

CS4445, CS4803, BCB4003  
Data Mining & Knowledge  
Discovery



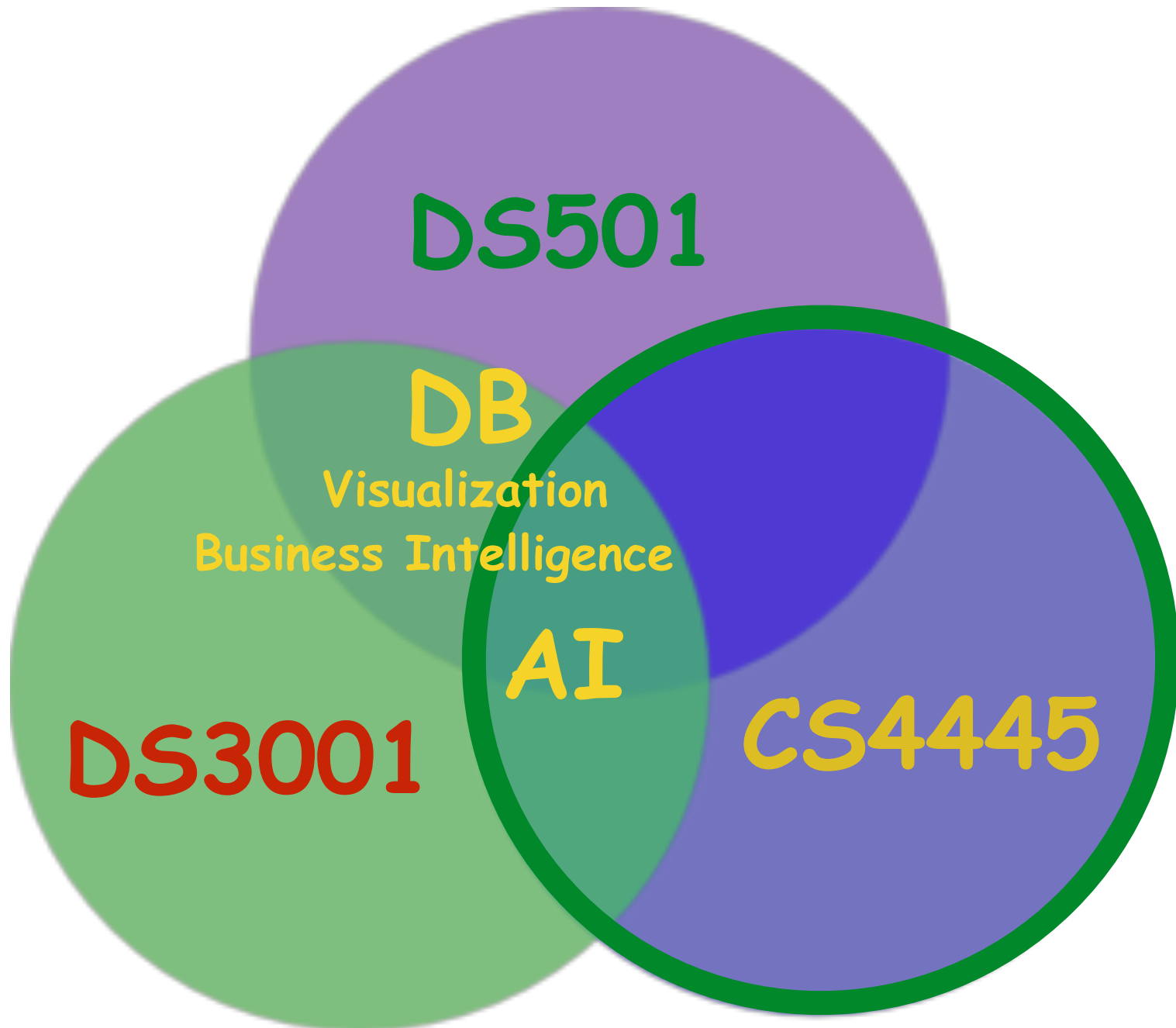
# Goal of this course

## A bird's eye view of Data Mining





# Overlaps in Course Material



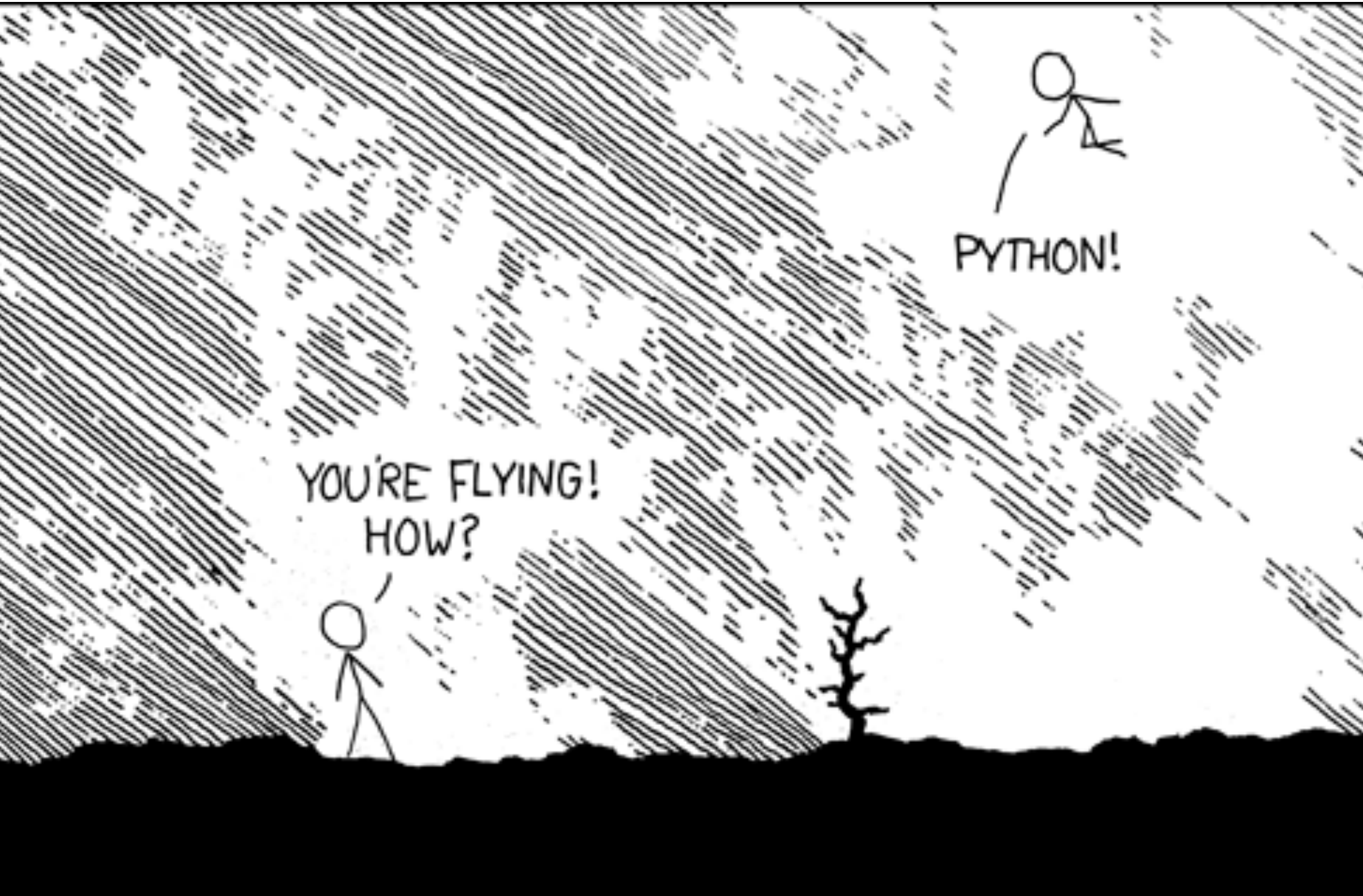
# Gear up in the era of Big Data



## Math + Tools

- Be comfortable with both

# Official Language for CS4445



IP[y]: IPython  
Interactive Computing



# Scores

- Homework Assignments: 60%
- Quizzes: 40%

# Homework ( $5 \times 12\% = 60\%$ )

- 6 Homework Assignments
- the lowest homework score will be dropped
- Need Python 3
- Homework 1 is released TODAY



