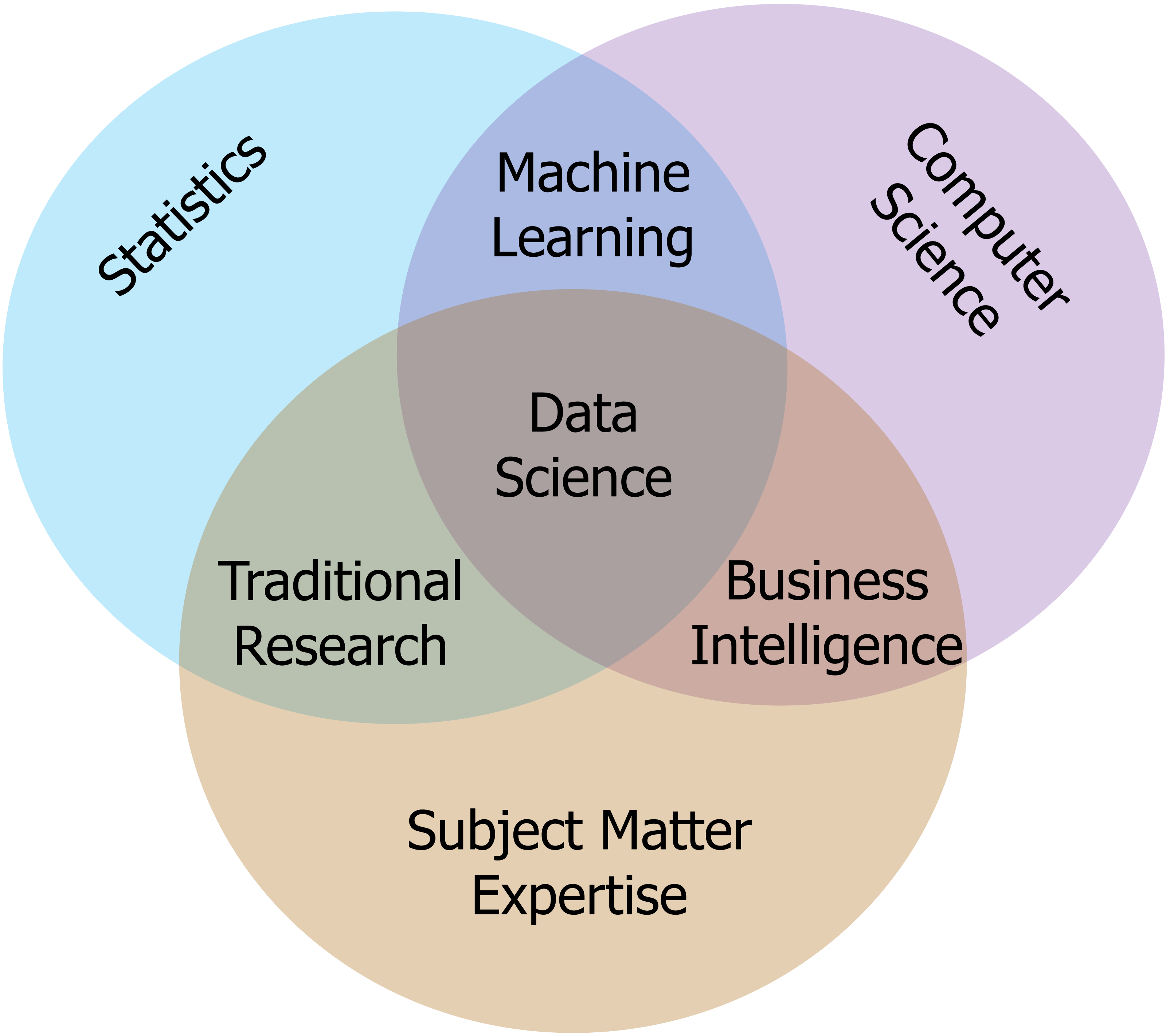
Data Science in Tech and Insurance Industries

March 5, 2018

Fabrizio Lecci

The Data Science Venn Diagram



Data Scientist: Type A or Type B?

> ***Type A Data Scientist:*** *The A is for Analysis. This type is primarily concerned with making sense of data or working with it in a fairly static way. The Type A Data Scientist is very similar to a statistician (and may be one) but knows all the practical details of working with data that aren't taught in the statistics curriculum: data cleaning, methods for dealing with very large data sets, visualization, deep knowledge of a particular domain, writing well about data, and so on.*
>
> ***Type B Data Scientist:*** *The B is for Building. Type B Data Scientists share some statistical background with Type A, but they are also very strong coders and may be trained software engineers. The Type B Data Scientist is mainly interested in using data "in production." They build models which interact with users, often serving recommendations (products, people you may know, ads, movies, search results).*

Robert Chang, Data Scientist at Twitter

Data Analysis Steps

A) Getting the data

B) Exploratory Data Analysis

C) Statistical Modeling / Machine Learning

D) Visualizing and interpreting results

E) Implementation

A) Getting the Data

The first step in Data analysis is getting the data!

A **database** is a collection of information that is organized so that it can easily be accessed, managed, and updated.

