XII

Porazil bys hloupého soupeře?

Po této kapitole nezačneš u žádného modelu ani oslnivého AI dema otázkou „jak je dobrý?“, ale otázkou „koho poráží?“ — a poznáš, že devadesátiprocentní přesnost může být bezcenná, když je nemoc vzácná.

Nejsou to proroctví ani předpovědi v hadačském smyslu; slovo forecast *přísluší jen takovému úsudku, jaký je výsledkem vědeckého skládání a výpočtu.*

— Robert FitzRoy, *„Prophecies and predictions they are not...“ · The Weather Book: A Manual of Practical Meteorology, 1863*

Předpověď, které se smáli

Roku 1854 zřídil britský parlament při ministerstvu obchodu meteorologický úřad — a do jeho čela postavil Roberta FitzRoye, bývalého kapitána lodi *Beagle*, na jejíž palubě kdysi vozil mladého Darwina kolem světa. Úkol zněl střízlivě: sbírat od lodí měření větru a tlaku, aby se z nich dala jednou počítat bezpečnější námořní cesta. Nikdo tehdy nemluvil o předpovídání počasí; to slovo ještě pořádně neexistovalo.

Pak přišla noc z pětadvacátého na šestadvacátý října 1859. Přes Britské ostrovy se přehnala pomalá, zničující bouře; jen za ty dvě noci poslala ke dnu na sto třiatřicet lodí, mezi nimi parní klipr *Royal Charter* s nákladem zlata z Austrálie, a připravila o život přes osm set lidí. FitzRoy věděl, že tlakoměry na pobřeží pokles ohlašovaly hodiny předem — jen to nikdo nespojil dohromady a nikomu to neřekl. Postavil proto síť pobřežních stanic, které mu telegrafem hlásily měření, a začal z nich v únoru 1861 rozesílat do přístavů *bouřková varování*: soustavu kuželů vytažených na stožár, když se blížil vichr.

A prvního srpna 1861 udělal krok, který dějiny zapamatovaly: v *The Times* vyšla jeho první veřejná předpověď počasí na světě. Pro slovo, které potřeboval, si musel dojít sám — razil termín *forecast*, „napřed-hled“, a v knize o dva roky později pečlivě vysvětlil,

co jím míní a co ne: proroctví to nejsou, jsou to úsudky z měření a výpočtu. Právě o tenhle rozdíl se pak vedl celý spor jeho života.

Námořníci mu žehnali — varování zachraňovala lodě i posádky. Ale tisk se mu vysmíval za každý den, kdy se předpověď netrefila, a vědecká honorace jím pohrdala: seriózní přírodověda přece nehádá, co bude zítra. Postoj té doby se dá shrnout do jediné věty, kterou nad ním visela jako soud — *hádání není věda*. FitzRoy nesl posměch, klesající zdraví i finanční ruinu; třicátého dubna 1865 si vzal život.

Po jeho smrti sestavila Královská společnost komisi — seděl v ní i mladý Francis Galton — a ta roku 1866 předpovědi zrušila jako nedostatečně vědecké. Bouřková varování obnovila veřejná poptávka už napřesrok; na veřejnou předpověď počasí si Británie musela počkat až do roku 1879. FitzRoye tak jeho vlastní obor napřed vysmál a pak posmrtně umlčel — a za pár let ho tiše vzkřísil. Zbyla po něm otázka, na kterou tehdy nikdo neuměl odpovědět a která je jádrem téhle kapitoly: *byla FitzRoyova předpověď vůbec k něčemu — nebo jen zpožděně opisovala včerejšek?*

Prameny: Met Office, „Robert FitzRoy and the early Met Office“ a „Our history“; bouře Royal Charter *v noci 25.–26. 10. 1859 (133 lodí, přes 800 mrtvých); první veřejná předpověď* The Times*, 1. 8. 1861; termín* forecast *a jeho vymezení: R. FitzRoy,* The Weather Book*, 1863.*

„Hádání není věda“ je parafráze dobového postoje vědeckého establishmentu, ne doslovný citát. FitzRoy † 30. 4. 1865. Galtonova komise 1866 předpovědi zrušila; obnoveny 1879.

