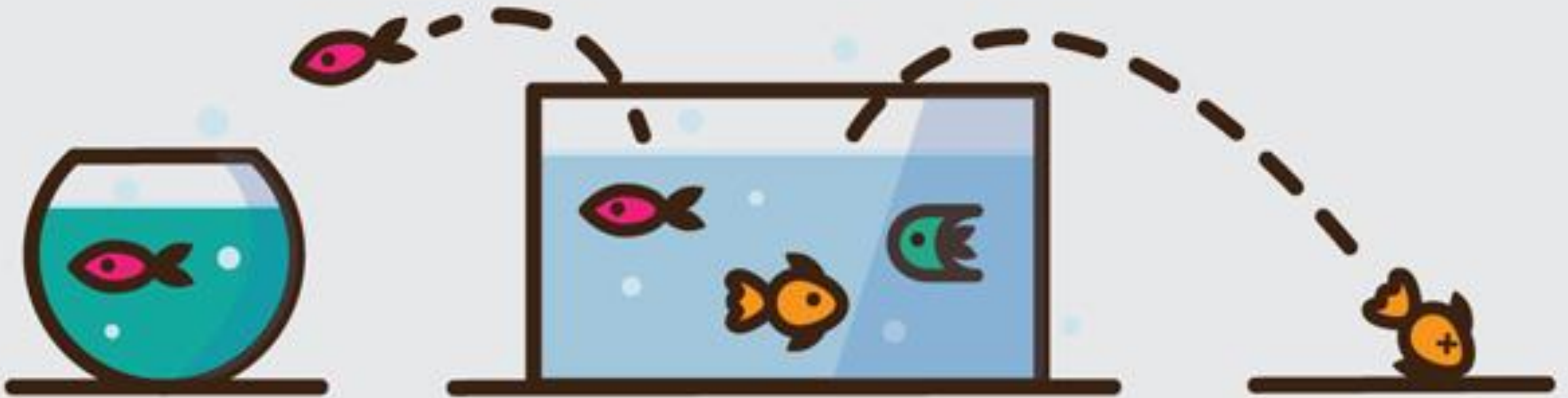
- Everyone!
 - *Not using the data is an operational **risk!***
 - *Using data is an operational **opportunity!***
 - *Winning the race is an operational **challenge!***
- Telecom
 - Understand **user needs**
 - **Optimize network** coverage/quality vs population habits
- Retail Chains
 - Optimize **advertising**
 - Optimize **product placement**
 - Make accurate **predictions**
- Banking System
 - Observe customer **behavior**
 - Suggest **relevant products** to **eligible customers**
 - Get **actionable** insights for existing customers
 - **Prevent** customer **churn**



Data Science



CUSTOMER CHURN



Customer Churn

- **Customer churn** refers to when a **customer** (player, subscriber, user, etc.) ceases his or her relationship with a company.
- Any business needs to minimize customer churn (attrition) to maintain a healthy and satisfied customer base.
- The need to minimize churn is even greater when the environment is heavily **challenging** as is the case with the **banking system** since **2007**
- **We will use data to predict (thus provide insights to prevent) customer churn in a banking program**



Introducing the problem



- A Banking Institution needs to gain insights about their affluent clients
- Given the financial and demographic data of a client, can we predict if the client will remain in the affluent program of the bank?



- Insight for each client
- Flag risky clients for further actions



Available information - Demographics

Id	Sex	Age	Postal	MaritalStatus	Education	HomeStatus
1	F	49	15126	ΠΑΝΤΡΕΜΕΝΟΣ/Η	ΤΡΙΤΟΒΑΘΜΙΑ ΕΚΠΑΙΔΕΥΣΗ	ΙΔΙΟΚΑΤΟΙΚΗΣΗ
2	F	60	19014	ΠΑΝΤΡΕΜΕΝΟΣ/Η	ΤΡΙΤΟΒΑΘΜΙΑ ΕΚΠΑΙΔΕΥΣΗ	ΙΔΙΟΚΑΤΟΙΚΗΣΗ
3	F	58	10445	ΠΑΝΤΡΕΜΕΝΟΣ/Η	ΤΡΙΤΟΒΑΘΜΙΑ ΕΚΠΑΙΔΕΥΣΗ	ΆΛΛΟ
4	M	49	17562	ΠΑΝΤΡΕΜΕΝΟΣ/Η	ΜΕΤΑΠΤΥΧΙΑΚΑ	UNKNOWN VALUE
5	M	54	14234	ΧΩΡΙΣΜΕΝΟΣ/Η	ΤΡΙΤΟΒΑΘΜΙΑ ΕΚΠΑΙΔΕΥΣΗ	ΙΔΙΟΚΑΤΟΙΚΗΣΗ
6	M	84	16232	ΠΑΝΤΡΕΜΕΝΟΣ/Η	ΤΡΙΤΟΒΑΘΜΙΑ ΕΚΠΑΙΔΕΥΣΗ	UNKNOWN VALUE