# Quizzes ( $10 \times 4\% = 40\%$ )

- 12 Quizzes
- 2 lowest quiz scores will be dropped
- No exam

# Course Calendar

Due Dates: homework, quiz

## Course Summary:

Date	Details	
Thu Oct 26, 2017	 Lecture 1	12pm to 1:50pm
Mon Oct 30, 2017	 Lecture 2	12pm to 1:50pm
	 Quiz 1	12pm to 12:10pm
Thu Nov 2, 2017	 HW 1 due at 12pm	12am
	 Lecture 3	12pm to 1:50pm
	 Quiz2	12pm to 12:10pm
Mon Nov 6, 2017	 Lecture 4	12pm to 1:50pm
	 Quiz 3	12pm to 12:10pm
Tue Nov 7, 2017	 Grading of HW1 due	12am
Thu Nov 9, 2017	 HW2 due at 12pm	12am
	 Lecture 5	12pm to 1:50pm
	 Quiz 4	12pm to 12:10pm
Mon Nov 13, 2017	 Lecture 6	12pm to 1:50pm
	 Quiz 5	12pm to 12:10pm



# What is Data Mining?

Data Mining:

Why should we care?



# Data in Silicon Valley

<http://ed.ted.com/on/MYodUMYe>



# Data in Wall Street





# Data in Biomedical Research

consensus	AGAC.tctT...ca.A...gctTATA.aGAG...gAATTT.aAGGA.ACAC...ggaa.....gca...cgcCAGcgtAca.....tac.gtgAg...AT.cgAGtaccGgAT.gACGta.AAATT.AcCt.Tagaag.a.....T.t...Aaga.gtctt	
Rc_hemC	AGACATCTTTCCAAACTC--GCTTATAGGAG--GTACTTGAAGGAAACAC--AGAAAG--ACTTGCCACC--CGACGTACAAAAACA--ATC-GCACA--GATGCAAGTACC--GGATTGACGTACAAATT--ACCCCTAGAAGTAGA--GTTTGGGAAGAGTGTCT	147
Rc_RP502h	AGACATCTTTGCCAAATTT--GCTTATAGAGAG--GAATTTAAAGGAAACAC--GGAACG--CAACA--CCACAGCGTACACTTTAG--TAC-GTGAG--GATGCGAGTACC--GGATTGACGTACAAATT--ACCTCTAGAAGTAGA--GTTTGGCAAGATGTCT	144
Rc_RR045h	AGACATCTTTGCCAAATTC--GCTTATAGAGAG--GAATTTAAAGGAGACAC--GAAACG--CAGCA--CGCAGCGTACACTTTAG--TAC-GTGAG--GATGCGAGTATCCGATTCGACGTATAAATT--ACCTCTAGAAGTAGA--GTTTGGCAAGATGTCT	144
Rc_ubiG	AGACATCTTTTAAACCA--GCTTATAGAGAG--GGAATCTAAAGGAGACAC--GGAACA--CAGCA--CGCAGCGTACACTTAG--TAC-GTGAG--GATGCGAGTACC--GGATTGACGTACAAATT--ACCTCTAGAAGCGAA--GTTTGGGAAGATGTCT	144
Rc_ubiH	AGACATCTTTGCCAAATTC--ACTTATAGAGAG--GAATTTGTAGGAAACAT--GCAGCA--CAGCA--CGCAGCGTACACTTTGG--TAC-GTGAG--GATGCGAGCAGCAGATTGACGCACAAATT--ACCTCTAGAAGCAGA--GTTTGGAAAAGATGTCT	144
Rc_RP507h	AGACATCTTTGCCAAATTC--GCTTATAGAGAG--GAATTTGTAGGAGACAC--GAAACG--CAGCA--CGCAGCGTACAAAGCG--TAC-GTGAG--GATGCGAGTATCCGATTCGACGTACAAATT--ACCTCTAGAAGTAGAAGATTGGGAAGATGTCT	147
Rc_RP167h	AGACATCTTTCCAAACCA--GCTTATAGAGAG--GAATTTAAAGGAGACAT--AGAATG--TAGCA--CGCAGCGTACAAAGAA--TAC-TTGAG--GATACGAGTACC--GGATTTCGACGTCTAAATT--ACCTCTAGAAGCGAA--GTTTGGAAAGATGTCT	144