Ta otázka — *je předpověď lepší než opsaný včerejšek?* — není o počasí. Je to nejdůležitější otázka, kterou položíš každému modelu, každému grafu a každému oslnivému AI demu, co ti kdy někdo ukáže. A skoro nikdo ji neklade.

Nejdřív najdi hloupého soupeře

Představ si, že ti někdo hrdě oznámí: „Můj model předpovídá zítřejší počasí správně v pětasedmdesáti procentech dnů!“ Zní to slušně. Jenže v Praze neprší zhruba tři dny ze čtyř — takže *kdokoli*, kdo bude každé ráno tupě tvrdit „zítra nebude pršet“, se strefí taky ve pětasedmdesáti procentech, a to bez jediného měření, bez modelu, bez elektřiny. Ten model tedy neporazil vůbec nic. Změřil jsi jeho nadšení, ne jeho inteligenci.

Tomuhle tupému, ale nesmírně užitečnému protivníkovi budeme v celé knize říkat **hloupý soupeř**. Je to laťka na zemi, přes kterou musí každý model nejdřív přeskočit, než ho pustíš mluvit o úspěchu. Hloupý soupeř nic neumí a schválně: jeho síla je právě v tom, že je triviální. Když ho tvůj chytrý model neporazí, není chytrý — je to jen dražší způsob, jak dělat totéž co on.

Hloupí soupeři bývají tři, podle toho, co předpovídáš:

hloupý soupeř (baseline) — triviální předpověď, kterou musí každý model nejdřív porazit: průměr, včerejšek, většina. Není to soupeř k výsměchu, ale k porážce; dokud ho nepřekonáš, nemáš co slavit.

→ v praxi: `DummyRegressor` *a* `DummyClassifier` *ve scikit-learn — hloupého soupeře postavíš jedním řádkem.*

TEENAGER · MECHANISMUS

Tři hloupí soupeři, které porazit dřív, než uvěříš jakémukoli modelu:

```
# 1) předpovídáš číslo (cenu, teplotu, tržbu)?
hloupý = pořád_stejně(průměr(trénovací_data))

# 2) předpovídáš vývoj v čase (počasí, kurz)?
hloupý = „zítra bude jako dnes“   # poslední
hodnota

# 3) předpovídáš třídu (spam? nemoc? podvod?)?
hloupý = pořád_stejně(nejčastější_třída)
```

Pravidlo je tvrdé a jednoduché: *hloupého soupeře postav a změř dřív, než postavíš cokoli chytrého.* Je to první krok, ne poslední — v scikit-learn na to jsou přímo `DummyRegressor` a `DummyClassifier`, aby ses nemohl vymluvit, že to dá práci.

FitzRoyův hloupý soupeř má jméno, které meteorologové používají dodnes: *persistence*, čili „zítra bude jako dnes“. A je to soupeř překvapivě zdatný — počasí má setrvačnost, takže „jako dnes“ se trefí častěji, než by sis myslel. Právě proto je tak dobrá laťka. FitzRoyova otázka, přeložená do řeči téhle knihy, zní: *porazil bys svou předpovědí prosté „zítra jako dnes“?* Kdo tohle neporáží, ten nepředpovídá — ten opisuje. A přesně tady se láme rozdíl mezi vědou a hádáním, o který FitzRoy přišel o pověst i o život.

→ kap. 3: šum na dashboardu, ve kterém šéf viděl signál, byl přesně tohle — neporažený hloupý soupeř. Náhodná čísla „předpovídala“ zrovna tak dobře jako on; nikdo je jen nepostavil vedle sebe a nezměřil.

