Predictive Modeling as AI

how it works

(Gowder, Policy lab, September 2015)

Modeling Trained Human Judgment

Examples

INPUT DATA	JUDGMENT RULES	OUTPUT DATA
Text of emails (incl headers with sender etc.)	If about X, relevant, if between Y and Z, privileged…	Classification: produce email/don't produce email
Violations found, facts found by jury	Judicial sense of severity of crime	Sentence: fine/term of years
Text of contracts	Apply rules of contract law	Classification: enforceable/ not enforceable

Computer judgment

predictive coding: the low-hanging legal fruit

HUMAN CODED INPUT/ OUTPUT PAIRS

From: Sally Johnson, CEO
To: Larry Lawyer
Re: Litigation plan

PRIVILEGED

Larry, we absolutely can't let Jake testify, he knows all the skeletons in our closet.

From: Jacob Aziz
To: Sally Johnson
Re: Product Safety

RESPONSIVE

Sally, engineering refuses to listen to me, but I'm telling you, this car is going to kill people if we don't do something about it quickly. We need to issue a recall.

From Alfred Accountant
To: Sally Johnson
Re: Taxes

NOT RESPONSIVE

Boss, our taxes are all filed. Everything looks ok.

x(a sufficiently large number)

Computer Representation

```
if to/from = larry lawyer, increase
probability e-mail is privileged

if subject contains “safety” or
“explosions,”  increase probability e-
mail is responsive

…etc.
```

Based on pool of human classification matching this pattern.

COMPUTER JUDGMENT ON NEW DATA

From: **Larry Lawyer**
To: Sally Johnson
Re: **Legal** Fees

PROBABILITY OF RESPONSIVE: .0031
PROBABILITY OF PRIVILEGED: .9208
EVALUATION: DO NOT PRODUCE

Last month, we spent 300 hours preparing a **motion** to **dismiss**, at $500/hour. Please remit $150,000 within thirty days.

From **Edward Engineer**
To: **Jacob Aziz**
Re: **Explosions**

PROBABILITY OF RESPONSIVE: .882
PROBABILITY OF PRIVILEGED: .0021
EVALUATION: PRODUCE

Jake, I don't give a damn about the **explosions**. It would cost us over ten thousand dollars a unit to shield the **gas tanks**. We'd never turn a profit again. I'm the head **engineer**, and I say let the babies die. We're going to market *next week*.

Upshot

Without machine learning: lawyers classify 80,000 documents at $500/hour.

With machine learning, lawyers classify 10,000 documents (or 5,000, or however much it takes to get a good model), the computer does the rest.

(There are inevitable errors, but if small enough then they can be acceptable.)

Digging into the details some

Statistical learning:

find decision rule that minimizes loss function between predicted and actual output.

Example:

Linear regression minimizes squared error between predicted and actual output values by finding a linear equation to map predictor (input) to predicted (output) values.

a stupid example
with some made-up data

Number of Hours Spent Studying in Law School	Annual Income
10	$2,000
100	$20,000
1000	$200,000
2000	$400,000
50	$10,000

Best fit: y = 0 + 200x
(x = study hours, y = annual income)

Then if we see another lawyer who studied 5,000 hours in law school, but we don't know how much they make, what's our best guess?

Machine learning is basically just that, but with more data, fancier mathematical models, and a bunch of computational power.

the “data science” process

1. Get and clean labeled data.
2. Create training and testing sets (n.b. “bias-variance tradeoff”).
3. Try feature representations and models until you get good outcomes on training set.
4. Test on testing set.
5. Repeat 3 & 4 until satisfied (automation helps*).
6. Apply to unlabeled data.
7. Profit!

* Don’t worry, before long the robots will get their jobs too.