Rc_mesJ	AGACATCTTTCCAAACCA--GCTTATAGAGAG--GAATGTACAGGAGACAC--GGAACG--CAGCA--CTACAGCGTATATGGACA--TAC-TTGAG--GATGTGAGTACC--GGATTCGACGTCTAAATT--ACCTCTAGAAGCGAA--GTTTGTAAAGATATCT	144
Rc_era	AGACATCTTTCCAAACCA--GTTTATAGAGAG--GAATTTAAAGGAGACAC--GGAACG--CAGCA--CGCAGCGTACACTTTAG--TAC-GTGAG--GATGCGAGTACC--GGAAACCGCGCTCTAAATT--ATCTCTAGAAGCGAA--GTTTGGAAAGATGTCT	144
Rc_orf1	AGACATCTTTGCCAAATTC--GCTTATAGAGAG--GAATTTAAAGGAGACACGGAAGCACT--TGCCG--CGCAG--TAC--TAC--GCGAG--GATGCGAGTACTGGATTCGACGTCTAAATT--ACCTCTAGAAGTGCA--GTTTGGGAAGATGTCT	135
Rc_gltX	AGACATCTTTTAAACCA--GTTTATAGAGAG--GAATTTAAAGAGATAC--GGAACA--CAGCA--CGCAG--ACA--T--AC--GTGAG--GATGCGAGTACC--GGATTCGACGTCTAAATT--ACCTCTAGAAGTGCA--GTTTGGGAAGATGTCT	135
Rc_mv1N	ACACTTCTTCGGCATCTC--CCTTATAGGAG--GAATTTGTAGGAAACAC--GGAACG--CAGCA--CGCAGCGTATATAGACA--TAC-GTGAG--GATGCGAGTACC--GGATTCGACGTACAAATT--ACCCCTAGCAGTAGA--GTTGCCGAAGAACTT	144
Rc_pcnB	CGACTTCT--TGCATAACGTAGCTAATAAAGAG--AAATTTGAAGGAAACAC--GGA--CGCAGCACCGCAGCGTACACTTTAG--TAC-GTGAG--ATGCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGCAAG--TTATGC--AAGAAGTGC	144
Rc_rlpA	AGATCTCT--TGCATAACCAAACTAATAAAGAG--GAATTTGAAGGAGACAC--GGA--CGCAGAACTCGCAATGTACATAAAGC--TAC-GTGAG--ATGCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGCAAG--TTATGC--AAGAAGTGC	144
Rc_orf2	AGACCTCT--TGCAGAAATGAAGCTAATAAAGAG--AAATTTGAAGGAAACAC--GGA--CGCAGAACTCGCAATGTACATAAAGC--TAC-GTGAG--ATGCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGCAAG--TTATGC--AAGAAGTGC	143
Rc_kdtA	AGACTGCT--TACAGAAATGAAGCTAATAAAGAG--GAATTTGAAGGAGACAC--GGA--CGCAGAACTCGCAATGTACATAAAGC--TAC-GTGAG--ATGCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGCAAG--TTATGC--AAGAAGTGC	144
Rc_orf3	AGACTGCT--TGCAGAACTCGCTAATAAAGAG--GAATTTGAAGGAGACAC--TT--CACTGCGCAACCAACGTATACATAA--TAC-GTGAG--ATGAGGGTTAGGATCAACGTCTAAATT--ACCTTTAGAAGCAAG--TGATGC--ACGAAGGCT	144
Rc_gmk	AGACTTCT--TACATAACCTATCTTATAAAGAG--GAATTTGAAGGAGACAT--GAAA--CGCAGCACCGCAGCTACAAAAGC--TAA-GTGAG--ATGCGAGTATTCGATTCGACGTATACG--GGATAT--TCAGAAG--TTGGGA--GTCAATAA	143
Rc_rpe20	AGAGTGCTTCAGAAATGC--GCTTATAGAGAG--GAATTTGAAGGAGACAT--GGAACG--TAGGA--CGCAGCG--ACGCTCTAAT--ACCTCTAGAAGTAGA--ACTTCGAGAGCACTCT	108
Rc_rpe21	AGACTTCC--TGTCACAACTTAGCTAATAAAGAG--GAATTTGTAGGAGACAC--GGA--CACAGCACCGCAGCGTACGCTTTGG--TA--CGTGAGG--ATACGAGTAA--CGGATTCGACGTACAAATT--ACCTCTAGAAGTAGA--TTATGC--AGAAAGTCT	144
