

FOUILLE DE DONNÉES ET AIDE À LA DECISION

Introduction à la data science
et au machine learning.

M2 Informatique, Université Paris Cité

Fabien Viger

fabien.viger@gmail.com

[TODOs](#)

DATA SCIENCE & MACHINE LEARNING

UNE SCIENCE À LA MODE

Mais pourquoi ?

- ChatGPT, DeepFakes, AlphaZero: gros progrès récents des réseaux de neurones.
- Stockage et traitement des données : de moins en moins cher.
- "Big Data": Volume + complexité des données, impossible de les comprendre "à la main" : SNCF, génétique, finance, ...
- **Data Scientist**, compétence **très** à la mode.
- Applications: pub, marketing, compréhension des clients → **optimisation**.

RENDRE LES ORDINATEURS INTELLIGENTS

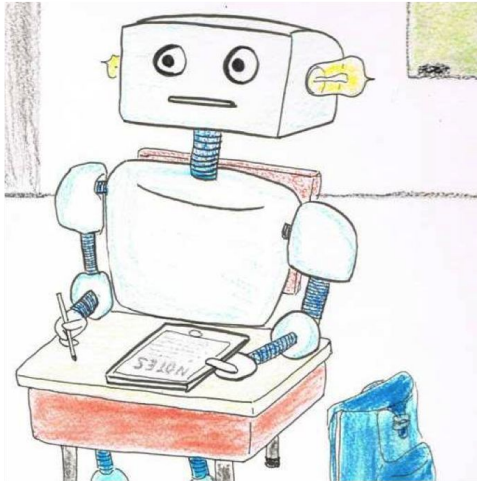


Figure: Tiré du blog du laboratoire "Computer and Cognition", NYU.

LA RENCONTRE DE PLUSIEURS DISCIPLINES

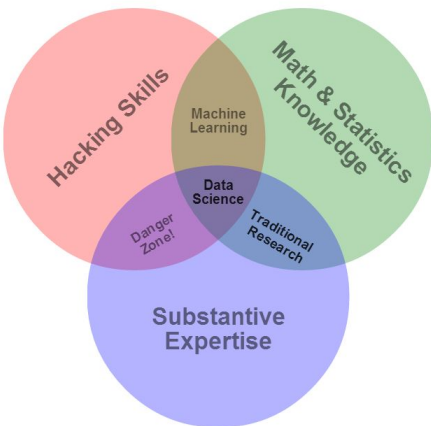


Figure: Tiré de
econometricsense.blogspot.fr

Changement de cap de plus en plus observé : des statistiques traditionnelles aux modèles **algorithmiques**.

Besoin de modélisation mais aussi de méthodes rapides et généralisables à la grande dimension.

QUELQUES APPLICATIONS

“Écris une introduction au cours de Data Science en français, 30 mots.”



“Bienvenue au cours de Data Science ! Dans ce programme, nous explorerons les méthodes et les outils pour extraire des insights précieux à partir des données.”

QUELQUES APPLICATIONS

“A fox flying in the sky, wearing a special backpack with four propellers rotating above like a helicopter”



QUELQUES APPLICATIONS



RUM RAISIN - BY AMRINDER



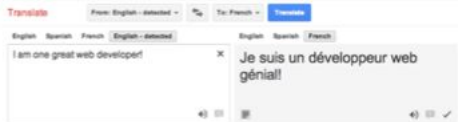
©OriginalArtist
Reproduction rights obtainable from
www.CartoonStock.com



"Let me guess. You had to review your diagnosis."