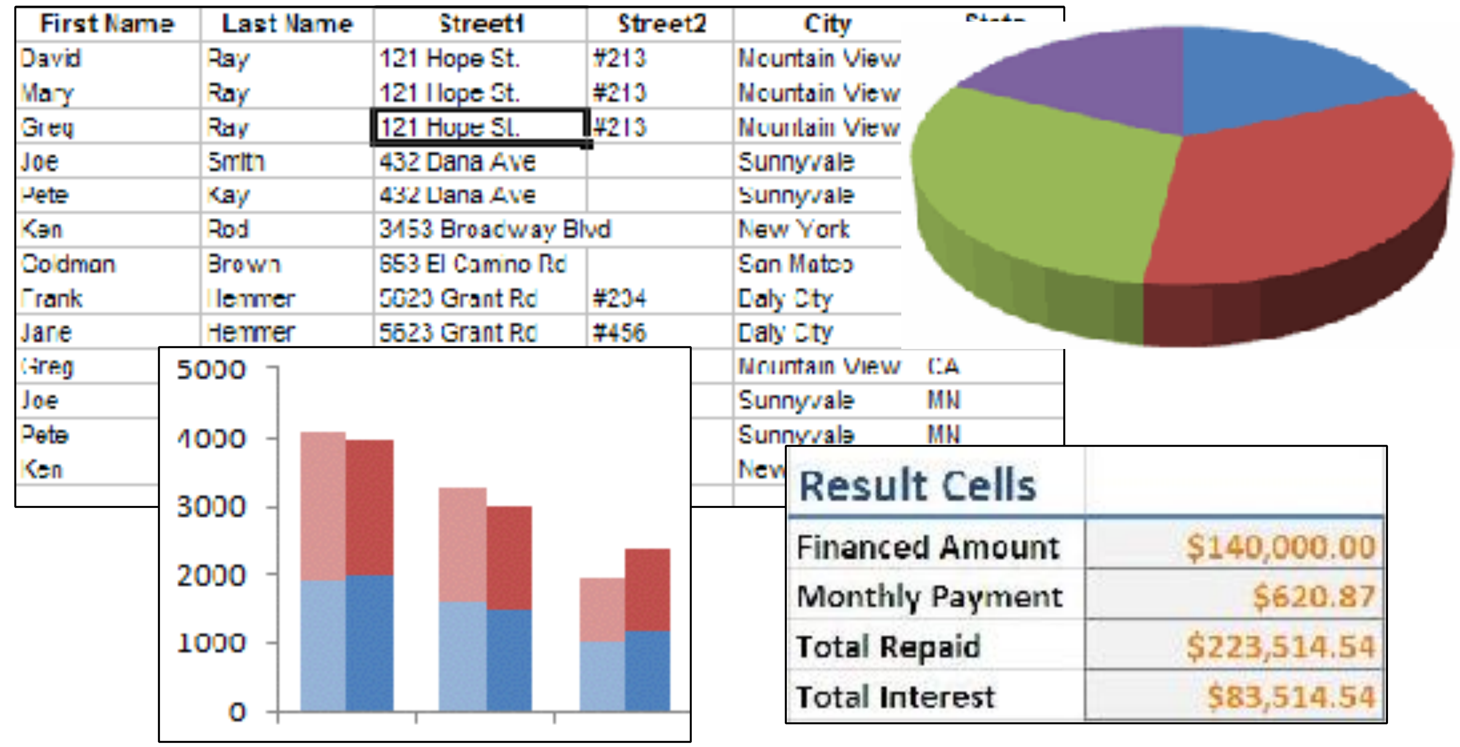
Database Management Systems

ORACLE
MySQL
IBM DB2
Microsoft Access
...
PostgreSQL

Data Architects/Engineers design databases, gather and collect the data, store it, and serve it to data scientists who can easily query it (for example using **SQL**).