Rc_rpe22	AGACTTCT--TGACTT--GCTAATAAAGAG--AAATTTGAAGGAGACAC--GGA--CGCAGAACTCGCAATGTACATAAAGC--TAC-GTGAG--ATGCGAGTACC--GGATTCGACGTACAAATT--ACCTTTATAGAAG--TTATGC--AAGAAGTCT	136
Rc_rpe23	AAACTTTTCTAAATACGCTAGCTAATAAAGAG--AAATTTGAAGGAGACAC--GGA--CGCAGAACTCGCAATGTACATAAAGC--TAC-GTGAG--ATGCGAGTACC--GGATTCGACGTACAAATT--ACCTTTATAGAAG--TTATGC--AAGAAGTCT	146
Rc_rpe24	AGACTTAT--TGCATAATATAGCTAATAAAGAG--GAATTTGAAGGAGACAC--GGA--CGCAGAACTCGCAATGTACATAAAGC--TAC-GTGAG--ATGCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGTAGA--TTATGC--AAGAAGTCT	148
Rc_rpe25	AGACTTCT--TGCATAACCTATCTTATAAAGAG--AGACTTGAAGGAGACAC--GGAAGTACTTGCACCGCGCAGCATACAGC--TAACTGAGG--ATGCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGTAGA--TTATGC--AAGAAGTCT	150
Rc_rpe26	AGACTTGT--TGCTGAACCTATCTTATAAAGAG--GAATTTGAAGGAGACAC--GGAAC--GCAGCACCGCAGCGTACACTTAGT--AC--GTGAGG--ATGCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGTAGA--TTATGC--AAGAAGTCT	143
Rc_rpe27	AGACTTCT--TGCATAACCTATCTTATAAAGAG--GAATTTGAAGGAGACAC--GGA--CATAGAACCAGCGCTAGCAACAAAT--CTAT--GTAAGG--ATTCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGTAGA--TTATGC--AAGAAGTCT	106
Rc_rpe28	AGACTTAT--TGCATAACCTATCTTATAAAGAG--GAATTTGAAGGAGACAC--GGA--CATAGAACCAGCGCTAGCAACAAAT--CTAT--GTAAGG--ATTCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGTAGA--TTATGC--AAGAAGTCT	142
Rc_rpe29	AGGCTTTT--TGCATAACCTATCTTATAAAGAG--GAATTTGAAGGAGACAC--GGA--CATAGAACCAGCGCTAGCAACAAAT--CTAT--GTAAGG--ATTCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGTAGA--TTATGC--AAGAAGTCT	144
Rc_rpe30	AGGCTTTT--TGCATAACCTATCTTATAAAGAG--GAATTTGAAGGAGACAC--GGA--CATAGAACCAGCGCTAGCAACAAAT--CTAT--GTAAGG--ATTCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGTAGA--TTATGC--AAGAAGTCT	144
Rc_rpe31	AGACTTCT--TGCATAACCTATCTTATAAAGAG--GAATTTGAAGGAGACAC--GGA--CATAGAACCAGCGCTAGCAACAAAT--CTAT--GTAAGG--ATTCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGTAGA--TTATGC--AAGAAGTCT	144
Rc_rpe32	AGACTTCT--TGCATAACCTATCTTATAAAGAG--GAATTTGAAGGAGACAC--GGA--CATAGAACCAGCGCTAGCAACAAAT--CTAT--GTAAGG--ATTCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGTAGA--TTATGC--AAGAAGTCT	144