0/1	1000_2 4/1	1001_2 2/1	1002_2 2/1
0/2	1003_2 3/1	1004_2 2/1	1005_2 3/1
0/3	1006_2 2/1	1007_2 8/1	1008_2 4/1
0/4	1009_2 7/1	1010_2 8/1	1011_2 1/1
0/5	1012_2 4/1	1013_2 4/1	1014_2 4/1
0/6	1015_2 3/1	1016_2 7/1	1017_2 5/1
0/7	1018_2 3/1	1019_2 7/1	1020_2 5/1
0/8	1021_2 4/1	1022_2 3/1	1023_2 0/1
0/9	1024_2 4/1	1025_2 5/1	1026_2 4/1
0/10	1027_2 4/1	1028_2 5/1	1029_2 4/1
0/11	1030_2 0/1	1031_2 0/1	1032_2 0/1
0/12	1033_2 0/1	1034_2 0/1	1035_2 0/1
0/13	1036_2 0/1	1037_2 0/1	1038_2 0/1
0/14	1039_2 0/1	1040_2 0/1	1041_2 0/1
0/15	1042_2 0/1	1043_2 0/1	1044_2 0/1
0/16	1045_2 0/1	1046_2 0/1	1047_2 0/1
0/17	1048_2 0/1	1049_2 0/1	1050_2 0/1
0/18	1051_2 0/1	1052_2 0/1	1053_2 0/1
0/19	1054_2 0/1	1055_2 0/1	1056_2 0/1
0/20	1057_2 0/1	1058_2 0/1	1059_2 0/1
0/21	1060_2 0/1	1061_2 0/1	1062_2 0/1
0/22	1063_2 0/1	1064_2 0/1	1065_2 0/1
0/23	1066_2 0/1	1067_2 0/1	1068_2 0/1
0/24	1069_2 0/1	1070_2 0/1	1071_2 0/1
0/25	1072_2 0/1	1073_2 0/1	1074_2 0/1
0/26	1075_2 0/1	1076_2 0/1	1077_2 0/1
0/27	1078_2 0/1	1079_2 0/1	1080_2 0/1
0/28	1081_2 0/1	1082_2 0/1	1083_2 0/1
0/29	1084_2 0/1	1085_2 0/1	1086_2 0/1
0/30	1087_2 0/1	1088_2 0/1	1089_2 0/1
0/31	1090_2 0/1	1091_2 0/1	1092_2 0/1
0/32	1093_2 0/1	1094_2 0/1	1095_2 0/1
0/33	1096_2 0/1	1097_2 0/1	1098_2 0/1
0/34	1099_2 0/1	1100_2 0/1	1101_2 0/1
0/35	1102_2 0/1	1103_2 0/1	1104_2 0/1
0/36	1105_2 0/1	1106_2 0/1	1107_2 0/1
0/37	1108_2 0/1	1109_2 0/1	1110_2 0/1
0/38	1111_2 0/1	1112_2 0/1	1113_2 0/1
0/39	1114_2 0/1	1115_2 0/1	1116_2 0/1
0/40	1117_2 0/1	1118_2 0/1	1119_2 0/1
0/41	1120_2 0/1	1121_2 0/1	1122_2 0/1
0/42	1123_2 0/1	1124_2 0/1	1125_2 0/1
0/43	1126_2 0/1	1127_2 0/1	1128_2 0/1
0/44	1129_2 0/1	1130_2 0/1	1131_2 0/1
0/45	1132_2 0/1	1133_2 0/1	1134_2 0/1
0/46	1135_2 0/1	1136_2 0/1	1137_2 0/1
0/47	1138_2 0/1	1139_2 0/1	1140_2 0/1
0/48	1141_2 0/1	1142_2 0/1	1143_2 0/1
0/49	1144_2 0/1	1145_2 0/1	1146_2 0/1
0/50	1147_2 0/1	1148_2 0/1	1149_2 0/1
0/51	1150_2 0/1	1151_2 0/1	1152_2 0/1
0/52	1153_2 0/1	1154_2 0/1	1155_2 0/1
0/53	1156_2 0/1	1157_2 0/1	1158_2 0/1
0/54	1159_2 0/1	1160_2 0/1	1161_2 0/1
0/55	1162_2 0/1	1163_2 0/1	1164_2 0/1
0/56	1165_2 0/1	1166_2 0/1	1167_2 0/1
0/57	1168_2 0/1	1169_2 0/1	1170_2 0/1
0/58	1171_2 0/1	1172_2 0/1	1173_2 0/1
0/59	1174_2 0/1	1175_2 0/1	1176_2 0/1
0/60	1177_2 0/1	1178_2 0/1	1179_2 0/1
0/61	1180_2 0/1	1181_2 0/1	1182_2 0/1
0/62	1183_2 0/1	1184_2 0/1	1185_2 0/1
0/63	1186_2 0/1	1187_2 0/1	1188_2 0/1
0/64	1189_2 0/1	1190_2 0/1	1191_2 0/1
0/65	1192_2 0/1	1193_2 0/1	1194_2 0/1
0/66	1195_2 0/1	1196_2 0/1	1197_2 0/1
0/67	1198_2 0/1	1199_2 0/1	1200_2 0/1
0/68	1201_2 0/1	1202_2 0/1	1203_2 0/1
0/69	1204_2 0/1	1205_2 0/1	1206_2 0/1
0/70	1207_2 0/1	1208_2 0/1	1209_2 0/1
0/71	1210_2 0/1	1211_2 0/1	1212_2 0/1
0/72	1213_2 0/1	1214_2 0/1	1215_2 0/1
0/73	1216_2 0/1	1217_2 0/1	1218_2 0/1
0/74	1219_2 0/1	1220_2 0/1	1221_2 0/1
0/75	1222_2 0/1	1223_2 0/1	1224_2 0/1
0/76	1225_2 0/1	1226_2 0/1	1227_2 0/1
0/77	1228_2 0/1	1229_2 0/1	1230_2 0/1
0/78	1231_2 0/1	1232_2 0/1	1233_2 0/1
0/79	1234_2 0/1	1235_2 0/1	1236_2 0/1
0/80	1237_2 0/1	1238_2 0/1	1239_2 0/1
0/81	1240_2 0/1	1241_2 0/1	1242_2 0/1
0/82	1243_2 0/1	1244_2 0/1	1245_2 0/1
0/83	1246_2 0/1	1247_2 0/1	1248_2 0/1
0/84	1249_2 0/1	1250_2 0/1	1251_2 0/1
0/85	1252_2 0/1	1253_2 0/1	1254_2 0/1
0/86	1255_2 0/1	1256_2 0/1	1257_2 0/1
0/87	1258_2 0/1	1259_2 0/1	1260_2 0/1
0/88	1261_2 0/1	1262_2 0/1	1263_2 0/1
0/89	1264_2 0/1	1265_2 0/1	1266_2 0/1
0/90	1267_2 0/1	1268_2 0/1	1269_2 0/1
0/91	1270_2 0/1	1271_2 0/1	1272_2 0/1
0/92	1273_2 0/1	1274_2 0/1	1275_2 0/1
0/93	1276_2 0/1	1277_2 0/1	1278_2 0/1
0/94	1279_2 0/1	1280_2 0/1	1281_2 0/1
0/95	1282_2 0/1	1283_2 0/1	1284_2 0/1
0/96	1285_2 0/1	1286_2 0/1	1287_2 0/1
0/97	1288_2 0/1	1289_2 0/1	1290_2 0/1
0/98	1291_2 0/1	1292_2 0/1	1293_2 0/1
0/99	1294_2 0/1	1295_2 0/1	1296_2 0/1
0/100	1297_2 0/1	1298_2 0/1	1299_2 0/1
0/101	1300_2 0/1	1301_2 0/1	1302_2 0/1
0/102	1303_2 0/1	1304_2 0/1	1305_2 0/1
0/103	1306_2 0/1	1307_2 0/1	1308_2 0/1
0/104	1309_2 0/1	1310_2 0/1	1311_2 0/1
0/105	1312_2 0/1	1313_2 0/1	1314_2 0/1
0/106	1315_2 0/1	1316_2 0/1	1317_2 0/1
0/107	1318_2 0/1	1319_2 0/1	1320_2 0/1