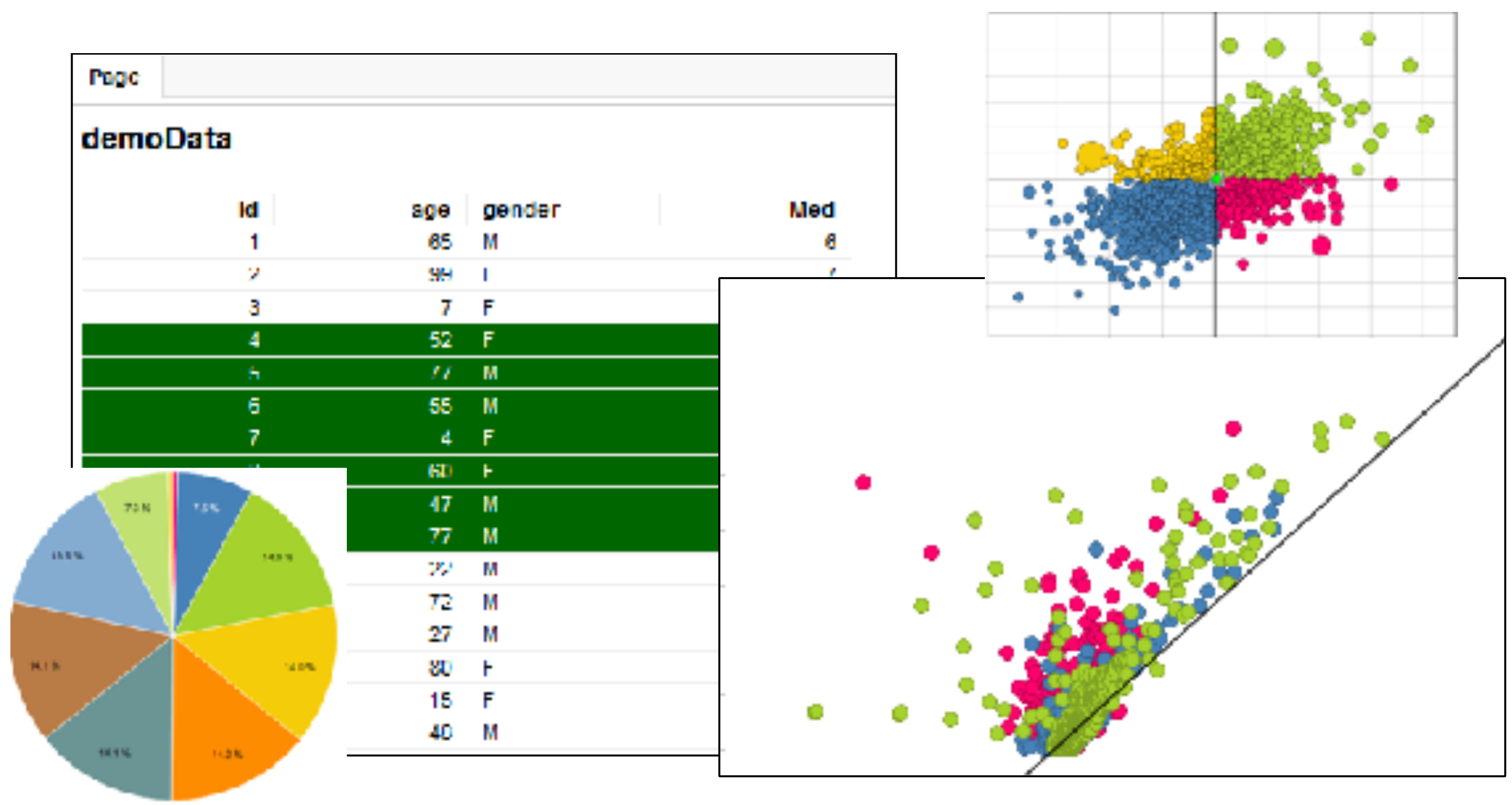
B) Exploratory Data Analysis

Excel

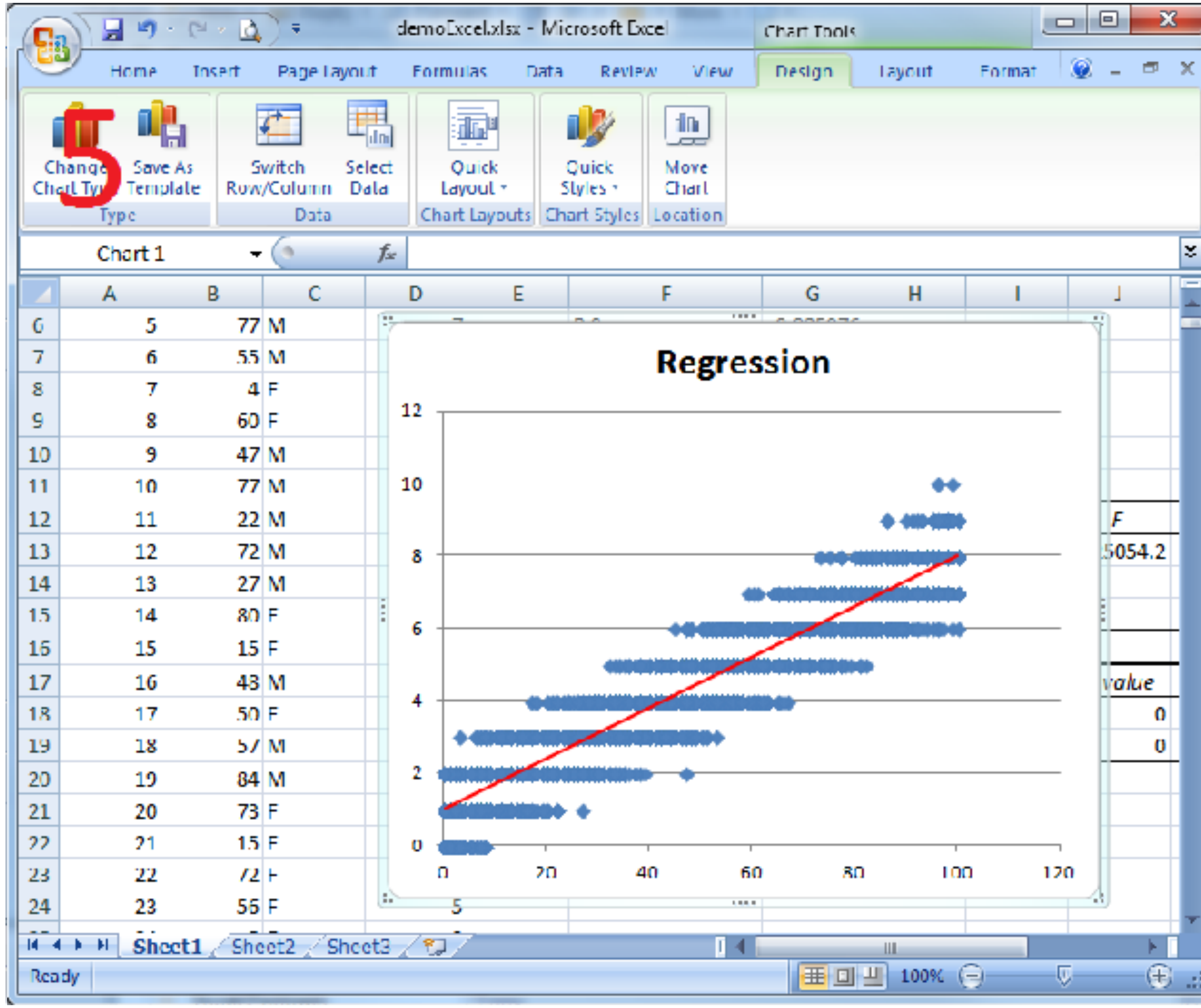


TIBCO Spotfire®

(.. and many more)



C) Statistical Modeling



Can we do
statistical modeling in

Excel ?

Pros

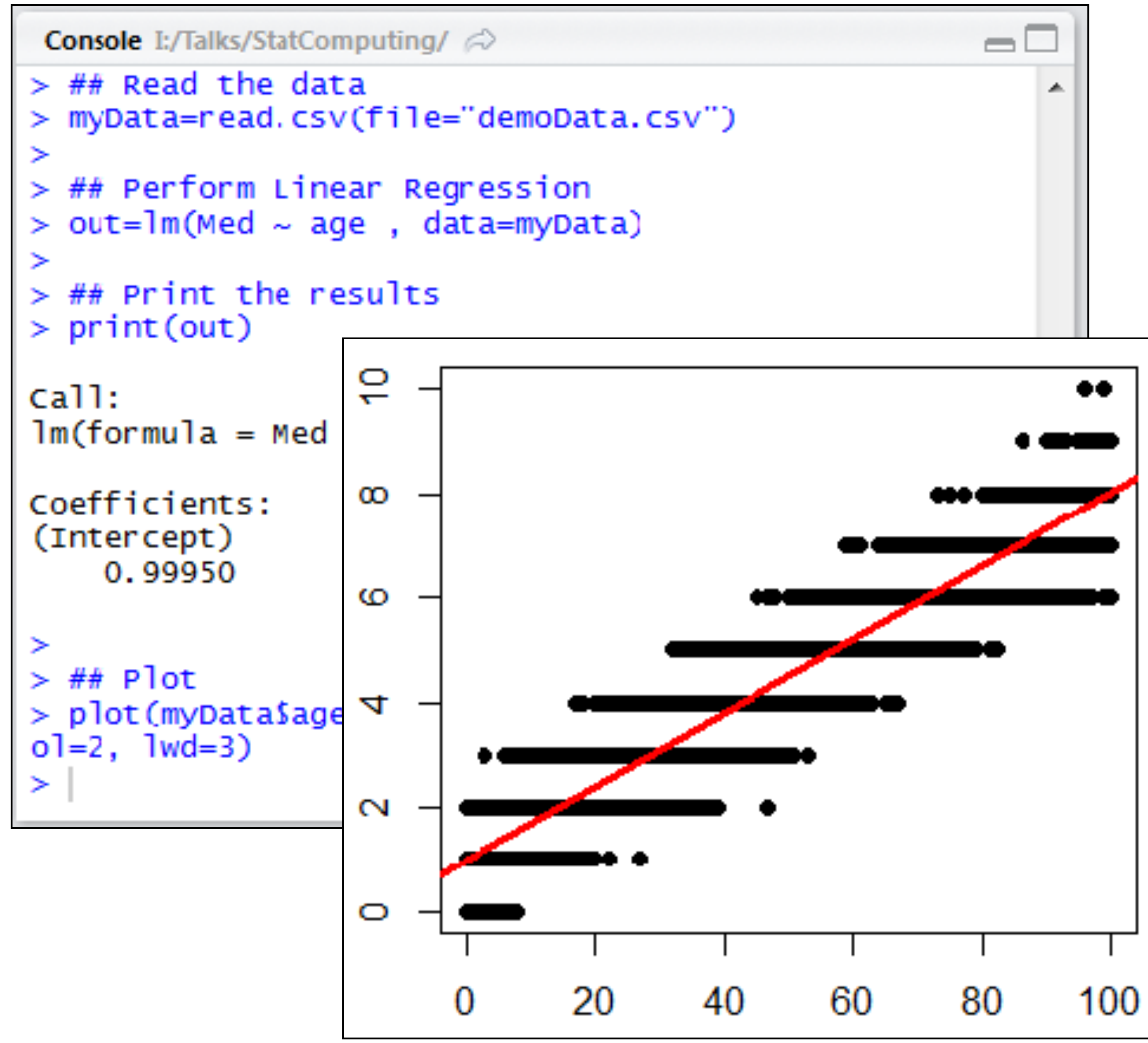
- Easy to use
- Quick and dirty

Cons

- Data Manipulation
- Max ~ 1 Million rows
- No flexibility (hard to correct mistakes on the fly)
- No reproducibility
- Limited automation
- No Advanced Models

C) Statistical Modeling

For Advanced Statistical Analysis we use scripting languages, like..



Pros

- Free and open source
- Relatively easy to learn
- Row limit depends on memory
- Thousands of packages (~ 6,000)
- Good for manipulating data
- Reproducible code
- Easy automation
- Advanced models

Cons

- Not as user friendly as Excel (no point and click)
- Sometimes slower than other tools (Python, Julia, C++, ..)