Profession	Email	InternetConn	Phone	Months	Stay
ΕΛ.ΕΠ.ΔΗΜΟΣΙΟΓΡΑΦΟΣ	0	1	1	5	FALSE
ΔΗΜ.ΥΠΑΛ.ΤΡΑΠΕΖΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	1	1	1	2	FALSE
ΙΔ.ΥΠΑΛ.ΤΡΑΠΕΖΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	1	1	1	19	FALSE
ΙΔ.ΥΠΑΛ.ΛΟΙΠΕΣ ΕΙΔΙΚΟΤΗΤΕΣ	1	1	1	3	FALSE
ΙΔ.ΥΠΑΛ.ΤΡΑΠΕΖΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	1	1	1	3	FALSE
ΣΥΝΤΑΞ.ΟΑΕΕ	0	0	0	1	FALSE



Available information - Demographics

- Sex (32% female)
- Age ([Q1, Median, Mean, Q2] = [49, 59.0, 59.57, 70])
- Postal Codes, (all over Greece)
- Marital Status(73% Married)
- Education (38% 3rd Lv, 31% HS)
- **Home Status (50% Unknown)**
- Profession (too many categories)
- Email (25 %)
- Internet (42 %)
- Phone flags (56%)
- Stay flag (50.7%)

➤ Total: 89698 clients with demographics



Available information - Economics

Id	Month	Immediate	Insurance	Investment	Business	Consumer	Closed	MB	Housing
1	2014.01	376.01	0	0	0	0	0	0	0.0
1	2014.02	497.30	0	0	0	0	0	0	0.0
1	2014.03	591.67	0	0	0	0	60000	3	0.0
1	2014.04	17035.37	0	0	0	0	0	0	-127749.1
1	2014.05	18035.37	0	0	0	0	0	0	-224712.6
1	2014.06	18028.70	0	0	0	0	0	0	-223748.9
Contributions		Sums	Status	Flag					
		0 60376.01	current	TRUE					
		0 60497.30	current	TRUE					
		0 60591.67	current	TRUE					
		0 77035.37	current	TRUE					
		0 78035.37	current	TRUE					
		0 78028.70	lost	FALSE					



Available information – Economics

[Q1, Median, Mean, Q3]

- **Immediate** [0, 2.8k, 15k, 13.6k]
- **Insurance** [0, 0, 4.9k, 0]
- **Investment** [0, 0, 15.5k, 0]
- **Business** [0, 0, -59, 0]
- **Consumer** [0, 0, -101, 0]
- **Closed** [0, 0, 11.9k, 0]
- **IR** [0, 0, 0.227, 0]
- **Housing** [0, 0, -4k, 0]
- **Contributions** [0, 0, 2.9k, 0]
- **Months in** [10, 20, 14.65, 20]