0/108	1321_2 0/1	1322_2 0/1	1323_2 0/1
0/109	1324_2 0/1	1325_2 0/1	1326_2 0/1
0/110	1327_2 0/1	1328_2 0/1	1329_2 0/1
0/111	1330_2 0/1	1331_2 0/1	1332_2 0/1
0/112	1333_2 0/1	1334_2 0/1	1335_2 0/1
0/113	1336_2 0/1	1337_2 0/1	1338_2 0/1
0/114	1339_2 0/1	1340_2 0/1	1341_2 0/1
0/115	1342_2 0/1	1343_2 0/1	1344_2 0/1
0/116	1345_2 0/1	1346_2 0/1	1347_2 0/1
0/117	1348_2 0/1	1349_2 0/1	1350_2 0/1
0/118	1351_2 0/1	1352_2 0/1	1353_2 0/1
0/119	1354_2 0/1	1355_2 0/1	1356_2 0/1
0/120	1357_2 0/1	1358_2 0/1	1359_2 0/1
0/121	1360_2 0/1	1361_2 0/1	1362_2 0/1
0/122	1363_2 0/1	1364_2 0/1	1365_2 0/1
0/123	1366_2 0/1	1367_2 0/1	1368_2 0/1
0/124	1369_2 0/1	1370_2 0/1	1371_2 0/1
0/125	1372_2 0/1	1373_2 0/1	1374_2 0/1
0/126	1375_2 0/1	1376_2 0/1	1377_2 0/1
0/127	1378_2 0/1	1379_2 0/1	1380_2 0/1
0/128	1381_2 0/1	1382_2 0/1	1383_2 0/1
0/129	1384_2 0/1	1385_2 0/1	1386_2 0/1
0/130	1387_2 0/1	1388_2 0/1	1389_2 0/1
0/131	1390_2 0/1	1391_2 0/1	1392_2 0/1
0/132	1393_2 0/1	1394_2 0/1	1395_2 0/1
0/133	1396_2 0/1	1397_2 0/1	1398_2 0/1
0/134	1399_2 0/1	1400_2 0/1	1401_2 0/1
0/135	1402_2 0/1	1403_2 0/1	1404_2 0/1
0/136	1405_2 0/1	1406_2 0/1	1407_2 0/1
0/137	1408_2 0/1	1409_2 0/1	1410_2 0/1
0/138	1411_2 0/1	1412_2 0/1	1413_2 0/1
0/139	1414_2 0/1	1415_2 0/1	1416_2 0/1
0/140	1417_2 0/1	1418_2 0/1	1419_2 0/1
0/141	1420_2 0/1	1421_2 0/1	1422_2 0/1
0/142	1423_2 0/1	1424_2 0/1	1425_2 0/1
0/143	1426_2 0/1	1427_2 0/1	1428_2 0/1
0/144	1429_2 0/1	1430_2 0/1	1431_2 0/1
0/145	1432_2 0/1	1433_2 0/1	1434_2 0/1
0/146	1435_2 0/1	1436_2 0/1	1437_2 0/1
0/147	1438_2 0/1	1439_2 0/1	1440_2 0/1
0/148	1441_2 0/1	1442_2 0/1	1443_2 0/1
0/149	1444_2 0/1	1445_2 0/1	1446_2 0/1
0/150	1447_2 0/1	1448_2 0/1	1449_2 0/1
0/151	1450_2 0/1	1451_2 0/1	1452_2 0/1
0/152	1453_2 0/1	1454_2 0/1	1455_2 0/1
0/153	1456_2 0/1	1457_2 0/1	1458_2 0/1
0/154	1459_2 0/1	1460_2 0/1	1461_2 0/1
0/155	1462_2 0/1	1463_2 0/1	1464_2 0/1
0/156	1465_2 0/1	1466_2 0/1	1467_2 0/1
0/157	1468_2 0/1	1469_2 0/1	1470_2 0/1
0/158	1471_2 0/1	1472_2 0/1	1473_2 0/1
0/159	1474_2 0/1	1475_2 0/1	1476_2 0/1
0/160	1477_2 0/1	1478_2 0/1	1479_2 0/1
0/161	1480_2 0/1	1481_2 0/1	1482_2 0/1
0/162	1483_2 0/1	1484_2 0/1	1485_2 0/1
0/163	1486_2 0/1	1487_2 0/1	1488_2 0/1
0/164	1489_2 0/1	1490_2 0/1	1491_2 0/1
0/165	1492_2 0/1	1493_2 0/1	1494_2 0/1
0/166	1495_2 0/1	1496_2 0/1	1497_2 0/1
0/167	1498_2 0/1	1499_2 0/1	1500_2 0/1
0/168	1501_2 0/1	1502_2 0/1	1503_2 0/1
0/169	1504_2 0/1	1505_2 0/1	1506_2 0/1
0/170	1507_2 0/1	1508_2 0/1	1509_2 0/1
0/171	1510_2 0/1	1511_2 0/1	1512_2 0/1
0/172	1513_2 0/1	1514_2 0/1	1515_2 0/1
0/173	1516_2 0/1	1517_2 0/1	1518_2 0/1
0/174	1519_2 0/1	1520_2 0/1	1521_2 0/1
0/175	1522_2 0/1	1523_2 0/1	1524_2 0/1
0/176	1525_2 0/1	1526_2 0/1	1527_2 0/1
0/177	1528_2 0/1	1529_2 0/1	1530_2 0/1
0/178	1531_2 0/1	1532_2 0/1	1533_2 0/1
0/179	1534_2 0/1	1535_2 0/1	1536_2 0/1
0/180	1537_2 0/1	1538_2 0/1	1539_2 0/1
0/181	1540_2 0/1	1541_2 0/1	1542_2 0/1
0/182	1543_2 0/1	1544_2 0/1	1545_2 0/1
0/183	1546_2 0/1	1547_2 0/1	1548_2 0/1
0/184	1549_2 0/1	1550_2 0/1	1551_2 0/1
0/185	1552_2 0/1	1553_2 0/1	1554_2 0/1
0/186	1555_2 0/1	1556_2 0/1	1557_2 0/1
0/187	1558_2 0/1	1559_2 0/1	1560_2 0/1
0/188	1561_2 0/1	1562_2 0/1	1563_2 0/1
0/189	1564_2 0/1	1565_2 0/1	1566_2 0/1
0/190	1567_2 0/1	1568_2 0/1	1569_2 0/1
0/191	1570_2 0/1	1571_2 0/1	1572_2 0/1
0/192	1573_2 0/1	1574_2 0/1	1575_2 0/1
0/193	1576_2 0/1	1577_2 0/1	1578_2 0/1
0/194	1579_2 0/1	1580_2 0/1	1581_2 0/1
0/195	1582_2 0/1	1583_2 0/1	1584_2 0/1
0/196	1585_2 0/1	1586_2 0/1	1587_2 0/1
0/197	1588_2 0/1	1589_2 0/1	1590_2 0/1
0/198	1591_2 0/1	1592_2 0/1	1593_2 0/1
0/199	1594_2 0/1	1595_2 0/1	1596_2 0/1
0/200	1597_2 0/1	1598_2 0/1	1599_2 0/1
0/201	1600_2 0/1	1601_2 0/1	1602_2 0/1
0/202	1603_2 0/1	1604_2 0/1	1605_2 0/1
0/203	1606_2 0/1	1607_2 0/1	1608_2 0/1
0/204	1609_2 0/1	1610_2 0/1	1611_2 0/1
0/205	1612_2 0/1	1613_2 0/1	1614_2 0/1
0/206	1615_2 0/1	1616_2 0/1	1617_2 0/1
0/207	1618_2 0/1	1619_2 0/1	1620_2 0/1
0/208	1621_2 0/1	1622_2 0/1	1623_2 0/1
0/209	1624_2 0/1	1625_2 0/1	1626_2 0/1
0/210	1627_2 0/1	1628_2 0/1	1629_2 0/1
0/211	1630_2 0/1	1631_2 0/1	1632_2 0/1
0/212	1633_2 0/1	1634_2 0/1	