Rc_rpe33	AGACTTCT--TGCATAACCTATCTTATAAAGAG--GAATTTGAAGGAGACAC--GGA--CATAGAACCAGCGCTAGCAACAAAT--CTAT--GTAAGG--ATTCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGTAGA--TTATGC--AAGAAGTCT	144
Rc_rpe34	AGACTTCT--TGCATAACCTATCTTATAAAGAG--GAATTTGAAGGAGACAC--GGA--CATAGAACCAGCGCTAGCAACAAAT--CTAT--GTAAGG--ATTCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGTAGA--TTATGC--AAGAAGTCT	144
Rc_rpe35	AGACTTCT--TGCATAACCTATCTTATAAAGAG--GAATTTGAAGGAGACAC--GGA--CATAGAACCAGCGCTAGCAACAAAT--CTAT--GTAAGG--ATTCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGTAGA--TTATGC--AAGAAGTCT	144
Rc_rpe36	AGACTTCT--TGCATAACCTATCTTATAAAGAG--GAATTTGAAGGAGACAC--GGA--CATAGAACCAGCGCTAGCAACAAAT--CTAT--GTAAGG--ATTCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGTAGA--TTATGC--AAGAAGTCT	144
Rc_rpe37	AGACTTCT--TGCATAACCTATCTTATAAAGAG--GAATTTGAAGGAGACAC--GGA--CATAGAACCAGCGCTAGCAACAAAT--CTAT--GTAAGG--ATTCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGTAGA--TTATGC--AAGAAGTCT	144
Rc_rpe38	AGACTTCT--TGCATAACCTATCTTATAAAGAG--GAATTTGAAGGAGACAC--GGA--CATAGAACCAGCGCTAGCAACAAAT--CTAT--GTAAGG--ATTCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGTAGA--TTATGC--AAGAAGTCT	144
Rc_rpe39	AGACTTCT--TGCATAACCTATCTTATAAAGAG--GAATTTGAAGGAGACAC--GGA--CATAGAACCAGCGCTAGCAACAAAT--CTAT--GTAAGG--ATTCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGTAGA--TTATGC--AAGAAGTCT	144
Rc_rpe40	AGACTTCT--TGCATAACCTATCTTATAAAGAG--GAATTTGAAGGAGACAC--GGA--CATAGAACCAGCGCTAGCAACAAAT--CTAT--GTAAGG--ATTCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGTAGA--TTATGC--AAGAAGTCT	144
Rc_rpe41	AGACTTCT--TGCATAACCTATCTTATAAAGAG--GAATTTGAAGGAGACAC--GGA--CATAGAACCAGCGCTAGCAACAAAT--CTAT--GTAAGG--ATTCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGTAGA--TTATGC--AAGAAGTCT	144
Rc_rpe42	AGACTTCT--TGCATAACCTATCTTATAAAGAG--GAATTTGAAGGAGACAC--GGA--CATAGAACCAGCGCTAGCAACAAAT--CTAT--GTAAGG--ATTCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGTAGA--TTATGC--AAGAAGTCT	144
Rc_rpe43	AGACTTCT--TGCATAACCTATCTTATAAAGAG--GAATTTGAAGGAGACAC--GGA--CATAGAACCAGCGCTAGCAACAAAT--CTAT--GTAAGG--ATTCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGTAGA--TTATGC--AAGAAGTCT	144
Rc_rpe44	AGACTTCT--TGCATAACCTATCTTATAAAGAG--GAATTTGAAGGAGACAC--GGA--CATAGAACCAGCGCTAGCAACAAAT--CTAT--GTAAGG--ATTCGAGTACC--GGATTCGACGTACAAATT--ACCTTTAGAAGTAGA--TTATGC--AAGAAGTCT	144