➤ Total: 89698 · 20 month-clients of data



Quick Reminder – Motivation

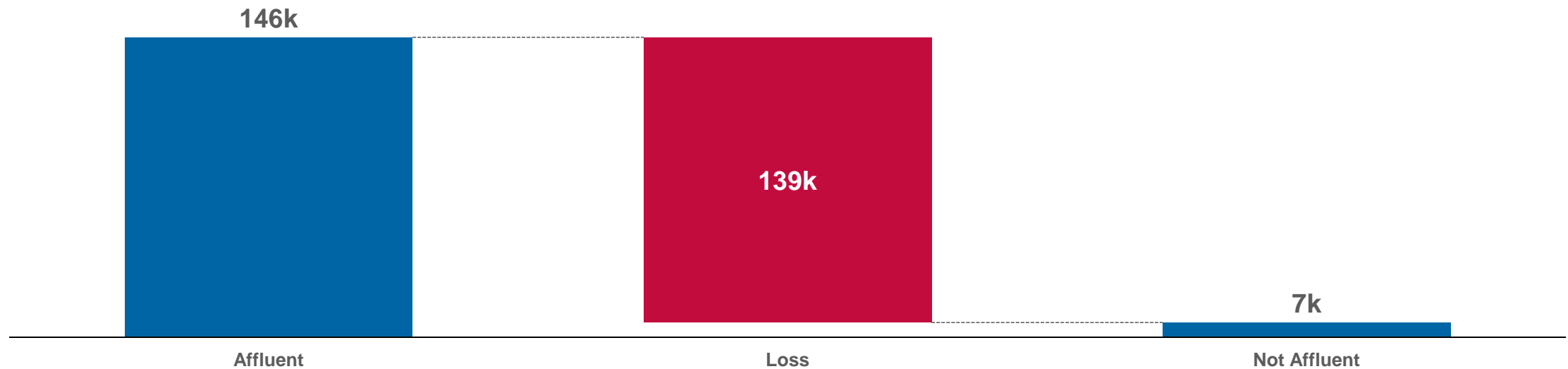
Affluent Clients

- Mean Position: 146k€ per client

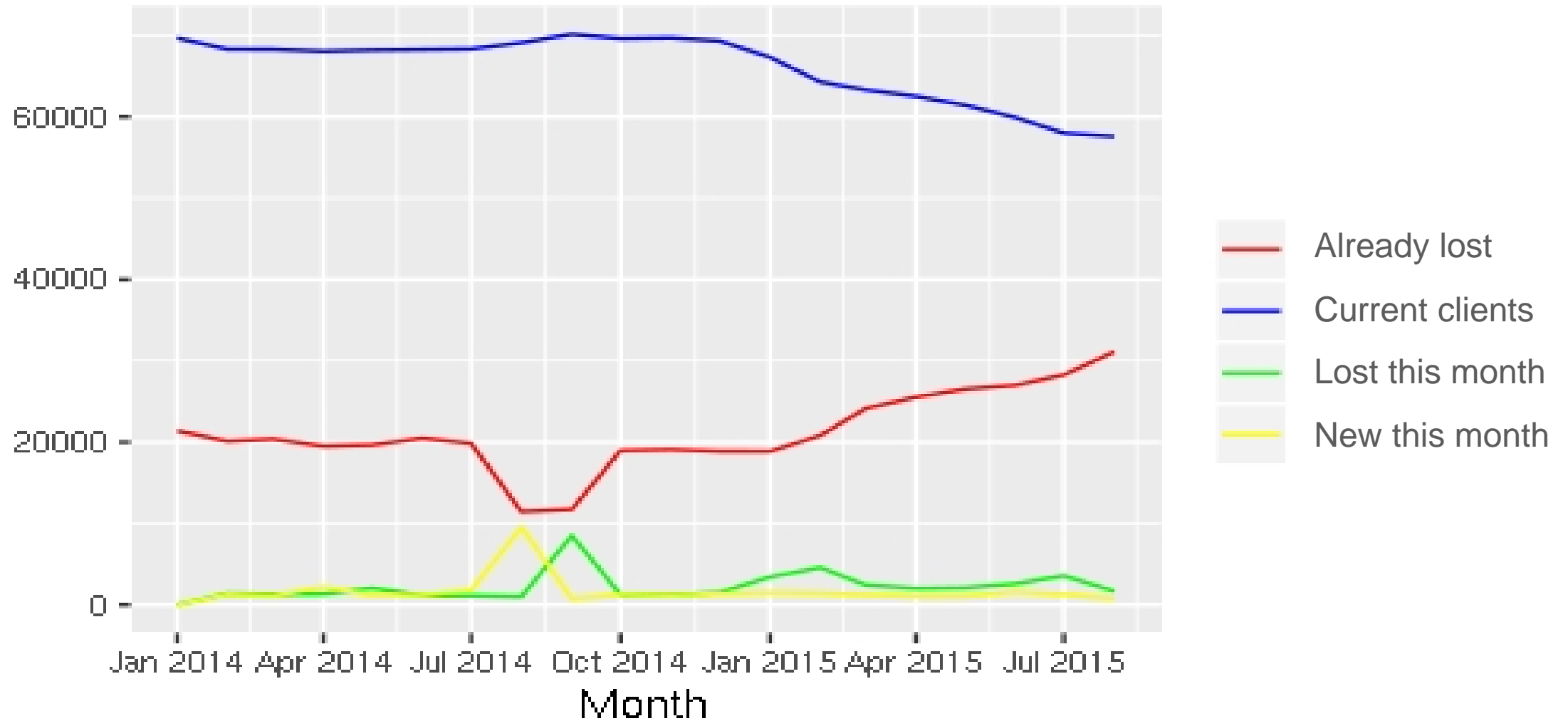
Lost Clients

- Mean Position 7.2k€ per client