C) Statistical Modeling

Other software / languages we use for advanced data analysis:

- Free and open-source
- Flexible programming languages
- Large communities of users

- Proprietary Commercial Software
- Older but with good specific properties
  (e.g. Matlab is fast with matrices,
  SAS is useful when dealing with large
  datasets)
- Customer support

- Proprietary Commercial Software
- User friendly (point and click)
- Less flexible than scripting languages
- Customer support

D) Visualization

Data Analysis Tools provide basic visualization capabilities

Excel

R

python

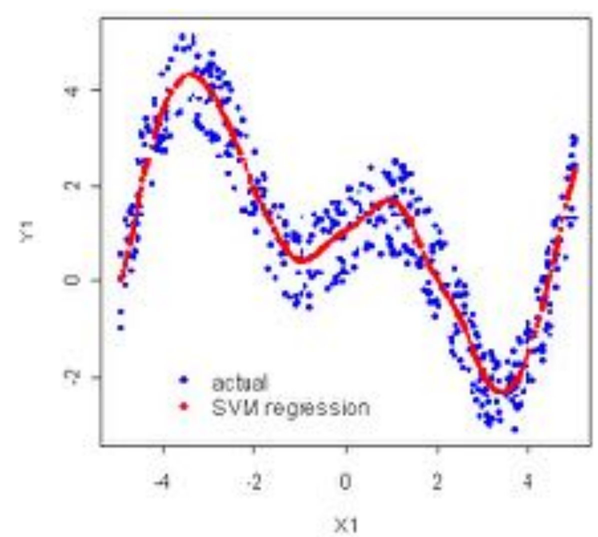
MATLAB

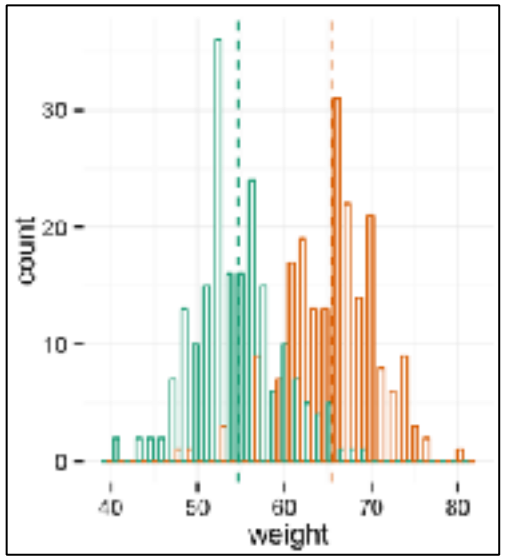
sas

SPSS

stata

…



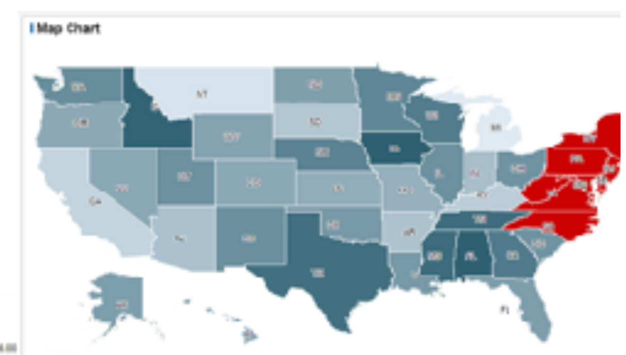


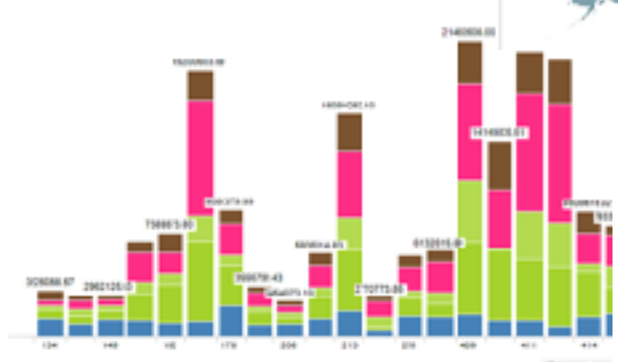
Data Visualization Tools provide advanced (interactive) visualizations

TIBCO Spotfire

tableau

…





E) Implementation

The final results are used for
• Sharing analytically derived insights
• Supporting the business by improving existing rules (data driven decisions)
• Producing new analytical tools (dashboards, applications, …)

General Purpose Programming Languages are designed to be used for writing software in a wide variety of application domains

C++ *Fortran*

Java Ruby

We can use these languages to:

• Speed up the algorithms written in R/Python/Julia.

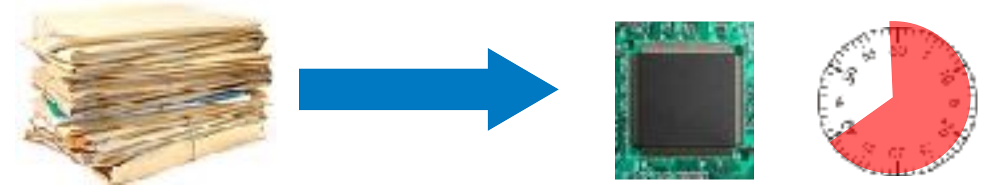
• Develop and manage:
  ○Apps
  ○Web applications
  ○GUI
  ○3D models
  ○Networks
  ○System Administration
  ○…

Technological Challenges and Big Data Solutions

As the amount of data grows and the algorithms become more sophisticated, we have to deal with technological challenges:

-**Memory**: how to efficiently store large datasets
(e.g. > 1 terabyte, hundreds of millions of rows)

- **Speed** how to efficiently read and analyze the data

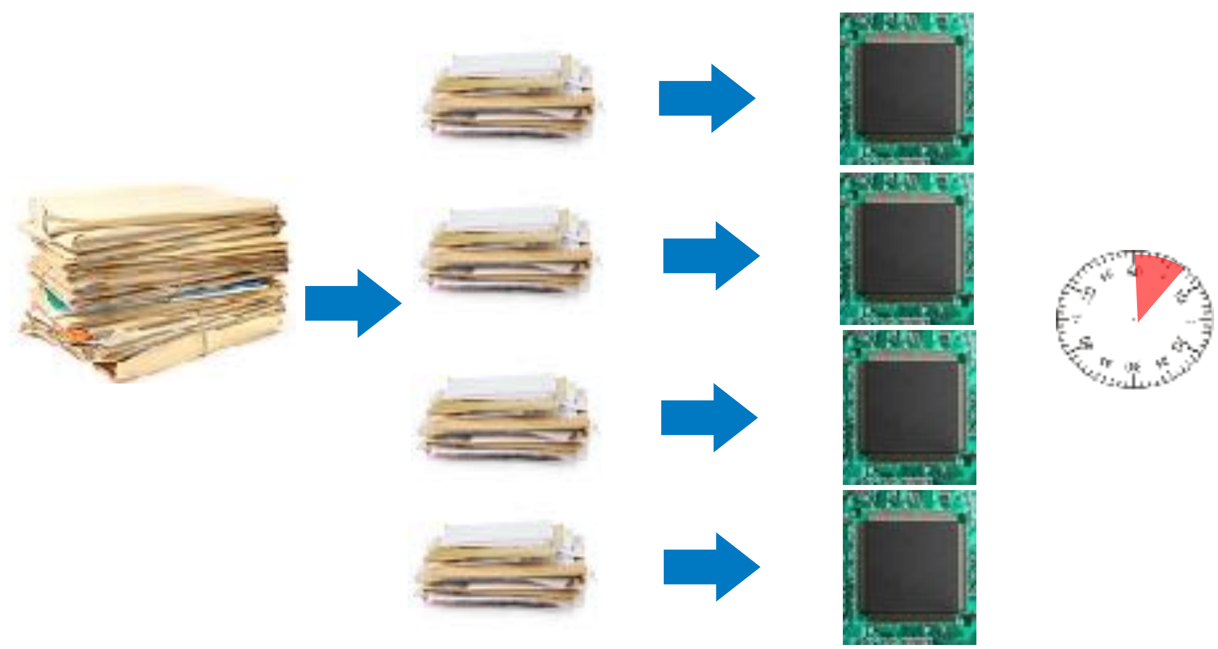


Parallel Computing

Some tasks can be sped up by employing multiple processors to tackle the problem.