Šéf a jeho náhodná čísla z třetí kapitoly? To byl celou dobu tenhle příběh, jen jsme ho neuměli pojmenovat. Do burzovních dashboardů jsme kdysi pustili generátor náhodných čísel a šéf v nich měsíce vysvětloval trendy. Jeho „předpovědi“ nikdy nikdo neporovnal s hloupým soupeřem — a kdyby to udělal, zjistil by, že náhoda „předpovídá“ zrovna tak dobře jako on. Šéfova čísla byla hloupý soupeř, kterého nikdo nezměřil; a bez toho měření vypadalo opisování včerejška jako vhled.

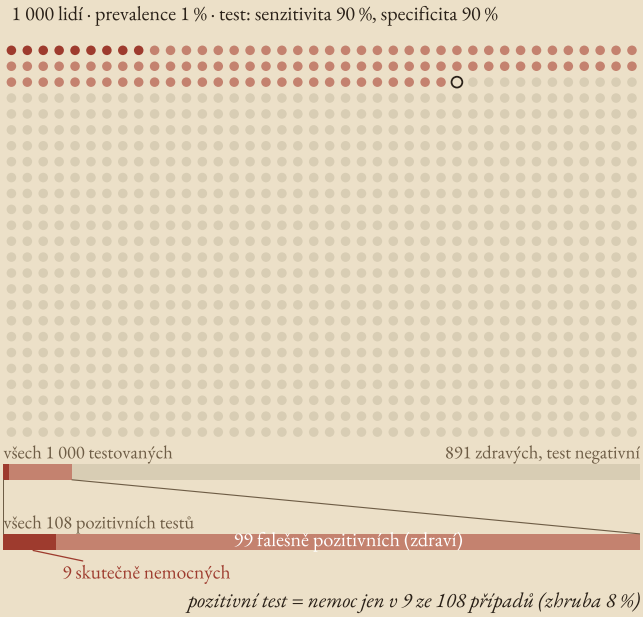
Krutější než přesnost: prevalence

Teď ale přijde past, která je horší než neporažený soupeř — protože se schová i za číslo, které vypadá skvěle. Vezmeme skutečný, každodenní příklad: test na vzácnou nemoc.

Řekněme, že máš test, který je z devadesáti procent přesný v obou směrech: z nemocných správně označí devadesát procent (těm říkáme jeho *úplnost*, lékařsky senzitivita), a ze zdravých správně propustí taky devadesát procent (jeho specificita). Zní to jako dobrý test. Nemoc, kterou hledá, je ale vzácná — má ji jeden člověk z sta. A teď ta otázka,

na kterou skoro každý odpoví špatně: *přišel ti pozitivní výsledek — jaká je pravděpodobnost, že jsi opravdu nemocný?*

Většina lidí — a nezřídka i lékařů — řekne kolem devadesáti procent. Pravda je pod deseti. Podívej se, proč, na tisícovce lidí, kde se nedá nic schovat:



Frekvenční model 1 000 lidí: prevalence 1 %, senzitivita i specificita 90 %. Deset nemocných, test najde devět (jeden mu unikne — prázdný kroužek). Devět set devadesát zdravých, ale 10 % z nich test označí omylem: devětadevadesát planých poplachů. Mezi 108 pozitivními testy je jen 9 skutečně nemocných.

Prameny čísel: model 1 000 pacientů dle materiálů kurzu (publikace Chyby, *bayes--paradox-medicinských-testů).*

Vezmi tisíc lidí. Nemocných je deset (to je ten jeden z sta). Test z nich najde devět — devadesátiprocentní úplnost. Zbývá devět set devadesát zdravých; test je z devadesáti procent správně propustí, ale těch zbylých deset procent, to je *devětadevadesát zdravých lidí, kterým test omylem zabliká „pozitivní“*. Sečti pozitivní výsledky: devět skutečně nemocných plus devětadevadesát planých poplachů, dohromady sto osm. A z těch sto osmi je nemocných jen devět. Devět ze sto osmi je zhruba osm procent. Tvůj pozitivní test tedy neznamená „skoro jistě nemocný“, ale „s pravděpodobností kolem osmi procent nemocný“ — a s pravděpodobností přes devadesát zdravý.