$$139k \times 10k = 1.4 \text{ bn€ loss}$$




Loss rate of clients



System Requirements

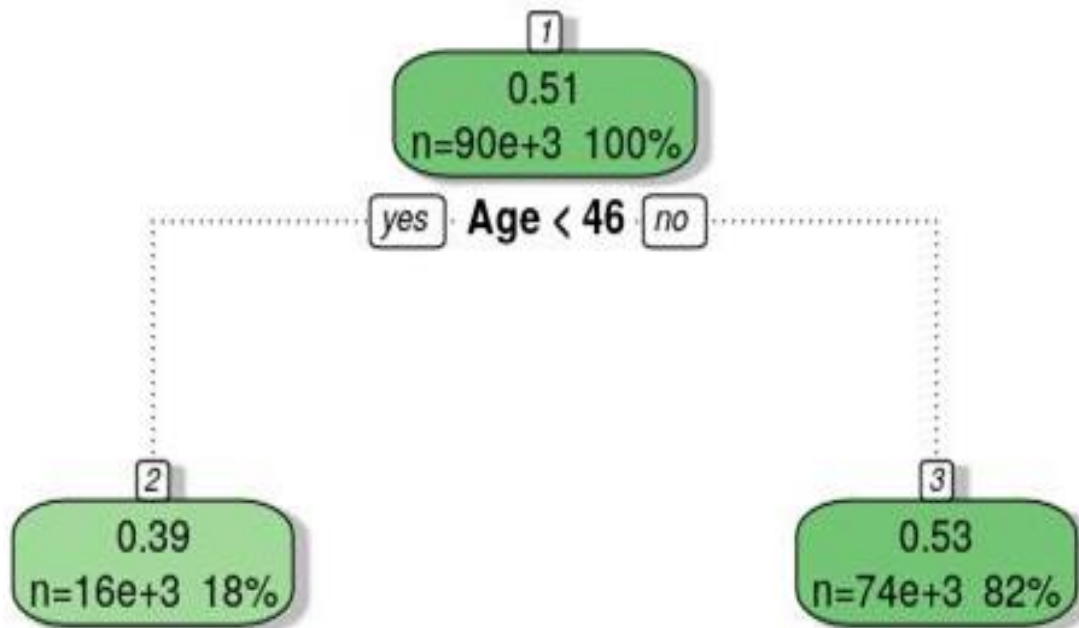
- Client data

- 
- A. Will the client leave the affluent program?
 - B. Does the client belong to a category that implies an action?