Rp_RP474	AGATATTTTTCTAAACCA--GCTTATAGAGAG--GCATTTAAAGCGAATAC--AATACG--TGCCA--CCACAGCGTATAAAG--TAT-TTAAT--GATCCTAGTCTAGGCT--CAACGTA--TCCTCTAGAAGTAGC--GTTTGGGAAGAGTGTCT	138
Rp_coxB	AGATATTTGGTCCAAACAA--TCTTATAAAGGTTAAAGTAGATCAAGAG--ATAACA--CCA--CA--GAATACACAAAG--TAT-GTAAT--AGCAGAGTCTAGGTTTCGACGTATACAT--ACCTCTAAAAATGGA--ATTATGGAAGATATAT	141
Rp_RP688	AAACTTTT--TGTTAATCACTAATTAAGAGAT--GGATTTAAAGGAGATAC--GGA--TGCAACGCTGCTGTACATAAAGC--TAT-GTAATG--ATTCAGAGTACCAAAATCAACGCTCTAAATT--AGCTTTAATAATAG--TTATGC--ACAAATCT	144
Rp_kdtA	AGATTTC--TACAAATGTCTGTTAATAAAGAG--GGATTTACAGACACAA--TGC--ATAGAA--ATTCAGAGTCTA--ATGTTAAT--ATGTTAAGAAAG--TTATGC--ACAGAGTCT	129
Rp_alr	AGAGGTTT--CGCTTAATCTGTCTTATAGAG--GAATTTGAAGGAGACAC--AGAGCTCTTACTACC--ACAGCGGTATATAA--AAT-GTCCAG--ATGTTAGTGTAGATGACTCTCAAAAT--ACCTTTAAAAAGGAG--TTATGC--TGTAAGTCT	147
Rp_RP545	AGATATCTCTCTAAACAA--ACTGATAGAGAG--GAATTTTACTACTCTG--AGAATG--GTA--TATAGGAATCTTAAGCA--TAA-GA--GTACTAATCAACTT--ATAAAT--ACCTCTAAAAAGGAGCA--ATTTGTATATGTCT	132
Rp_lyc	ATACCTCTTGTAAATCC--ACGTATATGGA--GAATCTGCACTAATAC--GAAACA--CAGCACTAAA--ATAGACA--TAC-CTGAG--GATGCTAGTGGAT--CAACATATAAAT--ACCATAGAAATAGC--ATTACAAAAACAGGTAT	135
Rp_pyrG	GCACATATTTCCAAATGT--ACTTATAGTGAA--GCATTTG--AATG--GAGGCT--T--CCACTGTATATCAAAAT--TAC-ATGAG--GATTCAAAATACTGCTCTACAGCAAAAT--GCAGATAGAAACAA--TTTATGAATATGTCT	135
Rp_RP404	AGACATTTTTCTAAACT--ACATATAGCAAG--AAATTT--GAAT--GCAATCTAGTAAGCA--TTT-G--CAACCACTGT--ACATT--ACCATTAATAATAG--ATTCTAGAAATGCACT	108
Rp_rpe55	AGACTTCT--TGCATAAGATAGCTAATAAGGT--AAATTTGTAATAAACC--AGCA--C--ATAGTAGATACGAATA--AC--GTGAGA--ATACGAGTACTGAATTGACATACAAATG--ACCTTTAGAAAGAGG--CTATAG--AGGAAGGCT	135
Rh_polA	AGGCCTCTTTCCAAACTC--GCTTATAGAGAG--GAATTTAAAGGAGACAC--GGAACG--CAGCA--CGCAGCGTACAAAAGCG--TAC-GTGAG--GATGCGAGTACC--GGATTCGACGTCTAAATT--ACCTCTAGAAGTAGA--GTTTGGGAAGAGGCGCT	144
Rf_polA	AGACATCTTTCTAAACCA--GCTTATAGAGAG--GAATTTAAAGGAGACACGGAAGCACT--TGCCA--CGCAGCGTACAAAAGCG--TAC-GTGAG--GATGCGAGTACC--GGATTCGACGTCTAAATT--ACCTCTAGAAGTAGA--GTTTGGGAAGAGTGTCT	147
1.....10.....20.....30.....40.....50.....60.....70.....80.....90.....100.....110.....120.....130.....140.....150.....160.....		

<https://www.youtube.com/watch?v=FzcTgrxMzZk>



# BBC documentary

- The Age of Big Data
  - Crime Prevention
  - ...