To číslo — kolik z pozitivních výsledků skutečně platí — má jméno: **přesnost** pozitivního testu, lékařsky pozitivní prediktivní hodnota (PPV). A tady musíme jednou provždy rozmotat past v jazyce: **přesnost** není totéž co **správnost**. Správnost je „kolik procent všech odpovědí je dobře“ — a ta je u našeho testu pořád vysoká, kolem devadesáti procent, protože zdravých je drtivá většina a ty test odbaví správně. Přesnost je „z toho, co jsi označil, kolik doopravdy platí“ — a ta spadla na osm procent. Jedno číslo oslňuje, druhé varuje; a rozhoduje to, které si zvolíš.

prevalence — jak časté to v populaci vůbec je (tady 1 %). Čím vzácnější jev, tím míň platí i dobrý test: planých poplachů z obrovské zdravé většiny je snadno víc než skutečných nálezů z hrstky nemocných.

Pozor na trojici, na které stojí celá kapitola: správnost (accuracy) = kolik ze všech odpovědí je dobře; přesnost (precision) = z označených kolik platí; úplnost (recall) = z platných kolik jsi našel. Tři různá čísla — a hloupý souper umí mít jedno z nich vysoké, aniž umí cokoli.

Vidíš, kdo tu past nastražil? Ne test — ten je slušný. Nastražila ji **prevalence**. Kdyby nemoc měl každý druhý, pozitivní test by znamenal skoro jistotu. Protože ji má jeden ze sta, drtí devětadevadesát planých poplachů z obří zdravé většiny těch devět skutečných nálezů z hrstky nemocných. To je věta, kterou si z téhle kapitoly odneseš pro celý zbytek knihy: **prevalence je krutější než přesnost**. Vzácnost jevu dokáže i vynikající test proměnit v generátor falešných poplachů — a žádné vylepšování senzitivity to nezachrání, dokud nezměníš, koho vůbec testuješ.

> *Tenhle jev je taky důvod, proč „test na vzácnou genetickou vadu byl pozitivní" obvykle znamená hlavně „jdi si to nechat potvrdit druhým, nezávislým testem" — a ne balit kufry. Bayes je milosrdnější než panika: druhý pozitivní test už startuje z prevalence osm procent, ne jedno, a to je úplně jiná hra.*

DEMO

`lesk-a-bida-vibecodingu.gaussalgo.com/d/paradox-medicinskych-testu`

Posuň prevalenci a přesnost testu a dívej se, jak pravděpodobnost nemoci po pozitivním testu padá — nebo roste, když totéž zopakuješ podruhé.

Matice záměn: kde se chyby počítají

Ať měříš cokoli — nemoc, spam, podvod —, dají se výsledky binárního testu srovnat do čtyř políček. Té tabulce se říká **matice záměn**, protože přesně počítá, které případy si tvůj model se kterými plete:

TEENAGER · MECHANISMUS

Matice záměn — čtyři osudy jedné předpovědi:

```
                    skutečnost: NEMOC
skutečnost: ZDRÁV
test říká POZITIV   správně pozitivní    falešně
pozitivní
                    (nález, TP)          (planý
poplach, FP)
test říká NEGATIV   falešně negativní    správně
negativní
                    (prošvihnuto, FN)    (v
pořádku, TN)
```

Z těch čtyř čísel spočítáš všechno ostatní:

```
úplnost  = TP / (TP + FN)   # z nemocných kolik
jsi našel
přesnost = TP / (TP + FP)   # z označených kolik
platí
správnost = (TP + TN) / vše # kolik z všeho je
dobře
```

Ta matice není účetnická nuda — je to místo, kde se rozhoduje, co je vlastně chyba. Protože *planý poplach* (FP) a *prošvihnutý případ* (FN) nejsou stejně drahé, a jejich cena se úlohu od úlohy obrací naruby.