For example the following statistical techniques are easily parallelizable: ensemble modeling (e.g. random forests), Monte Carlo Simulations, Markov Chain Monte Carlo, Variable Selection, …

NB: not everything is parallelizable! Some algorithms are inherently serial (e.g. boosting)



In the next slide.. two general approaches for managing large numbers of processors.

Technological Challenges and Big Data Solutions

Supercomputers

- Large systems sitting in 1 room
- Custom hardware

Pros: very powerful, relatively easy to use, good for real-time applications
Cons: cost, energy and heat management issues; scalability.

Examples:
Watson IBM (NY, USA)
2,880 cores, 16 terabytes of RAM
Cost: $3 million

Tianhe-2 (Guangzhou, China)
3,120,000 cores, 1.4 petabytes of RAM
Cost: $390 million

Distributed systems

- Systems of “regular” computers connected via a network

Pros: cheap(er), easy to maintain, scalable.
Cons: more complex to use than a single machine

hadoop

Open-source software framework for distributed storage and distributed processing of very large data sets on computer clusters.

Tools for Hadoop: HIVE Spark …

Applications:
-Computationally demanding **algorithms** (e.g. simulations)
- Handling **big data** from vendors (Acxiom, Experian, ..)
- Handling growing **internal data** (client/contract data, web traffic, …)

What else is out there?

Hundreds of software startups!

“ Sophisticated Machine Learning Made Easy “

“ Become a Data Scientist in seconds “

“ Big Data for everyone “

kyvos insights
SKYTREE THE MACHINE LEARNING COMPANY
Captricity
DataRobot
Dato
DOMO
RedPoint
CONTINUUM ANALYTICS
AYASDI
DataHero
TRIFACTA
H_2O
ARCADIA DATA
ZALONI
tamr

Pros

They try to facilitate our work, by solving technological problems:

- Efficient algorithms’ implementations
- User friendly interfaces
- Automated Signal Discovery
- Easy coding
- Integration of different languages

Cons

- Transparency and flexibility (additional layer between you and the data)
- Risk of using new tools as a black box
- Bugs out of your control!
- Need frequent updates to keep up with new methods
- Cost

Life Insurance

Life Insurance: the **insurer** promises to pay a designated **beneficiary** a sum of money in exchange for a premium, upon the death of an **insured person**.

Two main kind of contracts:

Whole Life and **Term**

We focus on three areas

Marketing

Customers Segmentation

A/B testing

Traditional marketing

Online marketing

Website Data

Agency

Agents segmentation

Agent retention

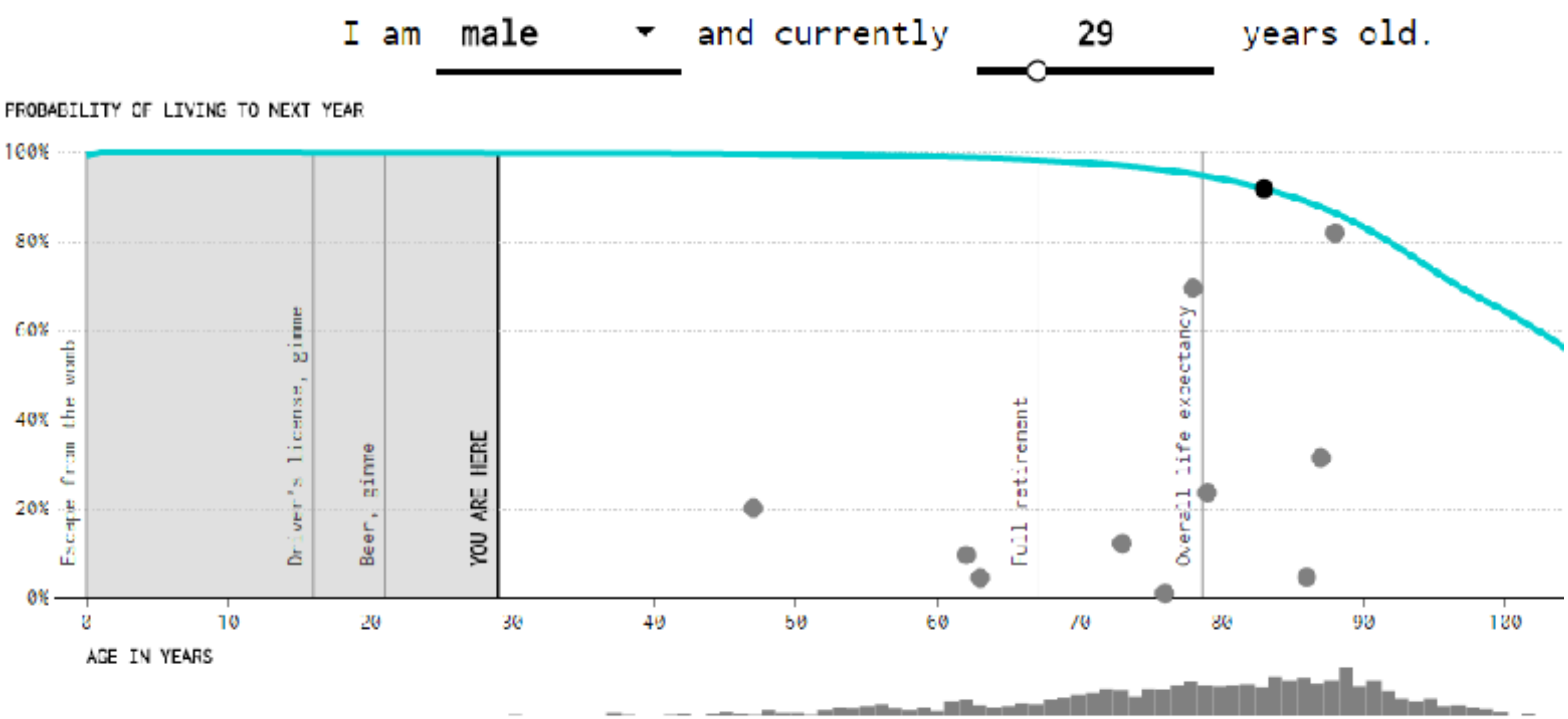
Tools for Agents:
-Customer prospecting
-Social Media
..

Underwriting

The process of collecting a client' information, estimating the risk and issuing the policy

Estimating The Risk

The objective is to **estimate individual mortality** in a precise and interpretable way.



The Problem in Statistical Terms

The data consists of clients and contracts issued in the past, and the corresponding mortality information

Client ID	Age at Issue	Gender	Risk Class	Face	Age at Death
1	45	F	Standard	100k	NA
2	13	M	Juvenile	50k	NA
3	30	M	Preferred	1M	84
4	67	F	Smoker	100k	78
...	...	...	...	...	...