A. Using Demographics

Regression Trees

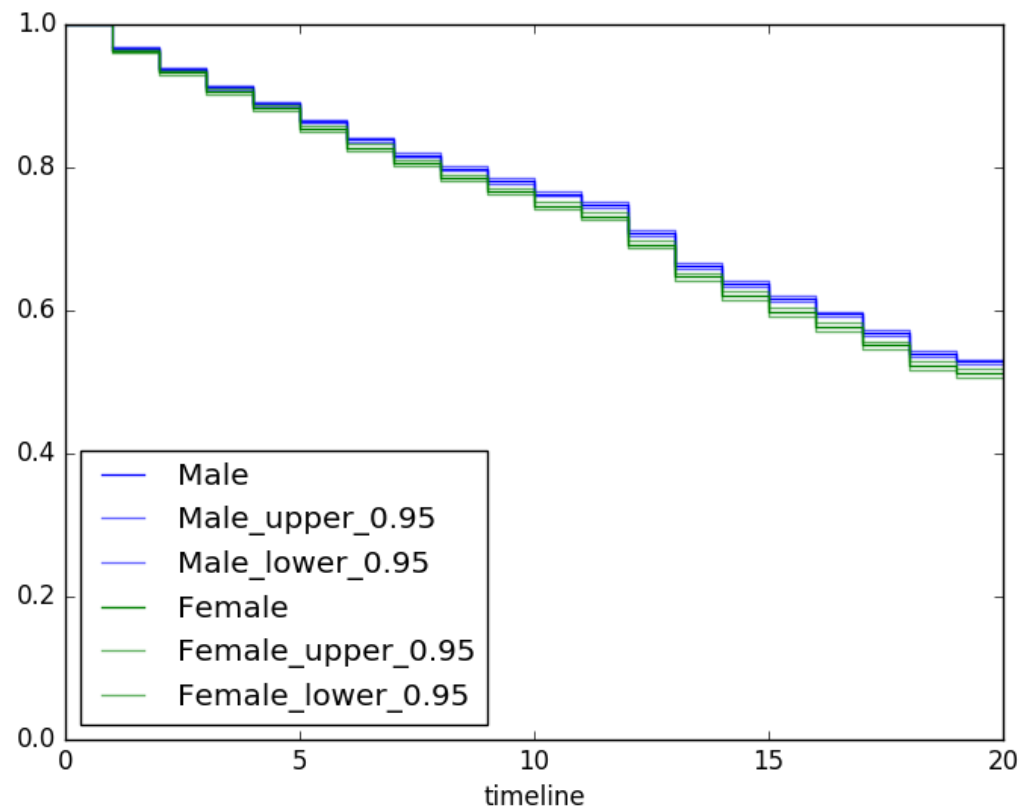


- Age proves to be the only relatively useful demographic
- Can provide an insight to frontliners to be more eager with younger affluent clients