Vezmi spamový filtr. Falešně negativní je otravný spam v doručené poště — mrzuté, ale přežiješ to. Falešně pozitivní je *důležitý e-mail, který ti filtr smetl do koše* — a možná jsi kvůli tomu přišel o práci nebo o zakázku. U spamu je dražší planý poplach; filtr proto radši pár spamů propustí, než aby smazal jediný dopis od tvého klienta.

falešně pozitivní (planý poplach) a falešně negativní (prošvihnutý případ) mají různou cenu, která se úlohu od úlohy obrací: u spamu je dražší FP, u nemoci FN. Metriku proto nevybírej podle vzorce — vybírej ji podle toho, která chyba tě víc bolí.

A teď to otoč. Test na rakovinu. Falešně pozitivní je nepříjemné leknutí a druhé, nepovinné vyšetření navíc. Falešně negativní je *přehlédnutý nádor*, který mezitím roste. Tady je nesrovnatelně dražší prošvihnutý případ; radši sto planých poplachů než jeden přehlédnutý nádor. Táž matice, opačná volba — a řídí ji cena chyby, ne matematika.

Proto neexistuje jediné „skóre modelu“, které by platilo všude. Kdo ti dá jedno číslo a řekne „model je z pětadevadesáti procent dobrý“, buď neví, na co se ptáš, nebo doufá, že se nezeptáš ty. Správná otázka nikdy nezní „jak je dobrý?“, ale „*které chyby dělá, a kolik mě každá z nich stojí?*“

Věž: přesnost, úplnost, ROC a kalibrace

Zatím jsme čísla počítali na prstech. Teď je postavíme na pevný základ — a přidáme dva nástroje, kterými se výkon testu měří napříč všemi možnými prahy najednou, a jeden, který se ptá na něco úplně jiného než správnost: jestli se dá modelově *jistotě* věřit.

EINSTEIN · MATEMATIKA — přeskočitelné

Přesnost, úplnost a jejich spor. Z matice záměn:

$$\text{přesnost} = \frac{TP}{TP + FP}, \quad \text{úplnost} = \frac{TP}{TP + FN}.$$

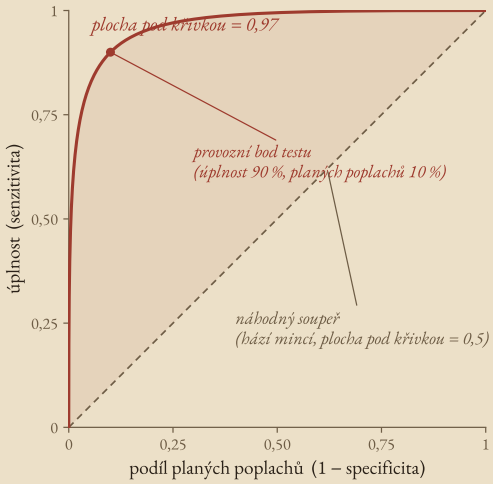
Přesnost trestá plané poplachy, úplnost trestá prošvihnuté případy — a táhnou proti sobě. Posuneš-li práh testu níž (označíš víc lidí za nemocné), chytíš víc skutečných nálezů (úplnost roste), ale nabereš i víc planých poplachů (přesnost klesá). Jedno číslo, které je smiřuje, je jejich harmonický průměr, **F1**:

$$F_1 = 2 \cdot \frac{\text{přesnost} \cdot \text{úplnost}}{\text{přesnost} + \text{úplnost}}.$$

Harmonický průměr schválně: je nízký, kdykoli je nízká *kterákoli* ze složek — nedovolí ti vykoupit mizernou přesnost skvělou úplností. Kdo chce vážit obě chyby jinak, sáhne po F_β a parametrem β řekne, kolikrát víc mu záleží na úplnosti než na přesnosti.