APPLICATIONS WEB



Now! Click the words above to edit and view alternate translations. [Dismiss](#)

Customers Who Bought This Item Also Bought



Above the Fold:
Understanding the ...
Brian Miller
★★★★☆ (15)
Paperback
\$17.49



Learning PHP, MySQL,
JavaScript, and CSS: A ...
Robin Nixon
★★★★☆ (21)
Paperback
\$23.99



Learning Web Design: A
Beginner's Guide to ...
Jennifer Niederst Robbi...
★★★★☆ (19)
Paperback
\$28.53



PROGRAMME DU COURS

- **Décision statistique:** statistiques 101, tests, intervalles de confiance.
- **Transformation des données**, encodage (aka *Feature Engineering*)
- Apprentissage supervisé: Naive Bayes, Régression Linéaire, Régression Logistique, Arbres de Décision, Random Forests, SVM, Réseaux de Neurones.
- **Apprentissage** non supervisé: clustering, réduction de dimension.
- **Moteurs de recommandation.**

VALIDATION DU COURS

VALIDATION DU COURS

Note finale: **50% projet**, 30% TD, 20% QCM

Pas de rattrapage possible.

La présence au cours/TD n'est pas *obligatoire*, mais très fortement recommandée

Ne trichez pas! Ce n'est pas l'esprit du cours. Si c'est trop difficile, dites-le moi, on s'adapte. Je suis **impitoyable** avec la triche et le plagiat: tout copier/coller ou plagiat dans les TD, ou projet, ou QCM = 0 partout, sans rattrapage possible, + signalement à l'UFR.

SITE WEB

<http://fabien.viger.free.fr/ml>

PROJET: EXEMPLES

<https://www.kaggle.com/datasets>

Inspirez-vous, choisissez et validez avec moi.

PROJET: ORGANISATION

1 ou 2 personnes par projet. Pas de **touriste** !...

Demi-page de “résumé des contribution personnelles” par écrit.

Soutenance: semaine avant exams, slides (.pdf): 15min, +15min de **questions** (10 + 10 si seul).

Les projets sont valorisables sur un CV !

Projet de A à Z, ou une bonne partie de:
Collecte des données + analyse + encodage
[+ visualisation?] + méthodo + résultats.

PROJET: CODE

Du code (un minimum documenté, au moins commenté) doit accompagner le projet.

Il doit être envoyé avec les slides.

Langage: Python, Java, C++ OK; pour tout autre me demander.

En fonction de votre projet: appli web, page html, executable, script... (Pas forcément d'interface.)

PROJET: À RENDRE

Liste contributions perso: 1/2-page par étudiant

- Ne doit laisser aucun doute sur qui a fait quoi
- Soyez honnête! Il est normal de réutiliser des choses faites par d'autres. **Dites-le.**

Support de soutenance: ~10-15 pages (1 page par minute). Exemple de Plan:

- Intro
- Données [Analyse? Visualisation?]
- Méthodologie
- Résultats
- Discussion, Conclusion.
- Code; Contributions

PROJET: PLANNING

Semaines 2/3: Choix du projet et formation des équipes.

Semaines 4 à 9: Analyse et rédaction. RDV de suivi par groupe, + suivi par mail.

Semaine 11 (?) (examens): **soutenance**

Obtention des notes de projet rapidement (1 jour ou 2).

EXAMEN(S) INDIVIDUEL: QCM

Examens sur machine ou sur table (sans documents), au cours du trimestre.

Nombre inconnu (au moins 1, peut-être 2)

Pendant les séances de TD (je préviens 1 semaine en avance)

TD/TP

Les TD sont sur machine, en **python**.

Ils doivent être rendus (Moodle), et seront **évalués** (automatiquement)

Rendu de TD: individuel, le jour même, à minuit.

Je fournirai des auto-tests dans la mesure du possible

Un corrigé de TD apparaîtra en ligne rapidement après le rendu.

L'ESPRIT DU COURS

Interactif

Travail d'équipe

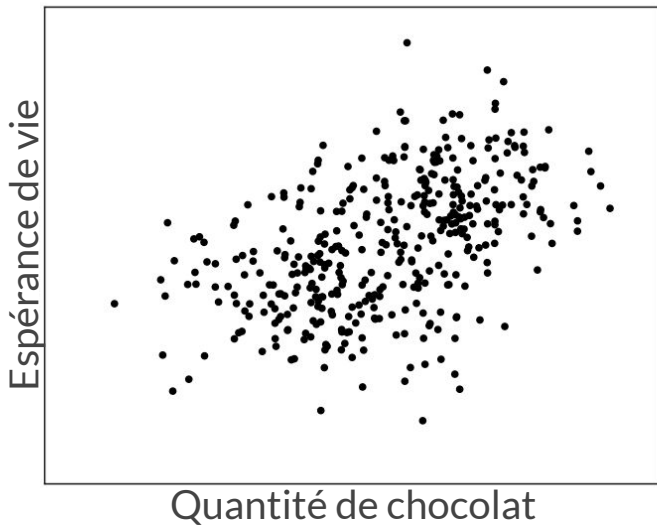
Appliqué

Toute proposition de thèmes à aborder est toujours la bienvenue.