We want to predict the **probabilities of death (mortality rates)** from historical data using features of clients and contracts

It is <u>a supervised learning</u> problem

The outcome variable is death (actually age at death)

Approaches

1- **Counting** claims within cells (**nonparametric**): for each cell (e.g. defined by age at issue and face amount), we count how many people die each year.

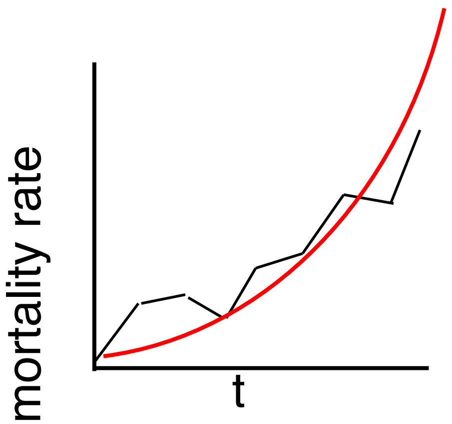
2- **Logistic Model** (**parametric model**): the model predicts the probability of death for each client at each given time t.

3- **Survival Model** (**semiparametric**): the model predicts the entire mortality curve for each client.

Approach 1: Counting

The **cells are defined by a few variables** like Age at issue, Gender, Risk Class and Face Amount.

Face \ Age	0-17	18-30	31-40	41-50
0-100k	n= 324,532	n=456,782	n=523,743	n=487,271
100-250k	n= 150,568	n=393,281	n=421,745	n=421,763
250-500k	n=42,657	n=273,129	n=367,664	n=376,561



For each cell we estimate mortality rates by computing the proportions of clients that die at each time t.

Ad-hoc adjustments are necessary to guarantee consistent rates within and among cells (smoothing)

Some cells might not have enough data (low credibility)

Low interpretability (what are the drivers of mortality?)

Approach 2: Logistic Regression

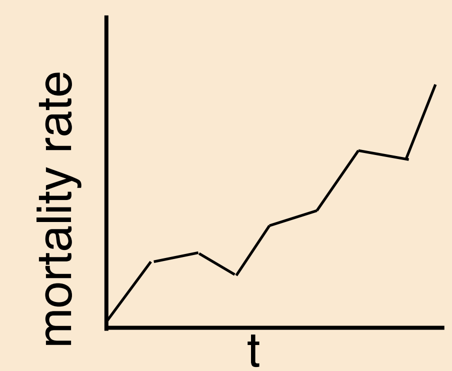
For each contract, for each year, we predict the probability of death.

The model treats <u>all the rows independently</u>, even if corresponding to the same client/contract. Nonetheless, with good modeling principles and feature engineering, it can perform well.

CLIENT_ID	POLICY_ID	YEAR	DEATH_IN (Event)	X... (Predictors)
1	a	2012	No	...
1	a	2013	No	...
1	a	2014	Yes	...
1	b	2013	No	...
1	b	2014	Yes	...
2	c	2014	No	...
2	c	2015	No	...
3	d	2011	No	...
3	d	2012	No	...
3	d	2013	No	...
3	d	2014	Yes	...

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots$$

The coefficients are estimated by maximum likelihood.



Approach 3: Survival Model

For each contract we estimate the **entire mortality curve** simultaneously.

Time of Event

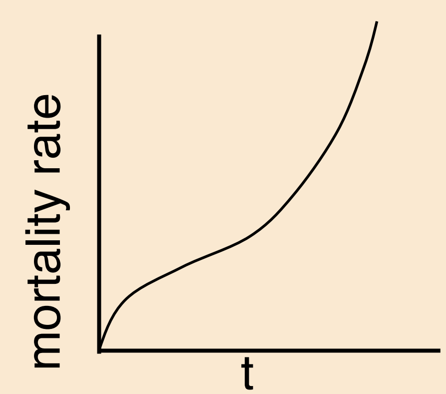
Predictors

CLIENT_ID	POLICY_ID	Age at Issue	Age at End	Death_at_End	X…
1	a	23	38	No	…
2	c	43	81	No	…
3	d	7	76	Yes	…

The hazard (think of it as the instantaneous mortality rate) is modeled as:

$$\lambda(t) = \lambda_0(t) \underbrace{\exp\{\beta_1 X_1 + \beta_2 X_2 + ...\}}_{\text{parametric}}$$

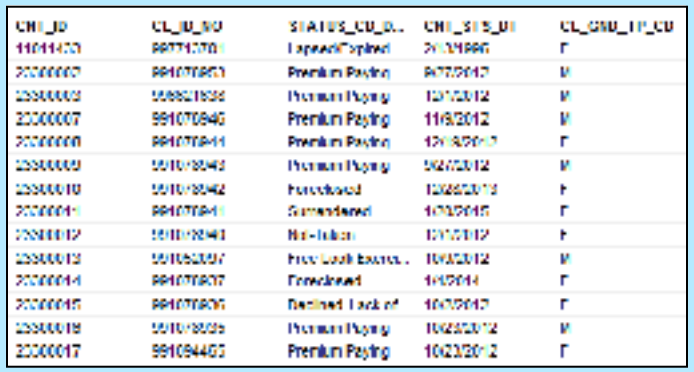
parametric or nonparametric



A survival model is the <u>natural choice for “time to event” data</u>.

Survival Model: Training Step

Input: the Data



<u>Outcome Variables</u>:
Age at End
Death Indicator

<u>Input variables:</u>
Age at Issue
Gender
Mortality Group
Risk Class
Birth Year
Face Amount
…

Survival Model

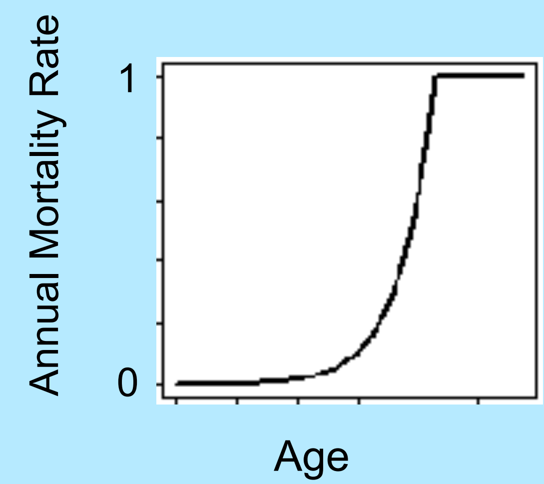
(Cox Proportional Hazards Model)

The model estimates one **nonparametric baseline mortality curve.** (more generally 1 baseline curve for each combination of the most important variables, e.g. Age at Issue)

It also estimates a **coefficient for each of the other predictors**. Positive (negative) coefficients are associated with increasing (decreasing) mortality, that is, the