EINSTEIN · MATEMATIKA — přeskočitelné

ROC křivka a plocha pod ní. Práh testu není dán osudem — je to páčka, kterou ty otáčíš. Pro každou polohu páčky dostaneš jeden pár čísel: úplnost (podíl správně chycených nemocných) a podíl planých poplachů (tj. 1 − specificita). Vyneseš-li je proti sobě pro všechny prahy, vznikne **ROC křivka**. Náhodný soupeř — ten, co hází mincí — leží na úhlopříčce: každé procento chycených nemocných zaplatí stejným procentem planých poplachů. Dobrý test se od úhlopříčky odklání k levému hornímu rohu.



Jedno číslo, které křivku shrne, je **plocha pod ní** (AUC): rovná se pravděpodobnosti, že náhodně vybraný nemocný dostane vyšší skóre než náhodně vybraný zdravý. Náhodný soupeř má AUC = 0, 5 (úhlopříčka), dokonalý test 1, 0. Všimni si ale léčky: AUC náš test hodnotí přes *všechny* prahy najednou a nezávisle na prevalenci — proto může být vysoká i tam, kde je pozitivní test skoro k ničemu. AUC měří, jak dobře test *řadí*; PPV z minulé sekce měří, co jeho pozitivní výrok *znamená* u konkrétní nemoci. Dvě různé otázky.

Provozní bod na křivce (senzitivita 90 %, planých poplachů 10 %) je jeden *práh — ten z hrdinského obrazu. Plocha pod celou křivkou je 0,96: poctivě jiný údaj než senzitivita 90 %, protože sčítá výkon přes všechny prahy. Vysoká AUC a bezcenný pozitivní test se u vzácné nemoci nevylučují.*

EINSTEIN · MATEMATIKA — přeskočitelné

PPV z prevalence — Bayes doslova. Pojmenujme, co jsme spočítali na tisícovce lidí. Označ D jev „nemocný“ a $+$ jev „pozitivní test“. Senzitivita je $P(+ \mid D)$, specificita $P(- \mid \neg D)$, prevalence $P(D)$. Ptáme se na $P(D \mid +)$ — pozitivní prediktivní hodnotu. Bayesova věta:

$$P(D \mid +) = \frac{P(+ \mid D) \cdot P(D)}{P(+ \mid D) \cdot P(D) + P(+ \mid \neg D) \cdot P(\neg D)}.$$

Dosaď hrdinská čísla — $P(+|D) = 0,9$, $P(D) = 0,01$, $P(+|\neg D) = 0,1$, $P(\neg D) = 0,99$:

$$P(D \mid +) = \frac{0,9 \cdot 0,01}{0,9 \cdot 0,01 + 0,1 \cdot 0,99}.$$

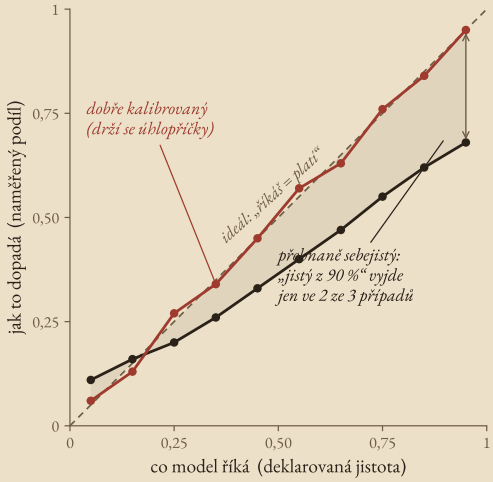
Nemocní v čitateli, plané poplachy v druhém členu jmenovatele:

$$P(D \mid +) = \frac{0,009}{0,009 + 0,099} = \frac{0,009}{0,108} \approx 0,083.$$

Osm celých tři. Celé kouzlo dělá jmenovatel: člen $0,1 \cdot 0,99$ — plané poplachy z obří zdravé většiny — utopí čitatel. Senzitivita se dá vyhnat na 99 %, a dokud je prevalence 1 % a specificita 90 %, PPV to skoro nehne, protože o výsledku rozhoduje ten druhý sčítanec ve jmenovateli. **Prevalence je krutější než přesnost** — teď víš, kterým členem to je.