QUESTIONS?

INTRODUCTION

CHOCOLAT ET ESPÉRANCE DE VIE

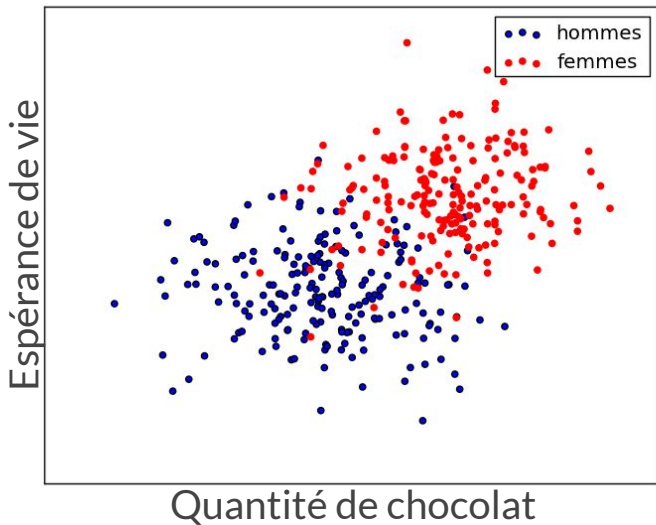


Exemple fictif emprunté à

Isabelle Guyon

Manger du chocolat **augmente**
l'espérance de vie !

CHOCOLAT ET ESPÉRANCE DE VIE

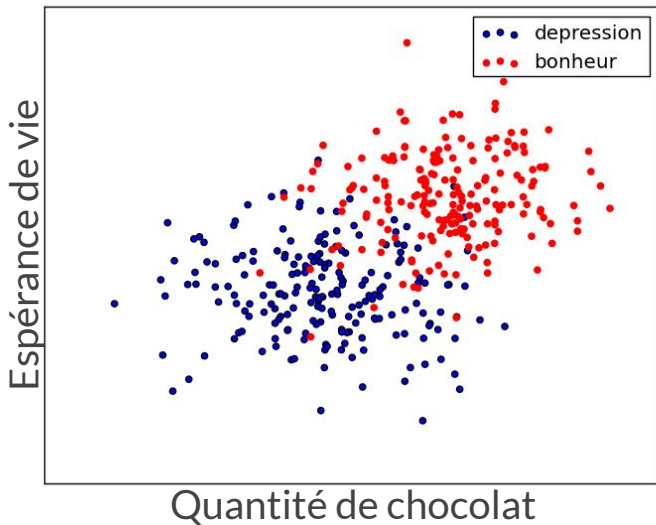


Exemple fictif emprunté à

Isabelle Guyon

Manger du chocolat **n'augmente pas**
l'espérance de vie !

CHOCOLAT ET ESPÉRANCE DE VIE

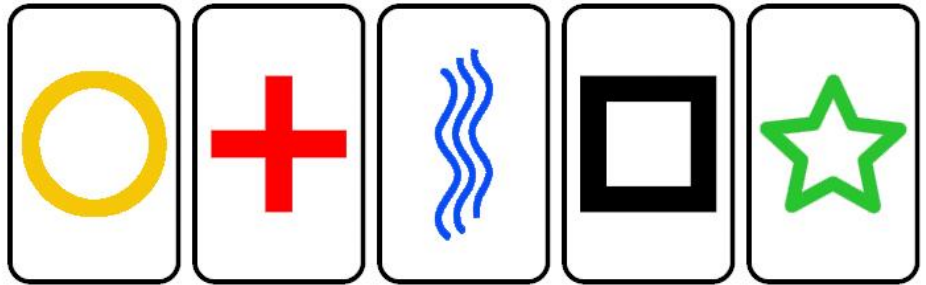


Exemple fictif emprunté à

Isabelle Guyon

Manger du chocolat **augmente peut-être**
l'espérance de vie !

LES EXPÉRIENCES DE RHINE



Source: Wikipedia

VOUS AVEZ UNE MAUVAISE INTUITION STATISTIQUE (SI SI!)



PEUT-ON FAIRE DIRE AUX CHIFFRES CE QUE L'ON VEUT ?

[Données fictives !]	Nombre de chômeurs	Nb. travailleurs potentiels
Année 1	1.000.000	10.000.000
Année 2	1.010.000	11.000.000

- VRAI** ● Le chômage a **augmenté** de 1%.
- OU** ● Le taux de chômage a **baissé** de 0.9 pts.
- FAUX** ● Il y a 10.000 chômeurs de **plus**.
- ?** ● Le taux de chômage a **baissé** de 10%.

PEOPLE VS COLLINS



1964. Un vol. Les témoins affirment avoir vu un homme noir barbu et moustachu et une femme blonde avec une queue de cheval s'enfuir dans une voiture jaune. Malcolm et Janet Collins correspondent à la description...

PEOPLE VS COLLINS

Raisonnement du procureur :

- Homme noir portant une barbe : 10%
- Homme noir portant une moustache : 25%
- Femme blanche avec queue de cheval : 10%
- Femme blanche aux cheveux blonds : 33%
- Voiture en partie jaune : 10%
- Couple "inter-racial" dans une voiture : 0.1%

Conclusion: probabilité que les Collins soient innocents < 1/12 millions. Ils sont **condamnés**.

La cour d'appel **annule** cela. **Pourquoi ?**

EXPLICATION

Admettons que les probabilités, bien qu'estimées sans doute arbitrairement, soient justes.

L'erreur principale est d'avoir ignoré les **dépendances** entre les événements.

Au lieu de multiplier les probabilités entre elles, comme les événements ne sont **pas** indépendants, il faut considérer les probabilités **conditionnelles**.