Model's Estimates

Baseline Mortality Curve



Coefficients

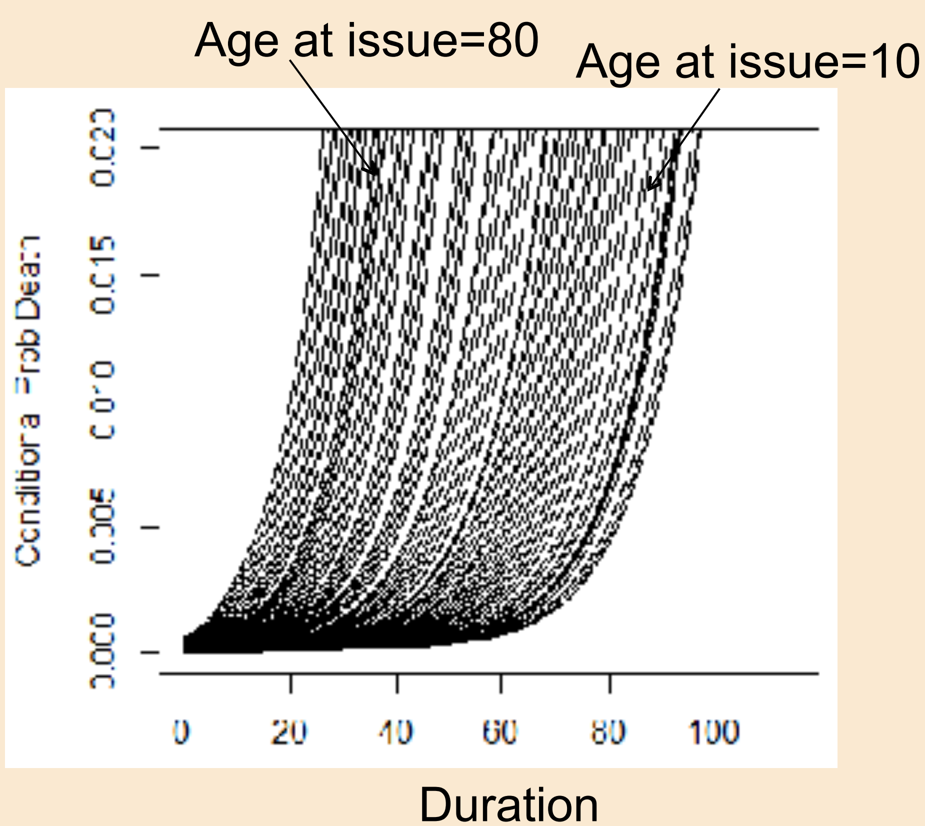
Variable	Coeff.	P<0.05
Female	-0.05	X
Risk Class 1	-0.08	X
Risk Class 2	0.07	
Risk Class …	…	
Birth Year	-0.04	X
Face	0.009	X

The training step is **fast and automated**. The **multivariate** nature of the model guarantees **interpretability** of the impact of each variable on mortality (through the corresponding coefficients).

Survival Model: Scoring Step

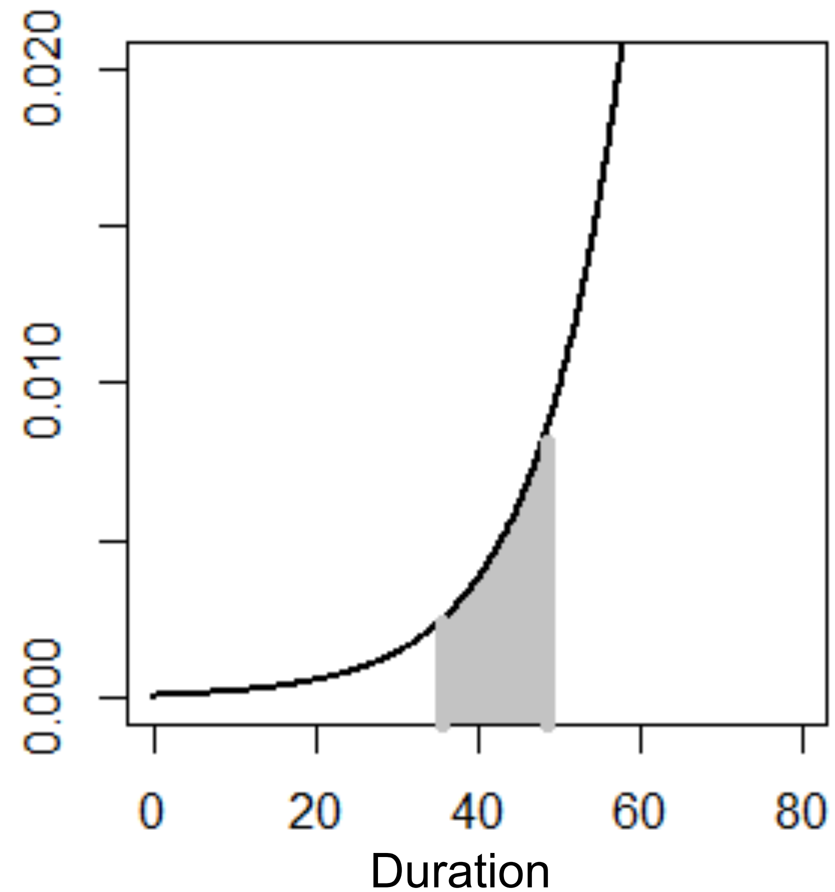
Scoring Output

The scoring step assigns a **mortality curve to each individual**.



These curves contain all the necessary information to study mortality experience.

These curves give us great **flexibility**.