A. Using Demographics

Survival



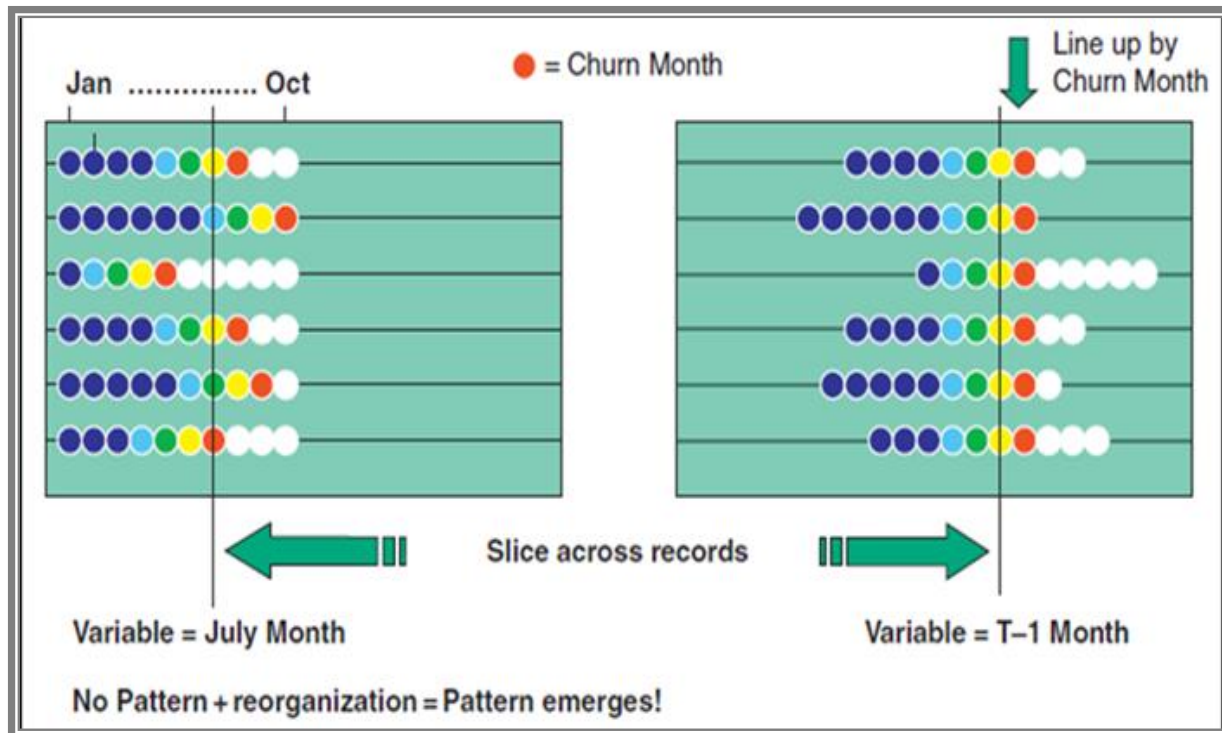
Other variables – e.g. Sex

Both sexes 'die' with the same rate, so this variable is independent



A. Using Economic Data

“Time Series” Analysis



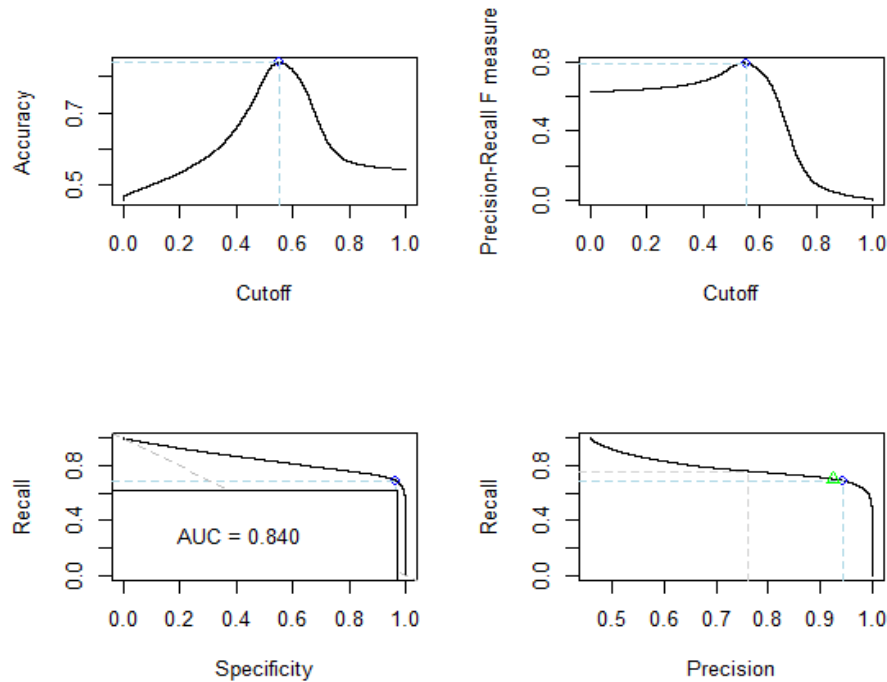
Methodology

- Align triplet of months
- Using $Month_0$ and $Month_1$ predict if $Month_2$ is churn month
- Failed!
- 1,174,496 records, only 2.52% losses
- Systems can predict ‘stay’ with 97.5% success, so they don’t even try