Exemple: la probabilité (à l'époque) d'avoir une moustache sachant que l'on a une barbe est très élevée, disons 80%. Donc la probabilité d'avoir une barbe ET une moustache devient $10\% \times 80\%$ au lieu de $10\% \times 25\%$. Idem pour les autres événements.

PARADOXE DE SIMPSON

100 étudiants (50 IMPAIRS et 50 DATA)
choisissent chacun entre 2 cours : **Fouille** et
Systèmes. Pourcentages de validation [fictif!] :

Fouilles de données		Systèmes avancés	
IMPAIRS	DATA	IMPAIRS	DATA
90%	84.5%	70%	60%

Les IMPAIRS réussissent mieux **chacun** des
cours.... Donc ils ont **forcément** un meilleur
taux de réussite globale (sur les 2 cours), non?

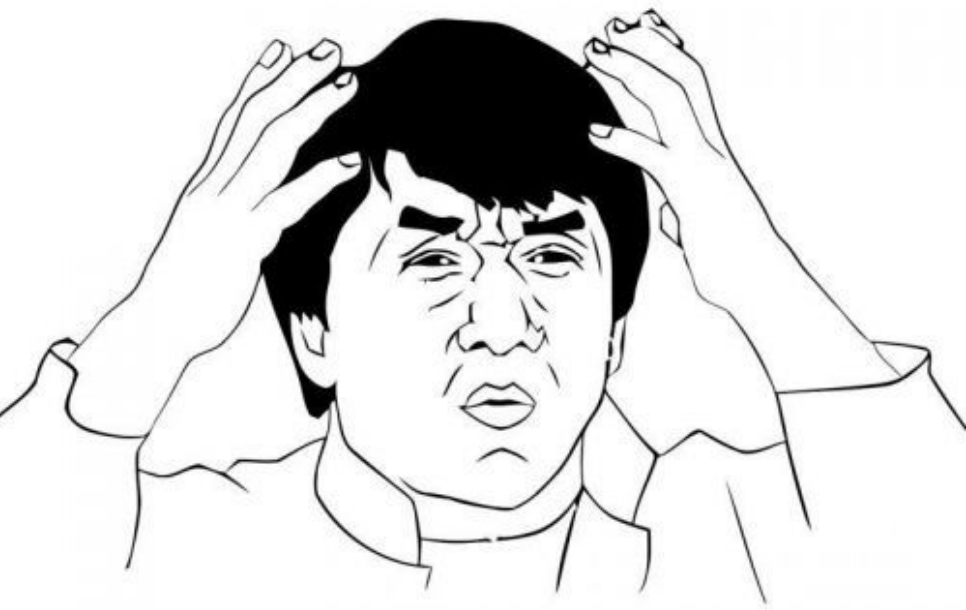
PARADOXE DE SIMPSON

100 étudiants (50 IMPAIRS et 50 DATA)
choisissent chacun entre 2 cours : **Fouille** et
Systèmes. Pourcentages de validation [fictif!] :

Fouilles de données		Systèmes avancés	
IMPAIRS	DATA	IMPAIRS	DATA
90%	84.5%	70%	60%

Réussite **globale**:
(sur l'ensemble
des 2 cours)

IMPAIRS	DATA
74%	82%



EXPLICATION

Les DATA préfèrent le cours où ils réussissent bien. Leur faible réussite en Systèmes ne compte presque pas: ils sont peu. Mot-clé: **Répartition**.

Fouilles de données		Systèmes avancés	
IMPAIRS	DATA	IMPAIRS	DATA
90% (9/10)	84.5% (38/45)	70% (28/40)	60% (3/5)

Réussite **globale**:
(sur l'ensemble
des 2 cours)

IMPAIRS	DATA
74%	82%

PARADOXE DES ANNIVERSAIRES

Quelle est la **probabilité** que deux personnes parmi vous aient la même date d'anniversaire ?

$N / 365$?

PARADOXE DES ANNIVERSAIRES

Quelle est la **probabilité** que deux personnes parmi vous aient la même date d'anniversaire ?

> 50% si $N \geq 23$

> 80% si $N \geq 35$

> 90% si $N \geq 41$

> 95% si $N \geq 47$

> 99% si $N \geq 58$

> 99.9% $N \geq 69$ (et dans ce cas, 1 chance sur 3 d'avoir 3 qui ont la même)

EXEMPLE AVEC 50 PERSONNES

Probabilité que 50 personnes aient des
anniversaires **tous différents**:

$$\begin{aligned} p &= \frac{365}{365} \times \frac{364}{365} \times \cdots \times \frac{365 - 50 + 1}{365} \\ &= \frac{365 \times 364 \times \cdots \times 316}{365^{50}} \\ &= 0.0296 \end{aligned}$$

→ Il y a donc **97%** de chances qu'au moins 2 personnes aient le même anniversaire.

INTUITION

Il y a $N \times (N-1) / 2 = O(N^2)$ **paires** de personnes dans un groupe de N personnes.

On peut imaginer que la probabilité d'avoir les anniversaires tous différents est similaire à la probabilité que **chaque paire**, de manière indépendante (en première approximation), ait un anniversaire différent: $P \approx (364/365)^{(N^2/2)}$

$$P \approx (1 - 1/365)^{(N^2/2)}$$

Quand $N^2/2 \approx 365$, i.e. $N=27$, ça ressemble à $\lim_{N \rightarrow \infty} (1 - 1/N)^N = e^{-1} \approx 0.368$ (vraie: 0.373..)

TEST: AU POKER (TEXAS HOLD'EM)

Sur un jeu de 52 cartes, quelle est la probabilité que j'aie une paire d'As ?

A - $2/52$

B - $1/52$

C - $1/221$

D - $3/1225$

E - $1/2652$

TEST: AU POKER (TEXAS HOLD'EM)

Sachant que j'ai As/Roi dans la main, quelle est la probabilité que mon adversaire ait une paire d'As?

A - $2/52$

B - $1/52$

C - $1/221$

D - $3/1225$

E - $1/2652$

CE QU'ON EN CONCLUT

Avoir de l'**information** change (parfois *drastiquement*) la donne!