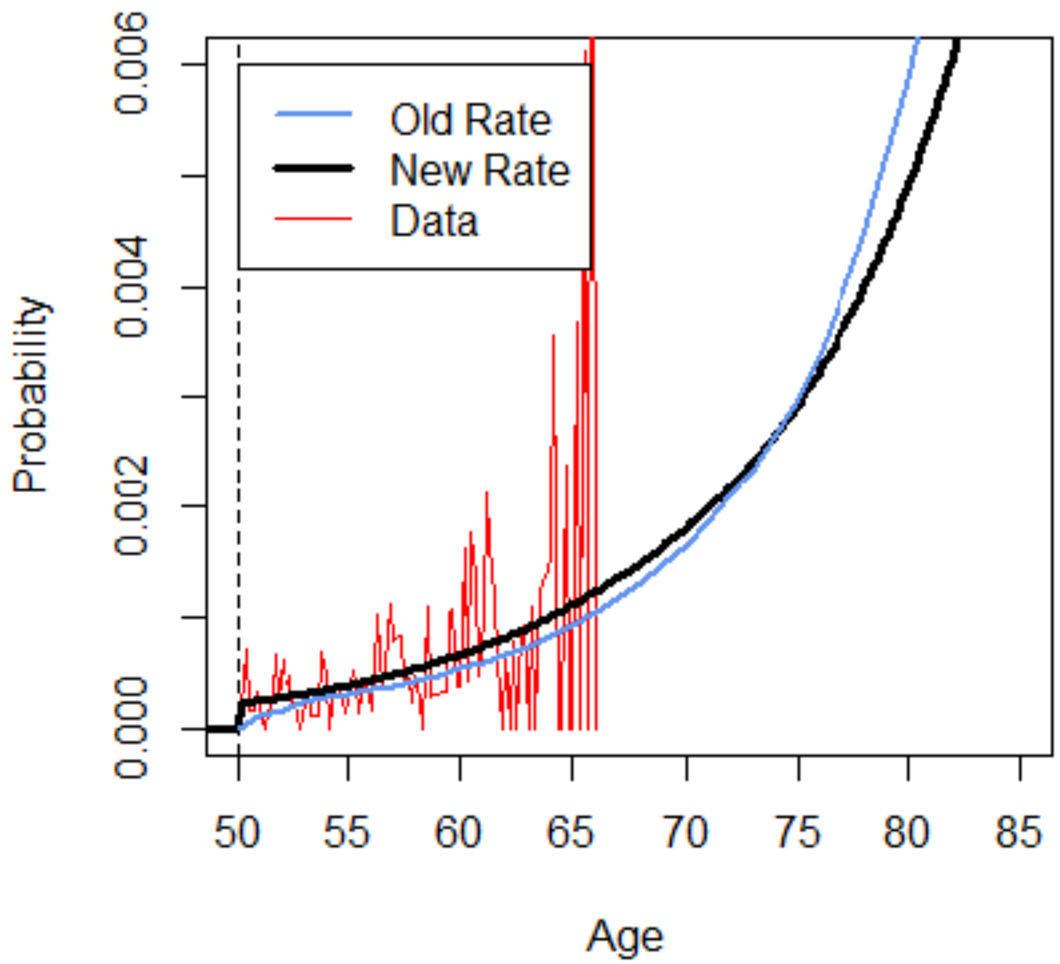


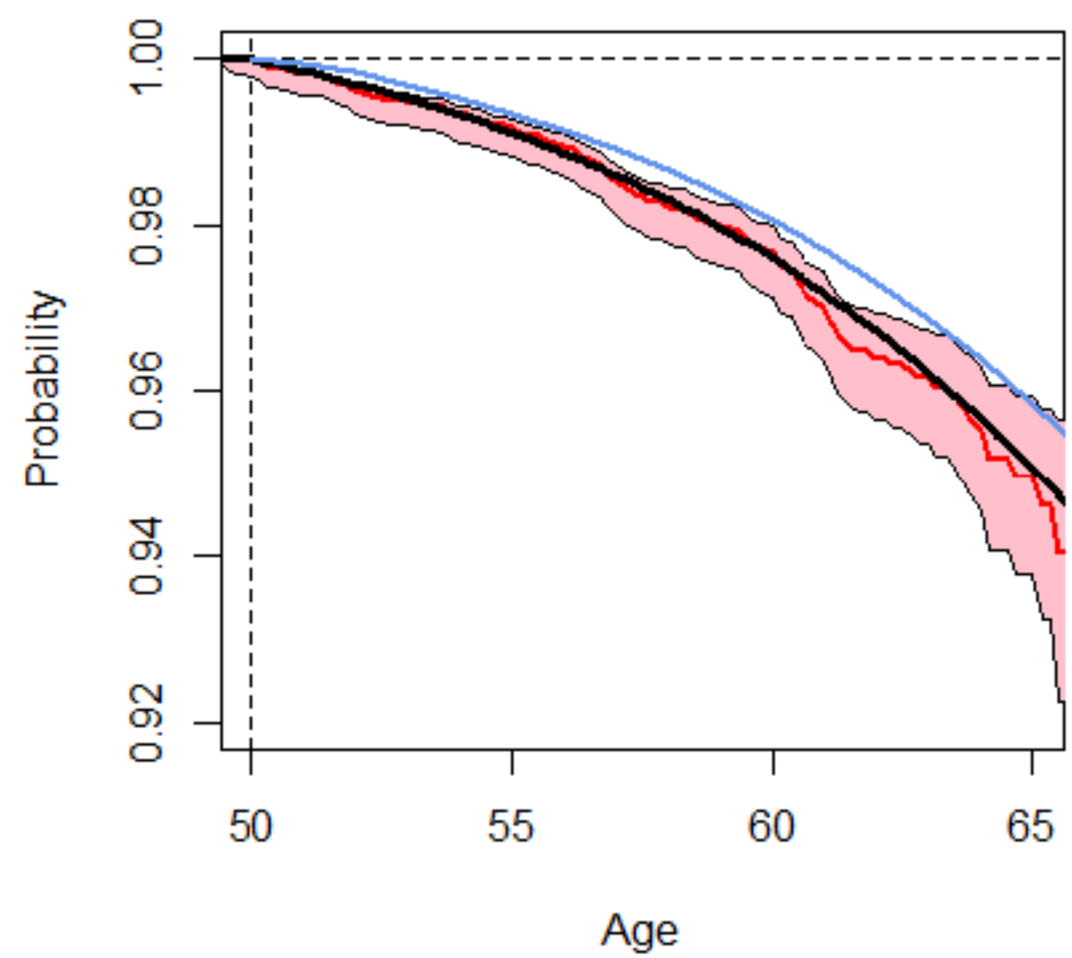
We can study mortality between any 2 given dates at the desired level of precision (year, month, even day)

Example

Male, Age at Issue=50, Born in 1955, Face=100k

Mortality Curve (Annual rates) 1-1 correspondence ⟷ Survival Curve (Cumulative)





The **red curve** is pure data - counting deaths for each year. (in-time, out-of-sample set)
The **blue curve** is the old mortality rate (from logistic model).
The **black curve** is the output of the survival model.
The new rate (black curve) is closer to the real data. The old rates (blue curve) underestimate mortality for the first few years of duration and then overestimate it.

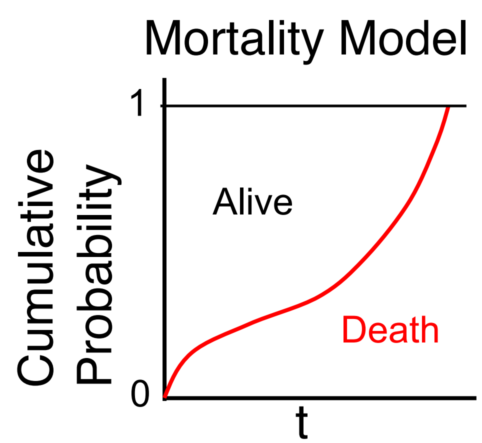
Extensions

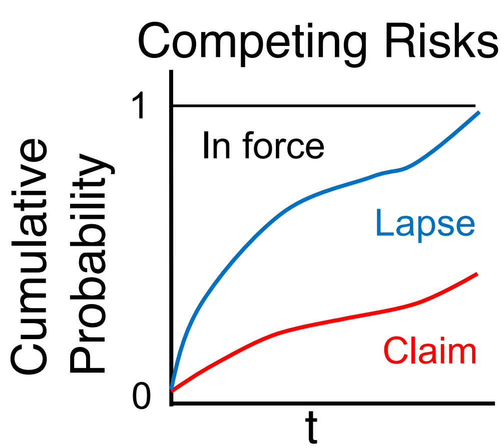
Research

We can add additional predictors (e.g. from Electronic Health Records) to estimate personalized mortality curves. New challenges: getting the data, dealing with many more variables and missing values.

Interaction between Lapse and Mortality

- We currently estimate 2 independent models, 1 for mortality and 1 for lapse.
- We can use the **"competing risks"** extension of the survival model to study their interaction.





Causes of death

In the framework of competing risks we can also study different causes of death.

Tools

Developers

SQL

Tables

CL_ID	CNT_ID	YEAR	RATE
1	a	2012	0.002
1	a	2013	0.0024
1	a	2014	0.003
1	b	2013	0.001
1	b	2014	0.002
2	c	2014	0.0015
2	c	2015	0.0019
3	d	2011	0.001
3	d	2012	0.002
3	d	2013	0.003
3	d	2014	0.004

Business Users

Thank you!