Let's pretend there are three causal drivers:

z_1: has sufficient funds to pay back loan at the time it's due?

$z_1 \in \{0,1\}$

z_2: unforeseen emergency?

$z_2 \in \{0,1\}$

z_3: criminal intent?

$z_3 \in \{0,1\}$

$$y = t(z_1, z_2, z_3) = z_1(1-z_2)(1-z_3)$$

Problems in practice?
(1) You don't know the z's because they are realized in the future.
(2) You may not know the function t which can be very complicated.

What is the next best thing since you have to make a decision now and you need a model that works now?

You obtain information that approximates the information in the z's and combine this information to approximate y. We denote these proxies that do this approximation the x's and we denote p to be the number of such proxies: $x_1, x_2, \ldots, x_p$. For example:

x_1: salary at the time of loan application $\in \mathbb{R}$
x_2: missing payments previously $\in \{0,1\}$
x_3: criminal charge in the past $\in \{0,1\}$
=> p = 3

x_j's are called features, characteristics, attributes, variables, independent variables, covariates, tregressors.

What is normally done in the real world? You use the features that are available.

To learn from data, you measure the x_j's on subjects i = 1...n.

Let $\vec{x}_i := [x_{i,1}, x_{i,2}, \ldots, x_{i,p}] \in \mathcal{X}$, the input space

Subjects are also called observations, settings, records, objects, inputs.

$x_2 \in \{0,1\}$ binary variable
$x_1 \in \mathbb{R}$ continous variable
x_3 is a binary variable
} types / names of variables

Let's consider measuring x_3 differently:

$x_3 \in$ {none, infraction, misdemeanor, felony} this is an ordinal categorical variable

How do we make this a metric?

(1) Code it in order of severity spacing by 1:

$x_3 \in \{0, 1, 2, 3\}$

$$\cancel{x \in \{0, 1, 5, 99\}}$$

Downside: coding is arbitrary

(2) Binarize / dummify this categorical variable:

$x_{3a} \in \{0,1\}$ infraction or not?
$x_{3b} \in \{0,1\}$ misdemeanor or not?
$x_{3c} \in \{0,1\}$ felony or not?

One variable became 3 variables. => p = 5
I had 4 levels (L = 4) but now I made L - 1 = 3 variables. Why?
You can capture the last category (called the reference category) by setting all "dummies" / binary variables to zero.

If the variable is "nominal categorical" meaning no inherent order, you must do #2 to be able to use it in a model e.g.

$X \in$ {red, blue, green, yellow, purple, brown,...}

Can we say that y = f(x_1, x_2, ..., x_p)?
"No! It is only approximating it at best" -Gabriel

y = t(z_1, ..., z_t) where you don't know t or the z's.

$$y \approx f(x_1, \ldots, x_p) \qquad f(x_1, x_2) = \frac{1}{1 + e^{\underset{\uparrow}{a}x_1 + \underset{\uparrow}{b}x_2}}$$

$$\text{or } y = f(x_1, \ldots, x_p) + \delta, \quad \text{s.t.} \quad \delta = t - f$$

What is delta? It's an error. It's error due to... ignorance. Ignorance of the true causal drivers. It's the error due to the fact that the proxies aren't the real thing. You're missing information.

How do we decrease delta? Increase p with more useful variables.

How do we get f? Note that there is no "analytical solution". The approach we use is "learning from data". This is an "empirical approach". There are many flavors. We will concentrate on "supervised learning" from "historical data". This requires three ingredients:

(1) Training Data

$$\mathbb{D} = \{\langle \vec{x}_1, y_1 \rangle, \langle \vec{x}_2, y_2 \rangle, \ldots, \langle \vec{x}_n, y_n \rangle\}$$

these are n historical examples of inputs / outputs.

Alternate notation:

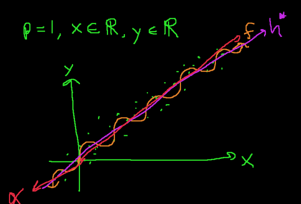
$$\mathbb{D} = \langle X, \vec{y} \rangle \text{ where } X = \begin{bmatrix} \leftarrow \vec{x}_1 \rightarrow \\ \leftarrow \vec{x}_2 \rightarrow \\ \vdots \\ \leftarrow \vec{x}_n \rightarrow \end{bmatrix}, \quad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

(2) $\mathcal{H}$:= a set of candidate functions with elements h that approximate f. We need this because the space of all functions is too large and too ill-defined to directly find the "best one". You need to limit this space!

(3) We need $\mathcal{A}$:= the algorithm that takes in $\mathbb{D}, \mathcal{H}$ and returns g, an approximation to f, g = $\mathcal{A}(\mathbb{D}, \mathcal{H})$.

Is is true that f $\in \mathcal{H}$? No. f is arbitrariliy complicated and unknown and the set curly-H contains usually simple functions that can be fit with curly-A.

However, there is a $h^* \in \mathcal{H}$ which is the candidate model that most closely approximates f. Here is an example:



$$f(x) = x + 0.1\sin(x)$$

$$\mathcal{H} = \{\text{all linear models}\} = \{b_0 + b_1 x : b_0 \in \mathbb{R}, b_1 \in \mathbb{R}\}$$

$$g = \mathcal{A}(\mathbb{D}, \mathcal{H})$$

$$y = h^*(\vec{x}) + \varepsilon$$

$$= h^*(\vec{x}) + \underbrace{\underbrace{(f(\vec{x}) - h^*(\vec{x}))}_{\text{model misspecification error}} + \underbrace{(t(\vec{z}) - f(\vec{x}))}_{\delta \text{ (ignorance error)}}}_{\varepsilon}$$

$$y = \underbrace{g(\vec{x})}_{\text{model}} + \underbrace{e}_{\text{residual (the "full error" the difference between predicted and observed)}}$$

$$= g(\vec{x}) + \underbrace{\underbrace{h^*(\vec{x}) - g(\vec{x})}_{\text{estimation error}} + \underbrace{\underbrace{f(\vec{x}) - h^*(\vec{x})}_{} + \underbrace{t(\vec{z}) - f(\vec{x})}_{\delta}}_{\varepsilon}}_{e}$$

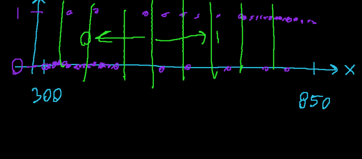
How to we decrease model misspecification error?
Expand the set of candidate functions $\mathcal{H}$ to be more complicated and thus more expressive of complex relationships.

How do we decrease estimation error?
Increase sample size n (more historical examples). The rows in $\mathbb{D}$.

Back to the loan example where y \in {0, 1}.
Let's say we have p = 1 feature, the credit score:
x \in [300, 850]. So your training data looks like:

$$\mathbb{D} = \langle X, \vec{y} \rangle = \left\langle \overbrace{\begin{bmatrix} 810 \\ 390 \\ 750 \\ \vdots \end{bmatrix}}^{X}, \overbrace{\begin{bmatrix} 1 \\ 0 \\ 1 \\ \vdots \end{bmatrix}}^{\vec{y}} \right\rangle \updownarrow n$$

Let's plot the training data:



What is the "null model" g_0 which is the model if you didn't have any x's whatsoever?

$$g_0 = \text{Mode}[\vec{y}]$$

What is the simplest possible candidate space $\mathcal{H}$?

$$\mathcal{H} = \{\mathbb{1}_{x \geq \theta} : \theta \in \mathcal{X}\} \text{ e.g. } g(x) = \mathbb{1}_{x > 600}$$