PARADOXE DES TROIS PORTES (MONTY HALL)



Jeu télé. Le candidat est devant 3 portes. 1 porte cache une voiture, les 2 autres **rien**.

- Il choisit une porte.
- L'animateur ouvre **systématiquement** 1 des 2 autres, **qui ne cache pas la voiture**.
- L'animateur propose au candidat de changer son choix de porte s'il le désire

Le candidat doit-il changer de porte ?

PARADOXE DES TROIS PORTES (MONTY HALL)



Jeu télé. Le candidat est devant 3 portes. 1 porte cache une voiture, les 2 autres **rien**.

- Il choisit une porte.
- L'animateur ouvre **systématiquement** 1 des 2 autres, **qui ne cache pas la voiture**.
- L'animateur propose au candidat de changer son choix de porte s'il le désire

Le candidat doit-il changer de porte ? OUI

EXPLICATION

Regardons les probabilités:.....

- Au départ, le candidat a 1 chance sur 3 de choisir la bonne porte. Donc $\frac{2}{3}$ que la bonne porte soit parmi les 2 autres.
- Lorsque le présentateur en ouvre une autre qui ne contient pas la voiture, il apporte une information : il élimine une des 2 portes restantes. Donc l'autre a 2 chances sur 3 de contenir la voiture.

Le candidat **doit donc changer de porte**, passant sa probabilité de gagner de $\frac{1}{3}$ à $\frac{2}{3}$.

ESPRIT STATISTIQUEMENT CRITIQUE

Les **absurdités** et **manipulations** à base de chiffres sont partout : politique, presse, et même recherche.

Les chiffres ont, pour la plupart des gens, une autorité intrinsèque ("c'est scientifique").

Les conclusions sont le fruit de **l'interprétation**. Il faut dissocier **résultats** et **conclusion**.

On ne fait rien dire du tout aux chiffres, mais on peut les utiliser pour faire passer ses opinions.

Un des objectifs de ce cours : **ne plus se faire manipuler !**

Préparation au TD(s) à venir

Les TDs sont en Python. **Sondage!**

- A) Je suis un assez bon codeur en python
- B) J'ai déjà codé en python, je m'en souviens
- C) Jamais
- D) C'est quoi un python?



Python a une **excellente** documentation en ligne. Exemple: “[python sort](#)” sur [Google](#)

STATISTICS 101

[TODOs](#)

NOUS PARLERONS DE

Variables aléatoires

Loi de probabilité

Echantillons

Significativité et intervalles de confiance

Tests statistiques

EXEMPLE 1

On cherche à répondre à ce type de questions:

Sur 1000 personnes (500 hommes, 500 femmes), on observe que les femmes gagnent en moyenne 2000 euros et les hommes 2300.

Peut-on en déduire que les femmes gagnent moins que les hommes?

Même question avec 10 femmes, 10 hommes.
Idem avec 100 femmes, 100 hommes.

EXEMPLE 2

On parle de **A/B** testing.

Sur une page web, on peut choisir 2 modèles pour le même bouton: un **rouge** et un **bleu**.

- Sur 1000 personnes l'ayant vu, 23 ont cliqué sur le **rouge**.
- Sur 500 personnes l'ayant vu, 17 ont cliqué sur le **bleu**.

Peut-on dire : “le bouton bleu marche mieux?”

EXEMPLE 3

Peut-on affirmer qu'il y a plus de garçons
que de filles en informatique ?

LA SIGNIFICATIVITÉ

On se pose donc la question de la **fiabilité**
d'un chiffre et de la validité des
conclusions.

DÉFINITION

Un résultat est **statistiquement significatif** à 5% (ou 95% selon les conventions de langage) si la probabilité qu'on l'observe par "chance" (eg. par hasard) est $\leq 5\%$.

Exemple: "La pièce est truquée: plus souvent face que pile" avec (3 face, 1 pile), avec (10 face, 0 pile), avec (103 face, 57 piles).

DÉFINITION

Un résultat est **statistiquement significatif** à 5% (ou 95% selon les conventions de langage) si la probabilité qu'on l'observe par "chance" (eg. par hasard) est $\leq 5\%$.

Exemple: "La pièce est truquée: plus souvent face que pile" avec (3 face, 1 pile), avec (10 face, 0 pile), avec (103 face, 57 piles).

Intuitivement, cela a un rapport avec:

- La taille de l'échantillon
- L'hétérogénéité de l'échantillon.

VARIABLES ALÉATOIRES

Définition (pas très mathématique)

Une variable aléatoire (v.a.) est une application définissant l'ensemble des résultats possibles pour une expérience donnée.

Une variable aléatoire est donc la chose que l'on **observe** et qui nous intéresse.

LES VARIABLES DISCRÈTES

Définition

Une variable aléatoire (v.a.) est dite **discrète** lorsqu'elle peut prendre un nombre **fini** (ou **infini dénombrable**) de valeurs.

LES VARIABLES DISCRÈTES

Exemples:

- X variable aléatoire liée à l'expérience "pile ou face" : les valeurs possibles de X sont 0 et 1.
- X variable aléatoire liée à l'expérience "réponse à un sondage": $X \in \{\text{"pas satisfait"}, \text{"plutôt satisfait"}, \text{"très satisfait"}\}$.
- X variable aléatoire liée à l'expérience "nombre de personnes dans un supermarché":
 $X \in \mathbb{N}$ c'est-à-dire $X \in \{0, 1, 2, \dots, \}$

LES VARIABLES CONTINUES

Définition

Une variable aléatoire (v.a.) est dite **continue** lorsqu'elle peut prendre un nombre **infini** **indénombrable** de valeurs.

LES VARIABLES CONTINUES

Exemples:

- X est la variable aléatoire liée à l'expérience "heure d'arrivée d'un ami": les valeurs possibles de X se trouvent dans $[15h, 16h]$.
- X représente la variable aléatoire liée à l'expérience "mon poids": les valeurs possibles de X se trouvent entre 0 et $+\infty$.
- X représente la variable aléatoire liée à l'expérience "Note moyenne en M2": les valeurs possibles de X sont entre 0 et 20.