EINSTEIN · MATEMATIKA — přeskočitelné

Kalibrace: dá se té jistotě věřit? Správnost, přesnost i AUC hodnotí *rozhodnutí* (nemocný / zdravý). Existuje ale druhá, jemnější ctnost: když model (nebo člověk) řekne „jsem si jistý na osmdesát procent“, má se to v osmdesáti procentech takových případů opravdu vyplnit. Tomu se říká **kalibrace** a měří se **diagramem spolehlivosti** (reliability diagram): na vodorovnou osu dáš deklarovanou jistotu, na svislou naměřený podíl úspěchů, a porovnáš s úhlopříčkou „říkáš = platí“.



Model pod úhlopříčkou je **přehnaně sebejistý**: tvrdí devadesát, platí sedmdesát. To není totéž co „dělá chyby“ — může mít i výbornou přesnost a přitom lhát o vlastní jistotě. A tady se poprvé v knize potkáváme s neduhem, který se vrátí ve velkém: jazykové modely bývají výmluvně, plynule a *nekalibrovaně* sebejisté. Diagram spolehlivosti je nástroj, kterým to poznáš — a poznamenej si ho, protože v kapitole 16 jím přistihneme model při podlézání a v kapitole 22 jím změříš sám sebe.

kalibrace — „když říkáš 80 %, musí to v 80 % případů vyjít“; diagram spolehlivosti (reliability diagram) srovnává, co model tvrdí, s tím, jak to dopadá.

→ kap. 16: RLHF kalibraci modelu spíš zhoršuje — sebejistý tón se odměňuje bez ohledu na to, jestli je oprávněný. → kap. 22: Brierovo skóre lidských expertů a vlastní kalibrační kvíz — odejdeš s číslem o sobě.

Než uvěříš čemukoli

FitzRoyova otázka a paradox medicínského testu jsou dvě tváře jedné lekce. FitzRoy musel svou předpověď postavit vedle prostého „zítra jako dnes“, jinak nevěděl, jestli něco umí, nebo jen opisuje včerejšek. Medicínský test zas ukázal, že devadesátiprocentní přesnost neznamená nic, dokud nevíš, jak vzácné je to, co hledáš. Obojí je táž věta: *číslo o modelu samo o sobě neznamená nic — význam dostane teprve srovnáním se správným soupeřem a zasazením do správného kontextu.*

Vezmi si z toho jediný, ale neúprosný reflex. Až ti příště někdo ukáže model s ohromující přesností, nebo AI demo, které jako by četlo

myšlenky, nezeptej se „jak je to dobré?“. Zeptej se dvě věci, a v tomhle pořadí: *Co je hloupý soupeř téhle úlohy — a porazil ho ten model vůbec?* A druhá: *jak vzácné je to, co hledá, a přežije ta přesnost prevalenci?* Většina oslnivých dem/čísel na jedné z těch dvou otázek tiše umře. Ta, která přežijí obě, stojí za tvou pozornost — a teprve za ní.

„Jak bych poznal, že se pletu?“

Konkrétní test k této kapitole: než uvěříš jakémukoli modelu nebo oslnivému AI demu, polož mu dvě otázky. První: *co je hloupý soupeř této úlohy — a porazil ho ten model vůbec?* Postav toho soupeře doslova (`DummyClassifier`, „zítra jako dnes“, průměr) a změř oba na týchž datech; jestli chytrý model neporazí toho hloupého, kapitola tvrdí, že jsi neměřil inteligenci, ale nadšení — a v tom případě chci vidět ta dvě čísla vedle sebe. Druhá: *jak vzácné je to, co hledáš?* Vezmi jeho úspěšnost, dosaď ji do Bayese s reálnou prevalencí a spočítej, co znamená jeden pozitivní výrok. Tvrdím, že u vzácného jevu ti PPV spadne hluboko pod deklarovanou přesnost; vyjde-li ti, že pozitivní výrok znamená skoro jistotu, buď je jev častý, nebo je test lepší, než se zdálo — a pak chci vidět tvůj výpočet.