A. Using Economic Data

Summary Statistics/Logistic Regression



Methodology

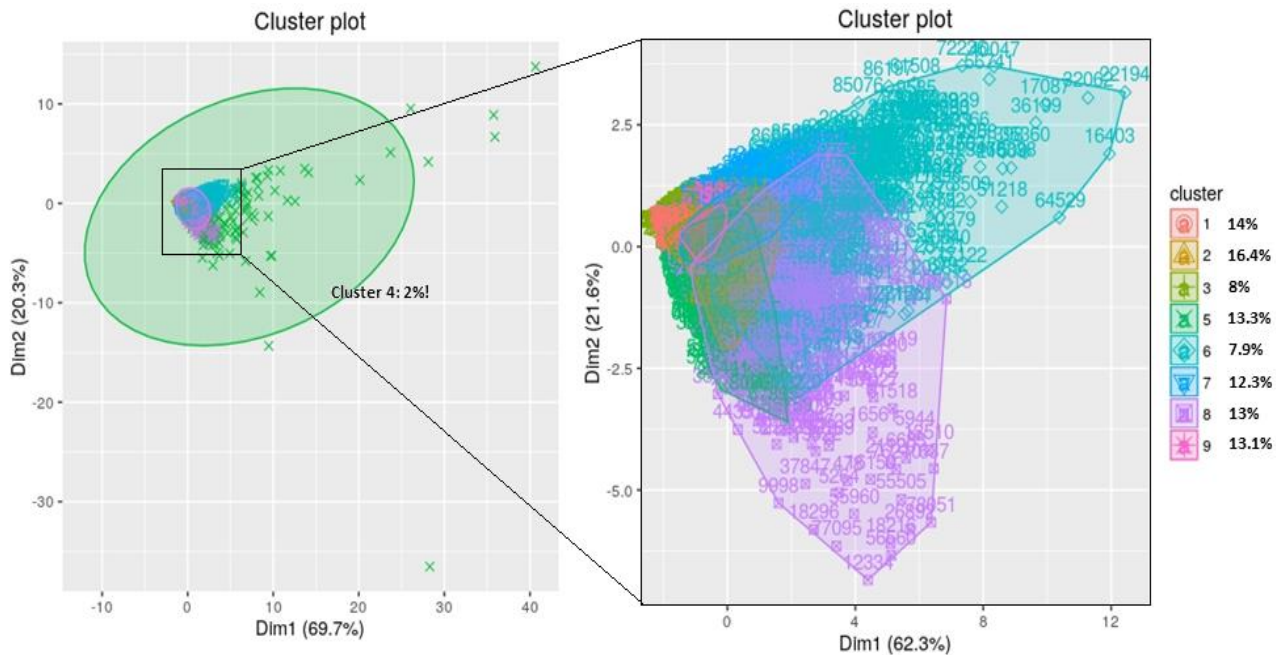
- Calculate summary statistics for the period a client is active
- Learn flag if client churned, any month
- Logistic regression using
 1. Sums_mean
 2. Sums_var
 3. Immediate_var
 4. Closed_var

These variables describe most clients – direction to use from business owner



B. Using Economic Data

Clustering Analysis



Methodology

- Run model clustering with the same variables as above
- Variables from business owner
- Explain cluster results



B. Using Economic Data

Clustering Analysis

Cluster	Stay	Leave	N	Verdict
1	27.2	72.8	11867	Leave
2	85.1	14.9	13870	Stay
3	1.9	98.1	6787	Leave
4	57.5	42.5	1684	Outliers
5	33.7	66.3	11278	Leave
6	60.3	39.7	6653	Stay?
7	42.5	57.5	10413	Leave?
8	65.2	34.8	11042	Stay?
9	91.1	8.9	11134	Stay

If client gets classified in

- Cluster 1, 3, 5, 7, will probably churn.
Actions:
 - storyline for each category
 - specific initiatives
 - flag as high risk, monitor each interaction
- Cluster 2, 6, 8, 9, will probably not churn
Actions:
 - ensuring trust with initiatives
- Cluster 4: Outlier
 - this cluster 'pulls' 0.05% of clients with outlier data



A Combining cluster and regression

The idea is to run both methods

- Classify client in cluster
- If classified in 2, 3, 4, 8, 9 return cluster verdict (from last slide)
- Otherwise return logistic regression result

- Improves performance by 2%

System Layout

Cluster	Sample Population	Model	Accuracy
1	14.0%	Regression	90.2%
2	16.4%	Clustering	85.1%
3	8.0%	Clustering	98.1%
4	2.0%	Clustering	57.5%
5	13.3%	Regression	75.8%
6	7.9%	Regression	74.2%
7	12.3%	Clustering/Regression	57.5%
8	13.0%	Regression	81.6%
9	13.1%	Clustering	91.1%



Achievements

- Understood & Analyzed data
- Provided actionable insights to client institution



Achievements

- Understood & Analyzed data
- Provided actionable insights to client institution
- Finished presentation



Achievements

- Understood & Analyzed data
- Provided actionable insights to client institution
- Finished presentation
- Ευχαριστώ!





Ευχαριστώ