LOI DE PROBABILITÉ

Définition (pas mathématique du tout!)

Une loi (ou distribution) de probabilité définit le comportement d'une variable aléatoire.

Autrement dit, la loi définit les probabilités qu'une variable aléatoire prenne une valeur ou un ensemble de valeurs. **Attention : elle ne dit pas quelle valeur va être prise !**

EX 1: LOI UNIFORME DISCRETE

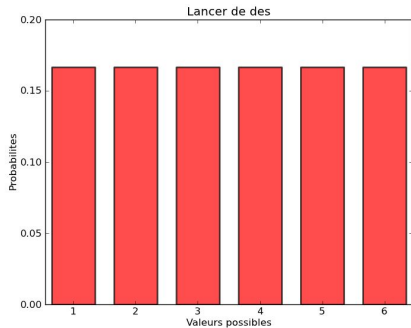
X représente la variable aléatoire liée à
l'expérience "**lancer de dé**": X prend ses
valeurs entre 1 et 6.

$$\begin{aligned} P(X=1) &= P(X=2) = P(X=3) \\ &= P(X=4) = P(X=5) = P(X=6) \\ &= \mathbf{1/6}. \text{ La somme vaut } \mathbf{1}. \end{aligned}$$

Loi est dite **uniforme**

Car toutes les probabilités sont les mêmes.

On parle aussi d'**équiprobabilité**.



EX 2: LOI DE BERNOULLI

X représente la variable aléatoire liée à
l'expérience "**lancer de tartine**": X prend les
valeurs 0 ou 1.

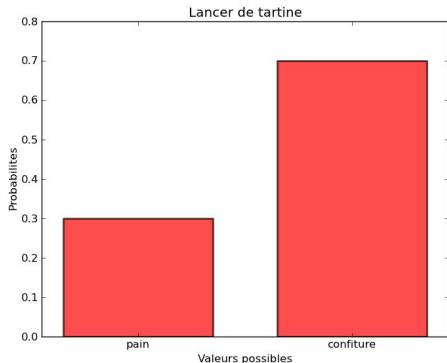
$$P(X=\text{côté pain}) = 0.3$$

$$P(X=\text{côté confiture}) = 0.7$$

La somme vaut **1**.

Le **paramètre** de la loi

est **p**. Ici, $p = 0.7$. Dans le cas d'un pile ou face
avec une pièce équilibrée, $p = \frac{1}{2} = 0.5$.



EX 3: LOI UNIFORME CONTINUE

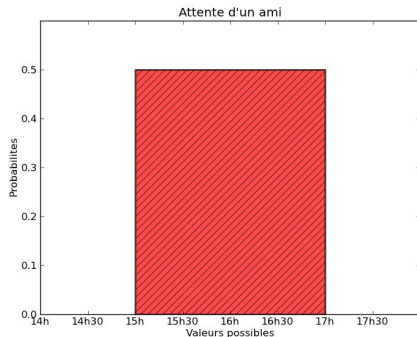
X représente la variable aléatoire liée à
l'expérience "**heure d'arrivée d'un ami**": X prend
ses valeurs entre 15h et 17h.

$$P(X = 15:00) = 0$$

$$P(X = 15:30) = 0$$

$$P(X \in [15:10, 15:40]) \\ = 1/4 = 0.25$$

$$P(X \geq 16:15) = 3/8 = 0.375$$



La "somme" des probabilités vaut 1. Ici, il s'agit
de l' **aire sous la courbe**.

EX 3: LOI UNIFORME CONTINUE

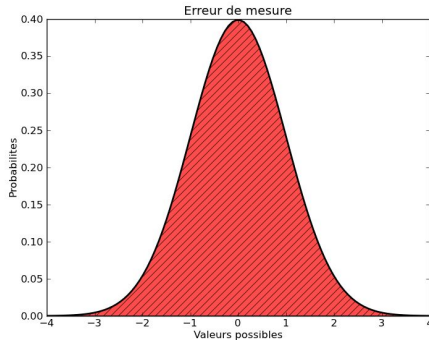
X représente la variable aléatoire liée à
l'expérience "**erreur de mesure**": X prend ses
valeurs entre $-\infty$ et $+\infty$.

$$P(X = 0) = 0$$

$$P(X = 3) = 0$$

$$P(X \in [-2, 2]) \approx 0.95$$

$$P(X > 2.33) = 0.01$$



"Somme" des proba = 1 (**aire sous la courbe**).

Ici, les deux **paramètres** de la loi sont : moyenne
= 0, écart-type = 1 (Nous y reviendrons)

MOMENTS D'UNE V.A.

L'**espérance** correspond à la moyenne *théorique* d'une variable aléatoire. On la note **E(X)**.

$$E(X) = \sum_{i \in \text{valeurs possibles}} P(X = i) * i$$

La **variance**, notée **V(X)** représente la *dispersion* d'une variable aléatoire autour de sa moyenne.

Variance grande \Rightarrow variable dispersée, imprévisible.

L'**écart-type** est la racine carrée de la variance.

$$\text{Variance}(X) = E((X - E(X))^2) = \sum_{i \in \text{valeurs possibles}} P(X = i) * (i - E(X))^2$$

MOMENTS D'UNE V.A.

(exemple au tableau)

PROBABILITÉS VS STATISTIQUES

Les **probabilités** permettent de représenter un état théorique des choses.

Les **statistiques** utilisent des données (souvent une grande quantité), expérimentales (ou simulées, dans le cadre d'un cours), et permettent de comprendre, d'**estimer les valeurs théoriques**.

ÉCHANTILLON STATISTIQUE

(X_1, \dots, X_n) est appelé **échantillon** si les variables aléatoires X_1, \dots, X_n sont **indépendantes** et suivent la même loi. On dit qu'elles sont **i.i.d.** : **indépendantes et identiquement distribuées**.

Exemples:

- $X_1 \dots X_n$ sont n lancers de pile ou face. Ils sont indépendants, et ont tous la même loi de probabilité.
- $X_1 \dots X_n$ représentent le QI de n personnes. Ces personnes sont indépendantes et leur QI suit la même distribution (assimilée à loi normale).

fabien.viger@gmail.com